Déchiffrement de l'activité des séquences cis-régulatrices chez la Drosophile basé sur la localisation des facteurs de transcription et la caractérisation de l'état de la chromatine

Charles Girardot

THESE DE DOCTORAT DE

L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité Bioinformatique

Ecole Doctorale Complexité du vivant

Présentée par

M. Charles Félix Béranger GIRARDOT

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Deciphering enhancer activity in Drosophila based on transcription factor occupancy and chromatin state characterization.

soutenue le 9 Juillet 2012

devant le jury composé de:

M. Denis THIEFFRY, Directeur de thèse

Mme. Eileen FURLONG, Co-directrice de thèse

M. Stein AERTS, Rapporteur

M. Paul BERTONE, Rapporteur

Mme. Nathalie DOSTATNI, Examinateur

M. Laurent PERRIN, Examinateur

M. Krzysztof JAGLA, Examinateur

Ce travail de thèse a été effectué dans le groupe du Dr. Eileen Furlong à l'European Molecular Biology Laboratory (EMBL) de Heidelberg (Allemagne) et au laboratoire de Biologie computationnelle des systèmes dirigé par le Pr. Denis Thieffry, Institut de biologie de l'Ecole Normale Supérieure de Paris (UMR ENS - CNRS 8197 - INSERM 1024 - 8197)

Remerciements

Je tiens tout d'abord à remercier de tout cœur Eileen Furlong et Denis Thieffry, sans qui cette thèse n'aurait pas vu le jour.

Je remercie profondément mes collègues et co-auteurs Robert, Julien, Guillaume, Mikhail, Martina, Hilary, Stefan, Nicolas, Bartek et Yad pour ces travaux communs effectués dans l'efficacité et la bonne humeur. Je remercie particulièrement Robert, Bartek et Julien pour leurs précieux conseils.

Un énorme merci à toi, Robert, pour ton temps passé à corriger mon anglais, deux bouteilles de vin n'y suffiront pas !

Je remercie mes collègues, présents et passés, du « Furlong Lab » pour tous ces bons moments partagés.

Je remercie chaleureusement tous les membres de mon jury.

A mes parents, sans qui rien ne serait possible !

A Céline, mon amour.

A Paul et Simon, mes enfants chéris.

Résumé

La caractérisation des modules cis-régulateurs (CRM) ainsi que de leur activité sont essentiels pour comprendre la régulation des gènes au cours du développement des métazoaires. La technique de l'immunoprécipitation de la chromatine suivie du séquençage à haut débit de l'ADN (ChIP-seq) constitue une approche puissante pour localiser les CRM. Afin de localiser des facteurs génériques au sein de tissus spécifiques, nous avons développé une approche ChIP-seq sur des noyaux triés par cytométrie de flux et localisons des modifications post-traductionelles de l'histone H3, ainsi que l'ARN polymérase II (PolII) dans le mésoderme de la Drosophile. Nous montrons que les CRM actifs sont caractérisés par la présence d'H3 modifiés (K27Ac et K79me3) et de PolII. De plus, la présence et la forme des signaux correspondants à ces marques corrèlent dynamiquement avec l'activité des CRM. Enfin, nous prédisons la présence de CRM actifs et confirmons leur activité *in vivo* à 89%. Parallèlement, nous étudions comment cinq facteurs essentiels au développement cardiaque se coordonnent en *cis* au sein du mésoderme dorsal, précurseur des mésodermes cardiaque (MC) et viscéral (MV). Nous démontrons que ces facteurs sont recrutés en tant que *collectif* au niveau des CRM cardiaques via un nombre limité de sites de fixation et en l'absence de contraintes architecturales. En outre, nous découvrons que ces facteurs cardiaques sont recrutés au niveau de CRM actifs dans le MV voisin et activement réprimés dans le MC, reflétant ainsi l'origine tissulaire commune de ces deux populations cellulaires. Nous concluons que les CRM impliqués dans le développement peuvent présenter une *empreinte développementale*.

Summary

The characterization of *cis*-regulatory modules (CRMs) and of their activity is central to understanding gene regulation and metazoan development. Chromatin immunoprecipitation followed by microarray or deep sequencing (ChIP-seq) against TFs are powerful approaches to map CRMs. To enable *in vivo* tissue-specific ChIP against ubiquitously expressed factors, we develop a ChIP protocol relying on the sorting of fluorescence activated cells, followed by deep sequencing. Using this protocol, we map histone modifications and RNA Polymerase II (PolII) occupancy in the *Drosophila* mesoderm, and subsequently study the chromatin state of active CRMs *in vivo*. We show that active CRMs are enriched for H3K27Ac, H3K79me3 and PolII, and that the presence and shape of these marks dynamically correlate with CRM activity timing and nucleosome positioning. Using Bayesian inference, we predict new CRMs to be active in the mesoderm and validate 89% of them *in vivo*. Next, we investigate how five TFs essential for cardiac specification operate in *cis* in the dorsal mesoderm, the developmental precursor of the visceral mesoderm (VM) and the cardiac mesoderm (CM). We demonstrate that they are recruited as a *TF collective* at cardiac CRMs without strong sequence requirements, thereby suggesting a novel mode for CRM activation. We further observe that cardiac TFs occupy CRMs that are active in the VM sibling lineage, echoing the fact that both cell populations derived from the dorsal mesoderm. We thus conclude that dormant TF binding signatures may reveal a developmental footprint of a cell lineage.

Mots Clés

Séquencage à haut débit, chromatine-immunoprécipitation spécifique d'un tissu, épigénomique, développement embryonnaire de la Drosophile, ChIP-chip, ChIP-seq, activité temporelle des séquences activatrices, état de la chromatine, mésoderme, réseaux bayésiens, régulation transcriptionnelle, modules cis-régulateurs, découverte de motifs, ARN polymérase II, positionnement des nucléosomes, prédiction de l'activité des modules cis-régulateurs, spécification du mésoderme dorsal, empreinte développementale, architecture des modules cis-régulateurs, co-localisation des facteurs de transcription.

Keywords

High-throughput sequencing, tissue-specific ChIP, epigenomics, Drosophila embryonic development, ChIP-chip, ChIP-seq, temporal enhancer activity, chromatin state, mesoderm, Bayesian network, RNA Polymerase II, nucleosome positioning, enhancer activity prediction, dorsal mesoderm specification, developmental history, enhancer architecture, motif discovery, TF co-localization.

Titre français

Déchiffrement de l'activité des séquences cis-régulatrices chez la Drosophile basé sur la localisation des facteurs de transcription et la caractérisation de l'état de la chromatine.

Résumé Long

Les Modules de Régulation en Cis (CRM) intègrent les effets des facteurs de transcription (TF) et les traduisent en profils d'expression spatio-temporels. Il est maintenant établit que les CRM sont les déterminants majeurs des profils d'expression géniques complexes observés au cours du développement. La caractérisation des CRM ainsi que de leurs profils d'activité sont donc des éléments essentiels pour comprendre la régulation des gènes, et plus largement le développement des métazoaires. A cet égard, la technique d'immuno-précipitation de la chromatine, suivie de l'hybridation sur puces à ADN (ChIP-chip) ou du séquençage à haut débit (ChIP-seq) de l'ADN immuno-précipité constituent de puissantes approches pour localiser les CRM à l'échelle du génome. Ces protocoles exploitent généralement la fixation à l'ADN d'un TF spécifique du tissu étudié et, dans ce cas, peuvent être conduits sur des embryons entiers. Dans le cas de TF exprimés dans plusieurs tissus et dans le cadre de l'étude d'un tissu particulier, les expériences doivent être réalisées à partir de chromatine extraite d'organes disséqués ou de cellules cultivées.

Nous avons précédemment localisé par ChIP-chip les sites de fixation des facteurs Twist (Twi), Myocyte enhancer factor-2 (Mef2), Bagpipe (Bap), Biniou (Bin) et Tinman (Tin), tous spécifiquement exprimés dans le mésoderme, au cours du développement embryonnaire de la Drosophile. A l'aide de ces données, nous avons montré que la manière avec laquelle ces facteurs se combinent localement permet de définir précisément la présence de CRM, et que ces données peuvent être intégrées (par apprentissage supervisé) pour prédire leurs profils d'expression spatio-temporels.

La présence simultanée de différents TF est en effet une caractéristique courante des

CRM et la présence de clusters de sites hétérogènes (mais aussi homogènes chez certaines espèces) a été exploitée par de nombreuses équipes pour prédire, *in silico*, la présence de CRM. Néanmoins, les règles qui régissent l'organisation des différents sites de fixation au sein des CRM (ou «grammaire des motifs») restent à découvrir, et les modèles existants («enhanceosome» et «billboard») doivent encore être validés. Nous nous intéresserons plus particulièrement à cet aspect dans la seconde partie de ce travail. La première partie de ce travail, quant à elle, porte sur la mise au point d'un protocole de ChIP-seq tissu-spécifique, ainsi que de l'analyse des données obtenues pour différentes marques de l'activité de la chromatine, ce qui nous amène à proposer une méthode bioinformatique de prédiction de modules de régulation actifs à partir de la caractérisation de l'état de la chromatine.

Partie 1 : L'analyse de l'état de la chromatine provenant d'un tissu unique mets en évidence des signatures temporelles liées à l'activité des séquences régulatrices au cours du développement embryonnaire (publication dans *Nature Genetics*)

La liaison à l'ADN des TF n'est possible qu'en l'absence de nucléosomes, à l'exception notable des TF pionniers qui auraient le potentiel de se lier à leurs sites de fixation en présence de nucléosomes afin de déplacer ces derniers et ainsi créer un environnement local propice au recrutement d'autres TF. Ainsi, l'accès des TF à l'ADN et donc la structure de la chromatine sont des facteurs cruciaux pour l'activité des CRM et de leurs gènes cibles. Différentes études ont suggéré que les complexes d'histones localisés au niveau des CRM, à l'instar de ceux localisés dans les gènes, portent des modifications post-traductionelles (PTM) spécifiques reflétant leur d'activité transcriptionelle. Des protocoles exploitant ces observations ont été développés afin de localiser les séquences régulatrices de manière globale (indépendamment de facteurs spécifiques à un tissu), en exploitant la présence de cofacteurs (p300/CBP), les PTM des histones, ou encore en caractérisant l'accessibilité de la chromatine (FAIRE, hypersensibilité à la DNase I). Néanmoins, ces approches identifient indifféremment des séquences régulatrices actives ou non et ne peuvent être conduites qu'à partir de cultures cellulaires ou d'échantillons provenant de tissus disséqués (du fait du caractère ubiquitaire de l'expression des protéines ciblées). L'hypothèse de l'existence d'un code basé sur la combinaison des PTM des histones a reçu une attention toute particulière. En effet, de nombreuses études s'intéressant à l'état de la chromatine ont

été conduites à l'échelle du génome, révélant des signatures spécifiques aux promoteurs actifs, aux gènes en cours de transcription, ou encore à diverses séquences régulatrices (séquences insulatrices, activatrices, régions réprimées par le complexe Polycomb).

Ces recherches ont clairement démontré la puissance des études de modifications de la chromatine pour caractériser différents éléments fonctionnels du génome. Néanmoins, ces études conduisent à des conclusions contradictoires quant aux PTM associées aux CRM et notamment à leurs états d'activité. En particulier la marque H3K4me1 est présente au niveau des CRM mais son association spécifique à des CRM actifs est maintenant contestée. De même, la marque H3K4me3 est considérée comme spécifique des promoteurs actifs et utilisée pour différencier les promoteurs des CRM, mais elle a récemment été mise en évidence au niveau de CRM actifs. Certaines de ces différences sont imputables, au moins en partie, à l'origine des échantillons utilisés (cultures cellulaires ou embryons entiers, type cellulaires ou organismes différents). Cependant, ces études reposent généralement sur certaines approximations pouvant influencer leurs conclusions. D'abord, la manière de définir les CRM est généralement basée sur la présence de cofacteurs ou de sites hypersensibles à la DNase I ; or la présence de cofacteur(s) et de sites hypersensibles ne sont pas spécifiques des CRM (ce sont aussi, par exemple, des caractéristiques des promoteurs) et définissent potentiellement soit une sous classe particulière de CRM (cofacteur), soit un ensemble de régions de fonctions différentes (sites hypersensibles à la DNase I). Ensuite, la manière d'évaluer l'état d'activité de ces CRM potentiels est effectuée en considérant l'état d'activité du gène le plus proche. Or, le gène le plus proche d'un CRM n'est pas forcement le gène cible (en particulier dans le cas de génome dense tel que celui de la Drosophile). Par ailleurs, même lorsque le gène le plus proche est bien le gène cible, la nature multiple de la relation liant CRM et gène(s) cible(s) ne garantit pas qu'un simple transfert d'activité soit pertinent. Enfin, ces études ont été réalisées à partir de cultures cellulaires, in vitro, et leurs conclusions doivent être confirmées dans le contexte du développement d'un organisme entier. En effet, des cellules souches embryonnaires de mammifères nécessitent entre 7 et 12 jours pour se différencier en culture alors que des transitions majeures sont réalisées en seulement quelques heures à l'échelle du développement embryonnaire (l'embryogénèse dure environ 18h chez la Drosophile). Il est donc essentiel d'étudier la dynamique de la chromatine et de comprendre comment celle-ci affecte ou est affectée par le recrutement des TF au sein d'un tissu particulier et dans le contexte du développement embryonnaire.

Immunoprécipitation de la chromatine spécifique d'un tissu à partir d'embryons entiers

Nous avons contribué à la mise au point d'un protocole de ChIP-seq utilisant des noyaux marqués spécifiquement pour leur tissu d'origine et triés par cytométrie de flux (BiTS-ChIP). En l'occurrence, nous utilisons une lignée transgénique de Drosophile ayant intégré de manière stable un transgène codant pour la protéine d'histone H2B fusionnée à la « Streptavidin Binding Peptide » et placé sous le contrôle de la séquence activatrice Twist-PEMK spécifique du mésoderme. Nous localisons ainsi, dans le mésoderme de la Drosophile (après 6-8 h de développement), la présence de l'ARN polymérase II (Pol II) et de H3 (afin de quantifier la densité des nucléosomes), ainsi que les PTM de cette histone, qui son associées aux promoteurs actifs (H3K4me3, H3K27ac), aux gènes activement transcrits (H3K79me3 et H3K36me3), aux CRM (H3K4me1 et H3K27ac), et aux régions réprimées par le complexe Polycomb (H3K27me3). Nous vérifions tout d'abord que notre nouveau protocole présente à la fois une haute sensibilité et une haute spécificité en comparant les sites de fixations du TF Mef2 (un TF spécifiquement exprimé dans le mésoderme) identifiés par BiTS-ChIP, ChIP-seq et ChIP-chip (NG Fig. 1). Plus de 81% des sites de fixation identifiés (pic de signal statistiquement élevé par rapport au signal de référence) sont partagés par ces trois méthodes prises deux à deux. La spécificité de notre nouvelle méthode est clairement démontrée en comparant les niveaux de signal obtenus pour H3K4me3, H3K27ac et Pol II au niveau des promoteurs de gènes exprimés exclusivement dans le mésoderme (fort signal) ou uniquement en dehors du mésoderme (absence de signal) (NG Fig. 2).

H3K27ac, H3K79me3 et Pol II sont enrichis dans les modules cis-régulateurs actifs

Afin d'évaluer les relations entre les PTM d'histones et l'activité des CRM, nous avons collecté les profils d'expression spatio-temporels de 465 CRM disponibles dans différentes bases de données et la littérature. Il est important de souligner que ces CRM et leurs profils d'expression ont tous été caractérisés in vivo (dans des animaux transgéniques) et vérifié *un à un* avant d'être pris en compte dans cette étude ; la base de données ainsi collectée est nommée CAD2 (NG Fig. 3a). Par ailleurs, les CRM situés dans des gènes ou à leur proximité immédiate (moins de 1 kb) n'ont pas été considérés plus avant dans cette étude

(ces gènes et CRM partageant un certain nombres des modifications étudiées, l'origine du signal ne pourrait être clairement établie) et tous les résultats mentionnés ci-dessous se reportent donc aux 144 CRM inter-géniques rassemblés dans CAD2.

Dans un premier temps, nous identifions les zones enrichies (ou pics) pour chacune des modifications étudiées, ainsi que pour Pol II, à l'aide du logiciel MACS, et les comparons avec les 144 CRM indépendamment de leur état d'activité : 111 CRM (77%) sont enrichis en H3K4me1, 23 (16%) en H3K27ac, 11 en Pol II (8%, un pourcentage similaire à ceux observés ailleurs) et 21 (15%) en H3K79me3, alors qu'aucun ne contient H3K36me3. Sur les 21 CRM couverts par H3K79me3 (une modification jusqu'à alors considérée comme spécifique des gènes transcrits), seuls 7 (33%) sont également enrichis en Pol II, ce qui suggére que la triméthylation de H3K79 au niveau des CRM s'effectue de manière indépendante de la présence de Pol II, ou que cette marque perdure un certain temps après que Pol II ait fini d'opérer. Quoiqu'il en soit, H3K79me3 représente une nouvelle signature des CRM impliqués dans le développement.

Nous divisons ensuite les 144 CRM en deux groupes en fonction de leur activité dans le mésoderme à 6-8h (ceux qui sont actifs dans le mésoderme au temps étudié (6-8h) *versus* ceux qui ne le sont pas) et évaluons si certaines modifications (et Pol II) sont liées à l'état d'activité des CRM (NG Fig. 3). Nous concluons que les CRM actifs sont caractérisés par la présence de PTM d'histones (H3K27ac et H3K79me3) et de la Pol II. A l'inverse, la marque H3K4me1 n'est pas enrichie dans une classe d'activité des CRM particulière et est présente sur une large majorité des CRM indépendamment de leurs classes d'activité. A la lumière de ces résultats nous pouvons d'ores et déjà conclure que H3K4me1 n'est pas une PTM associée spécifiquement aux CRM actifs et que ceux-ci présentent non pas une mais plusieurs caractéristiques : H3K27ac, H3K79me3 et Pol II.

Relations dynamiques des modifications de la chromatine et Pol II avec l'activité des modules de régulation et la présence des facteurs de transcription

Afin de préciser la relation entre la présence des PTM étudiées et Pol II avec l'activité temporelle des CRM, nous évaluons la présence de celles-ci au sein de trois classes d'activité temporelle : les CRM actifs dans le mésoderme seulement avant 6h ('<6h'), à 6-8h, et seulement après 8h ('>8h'). La présence des trois marques associées aux CRM actifs

(H3K27ac, H3K79me3 et Pol II) présente une forte corrélation avec le temps d'activité des CRM (NG Fig. 4). En effet, ces marques sont pratiquement absentes des CRM actifs uniquement avant ou après 6-8h . En particulier, la présence de Pol II n'est observée que sur les CRM actifs à 6-8h. Ici encore, H3K4me1 est présent sur une large majorité des CRM et ne présente aucune corrélation significative avec leur profil d'activité, confirmant notre conclusion antérieure.

Nous avons montré dans d'autres études que le temps d'activité des CRM actif dans le mésoderme est intimement lié à la présence de TF spécifiques du mésoderme. Nous avons récemment publié une large collection de CRM (que nous nommerons TF-Meso-CRM pour éviter toute confusion), dont la définition s'appuie sur la présence, déterminée par ChIP-chip, d'un ou plusieurs TF spécifiques du mésoderme (Twi, Mef2, Bap, Bin et Tin).

A nouveau, nous définissons trois groupes de TF-Meso-CRM, ceux liés par ces TF seulement avant le temps 6h, ceux liés pendant la période 6-8h (non exclusivement) et ceux liés seulement après le temps 8h (NG Fig. 4). Dans cette analyse, nous nous intéressons au profil moyen du signal, aligné sur les sites de fixation des TF, pour chacune des PTM (et Pol II) et pour chaque classe de TF-Meso-CRM définie. Cette approche nous permet d'évaluer, de manière précise, non seulement la quantité de signal présente, mais aussi sa forme relativement aux sites de fixation des TF. Les profils observés pour H3K4me1 et H3K27ac sur les TF-Meso-CRM occupés par des TF à 6-8h sont clairement bimodaux, avec une diminution nette du signal centrée sur les sites de fixation des TF, suggérant une absence de nucléosome au niveau des sites de fixation et la présence de nucléosome modifiés (H3K4me1 et H3K27ac) de part et d'autres des sites de fixation.

Cette hypothèse est confirmée par l'inspection du profil de densité de H3, qui présente une forte diminution au niveau des sites de fixation lorsque ceux-ci sont occupés par un TF. Par contre, lorsque les TF ne sont plus présents sur les TF-Meso-CRM (mais l'étaient antérieurement), le signal reflétant la densité de H3K4me1 et H3K27ac est unimodal et centré sur les sites de fixation des TF suggérant un repositionnement des nucléosomes au niveau des sites de fixation. Ainsi, la forme du signal plutôt que sa quantité semble être un meilleur indicateur de l'activité d'une séquence régulatrice. La densité de H3K79me3 ne présente pas tout à fait les caractéristiques décrites ci-dessus et, même si sa présence est un bon indicateur d'activité comme nous l'avons vu précédemment, sa densité reste basse autour des sites de fixation et s'élève en périphérie indiquant que les nucléosomes portant cette modification ne

sont pas les mêmes que ceux portant les modifications K4me1 et K27ac. A l'inverse des PTM d'histones observées au niveau des TF-Meso-CRM occupés par des TF à 6-8h, le signal de Pol II est lui de forme unimodal et le sommet de son pic coïncide parfaitement avec la position des sites de fixation des TF. Par contre, Pol II est absente lorsque les TF ne sont plus présents sur les TF-Meso-CRM (mais l'étaient antérieurement).

En conclusion, il apparaît que la présence de Pol II sur les CRM est étroitement liée à la position des sites de fixation des TF et à leur occupation, mais aussi au fait que ces CRM soient actifs (voir l'analyse effectuée avec les CRM de CAD2 plus haut). D'une manière générale, ces résultats suggèrent que la présence des TF sur les CRM est précurseur du recrutement de la Pol II et que ce recrutement pourrait être responsable de l'activation de certains CRM.

Prédiction de modules de régulation actifs à partir de l'état de la chromatine

Après avoir montré que la présence de H3K27ac, H3K79me3 et Pol II sur les CRM corrèlent individuellement avec l'activité de ceux-ci, nous voulons évaluer si une prise en compte combinée de ces marques permettrait de prédire efficacement la présence de CRM actifs dans le génome. Les analyses précédentes sont basées sur la définition préalable de régions enrichies en histones H3 modifiés ou en Pol II à l'aide du logiciel MACS, puis de leur comparaison par superposition avec différents jeux de CRM. Si cette approche à l'avantage de la simplicité, elle implique la binarisation des données: un CRM est enrichi en H3K27ac (par exemple) ou ne l'est pas. De plus, il n'est pas évident a priori d'établir une règle de prédiction : Faut-il considérer la présence d'une, deux ou des trois marques ? Devrions-nous avoir recours à des règles logiques plus complexes telle que par exemple « Pol II *ou* (H3K27ac *et* H3K79me3) » ? Doit on considérer l'union ou l'intersection des régions enrichies telles que définies par MACS ? En effet, une inspection précise des densités de PTM d'histone et de Pol II présentes au niveau des CRM de CAD2 actifs dans le mésoderme à 6-8h (NG Fig. 5b) révèle à la fois des densités et des combinaisons de marques hétérogènes.

Pour résoudre ce problème, nous avons décidé d'appliquer un modèle probabiliste quantitatif, l'inférence bayésienne (ou « réseaux bayésiens »), pour *apprendre* les dépendances existantes entre les densités des marques étudiées (PTM d'histones et Pol II) sur les CRM (de CAD2) et leurs activités dans le mésoderme. Dans ce réseau, nous modélisons

deux types d'activité dans le mésoderme: les CRM précisément actifs dans le mésoderme à 6-8h de développement (ils peuvent aussi être actifs à d'autres stades de développement et dans d'autres tissus) et les CRM actifs dans le mésoderme à n'importe quel stade de développement (ils peuvent aussi être actifs dans d'autres tissus). Les échantillons d'apprentissage sont construits avec les CRM de CAD2 et la performance du modèle reconstruit est estimée par validation croisée en utilisant des échantillons d'apprentissage composé de 75% des individus disponibles.

La performance est jugée satisfaisante pour les deux expressions modélisées (« actif dans le mésoderme à 6-8h » ou « actif dans le mésoderme »), avec des aires sous la courbe ROC (spécificité/sensibilité) de 0.82 et 0.76, respectivement. Le modèle identifie des dépendances conditionnelles positives entre la présence de H3K27ac, H3K79me3, et l'expression dans le mésoderme (à 6-8h et globalement), et entre la présence de Pol II et l'expression dans le mésoderme à 6-8h (NG Fig. 5b et Suppl. Fig. 11). Le modèle révèle aussi une dépendance conditionnelle négative entre la présence de H3K27me3 et l'expression globale dans le mésoderme.

Ces résultats sont équivalent à ceux observés précédemment, mais le réseau bayésien obtenu nous permet maintenant de scruter le génome afin de prédire des régions actives dans le mésoderme à 6-8h (l'expression globale dans le mésoderme ne sera pas étudiée plus avant). En se basant sur les courbes ROC, 112 régions, couvrant globalement ~303 kb de séquence génomique, sont prédites comme étant actives dans le mésoderme à 6-8h de développement avec une spécificité estimée à 100% (les prédictions ont été effectuées seulement dans les limites du génome intergénique). Il est intéressant de constater que les séquences prédites, à l'instar du jeux d'apprentissage, présentent une certaine hétérogénéité en terme de densité de H3K27ac, H3K79me3 et Pol II (NG Fig. 5c). Notons que 78% de ces prédictions contiennent un ou plusieurs TF-Meso-CRM occupés par au moins un TF à 6-8h. Les prédictions effectuées à partir du modèle bayésien correspondent donc effectivement à des séquences régulatrices actives à ce stade de développement dans le mésoderme.

Afin d'estimer la performance réelle de notre modèle, l'activité de 9 CRM prédits est examinée in vivo à l'aide de lignées transgéniques, dans lesquelles un gène rapporteur placé sous le contrôle d'un promoteur minimum précédé de la région à tester est intégré de manière stable. L'expression du gène rapporteur est alors évaluée par hybridation in situ au cours du développement embryonnaire. Précisons que la sélection des séquences testées a été réalisée

de manière à inclure des séquences présentant des profils hétérogènes en terme de présence et densité de H3K27ac, H3K79me3 et/ou Pol II ; alors que d'autres caractéristiques telles que la présence de TF, la conservation phylogénétique ou la prédiction de l'existence de sites de fixation de TF n'ont pas été considérées. Huit de ces neufs prédictions sont effectivement capables d'activer l'expression du gène rapporteur dans le mésoderme au stade de développement prédit. La taille moyenne des régions prédites (2.7 kb) est nettement supérieure à la taille moyenne des nombreux TF-Meso-CRM que nous avons pu tester dans des travaux antérieurs. Ceci suggère que les régions prédites représentent plutôt des régions actives contenant éventuellement plusieurs CRM. Etant donné que la présence de Pol II est hautement corrélée avec le stade précis d'expression et que sa position l'est avec la présence et le site de fixation des TF (NG Fig. 4), nous pensons que les pics de densité de Pol II (lorsqu'ils sont visibles) indiquent la position précise des CRM fonctionnels au sein des régions prédites. Afin de vérifier cette hypothèse, nous examinons le profil d'expression de deux sous-régions sélectionnées parmi les neufs régions déjà testées. Ces sous-régions sont centrées sur le pic de densité de Pol II (NG Fig. 6a,b). Notons qu'il s'agit généralement plutôt de traces de Pol II que de vrais pics de densité tels que ceux observés dans les promoteurs de gènes actifs. Les profils d'expression de ces sous-régions sont très largement similaires à ceux produits par les régions prédites originales (NG Fig. 6a,b) confirmant ainsi notre hypothèse. Par ailleurs nous testons, selon le même protocole, 4 CRM publiés par d'autres équipes dont l'expression dans le mésoderme à 6-8h n'a pas été décrite et pour lesquels la probabilité postérieure d'être actifs dans le mésoderme à 6-8h est minimale. Comme attendu, aucun de ces CRM n'est capable de diriger l'expression du gène rapporteur dans le mésoderme à 6-8h.

Partie 2 : Un collectif de facteur de transcription définit le devenir cardiaque des cellules et reflète l'histoire développementale de la lignée (publication dans *Cell*).

Au cours du développement embryonnaire, les cellules sont progressivement orientées vers leur destin final à travers l'intégration combinée des signaux provenant des tissus voisins (voies de signalisation) et des TF spécifiquement exprimés au sein de ces différentes cellules. Toutes ces informations convergent au niveau des CRM qui, en retour, promeuvent l'expression de gènes particuliers, qui, ensemble, définissent le devenir

développemental de la cellule. L'intégration de ces informations représentées par ces différents facteurs passe parfois par des interactions directes entre protéines, qui favorisent la liaison et la stabilisation de ces complexes au niveau des CRM cibles. Ce type de coopération dans la fixation à l'ADN requière souvent une organisation précise des sites de fixation des TF (orientations relatives, espacements des sites), ce qui correspond au modèle d'activation des CRM appelé « enhanceosome ». Un modèle alternatif, appelé « billboard », est plus permissif, sans architecture (ou « grammaire ») particulière. Dans ce dernier modèle, certains facteurs peuvent lier l'ADN de manière synergique, tandis que d'autres peuvent se lier de manière indépendante. Ainsi, les règles qui régissent l'organisation des différents sites de fixation au sein des CRM restent à découvrir, et les modèles existants («enhanceosome» et «billboard») doivent encore être validés.

Dans une étude récente, nous avons montré que des combinaisons différentes de facteurs de transcription mésodermiques (fixés sur différents CRM) pouvait engendrer une réponse transcriptionnelle similaire. Cela suggère que la position des sites au sein de CRM d'activité similaire est variable, mais également que l'identité des sites de fixation peut changer.

Le mésoderme cardiaque est spécifié au sein du mésoderme dorsal à l'intersection des voies de signalisation Wingless (Wg) et Dpp (Cell Fig. 1). La signalisation Dpp est requise pour maintenir l'expression du gène tinman dans le mésoderme dorsal. Tin et Dpp sont nécessaires pour former les trois types de cellules mésodermiques issus du mésoderme dorsal: le mésoderme cardiaque, viscéral et somatique dorsal. La voie de signalisation Wg permet de définir plus précisément le devenir de ces cellules en réprimant un gène clé du développement du mésoderme viscéral, bagpipe, dans le compartiment cardiaque, et en y activant l'expression d'une famille de gènes nécessaire à la spécification cardiaque, les gènes Dorsocross (Doc). Les facteurs de transcription Tin et Doc activent alors l'expression de pannier (pnr), et ces facteurs coopèrent pour spécifier un nombre correct de cellules cardiaques. De nombreuses études ont mise en évidence les diverses interactions génétiques (Cell Fig. 1c) qui existent entre ces facteurs et ont montré que celles-ci sont très conservées de la Drosophile à l'Homme. Néanmoins, la nature moléculaire de ces coopérations et les cibles directes de ces facteurs au cours de la spécification des cellules cardiaques sont encore mal caractérisées.

Dans la seconde partie de cette thèse, nous étudions comment ces cinq facteurs essentiels au développement cardiaque chez la Drosophile se coordonnent en *cis* au sein du mésoderme dorsal.

Les facteurs de transcription requis pour spécifier les cellules cardiaques se lient à l'ADN de façon collective

Nous procédons d'abord à l'identification, par ChIP-chip à partir d'embryons entiers agés de 4-6h et 6-8h, des sites de fixation des trois facteurs de transcription clés du processus de spécification cardiaque, Tin, Doc et Pnr, ainsi que des facteurs dTCF et pMad, les effecteurs des voies de signalisation Wg (dTCF) et Dpp (pMad). Il est important de noter que, à l'exception de Tin, ces facteurs sont présents dans plusieurs tissus (Cell Fig. 1b). L'analyse primaire des puces (normalisation, traitement du signal) est effectuée comme décrit dans une de nos études précédentes. Brièvement, nous identifions, pour chaque TF et chaque stade de développement indépendamment, les régions enrichies à l'aide du logiciel TileMap. Ensuite les positions exactes des sommets (de la courbe d'intensité du signal) au sein de chacune des régions TileMap définies sont déterminées ; comme nous avons pu le montrer ultérieurement, ces positions approximent les positions des sites de fixation des TF à 100 bp près. Finalement, toutes ces positions (tous facteurs et temps confondus) sont regroupées en clusters de sommets séparés par moins de 200 bp et chacun de ces clusters représente un CRM potentiel (comme les TF-Meso-CRM mentionnés précédemment), que nous nommerons TF-DM-CRM. Ainsi, chaque TF-DM-CRM est définit par une position génomique et un profil de fixation reflétant quels TF se lient à ce TF-DM-CRM et quand ils s'y lient. Finalement, les TF-DM-CRM sont regroupés en différentes classes (par *clustering*) en fonction de leur profil de fixation à l'aide du logiciel Autoclass (Cell Fig. 2).

Nous montrons d'abord que ces facteurs ont tendance à se fixer sur les mêmes TF-DM-CRM et ceci spécifiquement dans le mésoderme. En effet, les TF-DM-CRM liés par Tinman (facteur spécifique du mésoderme) sont majoritairement occupés par les quatre autres facteurs alors que les TF-DM-CRM Tin-négatifs sont liés par un seul facteur (Cell Fig. 1a,b). Près de 50% des TF-DM-CRM Tin-positifs identifiés sont occupés par les cinq facteurs. Les TF-DM-CRM restants représentent des classes de TF-DM-CRM montrant un niveau d'occupation enrichit pour seulement deux facteurs: « Tin + X » où X représente Doc, Pnr,

dTCF ou pMad. Autrement dit, les TF-DM-CRM occupés par Tin le sont soit avec un seul autre facteur, soit avec tous les quatre autres facteurs, mais pas avec deux ou trois des autres facteurs (Cell Fig. 1b). L'activité liée à des régions régulatrices caractéristiques de chaque classe (classes « Tin + X » ou classe « tous les 5 ») a été analysée in vivo au cours du développement embryonnaire à l'aide de lignées transgéniques (Cell Fig. 3). Les TF-DM-CRM co-occupés par les 5 facteurs (« tous les 5 ») sont fonctionnels et sont actifs dans le mésoderme dorsal ou ses dérivés, le mésoderme cardiaque et viscéral. Les classes de TF-DM-CRM « Tin + X » présentent des taux variés d'activité et, mis à part la classe Tin+dTCF, sont rarement actifs dans le mésoderme cardiaque. Ces résultats indiquent que les facteurs cardiaques se lient en tant que *collectif* pour réguler l'activité des gènes dans le mésoderme dorsal et ses dérivés.

Pannier et Dorsocross sont nécessaires à l'activité transcriptionnelle médiée par Tin, pMad et dTCF

Ayant observé que les cinq facteurs de la spécification cardiaque se fixent ensemble au niveau des TF-DM-CRM et compte tenu de leurs interactions génétiques et physiques, les CRM impliqués dans la spécification du mésoderme dorsal représentent un modèle idéal pour étudier la présence éventuelle d'une grammaire spécifique de sites de fixation. Pour ce faire, nous utilisons les régions enrichies (trouvées par les expériences ChIP-chip) et déterminons les modèles de fixation à l'ADN (matrices poids-position ou « PWM ») à l'aide de différents logiciels de découverte de motifs (RSAT, Weeder). Les sites de fixation pour les 5 TF sont ensuite prédits dans les TF-DM-CRM à l'aide des PWM obtenues et nous comparons la composition en sites de fixation des TF-DM-CRM occupés par les 5 facteurs (à fort potentiel coopératif) avec celle des TF-DM-CRM occupés par seulement 2 facteurs (classes « Tin + X »). Cette comparaison fait ressortir plusieurs différences en fonction des facteurs observés et suggèrent différents modes de recrutement au niveau de l'ADN. Ainsi les TF-DM-CRM occupés par les 5 facteurs contiennent un site de forte affinité pour Doc et Pnr plus fréquemment que les TF-DM-CRM « Tin + Doc » et « Tin + Pnr » (Cell Fig. 4). Cette observation s'inverse pour les 3 autres facteurs (Tin, pMad et dTCF) : un plus fort pourcentage de TF-DM-CRM « Tin + X » contient un site de forte affinité comparé aux TF-DM-CRM occupés par les 5 facteurs (Cell Fig. 4b). Si l'attractivité globale des TF-DM-

CRM, une mesure prenant en compte l'ensemble des sites de fixation potentiels et calculée par le logiciel TRAP, est maintenant considérée, ces observations sont globalement inversés avec des TF-DM-CRM « Tin + X » globalement plus attractifs (pMad et dTCF) et moins attractifs (Doc et Pnr) que les TF-DM-CRM occupés par les 5 facteurs (Cell Fig. 4c).

