This document contains Supplementary Notes 1 to 3 published as part of the following *Nature Methods* paper:

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

Stephen S. Gisselbrecht[1], Luis A. Barrera[1,2,3], Martin Porsch[1,4], Anton Aboukhalil[1,5], Preston W. Estep 3[rd 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[1,2,3,10]

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115.
[2]Harvard-MIT Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, MA 02115.
[3]Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, MA 02138.
[4]Institute of Computer Science, Martin Luther University of Halle-Wittenberg, 06099 Halle, Germany.
[5]Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139.
[6]TeloMe, Inc., Waltham, MA 02451.
[7]Bioengineering Department, Polytech Nice Sophia, University of Nice Sophia Antipolis, 06903, France.
[8]Laboratory of Developmental Systems Biology, Genetics and Developmental Biology Center, National Heart Lung and Blood Institute, National Institutes of Health, Bethesda, MD 20892.
[9]Systems Biology Graduate Program, Harvard University, Cambridge, MA 02138.
[10]Department of Pathology; Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115.

Correspondence should be addressed to M.L.B. (mlbulyk@receptor.med.harvard.edu).

Supplementary Note 1

>pEFS-Dest.fa -
GAACCAGATCGGCCGCGTCGACGATGTAGGTCACGGTCTCGAAGCCGCGGTGCGGGTGCCAGGGCGTGCCCTTGGG
CTCCCCGGGCGCGTACTCCACCTCACCCATCTGGTCCATCATGATGAACGGGTCGAGGTGGCGGTAGTTGATCCCG
GCGAACGCGCGGCGCACCGGGAAGCCCTCGCCCTCGAAACCGCTGGGCGCGGTGGTCACGGTGAGCACGGGACGTG
CGACGGCGTCGGCGGGTGCGGATACGCGGGGCAGCGTCAGCGGGTTCTCGACGGTCACGGCGGGCATGTCGACTCG
AGGCCGGTCTCCCTATAGTGAGTCGTATTAATTTCGATAAGCCAGGTTAACCTGCATTAATGAATCGGCCAACGCG
CGGGGAGAGGCGGTTTGCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCT
GCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAAC
ATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCC
CCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGC
GTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTC
CCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGC
TGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCC
GGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCT
ACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGC
CAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGT
TTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCT
CAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAA
ATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCAG
TGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGATAACTACG
ATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTAT
CAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTAT
TAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGC
ATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGAT
CCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTT
ATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGT
GAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATA
ATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGAT
CTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACC
AGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAA
TACTCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGA
ATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGACGTCTTCGCTAGT
TAAGATACATTGATGAGTTTGGACAAACCACAACTAGAATGCAGTGAAAAAAATGCTTTATTTGTGAAATTTGTGA
TGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTAACAACAACAATTGCATTCATTTTATGTTT
CAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTAAAACCTCTACAAATGTGGTATGGCTGATTATGATC
TAGAGTCGCGGCCGCTTTACTTGTACAAGTTAGCGTCTTCGTTCACTGCTGCGACTTCGGCTCCGTTGACGGCGTT
TTCGTTCACTGCTGCGACTTCGGCTGCGATTGCGGCTCCTAGAGCGCTGGCGATGCCTTCTGTGCCTGCTCTTGTA
CAGCTCGTCCATGCCGAGAGTGATCCCGGCGGCGGTCACGAACTCCAGCAGGACCATGTGATCGCGCTTCTCGTTG
GGGTCTTTGCTCAGGGCGGACTGGGTGCTCAGGTAGTGGTTGTCGGGCAGCAGCACGGGGCCGTCGCCGATGGGGG
TGTTCTGCTGGTAGTGGTCGGCGAGCTGCACGCTGCCGTCCTCGATGTTGTGGCGGATCTTGAAGTTCACCTTGAT
GCCGTTCTTCTGCTTGTCGGCCATGATATAGACGTTGTGGCTGTTGTAGTTGTACTCCAGCTTGTGCCCCAGGATG
TTGCCGTCCTCCTTGAAGTCGATGCCCTTCAGCTCGATGCGGTTCACCAGGGTGTCGCCCTCGAACTTCACCTCGG
CGCGGGTCTTGTAGTTGCCGTCGTCCTTGAAGAAGATGGTGCGCTCCTGGACGTAGCCTTCGGGCATGGCGGACTT
GAAGAAGTCGTGCTGCTTCATGTGGTCGGGGTAGCGGCTGAAGCACTGCACGCCGTAGGTCAGGGTGGTCACGAGG
GTGGGCCAGGGCACGGGCAGCTTGCCGGTGGTGCAGATGAACTTCAGGGTCAGCTTGCCGTAGGTGGCATCGCCCT
CGCCCTCGCCGGACACGCTGAACTTGTGGCCGTTTACGTCGCCGTCCAGCTCGACCAGGATGGGCACCACCCCGGT
GAACAGCTCCTCGCCCTTGCTCACCATGGTGGCGACCGGTGGATCGTTTAAACAGGCCTCTCGACTCTAGCGCGTA
CCCTAGAGGATCAGCTTGGCTGCAGATTGTTTAGCTTGTTCAGCTGCGCTTGTTTATTTGCTTAGCTTTCGCTTAG
CGACGTGTTCACTTTGCTTGTTTGAATTGAATTGTCGCTCCGTAGACGAAGCGCCTCTATTTATACTCCGGCGCTC
CTCGAGGGATCATCACCACTTTGTACAAGAAAGCTGAACGAGAAACGTAAAATGATATAAATATCAATATATTAAA
TTAGATTTTGCATAAAAAACAGACTACATAATACTGTAAAACACAACATATCCAGTCACTATGGTCGACCTGCAGA
CTGGCTGTGTATAAGGGAGCCTGACATTTATATTCCCCAGAACATCAGGTTAATGGCGTTTTTGATGTCATTTTCG

CGGTGGCTGAGATCAGCCACTTCTTCCCCGATAACGGAGACCGGCACACTGGCCATATCGGTGGTCATCATGCGCC
AGCTTTCATCCCCGATATGCGCCACCGGGTAAAGTTCACGGGAGACTTTATCTGACAGCAGACGTGCACTGGCCAG
GGGGATCACCATCCGTCGCCCGGGCGTGTCAATAATATCACTCTGTACATCCACAAACAGACGATAACGGCTCTCT
CTTTTATAGGTGTAAACCTTAAACTGCATTTCACCAGTCCCTGTTCTCGTCAGCAAAAGAGCCGTTCATTTCAATA
AACCGGGCGACCTCAGCCATCCCTTCCTGATTTTCCGCTTTCCAGCGTTCGGCACGCAGACGACGGGCTTCATTCT
GCATGGTTGTGCTTACCAGACCGGAGATATTGACATCATATATGCCTTGAGCAACTGATAGCTGTCGCTGTCAACT
GTCACTGTAATACGCTGCTTCATAGCACACCTCTTTTTGACATACTTCGGGTATACATATCAGTATATATTCTTAT
ACCGCAAAAATCAGCGCGCAAATACGCATACTGTTATCTGGCTTTTAGTAAGCCGGATCCTCTAGATTACGCCCCG
CCCTGCCACTCATCGCAGTACTGTTGTAATTCATTAAGCATTCTGCCGACATGGAAGCCATCACAGACGGCATGAT
GAACCTGAATCGCCAGCGGCATCAGCACCTTGTCGCCTTGCGTATAATATTTGCCCATGGTGAAAACGGGGGCGAA
GAAGTTGTCCATATTGGCCACGTTTAAATCAAAACTGGTGAAACTCACCCAGGGATTGGCTGAGACGAAAAACATA
TTCTCAATAAACCCTTTAGGGAAATAGGCCAGGTTTTCACCGTAACACGCCACATCTTGCGAATATATGTGTAGAA
ACTGCCGGAAATCGTCGTGGTATTCACTCCAGAGCGATGAAAACGTTTCAGTTTGCTCATGGAAAACGGTGTAACA
AGGGTGAACACTATCCCATATCACCAGCTCACCGTCTTTCATTGCCATACGGAATTCCGGATGAGCATTCATCAGG
CGGGCAAGAATGTGAATAAAGGCCGGATAAAACTTGTGCTTATTTTTCTTTACGGTCTTTAAAAAGGCCGTAATAT
CCAGCTGAACGGTCTGGTTATAGGTACATTGAGCAACTGACTGAAATGCCTCAAAATGTTCTTTACGATGCCATTG
GGATATATCAACGGTGGTATATCCAGTGATTTTTTTCTCCATTTTAGCTTCCTTAGCTCCTGAAAATCTCGACGGA
TCCTAACTCAAAATCCACACATTATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGCGGCCGCCA
TAGTGACTGGATATGTTGTGTTTTACAGTATTATGTAGTCTGTTTTTTATGCAAAATCTAATTTAATATATTGATA
TTTATATCATTTTACGTTTCTCGTTCAGCTTTTTTGTACAAACTTGTGATCTGTTGCCGAGCACAATTGATCGGCT
AAATGGTATGGCAAGAAAAGGTATGCAATATAATAATCTTTTATTGGGTATGCAACGAAAATTTGTTTCGTCAACG
TATGCAATATTTTTTATTAAAAGAGGGTATGCAATGTATTTTATTAAAAACGGGTATGCAATATAATAATCTTTTA
TTGGGTATGCAACGAAAATTTGTTTCGTCAAAGTATGCAATATTTTTTATTAAAAGAGGGTATGCAATGTATTTTA
TTAAAAACGGGTATGCAATAAAAAATTATTTGGTTTCTCTAAAAAGTATGCAGCACTTATTTTTTGATAAGGTATG
CAACAAAATTTTACTTTGCCGAAAATATGCAATGTTTTTGCGAATAAATTCAACGCACACTTATTACGTGATGCAG
CCAAGCTCGAAGGGGAAATACTTGTATTCTATAGGTCATATCTTGTTTTTATTGGCACAAATATAATTACATTAGC
TTTTTGAGGGGGCAATAAACAGTAAACACGATGGTAATAATGGTAAAAAAAAAAACAAGCAGTTATTTCGGATATA
TGTCGGCTACTCCTTGCGTCGGGCCCGAAGTCTTAGAGCCAGATATGCGAGCACCCGGAAGCTCACGATGAGAATG
GCCAGACCCACGTAGTCCAGCGGCAGATCGGCGGCGGAGAAGTTAAGCGTCTCCAGGATGACCTTGCCCGAACTGG
GGCACGTGGTGTTCGACGATGTGCAGCTAATTTCGCCCGGCTCCACGTCCGCCCATTGGTTAATCAGCAGACCCTC
GTTGGCGTAACGGAACCATGAGAGGTACGACAACCATTTGAGGTATACTGGCACCGAGCCCGAGTTCAAGAAGAAG
CCGCCAAAGAGCAGGAATGGTATGATAACCGGCGGACCCACAGACAGCGCCATCGAGGTCGAGGAGCTGGCGCAGG
ATATTAGATATCCGAAGGACGTTGACACATTGGCCACCAGAGTGACCAGCGCCAGGCAGTTGAAGAAGTGCAGCAC
TCCGGCCCGCAGTCCGATCATCGGATAGGCAATCGCCGTGAAGACCAGTGGCACTGTGAGAAAAAGCGGTAATTCG
GCAATCGTTTTGCCCAGAAAGTATGTGTCACAGCGATAAAGTCGACTTCGGGCCTCCCTCATAAAAACTGGCAGCT
CTGAGGTGAACACCTAAATCGAATCGATTCATTAGAAAGTTAGTAAATTATTAATATGCAAATGTATTCTAAACAA
GACTTACATTTATCGTGGCAAAGACGTTTTGAAAGGTCATGTTGGTCAGGAAGAGGAAGATGGCTCCGTTGATATT
CATCACGCCCACTTGCGTGAGTTGTTGGCCCAAAAAGATGAGGCCAATCAAGATGGCAACCATCTGCAAATTAAAA
TGTTACTCGCATCTCATTAATATTCATATCTTCAACATGTTCGCGAGTTAAATGAAATTTATTTATTTTCTGCAAA
ACTATAAACTATACATCTCATTGAAAAAAACTAAGAAGGGTGTGGAATCAGGCAATTCTAACTAAAATCTAGCGAA
TTTGTTTCCAAGAATTGTAAGCGTTATATCATTTGTTTCCACTGGAACCACTCACCGTTGTCTGAATAAGTCGCAC
TTTTACGAGGAGTGGTTCCTTGAGCACCGACAGCCAGGATCGCCACAGGACCGCCCGGAACTGCATGAACCAGGTG
GCCTTGTAGGTGTACCCATTCTCCGGCTGCTCCAGTGGCTTCTCCAAATTTTTGGTGGCCAACAACTGCTCCATAT
CCCGGGCTACTTTGCTAATAGCAAAATTGTCGCATATCTTGGCGATCCGATCACGGGACTCGATCTCCCGTCCGGG
CACAACGGCCAACACCTGTACGTAAAAGTCCGCCGGATTGTAGTTGGTAGGACACTGGGCACCCACGCTGGATAGG
AGTTGAGATGTAATGTAATGCTAGATACCCTTAATAAACACATCGAACTCACTAGGAAAAGAAGTCGACGGCTTCG
CTGGGAGTGCCCAAGAAAGCTACCCTGCCCTCGGCCATCAGAAGGATCTTGTCAAAGAGCTCAAACAGCTCGGAAG
ACGGCTGATGAATGGTCAGGATGACGGTCTTGCCCTTCTGCGACAGCTTCTTCAGCACCTGGACGACGCTGTGGGC
GGTAAATGAGTCCAGTCCGGAGGTGGGCTCATCGCAGATCAGAAGCGGCGGATCGGTTAGTGCCTCGGAGGCGAAT
GCCAGACGCTTCCTTTCTCCGCCGGACAGACCTTTCACCCTGCCGGGCACACCGATGATCGTGTGCTGACATTTGC
TGAGCGAAAGCTCCTGGATCACCTGATCCACGCGGGCCACTCGCTGCCGATAGGTCAGATGTCGTGGCATCCGCAC
CATGGCCTGGAAAATCAGGTGTTCCCTGGCCGTTAGGGAGCCGATAAAGAGGTCATCCTGCTGGACATAGGCGCAC
CTGGCCTGCATCTCCTTGGCGTCCACAGGTTGGCCATTGAGCAGTCGCATCCCGGATGGCGATACTTGGATGCCCT
GCGGCGATCGAAAGGCAAGGGCATTCAGCAGGGTCGTCTTTCCGGCACCGGAACTGCCCATCACGGCCAAAAGTTC

GCCCGGATAGGCCACGCCGCAAACTGAGTTTCAAATTGGTAATTGGACCCTTTATTAAGATTTCACACAGATCAGC
CGACTGCGAATAGAAACTCACCGTTCTTGAGCAAATGTTTCCTGGGCGCCGGTATGTGTCGCTCGTTGCAGAATAG
TCCGCGTGTCCGGTTGACCAGCTGCCGCCATCCGGAGCCCGGCTGATTGACCGCCCCAAAGATGTCCATATTGTGC
CAGGCATAGGTGAGGTTCTCGGCTAGTTGGCCGCTCCCTGAACCGGAGTCCTCCGGCGGACTGGGTGGCAGGAGCG
TGCCGTAGTTTTTGGCCTGCCCGAAGCCCTGGTTAATGCAGCTCTGCGAAGCGTCCGCTGTCACCCTGCAATGATA
GGGGATCTCAAATATCAACTACAAGCGTTATGCTCATCTAACCCCGAACAAAACGAAGTATCCTACGAAGTAGGTT
TATACTTTTATTTATTTTTTGTGCATCTAGGATCAGCTTAAAATATCTGGTTGTTATATTTTTTGTAAAAAAGAAT
GTAGTCGAAAATGAATGCCTTTAGATGTCTTGATCATGATATGATCTTAAAAATTGTCTTATATAGCGAGCACAGC
TACCAGAATAATCTGTTTCGTGTCACTATTTGTTTGTGCGATTGCGGTTTGGGATTTTTGTGGGTCGCAGTTCTCA
CGCCGCAGACAATTTGATGTTGCAATCGCAGTTCCTATAGATCAAGTGAACTTAAGATGTATGCACATGTACTACT
CACATTGTTCAGATGCTCGGCAGATGGGTGTTTGCTGCCTCCGCGAATTAATAGCTCCTGATCCTCTTGGCCCATT
GCCGGGATTTTTCACACTTTCCCCTGCTTACCCACCCAAAACCAATCACCACCCCAATCACTCAAAAAACAAACAA
AAATAAGAAGCGAGAGGAGTTTTGGCACAGCACTTTGTGTTTAATTGATGGCGTAAACCGCTTGGAGCTTCGTCAC
GAAACCGCTGACAAAGTGCAACTGAAGGCGGACATTGACGCTAGGTAACGCTACAAACGGTGGCGAAAGAGATAGC
GGACGCAGCGGCGAAAGAGACGGCGATATTTCTGTGGACAGAGAAGGAGGCAAACAGCCCGGACGTCTAAGAAACC
ATTATTATCATGACATTAACCTATAAAAATAGGCGTATCACGAGGCCCTTTCGTCTCGCGCGTTTCGGTGATGACG
GTGAAAACCTCTGACACATGCAGCTCCCGGAGACGGTCACAGCTTGTCTGTAAGCGGATGCCGGGAGCAGACAAGC
CCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTG
AGAGTGCACCATATGGACATATTGTCGTTAGAACGCGGCTACAATTAATACATAACCTTATGTATCATACACATAC
GATTTAGGTGACACTATA

Supplementary Note 2

This document contains discussion and technical details of the Online Methods.

Table of contents:

Nature Methods: doi:10.1038/nmeth.2558

A. Cloning and preparation of candidate CRMs (cCRMs) library

Selection of candidate CRMs (cCRMs).

We designed our cCRM library to comprise ChIP-CRMs [1] bound by a somatic mesoderm (SM) TF or by the transcriptional coactivator CBP [2,3], DNase I hypersensitive sites (DHSs) [4] near or not near a mesodermally expressed gene, regions tiling the noncoding DNA around 4 genes, and a number of control regions. Specifically, we selected 491 cCRMs, which fall into 5 categories: (1) 288 ChIP-CRMs [1] bound by a somatic mesoderm (SM) TF (Twi, Tin, or Mef2); (2) 38 regions bound by the transcriptional coactivator CBP [2,3] and located next to mesodermally expressed genes lacking adjacent ChIP-CRMs (note: 45 of the 288 selected ChIP-CRMs overlap CBP-bound regions); (3) 58 regions containing DHSs [4] located next to mesodermally expressed genes lacking adjacent ChIP-CRMs; (4) 41 cCRMs predicted by the PhylCRM algorithm on the basis of dense clusters of evolutionarily conserved TF binding site motif occurrences [5] adjacent to genes expressed in either somatic or cardiac mesoderm; and (5) 64 genomic windows tiling the noncoding DNA surrounding 10 genes (*cib, mud, nau, jumu, rgr, slou, CG3303, CG11202, CG13794, CG15319*) that we selected on the basis of biological interest and/or compact noncoding DNA. For each tiled gene, the intergenic sequences upstream and downstream of the gene and any introns longer than 1 kb were divided into overlapping segments of ~1 kb (*e.g.*, nt 1–1000, 501–1500, 1001–2000, etc.). Introns 200–1,000 bp long were included with additional coding DNA to extend them to a size of ~1 kb. All cCRMs were chosen to be 900-1,100 bp long to avoid potential PCR bias in later steps after cell sorting.

473 of the 491 cCRMs (96.5%) were successfully cloned into the eFS vector, as determined by Illumina sequencing of the resulting pooled, cCRM plasmid library (see below). Of the 473 injected cCRMs (see below), we detected (false discovery rate < 5×10^{-5}; see Section G below) 189 cCRMs in the *twi*:CD2 'input' samples, 213 in the *Mef2-I-*E_{D5}:CD2 input samples, and 411 in the *duf*:CD2 input samples; in total, 431 cCRMs were detected in at least one input sample. The remaining 42 cCRMs may have been missed due to underrepresentation in the injected cCRM plasmid library, stochastic lack of integration at the phiC31 genomic integration site, or stochastic underrepresentation among germline stem cells that led to the collected embryos.

Nature Methods: doi:10.1038/nmeth.2558

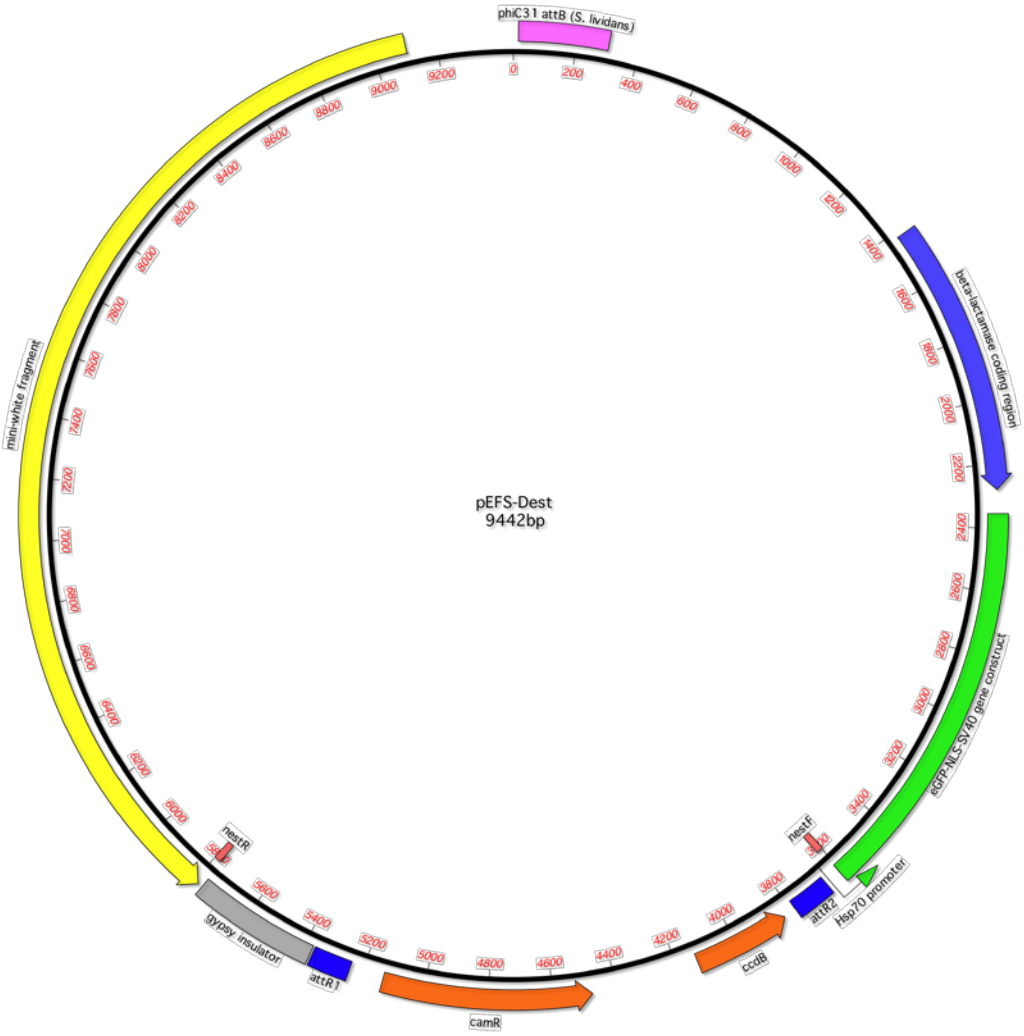
PCR amplification of cCRMs.

A two-step PCR amplification was used to include Gateway attB sites, and specific forward and reverse sequencing primers. We performed these PCRs in five 96-well plates, with a sixth plate to reattempt initially poor or failed PCRs; >95% (474/494) of the selected cCRMs were successfully amplified by the two-step procedure. First round PCR was performed in 25 μL reactions with Phusion enzyme (2X Master Mix with HF buffer) (New England Biolabs) using 50 ng *Drosophila melanogaster* OreR genomic DNA as template and 2.5 pmol of each PCR primer. Primer pairs were designed using the MacVector (v. 11.1) (MacVector, Inc.) primer design function, starting with default parameters and relaxing them stepwise until a suitable primer pair was found. All cCRM-specific primers are provided in Supplementary Table 1. Common PCR primer sequences were engineered onto the 5' ends of each forward (SEQ1: CAAGACGAGGCTATGCTCTAGC) and reverse (SEQ2: TAGAGTTGGCTTGCCATAGACC) PCR primer. Cycling conditions were as follows: 1 x 30" @ 98°C; 30 x [5" @ 98°C, 10" @ 60°C, 30" @ 72°C]; 1 x 5' @ 72°C; hold @ 4°C. Second round PCR was performed under identical conditions, using 1 μL of a 1:500 dilution of first round PCR (in water) as template and a common primer pair (attB1-SEQ1: GGGGACAAGTTTGTACAAAAAAGCAGGCTCAAGACGAGGCTATGCTCTAGC and attB2-SEQ2: GGGGACCACTTTGTACAAGAAAGCTGGGTAGCTAGAGTTGGCTTGCCATAGACC).

Design of reporter vector for eFS.

We created the vector for enhancer-FACS-Seq, pEFS-Dest, by blunt-end cloning the 1.8 kb HindIII-SpeI fragment of pPelican [6] (containing a nuclear-localized GFP reporter construct with a *gypsy* insulator element upstream of the MCS and minimal promoter) into pWattB, then replacing the Multiple Cloning Site with a cassette providing attR1 and attR2 sites for Gateway cloning. The sequence of vector pEFS-Dest is provided in Supplementary Note 1, and an image of the vector map of pEFS-Dest is provided below *(next page)*.

Nature Methods: doi:10.1038/nmeth.2558



A Gateway LR cloning reaction between pEFS-Dest and a donor plasmid containing a cCRM flanked by attL1 and attL2 sites results in the insertion of the cCRM proximal to the minimal promoter driving expression of the reporter gene (see below). pWattB was made by inserting (1) the ϕC31 attB site from *Streptomyces lividans* [7] and (2) the mini-white gene into the small cloning vector pSP73 (Promega). The ϕC31 attB site supports highly efficient, site-specific integration of plasmids derived from this vector into the genome of flies that carry an attP transgene when ϕC31 integrase is expressed; the site-specificity of the target means that a library of reporter plasmids can be injected and only a single plasmid per haploid genome will integrate.

Nature Methods: doi:10.1038/nmeth.2558

Additionally, site-specific integration permits the selection of an ideal insertion site conferring negligible position effects on reporter gene expression [8]. The mini-white gene permits selection of transformant adult flies by eye color. The reporter cassette comprises the Hsp70 minimal promoter driving expression of a nuclear localization signal-tagged EGFP gene with an SV40 polyadenylation sequence [6]. We have tested pEFS by inserting the previously characterized mesodermal enhancers for *even skipped* [9], *Ndg* [10], and the *Lbl*-expressing FC [10]. In each case, GFP expression recapitulated the characterized activity of the enhancer.

Purification, normalization and cloning of cCRM library into eFS reporter vector.

Aliquots of all PCR reactions were run on ethidium bromide agarose gels along with High DNA Mass Ladder (Invitrogen) and photographed and roughly quantified using Quantity One software (BioRad). Equal masses of each 900-1,100 bp band were calculated from this quantification, pooled, precipitated, and gel-purified, then cloned as a pool using Gateway BP Clonase II (Invitrogen) into pDONR221 (Invitrogen) according to the manufacturer's protocols. Cloning reactions were transformed into commercial competent *E. coli* Top10 cells (Invitrogen) and plated on LB+kanamycin agar, yielding ~30,000 colonies. These colonies were scraped up and a plasmid pool purified from them, from which the combined inserts were cloned using Gateway LR Clonase II (Invitrogen) into pEFS-Dest. Transformed cells were plated on LB+ampicillin agar, yielding ~30,000 colonies, from which the final library plasmid pool was prepared for embryo injection.

In a preliminary test of PCR insert retrieval and Illumina library preparation (described in detail in Section D below), libraries prepared from dilutions of this plasmid pool (here, diluted to a final estimated concentration of 50,000 plasmid molecules in the plasmid sample prior to PCR) were sequenced. According to the sequencing data and our read mapping algorithm and criteria described below (Section F), 473 of the 474 successfully amplified cCRMs were successfully cloned into the eFS vector and were detected in 4 out of 4 Illumina sequencing libraries.