Des études récentes ont montré que l'ajout de GATA4 et TBX5 (les protéines orthologues de Pnr et Doc chez les mammifères) et d'un autre facteur dans des cultures cellulaires était suffisant pour entrainer la trans-différenciation des cellules mésodermiques en cardiomyoblastes, ou de reprogrammer des fibroblastes en cellule cardiaques, établissant clairement le rôle central de ces facteurs dans l'acquisition de l'identité cardiaque. Compte tenu du fort potentiel coopératif des facteurs cardiaques et des interactions protéiques qui existent entre eux, nous avons développé un système de culture cellulaire permettant de tester le rôle de Pnr et Doc dans les TF-DM-CRM fixés par les 5 facteurs. Grâce à des expériences d'ARN interférence (Cell Fig. 5), nous montrons que Pnr et Doc sont essentiels à l'activité transcriptionnelle des TF-DM-CRM testés, mais que le niveau d'activation des gènes cibles est lié à la présence des trois autres facteurs. Ces résultats suggèrent que Pnr et Doc sont nécessaires au recrutement collectif des 5 facteurs. En outre, l'analyse des distances entre les sites ou la recherche d'une orientation stéréotypée entre des sites voisins est restée vaine. L'ensemble de ces analyses nous permet de proposer un nouveau modèle d'interaction protéine-ADN au niveau des CRM, basé sur le recrutement d'un *collectif* de TF par l'intermédiaire d'un nombre restreint de sites de fixation, en l'absence d'architecture cis-régulatrice complexe.

Slp1 réprime l'activité des CRM du mésoderme viscéral dans les cellules du mésoderme cardiaque

Les mésodermes cardiaque et viscéral ont une origine commune, le mésoderme dorsal chez la Drosophile et le mésoderme splanchnique (seulement en partie) chez les mammifères. Chez la Drosophile, l'équipe du Dr Frasch a pu mettre en évidence le rôle clé de la signalisation Wg pour réprimer dans le mésoderme cardiaque l'expression d'un facteur de transcription clé de l'identité mésoderme viscéral: Bap. Le gène bap est activé dans le mésoderme dorsal par la voie de signalisation Dpp et Tin. Dans le futur mésoderme cardiaque, la signalisation Wg va activer le facteur Slp1 qui va se lier à une région régulatrice

et à son tour réprimer bap. En analysant les lignées transgéniques générées avec des TF-DM-CRM occupés collectivement par les 5 facteurs étudiés, nous réalisons que 25% d'entre eux activent l'expression de gène rapporteur dans le mésoderme viscéral.

Une étude précédente nous a permis de définir les TF-Meso-CRM à partir des 5 facteurs clés du développement du mésoderme (Twi, Mef2, Bin, Bap et Tin). A l'aide de ces données, nous mettons en évidence que les TF-DM-CRM occupés par les 5 facteurs cardiaques et ayant une activité dans le mésoderme viscéral sont également occupés par Bin (Cell Fig. 6). La présence de Bin semble donc être un bon indicateur de l'activité dans le mésoderme viscéral de ces TF-DM-CRM. Afin de comprendre pourquoi ces TF-DM-CRM ne sont pas actifs dans le mésoderme cardiaque malgré la présence de tous les acteurs requis, nous localisons, par ChIP-chip, les sites de fixation du facteur répresseur Slp1. Slp1 se révèle très largement présent au sein des TF-DM-CRM actifs dans le mésoderme viscéral mais pas dans le mésoderme cardiaque. Ainsi, dans le mésoderme cardiaque, ces TF-DM-CRM seraient occupés par les 5 facteurs cardiaques mais leur activation serait bloquée par Slp1. Bin et Slp1 appartiennent à la même famille de TF (FoxF) et ont des caractéristiques de fixation à l'ADN très proches sinon identiques. Finalement, nous montrons, à l'aide d'expériences de mutagenèse des sites de fixation Bin/Slp1, que l'abolition des sites Slp1 dans ces TF-DM-CRM entraine la réactivation de l'activité dans le mésoderme cardiaque (Cell Fig. 7). Nous concluons que les CRM impliqués dans le développement peuvent présenter une *empreinte développementale*.

Table of Contents

ABBREVIATIONS

AEL, after egg laying
AP, anterior-posterior
AR, androgen receptor
atf-2, activating transcription factor 2
bap, bagpipe
bp, basepairs
BDGP, Berkeley Drosophila Genome Project
BEAF-32, boundary element-associated factor of 32kD
bin, biniou
BN, Bayesian network
CAGE, Cap Analysis of Gene Expression
CBP, CREB-binding protein
c-jun, jun proto-oncogene
bHLH, basic helix-loop-helix
ChIP, chromatin immunoprecipitation
ChIP-chip, ChIP followed by microarray hybridization
ChIP-seq, ChIP followed by deep sequencing
CM, cardiac mesoderm
CRM, cis-regulatory module
CTCF, CCCTC- binding factor
DBD, DNA Binding Domain
DHS, DNAse I hypersensitive sites
dif, dorsal-related immunity factor
dl, dorsal
DNA, deoxyribonucleic acid
DNAse I, Deoxyribonuclease I
dpp, decapentaplegic
DREF, DNA replication-related element factor
dTCF, pangolin
DV, dorso-ventral
ES, embryonic stem
eve, even-skipped
eve MHE, *eve* muscle and heart enhancer
FAIRE, Formaldehyde-Assisted Isolation of Regulatory Elements
FB, fat body
FDR, false discovery rate
FoxA1, Forkhead box protein A1 (or Hepatocyte Nuclear Factor 3a)
GRN, gene regulatory network
hh, hedgehog
HMM, hidden Markov models
HOT, highly occupied target
IgG, immunoglobulin G
INF-β, interferon beta
INFγ, interferon gamma
IRF-3, interferon regulatory factor 3
IRF-7, interferon regulatory factor 7
kb, kilobase pairs
LAD(s), lamina-associated domain(s)
mad, mothers against dpp
mef2, myocyte enhancing factor 2
mRNA, messenger RNA
meRNA, multiexonic poly(A)+ RNA
MZT, maternal-zygotic transition

NDR, nucleosome depleted region
NF-κB, nuclear factor kappa-light-chain-enhancer of activated B cells
NFR, nucleosome free region
p300, EP300 or E1A binding protein p300
PCA, principal component analysis
PCR, polymerase chain reaction
PEAT, Paired End Analysis of Transcription start sites
pMad, phosphorylated Mad
pnr, pannier
Pol II, RNA Polymerase II
PRC2, Polycomb Repressive Complex 2
PSSM, position specific scoring matrix
PTMs, post-translational modification(s)
PWM, position weight matrix
rel, relish
rho, rhomboid
RNA, ribonucleic acid
SELEX, Systematic Evolution of Ligands by Exponential Enrichment
sep, ventral veins lacking
slp, sloppy paired
SM, somatic muscle
sna, snail
sog, short gastrulation
TAF, TBP-associated factor
TBP, TATA-box-binding protein
TF(s), transcription factor(s)
TFBS(s), transcription factor binding site(s)
tin, tinman
TSS(s), transcriptional start site(s)
twi, twist
VM, visceral mesoderm
vnd, ventral nervous system defective
wg, wingless
zen, zerknullt

1 Introduction

1.1 Gene expression regulation and development

Embryonic development is a sequential process that ultimately leads to the formation of complex organs and tissues. All cells of an organism derive from a single cell – the fertilized egg, and thus share an almost identical genome. Nonetheless, cells exhibit a vast array of shapes and sizes, but also play very different roles regarding structural integrity and biochemical function. How can a single cell be at the origin of vastly different cell types and tissues like muscles, neurons or lymphocytes? How, when and why do pluripotent cells decide to specify into a particular cell type? The differentiation process, cell specification and ultimately cellular identity, as well as responses to environmental cues largely rely on the control of gene expression. Thus, different portions of the genome are selectively expressed in different cell types, and the assortment of gene products expressed in one cell type defines its specific characteristics. Precise control of gene expression, both in space and time, is therefore essential to ensure robust developmental programs and maintain tissue physiology.

Synthesis of RNA and proteins is regulated at different levels. In the particular case of coding genes, a gene first needs to be transcribed into a mature RNA molecule. This step requires several conditions to be fulfilled: (1) the DNA sequence must be accessible to allow the transcriptional machinery to load and assemble upstream the adequate transcription start site (TSS), (2) the presence (or absence) of activating (or repressing) transcription factors (TFs) might be necessary to activate transcription, (3) the nascent RNA might require proper splicing, and (4) 5'-capping and adequate 3'-polyadenylation should occur to prevent early RNA degradation and to allow for efficient export from the nucleus. Next, mature mRNAs associate with ribosomes and are translated into proteins, which must fold properly and may be subject to post-translational modification (PTM). Finally, cell state is also a function of individual RNA and protein stability and synthesis rates. Each of these steps may be exquisitely regulated, but transcription initiation is generally considered the major rate-limiting step in eukaryotic gene expression control.

1.1.1 Transcription initiation control and transcription factors

During transcription initiation, the transcriptional machinery first assembles upstream the TSS on the basal (or core) promoter, attracted by a TATA-box or other core motifs [1]. Core promoters are mandatory elements for transcription allowing for proper alignment of the transcription machinery; nevertheless, they cannot generate significant levels of mRNA by themselves and they are rarely the point of gene regulation. Hence, the transcriptional machinery needs additional support to initiate gene transcription. This role is played by TFs, which are proteins able to bind the DNA in a sequence-specific manner. Once bound to DNA, TFs recruit co-factors and the resulting complex is spatially brought into contact with the transcriptional machinery to initiate transcription (Figure 1). Promoters may contain transcription factor binding sites (TFBSs), organized in homo- and/or heterotypic clusters, to recruit TFs in close proximity (100-200 bp upstream) of the loaded transcriptional machinery[2]. TFs can also bind to enhancers, or cis-regulatory modules (CRMs) that may be located far away from their target genes. CRMs integrate cues from both signaling and transcriptional networks and are major actors in establishing the complex spatio-temporal patterns of gene expression[3].

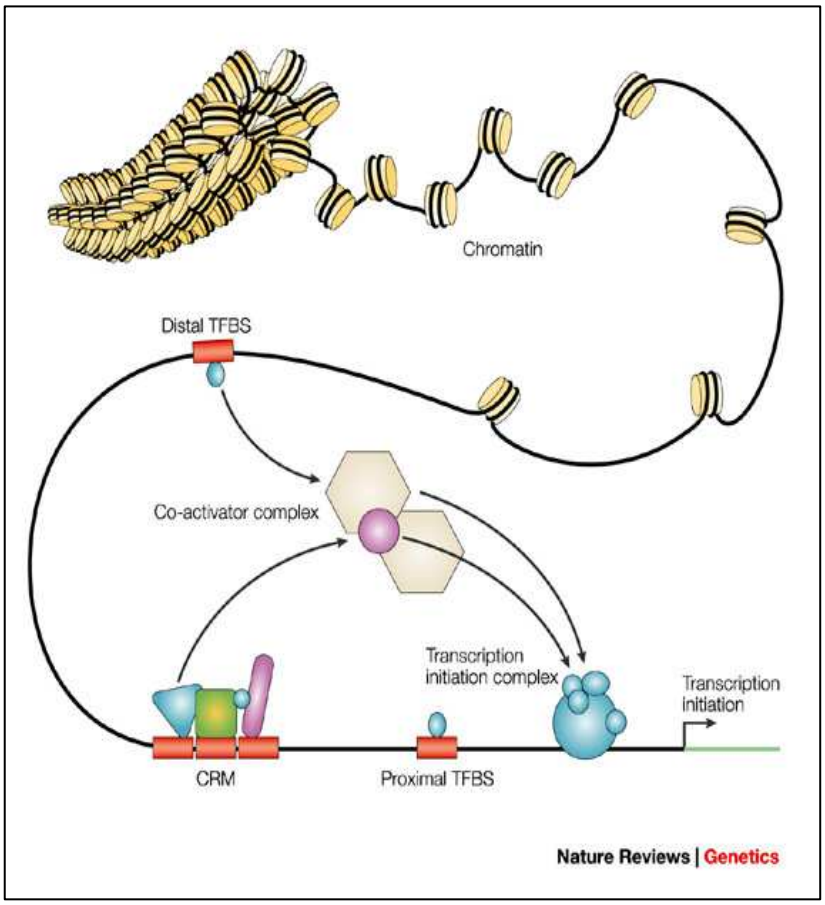


Figure 1. Components of transcriptional regulation.

TFs bind to specific TFBSs that are either proximal or distal to a TSS. Sets of TFs can operate in functional CRMs to achieve specific regulatory properties. Interactions between bound TFs and cofactors stabilize the transcription-initiation machinery to enable gene expression. The regulation that is conferred by sequence-specific TF binding is highly dependent on the three-dimensional structure of chromatin. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics, Wasserman W and Sandelin A, Applied bioinformatics for the identification of regulatory elements, 5, 276-287, copyright 2004.

TFs with DNA binding domains (DBD) bind to DNA in a sequence-specific manner, often in homo- or heterotypic clusters, and promote or repress expression of target genes by interacting with the transcriptional machinery. The sequences recognized by TFs show varying specificity, with some factors binding to very strict sequence motif (for example, the yeast TF Reb1 invariably binds TTACCCG [4]), while other factors binding a wider array of sequences (like the mouse TF Pax4 (JASPAR[5] entry MA0068.1). This recognition specificity is often formalized in terms of a consensus sequence (e.g. the TACCCG Reb1 signature), where the use of IUPAC code indicates flexible motif positions (for example the CAYRTG

Twist (Twi) signature where N is A, C, G, or T, and Y is either C or T; an exhaustive list of IUPAC symbols is available in annexes). A refined motif description makes use of a position-specific scoring matrix (PSSM) or position weight matrix (PWM) [6,7]. Such matrices are visualized using sequence logos[8]. Figure 2 presents how such matrices are built from identified footprints (TFBSs functional *in vivo*) and visualized as sequence logos.

TFs often have interaction domains allowing them to multimerize into homo-, or hetero-multimers and many TFs have been shown to interact with other TFs to cooperatively load onto the DNA (for example Tinman (Tin) with Mothers against dpp (Mad) [9] and with Pannier (Pnr) [10] in *Drosophila,* or Tbx5 with Gata4[11] and with Nkx2-5[12] in mouse). These interactions may modulate the sequence specificity of the TFs[13]. For example, the TF Twi binds the DNA through its basic helix-loop-helix (bHLH) domain that recognizes E-box motifs, CANNTG[14,15]. Twi readily forms homodimers to bind DNA with a strong preference for CACATG and CATATG, or more generally CAYRTG sites, and thereby promotes the expression of its target genes. Twi has also been shown to form heterodimers with a variety of other HLH-containing TFs, including Daughterless[14]; in this context, the complex preferentially binds the CASSTG motif and has a repressive role on its target genes.

TFs are classified into TF families based on the type of their DNA binding domains (Zinc fingers, Helix-Loop-Helix, Homeobox…) and members of the same family may have very similar sequence specificity. For example, Biniou (Bin) and Sloppy paired (Slp) are both members of the Forkhead TF family and have been shown to bind the same sequence profile. Thus, individual binding specificities of TFs of the same family (and resulting *in vivo* function) are thought to be largely acquired by multimerization partners and presence of co-factors in the protein complex that eventually binds the DNA[13].

In vivo, TFs can bind up to several thousands of sites and the overall binding landscape of a particular TF changes with time, thereby reflecting temporal progression during development, cell lineage identity or activation upon specific stimulation. For example, we profiled the genome-wide binding landscape of Twi, a mesoderm specific TF essential for early mesoderm development in *Drosophila*, at two consecutive time points of the early mesoderm development (2-4h and 4-6h after egg laying (AEL)) [15]. In this study we found that Twi binds to ~2000 TFBSs, of which 51% are continuously bound while 23% and 26% are specific to 2-4h and 4-6h conditions, respectively. Furthermore, we demonstrated that Dorsal (Dl) sites were enriched in the proximity of early bound TFBSs only while Tin

sites were enriched in the proximity of late bound TFBSs only, reflecting the collaboration of Twi with these two factors at distinct time points of mesoderm development (dorso-ventral patterning and mesoderm maturation, respectively).

Finally, TFs often bind DNA in absence or displacement of nucleosomes and therefore need the chromatin to be "open" first. How exactly the chromatin is opened in a time and lineage specific way remains a subject of intense research. Nevertheless, *pioneer* TFs like the human FoxA1 (Hepatocyte Nuclear Factor 3a) are able to bind nucleosome-dense DNA and to trigger chromatin remodeling to help recruiting lineage specific factors at CRMs[16].

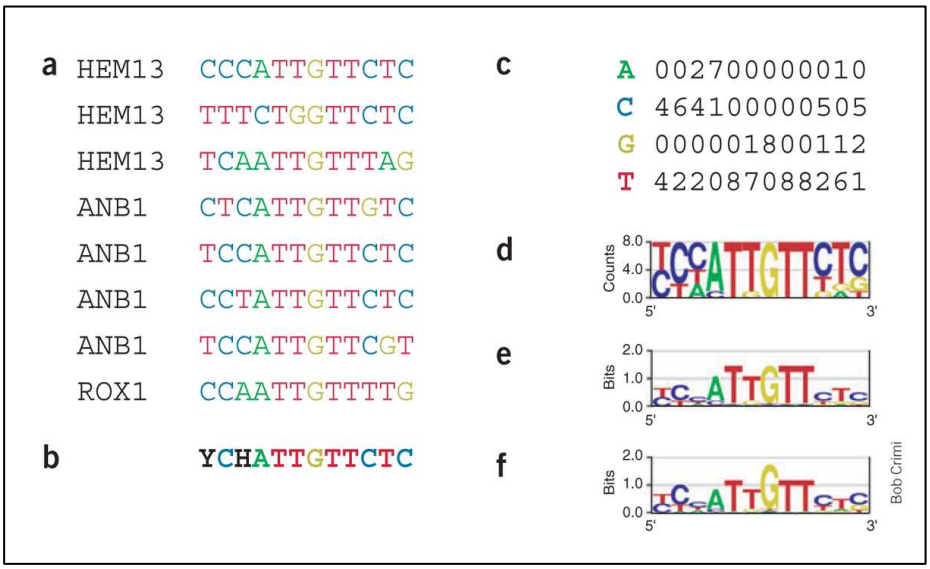


Figure 2. Representation of TF sequence specificity.

(a) Eight known genomic binding sites in three *S. cerevisiae* genes. (b) Degenerate consensus sequence. (c,d) Frequencies of nucleotides at each position. (e) Sequence logo showing the frequencies scaled relative to the information content (measure of conservation) at each position. (f) Energy normalized logo using relative entropy to adjust for low GC content in *S. cerevisiae*. Reprinted by permission from Macmillan Publishers Ltd: Nature Biotechnology, Patrick D'haeseleer, What are DNA sequence motifs?, **24**, 423 - 425, copyright 2006.

1.1.2 cis-Regulatory Modules

"Since the initial discovery of enhancers, it has been known that they are most often the dominant element in conferring tissue specificity to a linked gene. A hallmark of most

enhancers is their ability to activate transcription from any linked promoter in reporter gene constructs, even if promoter and enhancer originate from gene loci with completely different expression patterns *in vivo*. Although there are exceptions to the general principle, expression of the reporter gene follows the pattern governed by the enhancer, not the promoter" [3]. This excerpt from Bulger and Groudine underlines the key role that CRMs and enhancers play in gene expression program and during development in particular. Moreover it has now been shown that mutations in CRM sequence can cause or contribute to human disease. For example, such CRM mutations were associated with thalassaemias that result from deletions or rearrangements of enhancers of the β-globin gene, preaxial polydactyly resulting from sonic hedgehog limb-enhancer point mutations, and susceptibility to Hirschsprung's disease associated with a RET proto-oncogene enhancer variant[17]. So what are CRMs exactly?

CRMs are short regulatory elements (50-500 bp) driving a particular aspect of a gene expression in response to TFs[18] that bind TFBSs within the CRMs in a sequence-specific manner. CRMs can be found at large distances of their target genes (distal elements) as well as in introns and promoters (proximal elements) and they are generally considered to modulate gene expression regardless of their orientation or relative position to the TSS[19,20]. CRMs commonly have TFBSs for a variety of TFs[21]. The binding of TFs can have both positive and negative effects on the target gene expression depending on the activating or repressing nature of the TF(s). Though TFs are typically considered either as activating or repressing, several cases of TFs functioning as both, activator and repressor, depending on the specific context have been reported[22,23]. CRMs operate at different times during an organism's life, reflecting the transient presence of particular TFs, activator and repressor concentration balance, presence of co-factors or simply different accessibility of the genome. For example, Wilczynski and Furlong recently showed that the dynamic CRM occupancy by mesodermal TFs tightly reflects developmental progression [24]; in particular, the temporal changes in TF binding correlate with dynamic patterns of target gene expression. Thus, gene expression patterns are not only explained by the timing of TF availability, but also by their exact temporal occupancy. Overall, spatio-temporal expression of a gene is explained by the combination of all the CRMs acting on it throughout the organism's life. For example, the *Drosophila* gene encoding the TF Tin has at the very least 4 different CRMs controlling its expression in embryonic development, each driving a particular aspect of its spatio-temporal pattern (Figure 3). Housekeeping genes are not exempt of gene expression modulation, in

particular in term of expression level (expression can, for example, be fully turned off in response to extreme conditions such as heat shock), and are therefore also under CRM control. Nevertheless, the most complex spatio-temporal expression patterns are characteristics of developmental genes[2,3].

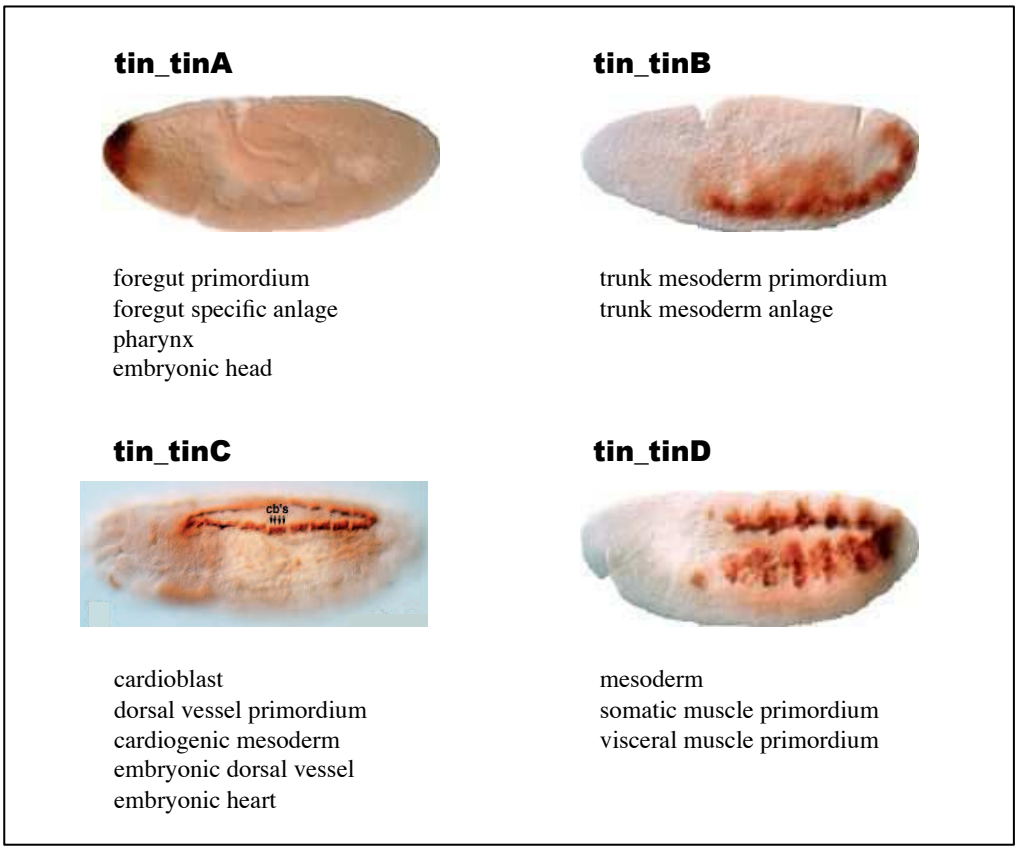


Figure 3. Each *Drosophila* Tinman enhancer drives a specific pattern of Tinman's expression.

Expression patterns of four Tin enhancers (tin_tinA, tin_tinB, tin_tinC, tin_tinD) in *Drosophila* embryos together with anatomical annotations. Embryos are oriented dorsal up, anterior left. Pictures and anatomical annotations were obtained from the REDfly[25] database and Yin *et al*.[26] for tin_tinC. This figure illustrates how complex spatio-temporal patterns are established by distinct enhancer elements (cb's : cardioblasts).

1.1.2.1 CRM architecture

As mentioned above, CRMs are often composed of multiple TFBSs for different TFs. In addition, numerous studies have shown that TFs frequently interact with each other. A fundamental question is therefore to understand if TFBSs need to be arranged in a specific manner to allow TFs to bind DNA cooperatively. Tentatively, a tight architecture, or

grammar, might promote protein-protein interactions between DNA binding domains themselves thereby enabling a mutual stabilization of the overall protein-DNA interactions. Protein-protein interactions may yield protein complexes where the DNA binding characteristics of complex components may be distinct from the DNA binding characteristics of component proteins in isolation. A multi-protein complex might also be constrained to bind to individual TFBSs arranged in a spatially defined way, which could impinge on interaction and coordination with the transcriptional machinery. X-Ray studies revealed that some TFs interact with DNA using the DNA major groove, the minor groove being the place of secondary contacts thought to modulate binding strength. Alternatively, other TFs interact mainly with the minor groove. It is therefore expected that cooperative occupancy would require topological features like specific relative orientation, helical phasing or spacing of TFBSs. On the other hand, DNA is rather flexible: in typical eukaryotic cells, it is coiled around a core of histones spanning only ~90 bp . It is therefore also possible to bring partners together simply by twisting DNA without the need of particular grammar. In fact, both situations have been observed and described in the literature.

The classic example of constrained architectural organization is the enhanceosome model of enhancer activation of the eukaryotic IFN-β gene[27], expression of which is induced upon viral infection. Activation of the IFN-β gene by its enhancer requires the coordinate activation and binding of the ATF-2/c-Jun, IRF-3, IRF-7 and NFκB (i.e. p50/RelA) TFs[28] in the enhancer region located from -102 to -47 bp upstream the TSS. In its active configuration, the enhancer is devoid of nucleosomes[27]. In this enhancer, the 8 individual TFBSs exhibit strict positional requirements and overlap each other substantially (Figure 4). This organization allows for cooperative binding and assembly of the activators into a protein complex called the 'enhanceosome'. Formation of this enhanceosome is only possible in the presence of all TFs and does not tolerate changes in TFBS spacing. The enhancer overall acts as a functional unit articulated around IRF-7, the master regulator of type-I interferon-dependent immune response[29]. In particular, individual TFs are not able to activate the IFN-β gene[27], mutations in any of the IRF TFBSs terminate the transcription[30] and absence of either IRF-3 or IRF-7 prevents induction of IFN-β[29]. These unique features explain why this enhancer is evolutionary conserved and why modification of virtually any nucleotide impacts on the enhancer's activity[27]. They also led Panne *et al.* to initially hypothesize that direct protein-protein interactions between adjacent DNA binding domains underly this

cooperation[31]. It is only recently that Panne *et al.* constructed a complete atomic model of the fully assembled enhancer[32]. This study revealed that this structure is largely devoid of major protein-protein interactions between adjacently bound DNA-binding domains. Rather, cooperative binding is mediated by local DNA conformation changes (induced by the binding of one activator) that favor binding of the different activators, including at nonconsensus sites. The enhanceosome protein complex is further stabilized by the interaction of each of the TFs with the coactivator CREB binding protein (CBP, or its paralog p300) through their activation domains.

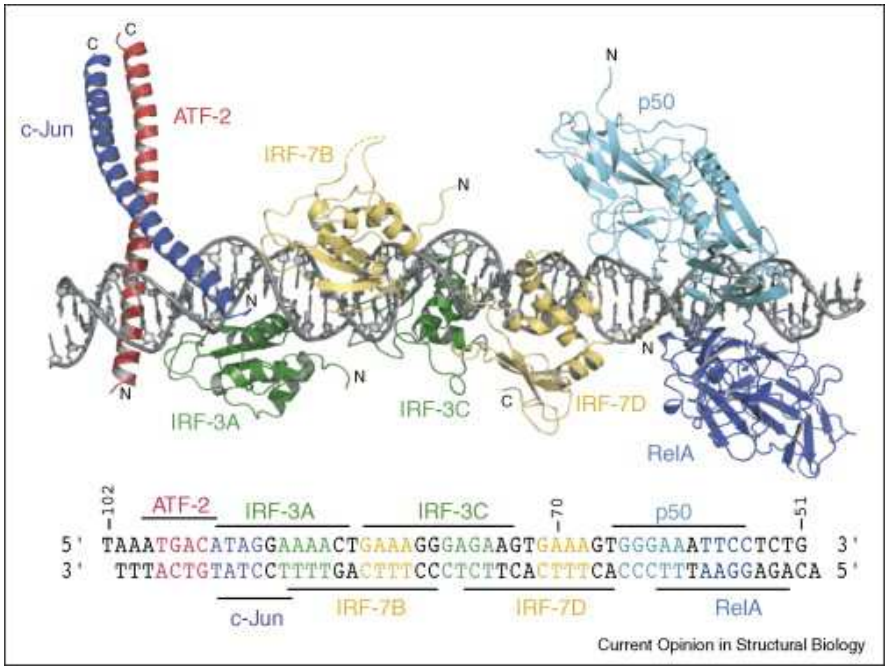


Figure 4. Atomic model of the INF-β enhanceosome. The p50 is in light blue and RelA in dark blue. IRF-7B and IRF-7D are in yellow and IRF-3A and IRF-3C are in green. ATF-2 is in red and c-Jun in blue. The DNA sequence is shown with the core-binding sites colored accordingly. Reprinted from Current Opinion in Structural Biology, Volume 18, Daniel Panne, The enhanceosome, Pages 236-242, Copyright (2008), with permission from Elsevier.

In 2004, Senger *at al.* suggested that the synergy between Rel-containing proteins (Dorsal, Dorsal-related immunity factor (Dif) and Relish (Rel)) and the GATA factor Serpent is essential for the activation of several immunity genes in the *Drosophila* fat body. The authors showed that about half of these immunity genes exhibit constrained structural features, similar in essence to the enhanceosome model, in which Rel and GATA binding sites are positioned in the same orientation. In addition, they showed that mutations that flip

either Rel or GATA site orientation abolish the reporter gene activity in transient transfection assays[33].

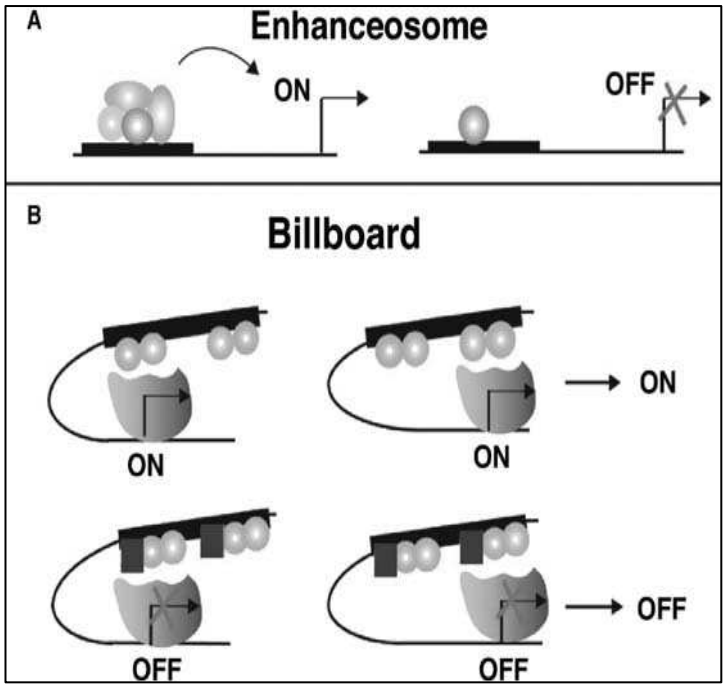


Figure 5. The enhanceosome and billboard models.
A: In the Enhanceosome model, the binding sites within the enhancer allow for a highly cooperative assembly of TFs (ovals), leading to gene activation. Disruption or displacement of a single binding site, or the absence of one regulatory protein, causes the element to be inactive. **B**: In the Billboard model, the enhancer contains multiple functional units that are able to independently regulate gene expression. Above, activators (colored ovals) located in separate portions of the enhancer are "sampled" by the basal machinery, and the integration of such interactions results in total gene output. Below, regulation by short-range repressors. Individual sub-elements of the enhancer are repressed by the action of short-range repressors (squares) located near each cluster of activators. Note that an intermediate, 'partially on' situation might be achieved when only one of the two sub-elements is repressed. Reprinted from Arnosti D. and Kulkarni M., Journal of Cellular Biochemistry, Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?, Volume 94, Issue 5, Pages 890-898, doi:10.1002/jcb.20352, http://onlinelibrary.wiley.com/doi/10.1002/jcb.20352/full, Copyright (2005), with permission from Wiley.

Taken together, the aforementioned studies and the concept of an enhancer grammar fits well with the hypothesis that enhancers work as information-processing devices, which integrate multiple inputs (both positive and negative) through TF binding into a single binary (on/off) output[34]. In this model, the enhancer is the active regulatory device, while the basal transcription machinery plays a more passive and permissive role. However, this strict organization within the enhancer sequence of the enhanceosome model (and its associated stringent binary mode of regulation) may represent only a small fraction of enhancers. Indeed, many developmental enhancers display no or much looser architectural constraints,

where a subset of factors may bind cooperatively while the remaining factors are recruited independently. In 2003, Kulkarni and Arnosti proposed an alternate model (Figure 5), in which the enhancer acts as an information display (as opposed to an information integration platform) potentially presenting both active and repressed states to the basal transcription machinery[34]. Using enhancer constructs containing different numbers of sites for activators and short-range repressors, the authors showed that (1) the transcriptional outcome varies with the number of activator sites given a fixed number of repressive sites, (2) repression and activation can happen simultaneously, and (3) displayed information might be redundant. In this 'billboard' model, the basal transcription machinery plays an active role in interpreting signals presented by the enhancers by sampling this displayed information. The fundamental difference between this model and the enhanceosome model lies in the inherent flexibility of the former, which is thought to allow for more diversity in gene expression (modulation of activation level) and evolutionary flexibility (the architectural flexibility allowing for the emergence of new patterns of activity).

In some cases, enhancer flexibility appears even more extreme than in the billboard model. In *Ciona intestinalis*, 19 muscle genes are coexpressed in the 36 muscle cells of the developing embryo. 17 of these 19 gene products participate in the same macro-molecular complex and are therefore under tight coexpression control. Brown *et al.* took advantage of this system to investigate the functional architecture of the 19 enhancers controlling these 19 muscle genes[35]. The authors systematically mutated the TFBSs found in these CRMs and assessed their individual *in vivo* activity using regression models, which allowed the authors to identify important TFBSs and quantify their activity. Focusing on functional TFBSs, the authors could not find any lexical features such as TFBS order, spacing or relative orientation. Overall, these CRMs are composed of TFBSs of widely varying quantitative activity, found in diverse arrangements and from different combinations of motif types. Strikingly, the authors showed that different *Ciona* muscle enhancers can achieve the same function with widely different architectures, yet that functional architectures are preserved in orthologous enhancers with important TFBSs being more conserved[35].

Using genome-wide binding maps for 5 key mesodermal TFs generated at 5 consecutive time points in the *Drosophila* developing embryo, we have recently

demonstrated that spatio-temporal activity of enhancers can be predicted from TF binding solely[36]. This study revealed an unanticipated plasticity in TF binding (in terms of TF identity and binding dynamics) leading to similar expression patterns. Along with those obtained in *Ciona*[35], our results question the generally assumed stringency of regulatory codes and suggest that architectural flexibility may represent an inherent property of developmental *cis*-regulatory modules.

1.1.2.2 CRM conservation

CRMs play a crucial role in the regulation of precise gene expression patterns both in space and time. A number of key TFs and gene regulatory networks (GRNs) are shared among organisms and are sometimes well-conserved, such as the cardiac regulatory network[37,38]. Furthermore, several complex gene expression patterns have been shown to be under evolutionary constraint. For example, the stripe pattern of the *pair rule* genes observed in the *Drosophila melanogaster* embryo are generally conserved among drosophilids[39]. Hence, many enhancers are likely to be conserved over the course of evolution to preserve fundamental gene regulatory interactions. This assumption is a central tenet of *in silico* CRM prediction methodology, where sequence conservation is used as a guide (discussed later in section 1.2.1.2). Indeed, 50% of 'ultraconserved' elements (perfect sequence identity of at least 200 bp between very distant organisms like human and mouse/rat) as well as 50% of 'extremely conserved' elements (sequences with slightly less-than-perfect extended identity) have been shown to be capable of driving expression during embryonic development[40]. The enhanceosome model of IFN-β enhancer is another example of near-perfect conservation[32]. Though sequence conservation can be used to detect regulatory sequences, it is unclear what fraction of enhancers could be discovered using this approach. For example, less than 2% of tested ultraconserved elements acted as heart enhancers, compared to ultraconserved elements acting as limb, midbrain, or forebrain enhancers (5%, 14% and 16%, respectively) [41]. Besides the technical hurdle of reliably aligning genomes at various phylogenetic distances, many reports indicate that CRMs are not necessarily under selection pressure. In fact, different studies reported CRM functional conservation without overall significant conservation at the sequence level[35,42-44].

In the previously mentioned *Ciano intestinalis* study[35], the authors showed that the 19 enhancers driving similar expression patterns (i.e. expression in muscle cells) have widely different architectures. Strikingly, individual CRM architectures are preserved in orthologous enhancers found in *C. savignyi* (note that the neutral sequence divergence between these 2 species is about that between mammals and birds), with important TFBSs being much more conserved than expected (with more than 79% pairwise sequence identity between orthologous functional TFBSs compared to a background sequence identity of less than 20%). Importantly, pairwise sequence identity quickly drops off outside the boundaries of the functional TFBSs to reach the background level within only 12 bp.

Hare *et al.* compared enhancers of the *even-skipped* locus between *Drosophila* and highly diverged scavenger flies (that diverged 100 million year ago). The authors could show that the *Sepsid* and *Drosophila eve* enhancers have almost identical expression patterns in transgenic *D. melanogaster* embryos, while no significant sequence similarity is observed, but for a small number of short (20-30 bp) sequences that are almost perfectly conserved[43]. Interestingly, the authors reported that these highly conserved short sequences are enriched for pairs of adjacent or overlapping TFBSs and might therefore represent key architectural elements.

In a different study, Ho *et al.* compared the CRMs for the *Abdominal-B* gene from different *Drosophila* species[44]. Similarly, these authors reported low levels of overall sequence conservation while enhancers remained fully functional and drove identical spatio-temporal expression patterns in transgenic *D. melanogaster* embryos. Again, functionally critical TFBSs were highly conserved.

Altogether, these results suggest that while CRMs usually have low levels of overall sequence conservation, the critical TFBSs or architectural features within CRMs are conserved. This is also in line with the conclusions of the report by Parker *et al.* in which the authors assessed the evolutionary conservation of DNA structure rather than DNA sequence[45]. The 'Chai' algorithm developed by these authors measures constraint on the basis of similarity of DNA topography among multiple species and is based on the 'hydroxyl radical cleavage pattern', a metric that quantifies the solvent-accessible surface area of duplex DNA[46]. Regions identified by Chai (i.e. regions that are highly constrained topographically) correlated with enhancers better than did regions identified solely on the basis of nucleotide sequence, indicating that local structure conservation might be critical for

enhancer function.

1.1.3 Gene expression and chromatin

1.1.3.1 Chromatin structure

In eukaryotic nuclei, DNA is associated with histone proteins in a structure called the 'chromatin'. The nucleosome is the fundamental repeating block composing the chromatin. Each nucleosome core is made of 145-147 bp of DNA wrapped in 1.7 superhelical turns around a core histone octamer and occurs, on average, every 200±40 bp throughout the genome[47]. The DNA between two of these nucleosomes is referred to as the 'linker DNA' and can be loosely associated with an additional H1 'linker histone'. The core nucleosome contains 2 copies of each of the H2A, H2B, H3 and H4 histone proteins assembled into two H3-H4 dimers bridged together as a stable tetramer, which is flanked by two separate H2A-H2B dimers[47]. In addition, histones have N-terminal tails that extend beyond the nucleosome particle, which are the place of specific covalent PTMs such as methylation, acetylation, phosphorylation, ubiquitination, SUMOylation, citrullination, and ADP-ribosylation. Positively charged histone tails interact with the DNA (negatively charged) and it has been suggested that H3 and H2A tails are important for nucleosome structure and stability[48].

Histones are essential for efficient packaging and compaction of the genomic DNA into a 3D organization that fits within the nucleus. The first layer of this packaging, composed of nucleosomes, is often described as "beads on a string" – a fiber with a diameter of 11 nm. The second level of compaction, the 30 nm fiber, requires the linker histone H1 (or H5), which stabilizes interactions between 11 nm fibers (see Figure 1 for an illustration). Finally, the higher order levels of chromatin organization still remain poorly understood. Chromatin organization and more precisely its compaction level is an important feature influencing different cellular processes, including replication, gene expression, and DNA repair. These processes usually need the chromatin to be 'open' (i.e. accessible) to allow access of various proteins to the DNA. Many studies have shown that chromatin compaction

is modulated by core histone PTMs and the presence of histone variants in the core nucleosome.

Transcription has been linked to histone acetylation more than 20 years ago[49] and chromatin cannot fold into the 30 nm fiber when histones are acetylated[50]. As transcription proceeds, nucleosomes are thought to be displaced from the DNA. Nucleosome dynamics are a function of various parameters, such as DNA methylation, core histone PTMs and incorporation of histone variants (reviewed in [51]). For example, the *Drosophila* H3.3 variant is preferentially incorporated instead of the canonical H3 into nucleosomes located around TSSs of active genes, at a rate proportional to gene activity, thereby reflecting nucleosome disruption and reassembly during transcription[52]. In the same study, the authors also reported that promoters of active genes are depleted of nucleosomes in a region of about 100-200 bp upstream the TSS[52]. The presence of such a nucleosome-depleted region (NDR) (initially referred to as nucleosome free region (NFR)) at the TSS of active genes is not systematic though. Indeed, data produced using the Paired End Analysis of Transcription Start Sites (PEAT) methodology[53] (an extension of Cap Analysis of Gene Expression (CAGE) allowing to map the exact transcription start position) revealed that both flies and mammals have two types of promoters differing by the variability of the transcription starting positions of a particular TSS: the ‘‘focused’’ promoters (narrowly defined transcription starts, transcription starts here refer to the different observed starts for different transcript molecules transcribed from the same TSS) and the ‘‘dispersed’’ promoters (transcription starts spreading over a larger window). Rachel *et al.* found that NDRs are not hallmarks of active genes in general, but that they are more typically associated with active genes that exhibit “dispersed” promoters, this in both *D. melanogaster* and *H. sapiens*[54].

As previously mentioned, cell types differ in their gene expression programs, which includes the transcriptional silencing or activation of large genomic regions. Accordingly, such regions tend to be associated with typical structural features. For example, silenced regions typically exhibit higher chromatin condensation than active regions, a feature that can be visualized beautifully in *Drosophila* polytene chromosome spreads. Polytene chromosomes from *Drosophila* salivary glands consist of 1024 copies of in-register aligned chromosomes, which can be extracted, stained and microscopically visualized. Any stain that associates with DNA can be used to demonstrate that different regions exhibit different densities, i.e. condensation states. Several studies have shown that the denser, more tightly

packaged regions tend to harbor silenced genes. More recently, it has been shown that condensation state and gene activity strongly correlate with chromatin state, that is particular combinations of histone PTM and association with regulatory proteins, such as enzymes that catalyze the addition or removal of PTMs. Additionally, active (like transcription factories[55]) and inactive[56] chromatin regions have even been shown to be structured in terms of their sub-nuclear localization. Large chromatin domains, in which the lysine 9 of the H3 tail are largely di- and trimethylated (conventionally written as H3K9me2 and H3K9me3), have been demonstrated to associate with the nuclear lamina[56,57]. These lamina-associated domains (LADs) are largely transcriptionally inactive and tethering experiments have further demonstrated that recruiting active genes to the nuclear lamina is causal in reducing their activity[58,59].

1.1.3.2 Histone post-translational modifications

Since the discovery of nucleosomes, our understanding of their role has expanded from simple static DNA-packaging elements to key dynamic components involved in a wide array of genomic functions. In the early 1990s, histone tail PTMs, in particular acetylation, was linked with both transient (e.g. local increase of acetylation upon activation of inducible genes) and long-term maintenance of transcription states (e.g. X chromosome inactivation in female mammals, dosage compensation in *Drosophila* males, Polycomb- and Trithorax-mediated maintenance of transcriptional states of individual loci) [49]. Since then, PTMs of histone tails are thought to represent a mechanism to encode and transmit information across cell generations; in other words an epigenetic code. The correlation of distinct PTMs with predictable functional outcomes[49], the large number of different PTMs and the existence of residue-specific enzymes to either add or remove these PTMs, as well as the regulated manner of PTM deposition led investigators to formally propose the existence of a histone code[60,61]. In their hypothesis, unique combinations of PTMs act together to form chromatin states and regulate unique biological outcomes by affecting the local structure of the chromatin. With the advent of genome-wide chromatin immunoprecipitation (ChIP) technologies (see Figure 6 for an overview of these protocols), these PTMs – in particular methylation and acetylation of histone tails – have been studied at an unprecedented level

over the last years[62-65]. These studies, and others, have allowed for the association of individual PTMs, or of particular PTM combinations with various genomic features (e.g., genes, promoters, enhancers), as well as with their transcriptional states (on/off).

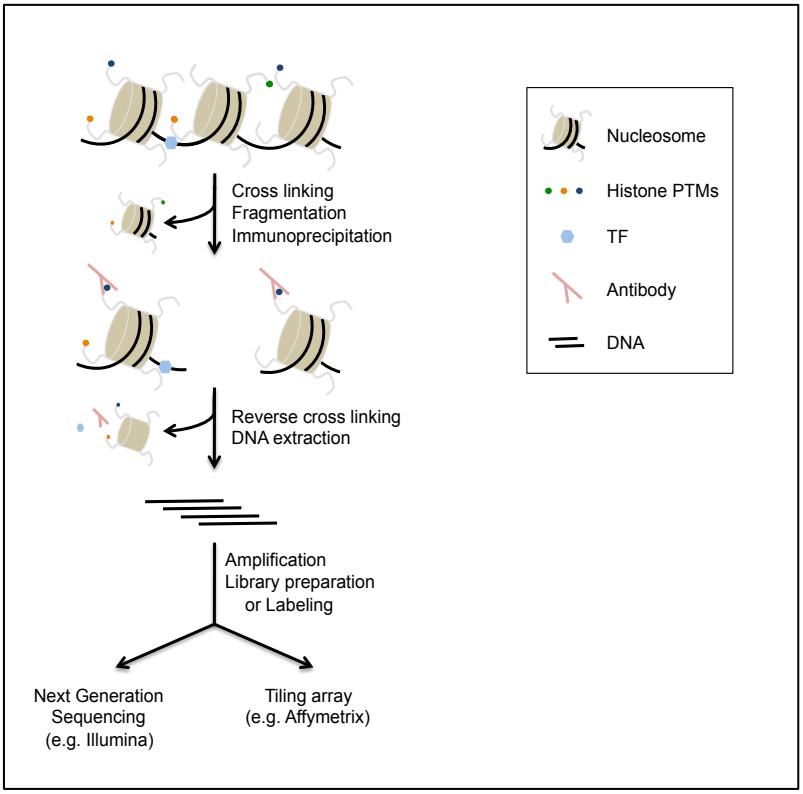


Figure 6. Overview of ChIP-chip and ChIP-seq protocols.

The chromatin is cross-linked (e.g. by formaldehyde), fragmented (e.g. by sonication) and immunoprecipitated using an antibody specific to an epitope of interest, such as a particular histone PTM (dark blue, green and orange spheres), or a particular TF (cyan hexagon). The nucleoprotein complexes are reverse cross-linked and the DNA is extracted. During library preparation, the DNA can be size-selected prior to PCR amplification (typically 200 and 500 bp fragments are selected for Illumina® sequencing and microarray hybridization, respectively). For ChIP followed by deep sequencing (ChIP-seq), sequence adapters are added to DNA fragments and fragment ends are sequenced. For ChIP followed by microarray hybridization (ChIP-chip), amplified fragments are labeled with a fluorophore and hybridized to a microarray (i.e. a microarray containing probes that tile across the genome). A reference sample is usually generated in parallel following the same protocol in which the immunoprecipitation step is simply omitted or in which the specific antibody is replaced with an non-specific antibody (like IgG) or, if available, with the pre-immune serum (also called a mock).

Transcriptionally inactive genomic regions, for example, tend to be enriched in nucleosomes carrying particular PTMs, such as H3K9me2, H3K9me3, and H3K27me3 [56,57,62]. H3K27 is trimethylated by Polycomb Repressive Complex 2 (PRC2). H3K27me3 is usually found on large regions spanning several dozens of kb that overlap silent genes and intergenic regions. Similarly to H3K9me2 domains that mark LADs, H3K27me3 regions

might group together within the transcriptionally silent structures called Polycomb bodies[66,67]. Studies in human embryonic stem (ES) cells and in differentiated cells suggest that H3K27me3 repressive domains are first seeded within ES cells during the initial phases of differentiation, but expand and are established differentially in concordance with cell type over the course of differentiation, thus reflecting their specific repression needs[68]. Dynamic changes in H3K27me3 marked domains during development has also been observed recently in plant by comparing H3K27me3 genome wide profiles between undifferentiated cells of the shoot apical meristem and differentiated leaf cells[69]. In addition to this typical pattern of broad enrichment domains, H3K27me3 has also been found as focused peaks at the TSSs of bivalent promoters[64,70,71], i.e. promoters holding both the repressive H3K27me3 and the activating H3K4me3 marks (see below). Bivalent promoters have been observed in mammalian ES cells[70] and correspond to important developmental genes with low or no expression in ES cells. The current understanding is that these promoters are in a poised chromatin state; upon differentiation, these promoters become either active or repressed but do not remain bivalent[72].

H3K4me3 associates with actively transcribed genes in all organisms studied so far [62,73-77]. H3K4me3 is found on the one to two nucleosomes downstream of active TSSs and therefore appears as very localized peaks, where enrichment levels tend to correlate with gene expression levels[62]. In yeast, this mark is deposited by the Set1 histone methyl transferase and it has been shown that mutants affecting the elongation but not the formation of the pre-initiation complex cannot recruit Set1 efficiently, suggesting that H3K4me3 is associated with transcriptional elongation[77]. In contrast, Guenther *et al.* have reported that H3K4me3 is present, together with RNA polymerase II (Pol II), H3K9Ac and H3K14Ac on promoters of most of the active and inactive genes in human undifferentiated ES cells, as well as in differentiated cells (primary hepatocytes and B cells). The authors showed that most of these inactive genes undergo transcription initiation without elongation and consequently linked H3K4me3 with transcription initiation[78]. Finally, H3K4me3 was recently reported to be present on active enhancers in mouse T-lymphoid cells[79]. These studies, however, disagree with studies performed in various human cell lines in which H3K4me3 is identified as a promoter specific mark and was used to differentiate between active and inactive promoters[62,63,65,80].

Other histone methylation[62,75,81,82] and acetylation[63,65] marks have been extensively profiled and studied in the context of gene transcription. H3K4me2 and H3K4me1 have been reported on active genes (with H3K4me2 signal downstream of the TSS proximal to H3K4me3, and with H3K4me1 located even further downstream) – their levels usually correlate with gene expression levels[62,65,75,81]. Interestingly, Bernstein *et al.* reported that, in human and mouse, H3K4me2 is also found in the vicinity of active genes, but not necessarily within gene bodies. They also reported that methylated profiles (H3K4me2 or H3K4me3) are more conserved between human and mouse than the corresponding genomic sequence, an observation that holds true even for intergenic methylated regions[82], suggesting H3K4me2 presence on regulatory regions. In all organisms studied so far, H3K36me3 is found within the body of transcribed genes and primarily occurs on exons[83], with a signal profile skewed towards the 3' end of genes (background level signal is frequently observed at the TSS, in particular for longer genes) [62,81]. Additional methylation marks have been surveyed in human T cells: H3K9me1, H3K20me1, H2BK5me1 and H3K27me1 are associated with active transcription, while H3K36me1, H3K79me1, H3R2me1, H3R2me2 and H3K20me3 are not[62]. H3K79me3 is enriched on coding genes in both yeast and human. However, while no correlation with gene activity has been observed in yeast [75,81], a clear positive correlation with the transcription rate was reported close by the TSS in human[62,84]. The H3K79me2 modification seems to be less universal: while H3K79me2 is present on almost all nucleosomes in yeast[75] and its presence does not correlate with transcription in humans[62], in *Drosophila* it correlates well with gene activity.

Acetylation of H3 and H4 had been shown to correlate with open chromatin and gene activity before individual acetylation modifications could be profiled[49,75]. Like H3K4me3, H3K9Ac and H3K14Ac are present at the TSS of active genes and their levels positively correlate with gene activity in yeast, human and mouse[81,82]. Lastly, Wang *et al.* profiled 18 distinct histone acetylation marks genome wide in human T cells and showed that H2AK9Ac, H2BK5Ac, H3K9Ac, H3K18Ac, H3K27Ac, H3K36Ac and H4K91Ac are mainly located around TSSs, whereas H2BK12Ac, H2BK20Ac, H2BK120Ac, H3K4Ac, H4K5Ac, H4K8Ac, H4K12Ac and H4K16Ac are enriched in the promoter and transcribed regions of active genes[63].

Many studies have convincingly demonstrated that histone PTMs correlate both with transcriptional states and genomic features. It is important to note that these correlations are sometimes organism-specific and that different marks can be associated with the same genomic feature e.g. H3K4me3, H3K9Ac and H3K14Ac. A natural question to ask is therefore how redundant are these marks? In other words, how many different combinations, i.e. *chromatin states*, do exist? In the next section, I will review recent studies in which the authors integrated histone PTMs, Pol II and TF maps together to uncover major chromatin states.

1.1.3.3 Uncovering the different chromatin states

Recent technological developments have allowed the study of chromatin components at an unprecedented scale. In the original publications looking at genome-wide distribution of histone PTMs, authors usually correlated the location of these marks with functional elements of the genome, including TSSs, genes, exons or enhancers. This showed that histone PTMs correlate with functional elements and that some redundancy exists between these marks (i.e. some marks correlate identically with investigated genomic features). With the accumulation of genome-wide maps of histone DNA-associated factors (histone PTMs, insulators, chromatin remodelers…), efforts were made to integrate these data sets in probabilistic and unsupervised frameworks. The aim was to discover the number of significant mark combinations – or *chromatin states* – present in a genome *de novo*. For clarity, it was important to reduce the dimensionality of the data by defining meaningful combinations of marks (i.e. ignoring marks harboring limited or no relevant information) that both correlate and distinguish functional features. Such *de novo* approaches also have the potential to *automatically* discover mark combinations corresponding to particular genomic features or even new features that one could not uncover using a feature-based supervised approach.

A very successful approach was the use of multivariate Hidden Markov Models (HMMs) to combine data from *Drosophila*[85] or human[86,87] cell lines. Practically, this approach divides the genome into nucleosome size intervals (200 bp) and each signal map is

converted into a binary vector representing the absence/presence of the mark in the defined intervals (absence/presence calls were made using a statistical test based on a Poisson distribution). The resulting matrix is then used to learn a multivariate HMM having a fixed number of hidden states. The downside of this approach is that the model does not discover the number of states by itself and the authors must therefore *find* the appropriate number of states to be used. To this end, the authors systematically learned HMMs with different number of states (for example, from 2 to 80 in [86]), with randomly selected initial parameters. Using model log likelihood, the authors selected the best HMM and iteratively removed states (removing the most redundant states first) and, guided by correlation with functional genome annotations, eventually chose the final HMM with N states (where N is the final number of states). Note that this final step might be slightly contradictory with a plain *de novo* approach. The learned model is used to give each 200 bp interval a posterior probability of belonging to each of the N states. Finally each interval is assigned to the state having the maximum posterior probability (note that more stringent criteria can be applied to avoid dubious assignments; for example when the best posterior probabilities are very close).

Using this approach, Ernst *et al.* integrated 18 acetyl modifications, 20 methyl modifications, H2A.Z, CTCF and Pol II ChIP-seq maps as assayed in human CD4+ T cells (published by[62,63]) into 51 chromatin states that correlated with promoters, transcribed regions, active intergenic regions, large-scale repressed domains, and repetitive sequences[86]. A closer look at the 11 promoter states revealed that they were all marked by H3K4me3, various acetylation marks and various combinations and levels of H3K79me2/3, H4K20me1, H3K4me1/2 and H3K9me1 (as a function of the promoter proximity to TSS). The 17 transcription-associated states were defined by various combinations of H3K79me3, H3K79me2, H3K79me1, H3K27me1, H2BK5me1, H4K20me1 and H3K36me3; some of these states specifically correlated with spliced exons, transcription end sites, or zinc finger genes. The 11 active intergenic states were associated with higher frequencies of H3K4me1 (other methylation mark frequencies were reduced), H2A.Z, numerous acetylation marks and/or CTCF. Interestingly, the authors noted that, in these active intergenic states, levels of acetylation marks and H2A.Z correlated with the expression of the closest gene. The 5 large-scale repressed states together covered 64% of the genome and were largely associated with H3K27me3 and H3K9me3. H3K9me3 and H3K20me3 were the major determinants of states associated with repeat elements.

Using a similar approach, Kharchenko *et al.* trained a 9-state HMM from 18 maps in *Drosophila* S2 and BG3 cell lines[85,88]. Of note, the authors also generated a finer-grained 30-state model following the same strategy as in[86] (see previous paragraph) but did not report major differences to the simple 9-state model and state that *"the final number of states was chosen for optimal interpretability"* (in Methods section of Kharchenko *et al.*). Mainly, these 9 states were associated with (1) active promoters and TSSs (H3K4me2, H3K4me3 and H3K9Ac, state 1), (2) transcriptional elongation with exonic preference (H3K36me3 and H3K79me1 and H3K79me2, state 2), (3) intron-biased regions with high enrichment of H3K27Ac, H3K4me1 and H3K18Ac as well as presence of H3K4me2, H3K9Ac and H3K16Ac (this state resembles active mammalian enhancer signature, state 3) or a similar combination without H3K27Ac but with H3K36me1 (state 4), (4) pericentromeric heterochromatin or hetechromatin-like domains marked by H3K9me2 and H3K9me3 (states 7 and 8), (5) Polycomb-mediated repression domains characterized by the presence of H3K27me3 (state 6) and (6) silent chromatin that exhibits low levels of H3K27me3 (state 9).

A striking difference between these two studies is the number of states that the authors selected: 51 versus 9 (or 30). Nevertheless, major functional domains are present in both situations: active TSSs, transcribed units, silent chromatin, repressed chromatin and heterochromatin. Surprisingly, the active intergenic states from Ernst *et al.* were not clearly represented in Kharchenko *et al.* who reported an intron-biased state 3 although the signatures present in these states resemble each other and match enhancer-like signatures (see next section). A number of aspects in the experimental setup accounts for these different state numbers: (1) the use of cell lines from different organisms, (2) the use of different technological platforms (ChIP-Seq[86] versus ChIP-chip[85]), (3) the different number and nature of the markers used (40[86] versus 18[85]), (4) the algorithm used, and (5) the level of supervision during the state number selection.

A comprehensive study by Filion *et al.* used the occupancy of 53 chromatin binding proteins in *Drosophila* Kc167 cells to segment the *Drosophila* genome into 5 major chromatin types (referred to as 'colors') [89]. In this study, the protein occupancy was assayed by DamID, an alternative to ChIP-chip in which the targeted protein is fused to the *E. coli* adenine methyltransferase Dam. Upon TF binding, the Dam protein specifically methylates nearby GATC palindromes; which methylation is eventually detected using microarrays. This is technically feasible as the *Drosophila* genome features little to no endogenous DNA

methylation. Importantly, the authors found that a subset of only five of these proteins (which collectively occupy 97.6% of the genome) can recapitulate the five chromatin states with 85.5% accuracy, thereby underlying the robustness of their approach. Technically, the authors first reduced the complexity of their data (53 dimensions), using principal component analysis (PCA), and found that the 3 principal components explained most of the variance. Projecting the data on the principal components revealed 5 classes. Filion *et al.* used this knowledge in a second phase to fit a five-state HMM onto the first three principal components and thus segmented the genome into 5 'colors'. The 'blue' chromatin represents a repressed state and is marked by H3K27me3 enrichment. The 'green' chromatin corresponds to classic heterochromatin that is prominent in pericentric regions and on chromosome 4 and is largely marked by H3K9me2. The 'black' chromatin corresponds to 48% of the genome and is a new type of silent chromatin; it is marked by the presence of histone H1 and a general absence of other chromatin modifications. Notably, the authors showed that this 'black' chromatin conforms to the LADs mentioned earlier. Active chromatin ('yellow' and 'red') is characterized by H3K4me3, H3K27Ac, H3K79me3 and H3K36me3 and can be readily distinguished from each other by splitting active regions (denoted by the active chromatin marks H3K4me3, H3K27Ac and H3K79me3) into those that have H3K36me3 ('yellow') and those that do not ('red'). Interestingly, comparing 'yellow' and 'red' chromatin, Filion *et al.* describe that (1) the nucleosome-remodeling ATPase Brahma and the Mediator subunit MED31 are exclusively found in 'red' chromatin, (2) that 'red' chromatin is characterized by the presence of H3K79me3 and a lack of H3K36me3 and (3) that 'red' chromatin contains genes with restricted expression domains and that are linked to more specific processes than genes found in the 'yellow' chromatin. Based on these results, the authors suggested that the intergenic 'red' chromatin may contain more regulatory chromatin complexes. In addition, the authors suggested that H3K36me3 is therefore not a universal marker of gene activity, as many genes in the 'red' chromatin (lacking H3K36me3) are active. These five major chromatin types do not directly match with particular genomic features like TSSs, exons or enhancers (as opposed – to some extend – to models with higher state numbers). Filion *et al.* commented that these states, in particular the active ones (red and yellow), could be further subdivided, depending on how fine-grained one wishes the classification to be. Nevertheless, this 5-state classification seems very robust, as extending the set of binding maps with 50 additional chromatin-related proteins does not

change the outcome of the classification[90].

As already mentioned, results gained from different organisms and cell lineages are not readily comparable. Nevertheless, in a recent review[90], Bas van Steensel noted a good agreement between the 9 states of Kharchenko *et al.* (gained in S2 cells) and the five colors of the chromatin (gained in Kc167 cells), where the difference mainly lies in further sub-dividing the active red and yellow chromatin states. On the other hand, the resolution attached to the DamID technology ranges from 2 to 5 kb[91] (i.e. the methylation by tethered Dam spreads over 2-5 kb from a discrete protein-binding sequence) and limits, *de facto*, the subdivision into shorter states spanning only few hundreds of bases.

Globally, these methods enable the integration of vast amounts of data and reducing the combinatory complexity into an interpretable number of states. They also have the advantage to potentially uncover novel genomic elements, decipher new functional associations and annotate functional elements in well-studied or new genomes. On the down-side, the use of statistical models require active selection of the final number of states. The resulting models might therefore represent a trade-off between the statistically optimal number of states representing the data and state selection for reasons of interpretability (i.e. the set of states that best correlates or distinguishes functional and known features). In addition, there is no possibility of ensuring that the resulting model will discern between similarly marked regions (e.g. enhancers versus promoters) or discriminate between the features of interest (e.g. active versus inactive enhancers). In each study, several states could be correlating with enhancers and/or their activity, but these states could easily be overlooked and get lost in the data. So what does resemble an active enhancer? In the next section, I will summarize what is known about histone PTMs found on enhancers – the genomic element central to our work.

1.1.3.4 Chromatin state at enhancers

Early studies clearly suggested that both methylation and acetylation marks are present on enhancers. Bernstein *et al*. reported conserved intergenic enrichment of H3K4me2 in human and mouse[82]. The same year, Roh *et al.* reported that islands of H3K9 and H3K14

acetylation colocalize with known regulatory elements in human T cells[92] and showed in a second publication that some of these islands can function as enhancers when transfected into human Jurkat T cells[93]. Nevertheless, these marks are also associated with active genes and promoters and more studies were required to characterize enhancers, particularly regarding enhancer activity. The most common strategy used to tackle this question was to evaluate what marks are found at active enhancers, which involves two key issues. The first is to define a set of enhancers. To this end, different proxies have been utilized: (1) mapping the binding of the co-factor p300 (or CBP), which has been shown to locate to enhancers *in vivo* in a tissue-specific fashion[17,41]; (2) identification of DNaseI hypersensitive sites (DHSs), which identify DNA regions devoid of nucleosome (i.e. accessible chromatin) that are found at some TSSs (NDRs of active genes) but also on distant regulatory sequences[94,95]; (3) monitoring of the binding of an inducible TF (before and after activation); and (4) presence of H3K4me1 in absence of H3K4me3. Note that authors always considered TSS-distal features to distinguish enhancers from promoters. The second key issue is to be able to discern active from inactive enhancers, as the simple presence of p300 or of a DHS is not indicative of enhancer activity [94,96]. The proxy chosen for enhancer activity was the activity of the closest gene (as assayed by expression profiling).

One of the first studies that evaluated the chromatin state at enhancers in a large scale fashion was conducted by Heintzman *et al.* [65]. In this work, the authors performed ChIP-chip against the core histone H3, several histone modifications (H3K9/14 acetylation, H4K5/8/12/16 acetylation, H3K4me1, H3K4me2, H3K4me3), Pol II, TBP-associated factor 1 (TAF1) and the transcriptional coactivator p300 in human HeLa cells, before and after treatment with INFγ, which induces p300 binding as part of its induced cellular response. Using known TSSs (to locate promoters) and TSS distal p300 binding (to define enhancers), the authors found that active promoters presented strong H3K4me3 enrichment and a bimodal enrichment of H3K4me1 around the nucleosome free region while enhancers were depleted in H3K4me3 and showed a strong mono-modal enrichment of H3K4me1 centered on the p300 binding site. The authors did not find a difference in the H3K4me2 and acetylation enrichments (and profiles). Of note, Birney *et al.*, using the same chromatin data, reported a decrease of H3 acetylation on putative enhancers (that were defined as TSS-distal DHSs, as opposed to p300 binding in Heintzman *et al.*) [97]. Heintzman *et al*. confirmed these conclusions in a second study that used 5 distinct human cell lines and in which they also

showed that (1) H3K27Ac was also frequently associated with enhancers, and (2) the chromatin state at enhancers is cell type specific with a minority of enhancers being shared between cell types[80]. The presence of H3K27Ac on enhancers (defined by p300 binding) was further described to distinguish between active and poised enhancers in human embryonic stem cells where active enhancers are marked by H3K27Ac while poised enhancers are enriched in H3K27me3[98]. This link between enhancer activity and H3K27Ac has also been reported in mouse embryonic stem cells[96]. In this study, the authors defined enhancers as regions of H3K4me1 enrichment combined with an absence of H3K4me3 (as in Heintzman *et al.*) and further verified that the link between H3K27Ac presence and enhancer activity (assessed by proximity to active genes, as already mentioned) was general by profiling H3K4me1, H3K4me3 and H3K27Ac enrichment in proB cells, neural progenitor cells and adult liver.

Other landmark investigations assessing chromatin state(s) on enhancers (and other genomic features) have been conducted using ChIP-seq by Barski *et al* and Wang *et al.* [62,63]. The first study focused on 19 methylation marks and the histone variant H2A.Z in human CD4+ T cells, Barski *et al* observed all three H3K4 methylations (mono-, di-, and trimethylation) and H2A.Z were found at TSS-distal DHSs. Wang *et al.* additionally sequenced and mapped 18 acetylation marks in the same cell line and assessed the mark combinations found at gene promoters and TSS-distal DHSs. The authors found that on both promoters and TSS-distal DHSs, only a tiny fraction of all possible mark combinations were actually observed underlying the non-random association of marks. Concerning enhancer states, H2A.Z, H3K4me1, H3K4me2, H3K4me3, H3K9me1 and H3K18Ac were found at more than 20% of the TSS-distal DHSs and significant presence of H3K36me3 and H4K20me1 were also reported on these putative enhancers. Of note is that H3K4me3 was also recently reported to be present on active enhancers in mouse T cells[79]. Importantly, the authors did not find a significant correlation between gene expression and the modification patterns.

Some of these results are in contrast to studies mentioned earlier[65,80,96,98] in which H3K4me3 was strictly associated with promoters. It was recently proposed that the difference might be due to the use of p300 binding versus DHSs as enhancer predictors[99]. Indeed, p300 is recruited by different sequence-specific DNA binding proteins and is found only at a subset of DHS sites. Promoter distal DHS sites certainly represent a more heterogeneous

population of CRMs, as compared to p300 binding sites, that includes enhancers (repressed and active) but also insulators; thereby explaining the heterogeneity of chromatin patterns found at these locations[63]. Finally, it remains unclear to what extent the activity status of the closest gene is an adequate proxy for enhancer activity.

Enhancers are not only characterized by patterns of histone modifications but also by nucleosome dynamics and presence of unstable histone variants H3.3 and H2A.Z[62,63,100]. He *et al*. compared H3K4me1, H3K4me2 and H3K4me3 profiles in LNCaP prostate cancer cells before, as well as 4h and 16h after stimulation with an androgen receptor (AR) agonist, which results in the activation of AR-responsive enhancers. The authors took advantage of FoxA1 and AR binding maps previously established[16] in LNCaP prostate cancer cells (FoxA1 is a pioneer factor that facilitates binding of activators like AR in prostate cells[16]) to define putative enhancers enriched for H3K4me2 and lacking H3K4me3, as regulated by FoxA1 and/or AR. The authors showed that FoxA1 sites are flanked by a H3K4me2 marked nucleosome at each side, both before and after stimulation. In contrast, at AR sites the H3K4me2 profile switches from a single peak centered on the AR site to a bimodal profile centered on the AR site, suggesting nucleosome displacement upon AR binding. Using quantitative PCR targeting five AR binding sites, the authors could also show that the histone variant H2A.Z is enriched in the central nucleosome as compared to the flanking nucleosomes, suggesting an intrinsic propensity of this nucleosome for displacement. Such nucleosome displacement was also observed upon binding of the E47 isoform of the E2A TF in B cell progenitors, using an H3K4me1 readout[101].

Altogether, enhancers are characterized by the presence of H3K4me1 and H3K4me2, while the presence of H3K4me3 is controversial. Recent studies have recognized H3K27Ac as a signature of *active* enhancers[96,98], while earlier studies relied on H3K4me1 enrichment properties[65,80]. The presence of H3K9me1, H3K18Ac, H3K36me3 and H4K20me1 have been reported in one study[63] and remain to be further validated. The histone variant H2A.Z also seems to be a common feature of enhancers. Lastly, several recent studies have shown that Pol II is present at a subset of enhancers and that non-coding transcription occurs at these enhancers[79,96,98,102-104]. It is feasible that different studies have reached slightly different conclusions with respect to the characteristics of specific histone PTMs due to the use of different cell lines, organisms, PTM sets, or variations in experimental procedures (e.g. use of

different antibodies for a specific PTM exhibiting different cross-reactivity characteristics, immunoprecipitation procedures, sequencing technologies, etc.), in analysis procedures, and yet the fact that there is no unity in how enhancers are defined in the first place.