Nature Methods: doi:10.1038/nmeth.2558

B. Generation of CD2 reporter vectors

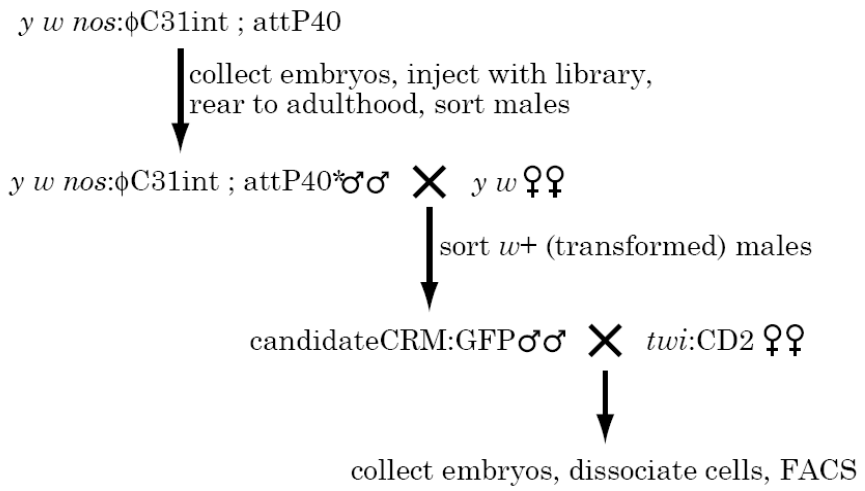
A minimal promoter was fused to rat CD2 and subsequently cloned into P-element transformation vectors by PCR-amplifying the TATA box from pUAST-NTAP with PCR primer pair ATGGCTAGCTAGCGAGCGCCGGAGTATAA and GGTGTCAATTCCCAATTCCCTATTCAG, and CD2 from twi-CD2 [11] with primer pair CTGAATAGGGAATTGGGAATTGACACCATGAGATGTAAATTCCTAGGGAGTTTCTTT and TAGGCTAGCTTAATTAGGGGGTGGCAACGA. These two PCR products served as templates for an assembly PCR reaction using the primer pair ATGGCTAGCTAGCGAGCGCCGGAGTATAA and TAGGCTAGCTTAATTAGGGGGTGGCAACGA. This PCR product was subcloned into pCR (Invitrogen), sequence-verified, digested with NheI and cloned into XbaI-digested pETWN [9], resulting in our CD2 vector pETWCD2.

The enhancer region for *Mef2-I-E_{D5}* was synthesized in vitro (Integrated DNA Technologies, Coralville, IA, USA) and subcloned into pETWCD2, while the enhancer region for *duf* [12] was PCR-amplified from genomic DNA isolated from a BAC clone (Children's Hospital and Research Center at Oakland, BACPAC Resources, Oakland, CA, USA) using the following primer pair (AATTCCCTTCCACATTGGTCCTCTC, GATCCAAGTGTGATATCGTGTTTGGCC), subcloned into pETWCD2 and sequence-verified.

C. Fly embryo injections and husbandry

We begin with a description of our strategy; experimental details follow. The resulting pooled plasmid cCRM clone library was injected into embryos. We are using a strain of flies containing an attP site that gives very low basal expression but very high induced expression in the embryonic mesoderm when an inducible promoter construct is integrated there [8]. This strain additionally expresses a nuclear-localized ϕC31 integrase under the control of the *nanos* promoter, which causes mRNA to be produced during oogenesis and deposited in the egg before fertilization. Injection of a plasmid containing an attB site into fertilized eggs of this line at the posterior pole, where the primordia of the germ line will form, results in the transmissible integration of the plasmid DNA into the genome. The recombination of an attP and an attB site,

Nature Methods: doi:10.1038/nmeth.2558



Genetic crosses for enhancer-FACS-Seq. In this example, *twi*:CD2 is used to identify all candidate CRMs with expression in the embryonic mesoderm.

mediated by the integrase enzyme, produces an attL and an attR site (distinct from and not cross-reacting with those used in the Gateway system, which are used by a different bacteriophage) which are not themselves substrates for the integrase; thus, integration is non-reversible and one integration event destroys the attP site used, preventing any further events at that genomic locus.

In principle, then, each primordial germ cell in an injected embryo can integrate two molecules of plasmid, one into each copy of the haploid genome. When gametes arising from these cells contribute to the progeny of these flies, each will receive only a single integrated transgene. As each embryo produces ~20–30 primordial germ cells before mitotic expansion, the gametes produced by an injected fly will carry a population of different library insertions. Progeny of injected flies that receive an integrated transgene can be recognized by the inclusion of the mini-white gene in the reporter vector, which causes a partial reversion of the adult eye color toward wildtype. *w+* (transformant) males are collected and crossed to virgin females of a strain homozygous for a transgene which drives expression of rat CD2 in the cell type of interest, producing the desired population of embryos.

To investigate whether competition for stem cell niches in the gonads of injected flies could limit the number of independent insertion events transmitted by each parent [13], we examined the transformant progeny of individual injected flies. For this investigation, we injected an earlier

Nature Methods: doi:10.1038/nmeth.2558

library of 275 cCRMs into embryos and reared survivors to adulthood. Males (to avoid confounding effects from mated females) were crossed to *y w* virgin females individually in vials. Up to 24 *w+* male progeny from each of several vials were isolated and processed for single-fly, insertion site targeted PCR, and the resulting PCR products sequenced to identify which transformed progeny of an injected fly had common inserts and which contained unique inserts. At one extreme, 24 flies from a cross with 47% transformant progeny had identical inserts; at the other, 10 flies from a vial with 22% transformant progeny had among them 7 different inserts. The average number of insertions per injected parent over all sequenced transformants was 3.5.

For our full-scale eFS experiments, after purification and normalization, our pooled plasmid cCRM clone library was diluted to 0.75 mg/mL in standard injection buffer (5 mM KCl, 0.1 mM PO_4, pH 7.8) and injected posteriorly into syncytial embryos carrying the *nos*-ϕC31\int.NLS transgene [14] on the X chromosome and the attP40 insertion [8] on the 2nd chromosome. Injections were performed either in-house (used for the *twi*:CD2 and *Mef2-I-*E_{D5}:CD2 sorts) or by Rainbow Transgenic Flies, Inc. (Camarillo, CA) (used for the *duf*:CD2 sorts). Surviving males were crossed to excess *y w* virgin females. Transformant male progeny were selected on the basis of eye color. In pilot experiments, roughly 10% of all progeny of injected flies were transformant; based on the uniformity of the resulting eye color (and variation expected due to differences in insertion site), we estimate that the vast majority (~99%) of insertions were into the intended target location.

We collected several thousand transformant males and, separately, several thousand virgin females from each tissue-specific CD2 line of interest (see main text). These flies were combined in population cages ~36 hours before the beginning of embryo collections. Additional control cages containing only wild type (*y w*) flies, or wild type males with CD2 virgin females, were generated to provide control cells for assessment of FACS parameters, as discussed in the next Section.

Population cages were collected from twice "pre-lays" to minimize the presence of older embryos due to retention of fertilized eggs by females, then two collections of 2 hours (for

Nature Methods: doi:10.1038/nmeth.2558

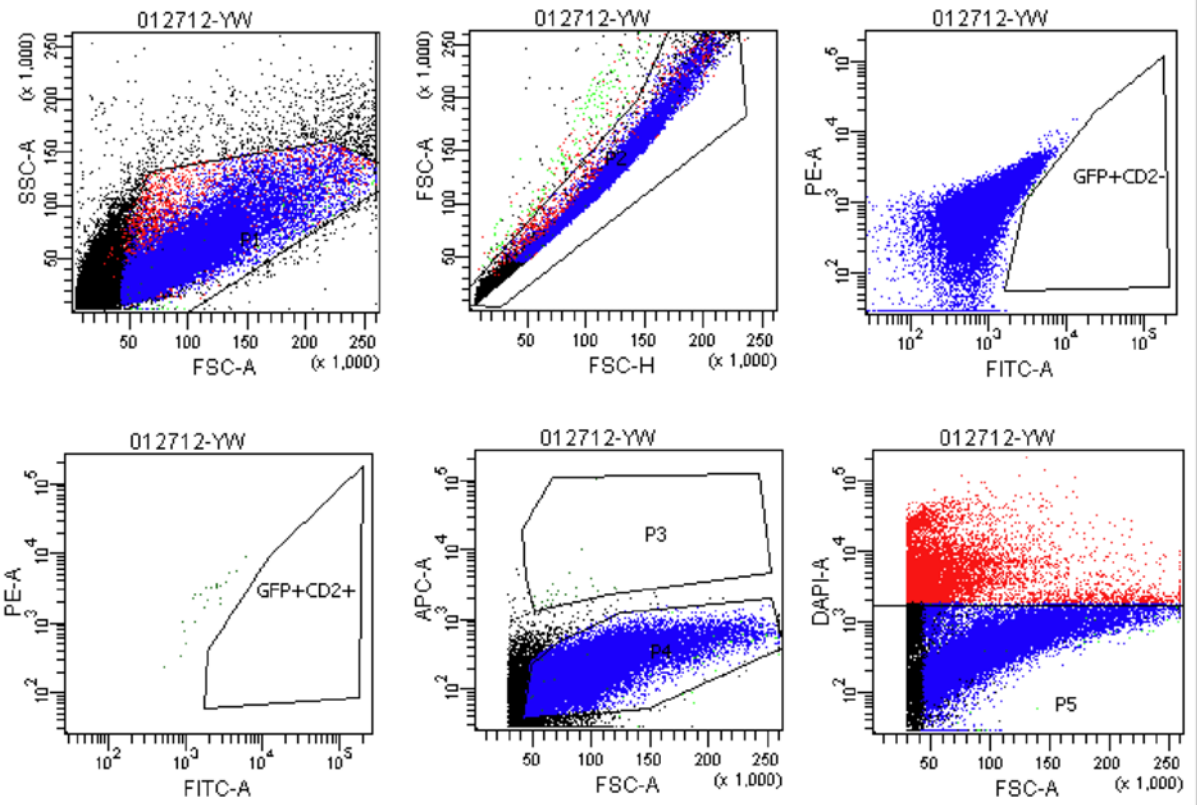
twi:CD2 sorting) or 2.5 hours (for *Mef2-I-E_{D5}*:CD2 and *duf*:CD2 sorting) were performed. These plates were aged 10–11 hours at 18°C, after which embryos were collected and dechorionated, and single cell suspensions were prepared for FACS.

D. Fluorescence activated cell sorting (FACS)

We previously described and used extensively the isolation of single cells for FACS from live *Drosophila* embryos at stage 11, when the SM FC and FCM populations are being specified [15]. We have modified the existing protocol by incorporating a step in which dissociated cells are resuspended in *Drosophila* cell culture medium and incubated on ice (10 minutes with moderate shaking) with a commercially available Alexa647-conjugated anti(rat CD2) antibody (AbD-Serotec, cat. #MCA154A647). The antibody is diluted 1:400 in Schneider medium + 8% FBS and preadsorbed by incubation with cells prepared from non-CD2-expressing embryos, then after removal of cells by centrifugation, samples were filtered with Nytex mesh and supplemented with 2 μg/mL DAPI to permit the detection and removal of dead cells and yolk granules. After brief washing (i.e. two cycles of centrifugation and resuspension in Schneider + 8% FBS), the cells are analyzed and separated by FACS.

FACS output is shown here *(next page)* for cells isolated from wild type (non-CD2-expressing, non-GFP-expressing) embryos ("YW"):

Nature Methods: doi:10.1038/nmeth.2558



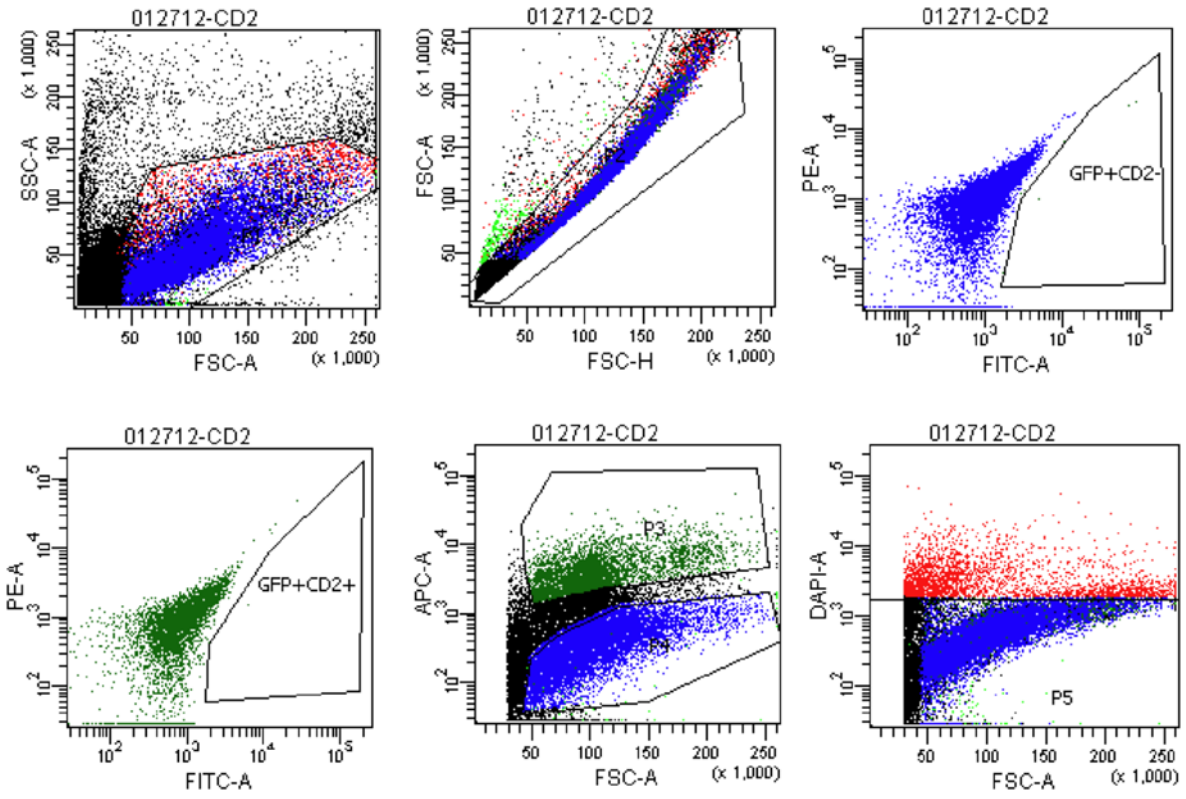
Tube: YW

Population	#Events	%Parent	%Total
All Events	100,000	####	100.0
P1	39,027	39.0	39.0
P5	31,318	80.2	31.3
P2	31,145	99.4	31.1
P3	23	0.1	0.0
GFP+CD2+	0	0.0	0.0
P4	23,407	75.2	23.4
GFP+CD2-	10	0.0	0.0

The populations P1 (chosen from the side scatter [SSC-A] vs. forward scatter [FSC-A] plot, upper left, which primarily selects cells over debris and yolk granules), P2 (chosen from the forward scatter amplitude [FSC-A] vs. height [FSC-H] plot, upper middle, which largely removes doublets), and P5 (chosen from the DAPI signal vs. forward scatter plot, lower right) are standard FACS filters to isolate viable single cells. The lower middle plot (APC signal, representing detection of the anti-CD2 antibody, vs. forward scatter) shows the robust detection of CD2-negative cells (P4) and the low background detected in the CD2-positive gate (P3).

Nature Methods: doi:10.1038/nmeth.2558

Identical analysis of cells prepared in parallel from embryos that do express CD2, here under the control of the *twi* promoter to identify all mesodermal cells [16], is shown here:



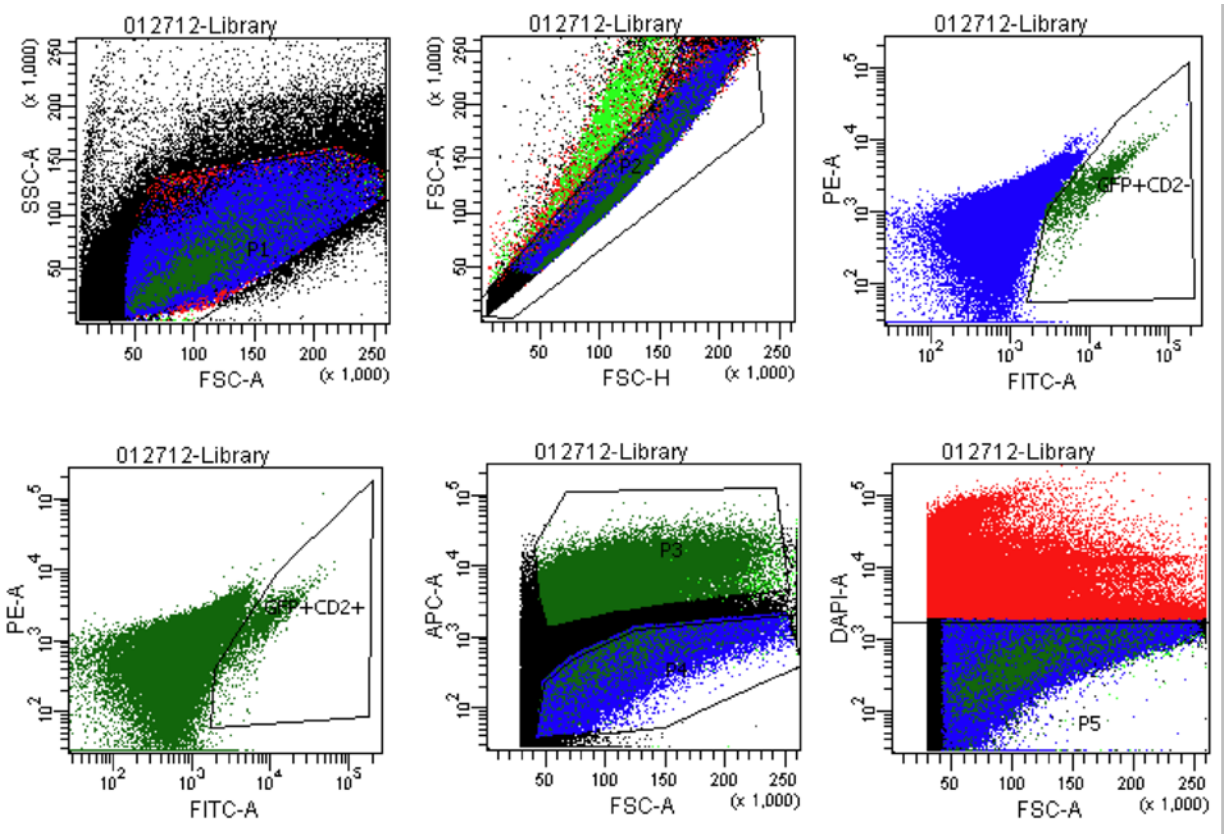
Tube: CD2

Population	#Events	%Parent	%Total
All Events	100,000	####	100.0
P1	24,921	24.9	24.9
P5	22,343	89.7	22.3
P2	22,084	98.8	22.1
P3	4,449	20.1	4.4
GFP+CD2+	2	0.0	0.0
P4	8,926	40.4	8.9
GFP+CD2-	7	0.1	0.0

A robust CD2-positive cell population is detected. The lower left plot shows analysis of this population for yellow (PE-A) vs. green (FITC-A) fluorescence. Autofluorescence of cells and yolk granules is considerable in the green band, but consistent, causing the particles observed in this non-GFP-expressing population to distribute along the diagonal. The $GFP^{+}CD2^{+}$ gate shown represents a higher green:yellow ratio, reflecting the shorter wavelength emission maximum of GFP relative to autofluorescence, and detects very few events (here 2 per 100,000) in this non-

Nature Methods: doi:10.1038/nmeth.2558

GFP-expressing population. A similar analysis is shown for CD2-negative cells in the upper right plot. Finally, we present the results of the same analysis performed on our experimental samples, for which pooled library transformant males were crossed to CD2-expressing females (the total number of events shown here is 10-fold higher):



Tube: Library

Population	#Events	%Parent	%Total
All Events	1,000,000	####	100.0
P1	349,223	34.9	34.9
P5	264,717	75.8	26.5
P2	262,266	99.1	26.2
P3	92,211	35.2	9.2
GFP+CD2+	1,617	1.8	0.2
P4	83,225	31.7	8.3
GFP+CD2-	1,516	1.8	0.2

Note the presence of robust GFP-expressing subpopulations in both CD2-positive and CD2-negative populations. These cells are sorted into culture medium from roughly 10^8 total cells per day for three days to give the GFP^+ cell populations analyzed further, while control cells

Nature Methods: doi:10.1038/nmeth.2558

(typically $2x10^5$) are collected from P3 and/or P4 (in cases where P3 is limiting, due to expression of CD2 in a more restricted population of embryonic cells); this gives an 'input' sample, reflecting the frequency with which each library insert is found in the population of sorted embryos. In principle, CD2-positive and CD2-negative cells prepared from the same embryo should not differ in their representation of library inserts in the absence of GFP sorting; we confirmed this prediction in pilot experiments, as we discuss below. Supplementary Table 2 lists the yield of cells obtained from each population.

Cells were concentrated by centrifugation of collection tubes (15 min at ~400g) and aspiration of all but ~0.5 mL of culture medium, then resuspended in the remaining culture medium and transferred to 0.6 mL PCR tubes and centrifuged 5 minutes @ 5,000g. This two-step collection process proved best for collecting very small numbers of cells in extensive pilot experiments. After complete removal of culture medium, the cell pellet is vortexed in 5 μL extraction buffer (10 mM Tris pH 7.5, 1 mM EDTA, 25 mM NaCl, 200 μg/mL proteinase K) and incubated 30 minutes at 37°, then 10 minutes at 95° to inactivate proteinase K and stored at -20°.

E. cCRM insert amplifications from collected cells, and Illumina library preparation including indexing

We designed our cCRM library to be compatible with Illumina sequencing. Each insert, as initially synthesized, has a common upstream end comprising, in 5' to 3' order, a Gateway attB1 site, and a common forward sequencing primer. The downstream end of each synthesized library insert comprises, in 5' to 3' order, a Gateway attB2 site, and a common reverse sequencing primer. The attB1 and attB2 sites facilitate the cloning of the inserts *en masse* first into a Gateway DONR vector and then, after amplification of the library, into the pEFS-Dest vector. After a population of cells containing inserts of interest is isolated, PCR using the common sequencing primers allows recovery of a very pure insert population for quantification by high-throughput sequencing. We designed these common primers using our UniPROBE database of TF DNA binding specificities [17]; we generated synthetic sequence that is not predicted to be bound by any known TF (considering hundreds of TFs from multiple organisms) to minimize chances of it modulating the activity of any inserted CRM, and then tested primers derived from

Nature Methods: doi:10.1038/nmeth.2558

it in pairs, first for their suitability as PCR primers and then in control reporter experiments with known enhancers, to ensure that they showed the expected pattern of embryonic reporter expression.

We recovered library inserts by nested PCR from sorted cell genomic DNA. Crude cell extracts were pooled according to sample where necessary to achieve sufficient numbers for accurate quantification of insert abundance, then split five-fold before PCR amplification (Supplementary Fig. 2). This ensures that ample PCR product will be produced for Illumina library preparation.

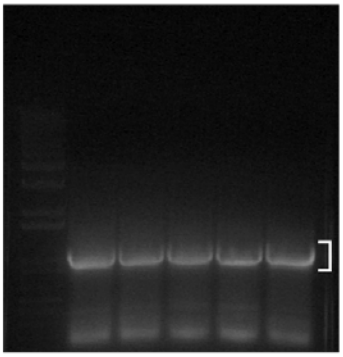
Each first-round PCR was performed in a 25 μL reaction using KAPA Hi-Fi HotStart ReadyMix (Kapa Biosystems), supplemented with 6% ethylene glycol and $MgCl_2$ to offset the EDTA introduced with the crude cell extract (typically 0.5 μL of 25 mM $MgCl_2$ when ~6 μL of template is added to each reaction). Primers for the first-round (outer) PCR are specific to our pEFS reporter vector (see Section A) and were used at 1 μM each: nestF = GAATTGAATTGTCGCTCCGTAGAC; nestR = CAAGTATTTCCCCTTCGAGCTTG. Cycling conditions (optimized for sensitivity of detection and amplification) were: 1 x 2' @ 98°C; 1 x [20" @ 98°C, 3' @ 63°C, 30" @ 72°C]; 1 x [20" @ 98°C, 2:30 @ 63°C, 30" @ 72°C]; 1 x [20" @ 98°C, 2' @ 63°C, 30" @ 72°C]; 1 x [20" @ 98°C, 1:30 @ 63°C, 30" @ 72°C]; 13 x [20" @ 98°C, 1:15 @ 65°C, 30" @ 72°C]; 1 x 1' @ 72°C; hold @ 4°C (17 cycles in all).

Second-round (inner) PCRs were performed in 50 μL reactions, again using KAPA Hi-Fi HotStart ReadyMix supplemented with 6% ethylene glycol. The template for each reaction was 2 μL of first-round PCR, and the reaction mixture was heated to 80°C before template was added in order to ensure a true hot start despite the addition of active enzyme with the template. Primers for the second-round PCR are the SEQ1 and SEQ2 sequences introduced at the ends of each cCRM during library construction (see Section A) and were used at 2 μM each. Cycling conditions (optimized for yield of full-length product and accurate representation of spiked-in controls) were: 1 x 2' @ 98°C; 21 x [20" @ 98°C, 1' @ 65°C, 30" @ 72°C]; 1 x [20" @ 98°C, 1' @ 65°C, 45" @ 72°C]; 1 x [20" @ 98°C, 1' @ 65°C, 1' @ 72°C]; 1 x [20" @ 98°C, 1' @ 65°C, 1:15 @ 72°C]; 1 x [20" @ 98°C, 1' @ 65°C, 1:30 @ 72°C]; 1 x [20" @ 98°C, 1' @ 65°C, 1:45 @

Nature Methods: doi:10.1038/nmeth.2558

72°C]; 1 x [20" @ 98°C, 1' @ 65°C, 2' @ 72°C]; 1 x [20" @ 98°C, 1' @ 65°C, 2:15 @ 72°C]; hold @ 4°C (28 cycles in all).

PCR products (900-1,100 bp) were agarose gel-purified as shown below, and quantified by NanoDrop. 1 μg each of the 5 PCR product pools from a given cell pool were combined to make the starting material for Illumina library preparation.



Nested PCR product from one cell sample (divided 5 ways before first-round PCR). The bracketed region was cut out for purification, sonication, and subsequent library preparation.

Spike-in and replicate control experiments.

We employed 'spike-in' controls to assess bias and replicate lanes to assess noise in amplifying cCRMs from a complex library; the known, qPCR-validated dilution of each control enabled us to accurately assess the representation of each control in the resulting sequencing lanes. Specifically, we amplified 900-1,100 bp regions of the *E. coli* chromosome (using the same 2-step PCR protocol described in Section A), and cloned them individually into pEFS-Dest. Purified control plasmids were spiked into pooled library plasmid at various known dilutions, and the dilutions of each control plasmid in this mixture were confirmed by qPCR with insert-specific primer pairs; we performed duplicate measurements of two pairs per insert. An aliquot of ~50,000 plasmids (1 μL of a 500 fg/μL dilution) was subjected to our library insert recovery PCR workflow, as described above, followed by Illumina library preparation and sequencing

Nature Methods: doi:10.1038/nmeth.2558

(see below). The following table shows the comparison of the frequency with which each control *E. coli* region was detected by PCR from a complex mixture followed by high-throughput sequencing, compared with the expected frequency assuming the qPCR results accurately reflect the composition of the library pool:

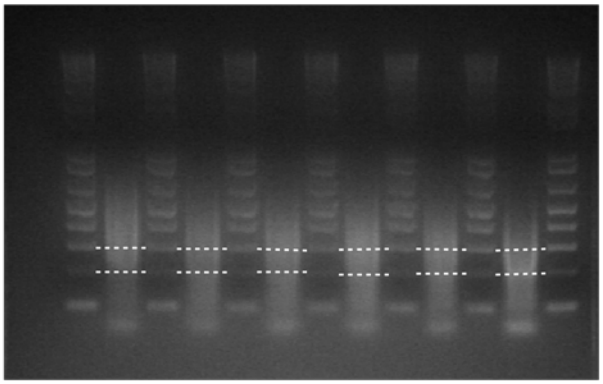
spike-in control	nominal dilution factor	measured dilution factor (qPCR)	expected counts / million reads	observed counts / million reads
E_coli_control_3	1:100	1:80	12,500	10,256.85
E_coli_control_5	1:100	1:87	11,494	12,243.19
E_coli_control_7	1:1,000	1:838	1,193	343.17
E_coli_control_11	1:1,000	1:829	1,206	77.08
E_coli_control_12	1:10,000	1:2,752	363	18.14
E_coli_control_16	1:10,000	1:2,726	367	21.23
E_coli_control_21	1:100,000	1:88,551	11	0.11
E_coli_control_22	1:100,000	1:968	1,033	0.77

These experiments show that, for very common inserts, our workflow is fairly accurate in recovering the abundance of insert in a complex mixture, but that accuracy suffers at very high dilutions. Since only ~50,000 plasmids were sampled, a 1:100,000 dilution would be expected to present a difficulty for accurate representation. To test whether the observed variation from expected frequencies represented bias in PCR amplification of different inserts based on their sequence (which would be expected to affect alternate samples equally, so that a comparison of observed abundance in different samples would still be informative) or noise introduced during the PCR amplification step, we performed the entire workflow on four different aliquots of a library pool and compared the resulting sequence counts. The entire sample preparation and sequencing process is highly reproducible: the number of reads mapped to each cCRM or control region was consistent across sequencing libraries prepared from four different aliquots of the plasmid library pool (mean Pearson $R^2 = 0.81$, ranging from 0.78 to 0.83 over six pairwise comparisons; Supplementary Fig. 3).