1.1.4 Overview of *Drosophila melanogaster* mesoderm development

1.1.4.1 Early development of the fertilized egg

The *Drosophila* egg is endowed from the outset (even prior to fertilization) with asymmetry along anterior-posterior (AP) and dorso-ventral (DV) axes due to maternal cues [105]. After fertilization, the zygote's nucleus undergoes eight fast nuclear divisions (8 minutes each) without cellular division. At the end of the eighth division cycle, the 256 nuclei slowly migrate from the center of the egg to its periphery, where nuclei divisions continue until division cycle 13. From cycle 9 on, divisions progressively slow down, taking c.a. 25 minutes at cycle 13. During these 13 nuclear division cycles, or *cleavage*, the embryo is made of a unique cell or 'syncytial blastoderm', containing all the nuclei. At this stage, all divisions are synchronous. Cellularization of the blastoderm occurs at nuclear cleavage cycle 14 (corresponding to developmental stage 5), thus forming the 'cellular blastoderm', in which each somatic nucleus is enclosed within cell membranes (Figure 7A). This occurs by invagination of the oocyte's plasma membrane, progressively enclosing the underlying nuclei to the 'cellular blastoderm', defined by a single layer of about 6000 cells. This stage also marks the maternal-zygotic transition (MZT), characterized by the transcriptional activation of the zygotic genome.

Very early on, the embryo is patterned by maternal cues: Genes known as 'gap genes' are transcribed only in particular compartments along the AP axis, while other genes are expressed in distinctive patterns along the DV axis (Figures 8 and 9). Among these are, for example, *twi* and *snail (sna)*, which are expressed only in the ventral-most cells of the embryo and are pivotal for the formation of the mesoderm (giving rise to the various muscle systems and the fat body) at the ventral side of the embryo. Initial gastrulation starts after cellularization by invagination (folding inwards) of the mesoderm at the ventral midline along the AP axis (stage 6, see Figure 7A), and by extension of the posterior pole anteriorly across the dorsal surface ('germband extension', complete by stage 8). Proper development requires tightly regulated and coordinated spatio-temporal control of gene expression from

the very beginning. The question of how are these very specific patterns of gene expression achieved is central to contemporary developmental biology.

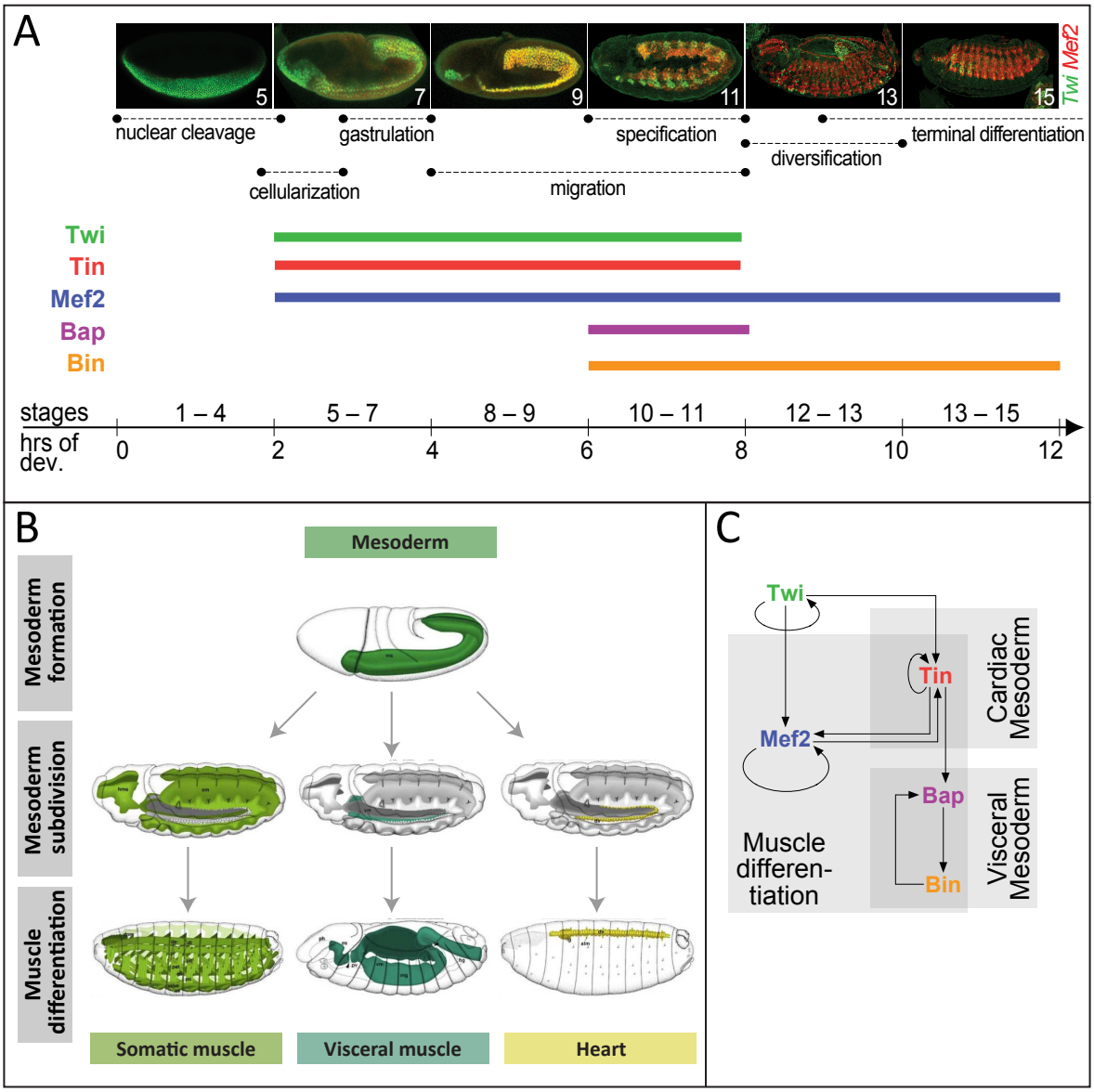


Figure 7. Major events in *Drosophila* early development and mesoderm specification.

(A) Top, major events in mesoderm specification and early embryo development (indicated by dashed lines). *In situ* RNA hybridization against twi in early development (far left) and immuno-staining against Twi (green) and Mef2 (red) later (other pictures) illustrate mesoderm development during the relevant developmental stages. Middle, ranges of expression for five central TFs in mesoderm specification. Bottom, developmental stages and corresponding developmental times (in hrs AEL). **(B)** Overview of the three major muscle types in the *Drosophila* embryo and of their formation. Embryo images are from[106]. **(C)** Myogenic network of five key TFs in mesoderm specification indicating their regulatory connections as determined by genetic interaction studies.

1.1.4.2 Patterning of the *Drosophila* blastoderm

Up to the MZT, maternally provided mRNAs and proteins govern all processes; in particular gradients of the TFs Bicoid and Hunchback are at the basis of the AP patterning, while nuclear gradient of Dl subdivides the DV axis.

Bicoid mRNA is deposited and anchored in the anterior pole of the embryo during oogenesis. Upon fertilization, mRNA translation is activated and the newly synthesized Bicoid protein (but not the mRNA) diffuses from this production source within the embryo, thereby establishing an anterior-posterior gradient of Bicoid protein concentration. Wherever the local concentration of Bicoid is above a certain threshold, early targets such as the *hunchback* gene can be activated (*hunchback* mRNA is also maternally deposited in the oocyte). Gradients of Bicoid and Hunchback along the anterior-posterior axis activate the gap genes *kruppel*, *knirps* and *giant,* whose products in turn help to delineate the expression of the pair-rule genes, e.g. *even-skipped* (*eve*), which are expressed in 7 stripes along the AP axis.

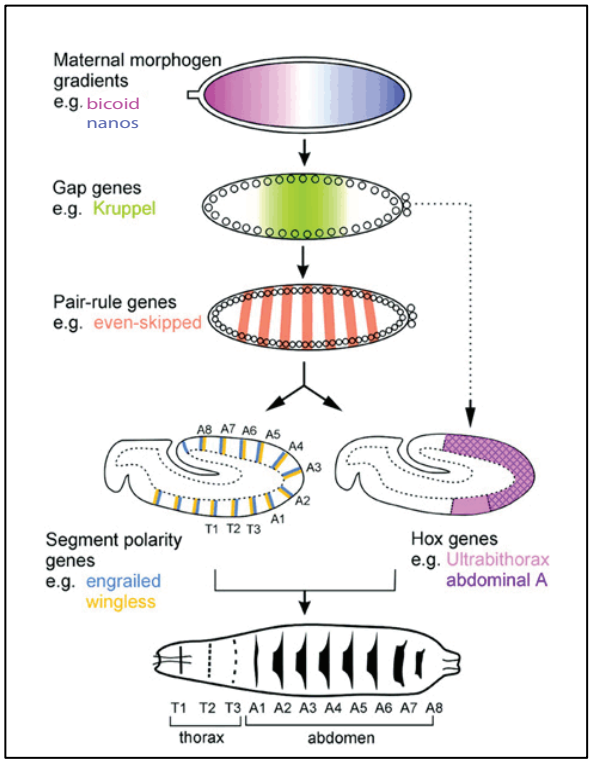


Figure 8. Patterning along the AP axis of the *Drosophila* embryo.

A cascade of maternal (nanos, bicoid) and zygotic genes is activated in the syncitial embryo to subdivide the ectoderm into smaller domains. The embryo cellularizes and undergoes gastrulation after activation of the pair-rule genes. The segment polarity genes and the *Hox* genes are activated by the pair-rule genes but a subset of gap genes also influences directly the *Hox* genes. Both segment polarity and Hox genes are thought to act in concert to control the differentiation of each segment of the future larvae. Reprinted by permission from Macmillan Publishers Ltd: EMBO Reports[107], copyright (2001).

A nuclear concentration gradient of the TF Dl is established along the DV axis by the time DV patterning genes are activated (stage 5, Figure 9). This is achieved by maternal cues that activate the Toll receptor only on the ventral side of the egg. Toll activation initiates a proteolytic cascade that ultimately leads to the regulated degradation of Cactus. Although Dl is maternally loaded and uniformly distributed throughout the egg, it remains inactive when forming a complex with Cactus (as Cactus prevents its translocation to the nucleus). Thus, ventral degradation of Cactus allows Dl to enter nuclei in a ventral-to-dorsal gradient, with Dl levels being highest in ventral regions, progressively lower in ventro-lateral and lateral regions, and absent from dorsal nuclei (Figure 9). Once in the nucleus, Dl can bind DNA and activate its target genes, in a concentration-dependent manner. While high Dl concentrations are required to activate genes such as *twi* and *sna* in ventral regions, lower levels in lateral

regions are sufficient to turn on genes such as *vnd, rho* and *sog* (Figure 9). Dl also contributes to the repression of various target genes, which delimits the expression of, for example, *zen* to the most dorsal regions. Positive regulations between TFs, including auto-activation and positive feed-forward motifs (e.g., the Dl target *twi* activates itself, as well as the Dl target *sna*), along with negative regulations (e.g., *sna* represses genes such as *vnd, rho*, and *sog*, thereby excluding them from ventral regions and limiting their expression to lateral domains) lead to characteristic expression domains defining the principle early *Drosophila* germ layers: (1) the mesoderm is established in the ventral-most domain in the presence of *twi* and *sna,* and will give rise to various muscle systems and the fat body; (2) expression of genes such as *vnd* and *sog* in more lateral regions define the neurogenic ectoderm, which gives rise to the peripheral and central nervous systems; and (3) the dorsal-most regions, which express genes like *zen* and *dpp*, form the dorsal ectoderm, which is the source of extra-embryonic tissues (the endoderm forms slightly later by invagination from the anterior and posterior parts of the gastrulated embryo).

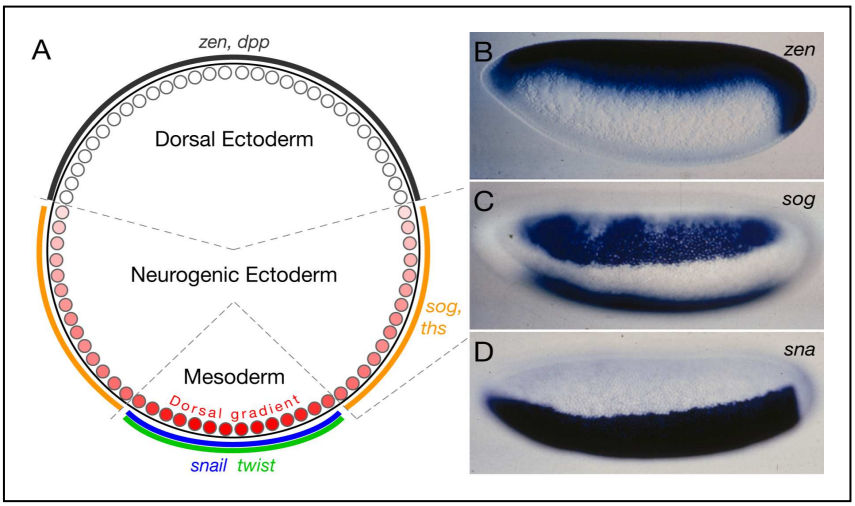


Figure 9. Dorsal establishes three primary tissue types in the embryo

(A) A schematic cross-section through the trunk of a nuclear cleavage cycle 14 embryo, ventral down, dorsal up. The nuclear concentration gradient of the TF Dl (A, red) sets up the three primary tissue types in the early *Drosophila* embryo. Highest levels of nuclear Dl lead to transcription of *twi* (green) and *sna* (blue) in the mesoderm. Lower levels in lateral regions establish the neurogenic ectoderm and allow for the transcription of genes such as *sog* and *ths* (orange), as well as for the transcription of neurogenic genes such as *vnd* in a ventral subset of the neurogenic ectoderm. Dl acts on genes such as *zen* and *dpp* as a repressor and thus confines their expression to the dorsal ectoderm, where Dl is not present in the nuclei. *In situ* hybridizations show the Dl threshold responses of *zen* (B), *sna* (C), and *sog* (D). Embryos are shown in lateral (B) or ventro-lateral (C, D) views with anterior to the left and dorsal up. Figure courtesy of Robert Zinzen.

1.1.4.3 Specification of the mesoderm

Between embryonic stages 5 and 15 of *Drosophila* embryonic development, the mesoderm is specified into several primordia (Figure 7B), including the three largest for cardiac mesoderm (heart muscle), the somatic mesoderm (analogous to vertebrate skeletal muscle) and the visceral mesoderm (gut muscle). The early *Drosophila* mesoderm (stage 5) is composed of a field of pluripotent cells[108,109]. After invagination of the mesoderm (stage 6), these pluripotent cells (now located inside the embryo) dissociate from each other, proliferate, and migrate dorsally along the overlying ectoderm, which then also acts as a signaling source for patterning of the underlying mesoderm. The specification of the mesoderm into the different tissue primordia requires that these pluripotent cells express the appropriate TFs and signaling proteins. This multilevel information converges on CRMs to elicit specific developmental programs. Genetic studies revealed that mesoderm specification requires the successive activation of key TFs[110,111](Figure 7A,C), such as *twi*, *tin*, *myocyte enhancing factor 2* (*mef2)*, *bin* and *bagpipe* (*bap)*.

At stage 5, in the ventral part of the blastoderm, high concentration of the maternally provided Dl activates *twi*, a basic helix–loop–helix TF. Twi then cooperates with its activator Dl to pattern the dorsoventral axis, as well as with its target Snail to drive the process of mesoderm gastrulation (~stage 6, Figure 9). Up to stage 11, Twi acts as a master regulator that is both essential and sufficient to initiate mesoderm development[112]. In particular, Twi directly regulates the expression of both Tin and Myocyte enhancing factor 2 (Mef2). Tin is co-expressed with Twi (stage 5 to 11) and is essential for the specification of the dorsal mesoderm into the heart, the visceral muscle and the dorsal somatic muscle[113,114]. Mef2 expression spans a wider range (stage 5 to 15) and initiates muscle differentiation. To better understand how Twi can regulate such a broad variety of processes, we used ChIP-chip analysis to map its genome-wide binding landscape at two time points (stages 5-7 and stages 8-9, see Figure 7A) of the early mesodermal development[15]. This study showed that Twi binds to thousands of CRMs and potentially directly regulates ~500 genes involved in cell proliferation, morphogenesis and cell migration. Strikingly, Twi directly targets about 25% of all annotated TFs, which might represent the complete subset of TFs regulating mesodermal early development.

Tin expression is restricted to the dorsal mesoderm by pMad, the effector of the Dpp signaling (*dpp* is a morphogen which concentration decreases along the DV axis). At the same time, the pair-rule genes *eve* and *slp* and the segment polarity genes *hedgehog* (*hh*) and *wingless* (*wg*) further subdivide the mesoderm along the AP axis[115]. In the *tin* expressing dorsal mesoderm, pluripotent cells that receive both ectodermaly derived Dpp and Wg signals (which effector proteins are pMad and dTCF, respectively) are specified to become the cardiac mesoderm (CM, Figure 10A). In particular, Tin acts together with Pannier (Pnr, a GATA factor) and Dorsocross (Doc, a T-box factor) to specify CM cell fate[116], whereas the visceral mesoderm (VM) fate is actively repressed in these cells by Slp, a repressor activated by Wg signaling. Neighboring cells that only receive Dpp signal specify into VM. In these cells, Bap is activated by Tin and its expression is restricted to stage 10-11. Tin and Bap activate Bin (stage 10), which remains expressed in the VM until stage 15 (Figure 7). Bin targets a large number of mesodermal genes and is a key TF of the VM specification [117]. The ventral region of the hemi-segment (Figure 10A) will become fat body (FB, in the Wg negative part) and somatic muscle (SM, in the Wg positive part). In the FB, Notch signaling actively represses Twi[118]; while high levels of Twi are essential for somatic mesoderm specification[112].

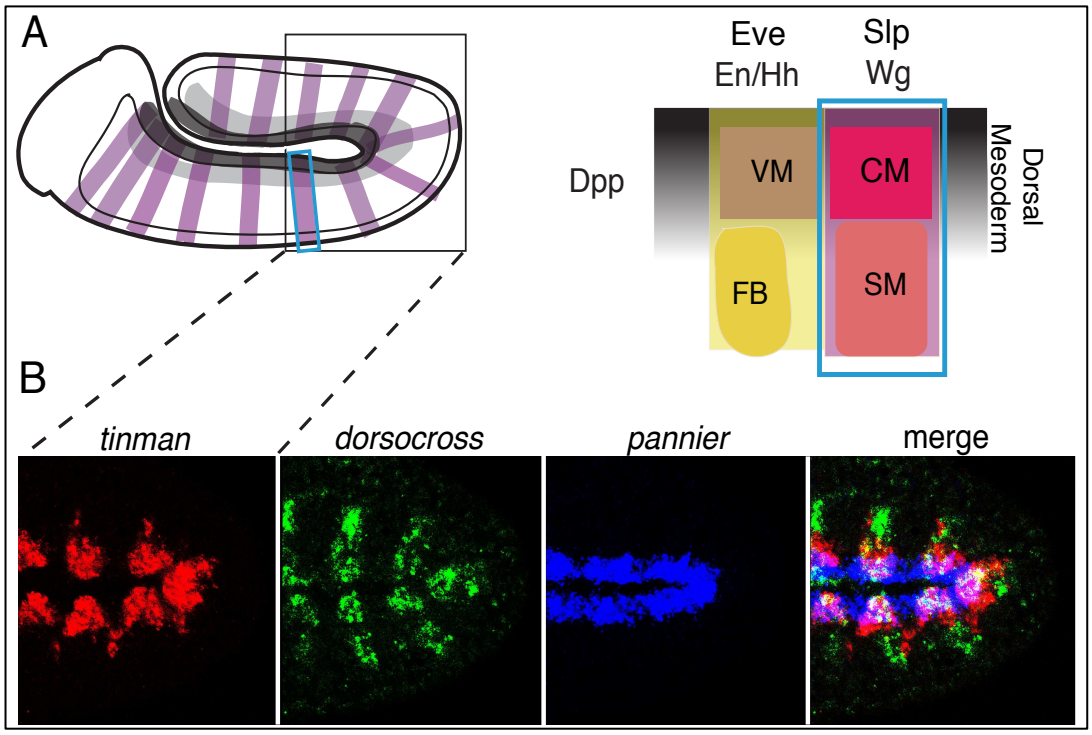


Figure 10. Dorsal mesoderm specification into cardiac and visceral mesoderm during *Drosophila* embryogenesis.

(A) Diagram of a *Drosophila* embryo showing wg expression in 14 parasegments. Area indicated by blue rectangle is enlarged in the right panel, showing a schematic representation of mesoderm subdivision in one hemisegment. The dorsal domain, which has high levels of Dpp signaling (black), gives rise to VM and CM, whereas ventral regions become FB and SM. CM is specified at the intersection of Wg (purple) and Dpp signaling in the posterior part of each parasegment. Wg activates *slp* expression, and together they promote CM and repress VM specification. (B) Triple-fluorescent in situ hybridization showing tinman, dorsocross, and pannier expression in the dorsal mesoderm during early stage 11, when cardiac specification takes place. All three genes are coexpressed exclusively in the cardiogenic mesoderm (pink-white area of coexpression). The region of the embryo shown is depicted by the black square in (A).

1.2 Prediction of CRM location and activity status

The expression of developmental genes changes during development and reflects commitment into particular cell fates or response to particular cellular events. These complex gene expression patterns are governed by CRMs, which translate TF binding and chromatin information into gene expression. Altogether, TFs, CRMs and the targeted genes form a gene regulatory network (GRN) that defines and explains the state of the cell, with CRMs being the bridges between regulators and *de facto* gene regulation. The characterization of these GRNs is fundamental for the understanding of gene regulation underlying metazoan development. This requires (1) the identification of the repertoire of CRMs present in genomes; and (2) the determination of when and where an enhancer is active. The next sections review the computational and experimental strategies used to find the location of CRMs and predict their activity.

1.2.1 *in silico* prediction of CRMs

The exponential accumulation of sequenced genomes since the release of the first draft of the human genome in 2000 has stimulated the development of computational methods to annotate the various features of the genomes. In particular, the lack of high throughput experimental methods to identify CRMs has pushed investigators to develop numerous *in silico* strategies to locate CRMs genome wide. Reviewing existing computational methods and tools addressing this task is beyond the scope of this thesis and I kindly point the reader to recent papers reviewing this extremely prolific field[119-121]. In the following sections, I give an overview of these different strategies without getting into the implementation and statistical details of individual algorithms; rather, I extract major principles, advantages and limitations of high-level strategies.

1.2.1.1 Predicting TFBSs and the futility theorem

CRMs are composed of TFBSs and predicting CRMs thus naively boils down to predicting TFBSs. As shown in Figure 2, the binding specificity of a TF can be represented using a PWM that, in turn, can be used to scan the DNA sequence to find and score sequences conforming to the PWM model, P_m, in contrast with a background model, P_b. A typical approach is to consider the log likelihood ratio of these two probabilities and keep sub-sequences yielding positive values i.e. $\log(P_m/ P_b) > x$ where $x > 0$. Technically, a PWM model gives the probability to find each base of {A,C,T,G} at the different positions of the binding site. The overall probability of a particular word to originate from the PWM model, P_m, is therefore the product of the individual probabilities of having base b at position i of the model. Different background models can be used to compute P_b, a simple one being to consider that the probability of finding base b (b in {A,C,T,G}) at position i equals the global frequency of b in the genome. This simple model corresponds to a Markov model of order 0, meaning that the probability of base b is independent from the preceding base(s), while a Markov model of order m implies that the probability of base b depends on the m preceding bases.

Several motif scanners have been developed, such as Patser[122], with the most recent ones, such as matrix-scan[123], being able to accommodate higher order Markov models. In practice, the number of sites predicted by such tools is huge (1 site every 500-5000 bp using common settings), with the vast majority of these predictions being non functional *in vivo* and therefore considered as false positives. Unfortunately, this situation cannot be solved by considering a higher threshold[119]. Wasserman and Sandelin termed this phenomenon the 'futility theorem', as virtually every gene harbors a binding site for any TF in its immediate proximity. As a result, single site detection using motif scanners cannot be considered as a viable approach to predict CRMs, especially in metazoans, and additional considerations must be used to better reflect the biology, such as sequence conservation, presence of additional sites (TFBSs clusters), presence of specific TFBS arrangements, or a combination of thereof.

1.2.1.2 Using sequence conservation to locate CRMs

Sequence conservation has been successfully used in many bioinformatics applications and reflects the assumption that mutations should accumulate more slowly in functional elements than in regions without sequence-specific functions. Sequence conservation can be considered at the TFBS level and at the CRM level. As mentioned before (section 1.1.2.2), conservation is far from systematic at the CRM level and is much more frequently observed at the TFBS level (reviewed in [119,121]). For example, we showed[36] that TFBSs for a particular TF in regions bound *in vivo* by the corresponding TF (as assessed by ChIP-chip) are much more conserved than the same TFBSs predicted in regions bound by other TFs (Figure 11). Using conservation to enrich TFBSs in functional prediction is therefore a valid and commonly used approach, and prediction of conserved PWM instances is implemented in a number a prediction tools[120,121]. Nevertheless, such approaches are *de facto* ignoring TFBSs that are species specific, or are weakly matching their PWM model (inherent to the use of stringent thresholds to predict TFBSs in the first place), or fall within an un- or misaligned sequence region due to technical limitations (duplications, repeats or low complexity sequence), alignment mistakes or even incomplete sequencing, or yet cases where exact TFBS position has moved over the course of evolution within the CRM[121].

An alternative way to locate CRMs using sequence conservation is to identify conserved blocks in the non-coding genome. This idea has been pushed to its paroxysm with the detection of ultraconserved elements[124], which are defined as a perfect sequence identity of at least 200 bp between very distant organisms[125]. Visel *et al.* tested the exact potential and uniqueness of these ultraconserved elements and compared them to the less evolutionary constrained 'extremely conserved' elements, which are defined as sequences with conservation properties similar to ultraconserved ones but lacking perfect extended identity. Strikingly, 50% of 'ultraconserved' elements as well as 50% of 'extremely conserved' elements have been shown to drive expression in transgenic animals during embryonic development[40], a rate identical to that obtained 2 years before by Pennacchio *et al.,* who tested 167 of these human-mouse-rat extremely conserved sequences in transgenic mouse enhancer assays[126] (note that the remaining elements might be functional at stages of development or under conditions not assayed). These results clearly demonstrate that high conservation of non-coding sequences points to functional cis-regulatory elements and

different algorithms have been developed to identify conserved blocks in multiple alignments (with much looser sequence conservation criteria) and to predict conserved TFBSs[120,121]. Unfortunately, identifying conserved TFBSs (which often results from arbitrary thresholds) is usually not sufficient to reliably identify functional sites and encompassing CRMs[121,127]. Identification of CRMs using overall sequence conservation, which is especially tricky in compact genomes like that of *Drosophila* and other invertebrates, still yields high false positive rates[3,121,127].

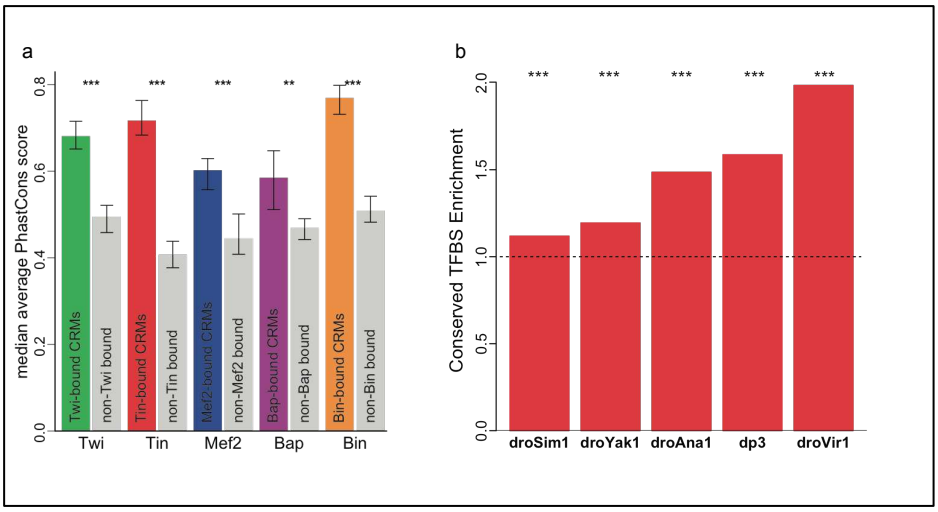


Figure 11. Conservation of TFBSs.

(a) TFBSs for Twi, Tin, Mef2, Bin and Bap were predicted using Patser in regions bound or unbound by the corresponding factor (unbound regions still had to be bound by at least one of the other four TFs). The average of the PhastCons[128] score over the bases of the TFBS was computed for the best scoring TFBS found in each bound and unbound regions. The histogram presents the median of average PhastCons[128] scores for motifs in bound (coloured bars) and non-bound regions for that TF (grey bars). Error bars represent equi-tailed 95% confidence intervals of the median. **P < 0.001; ***P< 10^{-6} (one-sided Wilcoxon's rank-sum test). **(b)** Enrichment of conserved Tin TFBSs in bound CRMs compared to random intergenic regions for 5 *Drosophila* species found at increasing phylogenetic distances from *D. melanogaster*: *D. simulans* (droSim1), *D. yakuba* (droYak1), *D. ananassae* (droAna1), *D. pseudoobscura* (dp3), and *D. virilis* (droVir1). Tin TFBSs predicted in *D. melanogaster* were used to extract the corresponding sequence from each pair-wise alignment (ungapped alignments only, alignments downloaded from UCSC). A TFBS prediction was scored as 'conserved' in a particular species if its aligned sequence triggered a match scoring above used cutoff, or was otherwise scored as 'not conserved' (unaligned TFBSs were also counted as 'not conserved'). Using the best TFBSs (found in each bound and random regions) shows significant increase in the proportion of conserved TFBSs in CRMs compared to background sequences; *=p<0.05, **=p<10^{-3}, ***=p<10^{-10} (one-tailed Exact Fisher test). Of note, repeating the analysis presented in (a) and (b) with all predicted TFBSs yielded similar results (with a reduced significance though). More details as well as results for Twi, Mef2, Bin and Bap (which are similar to the results obtained with Tin) can be found in the original publication[36].

1.2.1.3 High density of TFBSs improves CRM predictions

CRM organization, such as TFBS density (usually referred to as TFBS clustering) is another feature exploited in CRM prediction methods. Indeed, this feature of CRMs has been recognized in early studies[18,129] and various examples of CRMs harboring multiple TFBSs for several distinct TFs have been reported in fly, mouse and human[121]. For example, the *Drosophila* eve muscle and heart enhancer (MHE) contains 6 pMad, 4 Ets, 4 Tin, 2 Twi, and 1 dTCF binding sites in a stretch of only 312 bp, while the human β-globin locus control region (that contains binding sites for GATA1, EKLF, NF-E2, SOX6, BCL11A), and the IFN-β enhanceosome (model shown in Figure 4) contains 8 TFBSs for 6 factors in only 55 bp. It is therefore not surprising that the detection of clusters of heterotypic (TFBSs of several TFs) or homotypic (TFBSs of a single TF) sites is at the basis of numerous algorithms.

While most known CRMs fall in the heterotypic category, strong evidence suggests that homotypic clusters play functional roles in both vertebrates[130] and *Drosophila*[131]. Indeed, CRMs have been identified using homotypic clusters of Dl in *Drosophila* by simply searching for 3 or more Dl sites within a 400 bp window[132]. Practically, methods vary from simple sliding window approaches combined with user-defined criteria (TFBS number and diversity) to more sophisticated probabilistic models like HMMs (see e.g. Ahab[133], Cluster-Buster[134] or more recently SWAN[135]). Using the MHE enhancer mentioned above as a model, Halfon *et al.* enumerated all 500 bp windows harboring a similar TFBS composition (at least 1 dTCF site and 2 sites each for pMad, Ets, Tin, and Twi) and showed that one of the 33 predicted elements had the expected spatio-temporal expression pattern in transgenic animals[136]. Other studies adopted similar strategies and could validate a number of their predictions[132,137].

A common aspect of these studies that certainly explains part of their success is a promising starting point: availability of known CRMs that can serve as guides and models. This, however, prevents the application of such strategies to *cis*-regulatory problems where no clear combination of TFBSs is known. More sophisticated algorithms, like Ahab[133], especially address this question by identifying sub-sequences most likely to originate from a 'motif cluster model'; this abrogates the need of specifying thresholds on PWM predictions, for example. Still, the investigator is expected to operate a pre-selection of PWMs likely to cluster together, i.e. reflecting a particular biological system. Using PWMs of 9 maternal and

gap factors, Schroeder *et al.* ran Ahab on 0.75 Mb of sequence located around 29 genes selected for their gap and pair-rule expression patterns during gastrulation in *Drosophila* [138]. Remarkably, 13 of the 16 CRM predictions showed AP differential expression in transgenic flies. Although successful, these approaches are neither fully agnostic nor genome-wide in the sense that they require to select adequate PWMs and search limited spaces around pre-selected sets of genes. In other words, they cannot address the more general challenge of predicting *all* potential CRMs in a genome.

Approaches combining both conservation and TFBS clustering to produce an unbiased and genome-wide set of CRM predictions using large PWM sets have been developed by several groups. In particular, the PReMod database[139] centralizes genome-wide mammalian CRM predictions computed using the method developed by Blanchette *et al.* [140]. Nevertheless, these approaches still yield low specificity and therefore need to be combined with additional information[121].

1.2.1.4 Machine learning approaches

When a set of experimentally characterized CRMs is available, the dissection of the regulatory inputs allowed the investigators to select the features (what TFs should be present, site number and density, window size, TFBS organization) that characterized the CRMs the best. Using this set of features, possibly supplemented with sequence conservation filtering, authors often perform a space-oriented search to identify similar CRMs. This is typically what a supervised machine learning approach does but in a more systematic and probabilistic way. Provided a positive and a negative set of individuals (here CRMs) and features (or characteristics, e.g. TFBS presence, or TF binding), a machine learning approach will learn what features best discriminate the positive and negative individuals and offers a framework to estimate the performance of the trained classifier. The trained model is then used to predict new positive CRMs. Amongst the most popular supervised machine learning methods used in computational biology are artificial neural networks, generalized linear models (logistic regression in particular), support vector machines, Bayesian networks, decision trees, random forests, and Markov models like HMMs[141]. The success of machine learning approaches is conditioned by the availability of training sets (positive and negative individuals), by the

degree of similarity (or homogeneity) of these individuals (common traits should be shared by most CRMs), and the existence of discriminative features.

A number of studies have successfully applied machine learning approaches, in particular the pioneering study by Wasserman and Fickett[142]. Using 29 CRMs driving expression in human skeletal muscle and PWMs for 5 TFs acting in muscle development (Mef-2, Myf, Sp-1, SRF, Tef), the authors employed logistic regression to train a model able to predict skeletal muscle enhancers. The negative set mainly contained random sequences sampled from the primate genome and from a promoter database. Applying their model to the human genomic sequences available at this time (~ 2 Mb in total), authors could identify 91 regions (using a cut-off corresponding to a sensitivity of 66%) and evaluated that at least 50% of these were located in the immediate vicinity of genes with consistent tissue expression. Wasserman and colleagues used the same approach two years later to identify CRMs driving specific liver expression[143], using a different positive training set (16 CRMs) and a different collection of PWMs (HNF-1, HNF-3, HNF-4, and C/EBP). This time, the authors predicted CRMs in the complete human genome and used phylogenetic footprinting to post-filter their predictions, leading to the identification of 147 potential liver modules. Interestingly, of the 12 training set CRMs correctly identified by the model, only 4 survived the phylogenetic footprinting filter. This result again underlines that sequence conservation is not a general feature of functional enhancers.