Nature Methods: doi:10.1038/nmeth.2558

Illumina library preparation, including indexing

Illumina sequencing libraries were prepared using very minor modifications of standard protocols [18] and the Multiplexing Sample Preparation Oligonucleotide Kit (Illumina). Pooled, purified PCR product was sonicated by Covaris S2 exactly as described [18]. Samples were then end-repaired with the End-IT DNA End-Repair Kit (EpiCentre Biotechnologies) and A-tailed with Klenow exo- (New England Biolabs). Standard adapters (Index PE Adapter Oligo Mix) were ligated using Quick T4 DNA Ligase (New England Biolabs). The products of adapter ligation were run on 2% agarose gels and size-selected as indicated below:



Purified products were quantified and checked for concentration and size distribution by Agilent 2200 TapeStation. Enrichment PCRs were performed in 50 μL reactions using Phusion thermostable polymerase (New England Biolabs). Each reaction contained 25 ng of template DNA and 1 μL each of primer InPE2.0, InPE1.0, and a numbered index primer. Indices were chosen for the 36 total samples (Supplementary Table 3) in order to fill 7 lanes of a flow cell, and so that each statistical comparison (*e.g.*, 3 *twi*:CD2^{+}GFP^{+} samples versus 3 *twi*:CD2^{+} 'input' samples) would be between samples run in different lanes with the same index. Cycling conditions, chosen to balance adequate yield for sequencing with minimal variance introduced at the enrichment PCR step, were: 1 x 30" @ 98°C; 10 x [10" @ 98°C, 30" @ 65°C, 30" @ 72°C]; 1 x 5' @ 72°C; hold @ 4°C.

Nature Methods: doi:10.1038/nmeth.2558

F. Illumina DNA sequencing

Purified enrichment PCR products were again assessed by Agilent 2200 TapeStation and submitted to the Partners Center for Personalized Genetic Medicine for concentration measurement by PicoGreen fluorescence and qPCR, followed by equimolor index pooling and sequencing (50 base single-end read) on the Illumina HiSeq 2000.

G. Mapping Illumina sequencing reads

In principle, it would be faster (in terms of total CPU run-time) and potentially less noisy to map high-throughput sequencing reads directly to a reduced 'genome' constructed from our library of cCRMs. However, we anticipate that tiling large swaths of noncoding sequence to assess their regulatory potential will be a major application of eFS going forward, and as this requires the testing of overlapping sequences, it would create huge numbers of ambiguous mappings to repetitive regions. We have therefore instead chosen to pursue a two-step strategy, in which sequencing reads are first mapped to the *D. melanogaster* genome, and then mapped positions are assigned to members of the cCRM library, as described in detail below.

Comparison of sequencing read mapping algorithms.

We compared 5 different algorithms for mapping sequencing reads to the *Drosophila* genome, and optimized parameter settings for each algorithm. Specifically, we compared:

- segemehl [19] (version 0.0.9.4, current version is 0.1.3)
- bowtie [20] (version 0.12.7, current version is 0.12.8 or bowtie 2.0.0-beta7)
- BWA [21] (using current version 0.6.2)
- Stampy [22] (version 1.0.10, current version is 1.0.18)
- SOAP2 [23] (version 2.20, current version is 2.21 or Soap3 0.01 beta)

Bowtie, BWA, Stampy, and SOAP2 use the Burrows Wheeler Transformation (BWT) and segemehl internally uses an enhanced suffix array. We considered but did not compare the most popular algorithms using hash tables because they failed to process the input data (MAQ [24]) or

Nature Methods: doi:10.1038/nmeth.2558

were not able to process reads longer than 32 nucleotides (Eland, provided with Illumina facilities) at the time of our comparisons.

The final parameter settings we used for each algorithm were as follows:

- segemehl: *-M 100 -E 5 -D 2 -A 80*
- bowtie: *-l 20 -n 0 -e 70* (border reads) and *-l 27 -n 1 -e 150* (center reads)
- BWA: *-l 20 -k 0* (border reads) *-l 27 -k 1* (center reads)
- Stampy: builds upon BWA results and uses BWA parameters
- SOAP2: *-l 20* (border reads) and *-l 27* (center reads)

A description of these parameter settings follows below. We refer to reads that contained SEQ1 or SEQ2 primers (see Section A for a description of the SEQ1 and SEQ2 primers) as *'border' (or 'end') reads*, since they derive from the 5' or 3' ends of the cCRMs; reads between the end reads are referred to as *'center' reads*.

All algorithms except segemehl take the length of the seeding subsequence as an input parameter (*-l*). Since our reads effectively have different lengths, i.e. for some a 22-nt long primer sequence was removed, we set this for center reads and border reads to 27 and 20 nt, respectively. Bowtie and BWA count the mismatches in the seed subsequence separately (parameters *-n* and *-k*, respectively); we set them to 0 for border reads and 1 for the longer center reads. The possibility of 2 mismatches showed only minor improvement for mapping results but significantly longer runtime in bowtie. For segemehl, an edit distance of 2 (parameter *-D*) for the dynamic seeds is allowed and gave the best results in our comparisons. Bowtie additionally limits the sum over quality scores at mismatched positions. With a maximum added score of 30, setting *-e* to 70 and 150 allows up to 2.3 and 5, respectively, high-confidence mismatches in the entire read. All other alignment parameter settings not mentioned above were set at the default values of each algorithm. The parameters offered by segemehl, and the values we settled on, were:

- A maximum of two mismatch or indels in the seed area (*-D*)
- A maximum of an E-value (introduced in BLAST [25]) of 5 in the seed alignment (*-E*)
- At most 100 occurrences of the seed in the genome (*-M*)

Nature Methods: doi:10.1038/nmeth.2558

- 80% or higher matching of characters after extending the seed to the entire read (*-A*).

The criteria on which we judged the performance of the selected mapping algorithms were:

- The fraction of reads that could be mapped to the genome
- The fraction of reads that mapped to the cCRM library
- The number of cCRMs to which reads could be mapped
- A false discovery rate (FDR) for the detection of cCRMs

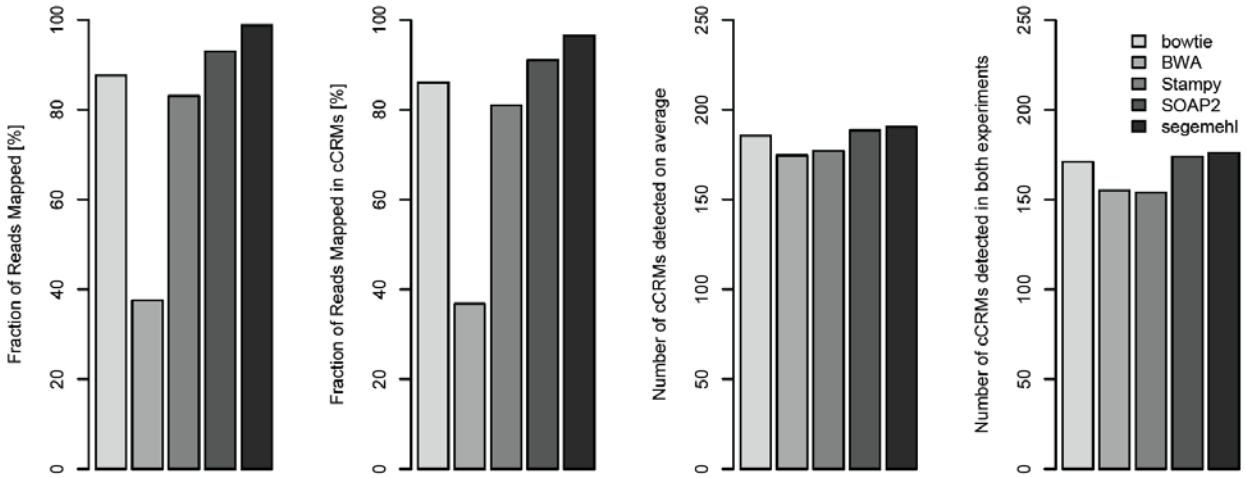
The fraction of reads that can be mapped to at least one position in the reference genome is the simplest and most obvious criterion. Reads that do not map to the regions that went into the cCRMs are most likely not mapped correctly, so we look at the fraction that maps to exactly these regions. If some reads map to a cCRM, then we call the cCRM as *"detected"*. As a threshold for cCRM detection, we required: (1) at least 1 read from each end (*i.e.*, 5' and 3' ends); (2) at least 5 positions covered by center reads (*i.e.*, reads that were sequenced without SEQ1 or SEQ2); and (3) at least 10 reads in total. Requiring at least 1 read at each end results in a very low FDR, since end reads map to at most 1 cCRM, and are only accounted to the cCRM if the end read is within 2 nt of the expected position (i.e., the beginning or end of the cCRM) thus allowing for potential indels. Requiring at least 5 positions covered by center reads does not improve FDR or detection, but in theory should filter out noise due to sequencing of primer-dimers, although we have not actually observed evidence that this problem occurs. Requiring at least 10 reads in total ensures some comparability between detections in different experiments.

We compared the algorithms on two different input samples, each comprising 200,000 cells (corresponding to 100,000 cCRM inserts), collected on two different days in FACS on *duf2*:CD2$^-$ cells from *D. melanogaster* embryos (these embryos, in which the *duf2 -5.1 kb* enhancer [12] was used to drive CD2 expression, were not used for eFS in this study, but rather served as an independent source of input samples); these two samples were barcoded with the same primer for indexing, and were sequences in different lanes of the same HiSeq flow cell. We generated and tested 48 sets of 420 cCRM-like 1-kb genomic windows matched for length, sequence context (*e.g.*, intronic, intergenic) and GC content to our foreground windows to

Nature Methods: doi:10.1038/nmeth.2558

determine a false discovery rate (FDR) for each mapping technique. Counting those FDR windows that are assigned enough reads to be called detected, gives an estimation of cCRMs that got reads mapped to only from noise. Using the parameter settings described above, for each of the 5 algorithms the FDR for cCRM detection was less than $5x10^{-5}$.

As shown below, BWA, Stampy, SOAP2, and segemehl all achieved a 0% FDR, with segemehl mapping the highest number of reads to the *D. melanogaster* genome (98.8%) and the most reads to the cCRMs (96.5%). Segemehl also detected the highest number of cCRMs in total (173.5 averaged over 2 Illumina HiSeq sequencing lanes), including 155 identical cCRMs in the two duplicate samples. Although comparing mapping algorithms highly depends on the data set and the parameter settings, we found segemehl to perform best in terms of the above-described four criteria in our experiments. Thus, we chose segemehl as the most suitable mapping algorithm for our enhancer-FACS-Seq data sets.



Pipeline for mapping sequencing reads.

We implemented our analysis pipeline for mapping sequencing reads with a Java program for each subtask and a Perl script connecting these modules to the complete pipeline, which can be run with one command line. The raw reads are processed in 5 smaller steps: (1) preprocessing; (2) mapping; (3) filtering; (4) stacking and (5) assignment to cCRMs. The mapping itself is provided by the segemehl program.

Nature Methods: doi:10.1038/nmeth.2558

The input file for the pipeline, i.e. the raw reads, are provided with command line parameter *-i*. Additionally a tag for the experiments has to be set with *-t name* to distinguish different analyses. Every file generated during processing will start with this tag. The output directory for these files can be specified with *-o path* or is otherwise equal to the input directory.

The preprocessing includes the search for and removal of PCR primer sequences (*i.e.*, SEQ1 and SEQ2, described in Section A) from the 5' ends of the reads (the path to the file containing these primers is set with parameter *-p path*). We look only for complete primer sequences starting at the first sequenced position, since the detection of the other cases would bring only minor improvement to the overall results but would take significantly longer in terms of analysis run time. To conserve the information that the presence of these primers gives, these reads are processed separately but in the same manner. Due to the origin of the two data sets with reads that had a primer and those that did not, we refer to them as 'border' (or 'end') and 'center' reads, respectively (see description above). If for some technical reason the sequencing reads are consistently offset by a few nucleotides from the PCR-amplified fragments, this can be handled by setting the *-1* parameter of the pipeline script (e.g. *perl eFS.pl ... -1 5* for a 5-nt offset).

Parameters used by segemehl are passed through the parameter setting *-2* (e.g. *perl eFS.pl ... -2 "-H 0 -M 100 -E 5 -D 2 -A 80"*). For convenience, the string *"opt"* will be automatically translated to the settings we used for our analysis of mapping algorithms, which are also the default values.

Segemehl can report multiple locations for one read (parameter *-H 0*, which is also the default setting). We use our knowledge about the experiment to decide which locations appear to be correct. If a read maps to exactly one legal position, *i.e.*, in the range of a cCRM for center reads or within a 5-nt window at a cCRM start or end for border reads, we accept this location. If a read maps to no reasonable position or fits ambiguously, it is left out. With command line parameter setting *-3* this filtering can be skipped.

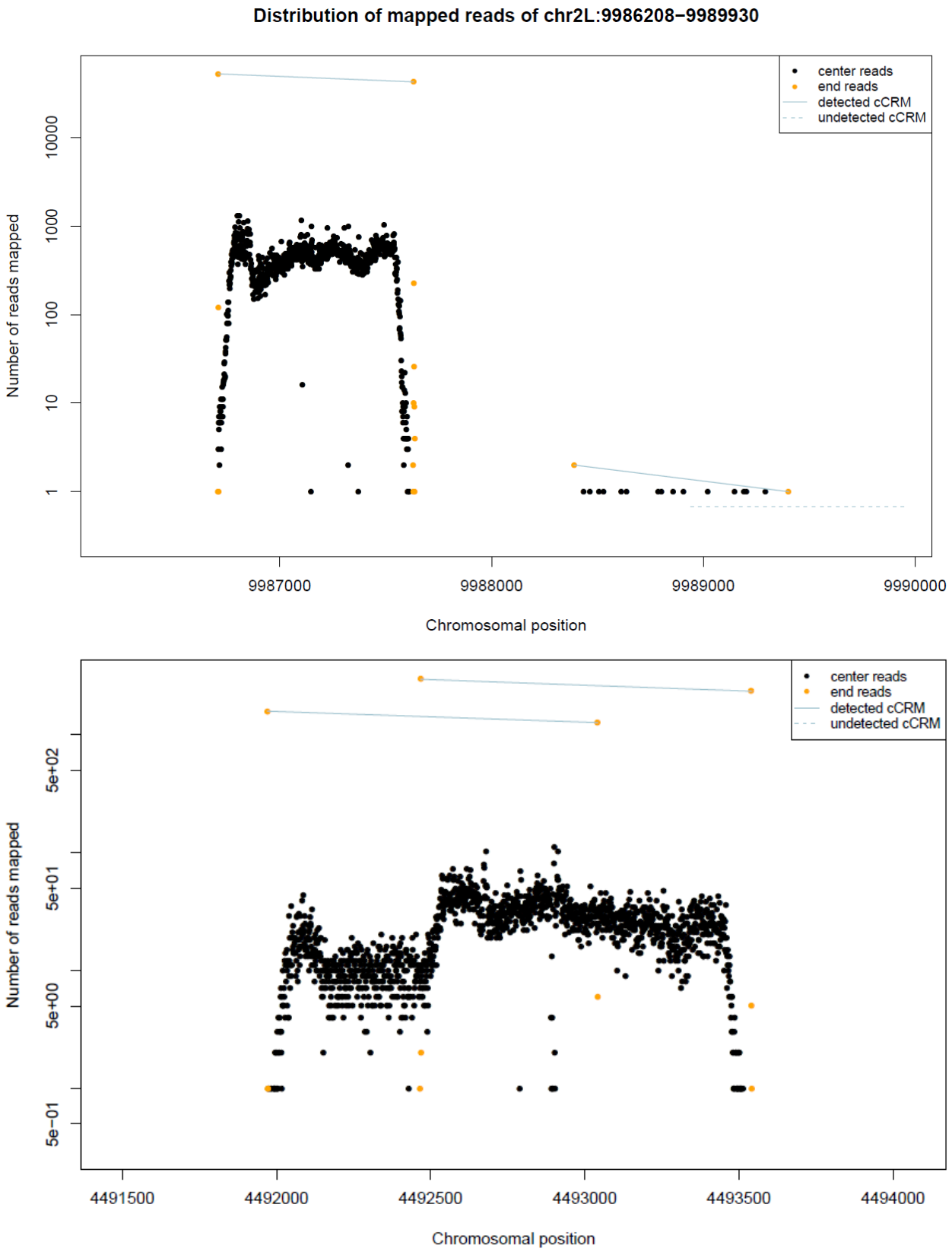
In the next step, similar reads, *i.e.*, reads starting at the same chromosomal position and with the same length, are collected. This reduces file sizes enormously to a point manageable by a PC and

Nature Methods: doi:10.1038/nmeth.2558

standard text editors. If the parameter setting *-4* in the pipeline is set, then more information on the reads is kept for further analysis, which means reads have to agree in their strand as well and the GC content of the fragments sequenced is estimated and stored in the files containing lists of the mapped reads.

In the last step, the pipeline assigns the reads to the cCRMs provided by the command line parameter setting *-c path*. One potential problem is that overlapping cCRM windows contribute indistinguishable reads to the same genomic regions. We found that this can be addressed in a relatively straightforward manner by using the unambiguous border reads as weights for dividing the reads that map to overlapping cCRM windows, since we found that in isolated cCRMs the number of border reads and the total number of reads are highly correlated (Pearson $R > 0.95$). The number of reads that have mapped per position is depicted below *(next page)* for two representative genomic regions; the upper panel shows two isolated, detected cCRMs and one overlapping, undetected cCRM, and the lower panel shows two overlapping cCRMs that were created for the purpose of tiling a genomic region in creation of the cCRM library.

Nature Methods: doi:10.1038/nmeth.2558



Nature Methods: doi:10.1038/nmeth.2558

The final output of our eFS sequencing reads mapping pipeline includes for each cCRM: its name and chromosomal location, the number of border reads that mapped to its start and end, the resulting weight, the number of center positions covered by the beginning of a mapped center read and the total count of reads mapped to the cCRM. From these data, one can then readily compute whether a cCRM has been detected and can compare the results quantitatively to those from replicate experiments.

H. Statistical analysis of enhancer-FACS-Seq (eFS) data

In order to detect and quantify enrichment of cCRMs in GFP^+ cell populations, we collected the number of reads mapped to each cCRM for each replicate population and control 'input' population, and then filtered out cCRMs not detected (as defined in Section G above) in any input sample replicate. Collected values for detected cCRMs were then read into R (http://www.r-project.org/), rounded, and coerced to integers.

The comparison of the frequency with which high-throughput sequencing reads mapping to a given cCRM are observed in a selected population with the frequency observed in an input population is fundamentally similar to the problem of detecting differential expression in an RNA-Seq experiment, or enrichment in a ChIP-Seq experiment if genomic occupancy regions are pre-defined. Thus, we utilized a previously developed software package intended for use with these data types. Specifically, enrichment and statistical significance were calculated using the DESeq package [26] with standard parameters and size factor estimation; we considered an adjusted p-value < 0.1 as evidence of statistical significance. Importantly, the DESeq software package permitted the comparison of unreplicated data using variance estimated from replicate controls, and it produced results that performed agree quite well with those obtained from validation by traditional reporter assays (see below). The distribution of results is depicted by plotting the $\log_2$(fold enrichment or depletion) against the abundance of reads in the input population (Fig. 2a, Supplementary Fig. 4).

For abundant populations (*twi*:$CD2^+$, *twi*:$CD2^-$, *Mef2-I-*E_{D5}:$CD2^-$ and *duf*:$CD2^-$), three or more GFP^+ replicates were compared to three or more input replicates (*e.g.*, GFP^+CD2^+ cell replicates

Nature Methods: doi:10.1038/nmeth.2558

were compared to $CD2^+$ cells collected from the same preparation). Supplementary Table 2 shows the number of cells in each sample from which a sequencing library was prepared. CD2 expression driven by *Mef2-I-*E_{D5}:CD2 or *duf*:CD2 is in a much more restricted population of cells (*i.e.*, 0.8% of viable cells on average for *Mef2-I-*E_{D5}:CD2); since a much smaller number of GFP^+CD2^+ cells could be collected, we were concerned that the noise introduced by sampling such small numbers of cells would reduce the sensitivity of our analysis. Therefore, we pooled all three or six days' collections of double-positive cells from these experiments into a single sample. To conservatively estimate variance and sampling noise, we collected input populations of twice the size of the double-positive cell population (50% of input cells are expected to carry inserts, as the male parents of the dissociated embryos were heterozygous for reporter transgenes, while in principle 100% of GFP^+ cells should carry inserts) or fewer (see Supplementary Table 2). As collection of relatively rare $CD2^+$ cells from this experiment would have been costly in terms of sorting time and wasted opportunity to collect GFP^+CD2^+ cells, we used $CD2^-$ cells as input controls. In principle, the abundance of each cCRM in $CD2^+$ and $CD2^-$ cell populations should be the same, as these are derived from the same embryos. Nevertheless, to test this assumption, we simultaneously assessed the false positive rate at which cCRMs would be called significantly enriched or depleted in identical populations, by performing an identical comparison of *twi*:$CD2^+$ (triplicate) to *twi*:$CD2^-$ (triplicate) cell samples. The results, shown in Fig. 1d, support that these populations are indeed essentially identical.

Supplementary Table 4 shows the complete list of cCRMs assayed for enrichment in GFP^+ cells versus input in *twi*:$CD2^+$ (mesoderm), *twi*:$CD2^-$ (non-mesoderm), *Mef2-I-*E_{D5}:$CD2^+$ (fusion-competent myoblast subset), *Mef2-I-*E_{D5}:$CD2^-$ (approximately whole embryo), *duf*:$CD2^+$ (muscle founder cells), and *duf*:$CD2^-$ (more approximately whole embryo) with enrichment/depletion values, *p*-values, and a description of the results of validation by conventional reporter assays, where such results are available.

Nature Methods: doi:10.1038/nmeth.2558

I. DNA sequence motif enrichment analysis

One of the goals of this study was to identify *cis* regulatory motifs and the corresponding transcription factors that act through transcriptional enhancers active in the *Drosophila* embryonic mesoderm. One approach for identifying putative transcriptional regulators is to look for enrichment or depletion of DNA sequence motifs associated with known TFs, within sets of transcriptional enhancers active in the same tissue. The DNA-binding specificities of hundreds of *D. melanogaster* TFs have been determined by a variety of experimental methods [1,10,27-29]. Therefore, we used the Lever algorithm [5] to identify significant over-representation of matches, scored according to their evolutionary conservation, to 567 publicly available *Drosophila* TF binding site motifs [1,10,27-29] as compared to matched, randomly selected noncoding sequence, as described in more detail below.

A dictionary of *Drosophila* TF binding site motifs was compiled from the Fly Factor Survey database [27] as well as from the literature [1,10,28,29]. Uninformative positions flanking the motifs were trimmed from both sides until either one position with information content (IC) ≥ 1 or two consecutive positions with IC ≥ 0.5 were reached.

To remove redundancy within the motif dictionary, motifs were clustered as in [30] using centroid-linkage hierarchical clustering and a Pearson correlation coefficient cutoff of 0.77. Then, for each of the resulting 139 clusters, a motif representative ('exemplar') was chosen as the motif that was closest to the cluster average and whose TF had evidence of mesodermal expression [15,31,32], except that in cases where a known mesodermal regulator was within a cluster, its motif was chosen as the exemplar (*e.g.*, Bap).

To identify the putative regulatory motifs of the *twi*:CD2^{+}, *Mef2-I-E*$_{D5}$:CD2^{+}, and *duf*:CD2^{+} foreground (FG) sequence sets, we used the previously described algorithm Lever [5]. Briefly, this analysis calculates the over-representation of individual motifs or combinations of motifs, according to their density and evolutionary conservation as quantified by the PhylCRM scoring scheme [5], in each FG sequence set as compared to a matched random set of background (BG) sequences. Rejection sampling was used to build a BG set for each FG sequence set; the BG set

Nature Methods: doi:10.1038/nmeth.2558

was about 20 times the size of the FG set and matched its length, GC content, and repeat content. Formatting of the 12 fly genome sequences and all settings were used as previously described [33], except that repeats were not masked and length correction was not used since all sequences were roughly of uniform length, approximately 1 to 1.5 kb.

We first ran Lever to score each of the 139 exemplar motifs individually. To reduce the loss of statistical significance due to a large number of hypotheses tested, any motif that did not have occurrences in at least one-quarter of the FG sequences was considered unlikely to contribute significantly to the overall FG set and thus was removed from the 139-exemplar motif dictionary, resulting in an 86-exemplar motif dictionary. We then used Lever to inspect the over-representation of all single and pairwise combinations of the resulting reduced dictionary of 86 motif exemplars. Motif PWMs are provided in Supplementary Table 6.

J. Classifier analysis

We sought to determine whether cCRMs could be classified as active or inactive in a given cell population based on the number and quality of motif matches in the cCRM sequence. 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). This scoring method allows weaker motif instances to be captured and avoids threshold effects. The results did not change noticeably when this threshold was varied as long as it remained permissive.

For classification we used the Gaussian Naïve Bayes implementation in the *scikit-learn* package [34] for Python. A Naïve Bayes (NB) classifier implements a simple probabilistic model where each feature (here, the score for a given motif) contributes independently to the posterior probability of belonging to a class. NB classifiers previously have been shown to outperform more complex models at predicting expression based on sequence in other systems [35]. Here, we compared the performance of the NB model to more complex classifiers (Random Forest and

Nature Methods: doi:10.1038/nmeth.2558

Support Vector Machine classifiers with Gaussian kernels), which can model interactions between features, and found that the more complex models did not improve classification performance (data not shown). In addition, there are theoretical justifications for why NB classifiers can achieve high accuracy even if the strong independence assumption does not hold [36].

Here, for each motif, our classifier separately fits a Gaussian distribution to the scores for that PWM in the active cCRM set and in the inactive cCRM set (see below). These distributions are then used to generate posterior class probabilities for unlabeled data. The level of background motif matches is modeled by a Gaussian distribution inferred from the inactive cCRMs, which compensates for the permissive motif match threshold that was used to generate the scores.

We assigned class labels to specific cCRMs as follows: 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 ranked list. To evaluate classification accuracy, we split the labeled cCRM feature vectors into training and test sets using stratified 10-fold 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 the classifier's performance by calculating the AUC statistic over all of the cross-validated predictions. We compared performance with different values of *k*, and found that the classifier for *Mef2-I-*E_{D5}:$CD2^+$ cCRMs performed better with lower values of *k* relative to the *twi*:$CD2^+$ classifier, which is consistent with expectations based on tissue heterogeneity.

We compared the performance of our feature selection method (i.e., in which we identified discriminatory motifs) relative to manually picking motifs of known regulators (see below) for that tissue type. Classifier models trained with our feature selection method outperformed those that used the motifs of known regulators for both *twi*:$CD2^+$ (AUC = 0.74 using feature selection method versus 0.59 using known regulatory motifs) and *Mef2-I-*E_{D5}:$CD2^+$ (AUC: 0.93 using feature selection method versus 0.72 using known regulatory motifs) eFS data.

Nature Methods: doi:10.1038/nmeth.2558

The known regulatory motifs used for *twi*:CD2^{+} eFS data (*i.e.*, whole mesoderm) are:
1258_twi_Zinzen
241_Tin_Cell_FBgn0004110
1256_mef2_Zinzen
1254_bap_Zinzen
1246_Lmd_Busser_Sec (called “lmd_sec” in Fig. 5a and Supplementary Fig. 7)
573_lmd_SOLEXA_5_FBgn0039039 (called “lmd_FFS” in Supplementary Fig. 7)

The known regulatory motifs used for *Mef2-I-E*$_{D5}$:CD2^{+} eFS are:
1258_twi_Zinzen
1256_mef2_Zinzen
1254_bap_Zinzen
1246_Lmd_Busser_Sec (called “lmd_sec” in Fig. 5a and Supplementary Fig. 7)
573_lmd_SOLEXA_5_FBgn0039039 (called “lmd_FFS” in Supplementary Fig. 7)
338_SuH_FlyReg_FBgn0004837

The discriminatory motifs learned in at least 50% of the ‘folds’ in the 10-fold cross-validation for *twi*:CD2^{+} eFS data (*i.e.*, whole mesoderm) are:
716_lola-PC_SANGER_5_FBgn0005630
659_CG8765_SANGER_5_FBgn0036900
658_CG8281_SANGER_5_FBgn0035824
652_CG3919_SANGER_5_FBgn0036423
623_kay_Jra_SANGER_5_FBgn0001297_SANGER_5_FBgn0001291
583_CG7928_SOLEXA_5_FBgn0039740
269_D_NAR_FBgn0000411
1258_twi_Zinzen
297_brk_FlyReg_FBgn0024250
700_lola-PO_SANGER_5_FBgn0005630
600_vfl_SOLEXA_5_FBgn0259789 4

Nature Methods: doi:10.1038/nmeth.2558

The discriminatory motifs learned in at least 50% of the 'folds' in the 10-fold cross-validation for *Mef2-I-E*$_{D5}$:CD2^{+} eFS (*i.e.*, FCMs) are:
573_lmd_SOLEXA_5_FBgn0039039 (called "lmd_FFS" in Supplementary Fig. 7)
702_lola-PW_SANGER_5_FBgn0005630
583_CG7928_SOLEXA_5_FBgn0039740

Discriminatory motifs learned in 10-fold cross-validation for *duf*:CD2^{+} eFS (*i.e.*, FCs) are not reported here, since the results of the classifier analysis were not statistically significant.