In both of these studies, the selection of the initial PWMs was driven by prior knowledge of the TFs active in the tissue of interest and, more importantly, their binding affinities (PWMs) could be built based on available footprints. Alternatively, starting with PWM collections (available in JASPAR[144], UniPROBE[33], FlyFactorSurvey[145] or the commercial TRANSFAC® database), motifs (k-mers or PWMs) overrepresented in the positive training set (as identified by *de novo* motif discovery), or other features like sequence composition (encoded in Markov chains), one can select the features that discriminate the positive from the negative set the best. For example, Narlikar *et al.* used the LASSO linear regression method to select 45 features from an initial set of 727 features[146]. This initial set of features was composed of (1) PWM match density using both existing PWMs and PWMs discovered in the positive set (that contained 50 heart enhancers), and (2) Markov models of orders 0–5 learned on both positive and negative sets (the feature used being the likelihood ratios). Technically, the LASSO method models the class (+1 and -1 for

the positive and negative sets, respectively) of each sequence as a linear combination of features, and learns the optimal weights associated with each feature. Features with no discriminative power are eliminated (i.e. their weight is 0). Finally, the authors predicted 42,000 putative human heart enhancers genome-wide (note that predictions were only performed in human-mouse conserved non-coding sequence) and validated 16 of 26 predictions *in vivo* (they also tested 20 negative predictions of which only 2 drove heart expression).

Careful evaluation of the contribution of each selected feature showed that the Markov model based features increased the overall classifier accuracy by 7%, suggesting that sequence features other than PWMs must be considered. Indeed, it is not always known what TFs are relevant to the specific regulatory network of interest; in addition, the binding affinities of known ones might not be available and successful *de novo* motif discovery on the positive CRM set is not guaranteed to yield results, especially when the available training set is small. To address these limitations, Kantorovitz *et al.* have applied supervised learning in a 'motif-blind' way. The authors defined 8 scores based on different sequence composition features: Markov chains (exactly as in Narlikar *et al.*, see above), dot products and sets of k-mers overrepresented in the positive sets. Each of these 8 metrics was evaluated independently using 31 enhancer sets (catalogued in the REDfly[25] database), each set representing a different regulatory subnetwork in *D. melanogaster*. Using extensive cross-validation, they found that 15 of these 31 data sets were amenable to supervised learning (and therefore to CRM prediction) and could correlate prediction accuracy with (1) the extent of homotypic clustering (of k-mers) in the training set, (2) the GC content of the training set, and (3) the extent of nucleotide-level conservation with orthologous sequence. In addition, the authors showed that their 'motif-blind' approach outperformed a 'motif-aware' approach, and that integrating orthologous information further improved accuracy. Genome wide predictions in the fly and the mouse (the learning/prediction pipeline was also applied to 8 sets of tissue-specific mouse enhancers[147]) were performed using a 'fusion' score that combined 3 of the 8 metrics evaluated. Finally, the authors validated *in vivo* 5/5 predictions in the fly and 2/2 in the mouse. Given the different criteria and post-filters used to select the predictions for *in vivo* testing, this astonishing success rate (100%) should be regarded with caution. For example, all tested fly predictions originated from the sub-network with the

highest accuracy (the blastoderm) and were located in the vicinity of genes with likely expression profiles.

Altogether these machine-learning approaches proved to be extremely powerful and represent a natural choice when a training set is available, which also restricts their use to coherent subnetworks.

Besides motif-blind machine-learning approaches, *in silico* methods heavily rely on the availability of, at least, the PWM model for your TF of interest (e.g. localization of homotypic clusters); although the use of multiple and functionally related PWMs generally performed much better (e.g. localization of heterotypic clusters). Alternatively, a set of CRMs driving similar expression patterns can be used as a training set to learn key PWMs. However, a number of limitations are associated with PWM-based *in silico* methods described in the previous sections. First, they require a prior knowledge of the different TFs acting in the regulatory network of interest (as using a unique PWM would likely fail, a consequence of the futility theorem). Second, TF PWMs are often missing and, when available, they might be of poor quality. Indeed, until recently, PWMs were constructed (Figure 2) from few experimentally determined footprints typically generated from *in vitro* experiments using purified protein and naked DNA, possibly supplemented with orthologous sequences to increase information content. Although the number of available binding models has significantly increased with the development of novel experimental methods aiming at determining TF binding specificities (bacterial-1-hybrid[148], protein-binding microarrays[149], SELEX[150], MITOMI[151]), or at locating TF binding *in vivo* by ChIP-chip or ChIP-seq (coupled with the development of adapted motif finders like MEME-ChIP[152], DREME[153] or Peak-Motifs[154]), PWMs are available for only a fraction of all existing TFs (a situation that might rapidly change). As a result, it is not always possible to assemble a coherent set of PWMs to predict CRMs using an *in silico* approach. Moreover, it has been shown that TF affinity can vary with co-factors[13] suggesting that TF binding specificity might be better represented by more than one PWM. Finally, TF "binding" *in vivo* does not necessarily implies the presence of the relevant TFBSs, as TFs can be part of larger protein complexes. For all these reasons, *in vivo* approaches probing TF occupancy along the genome will always be superior to computational methods that predict TF occupancy.

1.2.2 Predicting CRMs from experimental data

ChIP coupled with microarray or, more recently, high-throughput sequencing (see Figure 6) has quickly become the method of choice to study protein-DNA interactions, in particular TF binding, co-factor localization and histone modifications. For example, a simple search in PubMed for articles published after 2000 which abstract contains 'ChIP-(on-)chip' or 'ChIP-seq' yielded more than 1300 results at the time of writing, demonstrating the wide impact of high-throughput ChIP approaches.

1.2.2.1 ChIP against transcription factors

Chromatin Immuno-Precipitation (ChIP) relies on an antibody that specifically recognizes the targeted TF (or chromatin mark) in order to immuno-precipitate bound DNA fragments. Importantly, the targeted TF might be expressed in multiple tissues or even ubiquitously. In such a situation, the use of whole embryos can be problematic as it yields mixed signals from a non-uniform pool of cells. Consequently, investigations have usually been conducted in cell lines[88,97,155], with dissected organs[17,156], in whole embryo with tissue-specific factors[15,36,117,157,158], or at very early developmental stages when the embryo is still composed of a homogenous population of cells[159].

Early ChIP studies demonstrated the ability of ChIP approaches to identify regulatory regions in a genome-wide and unbiased manner. An important and fundamental question concerns the specificity of ChIP: Are all identified binding locations, usually named *peaks*, biologically functional? After quality assessment and validation of the assay, the first step of the data processing workflow is to extract the signal from the noise, a task usually performed using a 'peak finder' (e.g. TileMap[160] for tiling arrays, or MACS[161] for high-throughput sequencing). The methods converting raw signal into peaks vary substantially between platforms (i.e. microarray versus sequencing, but also between different sequencing or microarray technologies), but commonly associate a confidence value to each potential peak, as well as a false discovery rate (FDR) tied to a particular threshold. Of note, the FDR

computed by peak finders is empirical in most cases (determined by finding peaks in the control data using the real ChIP sample as control) and its caveats should be taken into account. The nature of the sample (Figure 6) used as control is also an important aspect, as controls based on genomic DNA, ChIP with IgG, or yet the pre-immune serum, will not identify the same potential artifacts. For example, peaks identified by both the serum and its corresponding pre-immune serum are not *bona fide* binding locations of the TF under study, but result from the presence of another antibody and are therefore different from technical noise. When plain genomic DNA fragments or fragments resulting from ChIP with an IgG are used as control, these peaks would not be filtered out and would thus affect the FDR estimates. An efficient way to minimize false positives is to use different antibodies (e.g., targeting non-overlapping portions of the TF) for biological replications. Aware of these potential pitfalls, Li *et al.* evaluated by ChIP-chip the binding landscape of 6 maternal and gap genes in the *Drosophila* blastoderm using two different antibodies against each TF (rabbit serum) [162]. In addition, both genomic DNA and ChIP with IgG were used as controls, and the FDR was estimated with two separate methods. The authors then considered two different levels of confidence, 1% and 25% FDR, which resulted in the identification of thousands of peaks per TF. The authors further confirmed by quantitative PCR that regions selected from the bottom half of the 25% FDR list were indeed bound (11 out of 16 tested regions). The analysis of these different sets showed that highly bound regions (found in the 1% FDR set) were enriched in the proximity of genes transcribed in the blastoderm, contained most of the known CRMs targeted by these TFs, were largely located within intergenic regions and intronic sequences (as expected for CRMs), and showed higher conservation than other non-coding sequences. Conversely, in regions with lower confidence, all these associations dissipated, suggesting that the poorly bound regions (1-25% FDR set) were not functional. Consequently, very stringent cut-offs must be used to identify functional binding, while the exact role of lowly bound regions remains unclear.

Another aspect is that TFs tend to bind to CRMs in a combinatorial and dynamic manner, a property that can be exploited to improve CRM prediction. First, TF bound regions can be used to identify the 'collaborative tendencies' of TFs [15,157]. For example, we profiled the genome-wide binding landscape of Twi, a mesoderm specific TF essential for early mesoderm development in *Drosophila*, at two early developmental time points (2-4h and 4-6h AEL) [15]. Consequently, we found that Twi binds to ~2,000 TFBSs, of which 51% are

continuously bound, while 23% and 26% are specifically bound at 2-4h and 4-6h AEL, respectively, indicating that Twi binds to CRMs in a dynamic manner. Using motif enrichment analysis, we found that Dl sites were only enriched in the proximity of most early bound TFBSs, while Tin sites were only enriched in the proximity of later bound TFBSs, presumably reflecting the collaboration of Twi with these two different factors in DV patterning and mesoderm maturation in a temporally dependent manner. Importantly, we confirmed 7/7 and 11/11 predictions (for Dl and Tin, respectively) by ChIP and quantitative PCR.

Such combinatorial binding can be used to decipher high-order cis-regulation codes. We recently generated genome-wide binding maps for 5 key mesodermal TFs (Twi, Mef2, Tin, Bin and Bap) at 5 consecutive time points in the *Drosophila* developing embryo (2-4h, 4-6h, 6-8h, 8-10h and 10-12h) using ChIP-chip[36]. We found that these TFs bind near each other at specific developmental stages, indicating that these TFs co-occupy CRMs. Notably, this enrichment in TF binding proximity (~100 bp) is not observed for TFs functioning at different developmental stages, (e.g. Twi 4-6 hours and Mef2 10-12 hours). We exploited this property to delineate 8,008 putative CRMs, of which more than 46% involve more than a single binding event. Using experimentally validated CRMs of known expression and machine learning, we finally demonstrated that the spatio-temporal activity of these putative CRMs can be predicted solely on the basis of their binding profile (i.e. the combination of TFs and times at which the CRMs is bound), and we could validate more than 71% of our predictions by *in vivo* transgenic reporter assays. Importantly, 35 of out of 36 (97%) putative CRMs tested during this study were sufficient to function as discrete regulatory modules *in vivo*, demonstrating the power of such combinatorial approach. The propensity of TFs (that are not necessarily functionally related) to bind to common places has been also been shown in *Drosophila* by the modENCODE consortium[88]. Using the binding profiles of 41 TFs in early embryonic development, the authors identified 1962 highly occupied target (HOT) regions (defined as regions bound by ~10 TFs), which have recently been shown to be *bona fide* enhancers, with 94% (of the 108 tested regions) being active during embryogenesis[163].

Associated with stringent cut-offs, ChIP approaches provide a straightforward means to identify enhancers in a genome-wide and unbiased manner with impressive success rates. The timing of enhancer activity might not correspond to the first observed binding event, as the

presence of more than one TF could be required for activity. Thus, the notion of CRM activity should be clearly distinguished from that of CRM identification. However, ChIP approaches are not always feasible, as a specific antibody must be available. Furthermore, biological material should be available in sufficient amounts, which may not be possible. Finally, targeted proteins might be expressed in several tissues or ubiquitously, thereby complicating tissue specific analysis in whole organisms. Even when technically feasible, a ChIP approach might reveal quickly unaffordable in terms of cost or time when the number of TFs, experimental conditions and replicates becomes too high. These limitations have encouraged the development of alternative approaches aiming at determining the complete repertoire of regulatory regions in the genome.

1.2.2.2 ChIP against co-factors and methods exploiting chromatin structure

Regulatory elements are characterized by the presence of sequence-specific TFs and co-factors. To bind their target TFBSs, TFs need to access the DNA and therefore require both the chromatin to be open and their TFBSs to be devoid of nucleosome (excepting the pioneer factors mentioned earlier). This phenomenon has been initially observed in *Drosophila,* where it has been shown that the TSSs of active genes are hypersensitive to both DNaseI and micrococcal nuclease[164], in correlation with a loss or a destabilization of nucleosomes. Nagy *et al.* demonstrated that, following phenol-chloroform extraction of formaldehyde-crosslinked yeast chromatin, the genomic regions immediately upstream of genes were preferentially segregated into the aqueous phase[165]. This phenomenon was interpreted to indicate relatively inefficient cross-linking between proteins and DNA at these regions, and further linked to an absence of nucleosomes. Called FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements), this protocol enabled to confirm that FAIRE-enriched regions exhibit a strong negative correlation with nucleosome occupancy [166].

DNAseI digestion has been combined with tiling arrays (DNAse-chip) [167,168] and with sequencing (DNAse-seq) [169] to identify all open chromatin locations under a particular condition. Similarly, FAIRE-chip and FAIRE-seq have been developed[170]. Both DNAseI and FAIRE based assays were used to isolate a variety of regulatory regions (promoters,

insulators, enhancers, locus control regions, silencers, etc.) independently of the specific proteins responsible for the absence of nucleosomes. Importantly, the capacity to identify cell-type specific regulatory regions has been suggested for both DNaseI [169] and FAIRE[166] assays. Comparing regions identified by the two approaches, Giresi *et al.* reported that FAIRE-isolated regions are largely coincident with the location of DHSs[166,170]. Recently, Song *et al.* performed both DNAse-seq and FAIRE-seq in seven human cell lines and identified altogether more than 870,000 DHSs covering nearly 9% of the genome[171]. The authors reported that the combination of DNaseI and FAIRE is more effective than either assay alone in identifying likely regulatory elements. As suggested previously, open chromatin common to all seven cell types tended to be at or near TSSs and to be coincident with CTCF binding sites, whereas open chromatin sites found in only one cell type were typically located away from TSSs and contained DNA motifs recognized by regulators of cell-type identity (i.e. putative CRMs).

More recently, investigators took a slightly different approach and mapped co-factors like p300 (or CBP) that are recruited by sequence-specific TFs, including at enhancers[65]. Using dissected mouse tissues (embryonic forebrain, midbrain and limb at stage E11.5), Visel *et al.* demonstrated that p300 *in vivo* binding reflects enhancer activity in a tissue specific manner[17]. Of the 86 putative enhancers predicted based on p300 binding (in more than one tissue for 32 of these predictions) tested in transgenic mouse embryos, 88% showed enhancer activity and 80% were active at stage E11.5 in the predicted tissue (i.e. in the tissue where p300 was assayed). Notably, 22 of the 32 enhancers (69%) identified by p300 peaks in more than one tissue perfectly recapitulated the predicted expression patterns. Blow *et al.* have used a similar approach to locate heart enhancers using p300 in mouse embryonic heart tissues (also at stage E11.5) and tested 130 candidate enhancers in transgenic mouse embryos[41]. The authors further demonstrated that 97 (75%) of them drive expression in E11.5 embryos, of which 81 (84%, or 62% of the initial 130) are active in the developing heart. Interestingly, identified heart enhancers exhibited much less evolutionary constrains than forebrain, midbrain and limb enhancers identified by Visel *et al*. These results indicate that tissue specific mapping of p300 provides an accurate means for identifying enhancers and their associated tissue-specific activity (although to a lower extend).

As mentioned earlier, p300 is recruited by different sequence-specific DNA binding proteins and is thus found only at a subset of DHS sites[99]. Thus, promoter distal DHS sites represent a more heterogeneous population of CRMs, as compared to p300 binding sites. Note that only TSS-distal DHSs and p300 peaks should be considered for enhancer identification but that, in the absence of additional support, these locations might represent alternative unannotated promoters (e.g. H3K4me3 marked regions are typically excluded). It is important to stress that these methods inherently probe general and ubiquitous features and are thus most useful when done in a tissue-specific context. Hence, their application remains limited to studies with cell lines or dissected organs until *in vivo* tissue-specific methods become available.

1.2.2.3 ChIP against histone post-translational modifications

The last approach used to locate CRMs is based on histone post-translational modifications (cf. sections 1.1.3.3 and 1.1.3.4). Some of these studies exploit the identified CRM signatures to predict enhancers, or to validate their predictions.

Heintzman 2007 et al. performed ChIP-chip against the core histone H3, 5 histone modifications, Pol II, TAF1 and p300 in human HeLa cells before and after treatment with INFγ, which induces p300 binding as part of its induced cellular response [65]. Using known TSSs (to locate promoters) and distal p300 binding (to define enhancers), these authors used a supervised approach and built a model based on H3K4me3 and H3K4me1 profiling to predict 389 regions in untreated cells and 324 regions in treated cells (89% of regions in common). They assessed the validity of these predictions by indirect means, such as the distance from the TSS (85% of predictions being more than 2.5 kb from a TSS), the presence of strongly conserved sequence in 53% of the predictions, the overlap with p300 or TRAP220 (also a transcriptional co-activator) bound regions or with DHSs for 63.5% of the predictions, or with independently computationally predicted CRMs (PReMods, based on clustering of conserved TF binding motifs). The authors further tested 4 regions using *in vitro* luciferase assays, where 3 of them gave some activity. Importantly, these 4 regions were selected based on their overlap with STAT1 binding (observed by ChIP-chip after INFγ treatment) and therefore do not constitute an unbiased set to assess the method accuracy.

Two years later, Heintzman *et al.* used the same methodology to predict enhancers in 5 different human cell lines and could demonstrate that 7 out of 9 (78%) of the regions tested function as regulatory elements *in vitro* (luciferase assay) [80].

Comparing unstimulated and activated mouse macrophages, De Santa *et al*. used a supervised approach to train a Support Vector Machine (SVM) to discriminate between promoter and enhancers based on the H3K4me1 and H3K4me3 signal[103]. Extragenic p300 binding was used to define the enhancers of the SVM training set. The model was then used to classify 4,588 extragenic Pol II peaks (identified by ChIP-Seq) as putative enhancers or promoters. The authors first verified that (1) predicted regions have a significantly higher conservation than random genomic sequence, and (2) 84% of predicted enhancers overlap with PU.1 bound regions. Finally, the authors tested 7 regions (associated with Pol II occupancy) by *in vitro* luciferase assays. Based on published data, 5 of these regions (71%) presumably correspond to *bona fide* enhancers (cf. the error bars shown in Figure 6C of [103]).

Ernst *et al.* defined chromatin states using ChIP-seq data for CTCF and 8 histone modifications in 9 human cell types[87]. Here the authors start from a chromatin-centric view and use a HMM to segment the genome into regions with different chromatin states. They then correlate each set of genomic regions linked to a specific chromatin state to known annotations (gene bodies, promoters, enhancers and insulators…). Correlating the putative enhancer predictions to gene expression data, they separated the enhancer predictions into 4 classes, based on their proximity to genes that are (1) highly expressed (referred to as strong enhancers), (2) intermediately expressed, (3) lowly expressed, and (4) not expressed ('inactive enhancers'). Experimentally, the authors selected 18 regions corresponding to "strong enhancers" and tested them using *in vitro* luciferase assays. Importantly, these "strong enhancers" are also enriched for H3K4me3, a mark typically found on active promoters[65,103]. It is of note that a luciferase assay cannot distinguish between the activities of an enhancer or a promoter, as both can lead to luciferase expression, with a strong promoter potentially having a higher chance of doing so. Based on published figures, we estimate that between 50% and 75% of the regions tested function as regulatory elements *in vitro*.

Finally, Nègre *et al.* generated *Drosophila* genome-wide maps for 6 histone modifications, CBP and Pol II across twelve stages of development[155]. Importantly, these maps were generated using whole animals. To identify putative enhancers, the authors required the combined presence of CBP and H3K4me1 and tested 33 sequences using

reporter assays in transgenic *Drosophila*. Thirty of these produced specific expression patterns during embryonic development. Unfortunately, the authors do not discuss the potential concordance between the timing of CBP binding and that of enhancer activity. As mentioned earlier, p300 is also found at poised enhancers[98]. Indeed, Rada-Iglesias *et al.* showed that p300 bound regions enriched in H3K27me3 (and lacking H3K27Ac, which represented ~30% of p300 binding in human ES cells) can function as enhancers at distinct developmental stages and anatomical locations, using zebrafish embryo transgenic reporter assays, for 8 out of 9 of the tested sequences[98].

2 Aim of the PhD

CRMs integrate and translate the input of multiple factors into spatio-temporal patterns of gene expression. The characterization of CRMs is therefore central to understanding gene regulation and metazoan development. In previous studies, we demonstrated that *in vivo* binding profiles of TFs could not only be used to locate enhancers, but also to predict their spatio-temporal activity. It is unfortunately not feasible to profile the hundreds or thousands of TFs, in all different tissues, at the different developmental stages of an organism's life. We therefore need alternative approaches to identify comprehensive sets of active enhancers *in vivo* at high accuracy in a TF agnostic manner. Recent studies used DHSs, p300/CBP or FAIRE to globally locate enhancers. Nevertheless, these approaches indifferently identify various types of regulatory regions (enhancers, insulators, promoters…) and are not necessarily informative regarding the activity state of the putative CRMs (in particular for DHS and FAIRE). Other studies have used histone PTMs to define different chromatin states that associate with genomic features and their activity state. Importantly, these approaches could only be conducted in cell lines or with dissected tissues, as the signal from whole embryo experiments is not tissue specific.

In this context, our first objective was to study chromatin state at enhancers within the developing embryo in a tissue specific way. Toward this goal, we needed first to develop a protocol enabling tissue-specific ChIP. Next, using *Drosophila* mesoderm as a model, our goal was to identify a subset of chromatin marks specific to active enhancers. Finally, we aimed at using this information to predict enhancers active in the mesoderm using a machine learning approach.

A second objective of this work was to increase our understanding of how CRMs function and, in particular, whether TFBSs found in CRMs obey to specific architectural rules. To address this question, we chose the specification of the dorsal mesoderm into the cardiac and visceral mesoderm. Although cardiac enhancers display relatively weak sequence conservation, the heart *cis*-regulatory network is one of the best-conserved networks from fly to human. Importantly, essential TFs of this network have been shown to cooperate genetically and to form protein-protein interactions. This system is thus particularly relevant to study potential *cis*-regulatory constraints. To address this challenge, we analysed the

binding profiles of five TFs essential for *Drosophila* heart development by ChIP-Chip and deciphered the organization of the CRMs predicted based on binding correlations.

3 Results

3.1 Article 1. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development.

3.1.1 Introduction

Previous studies aiming at deciphering the chromatin state on enhancers have been conducted in cell lines[62,63,65,80,88,96,98,103], or with whole organisms[155]. Approaches based on tissue culture allow probing a (mostly) uniform cell population, but remain essentially akin to *in vitro* assays, as the cells are cultured outside the living and developing organism. In contrast, ChIP against ubiquitous factors in whole embryos yields mixed and overlaid signals from various tissues and cell types, which severely limits their interpretability. Consequently, a major challenge remained to extract the cell type specific signatures of otherwise ubiquitous (or non-tissue specific) factors from complex tissues and organism.

Notably, the conclusions of different cell-culture and dissection studies appear contradictory at various levels. First H3K4me1 has been considered as an indicator of active enhancers[65], while two recent studies concluded that its presence did not correlate with activity[96,98]. Second, the presence of H3K4me3 has been recently described at enhancers[79], while the vast majority of studies specifically associated this modification with active promoters and its depletion as an indicator of enhancers. Third, all previous studies used sets of putative enhancers like TSS distal regions identified by p300/CBP binding or using DHSs. Finally, the activity status of a putative enhancer was assessed using the expression of the closest gene, which may be sometime misleading. Indeed, genes are regulated by multiple enhancer elements, both distal and proximal, that have independent or partially overlapping effects on gene activity. Several studies have shown that enhancers can be located in the body of other genes, while genes can be found between enhancers and their target genes.

Our aim was to overcome these limitations and evaluate the chromatin state at *bona fide* enhancers *in vivo*. Towards this goal, we first developed a novel *in vivo* tissue-specific ChIP-seq protocol and used it to map nucleosomes marked by H3K4me1, H3K4me3, H3K36me3, H3K79me3, H3K27Ac and H3K27me3 and Pol II occupancy, in *Drosophila* mesoderm at 6-8h AEL. We then used a set of enhancers characterised *in vivo*, which we have curated for their exact spatio-temporal activity, and used it to discover what differentiates an active from an inactive enhancer, both spatially (between tissues) and temporally (activity switches within the same tissue over time). Using Bayesian inference, we subsequently predicted locations of regulatory regions and their activity status in the mesoderm at 6-8h, and validated 89% of them to be active *in vivo* at the predicted time. Finally, we integrated temporal binding maps of 5 key mesodermal TFs and identified temporal signatures of enhancer activity in terms of chromatin state, TF binding, and nucleosome displacement.

3.1.2 Personal contributions to this work

In this work, I participated in the design of the study, conceived and implemented the complete ChIP-seq analysis pipeline (with the exception of the quality control of sequencing results and read mapping, which were performed by Nicolas Delhomme), and applied it to the datasets generated. I further assembled all gene and CRM lists used, conceived the Bayesian modelling approach and generated the subsequent CRM predictions. Finally, I contributed to the writing of the manuscript (main text, methods, figures, supplementary materials, rebuttal, revisions, and proofing process).

3.1.3 Article

Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development

Stefan Bonn[1,2], Robert P Zinzen[1,2], Charles Girardot[1,2], E Hilary Gustafson[1], Alexis Perez-Gonzalez[1], Nicolas Delhomme[1], Yad Ghavi-Helm[1], Bartek Wilczyński[1], Andrew Riddell[1] & Eileen E M Furlong[1]

Chromatin modifications are associated with many aspects of gene expression, yet their role in cellular transitions during development remains elusive. Here, we use a new approach to obtain cell type–specific information on chromatin state and RNA polymerase II (Pol II) occupancy within the multicellular *Drosophila melanogaster* embryo. We directly assessed the relationship between chromatin modifications and the spatio-temporal activity of enhancers. Rather than having a unique chromatin state, active developmental enhancers show heterogeneous histone modifications and Pol II occupancy. Despite this complexity, combined chromatin signatures and Pol II presence are sufficient to predict enhancer activity *de novo*. Pol II recruitment is highly predictive of the timing of enhancer activity and seems dependent on the timing and location of transcription factor binding. Chromatin modifications typically demarcate large regulatory regions encompassing multiple enhancers, whereas local changes in nucleosome positioning and Pol II occupancy delineate single active enhancers. This cell type–specific view identifies dynamic enhancer usage, an essential step in deciphering developmental networks.

Distinct chromatin modifications are associated with many aspects of gene expression; for example, trimethylation of histone H3 on lysine 4 (H3K4me3), trimethylation of histone H3 on lysine 79 (H3K79me3) and trimethylation of histone H3 on lysine 36 (H3K36me3) reflect promoter activity, gene-body transcription and, to some degree, exon-intron usage[1,2] and are highly correlated with gene expression levels[2–4]. Other histone modifications, in particular monomethylation of histone H3 on lysine 4 (H3K4me1) and acetylation of histone H3 on lysine 27 (H3K27ac), have proven to be a very effective means to determine the location of *cis*-regulatory elements (CRMs)[2,5]. However, linking chromatin modification to the activity of enhancers remains a key challenge. Studies in embryonic stem (ES) cells found a positive correlation between the presence of H3K27ac on putative enhancers and the activity of the closest proximal gene, but opposing results were reported for the presence of trimethylation of histone H3 on lysine 27 (H3K27me3) on regulatory elements[6,7]. In contrast, a study in human $CD4^+$ T cells, investigating a much more extensive collection of chromatin marks, found no significant correlation between any chromatin modification and enhancer activity[8]. These discrepancies may have arisen from the different methods used to define large sets of putative enhancer elements, using either a collection of chromatin marks in noncoding regions[6,7] or DNase I hypersensitive sites[8]. Even with a *bona fide* set of enhancers at hand, the activity of the closest proximal gene may be a poor proxy for enhancer activity, as genes are regulated by multiple enhancer elements, both distal and proximal, that have independent or partially overlapping effects on gene activity.

Much of our knowledge on the role of chromatin modification has come from cell culture studies[2,4,6,7,9,10], but there is little information on how they reflect transcriptional networks driving embryonic development[11]. For example, histone modifications undergo dramatic changes over the 7–12 d of ES cell differentiation[4,6,7,9,10], reflecting changes in promoter and enhancer usage similar to those observed for transcription factor occupancy[12–15]. In contrast, many cell fate transitions during embryonic development occur on the order of hours, yet it is not known how this relates to dynamic changes in chromatin state. More fundamentally, it is currently not clear how accurately changes in chromatin modification reflect the precise timing of enhancer, promoter or gene activity. Within an *in vivo* context, available chromatin data comes from dissected tissues[16] or whole embryos, yielding mixed signals from heterogeneous cell types[1,17–20]. The latter studies[1,17–20] form part of an important effort to annotate the genome, but it is essential to move beyond whole-embryo data[21] to understand the dynamic interplay between chromatin modification and transcription factor occupancy at cell type–specific resolution during embryonic development.

[1]Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. [2]These authors contributed equally to this work. Correspondence should be addressed to E.E.M.F. (eileen.furlong@embl.de).

Received 31 May 2011; accepted 7 December 2011; published online 8 January 2012; doi:10.1038/ng.1064

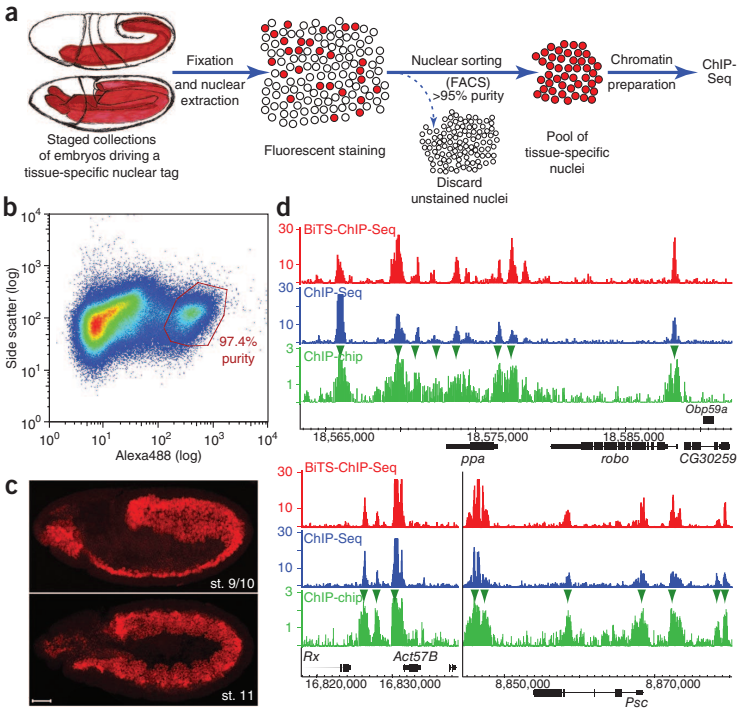


Figure 1 BiTS-ChIP facilitates cell type–specific ChIP in a multicellular context. (**a**) Method outline. Embryos with a transgene encoding a tagged nuclear protein expressed in a specific tissue (SBP-H2B) are collected and aged to the desired stage (6–8 h) and then cross-linked with formaldehyde. Fixed nuclei are extracted, fluorescently stained for the tag and sorted by FACS to >95% purity. Chromatin is extracted, sheared, immunoprecipitated and subjected to Solexa sequencing. (**b**) FACS sorting of nuclei results in very high purity. Typical FACS scatter graph relating side scatter (*y* axis) to fluorescent intensity (Alexa488; *x* axis). The red gate indicates the sorting events that were isolated (only events containing single fluorescent particles were selected) and processed further. This representative sample yielded ~97.4% purity from a single sort, as estimated by epifluorescent inspection of DAPI-counterstained sorted nuclei. (**c**) Transgenic embryos encoding a tagged nuclear protein: the transgenic *twi*^{IPEMK}*::SBP-His2B* line directs expression of tagged histone H2B throughout the mesoderm, representing ~20% of the embryo at the indicated stages. Shown are embryos stained for SBP (red) at stages 9/10 (top) and stage 11 (bottom). Left, anterior; up, dorsal, st., stage. Scale bar, 50 µm. (**d**) Sorting fixed nuclei does not affect the regulatory context. Representative loci showing Mef2 binding data determined by three methods: BiTS-ChIP followed by Solexa sequencing (BiTS-ChIP-Seq, red), conventional ChIP-Seq (blue) and ChIP-chip[27] (green). Shown are the background-subtracted read per genomic coverage (RPGC) values (BiTS-ChIP-Seq and ChIP-Seq) and the mean $\log_2$ ratios for ChIP-chip (computed peaks are indicated by arrow heads).

RESULTS

Cell type–specific ChIP in developing embryos

We developed a method to batch isolate tissue-specific chromatin for immunoprecipitation (BiTS-ChIP), which uses a transgene to express a tagged nuclear protein specifically in the cell type of interest. Entire embryos are covalently cross-linked[22], and intact, fixed nuclei are isolated and sorted by FACS to obtain pure populations of nuclei from specific cell types (**Fig. 1a**); the average purity of all samples used in this study was 97.4% (**Fig. 1b**). To generate a widely applicable protocol, we optimized our ChIP procedure[22] to use less chromatin, thereby allowing multiple ChIP experiments to be performed from a single FACS sort (see Online Methods).

We applied BiTS-ChIP to examine six chromatin marks and RNA polymerase II (Pol II) occupancy in mesodermal cells during *Drosophila* development, for which extensive transcription factor occupancy data are available[14,23–27]. Transgenic *Drosophila* strains expressing a tagged histone under the control of a mesodermal enhancer (**Fig. 1c**; see Online Methods) were used for staged embryo collections at 6–8 h of development (stages 10–11) and processed by fixation, FACS nuclear sorting and ChIP-sequencing analysis (ChIP-Seq) to examine chromatin modifications at promoters (H3K4me3 and H3K27ac), gene bodies (H3K79me3 and H3K36me3), *cis*-regulatory elements (H3K4me1 and H3K27ac) and repressed regions (H3K27me3), as well as Pol II occupancy and histone H3 density. These six chromatin marks, in addition to histone H3 density, represent four of the five major chromatin types recently defined in *Drosophila*[28], with the exception of silent heterochromatic regions (**Supplementary Note**).

BiTS has high sensitivity and specificity

The dissociation of cells from tissues and embryos leads to a transcriptional stress response, which is typically observed with FACS sorting of live cells. Covalent cross-linking before embryo dissociation avoids this problem by blocking all transcriptional activity. This key feature of the BiTS-ChIP protocol preserves the transcriptional context during nuclear sorting and facilitates cell type–specific analysis of transcription factor binding, which is not possible with native ChIP. We directly confirmed this by performing ChIP experiments on a mesoderm-specific factor, Mef2, which has a conserved role in myogenesis in insects and vertebrates[29]. Mef2 occupancy in sorted nuclei (BiTS-ChIP) was remarkably similar to that observed with standard ChIP-Seq and ChIP-chip[27] analyses (**Fig. 1d**), with >81% of peaks being called by any two methods (**Supplementary Fig. 1**), thus validating the reliability of the BiTS-ChIP method.

A second important feature of BiTS-ChIP is the high specificity of the data it generates. Genes that are known to be expressed exclusively in mesoderm at 6–8 h of development showed high enrichment for H3K4me3 and H3K27ac at their promoters and H3K79me3 on their gene bodies (**Fig. 2a**, *lmd*, and **Supplementary Fig. 2a–d**), whereas mesodermally inactive genes typically showed no sign of transcription (**Fig. 2b**, *tll*, and **Supplementary Fig. 2c–f**). To evaluate tissue specificity more globally, we used annotated data of the spatio-temporal expression patterns of over 6,000 *Drosophila* genes[30]. Genes that are mesodermally (but not ubiquitously) expressed at 6–8 h of development had high levels of chromatin modifications associated with active transcription (H3K4me3, H3K27ac, H3K79me3 and H3K36me3) and Pol II occupancy (**Fig. 2c–f** and **Supplementary Fig. 3**,

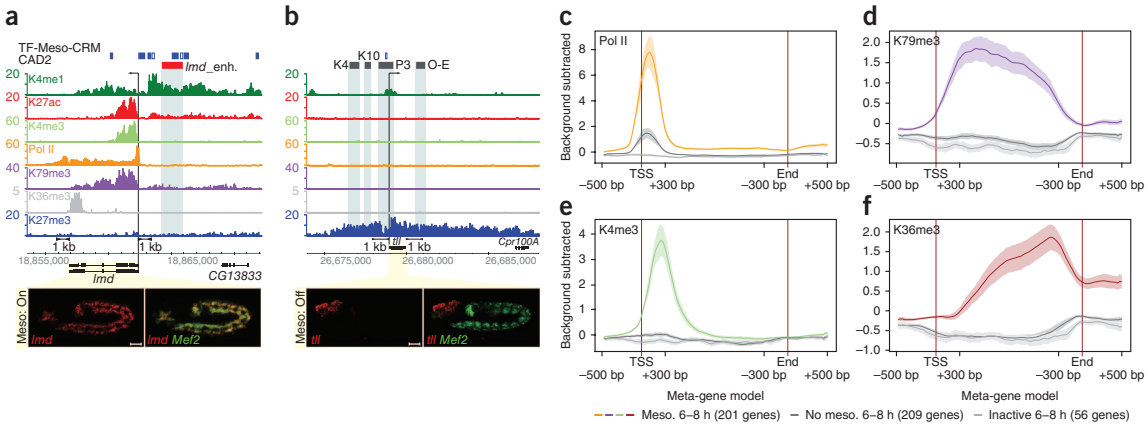


Figure 2 BiTS-ChIP has high tissue specificity and sensitivity. (**a**,**b**) BiTS-ChIP signal enrichment for histone modifications (RPGC, H3 subtracted) and Pol II occupancy (RPGC, input subtracted) for a mesodermally (*lmd*) (**a**) and non-mesodermally expressed (*tll*) (**b**) gene. Promoter arrows indicate transcription direction. Blue boxes show CRMs defined by mesoderm transcription factor occupancy (TF-Meso-CRMs: filled symbol, bound at 6–8 h; unfilled symbol, not bound at 6–8 h). Red and gray boxes show CAD2 enhancers (red, mesodermally active; gray, mesodermally inactive). Gray shading indicates CAD2 regions. A distance filter (double-tailed arrows) served to exclude enhancers within 1 kb of genes. Bottom, embryos showing the expression pattern of the indicated gene (red) by double *in situ* hybridization with a mesodermal marker (*Mef2*, green). Scale bars, 50 µm. (**c**–**f**) Global assessment of tissue specificity. Large-scale *in situ* hybridization data (from the Berkeley *Drosophila* Genome Project (BDGP)) were used to identify mesodermally (but not ubiquitously) expressed genes at 6–8 h of development ('meso. 6–8 h', colored line, 201 genes), genes expressed only non-mesodermally at 6–8 h ('no meso. 6–8 h', dark gray line, 209 genes) or those expressed only at later stages non-mesodermally ('inactive 6–8 h', light gray line, 56 genes). Background subtracted signal is shown for Pol II (**c**), H3K79me3 (**d**), H3K4me3 (**e**) and H3K36me3 (**f**); shading indicates 95% confidence intervals. 'Meta-gene models' were calculated for genes of ≥850 bp to ensure reliable signal interpolation (see Online Methods). Whereas Pol II and H3K4me3 signals were highly enriched around the TSSs and H3K79me3 and H3K36me3 were enriched across the transcription units of mesodermally active genes, there was very little signal at non-mesodermal and inactive genes.

colored lines). In contrast, genes that are not mesodermal but are active in other cells at this stage of development showed very low levels of chromatin signatures linked to active transcription (**Fig. 2c–f** and **Supplementary Fig. 3**, dark lines). The remaining Pol II signal at non-mesodermal genes (**Fig. 2c**, dark gray line) in the absence of active chromatin marks suggests Pol II pausing[31] and was absent at genes not expressed in any tissue at this stage (**Fig. 2c**). The ability to detect tissue-specific gene regulation was markedly reduced when using whole-embryo data[32] (**Supplementary Fig. 4**); the lower sensitivity and general lack of any spatial information in whole-embryo chromatin data highlight the limitations in using this approach to dissect regulatory programs driving tissue development.

A new role for H3K79me3 on developmental enhancers

To directly assess the relationship between chromatin modifications and enhancer activity, we assembled publicly available information on the activity of 465 characterized *Drosophila* enhancers examined *in vivo* using transgenic reporter assays (CRM Activity Database 2 (CAD2); see Online Methods and **Supplementary Table 1**), in which reporter gene expression provided a direct transcriptional read-out of the spatio-temporal activities of the enhancers (**Fig. 3a**). Each literature-annotated enhancer was manually curated and its activity mapped to *Drosophila* embryonic tissues. Enhancers were broadly grouped into those with mesodermal or non-mesodermal activity at each developmental stage (**Fig. 3a** and Online Methods). To avoid potentially confounding signatures from the transcription of genes, enhancers within 1 kb of gene boundaries were excluded (see Online Methods and **Supplementary Table 2**). The remaining 144 intergenic enhancers were used for all subsequent analyses.

We first examined the general distribution of chromatin marks on developmental enhancers, without considering their activity status. Of the 144 intergenic enhancers, 111 (77%) were enriched for H3K4me1, and 23 (16%) were enriched for H3K27ac (**Supplementary Fig. 5**). Pol II occupancy was seen at 11 (8%) of the developmental enhancers, in line with recent observations in mice[33–35] and human ES cells[7]. We also observed H3K79me3, a modification previously only associated with active gene transcription, on 21 (15%) of the gene-distal enhancers, indicating a potentially new role for this chromatin mark. Although the presence of H3K79me3 on gene bodies is associated with Pol II elongation, only 7 (33%) of the enhancers containing H3K79me3 also had Pol II binding, suggesting either Pol II–independent trimethylation of H3K79 at enhancers or H3K79me3 perdurance after transient Pol II occupancy. In contrast, H3K36me3, another mark associated with active transcription, was not present at any enhancer element examined[1,2]. H3K27me3, a modification associated with Polycomb-mediated repression, was present at 95 (66%) of all the developmental enhancers examined (**Supplementary Fig. 5**). The general presence of these chromatin modifications (H3K27me3, H3K4me1, H3K27ac and H3K79me3) and of Pol II is significantly greater on developmental enhancers than expected by chance (**Supplementary Fig. 5**), suggesting an association with enhancer function.

We note that chromatin marks typically spanned large genomic regions that contained several enhancer elements; for example, H3K79me3 (**Fig. 2a** and **Supplementary Fig. 6a**,**c**) and H3K27me3 (**Fig. 2b** and **Supplementary Fig. 6e**,**f**) often spread from the gene body into upstream enhancer regions. This is in contrast to Pol II occupancy, which was restricted to small local regions within known enhancer elements (**Supplementary Fig. 6a–c**).

Diverse chromatin marks and Pol II indicate active enhancers

To assess the relationship between chromatin marks and the activity status of an enhancer, we divided the developmental enhancers into

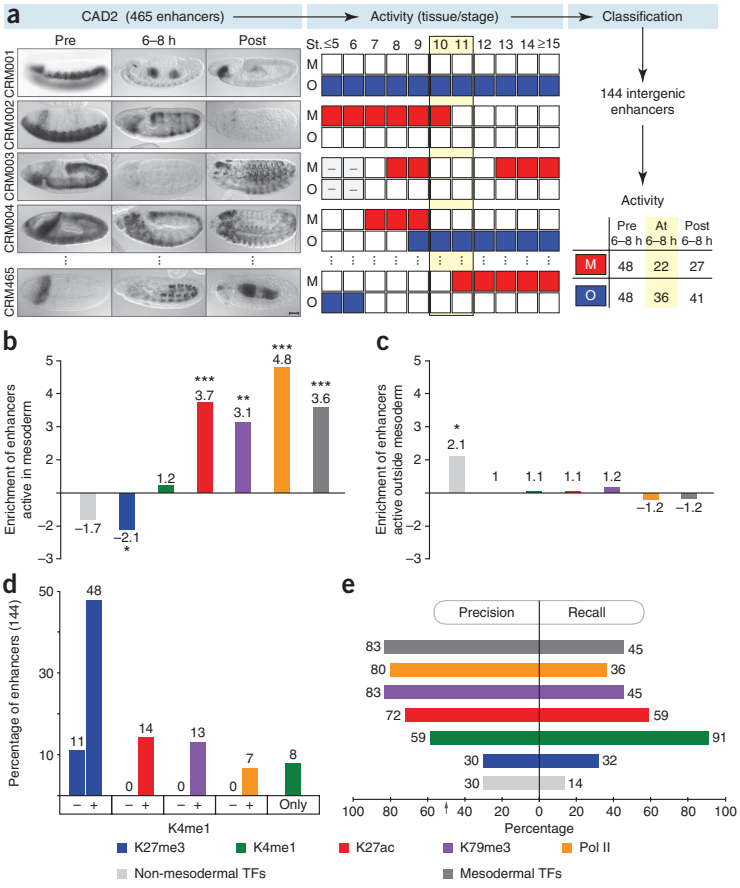


Figure 3 Chromatin marks and Pol II presence are highly correlated with enhancer activity. (**a**) CAD2 enhancer activity annotation and filtering. CAD2 contains literature-based enhancer activity information. Left, reporter gene expression directed by CRMs in transgenic embryos. Middle, reported activity was evaluated by stage for activity in mesoderm (M, red), activity elsewhere (other: O, blue) or no activity (white). –, no information (gray). Right, 144 of 465 enhancers are located >1 kb away from genes and do not overlap with H3K4me3 peaks. Activity data (bottom right) tabulated nonexclusively for tissue and stage. Yellow shading indicates the investigated developmental stages. (**b**,**c**) Correlating chromatin marks and Pol II occupancy with enhancer activity. Enrichment of enhancers active at 6–8 h mesodermally (**b**) or non-mesodermally (**c**) within regions marked by H3 modifications or transcription factor or Pol II occupancy. The *y* axes show fold change relative to background (where 22 of 144 enhancers are mesodermally active and 31 of 140 are active exclusively outside mesoderm—enhancers active in both were removed). Significance was estimated using a two-sided Fisher's exact test: $*P \leq 0.05$; $**P \leq 0.001$; $***P \leq 0.0001$. (**d**) Enhancers with H3K27ac, H3K79me3 or Pol II are co-marked by H3K4me1. Enhancers were grouped by H3 modifications or Pol II occupancy and inspected for H3K4me1 presence (+) or absence (–). The green bar represents enhancers carrying H3K4me1 only. (**e**) Precision and recall for active mesodermal enhancers at 6–8 h by chromatin marks, Pol II presence and transcription factor occupancy. Left, precision of mesodermal enhancers (gray arrow indicates the baseline; 22 of 43 enhancers had mesodermal activity at 6–8 h). Right, recall of mesodermal enhancers active at 6–8 h.

two groups on the basis of their reported activities: enhancers active in mesoderm at 6–8 h of development and those without reported mesodermal activity (**Fig. 3** and Online Methods). Examining the presence of chromatin marks on active enhancers, we observed that although H3K4me1 is present on regulatory elements as previously reported[5,8,9], it provides no information on their activity status—enhancers marked by H3K4me1 were not significantly enriched for activity over background (**Fig. 3b**). This finding is in line with a recent report showing H3K4me1 enrichment in the vicinity of both active and inactive genes[6].

H3K27me3 was significantly depleted on active mesodermal enhancers (2.1 times; **Fig. 3b**). A recent study proposed that the presence of H3K27me3 at H3K4me1-defined regulatory regions indicates enhancers in a poised state, ready for subsequent activation during ES cell differentiation[7]. This is in contrast to what we observed in the context of embryonic development, as many enhancers marked by H3K27me3 in the mesoderm were active in other cell types at this stage of development but did not become active in mesodermal cells (**Supplementary Figs. 6e**,**f** and **7**), indicating that these enhancers were in a repressed rather than a poised state.

In contrast, enhancers with H3K27ac and H3K79me3 marks and Pol II occupancy were significantly enriched for mesodermal activity (by 3.7-, 3.1- and 4.8-fold, respectively; **Fig. 3b**). This enrichment was not seen when examining mesodermally inactive enhancers (**Fig. 3c**), indicating that H3K27ac and H3K79me3 marks and Pol II binding distinguish active and inactive enhancers with high precision. However, their recovery varied substantially, with H3K27ac recalling 13 (59%), H3K79me3 recalling 10 (45%) and Pol II recalling 8 (36%) of the active mesodermal enhancers (**Fig. 3e**). Of note, two active enhancers did not contain significant levels of any of the six chromatin marks studied here, suggesting that these regulatory regions may be marked by other chromatin signatures[8,36] or that covalent nucleosome modifications are not required for their activity.

Taken together, our results show that there is not just one specific chromatin mark associated with active enhancers, such as H3K27ac[6,7], but instead active regulatory regions are enriched for multiple chromatin modifications and Pol II occupancy.

Chromatin modifications and the timing of enhancer activity

Development requires very rapid transitions from one regulatory state to another, especially in *Drosophila*, in which the entire process of embryogenesis occurs within ~18 h. Taking advantage of this rapid pace, we assessed the relationship between temporal changes in enhancer activity and chromatin modification within a 2-h window (6–8 h) by dividing the enhancers into three temporal classes: those that are mesodermally active during the 2-h time period (at 6–8 h), those that are only active earlier (<6 h) and those that are active only

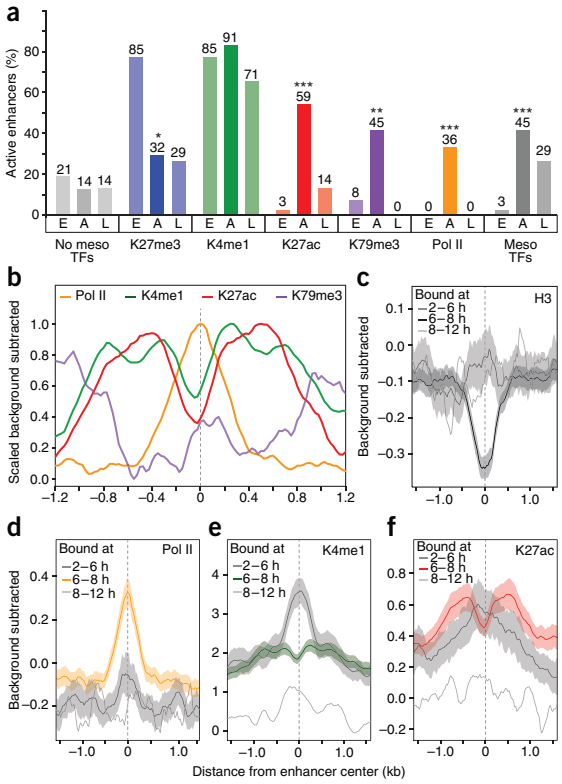


Figure 4 Pol II occupancy and local nucleosome positioning identify temporal enhancer activity. (**a**) The presence of chromatin marks and Pol II was highly correlated with the timing of enhancer activity. Analysis of three temporal classes of mesodermal enhancers: active only early (<6 h AEL, $n = 39$; E), late (>8 h AEL, $n = 7$; L) or at 6–8 h ($n = 22$; A). Bar graphs show the percentage of enhancers containing the indicated chromatin marks or having Pol II or transcription factor binding. Presence of H3K27ac, H3K79me3 and Pol II at enhancers at 6–8 h was highly correlated with the timing of enhancer activity, whereas H3K4me1 was present irrespective of activity. Significance was calculated using a two-sided Fisher's exact test: $^{*}P$ 0.05; $^{**}P$ 0.001; $^{***}P$ 0.0001. (**b**–**f**) Distribution of Pol II and chromatin mark quantitative signals across TF-Meso-CRMs. *x* axes show distance from CRM center defined by transcription factor binding; *y* axes show background-subtracted signal at 6–8 h. (**b**) Spatial distribution of Pol II, H3K4me1, H3K27ac and H3K79me3 on enhancers with signal normalized to [0,1]; Pol II signal is centered, and chromatin modifications show bimodal distributions around Pol II. Signals for H3 (**c**), Pol II (**d**), H3K4me1 (**e**) and H3K27ac (**f**) on intergenic CRMs bound by transcription factors at 6–8 h (colored line, $n = 293$), enhancers bound only early (2–6 h, dark gray, $n = 72$) or only later (8–12 h, light gray, $n = 8$). Shading indicates 95% confidence intervals. Pol II signal peaks at the time of transcription factor binding but not when transcription factors are no longer bound (orange versus dark gray lines in **d**. H3K4me1 and H3K27ac signals exhibit bimodal distributions at the time of transcription factor binding but peak centrally thereafter (green and red versus dark gray lines in **e** and **f**, respectively).

at later stages of development (>8 h) (see Online Methods). The presence of H3K27ac and H3K79me3 marks and Pol II on enhancers at 6–8 h was highly correlated with the precise timing of enhancer activity (**Fig. 4a**). Pol II occupancy was particularly transient, being absent from enhancers that were only active at slightly earlier stages of development and from those that had just become inactive in mesodermal cells (**Fig. 4a**). This result is in contrast with our observations for H3K4me1, which persisted at 6–8 h on early enhancers, even though these enhancers were no longer active in the later time frame (**Fig. 4a**).

The timing of enhancer activity is highly correlated with the timing of transcription factor occupancy, as has been shown for mesoderm-specific factors[14,15,26]. We therefore assessed the relationship between chromatin marks and temporal transcription factor occupancy using a large collection of enhancer elements defined by the binding of mesoderm-specific transcription factors at multiple stages of development (TF-Meso-CRMs[27]; Online Methods). In examining the quantitative signals across these TF-Meso-CRMs, we observed very different spatial distributions for chromatin modifications compared to Pol II occupancy (**Fig. 4b**–**f**). H3K27ac and H3K4me1 marks exhibited a bimodal distribution on enhancer elements at the time of transcription factor binding (**Fig. 4b**), presumably as a result of nucleosome displacement by transcription factors at their site of occupancy, as evidenced by the positioning of histone H3 (**Fig. 4c**). In contrast, when transcription factors were no longer bound (CRMs bound before 6 h), histone modifications peaked around the earlier transcription factor binding site(s), suggesting nucleosome remodeling at these developmental enhancers (**Fig. 4e**,**f**, dark gray lines). Therefore, the local distribution of chromatin modifications around transcription factor binding sites rather than the simple presence or absence of these marks may better reflect enhancer activity, with a bimodal distribution being indicative of an active enhancer, whereas a single peak indicates a switch to an inactive state[37,38].

A notable exception to this rule was H3K79me3. Although its presence was highly correlated with the activity of developmental enhancers (**Fig. 3b**), its distribution was much broader than those of H3K27ac and H3K4me1, and it seemed to be present on different nucleosomes located at a greater distance from the region of transcription factor occupancy (**Fig. 4b**).

In contrast to chromatin modifications, Pol II occupancy was enriched in a discrete peak centered on the region of transcription factor binding (**Fig. 4b**). When transcription factors were no longer bound to an enhancer but had been bound at a slightly earlier stage of development, Pol II was no longer present, suggesting that it is recruited to the enhancer by transcription factors (**Fig. 4d**). Thus, Pol II occupancy is tightly correlated with both the timing and location of transcription factor binding (**Fig. 4d**) and the precise timing of an enhancer's activity (**Fig. 4a**, CAD2 enhancers). Taken together, these results suggest that transcription factor occupancy facilitates Pol II recruitment, which may represent a crucial switch in the activation of some enhancer elements.

H3K4me1 constitutively marks enhancer elements

H3K4me1 was present on the vast majority of developmental enhancers in mesodermal cells (111 of 144 enhancers; **Supplementary Fig. 5**), being similarly distributed on mesodermally active (20 of 22, 91%) and inactive (14 of 21, 67%) enhancers, as well as on enhancers active in other tissues at these stages of development (24 of 31, 77%) (**Supplementary Fig. 7**). These findings indicate that the placement of this mark is not cell type specific during embryonic development, in contrast to what has been observed in tissue culture models[5,7,9]. On mesodermally inactive enhancers, its presence often coincided with the repressive H3K27me3 mark (**Fig. 3d** and **Supplementary Figs. 6f**, **7** and **8**).

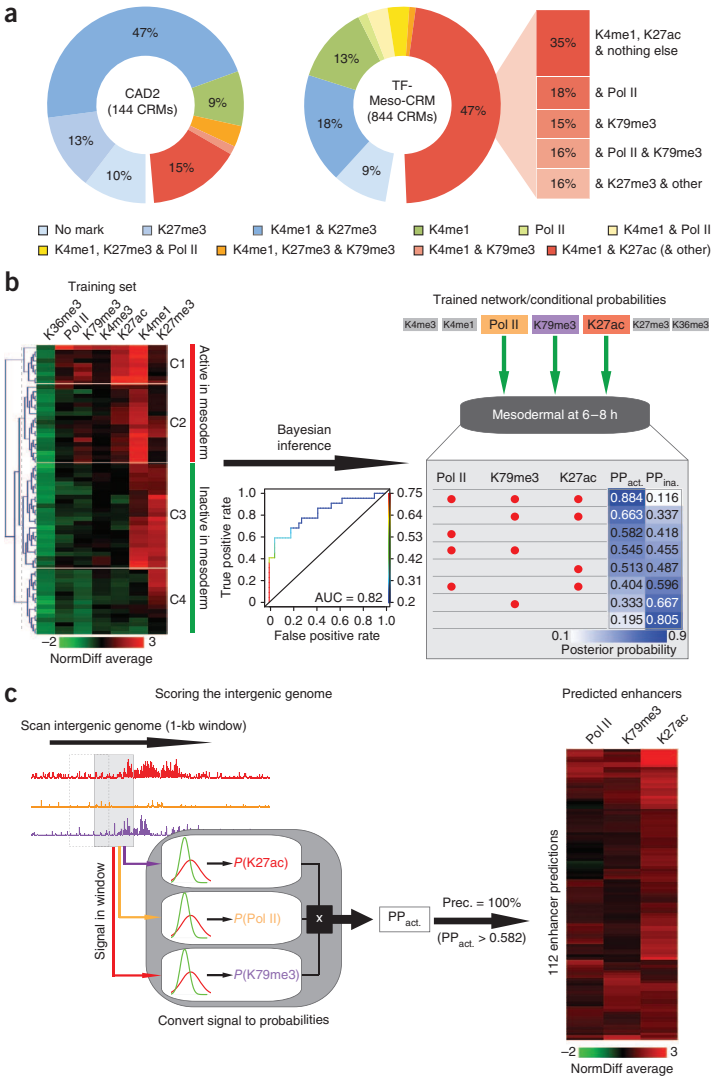


Figure 5 Modeling chromatin state on enhancers to predict enhancer activity. (**a**) Heterogeneous combinations of Pol II occupancy and chromatin marks on enhancers. Left, CAD2, containing active and inactive CRMs. Right, TF-Meso-CRMs, defined by mesodermal transcription factor binding, show higher incidence of activating marks (H3K27ac and H3K79me3) and Pol II occupancy. (**b**) Bayesian modeling of mesodermal enhancers at 6–8 h of development. Left, hierarchical clustering of ChIP-Seq signals on the training set (top row enhancer represents an unannotated promoter and was eliminated). Clusters (C1–C4) contain 9, 18, 24 and 15 enhancers: C1 and C2, active clusters (89% and 69% active mesodermal enhancers, respectively); C3 and C4, more repressed states (17% and 13% active mesodermally, respectively). Top right, Bayesian network trained to predict the activity state of developmental enhancers (dark gray box) from quantitative histone modification and Pol II levels. Green arrows indicate positive conditional dependencies. Bottom right, conditional posterior probabilities (PP) of an enhancer being active/inactive (PP_{act}/PP_{ina}) mesodermally at 6–8 h given H3K27ac, H3K79me3 and Pol II presence (red dot). Receiver operating characteristic (ROC curve; middle) shows classifier quality. (**c**) Predicting mesodermal regulatory regions active at 6–8 h *de novo*. Left, quantitative H3K27ac, H3K79me3 and Pol II levels for each intergenic 1-kb window (50-bp steps) were converted to probabilities of being present (*P*(K27ac), *P*(K79me3) and *P*(Pol II)) using learned mixture models (red and green Gaussians) and used to compute the probabilities of each window being in each of the eight possible states (bottom right in **b**. These were multiplied by the corresponding PP_{act} to compute the final PP_{act} of evaluated windows; 1-kb windows of PP_{act} 0.582 (corresponding to 100% precision, 36% recall) were merged into 112 predicted active regions and hierarchically clustered, showing heterogeneity in quantitative signals (right). Prec., precision.

Examining active chromatin marks, we observed no characterized enhancers that contained H3K79me3 or H3K27ac marks in the absence of H3K4me1 (**Fig. 3d**). This is in contrast to reported *de novo* searches for enhancers using the presence of either H3K4me1 or H3K27ac, in which these marks were found to occur separately in noncoding regions[6,7,9]. To assess this discordance between the strict co-occurrence of H3K27ac with H3K4me1 on characterized enhancers versus their separate occurrence in global searches of noncoding regions, we examined the co-occurrence of H3K27ac with H3K4me1 on the TF-Meso-CRMs[27]. Of the 844 intergenic and transcription factor–occupied CRMs at 6–8 h of development, only one (0.12%) was marked by H3K27ac in the absence of H3K4me1 (**Supplementary Fig. 8**). The strikingly low percentage suggests that, in the context of *bona fide* enhancer elements, H3K27ac rarely occurs in the absence of H3K4me1. Performing *de novo* searches for H3K4me1 or H3K27ac regions throughout the *Drosophila* genome further confirmed this observation: 96% of H3K27ac regions (covering ~17.3 Mb) overlapped with H3K4me1 (covering ~29.4 Mb) ($P = 0.001$; Online Methods). As H3K4me1 did occur in the absence of H3K27ac, our results suggest a sequential order of H3 modifications in which Lys4 is monomethylated first, and Lys27 can then be acetylated.

This tight association between these marks may have been missed in previous studies because of undersampling of the H3K4me1 signal in organisms with large genomes. Subsampling our data (**Supplementary Fig. 9** and **Supplementary Table 3**) indicated that ~26 million mapped reads are required to reach saturation of H3K4me1 peaks in the *Drosophila* genome, which is ~16 times smaller than the human genome. Although the presence of H3K4me1 *per se* did not correlate with enhancer activity, the quantitative levels of signal were higher on active versus inactive enhancers (**Supplementary Fig. 10**). Undersampling H3K4me1 would therefore tend to detect regulatory regions enriched in active enhancers while missing many repressed regions, which may explain the observed differences in the presence of this mark between different cell types[7,39] in contrast to its general presence when sampling the entire regulatory landscape (**Supplementary Fig. 7**).

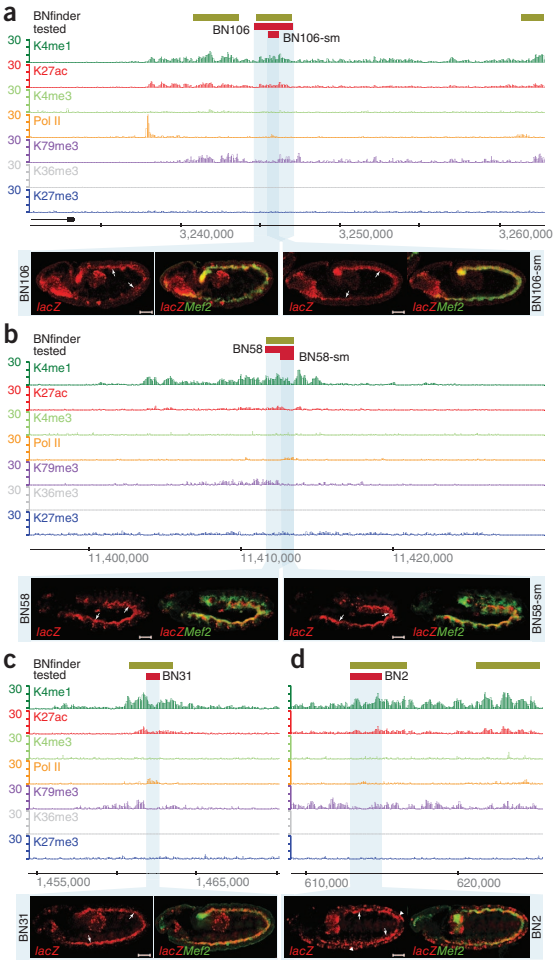


Figure 6 Predicted active regulatory regions function as enhancers *in vivo*. (**a**–**d**) Genomic regions predicted by the trained Bayesian model to direct mesodermal expression at 6–8 h of development. Top, BiTS-ChIP-Seq signal enrichment for histone modifications (RPGC, background subtracted using H3) and Pol II (RPGC, background subtracted using input) are shown. Green boxes, regions predicted by the Bayesian network (BNFinder); red boxes, the regions tested for enhancer activity *in vivo* by transgenic reporter assays. Bottom, embryos showing the expression pattern of the *lacZ* reporter gene driven by the genomic region (red), detected by double *in situ* hybridization with a mesoderm-specific marker (*Mef2*, green). Left, anterior; up, dorsal. Scale bars, 50 μm. The tested regions function as enhancers *in vivo*, regulating expression in the mesoderm (arrows) at 6–8 h, with or without expression in other tissues (arrow heads, ectodermal). Note that two smaller cloned regions centering on the location of Pol II (in **a** and **b**) are sufficient to give the same activity as the larger encompassing chromatin domains. Although assayed regions had different chromatin signatures, they functioned as enhancers at the predicted stages of development. Other regions tested are shown in **Supplementary Figure 13**.

Chromatin signatures can predict enhancer regulatory state

Having shown that individual chromatin marks and Pol II occupancy are highly correlated with active enhancers (**Fig. 3**), we asked whether combined signatures more accurately reflect enhancer activity and could therefore predict activity state *de novo*. Active and inactive enhancers contained heterogeneous combinations of chromatin marks and Pol II binding (**Fig. 5a**), where the relative level and presence of each varied between different enhancer elements (**Fig. 5b**). As it is not known *a priori* which combinations of marks and/or Pol II occupancy are important and to what degree, we moved from threshold-based correlations of single features to a probabilistic quantitative model that could directly assess which signatures were informative and in which combinations.

Recent studies have successfully addressed the challenge of predicting enhancer location in cell culture systems. Supervised learning strategies could distinguish between promoters and enhancer elements on the basis of the presence of H3K4me3 and H3K4me1, respectively[5,33]. Similarly, a Hidden Markov Model (HMM) was used to segment the genome into 15 chromatin states that correlated with known annotations for gene bodies, promoters and putative enhancer elements (their location and activity was based on DNase hypersensitive sites and the expression of the closest gene)[39].

Here, we took a contrary enhancer-centric approach, starting with a collection of well-characterized enhancers and used a model to directly learn which features distinguish active and inactive enhancers.

To obtain information on the dependency structure between chromatin marks, Pol II occupancy and enhancer activity, we employed Bayesian network inference, which has successfully identified probabilistic dependencies in many biological contexts[40–44]. A Bayesian network topology was reconstructed to discover dependencies between the quantitative signals of chromatin marks and Pol II occupancy in enhancers with two different activity states: a restricted set of enhancers active in mesoderm at 6–8 h of development and a broader set of mesodermally active enhancers (**Fig. 5b** and **Supplementary Fig. 11**). The trained Bayesian network was validated using a fourfold cross-validation scheme (see Online Methods and **Supplementary Fig. 11a**) and accurately represents both activity states (area under the receiver operating characteristic (AUC) of 0.82 and 0.76, respectively; **Supplementary Fig. 11b**), independent of enhancer distance from the transcription start site (TSS) (**Supplementary Fig. 12**). The model identified a conditional link between the presence of H3K79me3 and H3K27ac marks and enhancer activity, whereas H3K27me3 was contraindicative and H3K4me1, H3K4me3 and H3K36me3 had no predictive value for activity. Pol II presence was identified as a causal dependency for enhancers that were active specifically at 6–8 h but not for the broader group active in mesoderm at any time during development (**Supplementary Fig. 11a**), indicating that Pol II presence is predictive of the precise timing of an enhancer's activity, consistent with its very transient presence on active enhancers (**Fig. 4a**) and its relationship to transient transcription factor occupancy (**Fig. 4d**).

We applied the trained Bayesian network to the *Drosophila* intergenic genome (see Online Methods) and identified >303 kb of sequence predicted to direct mesodermal activity at 6–8 h of development, using a posterior probability threshold of ≥0.582 (corresponding to an estimated precision of 100% and recall of 36%; **Supplementary Fig. 11**). Even at this stringent threshold, the 112 predicted regions (**Supplementary Table 4**) had diverse levels of H3K79me3, H3K27ac and Pol II enrichment, with some regions containing all three to varying degrees and other regions lacking one component (**Fig. 5c**, heat map). Of the predicted regions, 78% overlapped with TF-Meso-CRMs[27] ($P = 0.001$; see Online Methods), indicating that these regions recruit mesodermal transcription factors at exactly these stages of development, suggesting that the majority of predicted regions are likely to function as active regulatory regions *in vivo*.

Predicted regions function as mesodermal enhancers *in vivo*

To assess the true accuracy of the predictions, we examined the activity of putative regulatory elements *in vivo* using transgenic reporter assays in the developing embryo. Predicted regions were cloned upstream of a minimal promoter and reporter gene, and these constructs were stably integrated into the *Drosophila* genome and assessed for spatio-temporal activity by *in situ* hybridization (see Online Methods and **Supplementary Table 5**). We selected for analysis nine regions featuring different chromatin states and/or Pol II occupancy, while other information, such as the presence of transcription factor motifs, motif conservation or transcription factor occupancy, was not considered (**Fig. 6** and **Supplementary Fig. 13**). Five tested regions had both H3K79me3 and Pol II, with varying levels of H3K27ac (**Fig. 6a**,**b** and **d** and **Supplementary Fig. 13b**,**c**), three regions had Pol II occupancy without H3K79me3 (**Fig. 6c** and **Supplementary Fig. 13a**,**e**) and one region had H3K79me3 without Pol II occupancy (**Supplementary Fig. 13d**). Despite this heterogeneity, eight of the nine regions tested were sufficient to function as mesodermal enhancers *in vivo* and, notably, directed expression at the predicted developmental stage of 6–8 h (**Supplementary Table 5**).

Chromatin marks typically span large intergenic regions, encompassing multiple known enhancers (**Fig. 2** and **Supplementary Fig. 6**), which is reflected in the large size of the regions predicted by the Bayesian networks (average size of 2.7 kb). Given that Pol II occupancy is highly correlated with the timing and location of transcription factor binding (**Fig. 4**), we reasoned that the location of Pol II occupancy might pinpoint the precise location of the functional enhancer element within larger chromatin domains. To test this hypothesis, we examined the activity of smaller regions within two larger Bayesian network–predicted domains (**Fig. 6a**,**b**). In both cases, the spatio-temporal expression driven by the smaller Pol II–bound region was largely indistinguishable from that of the entire region (**Fig. 6a**,**b**). These results indicate that Pol II, when present, is a good predictor of the precise location and temporal activity of enhancers.

We also examined the activity of four enhancers with very low posterior probability scores, indicating that they are very unlikely to have mesodermal activity at 6–8 h of development. The four regions are previously published enhancers for which expression information is lacking at this developmental stage. None of these enhancers directed observable activity in the mesoderm within 6–8 h of development (**Supplementary Fig. 13f–i**).

In summary, of the nine tested regulatory regions predicted to be active by the Bayesian networks, all but one functioned as developmental enhancers in the mesoderm at 6–8 h, showing that chromatin modifications and Pol II enrichment can serve as powerful read-out of an enhancer's activity during development.