Motif PWMs are provided in Supplementary Table 6.

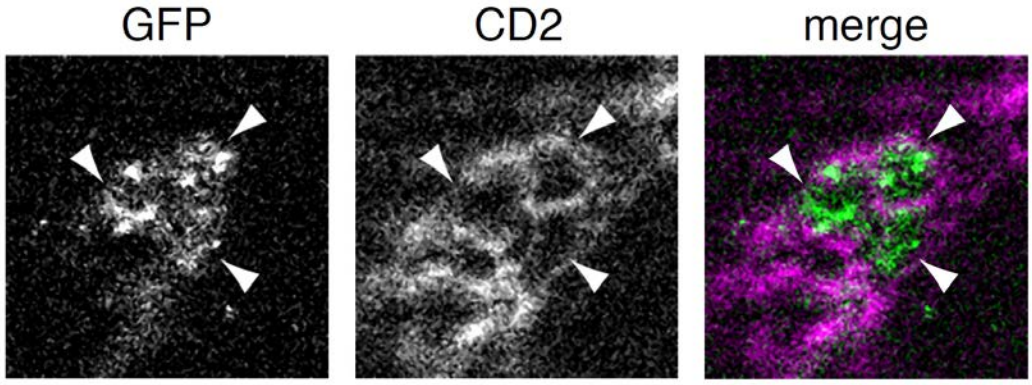
K. Traditional reporter assays

104 cCRMs assayed by eFS in any cell population were individually validated by more traditional means (Supplementary Fig. 5, Supplementary Table 4). 37 of these had been included in the library on the basis of earlier pilot experiments and sequence-validated clones in pEFS-Dest were available for them, while the remaining 67 were recovered from single transformant male flies (and identified by PCR and subsequent sequencing of their library inserts) after library injections. These latter 67 lines could in principle contain PCR-induced mutations which might affect the pattern in which they drive GFP expression, since they derived from library clones which were not sequence-verified. For 25 of these lines, we were able to ascertain the complete sequence of the library insert by sequencing two independent PCR reactions using vector-specific primers and the complete sequence of the corresponding genomic regions in the wild type flies (OreR) from which our library clones were initially amplified. In 5 of the 25 cases, the library insert matched the published *D. melanogaster* sequence (dm3) at every position, while in the remaining 20 cases all differences from the published sequence were found to be present in our wild type population. We thus conclude that we can neglect PCR-induced mutations as a significant source of error in subsequent library generation using this PCR protocol.

Homozygous (or, where unavailable, balanced heterozygous) transformant males were crossed to homozygous *twi*:CD2 females in small population cages, and broad collections (~2-17 hours

Nature Methods: doi:10.1038/nmeth.2558

after egg deposition) of embryos were fixed and stained for immunofluorescence by standard protocols [9]. Antibodies used were mouse anti-rat CD2 Alexa647 conjugate (MCA154A647, AbD Serotec) at a final dilution of 1:200, and rabbit anti-GFP, goat anti-rabbit Alexa488 conjugate, and goat anti-mouse Alexa647 conjugate (A-11122, A-11034, and A-21236, Molecular Probes), all at 1:500, and were preadsorbed to wild type embryos before use. Stained embryos were resuspended in Vectashield (Vector Labs) and, for those lines in which GFP was expressed at stage 11–12, imaged with a Zeiss Imager Z1 with Apotome in optical sectioning mode. Coexpression of GFP with CD2 (described in detail in Supplementary Table 5) was evaluated in individual optical sections with the annotator being blind to the predicted activity of the cCRMs. Coexpression is observed as GFP and CD2 being present in the same cells, since the GFP in these embryos is nuclear, while CD2 is expressed on the cell surface; in optical sections, the resulting appearance is of CD2 staining surrounding GFP staining when coexpression is present:



Maximum Image Projections were constructed for display (Fig. 3, Fig. 5c, and Supplementary Fig. 5). Coexpression was additionally subjectively assessed as "widespread" (roughly, a majority of high-CD2-expressing cells express GFP), "limited," or "no coexpression".

L. Accession ID for Illumina sequencing data

Illumina sequencing data reported in this paper have been deposited in the NCBI Gene Expression Omnibus (GEO) database with Series ID #GSE41503.

Nature Methods: doi:10.1038/nmeth.2558

Supplementary References:

1. 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).
2. Roy, S. *et al.* Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* 330, 1787-97 (2010).
3. Contrino, S. *et al.* modMine: flexible access to modENCODE data. *Nucleic Acids Research* 40, D1082-8 (2012).
4. Thomas, S. *et al.* Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. *Genome Biology* 12, R43 (2011).
5. Warner, J. *et al.* Systematic identification of mammalian regulatory motifs' target genes and functions. *Nature Methods* 5, 347-53 (2008).
6. Barolo, S., Carver, L.A. & Posakony, J.W. GFP and beta-galactosidase transformation vectors for promoter/enhancer analysis in Drosophila. *Biotechniques* 29, 726, 728, 730, 732 (2000).
7. 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, 1775-82 (2004).
8. 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, 476-83 (2008).
9. 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).
10. Philippakis, A.A. *et al.* Expression-guided *in silico* evaluation of candidate *cis* regulatory codes for *Drosophila* muscle founder cells. *PLoS Computational Biology* 2(2006).
11. Dunin-Borkowski, O.M. & Brown, N.H. Mammalian CD2 is an effective heterologous marker of the cell surface in Drosophila. *Dev Biol* 168, 689-93 (1995).
12. 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).
13. Drummond-Barbosa, D. Stem cells, their niches and the systemic environment: an aging network. *Genetics* 180, 1787-97 (2008).
14. Bischof, J., Maeda, R.K., Hediger, M., Karch, F. & Basler, K. An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases. *Proceedings of the National Academy of Sciences of the United States of America* 104, 3312-7 (2007).
15. Estrada, B. *et al.* An integrated strategy for analyzing the unique developmental programs of different myoblast subtypes. *PLoS Genet* 2, e16 (2006).
16. Thisse, C., Perrin-Schmitt, F., Stoetzel, C. & Thisse, B. Sequence-specific transactivation of the Drosophila twist gene by the dorsal gene product. *Cell* 65, 1191-201 (1991).
17. Newburger, D.E. & Bulyk, M.L. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res* (2008).
18. Quail, M.A., Swerdlow, H. & Turner, D.J. Improved protocols for the illumina genome analyzer sequencing system. *Current Protocols in Human Genetics* Chapter 18, Unit 18 2 (2009).
19. Hoffmann, S. *et al.* Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Computational Biology* 5, e1000502 (2009).

Nature Methods: doi:10.1038/nmeth.2558

20. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10, R25 (2009).
21. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754-60 (2009).
22. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research* 21, 936-9 (2011).
23. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25, 1966-7 (2009).
24. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* 18, 1851-8 (2008).
25. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* 215, 403-10 (1990).
26. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* 11, R106 (2010).
27. Zhu, L.J. *et al.* FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Research* 39, D111-7 (2011).
28. Busser, B.W. *et al.* Molecular mechanism underlying the regulatory specificity of a Drosophila homeodomain protein that specifies myoblast identity. *Development* 139, 1164-74 (2012).
29. 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. U.S.A.* 109, 20768-73 (2012).
30. Kheradpour, P., Stark, A., Roy, S. & Kellis, M. Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Res* 17, 1919-31 (2007).
31. Tomancak, P. *et al.* Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biology* 3, RESEARCH0088 (2002).
32. Tomancak, P. *et al.* Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biology* 8, R145 (2007).
33. Aboukhalil, A. & Bulyk, M.L. LOESS correction for length variation in gene set-based genomic sequence analysis. *Bioinformatics* 28, 1446-54 (2012).
34. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825-2830 (2011).
35. Yuan, Y., Guo, L., Shen, L. & Liu, J.S. Predicting gene expression from sequence: a reexamination. *PLoS Computational Biology* 3, e243 (2007).
36. Zhang, H. The Optimality of Naive Bayes. in *American Association for Artificial Intelligence, FLAIRS* (2004).

Nature Methods: doi:10.1038/nmeth.2558

Supplementary Note 3

We examined the eFS-identified enhancers for enrichment of previously described enhancer-associated chromatin marks, which were not used in the selection of cCRMs tested by eFS. Comparison to data obtained by batch isolation of tissue-specific chromatin for immunoprecipitation (BiTS-ChIP) for mesodermal cells from stages 10-11 during *Drosophila* embryonic development [1] showed that acetylation of histone H3 on lysine 27 (H3K27ac) and monomethylation of histone H3 on lysine 4 (H3K4me1), both of which previously have been found to be associated with active enhancers [2,3], are enriched (AUC = 0.730 and 0.781, $P = 1.17 \times 10^{-4}$ and 3.32×10^{-6}, respectively, by Wilcoxon-Mann-Whitney U-test) among the enhancers found to be active, as compared to inactive, in mesoderm by eFS (Fig. 4a). Consistent with the BiTS-ChIP data of Bonn *et al.*, who found Pol II occupancy to be tightly correlated with the timing and location of TF binding at active enhancers [1], we found Pol II to be enriched (AUC = 0.647, P =0.0033) among the eFS-identified, active mesodermal enhancers. However, in contrast to the finding by Bonn *et al.* that H3K27 trimethylation (H3K27me3) was depleted among 22 active mesodermal enhancers [1], we found H3K27me3 to be slightly enriched (AUC =0.647, P = 0.0093) among the enhancers identified by eFS as being active in mesodermal cells (Fig. 4a); this enrichment is unlikely to be due to eFS artifacts, since we have individual validation data for 8 eFS-identified mesodermal enhancers with positive ChIP signal for H3K27me3 and found 6 out of 8 to be active mesodermal enhancers. It is possible that H3K27me3 might be enriched in a different subset of mesodermal cells as those in which these enhancers are active. We found a slight enrichment of H3K79me3 (AUC = 0.612, P = 0.036) among the eFS-identified active mesodermal enhancers, which is consistent with reports by Bonn *et al.* that this mark occurs at approximately 15% of active distal enhancers [1]. We observed modest enrichment (AUC = 0.657, P = 0.006) of H3K4me3 among eFS-identified active mesodermal enhancers, which supports the association between H3K4me3 and enhancer activity [4].

As expected, overall there is a greater enrichment of DHSs identified from stage 5, 9, or 10 embryos, as compared to DHSs from later stages of development, among our enhancers, which were generated from stage 11-12 embryos (see Supplementary Figure 6). CBP occupancy was

least enriched among these various enhancer-associated genomic features; these results suggest that CBP does not play a primary role in enhancer activity at this stage of embryonic development, and are consistent with a recent report of p300-independent enhancers in human cells [5].

References

1. Bonn, S. *et al.* Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature Genetics* **44**, 148-56 (2012).
2. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-12 (2009).
3. Heintzman, N.D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-8 (2007).
4. Pekowska, A. *et al.* H3K4 tri-methylation provides an epigenetic signature of active enhancers. *EMBO J.* **30**, 4198-210 (2011).
5. Krebs, A.R., Karmodiya, K., Lindahl-Allen, M., Struhl, K. & Tora, L. SAGA and ATAC histone acetyl transferase complexes regulate distinct sets of genes and ATAC defines a class of p300-independent enhancers. *Molecular Cell* **44**, 410-23 (2011).

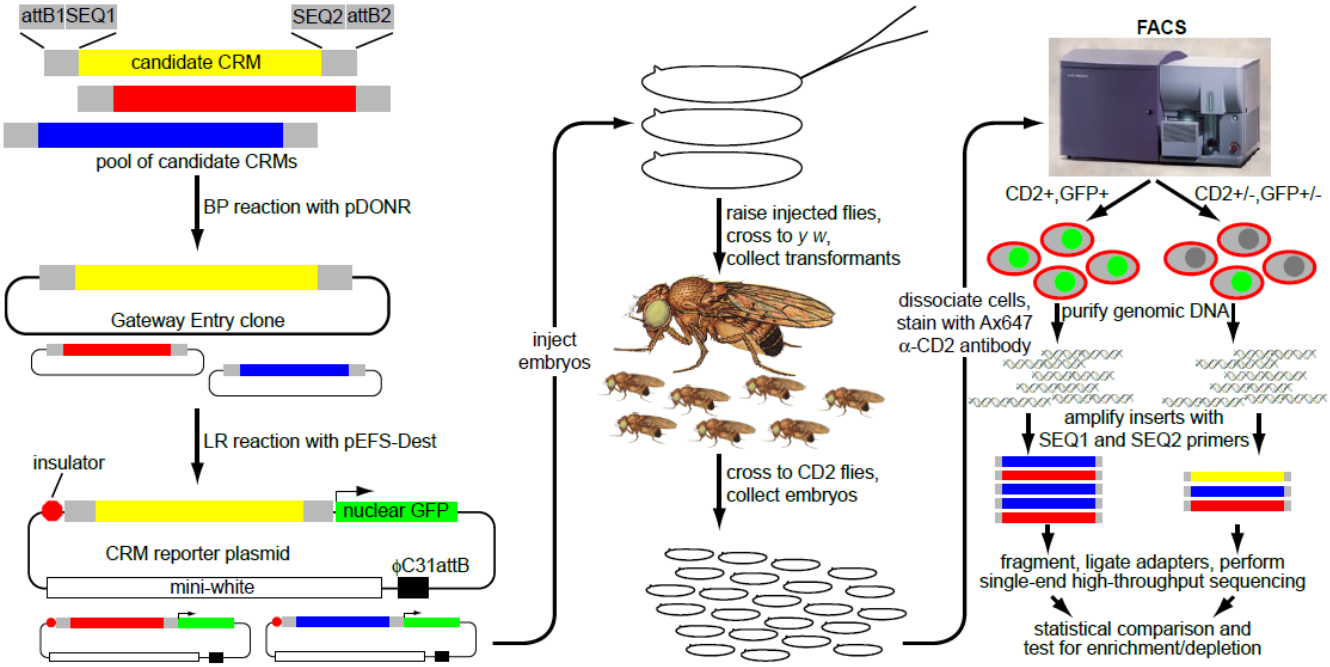
This document contains Supplementary Figures 1 to 9 published as part of the following *Nature Methods* paper:

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

Stephen S. Gisselbrecht[1], Luis A. Barrera[1,2,3], Martin Porsch[1,4], Anton Aboukhalil[1,5], Preston W. Estep 3[rd 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[1,2,3,10]

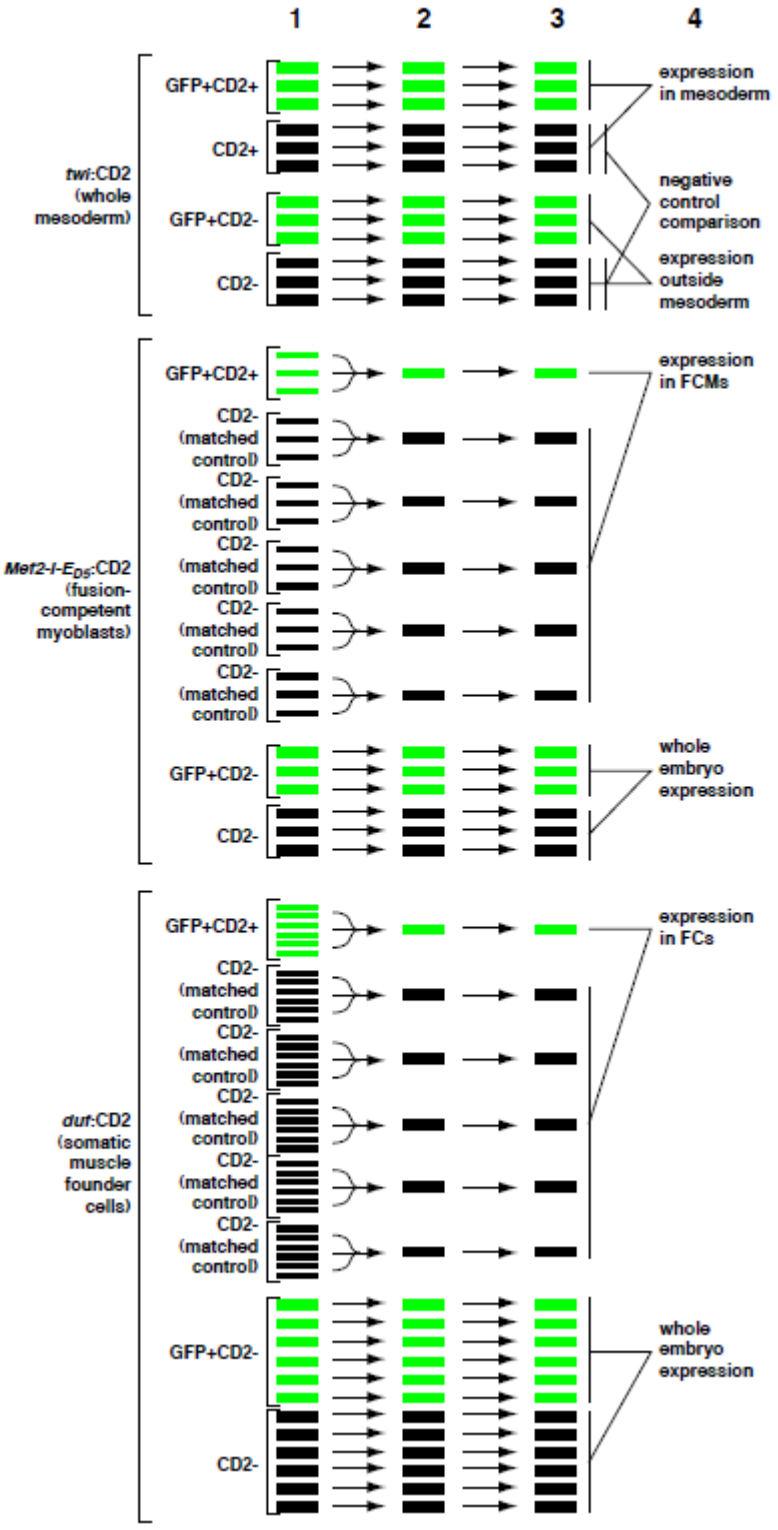
[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115.
[2]Harvard-MIT Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, MA 02115.
[3]Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, MA 02138.
[4]Institute of Computer Science, Martin Luther University of Halle-Wittenberg, 06099 Halle, Germany.
[5]Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139.
[6]TeloMe, Inc., Waltham, MA 02451.
[7]Bioengineering Department, Polytech Nice Sophia, University of Nice Sophia Antipolis, 06903, France.
[8]Laboratory of Developmental Systems Biology, Genetics and Developmental Biology Center, National Heart Lung and Blood Institute, National Institutes of Health, Bethesda, MD 20892.
[9]Systems Biology Graduate Program, Harvard University, Cambridge, MA 02138.
[10]Department of Pathology; Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115.

Correspondence should be addressed to M.L.B. (mlbulyk@receptor.med.harvard.edu).



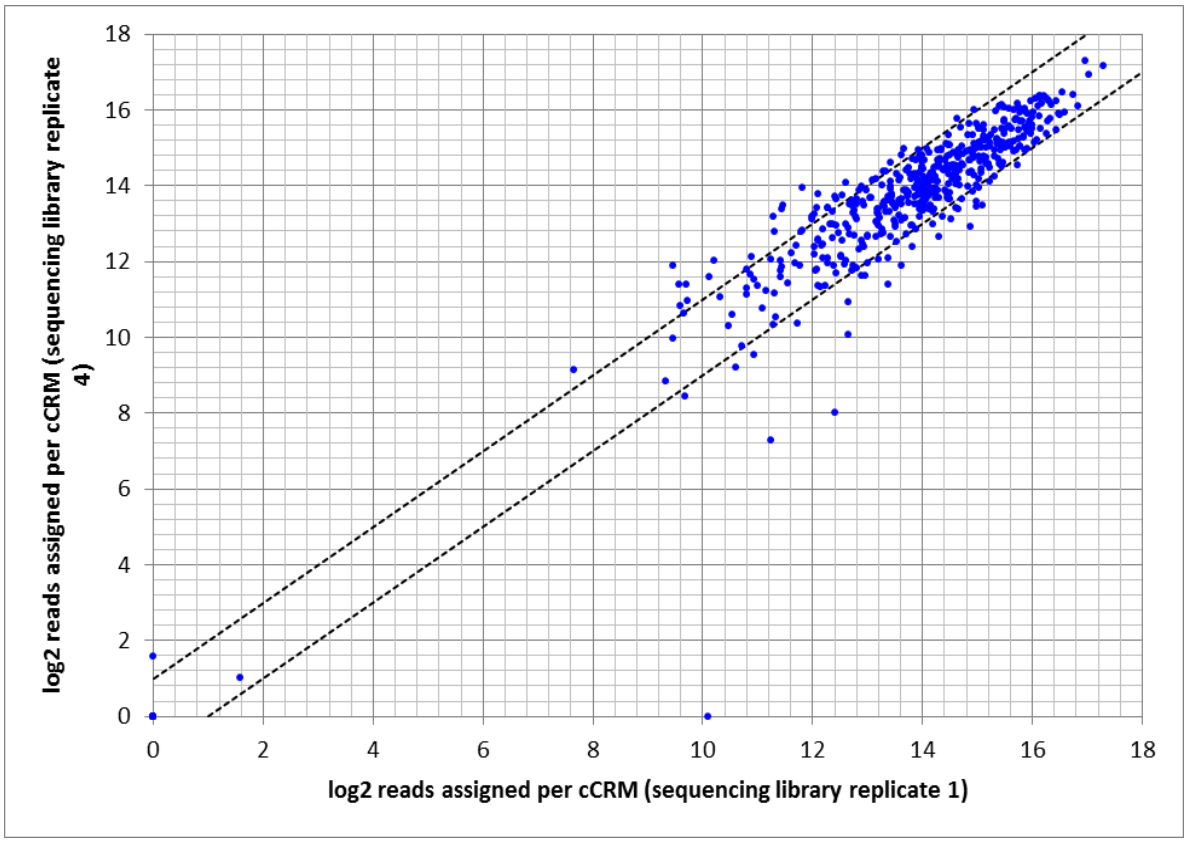
Supplementary Figure 1. Detailed eFS schema.

Nature Methods: doi:10.1038/nmeth.2558



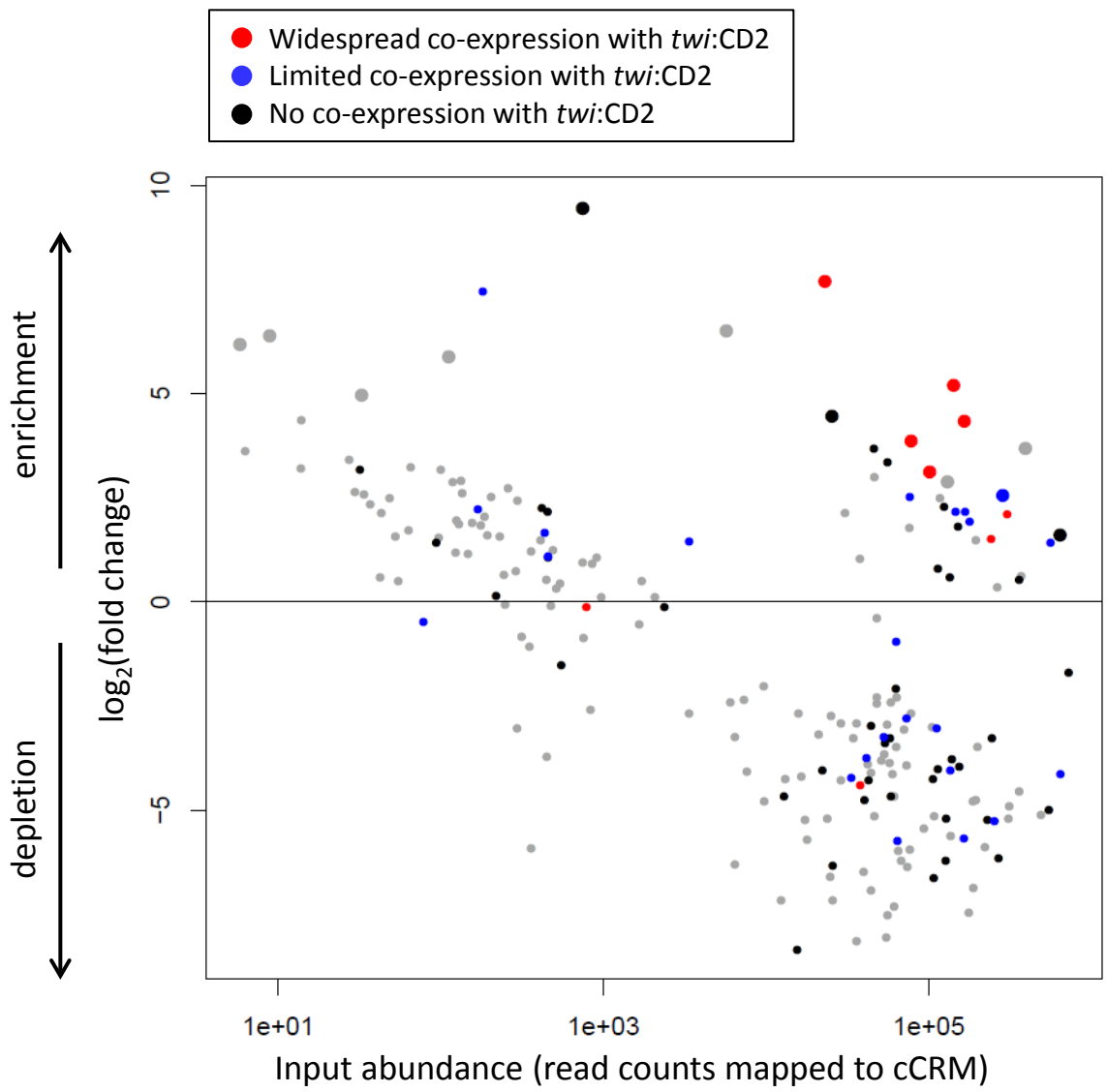
Supplementary Figure 2. Each population of cells is collected on three consecutive days (**1**). Green bars represent GFP+ populations, black bars input controls; thick bars represent cell numbers adequate for individual analysis (>10^4). Cell collections are pooled as necessary to produce populations for analysis (**2**) and processed by PCR, library prep, and high-throughput sequencing (**3**). Read counts mapped to individual windows are then compared, in various combinations, with statistically significant enrichment representing evidence for enhancer activity (**4**).

Nature Methods: doi:10.1038/nmeth.2558



Supplementary Figure 3. Reproducibility of sample preparation and sequencing process. The plot shows $\log_2$ of the total reads assigned to each cCRM in two representative sequencing libraries. only 72 out of 491 cCRMs show a greater than 2-fold difference in read count (dashed lines)

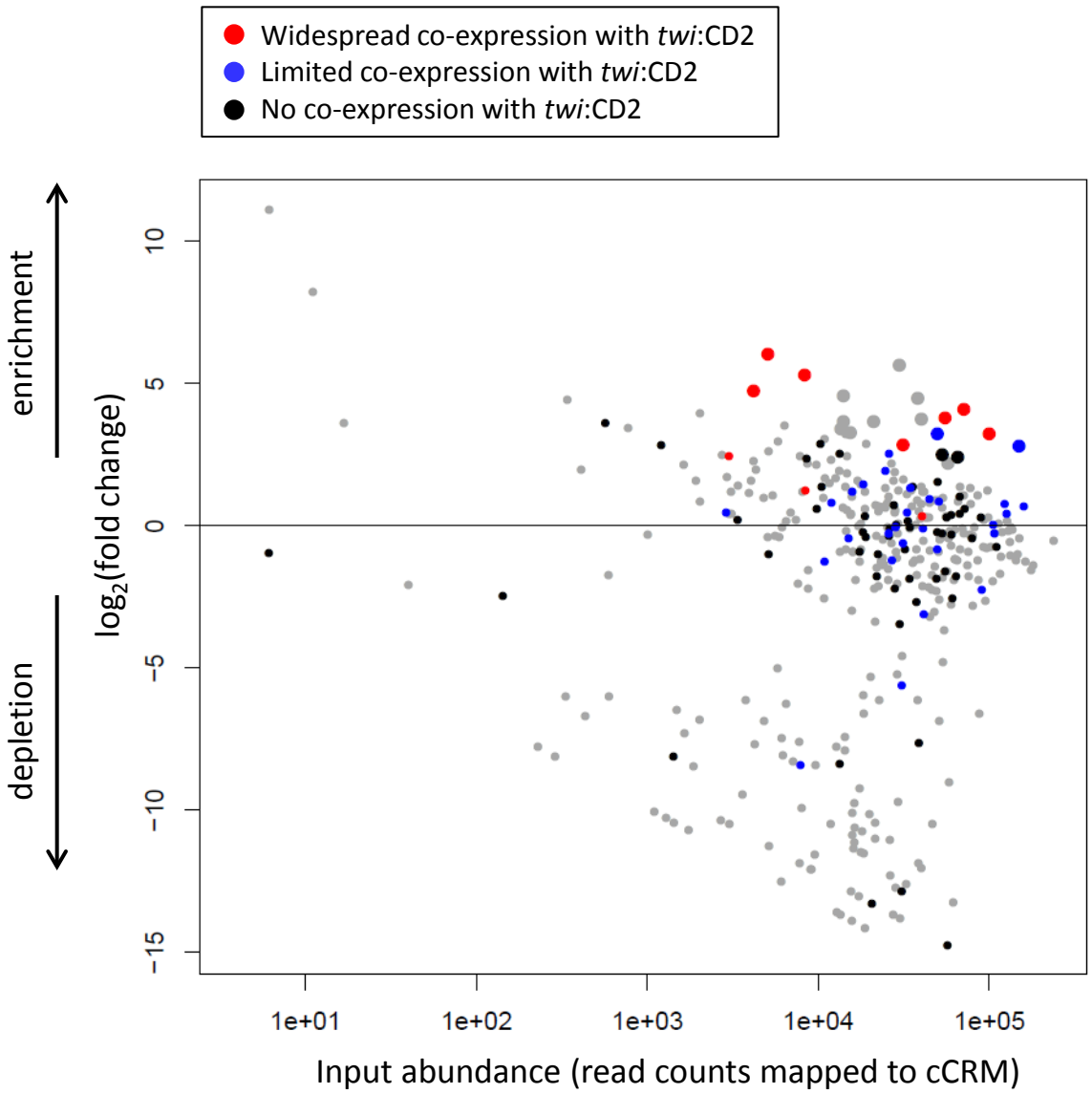
Nature Methods: doi:10.1038/nmeth.2558



Supplementary Figure 4. Enrichment ratios for cCRMs in **(a)** *Mef2-I-E_{D5}*:CD2$^+$GFP$^+$ cells as compared to *Mef2-I-E_{D5}*:CD2$^-$ Input cells, and **(b)** *(next page) duf*:CD2$^+$GFP$^+$ cells, as compared to *duf*:CD2$^-$ Input cells. *Large points:* significantly enriched (*P_adj* < 0.1), *small points: P_adj* > 0.1. Results from traditional reporter assays revealed cCRMs whose GFP expression shows widespread *(red)*, limited *(blue)*, or no *(black)* mesodermal (*twi*:CD2$^+$) expression.