DISCUSSION

We present a method that facilitates tissue-specific analysis of chromatin state, Pol II occupancy, general transcription factor binding, insulator recruitment and other aspects of transcriptional regulation within the context of a multicellular developing organism. The goal of BiTS-ChIP is similar to that of INTACT[45], an affinity-based method developed in plants for the acquisition of cell type–specific nuclei[45]. Although the dependency on FACS sorting may potentially be a limitation, the BiTS method has two crucial advantages: (i) it does not require *a priori* transgenesis if a good antibody for a cell type–specific nuclear protein is available; and (ii) although FACS has previously been used to sort unfixed cells or nuclei for native ChIP-Seq[46], BiTS carries the benefit that the covalent cross-linking before embryo dissociation freezes the chromatin state at the intended moment and facilitates an analysis that goes beyond nucleosomal features to include chromatin-binding proteins. Combining BiTS with new approaches to amplify ChIP signals[47,48] should facilitate cell type–specific analysis of very small populations of cells. Applying this method to the developing *Drosophila* embryo, we have identified multiple chromatin modifications associated with enhancer activity in a specific subpopulation of cells—signatures that would be largely obscured if assayed in the whole embryo.

Our results support a multistep model for enhancer activation: H3K4me1 may indicate enhancers in an intermediate state, in which they are susceptible to subsequent repression via H3K27 trimethylation or activation via H3K79 trimethylation, H3K27 acetylation and Pol II recruitment. Chromatin modifications typically cover relatively large regions of >2 kb that encompass multiple enhancer elements. The deposition of H3K79me3 or H3K27ac may therefore place an entire regulatory region in a permissive state, and the activity of individual enhancer elements contained within these regions appear to be determined by the timing of transcription factor occupancy, nucleosome remodeling and Pol II association.

We previously reported that transcription factor occupancy alone is sufficient to predict spatio-temporal enhancer activity[27]; here, we show that histone modifications and Pol II occupancy information alone can accurately predict the activity status of regulatory regions. Integrating chromatin modification and transcription factor occupancy data within the same cell type and at the same stage of development will provide a very accurate way to distinguish functional transcription factor binding events from nonfunctional occupancy and should facilitate better modeling of tissue-specific gene expression and the underlying *cis*-regulatory networks during development. Considering the high resolution, precision and sensitivity of the data afforded by BiTS-ChIP, this method provides a powerful approach to decipher transcriptional networks and should be widely applicable to other species and complex tissues.

URLs. Study data, http://furlonglab.embl.de/data/download; and http://www.ebi.ac.uk/ena/data/view/ERP000560; BNFinder, http://bioputer.mimuw.edu.pl/software/bnf/.

METHODS

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

Accession numbers. Raw sequences and mapping to the *D. melanogaster* reference genome dm3 (BAM files) of mesoderm-specific BiTS-ChIP-Seq data (for Mef2, H3, H3K4me3, H3K4me1, H3K27ac, H3K27me3, H3K36me3, H3K79me3 and Rpb3-Pol II) and whole-embryo Mef2 ChIP-Seq and input-Seq data generated in this study are accessible at the Sequence Read Archive (SRA) under the study accession number ERP000560. Two biological replicates at 6–8 h after egg laying (AEL) were generated for each condition, with the exception of Mef2 BiTS-ChIP-Seq. Processed data are available at the Furlong laboratory webpage (see URLs).

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

This work was technically supported by the EMBL Genomics Core facility for Solexa sequencing and by the IT service unit. We thank S. Müller for performing *Drosophila* injections and all members of the Furlong laboratory for discussions and comments on the manuscript. We thank J. Lis (Cornell University) for the Rpb3 (Pol II) antibody. This work was supported by grants to E.E.M.F. from ERASysBio (Mod Heart) and the Human Frontiers Science Program Organization

and by a long-term fellowship to R.P.Z. from the International Human Frontiers Science Program Organization. S.B. was funded by the EMBL Interdisciplinary Postdoc (EIPOD) programme.

AUTHOR CONTRIBUTIONS

S.B., R.P.Z. and E.E.M.F. designed the study. S.B., R.P.Z., E.H.G., Y.G.-H., A.P.-G. and A.R. conducted experiments. S.B., R.P.Z., A.P.-G. and A.R. performed FACS sorting. S.B. and R.P.Z. performed the BiTS-ChIP-Seq. C.G., S.B. and N.D. performed computational analyses. B.W. helped with the Bayesian modeling. S.B., C.G., R.P.Z. and E.E.M.F. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Kolasinska-Zwierz, P. *et al.* Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat. Genet.* **41**, 376–381 (2009).
2. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
3. Karlić, R., Chung, H.R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. USA* **107**, 2926–2931 (2010).
4. Mikkelsen, T.S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
5. Heintzman, N.D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
6. Creyghton, M.P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. USA* **107**, 21931–21936 (2010).
7. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
8. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.* **40**, 897–903 (2008).
9. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell-type–specific gene expression. *Nature* **459**, 108–112 (2009).
10. Riddle, N.C. *et al.* Plasticity in patterns of histone modifications and chromosomal proteins in *Drosophila* heterochromatin. *Genome Res.* **21**, 147–163 (2011).
11. Zhou, V.W., Goren, A. & Bernstein, B.E. Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.* **12**, 7–18 (2011).
12. Gaudet, J. & Mango, S.E. Regulation of organogenesis by the *Caenorhabditis elegans* FoxA protein PHA-4. *Science* **295**, 821–825 (2002).
13. Cao, Y. *et al.* Global and gene-specific analyses show distinct roles for Myod and Myog at a common set of promoters. *EMBO J.* **25**, 502–511 (2006).
14. Jakobsen, J.S. *et al.* Temporal ChIP-on-chip reveals Biniou as a universal regulator of the visceral muscle transcriptional network. *Genes Dev.* **21**, 2448–2460 (2007).
15. Wilczyński, B. & Furlong, E.E. Dynamic CRM occupancy reflects a temporal map of developmental progression. *Mol. Syst. Biol.* **6**, 383 (2010).
16. Soshnikova, N. & Duboule, D. Epigenetic temporal control of mouse Hox genes *in vivo*. *Science* **324**, 1320–1323 (2009).
17. Akkers, R.C. *et al.* A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in *Xenopus* embryos. *Dev. Cell* **17**, 425–434 (2009).
18. Liu, T. *et al.* Broad chromosomal domains of histone modification patterns in *C. elegans*. *Genome Res.* **21**, 227–236 (2011).
19. modENCODE Consortium *et al.* Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**, 1787–1797 (2010).
20. Wardle, F.C. *et al.* Zebrafish promoter microarrays identify actively transcribed embryonic genes. *Genome Biol.* **7**, R71 (2006).
21. Bulger, M. & Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327–339 (2011).
22. Sandmann, T., Jakobsen, J.S. & Furlong, E.E. ChIP-on-chip protocol for genome-wide analysis of transcription factor binding in *Drosophila melanogaster* embryos. *Nat. Protoc.* **1**, 2839–2855 (2006).
23. Cunha, P.M. *et al.* Combinatorial binding leads to diverse regulatory responses: Lmd is a tissue-specific modulator of Mef2 activity. *PLoS Genet.* **6**, e1001014 (2010).
24. Liu, Y.H. *et al.* A systematic analysis of Tinman function reveals Eya and JAK-STAT signaling as essential regulators of muscle development. *Dev. Cell* **16**, 280–291 (2009).
25. Sandmann, T. *et al.* A core transcriptional network for early mesoderm development in *Drosophila melanogaster*. *Genes Dev.* **21**, 436–449 (2007).
26. Sandmann, T. *et al.* A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Dev. Cell* **10**, 797–807 (2006).
27. 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).
28. Filion, G.J. *et al.* Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell* **143**, 212–224 (2010).
29. Black, B.L. & Olson, E.N. Transcriptional control of muscle development by myocyte enhancer factor-2 (MEF2) proteins. *Annu. Rev. Cell Dev. Biol.* **14**, 167–196 (1998).
30. Tomancak, P. *et al.* Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* **3**, RESEARCH0088 (2002).
31. Zeitlinger, J. *et al.* RNA polymerase stalling at developmental control genes in the *Drosophila melanogaster* embryo. *Nat. Genet.* **39**, 1512–1516 (2007).
32. Nègre, N. *et al.* A *cis*-regulatory map of the *Drosophila* genome. *Nature* **471**, 527–531 (2011).
33. De Santa, F. *et al.* A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol.* **8**, e1000384 (2010).
34. Kim, T.K. *et al.* Widespread transcription at neuronal activity–regulated enhancers. *Nature* **465**, 182–187 (2010).
35. Koch, F. *et al.* Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat. Struct. Mol. Biol.* **18**, 956–963 (2011).
36. Cui, K. *et al.* Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell* **4**, 80–93 (2009).
37. He, H.H. *et al.* Nucleosome dynamics define transcriptional enhancers. *Nat. Genet.* **42**, 343–347 (2010).
38. Hoffman, B.G. *et al.* Locus co-occupancy, nucleosome positioning, and H3K4me1 regulate the functionality of FOXA2-, HNF4A-, and PDX1-bound loci in islets and liver. *Genome Res.* **20**, 1037–1051 (2010).
39. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
40. Beer, M.A. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117**, 185–198 (2004).
41. Friedman, N. Inferring cellular networks using probabilistic graphical models. *Science* **303**, 799–805 (2004).
42. Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449–453 (2003).
43. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A. & Nolan, G.P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 523–529 (2005).
44. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176 (2003).
45. Deal, R.B. & Henikoff, S. A simple method for gene expression and chromatin profiling of individual cell types within a tissue. *Dev. Cell* **18**, 1030–1040 (2010).
46. Cheung, I. *et al.* Developmental regulation and individual differences of neuronal H3K4me3 epigenomes in the prefrontal cortex. *Proc. Natl. Acad. Sci. USA* **107**, 8824–8829 (2010).
47. Shankaranarayanan, P. *et al.* Single-tube linear DNA amplification (LinDA) for robust ChIP-seq. *Nat. Methods* **8**, 565–567 (2011).
48. Adli, M., Zhu, J. & Bernstein, B.E. Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. *Nat. Methods* **7**, 615–618 (2010).

ONLINE METHODS

***Drosophila* lines, staining and imaging.** The *twi*^{PEMK}*::SBP-His2B* *Drosophila* strain was constructed by injecting *w*^{1118}, using standard procedures, with a P-element transformation vector bearing a gene encoding *D. melanogaster* histone H2B tagged N-terminally with two copies of the strepdavidin-binding peptide (SBP)[49] separated from the H2B protein by three TEV protease cleavage sites. Expression of the transgene is directed by four copies of a compound CRM consisting of the *twi* proximal element (PE) enhancer (chr. 2R: 18,933,349–18,933,739, dm3)[50] and the *twi* 3′ MK enhancer (chr. 2R: 18,937,023–18,937,922) using a minimal *eve* promoter. Several independent homozygous P-element insertion lines were assayed for transgene expression using an antibody to the SBP tag (sc-101595, Santa Cruz Biotechnology) and a secondary Alexa555-conjugated donkey anti-mouse antibody (A-21127, Invitrogen,). Expression was initially detected in the presumptive mesoderm at late stage 5 and remained detectable in mesodermally derived tissues past stage 16.

Endogenous gene expression patterns were detected by fluorescent *in situ* hybridization (FISH)[51] using an antisense digoxygenin (DIG)-labeled RNA probe directed against the gene of interest. Embryos were counterstained for *Mef2* RNA with a fluorescein-labeled probe (mesoderm reference). Probes were made from ESTs obtained from the BDGP EST collections (*vvl*, RE27192; *Ama*, LD39923; *CG9650*, LD11946; *Him*, RE70039; *Act57B*, LD04994; *lmd*, LD47926) or from PCR products (primer information is provided in **Supplementary Table 6**) cloned into the pCRII Dual Promoter vector (Invitrogen).

Transgenic lines to assay enhancer activity were constructed as described[27] using the indicated primers (**Supplementary Table 6**). PCR-amplified fragments were cloned into the pDuo2n-attB site-specific transformation vector and injected into the VK33 landing site strain[52]. CRM activity was assayed by *in situ* hybridization (ISH)[51] for the *GFP* (CG32150) or *LacZ* (all others) reporter genes using DIG-labeled antisense RNA probes, and embryos were counterstained for *Mef2* expression as above. Additionally, we tested previously published enhancers for which no activity information at 6–8 h of development was available (*aopJ*[53], *run_neur_6kb*[54], *gt-10* (ref. 55), *CG32150_PE*[56] and *salm_IRU22* (ref. 57)). Transgenic *Drosophila* lines (*Ser_II-1.3*, *gt-10*, *CG32150_PE* and *salm_IRU22*) or P-element transformation vector (*run_neur_6kb*) were kindly supplied by the fly community[53–57]. Of the five published enhancers, only *aopJ* overlapped with a Bayesian Network Finder (BNFinder)-identified region and was re-cloned, as neither strains nor vector were available.

Images were acquired on a Leica SP2 or Zeiss LSM510Meta confocal microscope and were processed using ImageJ software.

Preparation of nuclear extracts and FACS sorting. Collections of *twi*^{PEMK}*::SBP-His2B* embryos at the 6–8 h stage of development were collected and fixed as described[22]. All subsequent steps were carried out at 4 °C. For the preparation of nuclei, 1 g of snap-frozen embryos was transferred to a 15-ml Wheaton Scientific Homogenizer and thawed for 5 min in 10 ml of homogenization buffer (HB) (15 mM Tris, pH 7.4, 0.34 M sucrose, 15 mM NaCl, 60 mM KCl, 0.2 mM EDTA, 0.2 mM EGTA and Roche Complete protease inhibitors). After douncing the embryos 20 times with a loose pestle and 10 times with a tight pestle, we filtered the resultant suspension through two layers of Miracloth (Calbiochem) into a 15-ml conical vial and centrifuged at 3,500*g* for 5 min. The supernatant was carefully decanted, and the nuclear pellet was washed in 10 ml of HB and centrifuged again. The nuclear pellet was resuspended in 3 ml of chilled PBT buffer (0.1% Triton X-100 in PBS) and dissociated by passage through a 20-gauge and then a 22-gauge needle, ten times each, using a 5-ml syringe, and samples were then filtered through a 20-µm Nitex membrane to clear debris and nuclear aggregates. The dissociated nuclei were stained with antibody to the SBP tag (1:100) for 1 h, washed with 6 ml of PBT for 10 min, resuspended in 3 ml of PBT and incubated for 1 h in the dark with Alexa488-conjugated donkey anti-mouse secondary antibody (1:100; A-21202, Invitrogen). Nuclei were centrifuged for 1 min at 1,000*g* and were resuspended in 3 ml of PBTB (5% BSA and 0.2% NP-40 in PBT). The suspension was then divided into 300-µl aliquots in 5-ml tubes (352063, Falcon) and brought to 3 ml by adding PBTB. Immediately before sorting, the nuclei were redissociated by passing the samples through a 22-gauge needle ten times and then filtering them through a cell strainer (322350, Falcon) to avoid clogging of the FACS machine. Nuclear samples were run on a Beckman Coulter MoFlo cell sorter using Summit software version 4.3 (detailed FACS sorting conditions are provided in the **Supplementary Note**).

Chromatin immunoprecipitation and Solexa sequencing. Immediately after sorting, the nuclei were centrifuged at 3,500*g* for 10 min, and the pellet was resuspended in 300 µl of RIPA buffer (10 mM Tris-HCl, pH 8.0, 1 mM EDTA, 1% Triton X-100, 140 mM NaCl, 0.1% SDS, 0.1% sodium deoxycholate and Roche Complete protease inhibitors) and transferred into a 1.5-ml tube. After incubation for 10 min on ice, the chromatin was sheared into 200-bp fragments using a Diagenode BioRuptor (18 cycles, high intensity, 30 s on/30 s off, 4 °C). The chromatin was then centrifuged for 2 min at 18,000*g*, and the supernatant was transferred to a low-binding tube (710176, Biozym Scientific). The majority of sheared chromatin was subsequently snap-frozen in liquid nitrogen, but a small aliquot was used to measure DNA concentration and fragment size[22].

Chromatin was prepared as described previously[22], with small modifications (**Supplementary Note**) and used for immunoprecipitation, after optimizing the conditions for each antibody by real-time PCR (**Supplementary Note** and **Supplementary Table 6**). Antibodies detecting H3 (ab1791), H3K4me3 (ab71998), H3K4me1 (ab8895), H3K27ac (ab4729), H3K36me3 (ab9050) and H3K79me3 (ab2621) were purchased from Abcam and antibody to H3K27me3 was from Active Motif (39155). The Rpb3 (RNA Pol II) antibody was a generous gift from John Lis[58], and the antibody to Mef2 was generated in the Furlong laboratory[26].

Solexa libraries were prepared according to the manufacturers' recommendations, with small modifications (**Supplementary Note**). Library quality was assessed on a 2100 Bioanalyzer system (Agilent). Two biological replicates of every mark were single-end sequenced with 36-bp reads using an Illumina Genome Analyzer IIX.

ChIP-Seq data processing. Quality assurance was performed using the Bioconductor package ShortRead[59,60] (**Supplementary Table 3**). Reads of 36 bp were aligned against the dm3 genome (obtained from FlyBase[61]) using bowtie[62] (reads aligning to more than one locus were discarded). Peaks were called with MACS (v1.3.7.1)[63], where for histone marks, H3 was used as control data and, for Pol II and Mef2, input was used. Saturation in peak calling (**Supplementary Fig. 9**) was assessed by calling peaks (using MACS) on increasing amounts of duplicate-filtered data (from 10–100%, in increments of 10%). Finally, samples were corrected to the mappable genome size (135 Mb), generating RPGC scores that were summarized (by median value) into adjacent non-overlapping bins of a defined size (25 or 50 bp). Two described approaches (NormDiff and Background Subtracted)[64] were used to independently perform the background correction (using H3 data as the background model for histone modifications and input otherwise). For details, see the **Supplementary Note**.

Comparison of Mef2 peaks from sorted (BiTS-ChIP) versus unsorted nuclei. ChIP-chip Mef2 peaks were obtained from previously reported results[27] and were remapped to dm3 using the UCSC LiftOver tool[65]; regions of 400 bp centered on the ChIP-chip peak were further considered. Mef2 peaks from BiTS-ChIP-Seq and ChIP-Seq were called using MACS and compared by overlap (by at least one base).

Gene lists using the BDGP ISH database. Gene lists (**Supplementary Table 7**) used in the analysis shown in **Figure 2c–f** and **Supplementary Figures 3** and **4** were assembled using the BDGP ISH database[66] (downloaded in July 2010) and the ontological term mapping in **Supplementary Table 8**. The 'active 6–8 h' list contains 572 genes expressed ubiquitously and in the mesoderm at 6–8 h of development; the 'meso. 6–8 h' list contains 267 genes expressed in mesoderm but not ubiquitously at 6–8 h; the 'only meso. 6–8 h' list contains 38 genes expressed exclusively in the mesoderm at 6–8 h; the 'no meso. 6–8 h' list contains 275 genes expressed only outside of the mesoderm at 6–8 h; and the 'inactive 6–8 h' list contains 78 genes inactive at 6–8 h but active later in development. To avoid confounding overlapping signatures from multiple TSSs, we selected genes that contain a single annotated TSS and are further than 1 kb from another TSS. Of note, as the data shown in **Figure 2c–f** and **Supplementary Figures 3** and **4** were summarized using a trimmed mean and filtered for genes larger than 850 bp, the total number of genes plotted was

smaller (427 'active 6–8 h', 201 'meso. 6–8 h', 30 'only meso. 6–8 h', 209 'no meso. 6–8 h' and 56 'inactive 6–8h'). See the **Supplementary Note** for details.

CAD2 CRMs and TF-Meso-CRMs. CAD2 (**Supplementary Table 1**) is based on CAD[27] but was updated to include new *Drosophila* enhancers reported since 2009 in REDfly[67] and elsewhere (**Supplementary Table 1**). Entries were filtered for size (≤2 kb) and remapped to dm3 using the UCSC LiftOver tool[65]. Stage-specific annotation encompassed enhancer activity from embryonic stage 5 or earlier, stages 6–14, individually, and stage15 or later; annotations also distinguish between expression in the mesoderm (and its derivatives) and/or expression in non-mesodermal tissues. In all analyses, CAD2 enhancers residing within 1 kb of gene annotations (both 5′ and 3′ of genes and within gene bodies) or that overlapped by at least one base with an H3K4me3 peak (as defined by MACS) were ignored to avoid confounding chromatin signatures coming from active promoters and gene transcription. After filtering, CAD2 contained 144 intergenic enhancers. Stage-specific annotations were used to define groups of enhancers that are active in the mesoderm at 6–8 h of development (class A, 22 enhancers), strictly inactive in the mesoderm at 6–8 h (class I, 21 enhancers), active exclusively outside of the mesoderm at 6–8 h (class O, 31 enhancers) and are active in the mesoderm before 6 h (class E, 39 enhancers) or after 8 h (class L, 7 enhancers).

The Mesodermal ChIP-CRM atlas (TF-Meso-CRMs) was obtained from a previous report[27] and was remapped to dm3 using the UCSC LiftOver tool[65]. The Mesodermal ChIP-CRM atlas was filtered for gene and H3K4me3 proximity (as above), and high-confidence TF-Meso-CRMs (bound by at least two transcription factors at the same developmental stage) were further used to define temporal groups on the basis of Mef2, Twi, Tin, Bin and Bap binding information: CRMs bound at 6–8 h (6–8 h class, 297 CRMs), CRMs bound only after 8 h (8–12 h class, 10 CRMs) and CRMs bound only before 6 h (2–6 h class, 88 CRMs). See the **Supplementary Note** for details.

Enrichment, precision and recall of enhancer activity. Enrichment of active enhancers (**Fig. 3b,c**) is defined as the fraction of active enhancers containing a specific modification compared to the fraction of active enhancers in the full enhancer dataset. The significance of enrichment was calculated using a two-sided Fisher's exact test. Precision and recall (**Fig. 3e**) were calculated according to their usual definitions. See the **Supplementary Note** for details.

Gene and CRM intensity profiles. Gene-centered intensity profiles (**Fig. 2c–f** and **Supplementary Figs. 3** and **4**) were computed using background-subtracted counts per 25-bp bins from 500 bp upstream of the TSS to 500 bp downstream of the gene end and show the smoothed (5 bins) trimmed signal mean (10–90%). Genes smaller than 850 bp were not considered. To account for variable gene sizes, signal between 300 bp downstream of the TSS to 300 bp upstream of the gene end was represented by 100 values obtained by cubic spline interpolation. Similarly, TF-Meso-CRM–centered intensity profiles (**Fig. 4b–f**) show the smoothed trimmed mean of background-subtracted counts per 25-bp bin from 1.6 kb downstream to 1.6 kb upstream of the CRM center. To obtain scaled values between 1 and 0 (**Fig. 4b**), the values were shifted by the difference between the minimum value and 0 and then divided by the maximum intensity for a given mark within the region 1.6 kb on either side of the TF-Meso-CRM. See the **Supplementary Note** for details.

Significance of overlap between two region sets. The significance of overlap between H3K27ac and H3K4me1 and BNFinder enhancer predictions and TF-Meso-CRMs (and analysis in **Supplementary Fig. 5**) was estimated by bootstrap (using 999 random sequence sets) following previously described recommendations[68]. The estimated *P* value was determined by ranking the observed overlap percentage within a set of randomly obtained overlap percentages (**Supplementary Note**).

Clustering and Bayesian modeling. NormDiff intensity values (using 50-bp bins) were summarized into a unique intensity value for each CAD2 enhancer using a moving-average approach. The maximum observed average (200-bp or 1-kb windows for single-bin step size) was used as the summarized enhancer intensity. Hierarchical clustering (**Fig. 5b,c** and **Supplementary Table 9**) was performed using MeV[69] and the NormDiff-summarized intensities. BNFinder (version 3.3)[70] was used to train a Bayesian network to understand the relationship between H3 histone modifications, Pol II occupancy and enhancer activity state. Summarized intensities (using a 200-bp window for Pol II and H3K4me3, 1 kb otherwise) of 65 CAD2 enhancers (for which activity at 6–8 h of development is known; **Supplementary Table 10**) were used to model two different activity states: expression in mesoderm (no time constraint) and expression in mesoderm at 6–8 h. The accuracy of the trained Bayesian network was assessed using a fourfold cross-validation scheme. Finally, each 1-kb window (by steps of 50-bp) of the intergenic genome was scored using the trained Bayesian network, and windows with a posterior probability of being active in the mesoderm at 6–8 h of >0.582 were selected and merged for overlap. These Bayesian network predictions were afterwards filtered for a minimum NormDiff level of H3K4me1 of 0.5 and a maximum NormDiff level of Pol II of 2. Candidate regulatory regions were selected from BNFinder predictions covering the range of posterior probabilities above the cutoff (0.582) to the maximum (0.884). See the **Supplementary Note** for details.

49. Keefe, A.D., Wilson, D.S., Seelig, B. & Szostak, J.W. One-step purification of recombinant proteins using a nanomolar-affinity streptavidin-binding peptide, the SBP-Tag. *Protein Expr. Purif.* **23**, 440–446 (2001).
50. Jiang, J., Kosman, D., Ip, Y.T. & Levine, M. The dorsal morphogen gradient regulates the mesoderm determinant *twist* in early *Drosophila* embryos. *Genes Dev.* **5**, 1881–1891 (1991).
51. Kosman, D. *et al.* Multiplex detection of RNA expression in *Drosophila* embryos. *Science* **305**, 846 (2004).
52. Venken, K.J., He, Y., Hoskins, R.A. & Bellen, H.J. P[acman]: a BAC transgenic platform for targeted insertion of large DNA fragments in *D. melanogaster*. *Science* **314**, 1747–1751 (2006).
53. Ramos, E., Price, M., Rohrbaugh, M. & Lai, Z.C. Identifying functional *cis*-acting regulatory modules of the yan gene in *Drosophila melanogaster*. *Dev. Genes Evol.* **213**, 83–89 (2003).
54. Klingler, M., Soong, J., Butler, B. & Gergen, J.P. Disperse versus compact elements for the regulation of runt stripes in *Drosophila*. *Dev. Biol.* **177**, 73–84 (1996).
55. Schroeder, M.D. *et al.* Transcriptional control in the segmentation gene network of *Drosophila*. *PLoS Biol.* **2**, E271 (2004).
56. Reeves, N. & Posakony, J.W. Genetic programs activated by proneural proteins in the developing *Drosophila* PNS. *Dev. Cell* **8**, 413–425 (2005).
57. Barrio, R., de Celis, J.F., Bolshakov, S. & Kafatos, F.C. Identification of regulatory regions driving the expression of the *Drosophila* spalt complex at different developmental stages. *Dev. Biol.* **215**, 33–47 (1999).
58. Adelman, K. *et al.* Efficient release from promoter-proximal stall sites requires transcript cleavage factor TFIIS. *Mol. Cell* **17**, 103–112 (2005).
59. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
60. Morgan, M. *et al.* ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* **25**, 2607–2608 (2009).
61. Tweedie, S. *et al.* FlyBase: enhancing *Drosophila* Gene Ontology annotations. *Nucleic Acids Res.* **37**, D555–D559 (2009).
62. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
63. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
64. Nix, D.A., Courdy, S.J. & Boucher, K.M. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics* **9**, 523 (2008).
65. Kuhn, R.M. *et al.* The UCSC genome browser database: update 2007. *Nucleic Acids Res.* **35**, D668–D673 (2007).
66. Tomancak, P. *et al.* Global analysis of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* **8**, R145 (2007).
67. Gallo, S.M. *et al.* REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in *Drosophila*. *Nucleic Acids Res.* **39**, D118–D123 (2011).
68. Phipson, B. & Smyth, G.K. Permutation *P*-values should never be zero: calculating exact *P*-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.* **9**, Article39 (2010).
69. Saeed, A.I. *et al.* TM4 microarray software suite. *Methods Enzymol.* **411**, 134–193 (2006).
70. Wilczyński, B. & Dojer, N. BNFinder: exact and efficient method for learning Bayesian networks. *Bioinformatics* **25**, 286–287 (2009).

3.1.4 Discussion

Enhancers active in the mesoderm at 6-8h are enriched for H3K27Ac, H3K79me3 and Pol II (Figure 3b of the enclosed *Nature Genetics* paper, page 87), and the presence of these marks tightly correlates with the timing of enhancer activity (Figure 4a of the enclosed *Nature Genetics* paper, page 88). Conversely, although H3K4me1 presence constitutively marks enhancers, it is not indicative of enhancer activity (Figure 3b of the enclosed *Nature Genetics* paper, page 87). Technically, these conclusions were reached by analysing the overlapping of peaks defined by MACS with enhancers encompassed in CAD2 databases. While popular and simple, this approach distinguishes enhancers into those marked versus not marked by a particular histone modification and might thus be sensitive to the peak calling thresholds employed.

A different approach was necessary to predict active enhancers *de novo* for several reasons. First, it is unclear what rule should be implemented to define new enhancers using MACS defined peaks. Should enrichment for all three features (H3K27Ac, H3K79me3 and Pol II) be required, or should specific sub-combinations with potentially more complex logical rules (e.g., like H3K27Ac AND (H3K79me3 OR Pol II)) be considered? And what minimum overlapping percentage should be required between discrete peaks? Second, even if such subjective rules could be defined, enhancer predictions would not be scored or even ranked since a given region would merely fulfil the rule or not (and is subsequently kept as an enhancer prediction or not). Third, the level of Pol II observed on active enhancers is generally too low to be confidently detected by MACS. However a clear, Pol II signal increase where TFs bind (Figure 4b,d of the enclosed *Nature Genetics* paper, page 88) at active enhancers suggests that the presence of Pol II at active enhancers is much more common than evaluated by peak calling (36%, Figure 4a of the enclosed *Nature Genetics* paper, page 88). Finally, signal levels for different marks observed at enhancers vary significantly (see hierarchical clustering in Figure 5b of the enclosed *Nature Genetics* paper, page 89). To fully exploit this important information, a quantitative approach is required, in order to assess which signatures are informative, to what degree, and in which combinations.

We chose Bayesian network (BN) inference, a popular framework to model complex probabilistic dependencies between variables. The model is represented as a graph with nodes

denoting variables and edges conditional dependencies of probability distributions between them (as presented in Supplementary Figure 11a of the enclosed *Nature Genetics* paper, annexe 2). Provided that the graph is acyclic, it is possible to find the optimal dependency structure effectively[172]. In our context, the model consists of two types of variables: (1) quantitative levels of observed enrichment of histone modifications and Pol II occupancy, used in input, and (2) binary enhancer activity classification variables, used as output. Since we are only interested in recovering connections linking "input" variables to "output" variables, only acyclic graphs can be generated thereby allowing us to use the efficient algorithm as implemented in the BNFinder package[173].

BNs offer the advantage of potentially uncovering non-additive interactions compared to simpler linear models (for example logistic regression). This is exactly what we see in our case as shown in the table of conditional probabilities (Figure 5b of the enclosed *Nature Genetics* paper, page 89). This table presents the posterior probabilities of the eight theoretical combinations that can be made using the three significant associations learned by BNFinder to model the "active in the mesoderm at 6-8h" output. The posterior probability of a region to be an enhancer active in the mesoderm at 6-8h is lower if Pol II is found with only H3K27Ac (0.404) or H3K79me3 (0.545), compared to when Pol II is found either alone (0.582) or with both H3K27Ac and H3K79me3 (0.884).

More precisely, at the BN training stage, each variable (quantitative histone mark and Pol II enrichment) is assumed to have two possible states: "high" and "low", each of which gives rise to experimental observations following a specific Normal distribution. Any subsequent observation (a numerical value representing, e.g. H3K27Ac occupancy) can then be converted to the probability of this observation coming from "high" signal based on the estimated mixture model for this variable (H3K27Ac in our example). For simplicity, we refer to these "high" and "low" states as the "present" and "absent" states, which is acceptable in the current situation but might be misleading in situations where the "low" states still have significant levels of signal (as exemplified below with the H3K4me1). If the fitting of the mixture model (for a particular input variable) results in two well-separated Normal distributions that naturally split active from inactive enhancers (as exemplified by the green and red Gaussians in Figure 5c of the enclosed *Nature Genetics* paper, page 89), an edge representing this conditional dependency can be drawn. On the contrary, non-informative inputs result in highly superimposed Gaussians or well-separated Gaussians that

do not reflect enhancer activity states. Given these mixture models, the BNFinder algorithm scans all possible networks and selects those maximising the posterior probability of the network given the data. In our situation, the BNFinder learned 3 conditional dependencies (between chromatin marks/Pol II occupancy at 6-8h and the state of being an enhancer active in the mesoderm at 6-8h) that translate into the 8 hidden states presented in Figure 5b (of the enclosed *Nature Genetics* paper, page 89). During the prediction phase, for each inspected 1 kb window evaluated, the signal of H3K79me3, H3K27Ac and Pol II found in the window are converted into the probability of H3K79me3, H3K27Ac and Pol II to be 'present' (using the learned mixture models). Using these three probabilities, the Bayesian model then computes the probabilities of the inspected window to be in each of these 8 hidden states (i.e. the eight possible discrete combinations of using H3K79me3, H3K27Ac and Pol II presented as rows in the probability table). Finally, these 8 probabilities are combined with the posterior probability of each state (to be active) into the final posterior probability, which reflects the probability of the inspected window to be an active enhancer in the mesoderm at 6-8h.

Based on an analysis of overlaps with MACS regions, we conclude that H3K4me1 presence is not indicative of enhancer activity and constitutively marks enhancers (Figure 3b of the enclosed *Nature Genetics* paper, page 87). We were surprised to then find that the level of H3K4me1 signal on active and inactive enhancers is significantly different (supplementary Figure 10 of the enclosed *Nature Genetics* paper, annexe 2), however the trained BN did not report predictive associations involving H3K4me1. Two different reasons could explain this apparent contradiction. First, as explained above, the mixture model might not naturally follow the active/inactive split that we implemented in the analysis shown in supplementary Figure 10 (of the enclosed *Nature Genetics* paper, annexe 2). Second, the BNFinder did not report this association because it does not increase the maximum likelihood criterion. Indeed, when different options are available, BNFinder selects the minimum number of edges, which might in some cases result in hiding associations of equivalent importance. Here, the correct explanation is likely to be the first one. Indeed, two populations of inactive enhancers co-exist with respect to H3K4me1 signal level, a low and a high signal class, while active enhancers only fall in the high-level signal class. In other words, this means that the "high" (or "active") Normal distribution of the learned mixture model for H3K4me1 encompasses both active and inactive enhancers. H3K4me1 level is therefore not a discriminative feature,

in particular when its mere presence/absence is evaluated (i.e., using a threshold-based peak calling approach), but it remains unclear why some inactive enhancers have high levels of H3K4me1. A potential explanation lies in the dynamics of the mark as shown in Figure 4e (of the enclosed *Nature Genetics* paper, page 88)). Enhancers that were bound by mesodermal TFs before 6-8h (but are no longer bound at 6-8h anymore) show a nucleosomal repositioning that exhibits a significant signal increase for H3K4me1 centred on the TF binding location. How long this high H3K4me1 signal might last is unknown, but H3K4me1 level at inactive enhancers could be proportional to the time elapsed between activity and observation time. This difference of H3K4me1 level at inactive enhancers might also reflect a chromatin state difference between actively repressed enhancers (e.g. repressed by Polycomb complex and thus marked by H3K27me3) and passively inactive enhancers (e.g. devoid of H3K27me3).

As mentioned above, when two associations impact the network equally, BNFinder keeps only one of them. To avoid missing important associations, one can learn new BN(s) upon omitting one or more input variables (one by one, the most important variables first) to check whether new dependencies emerge. We performed this procedure and, surprisingly, a BN learned in absence of H3K27Ac and Pol II reported a significant correlation between H3K36me3 signal and mesodermal activity. This BN, in which the state "active in the mesoderm at 6-8h" is now conditioned only by H3K79me3 and H3K36me3 occupancy (two marks associated with Pol II elongation), still cross-validates, but shows somewhat degraded predictive performance (maximum posterior probability of 0.843, AUC of 0.77 and false positive predictions appearing at 25% of true positive rate). Consistently, the H3K36me3 level on active/inactive enhancers is significantly different (supplementary Figure 10 of the enclosed *Nature Genetics* paper, annexe 2)). If the absence of exonic sequence explains the net depletion of H3K36me3 signal compared to H3 density, the increase in H3K36me3 signal might indicate the presence of an H3K36 methyltransferase in the transcription machinery assembled at active enhancers, resulting in very low, but consistent H3K36 trimethylation. It is noteworthy that no association was uncovered for H3K4me1 (supporting the conclusions presented in the previous paragraph that this histone PTM is not indicative of activity state), neither for H3K4me3, which nevertheless does exhibit statistically significant differences in enrichment levels on active vs. inactive enhancers, according to the test presented in supplementary Figure 10 (of the enclosed *Nature Genetics* paper, annexe 2). Although the

H3K4me3 signal levels at active enhancers is not comparable with that reported in the Pekowska study[79], this slight trimethylation of H3K4 might be a consequence of Pol II recruitment. Here again, further investigations will be required to confirm these observations.