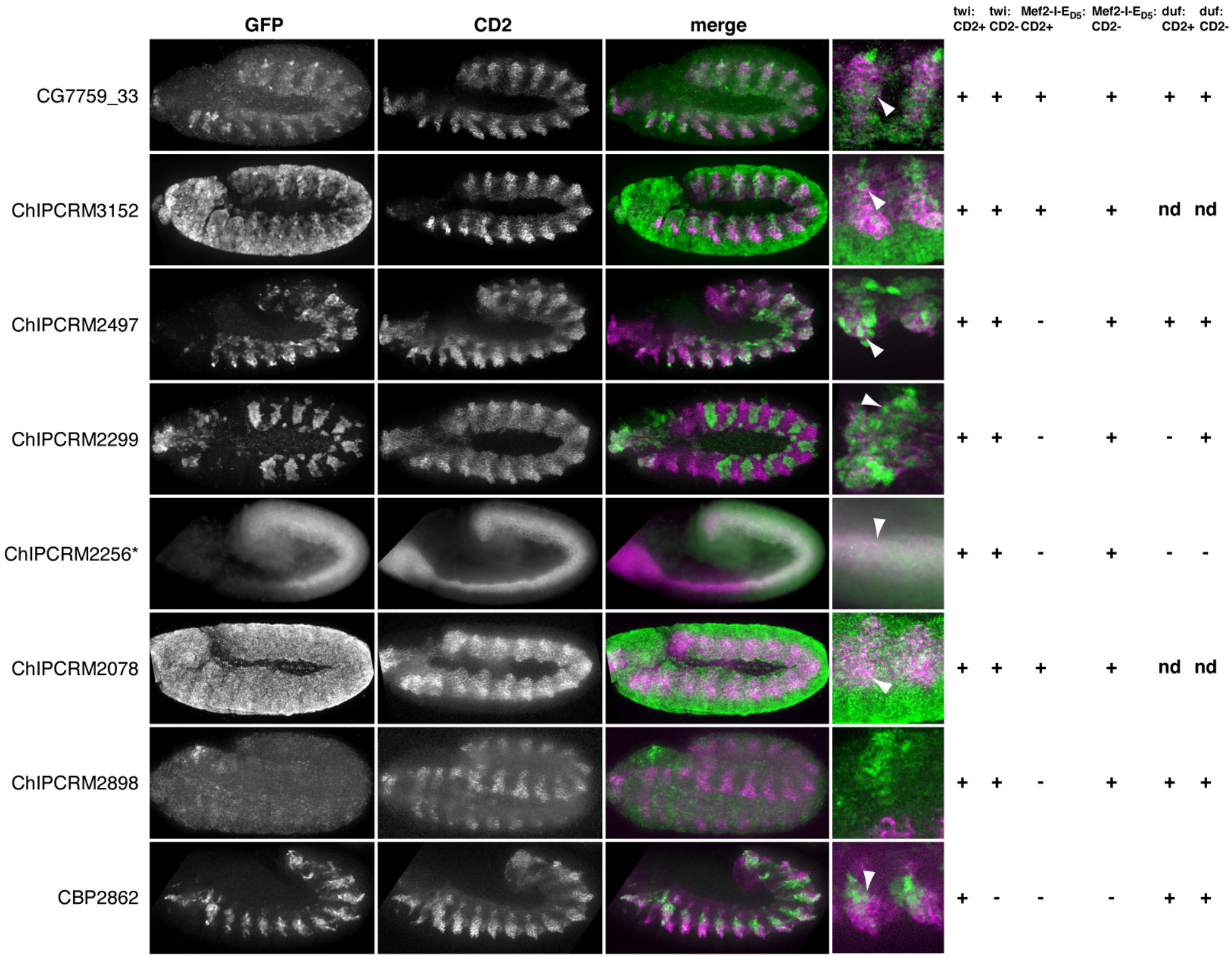
Nature Methods: doi:10.1038/nmeth.2558

(b) cCRM enrichment in *duf*:CD2^{+}GFP^{+} cells

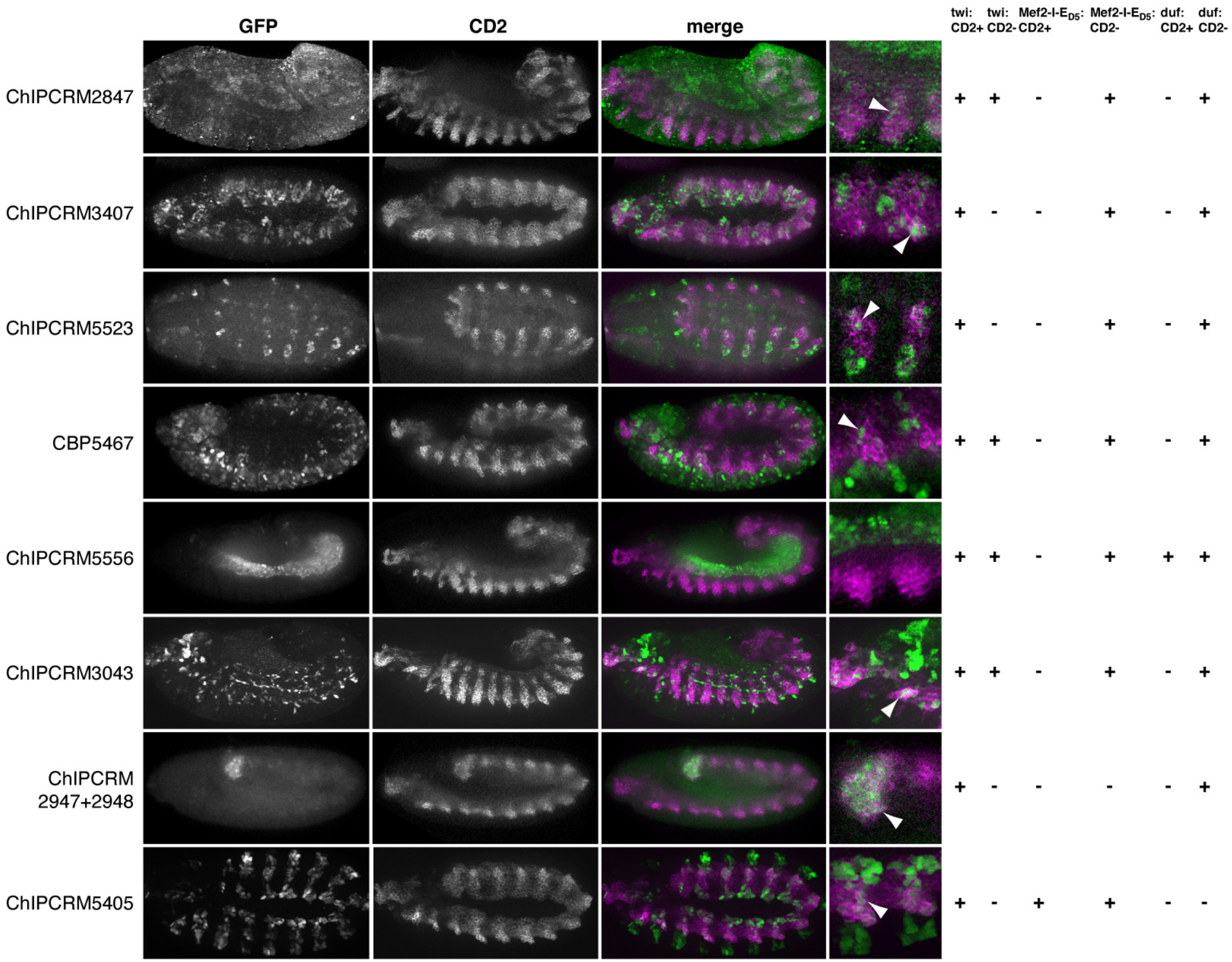


Nature Methods: doi:10.1038/nmeth.2558

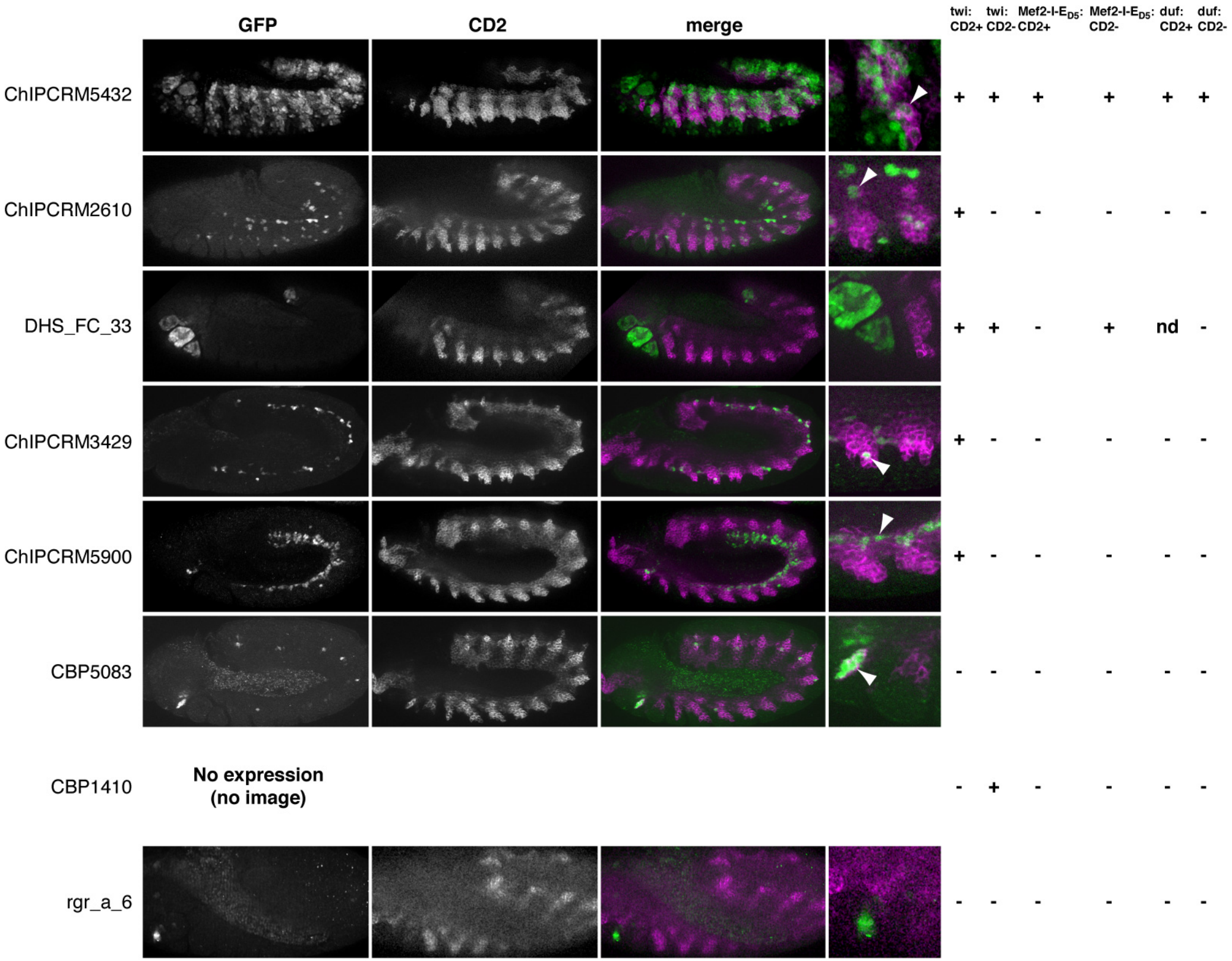
Supplementary Figure 5. Embryo images from individual validation of eFS findings.
Representative images are shown for each cCRM report line tested. Fixed embryos were stained with antibodies to GFP and CD2 (expressed pan-mesodermally under the control of the *twi* promoter) as described in Online Methods and Supplementary Note 2. Images displayed here are Maximum Intensity Projections of structured illumination optical sections except where noted by an asterisk. Expanded views of informative regions are shown to the right; arrowheads highlight a $GFP^{+}CD2$ coexpressing cell where such exist. Embryos shown are usually at or just before stage 11–12, except in the small minority of cases where no expression was visible at these stages and interesting patterned expression was seen in older embryos; in these cases (mid_30, cib_up_4, and nau_a_4) we display the later expression pattern. Tested cCRMs are displayed in the order in which they appear in Supplementary Table 4; complete descriptions of the pattern of detected GFP across embryogenesis in each line appear therein. 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 is observed as GFP and CD2 being present in the same cells, since the GFP in these embryos is nuclear, while CD2 is expressed on the cell surface (see Supplementary Note 2).



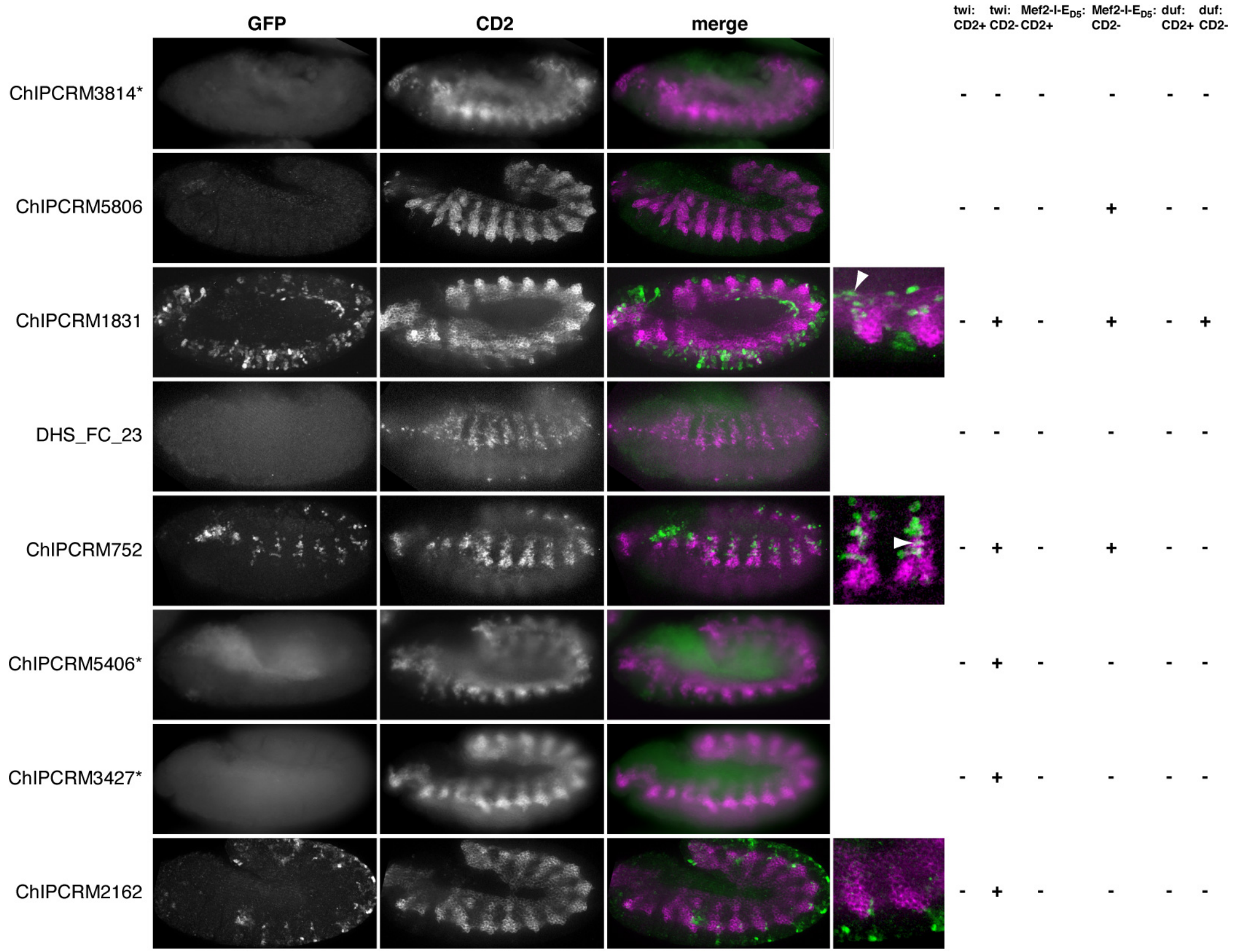
Nature Methods: doi:10.1038/nmeth.2558



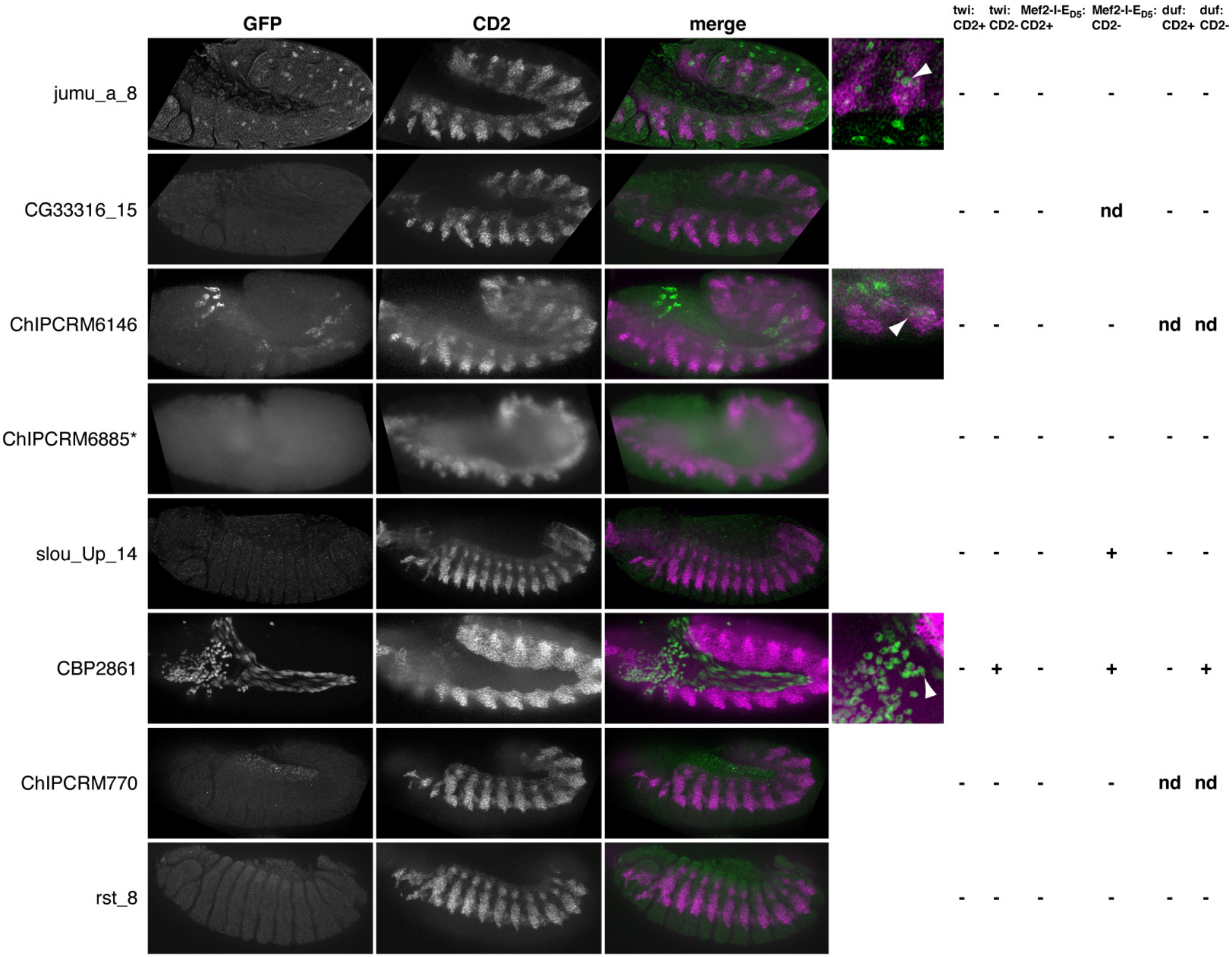
Nature Methods: doi:10.1038/nmeth.2558



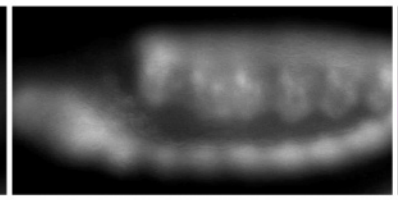
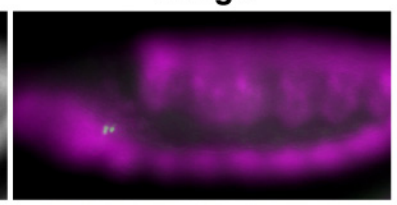
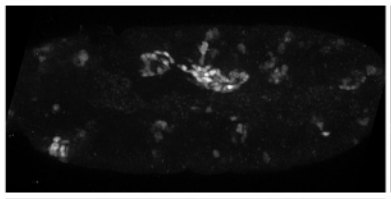
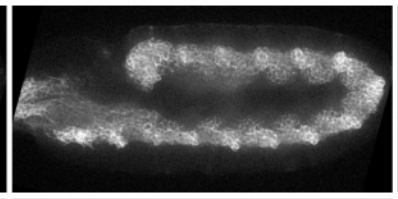
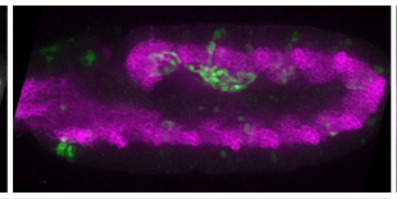
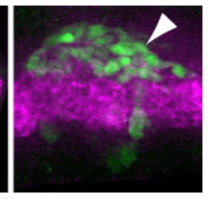
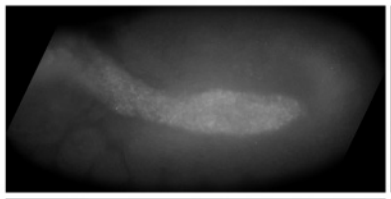
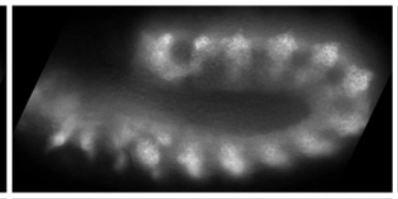
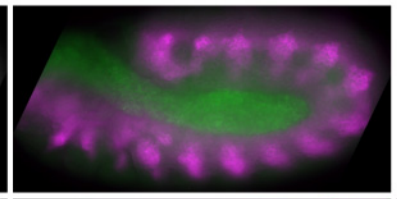
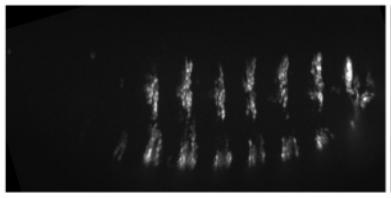
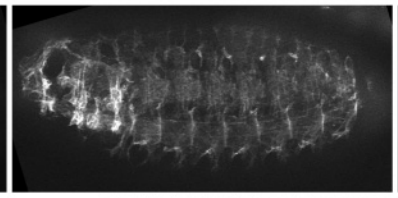
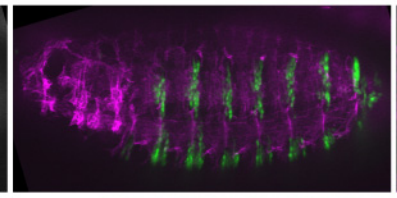
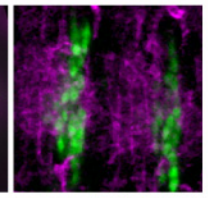
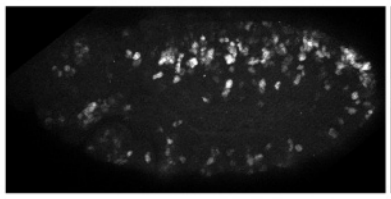
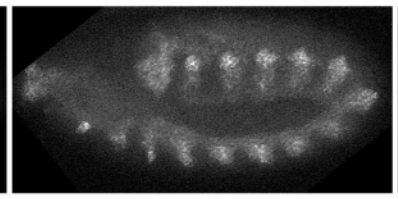
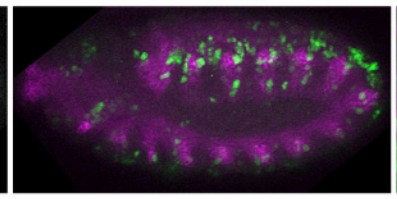
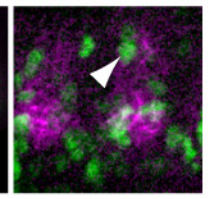
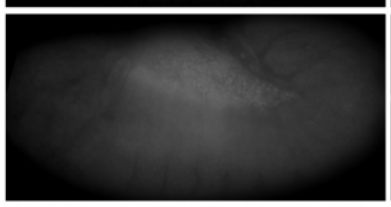
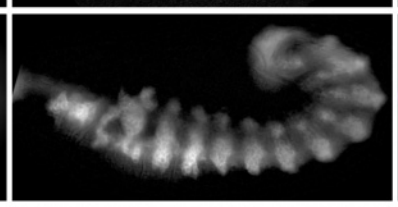
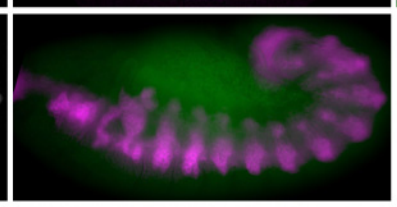
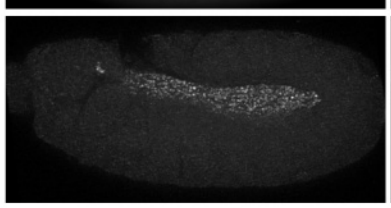
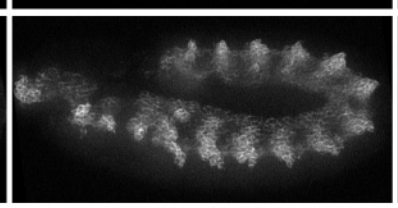
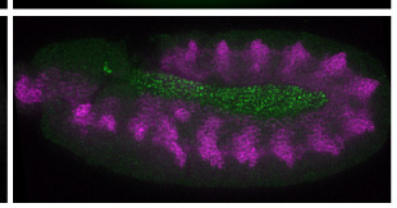
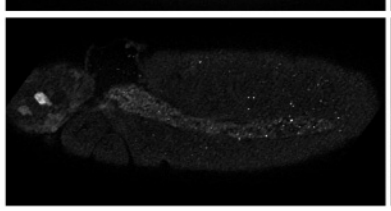
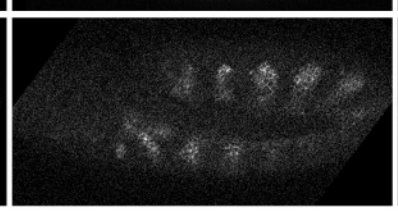
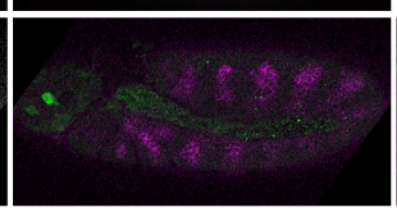
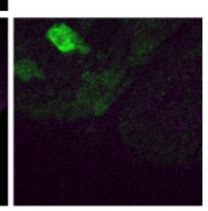
Nature Methods: doi:10.1038/nmeth.2558



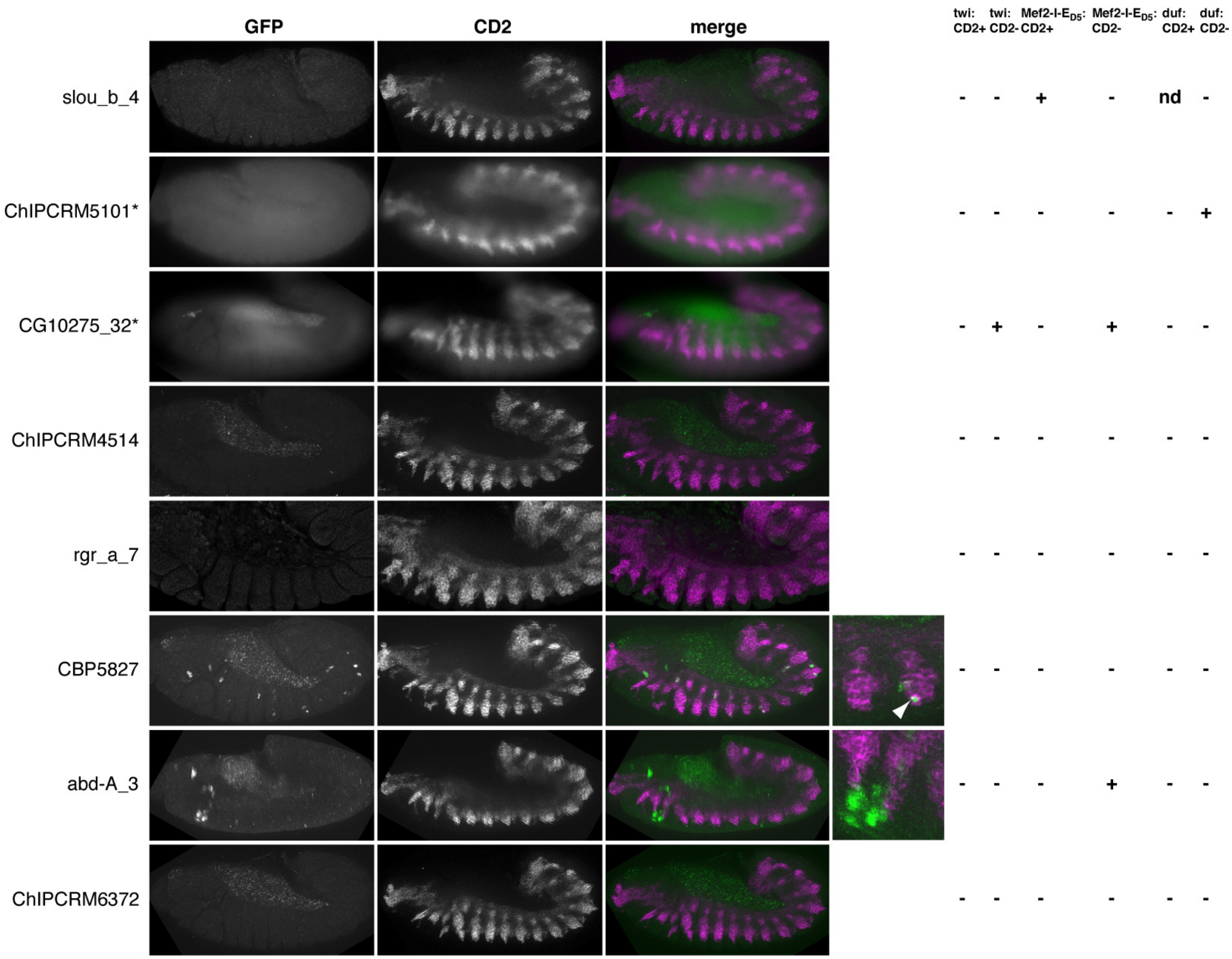
Nature Methods: doi:10.1038/nmeth.2558



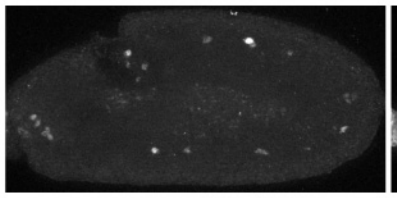
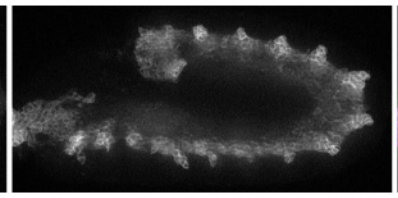
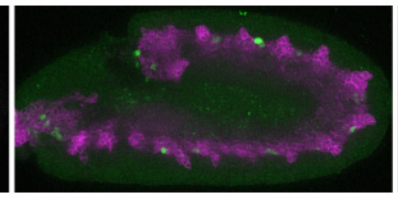
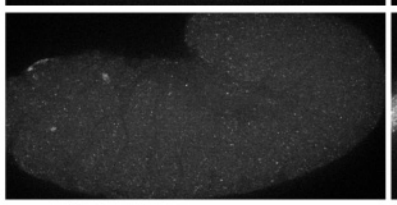
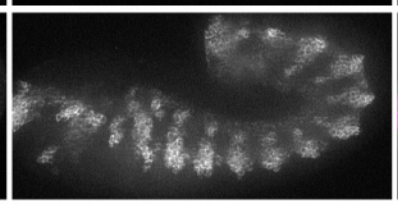
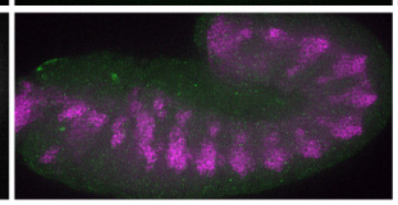
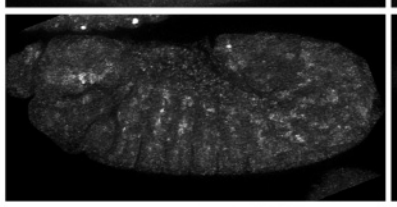
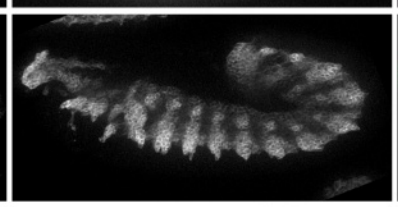
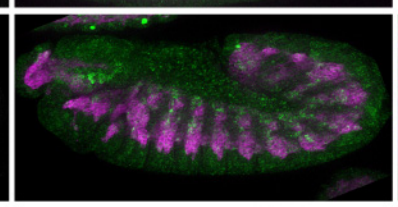
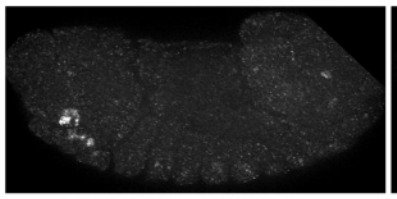
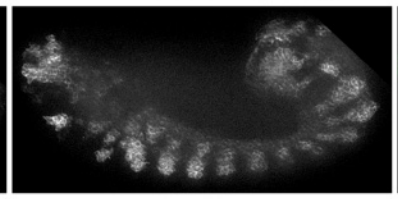
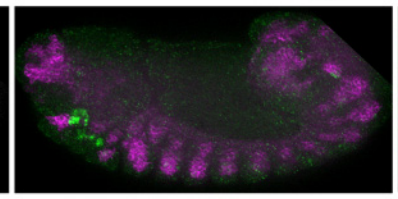
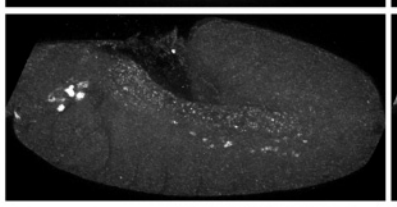
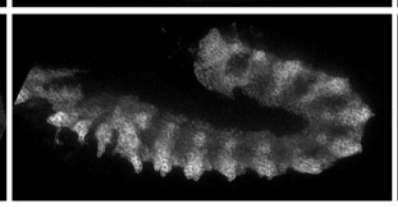
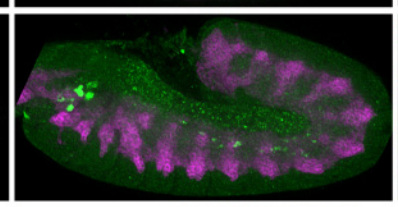
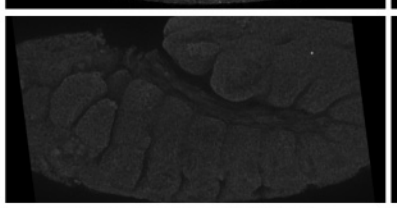
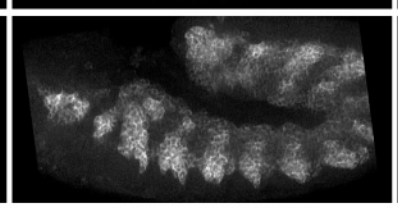
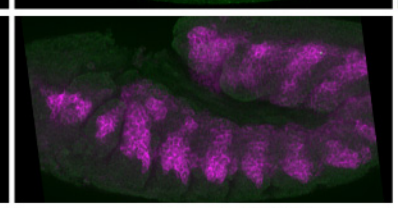
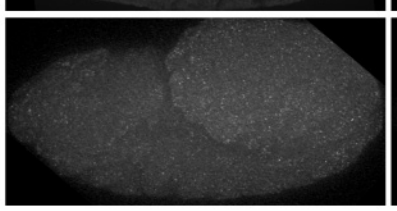
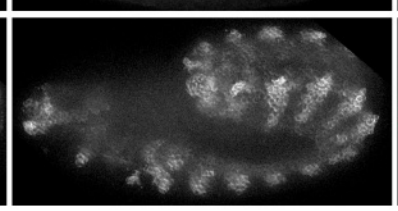
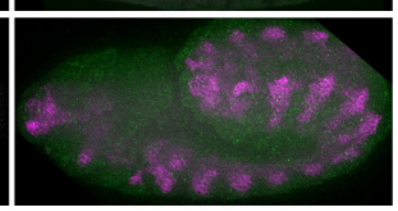
Nature Methods: doi:10.1038/nmeth.2558

	GFP	CD2	merge		twi: CD2+	twi: CD2-	Mef2-I-E_{D5}: CD2+	Mef2-I-E_{D5}: CD2-	duf: CD2+	duf: CD2-
eg_36*					-	-	+	-	-	-
ChIPCRM5561					-	+	-	+	nd	+
ChIPCRM5439*					-	-	-	+	-	-
mid_30					-	-	-	-	nd	nd
ChIPCRM4222					-	+	+	+	-	-
ChIPCRM3899*					-	+	-	+	-	-
ChIPCRM4235					-	-	-	+	-	-
slou_Up_30*					-	+	-	-	-	-

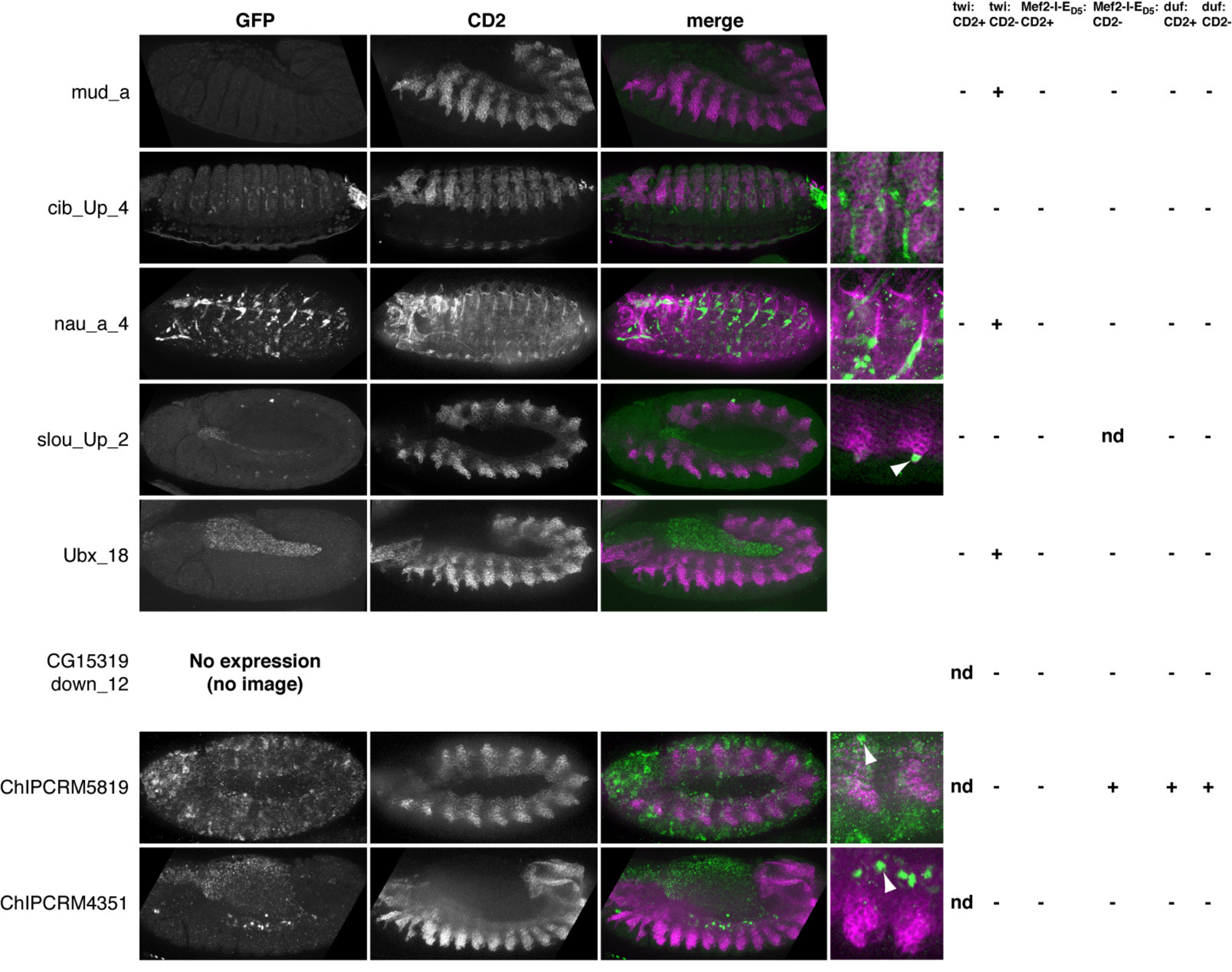
Nature Methods: doi:10.1038/nmeth.2558



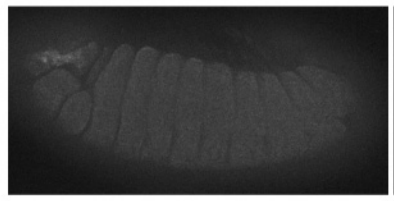
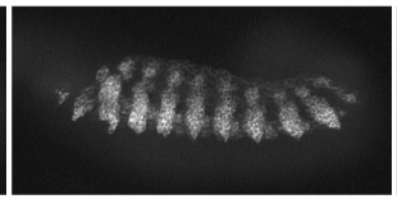
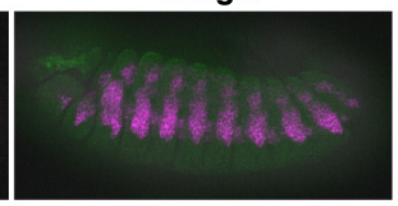
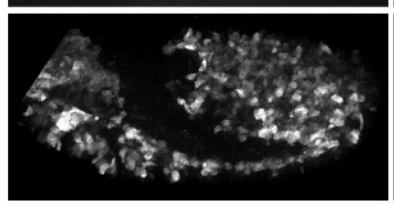
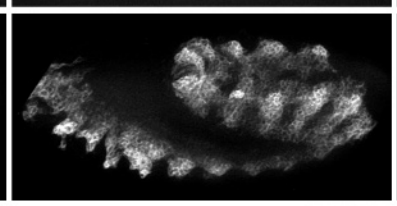
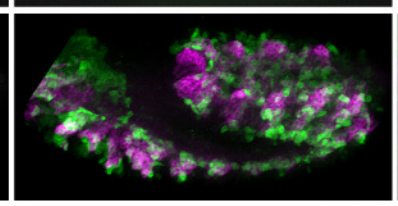
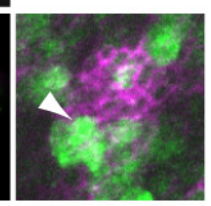
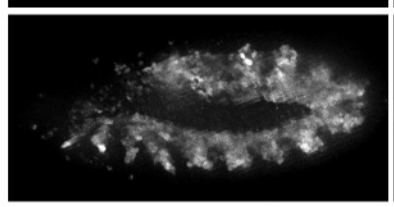
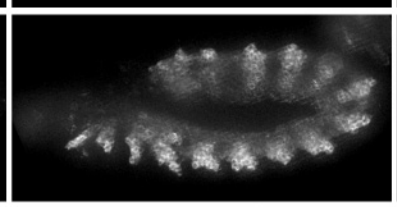
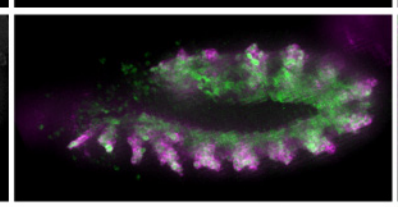
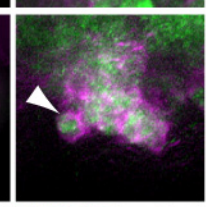
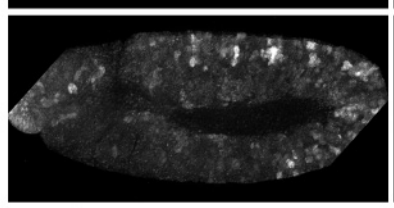
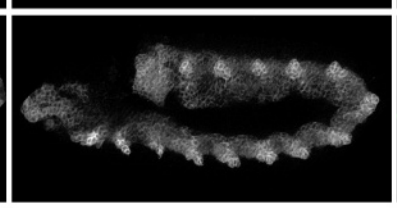
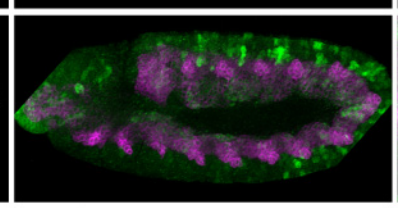
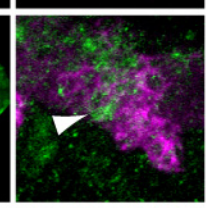
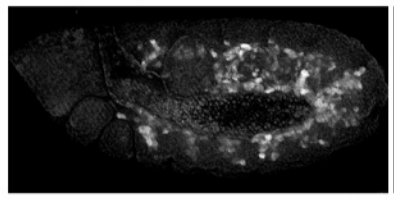
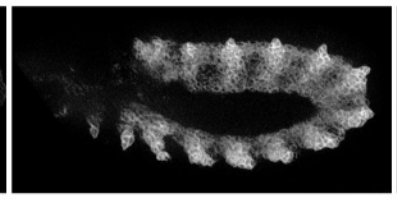
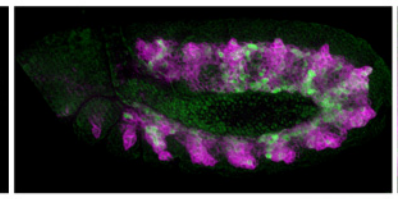
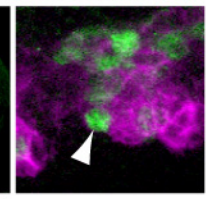
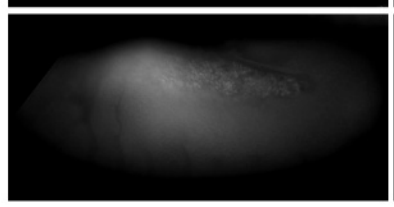
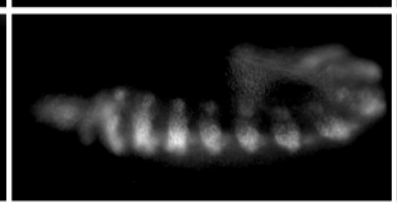
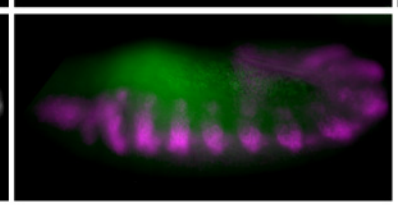
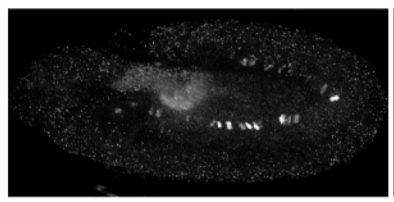
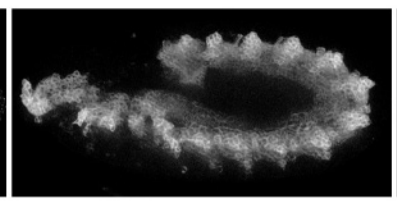
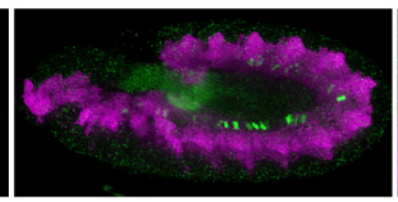
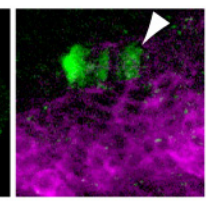
Nature Methods: doi:10.1038/nmeth.2558

	GFP	CD2	merge	twi: CD2+	twi: CD2-	Mef2-I-E_{D5}: CD2+	Mef2-I-E_{D5}: CD2-	duf: CD2+	duf: CD2-
DHS_FCM_1	No expression (no image)			-	-	-	-	-	-
ChIPCRM5397				-	-	-	-	-	-
ChIPCRM4236				-	-	-	+	-	-
ChIPCRM1739				-	-	-	-	-	-
DHS_FC_15				-	-	+	-	-	-
ChIPCRM6845				-	+	-	-	-	-
slou_Dn_9				-	-	-	-	-	-
CG15319 down_1				-	-	-	-	-	-

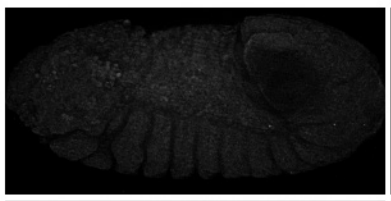
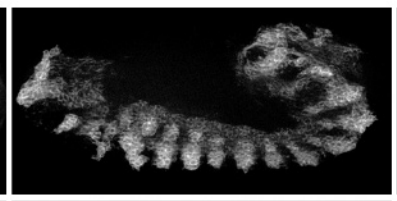
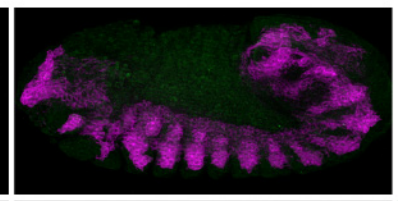
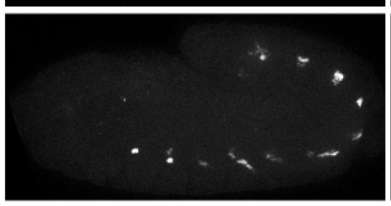
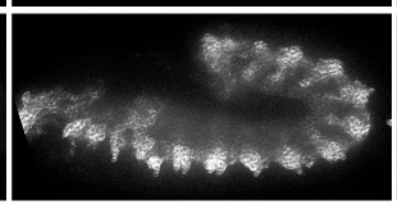
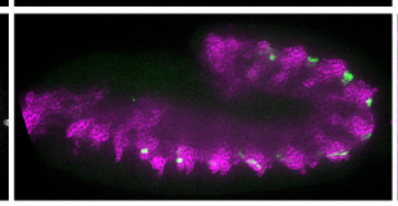
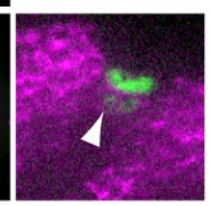
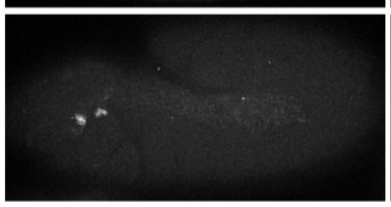
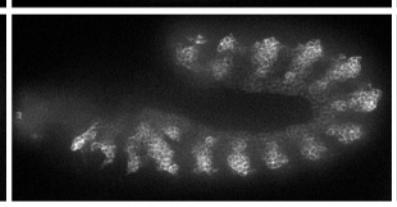
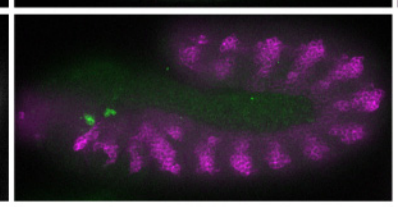
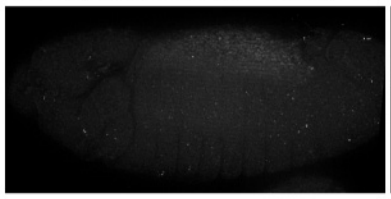
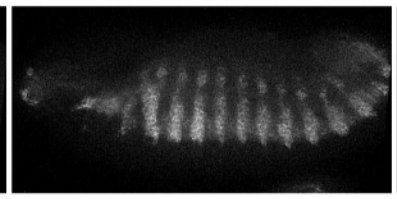
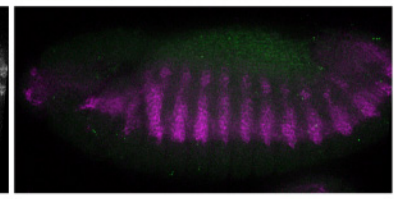
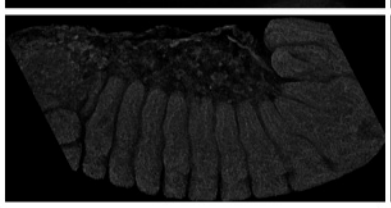
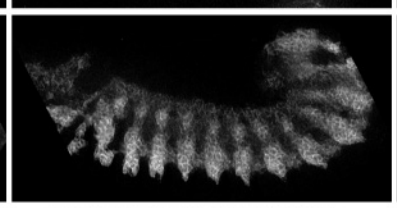
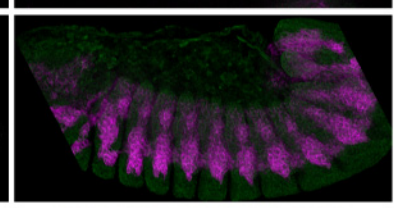
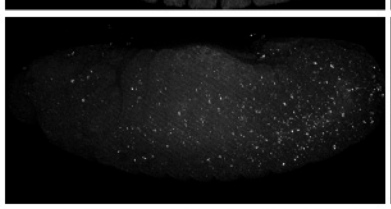
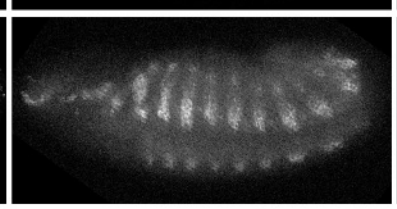
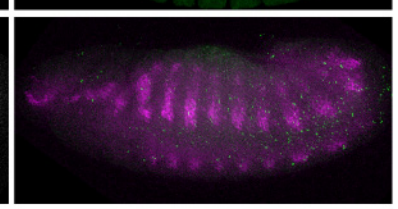
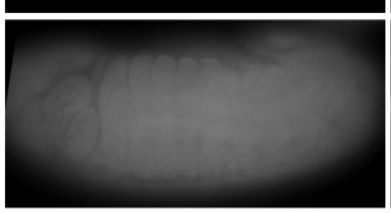
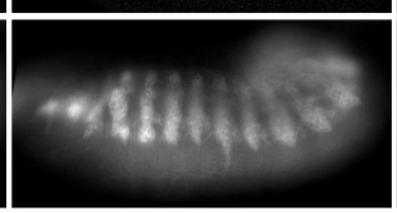
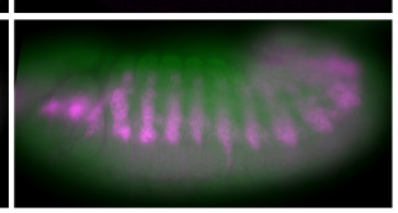
Nature Methods: doi:10.1038/nmeth.2558



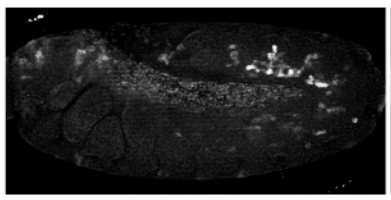
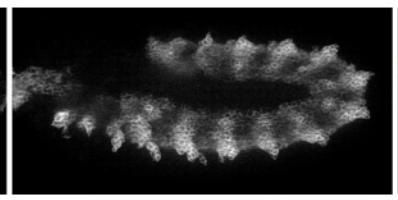
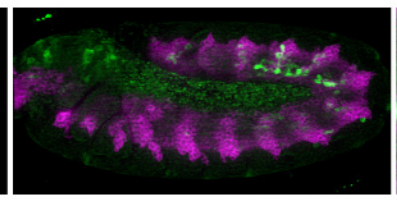
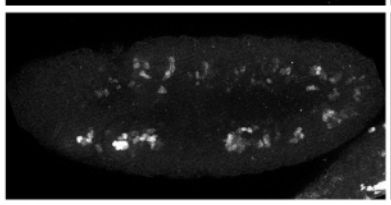
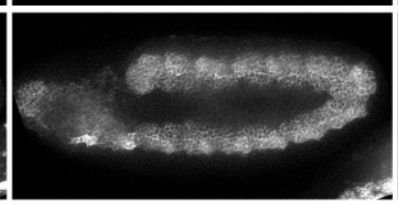
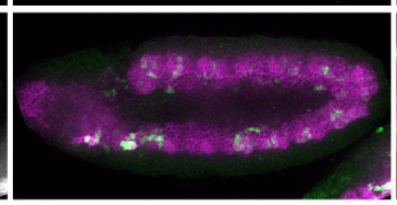
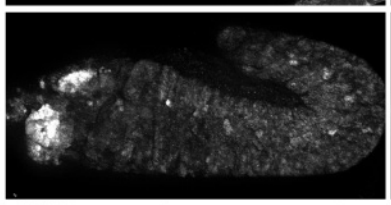
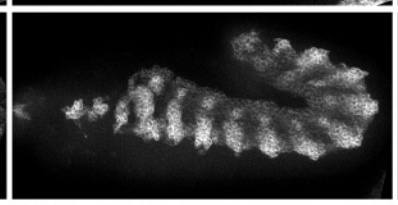
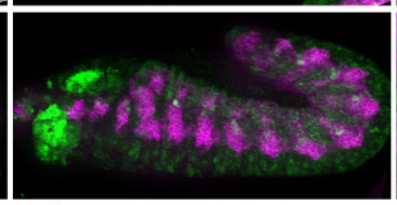
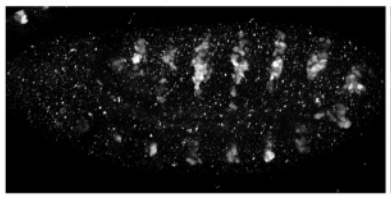
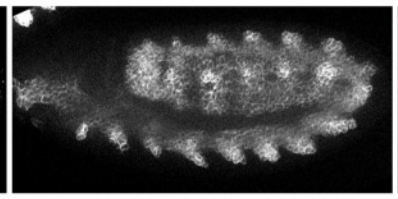
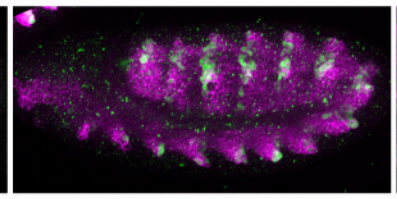
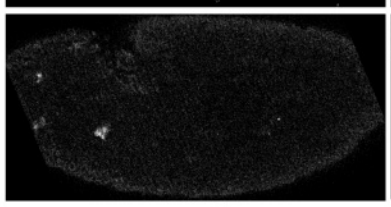
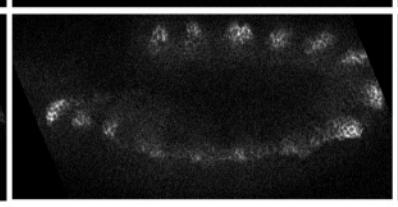
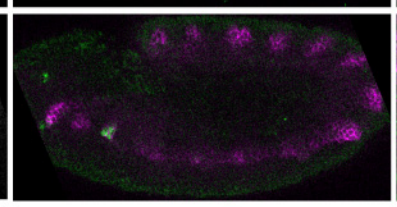
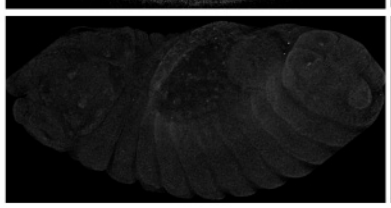
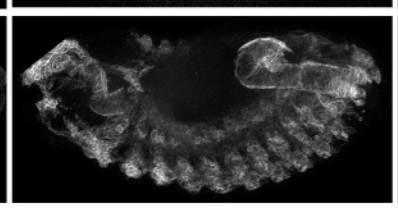
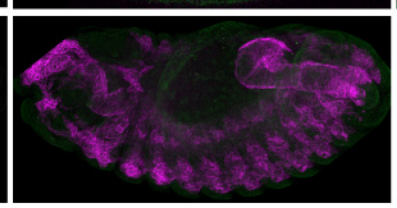
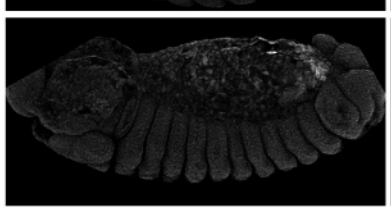
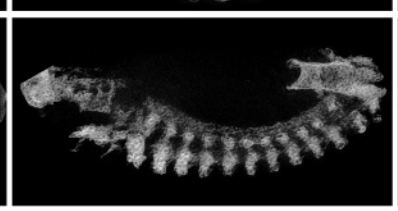
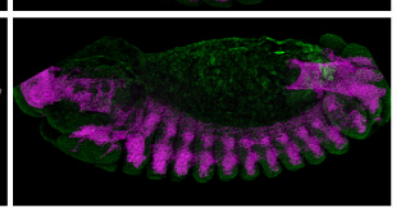
Nature Methods: doi:10.1038/nmeth.2558

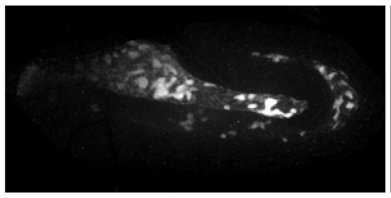
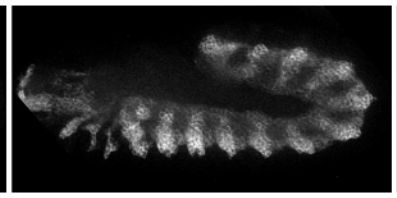
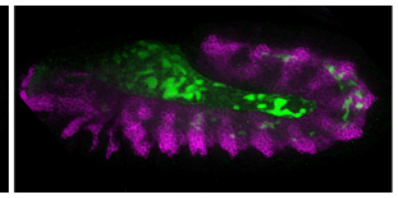
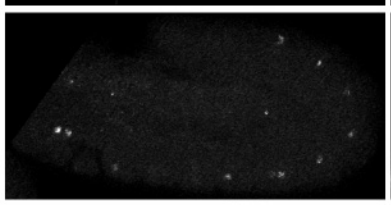
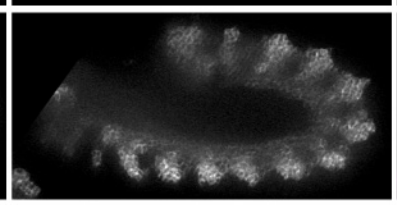
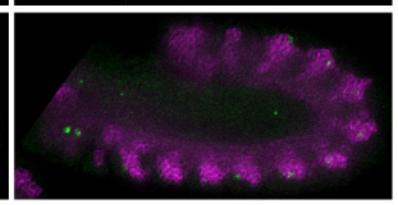
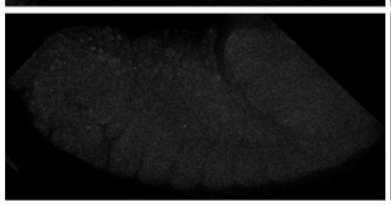
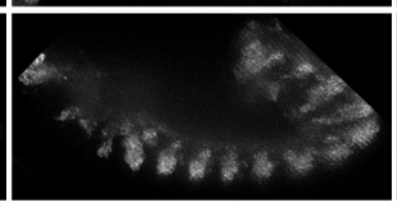
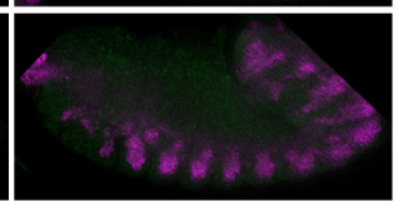
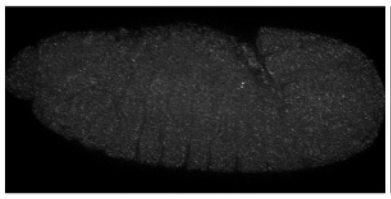
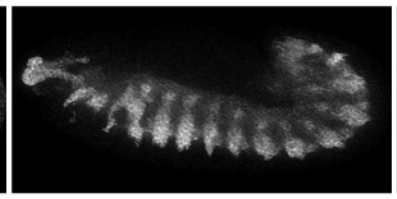
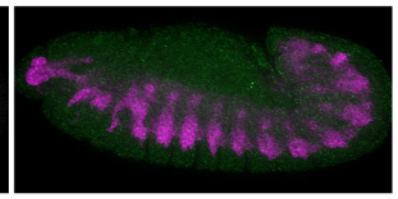
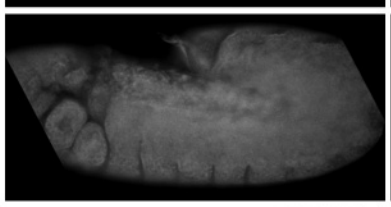
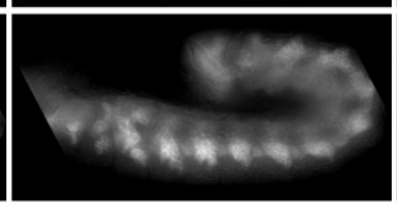
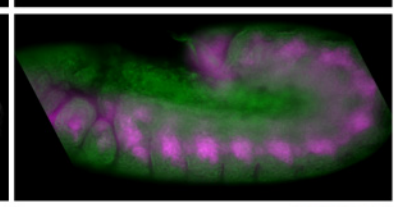
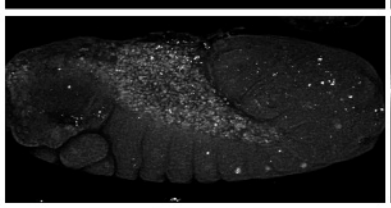
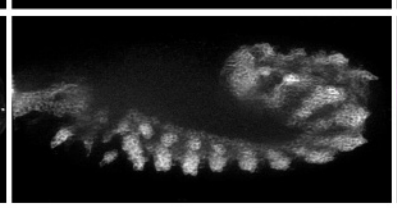
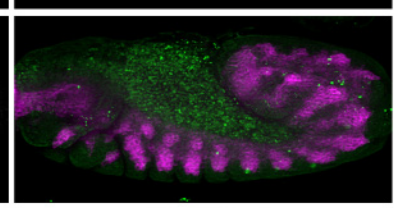
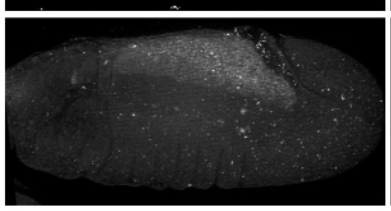
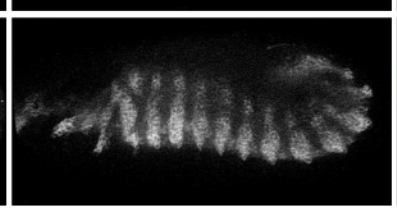
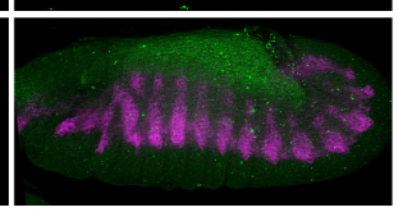
	GFP	CD2	merge		twi: CD2+	twi: CD2-	Mef2-I-E_{D5}: CD2+	Mef2-I-E_{D5}: CD2-	duf: CD2+	duf: CD2-
Sur_39					nd	-	-	nd	-	-
ChIPCRM5792					nd	nd	nd	nd	+	+
ChIPCRM2613					nd	nd	nd	nd	+	+
ChIPCRM2000					nd	nd	nd	nd	+	+
hth_3					nd	nd	nd	nd	+	+
mud_c_d*					nd	nd	nd	nd	-	-
numb_34	No expression (no image)				nd	nd	nd	nd	-	-
ChIPCRM6902					nd	nd	nd	nd	-	-

Nature Methods: doi:10.1038/nmeth.2558

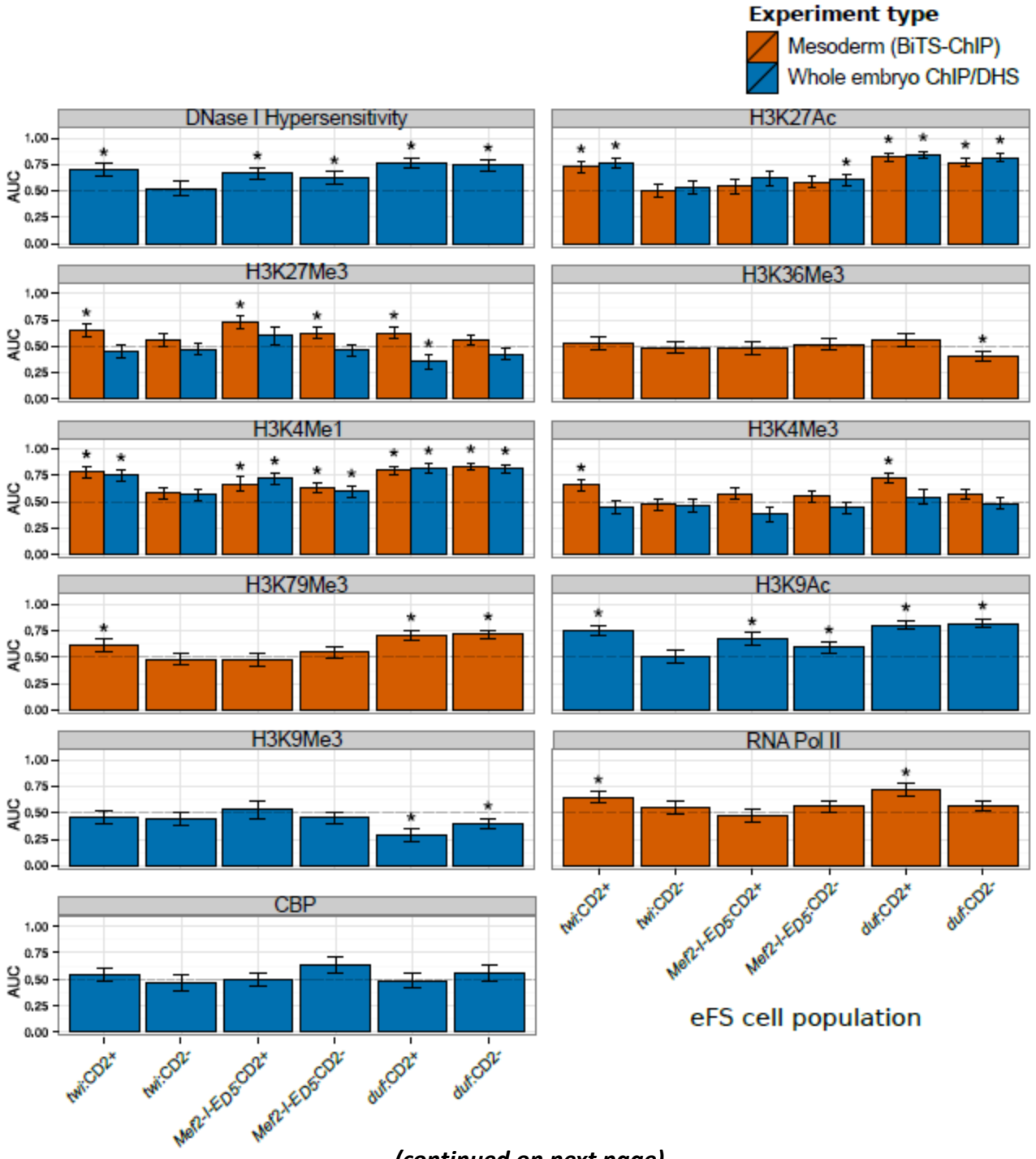
	GFP	CD2	merge	twi: CD2+	twi: CD2-	Mef2-I-E_{D5}: CD2+	Mef2-I-E_{D5}: CD2-	duf: CD2+	duf: CD2-
ChIPCRM4505	**No expression (no image)**			**nd**	**nd**	**nd**	**nd**	**-**	**-**
slou_Up_1				**nd**	**nd**	**nd**	**nd**	**-**	**-**
ChIPCRM3430			 	**nd**	**nd**	**nd**	**nd**	**-**	**-**
CG15319 down_7				**nd**	**nd**	**nd**	**nd**	**-**	**-**
Mef2_29				**nd**	**nd**	**nd**	**nd**	**-**	**-**
noc_6				**nd**	**nd**	**nd**	**nd**	**-**	**-**
slou_Dn_4				**nd**	**nd**	**nd**	**nd**	**-**	**-**
CBP4280				**nd**	**nd**	**nd**	**nd**	**-**	**-**

Nature Methods: doi:10.1038/nmeth.2558

	GFP	CD2	merge	twi: CD2+	twi: CD2-	Mef2-I-E_{D5}: CD2+	Mef2-I-E_{D5}: CD2-	duf: CD2+	duf: CD2-
CG13830_19	**No expression (no image)**			nd	nd	nd	nd	-	-
pnut_13				nd	nd	nd	nd	-	+
ChIPCRM6084				nd	nd	nd	nd	-	-
ChIPCRM5617				nd	nd	nd	nd	-	-
ChIPCRM5857				nd	nd	nd	nd	-	+
ChIPCRM2929*				nd	nd	nd	nd	-	-
sns_6				nd	nd	nd	nd	-	-
G-oalpha47A_5				nd	nd	nd	nd	-	-

	GFP	CD2	merge	twi: CD2+	twi: CD2-	Mef2-I-E_{D5}: CD2+	Mef2-I-E_{D5}: CD2-	duf: CD2+	duf: CD2-
slou_Up_5	**No expression (no image)**			**nd**	**nd**	**nd**	**nd**	**-**	**-**
ChIPCRM6679				**nd**	**nd**	**nd**	**nd**	**-**	**-**
ChIPCRM4231				**nd**	**nd**	**nd**	**nd**	**-**	**-**
TyrR_33				**nd**	**nd**	**nd**	**nd**	**-**	**-**
slou_Dn_1				**nd**	**nd**	**nd**	**nd**	**-**	**-**
CG2875_7*				**nd**	**nd**	**nd**	**nd**	**-**	**-**
mew_35				**nd**	**nd**	**nd**	**nd**	**-**	**-**
nau_Dn_2				**nd**	**nd**	**nd**	**nd**	**nd**	**-**

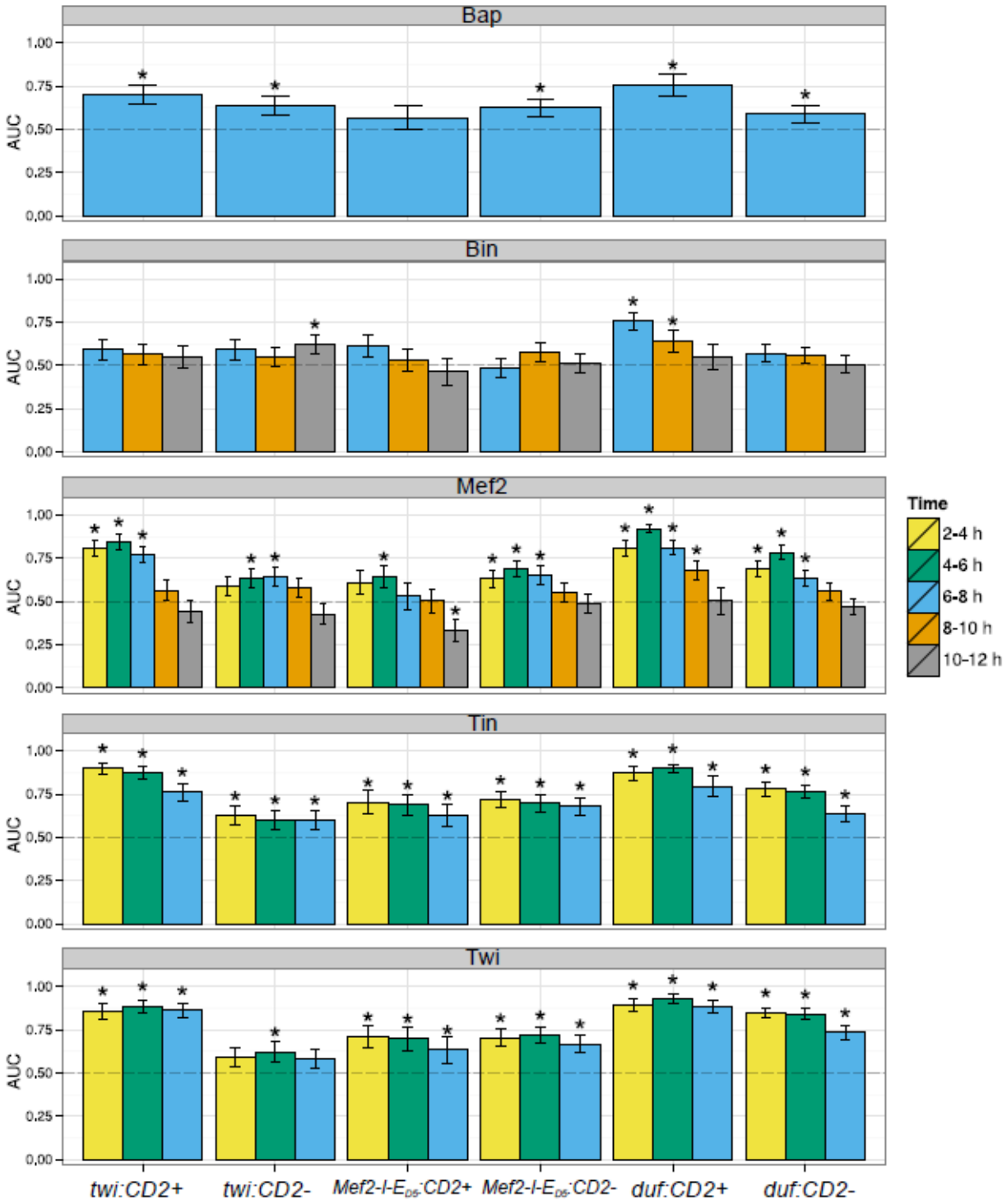
Nature Methods: doi:10.1038/nmeth.2558



(continued on next page)

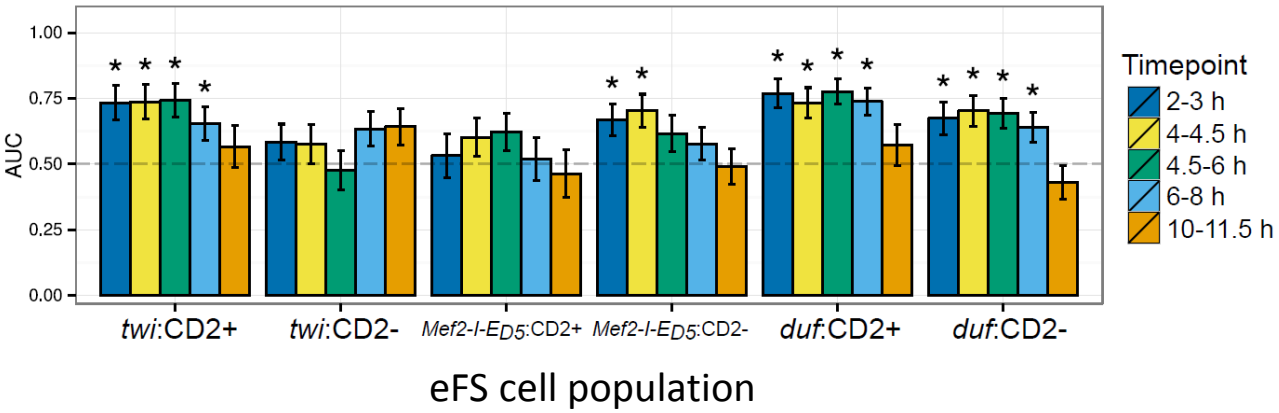
Supplementary Figure 6. Enrichment of various genomic features (*e.g.,* DHS, histone modifications, CBP binding, TF ChIP-binding) associated with active enhancers in whole mesoderm (*twi*:CD2$^+$), FCMs (*Mef2-I-E*$_{D5}$:CD2$^+$), FCs (*duf*:CD2$^+$) and each of the three corresponding CD2$^-$ cell populations. Genomic feature data were collected from either mesoderm by BITS-ChIP (Bonn *et al*., *Nature Genetics*, 2012) or whole embryos (*this page*, 4-8 hr embryos; *next page*, age of embryos ("Time") is indicated in color bar). AUC: area under receiver operating characteristic curve (see Online Methods). * indicates statistically significant ($p < 0.05$, Wilcoxon-Mann-Whitney U-test) enrichment (AUC > 0.5) or depletion (AUC < 0.5). Error bars indicate 1 standard deviation.

(continued from previous page)



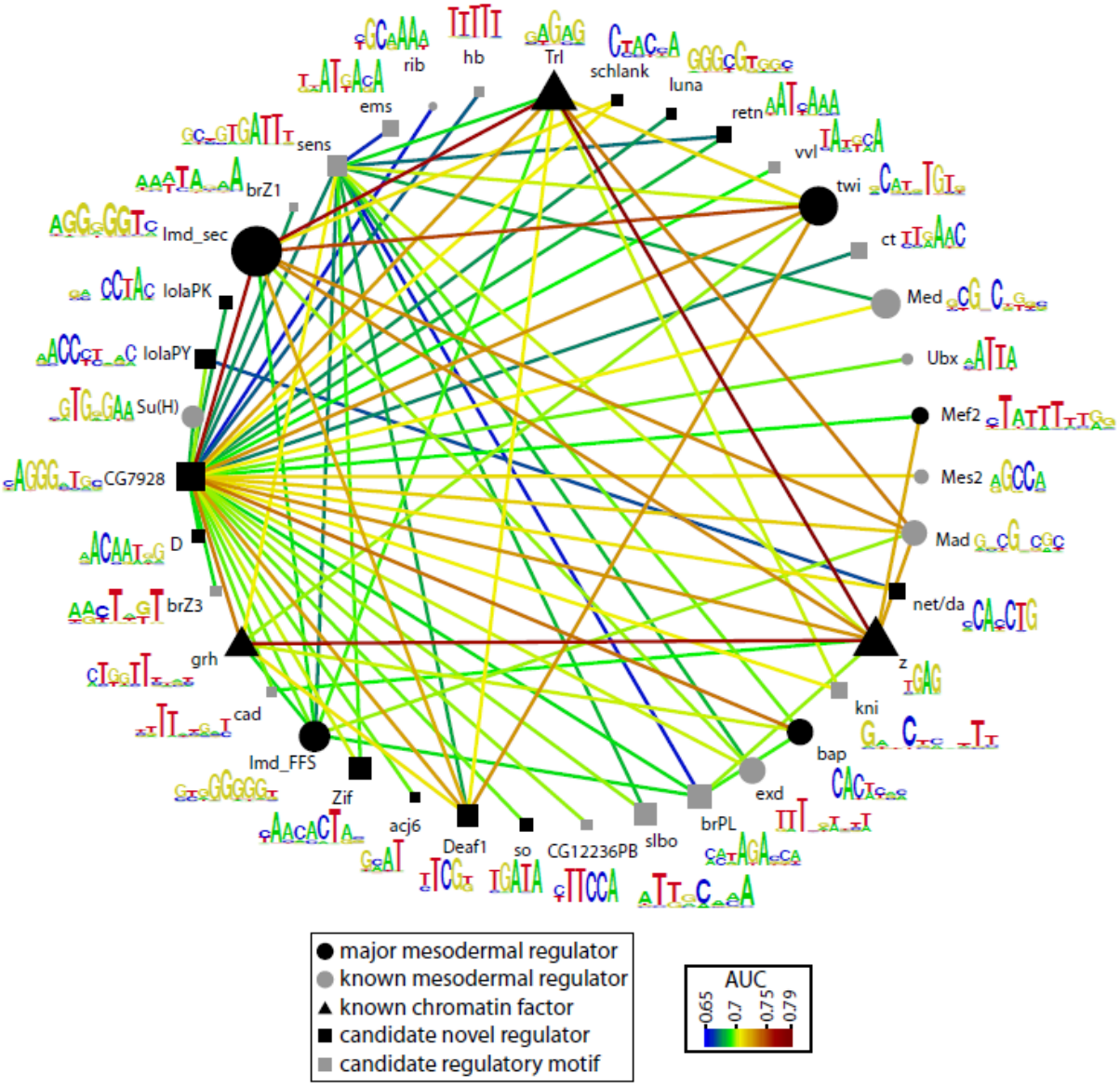
(continued on next page)

(continued from previous page)

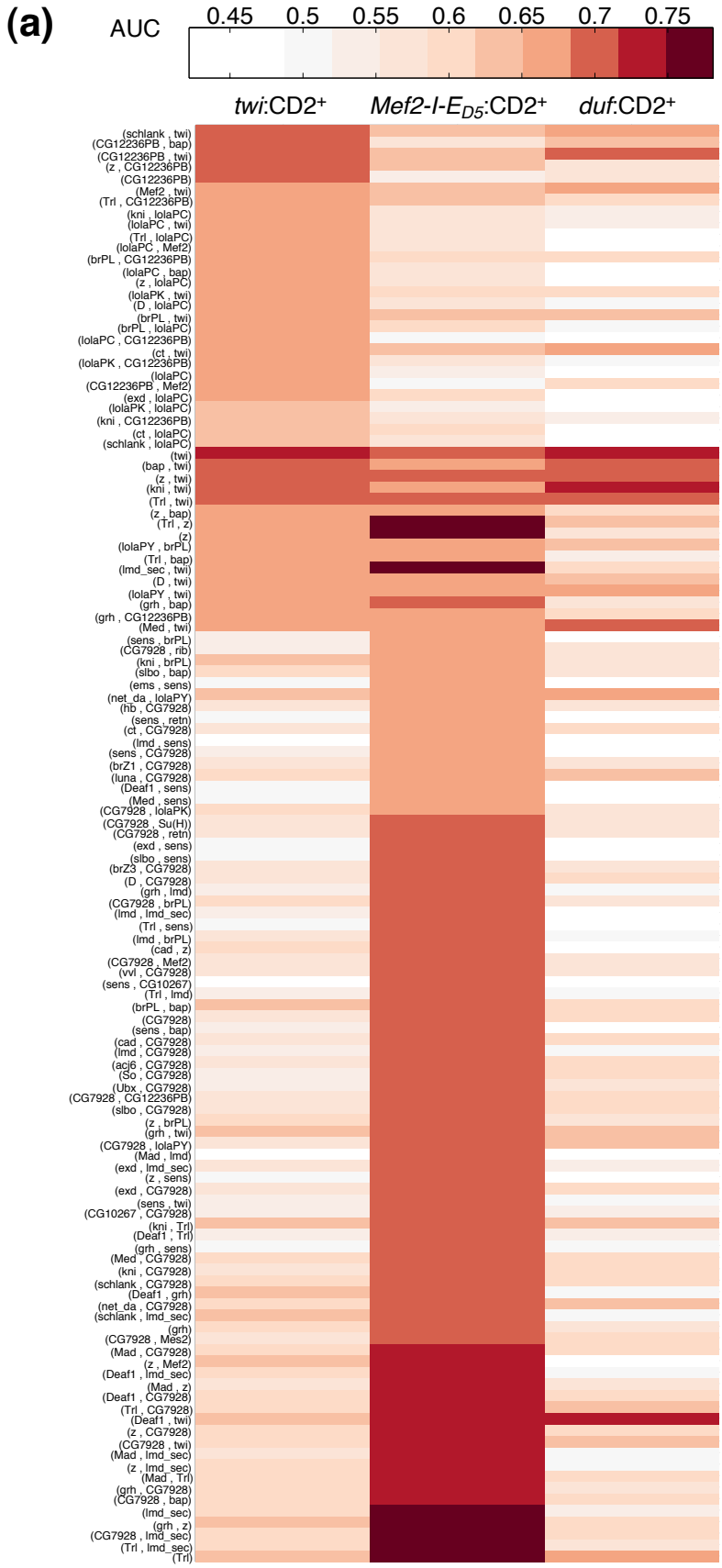


eFS cell population

Nature Methods: doi:10.1038/nmeth.2558

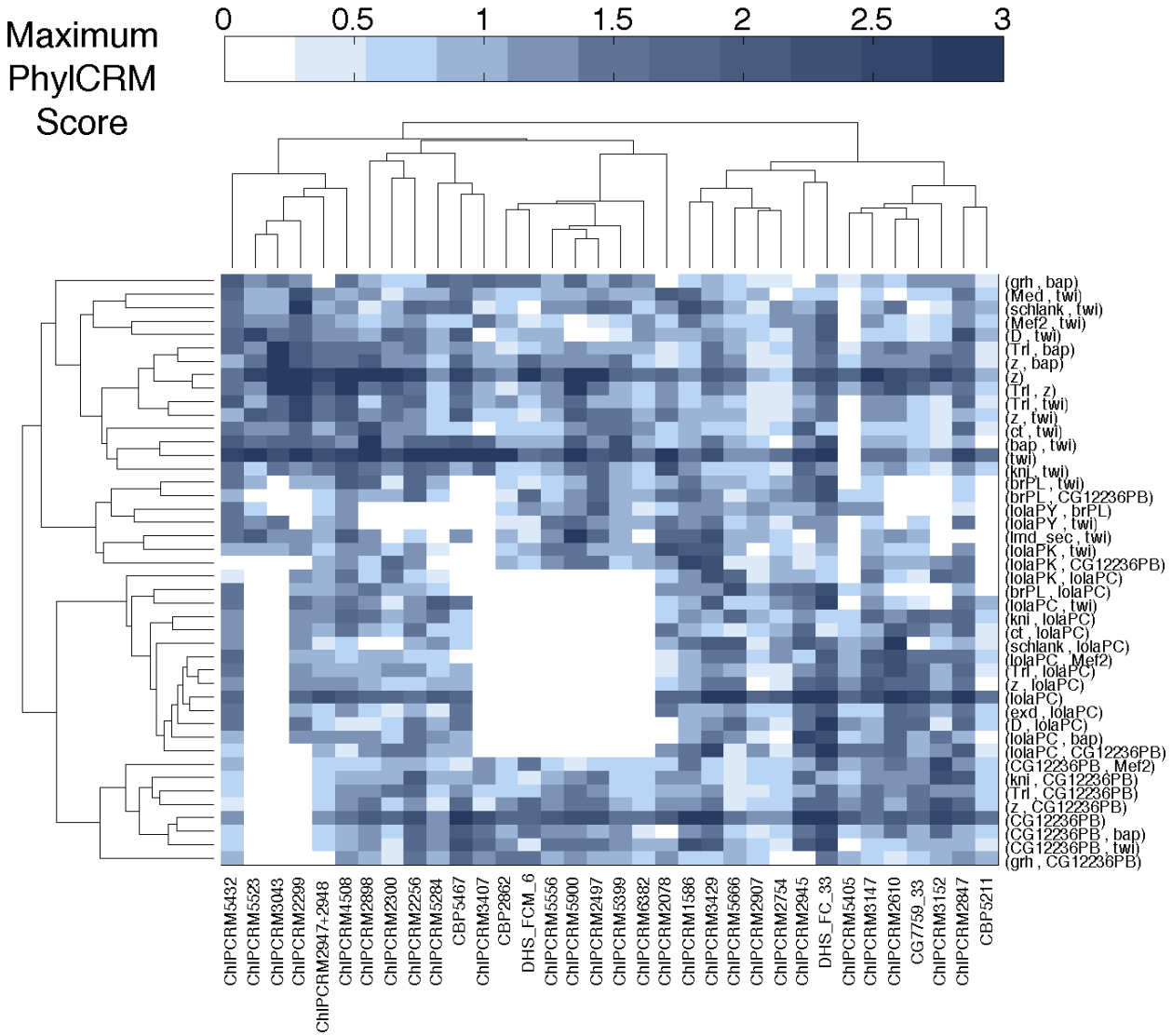


Supplementary Figure 7. TF binding site motifs or motif combinations significantly enriched (AUC ≥ 0.65, FDR ≤ 0.1) among eFS-identified active enhancers in FCMs (*Mef2-I-E_{D5}*:CD2$^+$ cells). Colors and shapes are as in Fig. 4a. Here, node diameter is proportional to (AUC-0.4)^2 considering the Lever AUC for the individual motif.



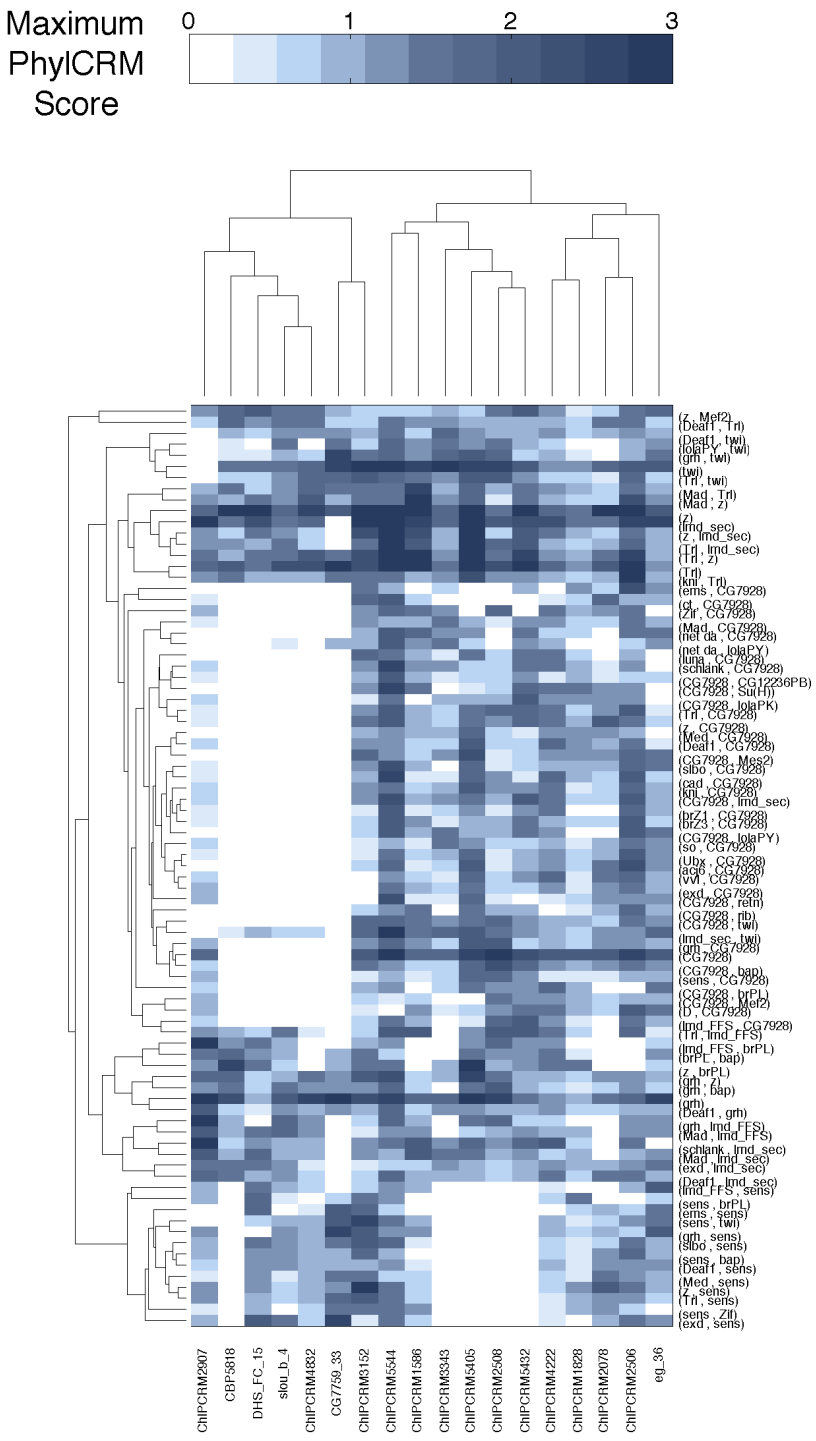
Supplementary Figure 8. Motif enrichment within enhancers active in mesoderm (*twi*:CD2+), FCMs (*Mef2-I-E*$_{D5}$:CD2+), or FCs (*duf*:CD2+). (**a**) Heatmap of Lever AUC values. (**b,c**) *(see next page for legend)*.

(b) ***twi*:CD2+**

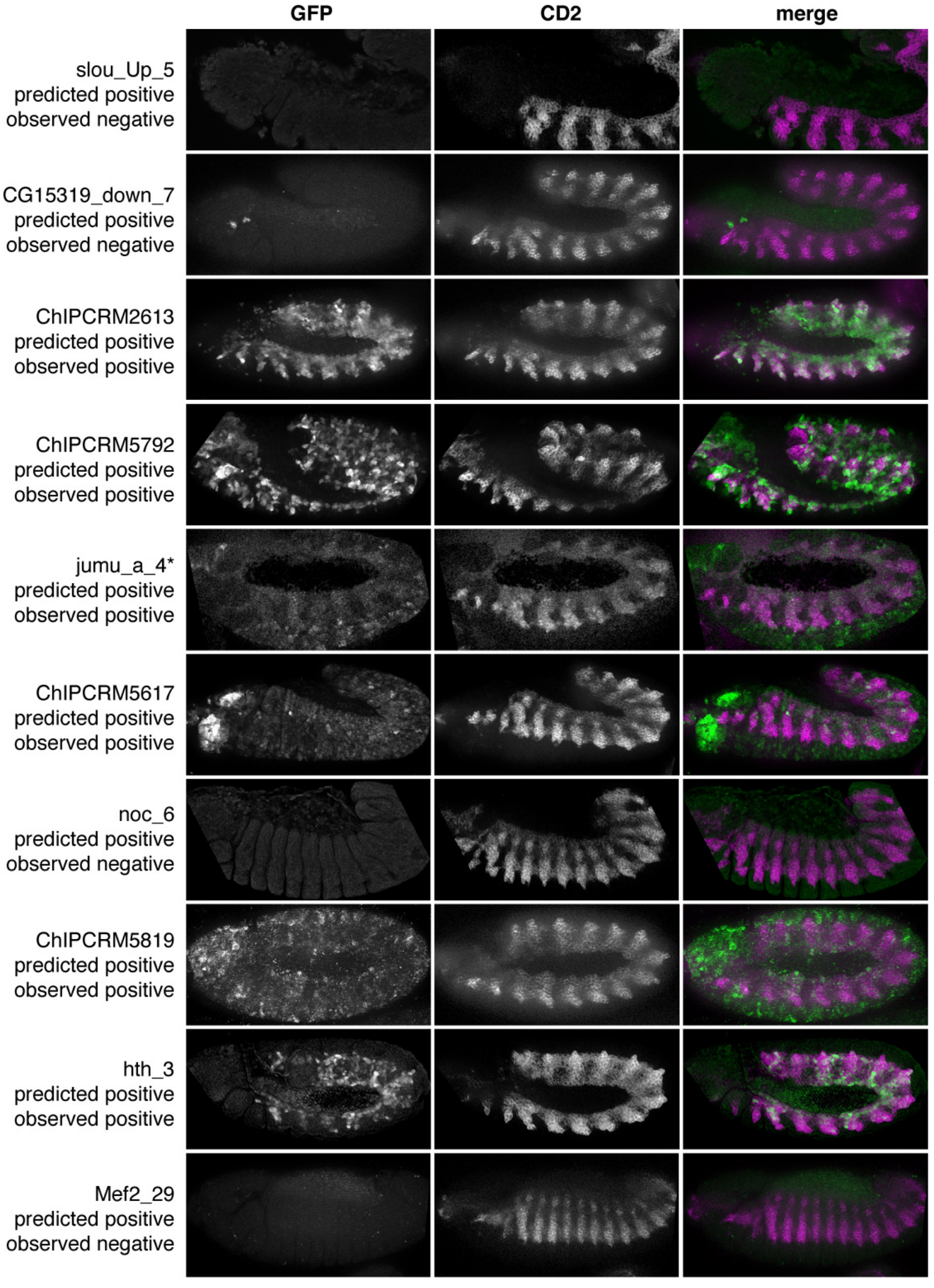


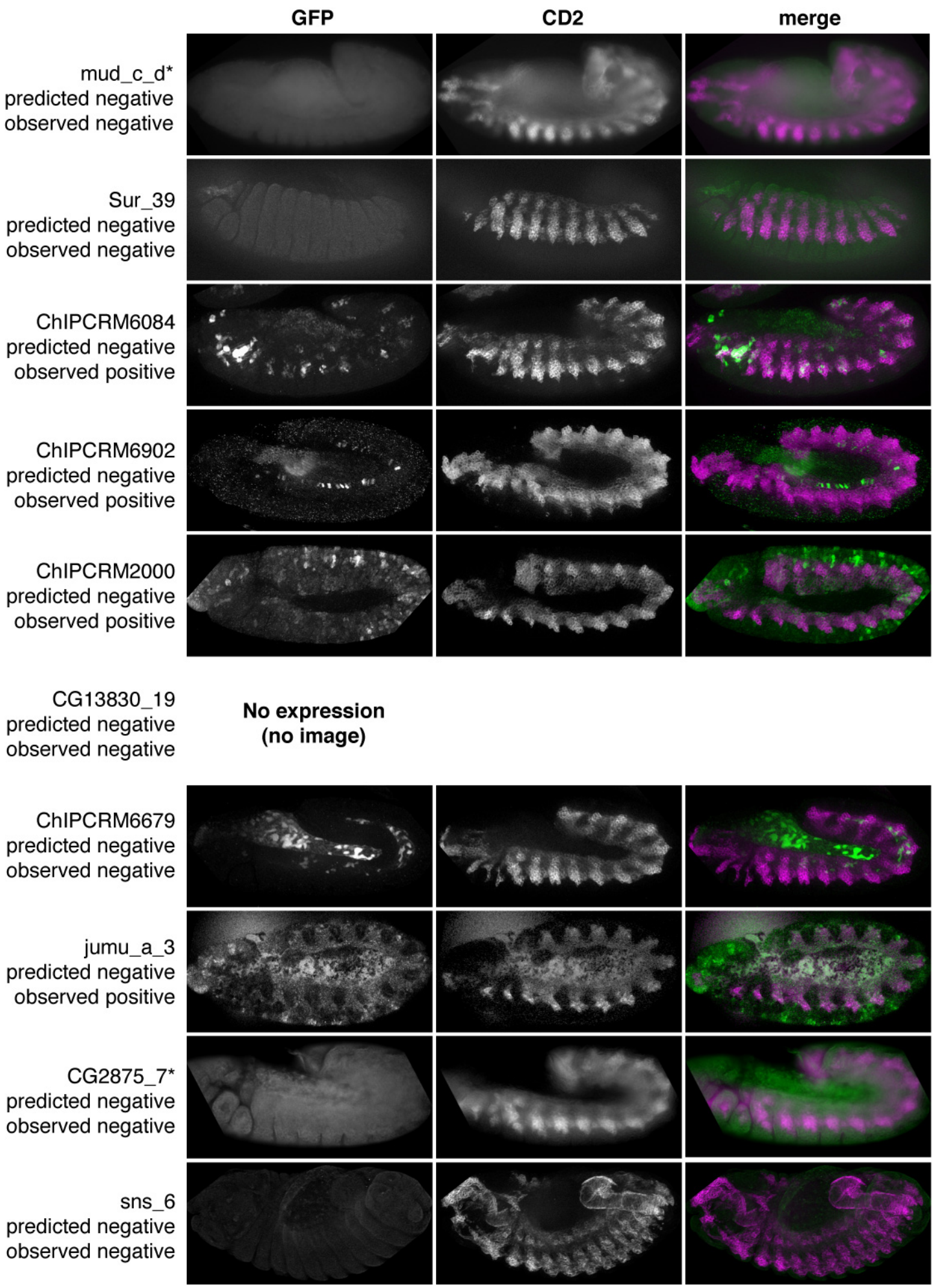
Supplementary Figure 8. (**b,c**) Heatmap of PhylCRM scores for individual cCRMs active in (**b**) mesoderm (*twi*:CD2+) or (**c**) FCMs (*Mef2-I-E*$_{D5}$:CD2+). Rows were hierachically clustered by average linkage according to Pearson correlation coefficient. Regions colored in white indicate 0-scoring motif-cCRM pairings that are not targeting a subset of the cCRMs. Darker blue rows indicate high scoring windows for the particular motif or motif combination.

(c) ***Mef2-I-E$_{D5}$*:CD2+**

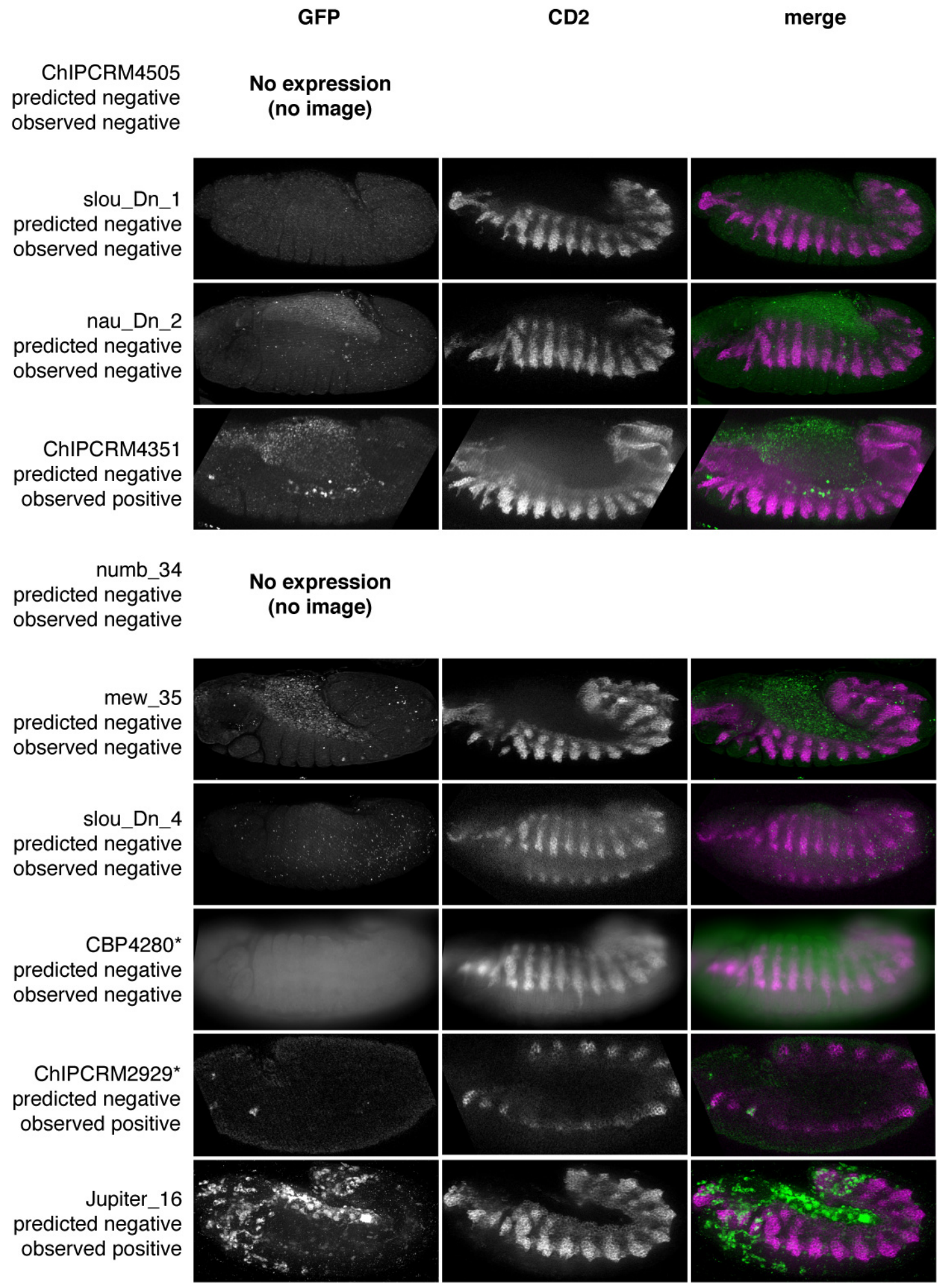


Supplementary Figure 9. Validation of classifier predictions. Activity of individual cCRMs in *twi*:CD2 cells was predicted by naive Bayes classifier (see text) and tested by conventional reporter assay. Individual cCRMs shown here were not detected in the *twi*:CD2 eFS experiment and therefore did not contribute to the training set for the classifier; several of them were detected in the *Mef2-I-*E_{D5}:CD2 and/or *duf*:CD2 eFS experiments, and these appear also in Supplementary Figure 5. Images shown are Maximum Intensity Projections except where noted by asterisk (see Oneline Methods and Supplementary Note 2). Co-expression 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. Co-expression is observed as GFP and CD2 being present in the same cells, since the GFP in these embryos is nuclear, while CD2 is expressed on the cell surface (see Online Methods and Supplementary Note 2).

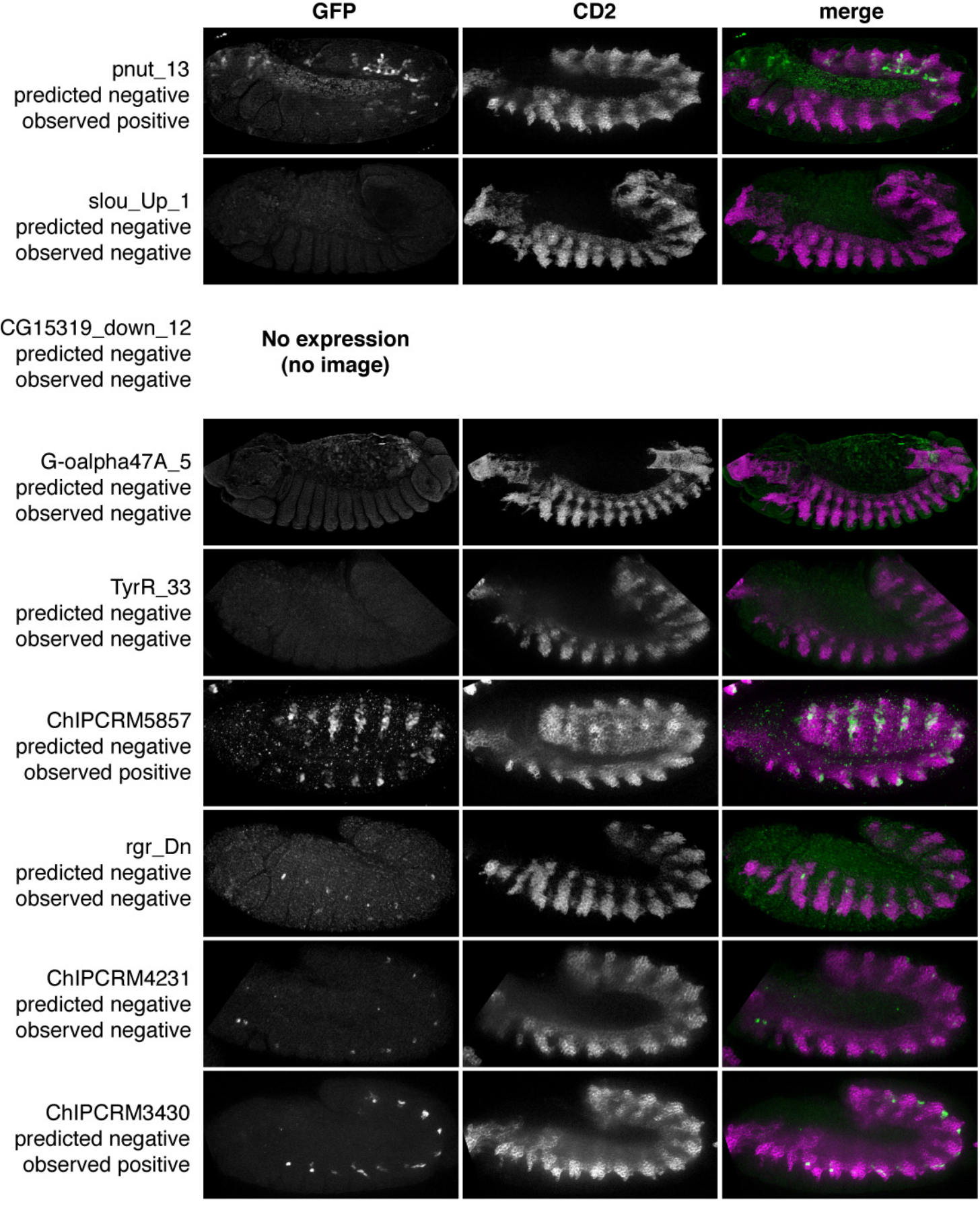

GFP
CD2
merge
slou_Up_5
predicted positive
observed negative
CG15319_down_7
predicted positive
observed negative
ChIPCRM2613
predicted positive
observed positive
ChIPCRM5792
predicted positive
observed positive
jumu_a_4*
predicted positive
observed positive
ChIPCRM5617
predicted positive
observed positive
noc_6
predicted positive
observed negative
ChIPCRM5819
predicted positive
observed positive
hth_3
predicted positive
observed positive
Mef2_29
predicted positive
observed negative



Nature Methods: doi:10.1038/nmeth.2558



Nature Methods: doi:10.1038/nmeth.2558



Nature Methods: doi:10.1038/nmeth.2558