It is difficult to compare our results with concurrent studies, as thorough statistical assessment is often lacking (see section 1.2.2.3). Furthermore, comparisons across different biological systems are delicate. Of note, beyond enhancer activity (i.e. if an enhancer is on or off), we predict activity at a given stage of development, which is assessed in transgenic animals, something that is not possible in the case of cell-culture based systems used in most other studies. However, based on published data, numbers speak in our favour (Table 1). Indeed, in their first study, Heintzman *et al.* tested 4 regions by luciferase assays, with 3 of the 4 regions revealing *some* activity [65]. Importantly, these 4 regions were selected based on their overlap with STAT1 binding (observed by ChIP-chip after INFγ treatment) and therefore do not constitute an unbiased selection to assess method accuracy. In the following study by the same group[80], the 2010 De Santa et *al.* study[103], and the 2011 Ernst et *al.* study[87], the authors respectively validated 78%, 71% and between 50 and 75% of the tested predictions (using cell-based luciferase assays). On their side, Nègre *et al*. validated 30 out of the 33 (90%) regions tested (expression in transgenic embryos at some stage of development) [155]. Of note, these predictions were made on the combined presence of CBP and H3K4me1. A similar rate (8/9 or 89%) was obtained by Rada-Iglesias *et al.,* although they validated predictions of poised enhancers (based on p300 binding and H3K27me3 presence) by showing that these sequences were able to drive expression of a reporter gene in transgenic zebrafish embryos at some stage[98]. In our case, we tested the activity of 12 constructs corresponding to 9 predicted regions in transgenic embryos. Here, the regions were selected based only on information of their chromatin status and Pol II occupancy. 11 constructs corresponding to 8 of the 9 regions (89%) function as enhancers *in vivo*, and – importantly – yielded expression in the mesoderm at the precise stage of development predicted.

Study	Bio. Source	Marks	Assay	Val.	Notes
Heintzman 2007	Cell line	H3K4me1, H3K4me3	Luc	75% (3/4)	Selection based on additional TF (STAT1) binding
Heintzman 2009	Cell line	H3K4me1, H3K4me3	Luc	78% (7/9)	
De Santa	Cell line	H3K4me1, H3K4me3, Pol II	Luc	71% (5/7)	
Ernst	Cell line	Histone PTMs, CTCF, Pol II, H2A.Z	Luc	50-75% (9-13/18)	CRMs are marked by H3K4me3
Nègre	*Drosophila* embryo	H3K4me1, CBP	trans	90% (30/33)	Expression at any developmental stage
Rada-Iglesias	Cell line	p300 , H3K27me3	trans	89% (8/9)	Poised CRMs were validated for any later expression
This work	Tissue specific nuclei	Histone PTMs, H3, Pol II	trans	89% (8/9)	Active at predicted dev. stage

Table 1. Summary of validation results in different studies.
The table presents 6 studies and ours (last row) in which some of the predicted CRMs were evaluated for their ability to drive expression. The table presents the biological source ('Bio. Source') and the occupancy marks ('Marks') used in the studies, the assay type used for CRMs expression validation ('Assay', Luc: *in vitro* cell-based luciferase assay, trans: *in vivo* transgenic reporter assay) and the validation results ('Val.') with exact numbers in brackets.

Lastly, several studies[101,174], including ours, showed that the bimodal shape of the chromatin signal at active enhancer clearly correlates with activity and TF binding. In particular, He *et al.* used this property to predict FoxA1 binding in LNCaP prostate cancer cells by monitoring a change in the shape of the H3K4me2 signature before vs. after stimulation with an androgen receptor agonist[174]. Indeed, plots of chromatin modifications or nucleosome density indicate a depletion of nucleosome centrally, independent of the chosen anchoring point (TFBS, DHS, p300 or Pol II). Although we did not include features representing the signal shape in our Bayesian model, we think that this is largely compensated for by the presence of marks specifically associated with activity, in particular H3K27Ac, as shown in our and other studies[96,98], or yet indicators of transcription like H3K79me3 or presence of Pol II. Several studies[79,96,98,102-104] and our results strongly suggest that Pol II loading and subsequent transcription is a common feature of active enhancers. This hypothesis is further corroborated by Kowalczyk *et al.* who very recently showed that, in mouse primary erythroid cells, intragenic enhancers act as alternative tissue-specific promoters producing a class of abundant, spliced, multiexonic poly(A)+ RNAs (meRNAs) [175]. Importantly, in the case of our predicted and tested regulatory regions, we showed the

spatio-temporal expression driven by smaller sub-regions encompassing the domain of Pol II enrichment was largely indistinguishable from that of the larger regions (Figure 6a,b of the enclosed *Nature Genetics* paper, page 90). Although the Pol II signal is low and not always readily apparent, the nucleosome displacement might be more readily apparent, as indicated by the bimodal distributions of enhancer-associated histone PTMs. Thus, integrating signal shape information in our model would certainly increase the overall accuracy, improve the enhancer location prediction, and might even directly point to active TFBSs.

In this work, we concentrated on enhancers found in *intergenic* regions, as many chromatin marks found on enhancers are also present on (active) genes (which could lead to problematic signal interpretation if genic regions are not carefully excluded). However, enhancers are frequently found within gene bodies (about 54% according to our own results[36], in agreement with other studies[65]), in particular in the first introns. Our model may be directly applicable to enhancers within silent genes (as defined for example by an absence of H3K4me3 peak at the TSSs), but a different approach must be used to analyze enhancers in introns of expressed genes. Tentatively, combining chromatin signal shape with tissue-specific DHSs or p300/CBP binding could provide a good approach. Another alternative would be to couple chromatin signal shape with genome-wide prediction of CRMs using PWM collections found in dedicated databases. Luckily, a complete collection of key mesodermal factor binding locations at 6-8h (Twi, Bin, Mef2, Bap and Tin) is available[36]. This PWM collection could be used along with our tissue specific chromatin mark profiles to learn how to distinguish between active and inactive enhancers.

3.2 Article 2. A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History

3.2.1 Introduction

In *Drosophila*, the VM and the CM derive from the dorsal mesoderm. Heart development is presumably controlled by one of the most conserved developmental GRNs, with several key TFs, and their wiring, being conserved from fly to man[37]. For example, the *Drosophila* TFs Mad, Tin, Doc2 and Pnr all have known orthologs in vertebrates (Smad, Nkx2.5, Tbx5 and Gata4, respectively) and are all required to drive the cardiogenic program in both flies and mice. Moreover, many of these TFs are involved in conserved protein-protein interactions, like Tin with Mad[9] and with Pannier[10] in the fly, or Tbx5 with Gata4[11] and with Nkx2-5[12] in mice. Despite this conservation at the network level, several studies have emphasised the lack of sequence conservation of vertebrate heart enhancers compared to enhancers governing expression in other tissues[17,40,41,176]. This finding is surprising and suggests that *cis*-regulation might be conserved in a more subtle way (e.g. in terms of TFBS grammar) or at the protein-protein interaction level. Here, we used genome-wide ChIP-chip to study how pMad, dTCF, Doc, Pnr, and the mesoderm-specific factor Tin cooperate in *cis* during dorsal mesoderm specification in *Drosophila* embryos. Although 4 of them are expressed in multiple tissues, we show that these 5 TFs non-randomly co-localize at enhancers with Tin, suggesting that they do so in a mesoderm-dependent manner. Surprisingly, this occurs under very relaxed sequence requirements and suggests that the TFs are loaded onto the DNA as a collective in which diverse sets of TFs specifically interact with the DNA.

3.2.2 Personal contributions to this work

In this work, I have performed the analysis of ChIP-chip data up to the prediction of putative CRMs. This includes quality control, peak finding, peak summit fitting, CRM database computation, and production of CRM binding logos. I further performed all *de novo* motif analysis and provided conceptual and technical support in other computational analyses. Finally, I contributed to the final manuscript preparation (correction of the manuscript, creation of figures, writing of the methods, supplements, as well as revision, rebuttal, and proofing).

3.2.3 Article

Cell

A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History

Guillaume Junion,[1,3] Mikhail Spivakov,[1,2,3] Charles Girardot,[1] Martina Braun,[1] E. Hilary Gustafson,[1] Ewan Birney,[2] and Eileen E.M. Furlong[1,*]
[1]Genome Biology Unit, European Molecular Biology Laboratory, D-69117 Heidelberg, Germany
[2]European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[3]These authors contributed equally to this work
*Correspondence: furlong@embl.de
DOI 10.1016/j.cell.2012.01.030

SUMMARY

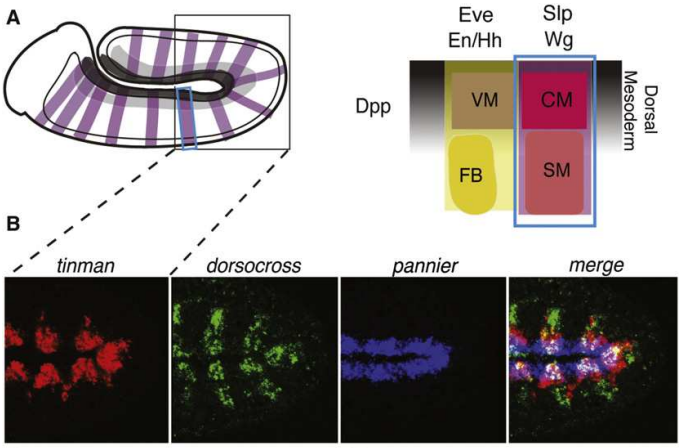
Cell fate decisions are driven through the integration of inductive signals and tissue-specific transcription factors (TFs), although the details on how this information converges in *cis* remain unclear. Here, we demonstrate that the five genetic components essential for cardiac specification in *Drosophila*, including the effectors of Wg and Dpp signaling, act as a collective unit to cooperatively regulate heart enhancer activity, both in vivo and in vitro. Their combinatorial binding does not require any specific motif orientation or spacing, suggesting an alternative mode of enhancer function whereby cooperative activity occurs with extensive motif flexibility. A fraction of enhancers co-occupied by cardiogenic TFs had unexpected activity in the neighboring visceral mesoderm but could be rendered active in heart through single-site mutations. Given that cardiac and visceral cells are both derived from the dorsal mesoderm, this "dormant" TF binding signature may represent a molecular footprint of these cells' developmental lineage.

INTRODUCTION

Pluripotent cells become progressively restricted in their cell fate through the action of inductive signals from surrounding tissues and specific cohorts of transcription factors (TFs). This multilevel information converges on *cis*-regulatory modules (CRMs, or enhancer elements) to elicit specific developmental programs. Information at some CRMs is integrated through cooperative TF binding, mediated via direct protein-protein interactions between TFs or common cofactors. Cooperative occupancy often requires a specific orientation, relative spacing, and helical phasing of TF-binding sites (Senger et al., 2004), referred to as motif grammar, to facilitate the appropriate protein interactions. A classic example of this is the enhanceosome model of enhancer activation (Panne, 2008). However, this stringent enhanceosome mode of regulation may represent only a small fraction of enhancers. Many developmental enhancers operate under more flexible conditions in which a subset of factors may bind cooperatively while the remaining factors are recruited independently and thus require little or no motif grammar. The billboard model, for example, suggests that TFs do not function in a single concerted manner at enhancers; rather, submodules interact independently and/or redundantly with the basal transcriptional machinery (Kulkarni and Arnosti, 2003). In some cases, enhancer flexibility appears even more extreme—not only can the relative location of binding sites vary, but also the identity of the TFs that are involved in regulating a specific pattern of expression (Brown et al., 2007; Zinzen et al., 2009).

The specification of the *Drosophila* dorsal mesoderm into visceral mesoderm (VM) and cardiac mesoderm (CM) cell fates represents an excellent paradigm for complex enhancer integration (Halfon et al., 2000; Kelly and Buckingham, 2002; Xu et al., 1998; Zaffran and Frasch, 2002). Here, cell fate decisions are induced through the intersection of ectodermal Wingless (Wg, a Wnt protein) and Decapentaplegic (Dpp, a TGF-β family protein) signaling (Figure 1A). Pluripotent cells that receive both signals within the underlying dorsal mesoderm are specified to become CM, and the neighboring cell population that only receives Dpp signal becomes VM (Lee and Frasch, 2000; Lockwood and Bodmer, 2002) (Figure 1A). Tinman (Tin, an Nkx factor) and pMad (the effector of Dpp signaling) provide the competence for these "precursor cells" to acquire either a VM or CM cell fate (Xu et al., 1998). In particular, Tin acts together with Pannier (Pnr, a GATA factor) and Dorsocross (Doc, a T box factor) to specify CM cell fate (Reim and Frasch, 2005), whereas the VM fate is actively repressed in these cells (Lee and Frasch, 2005) (Figures 1A and 1B).

Genetic studies in both *Drosophila* and mice suggest that the *cis*-regulatory network driving cardiac specification is highly cooperative. For example, although Nkx, GATA and T box factors are essential for heart development in all species studied to date (Cripps and Olson, 2002; Frasch, 1999; Olson, 2006; Reim and Frasch, 2005), neither factor alone is sufficient to



Tin	Doc	Pnr	Cardiac Cells (stage 16)
GOF			WT
	GOF		+
		GOF	+
GOF	GOF		++
GOF		GOF	+
	GOF	GOF	+++
GOF	GOF	GOF	+++
LOF -/+	LOF -/+		--
	LOF -/+	LOF -/+	--
LOF -/+	LOF -/+	LOF -/+	---

Figure 1. Dorsal Mesoderm Specification into Cardiac and Visceral Mesoderm during *Drosophila* Embryogenesis

(A) Diagram of a *Drosophila* embryo showing *wg* expression in 14 parasegments. Area indicated by blue rectangle is enlarged in the right panel, showing a schematic representation of mesoderm subdivision in one hemisegment. The dorsal domain, which has high levels of Decapentaplegic (Dpp) signaling (black), gives rise to visceral mesoderm (VM) and cardiac mesoderm (CM), whereas ventral regions become fat body (FB) and somatic muscle (SM). CM is specified at the intersection of Wingless (Wg, purple) and Dpp signaling in the posterior part of each parasegment. Wg activates *sloppy paired* (*slp*) expression, and together they promote CM and repress VM specification.

(B) Triple-fluorescent in situ hybridization showing *tinman*, *dorsocross*, and *pannier* expression in the dorsal mesoderm during early stage 11, when cardiac specification takes place. All three genes are coexpressed exclusively in the cardiogenic mesoderm (pink-white area of coexpression). The region of the embryo shown is depicted by the black square in (A).

(C) Summary of the genetic interaction between Tinman (Tin), Dorsocross (Doc), and Pannier (Pnr) to CB specification (Reim and Frasch, 2005). GOF, gain-of-function; LOF, loss-of-function (−/+ = heterozygous) genetic backgrounds. + and − denote an increase or decrease in the number of cardioblasts, respectively.

(D) Recursive regulation between the key factors essential for CM specification. Solid lines indicate direct regulation; dashed lines represent a genetic interaction (direct or indirect).

See also Figure S1.

induce a cardiac cell fate. Rather, the ectopic expression of combinations of TFs is required to drive the cardiogenic program in both flies (Reim and Frasch, 2005) and mice (Durocher et al., 1997; Sepulveda et al., 1998) (Figure 1C). Moreover, combinations of GATA4, Tbx5, and a third factor are sufficient to drive transdifferentiation of cell types into a CM cell fate (Takeuchi and Bruneau, 2009) and to direct reprogramming of fibroblasts into cardiomyocytes (Ieda et al., 2010), and yet, the molecular nature of this cooperativity is very poorly understood. Despite the extensive genetic characterization of CM specification, only a handful of enhancers are known to regulate early stages of heart development (Figures S1A–S1G available online), precluding any general hypotheses on how the input from multiple TFs (Tin, Pnr, Doc, and the effectors of Wg and Dpp signaling) converges in *cis*. For example, it is not known if the cooperativity observed between these factors at a genetic level (Figures 1C and 1D) is reflected at the *cis*-regulatory level and requires a specific motif grammar at the sequence level.

To address these issues, we examined the genome-wide occupancy of pMad, dTCF, Doc, Pnr, and the mesoderm-specific factor Tin during dorsal mesoderm specification in *Drosophila* embryos. We find that all five TFs are recruited to shared enhancers to a much higher degree than expected by chance and do so in a mesoderm-specific context, matching their only domain of coexpression (Figure 1B). These regions function as heart enhancers in vivo and require the presence of all five TFs for their cooperative regulation and maximal enhancer activity in vitro. The collective enhancer occupancy, which we further confirm using a cell culture model and mutagenesis analysis in vivo, occurs in the absence of any consistent motif grammar, revealing an alternative mode of cooperative regulation using very flexible motif content. Our analysis also uncovered an additional property of developmental enhancers, whereby dormant TF binding signatures reflect a developmental footprint of a cell's lineage. "Cardiac" TFs occupy enhancer elements that are active in the neighboring VM, echoing the fact that both cell populations are derived from the dorsal mesoderm.

RESULTS

Building a TF Binding Atlas for Enhancers Active in the Dorsal Mesoderm

To generate a TF binding atlas of regulatory regions active in the dorsal mesoderm, we performed genome-wide ChIP-on-chip experiments with antibodies directed against Doc, Pnr, dTCF, and pMad, the activated phosphorylated form of the Dpp effector Mad. The experiments were performed at two

consecutive stages of development: 4–6 hr after egg lay (stages 8 and 9) and 6–8 hr (stages 10 and 11), corresponding to the subdivision of the dorsal mesoderm and its subsequent specification into CM and VM (Campos-Ortega, 1997). A high confidence set of TF-bound regions was defined for each factor, identifying thousands of occupied sites per TF (Table S1 and Extended Experimental Procedures). These data were combined with Tin occupancy data generated under the same conditions at the same stages of development (Zinzen et al., 2009). The pairwise occupancy patterns of all five TFs showed highly significant overlap (Figures S2E and S2F), providing an initial indication that these factors occupy common *cis*-regulatory elements.

To convert TF binding peaks into co-occupied enhancers, binding events that clustered in close proximity to each other were merged to define putative *cis*-regulatory regions, as described previously (Zinzen et al., 2009). In this way, the combined 55,423 significant TF binding peaks clustered into 11,286 nonredundant CRMs, approximately one-third of which (4,041) have significant levels of Tinman binding. Though *tinman* expression is restricted to the mesoderm, the expression of the other four TFs is not, even within these narrow time windows of development (Figure 1B). We used Tin binding to limit our analysis to CRMs more likely to be active in mesodermal lineages and therefore focused for the remainder of this study on the 4,041 Tin-bound CRMs, the majority of which also recruited other factors (Table S2).

The TF occupancy patterns fully recapitulated *all* known binding to previously characterized dorsal mesoderm enhancers and the four known early cardiac enhancers active at the analyzed stages (Figures S1A–S1H) and in many cases revealed additional regulatory connections, demonstrating the sensitivity and resolution of the data. To extend this further, we selected 50 genes with at least Tin, Doc, and Pnr binding in their vicinity and examined their expression patterns by double-fluorescent in situ hybridization. Forty-two of these genes gave specific spatiotemporal expression, 38 of which (90%) are expressed in the dorsal mesoderm (26 genes) and/or cardiac mesoderm (20) and/or visceral mesoderm (7) (Figures S1I–S1K and Table S3). This supports our reasoning that the integration of binding signatures for four nonmesoderm-specific TFs (pMad, dTCF, Pnr, and Doc) with Tin is a valid approach to focus on transcriptional regulation within the dorsal mesoderm and its derivatives.

A Regulatory Collective of Cardiogenic TFs Is Recruited to Tin-Bound Enhancers

To relate TF binding signatures to specific *cis*-regulatory function, we applied an unbiased clustering approach to assess general TF preferences for enhancer co-occupancy, followed by extensive in vivo transgenic reporter analyses to assess enhancer activity. The maximum moving average ChIP signal for each TF at each enhancer was used as a quantitative input for enhancer classification. As enhancers were defined based on high-confidence binding signals for at least one TF, this procedure ensured that subthreshold signals for all other TFs were taken into account for enhancer classification (Extended Experimental Procedures). The Bayesian clustering algorithm Autoclass (Cheeseman, 1996) was used to partition enhancers based on their similarity in TF binding signatures across all experiments, computing a probability score for each enhancer to belong to each cluster. This approach produced confident single-cluster assignments for 77% (3,099) of the 4,041 Tin-bound enhancers (Figure 2A, left heatmap, and Table S4), and the robustness of this classification was confirmed by bootstrap analysis (Extended Experimental Procedures).

Examining the signal distribution of TF occupancy in each cluster revealed six broad enhancer classes that are qualitatively distinct from each other (Figures 2A, left, S2A, and Extended Experimental Procedures). The first class harbors enrichment for all five TFs (Figures 2A, left, "All TF" CRMs labeled with shades of red, and S2A, upper-left). The second class, in contrast, is depleted in binding signal for all TFs except Tin and represents ~20% of CRMs ("Tin only," labeled with shades of gray in Figure 2A). The four remaining classes are defined by elevated signals for Tin and one additional TF, with generally medium to low signals for other factors. We loosely refer to these as "two TF" classes as follows: "pMad+Tin" (~2% CRMs), "dTCF+Tin" (~8%), "Doc+Tin" (~4%), and "Pnr+Tin" (~20% CRMs) (Figures 2A, left, and S2A). Individual clusters within each of these classes differ in the quantitative levels of TF binding signals but generally not in the identity of the TFs themselves (Figures 2A and S2A and Extended Experimental Procedures). CRM clusters with the most prominent binding profiles from each class were used for further analysis (Figures 2C and S2A, boxed histograms, and Extended Experimental Procedures).

This unbiased grouping of enhancers, based on their similarity in TF occupancy, revealed two unexpected findings. First, the most prominent binding signature at enhancers is the recruitment of all five TFs. Depending on the threshold of the mean TF binding signal per class (Extended Experimental Procedures), between 22% and 46% of classified enhancers have highly correlated signals for Doc, dTCF, pMad, Pnr, and Tin across one or both developmental times (Figures 2A, left, labeled with shades of red from high TF binding signal [top] to low [bottom], and S2A). Second, there are few enhancers bound at high levels by three or four TFs; instead, the majority of regions are either occupied by all five factors or have high enrichment for only two factors (TF+Tin). This suggests that all five factors bind to these elements as a collective unit, which may require a specific mesodermal context to anchor their binding. To test this further, we applied the same clustering procedure to enhancers that are significantly bound by one or more TF but are not bound by Tin (the mesoderm-specific factor) at the analyzed stages of development (1,209 CRMs with near-zero Tin signal; Table S2). On these Tin-negative regions, there is very little correlated co-occupancy of the other four TFs (Figure 2A, right). This was further confirmed on a stringent set of enhancers that are highly enriched for two or more TFs other than Tin, whereas the signal for the remaining analyzed TFs is below the lower 50% of the background signal distribution (Figure 2B). The occupancy of Doc, dTCF, pMad, and Pnr at these "all or nothing" CRMs is strikingly different depending on the presence of Tin binding (Figure 2B). More globally, the degree of TF co-occupancy is significantly higher at Tin-bound regions, but not Tin-negative regions, compared to that expected at random (Figures S2B–S2D).

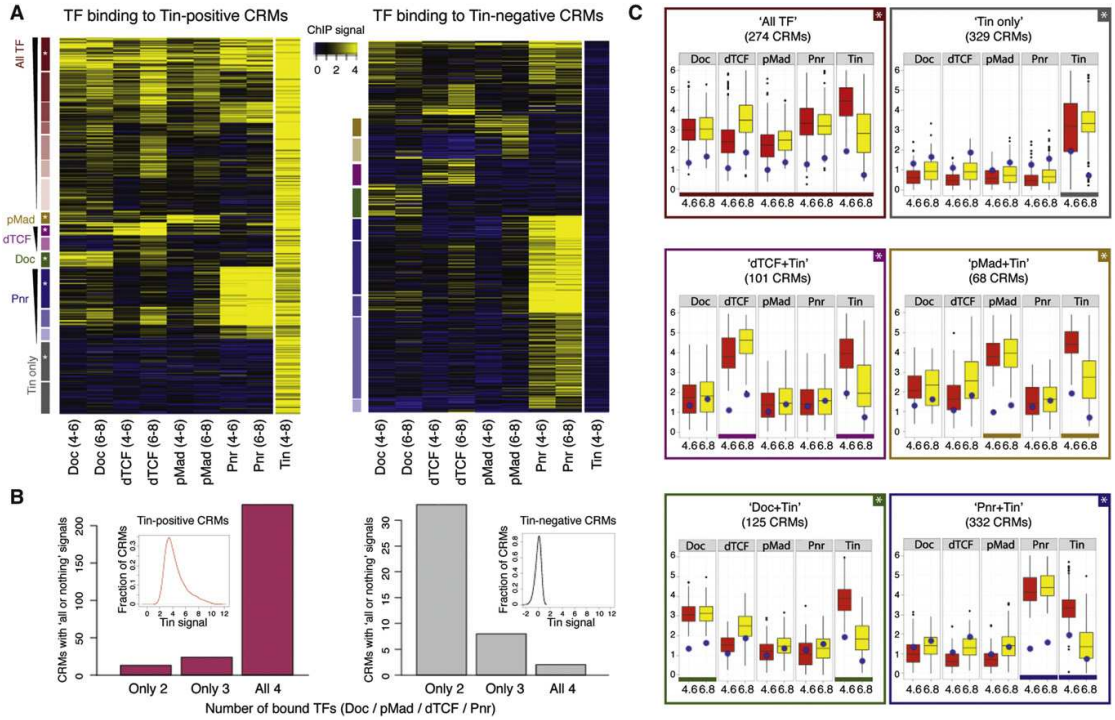


Figure 2. Co-Occupancy of Cardiogenic TFs at Tin-Bound CRMs

(A) Unsupervised classification of Tin-positive (left) and Tin-negative (right) CRMs using Autoclass Bayesian clustering. Rows correspond to defined CRMs, and columns correspond to the maximum moving average ChIP signal for a transcription factor (TF) at the indicated time points. Yellow represents high and blue background signal. Rectangles to the left of the heatmaps indicate subclasses of CRMs with related TF binding signals; white asterisks indicate subclasses with the most prominent TF binding signal from each class, which was selected for further analysis (shown in Figure 2C).

(B) Assessment of TF co-occupancy on a subset of CRMs that have either strong or background ("all or nothing") signals for each of the four analyzed TFs. Bar charts show the number of such CRMs occupied by two to four TFs on Tin-positive (left) or Tin-negative (right) CRMs. Density plots show the distribution of Tinman signal in each of the two CRM subsets (inset).

(C) Representative subclasses of each binding signature (marked with asterisks in Figure 2A) used for further analysis. Boxplots show the distributions of ChIP signals for the five TFs at two time points (4–6 and 6–8 hr). Blue dots show median signal for each TF across all CRMs. Comparing the position of the blue dot to the median area of the box plot indicates whether a TF's binding is specifically enriched on this group of CRMs.

See also Figure S2.

Taken together, these data indicate that all five TFs tend to be corecruited to regulatory regions in a concerted manner (as further confirmed using a cell-based system below), which occurs in a mesoderm-specific context (Tin-bound CRMs), in keeping with their only domain of coexpression (Figure 1B).

Enhancers Occupied by All Five TFs Regulate Expression in the Dorsal Mesoderm or Its Derivatives

Having defined specific classes of enhancers with qualitatively different TF occupancy patterns, we assessed which of these represent active enhancers in vivo and drive expression in the dorsal mesoderm and/or in cardiac cells. ChIP-defined enhancers (average size 550 bp) were cloned upstream of a GFP reporter gene and stably integrated into the *Drosophila* genome. Enhancer spatiotemporal activity throughout embryonic development was assayed by in situ hybridization in transgenic embryos to provide accurate temporal resolution for when the enhancer is active. Importantly, the selection of enhancers was based purely on representative binding signatures, without prior knowledge concerning the function of neighboring genes or the motif content of the enhancers. In total, the activities of 55 regions were examined in transgenic embryos, almost half of which correspond to the All TF binding class (47%), as this represents the most predominant TF binding signature (Figures 3, 6, and S3 and Table S5).

A striking 92% of enhancers tested from the All TF class (24 of 26 regions) were sufficient to function as enhancers in vivo. The vast majority of these (91.6%; 22/24) regulate expression in mesodermal lineages (Figure 3A), of which the most prominent expression signature (50%; 12 CRMs) is activity within the cardiogenic mesoderm (Figures 3B and S3A). These complex spatial patterns of enhancer activity cannot be achieved through

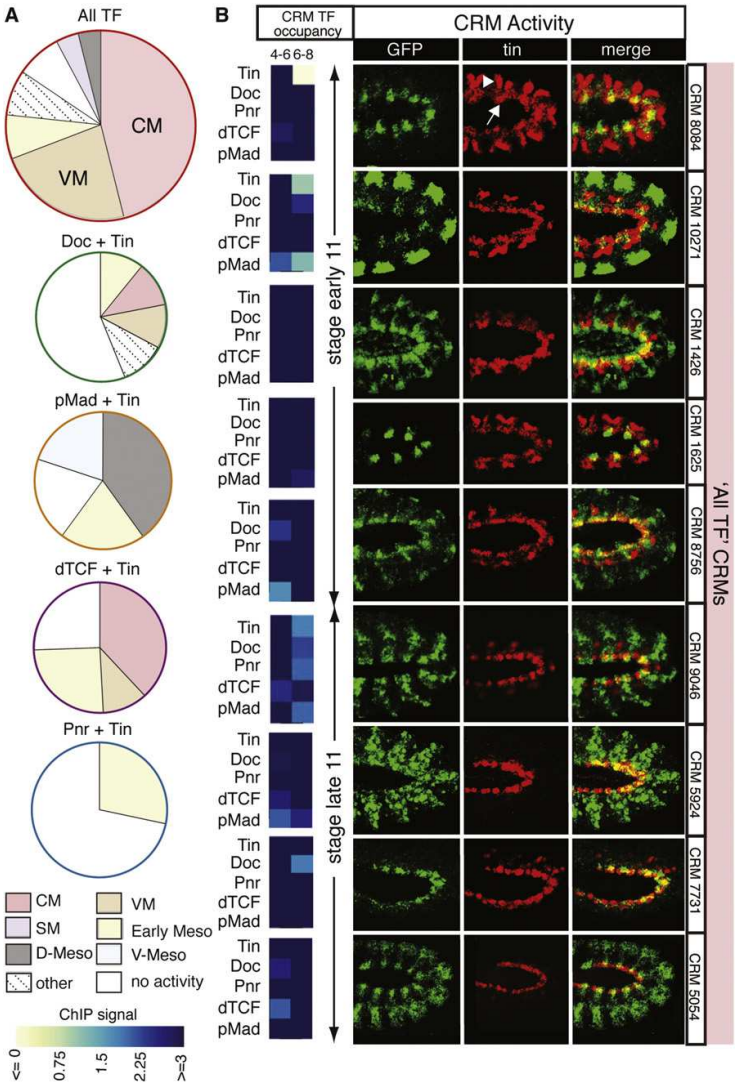


Figure 3. Collective TF Occupancy Correlates with Enhancer Activity in Cardioblasts

(A) Summary of the activity of 55 CRMs tested in vivo by transgenic reporter assays. Pie charts represent the proportion of CRMs driving expression in different tissues for each Autoclass-derived subclass. CRMs active in two (or more) mesodermal tissues are indicated in both. All TF CRMs had the highest percentage of regions that functioned as enhancers in vivo; 84.6% regulate expression in the mesoderm and/or its derivatives, with cardiac mesoderm (CM) expression being predominant (46%). VM, visceral mesoderm; SM, somatic muscle; Early Meso, early mesoderm; D-Meso, dorsal mesoderm; V-Meso, ventral mesoderm; other, nonmesodermal tissues.

(B) CRM spatiotemporal activity assayed by in situ hybridization of embryos with a transgenic reporter. (Left) The TF binding signals for each factor on each CRM at both time points (mean moving average ChIP signal per CRM; blue represents high levels). (Right) In situ hybridization using antisense RNA probes directed against the *GFP* reporter (green) and *tin* (red) as a marker of dorsal mesoderm and its derivatives. At stage early 11, *tinman* is expressed in visceral (arrowhead) and cardiac mesoderm (arrow); by late 11, only cardiac expression remains. CRMs show activity restricted to CM (1625, 7731) or more complex patterns in CM and other cell types (1426, 9046, 5054), reflecting the nonexclusive expression of many heart genes. All embryos shown laterally; anterior, left; dorsal, up; region depicted is indicated by the black square in Figure 1A. The remaining tested CRMs are shown in Figures 5, 6, and S3.

See also Figure S3.

the action of any one TF alone but, rather, reflect the intersection in expression domains of many of these factors, in line with their observed collective occupancy. The second most prominent activity (25%) was VM expression, which was surprising given the collective occupancy of all five "cardiogenic" TFs at these CRMs and is dissected in detail below.

The activity of approximately seven CRMs was tested for each of the four two-TF classes, of which 59% (17/29) function as enhancers in vivo (detailed results are shown in Figure S3). Eighty-eight percent (15/17) of active regions regulate activity in mesodermal tissues, including the early trunk, ventral, dorsal, and visceral mesoderm. However, in contrast to the All TF CRMs, only four regions (23%) regulate activity in CM, with all but one CRM belonging to the dTCF+Tin class (Figures 3A and S3).

In summary, regions co-occupied by all five TFs were much more likely to direct expression in the dorsal mesoderm (or its derivatives) compared to any of the two-TF classes, with 75% (18/24) of active All TF CRMs driving specific expression in CM or VM. It is important to note that Tin binding in the absence of Doc, Pnr, pMad, and dTCF is not sufficient to regulate enhancer activity in cardiac cells, as demonstrated by extensive analyses of enhancers bound by Tin in combination with other TFs (Liu et al., 2009; Zinzen et al., 2009). Therefore, the activity of Tin within the cardiogenic TF collective has unique properties in terms of its functional output.

Relaxed Sequence Requirements at Enhancers Occupied by All Five Factors

Given the extensive corecruitment of the five TFs, we asked whether the motif content of the All TF enhancers explains their collective occupancy and activity. We first determined the

general sequence preferences of each TF using de novo motif discovery on all regions bound by that factor (Extended Experimental Procedures). The identified position weight matrices (PWMs), which were similar to published models (Figure S4A), were then used to assess differential motif enrichment between All TF CRMs and two-TF CRMs (Figure 4). This analysis revealed two classes of TFs, suggesting different modes of their recruitment to DNA. Doc and Pnr transcription factor binding sites (TFBS) are preferentially found in All TF CRMs compared to their respective two-TF CRMs, whereas, in contrast, the numbers of TFBSs for pMad, dTCF, and Tin are lower in All TF CRMs compared to their respective two-TF CRMs (Figure 4B). This holds true regardless of which specific PWM score threshold is used for the motif detection (data not shown) or when using an unthresholded approach summarizing both high- and low-affinity TFBSs (TRAP, Roider et al., 2007; Figure 4C). The differential motif enrichment of Doc and Pnr compared to pMad, dTCF, and Tin was further confirmed using de novo motif analysis (Figure S4B).

The enrichment of Doc and Pnr TFBSs suggests that these factors are preferentially recruited to All TF CRMs in a sequence-specific fashion (Figures 4B and 4C), and consistent with this, their motifs are more conserved in All TF CRMs compared to their respective two-TF classes (data not shown). Conversely, the number of pMad, dTCF, and Tin sites are lower in All TF CRMs compared to their respective two-TF CRMs, which is particularly striking for dTCF (Figures 4B and 4C), suggesting that heterotypic cooperative binding may play a role in their recruitment to All TF CRMs. A role for cooperativity in this system is supported by direct protein-protein interactions between almost all of these TFs in both *Drosophila* and vertebrates (Brown et al., 2004; Bruneau et al., 2001; Durocher et al., 1997; Gajewski et al., 2001; Garg et al., 2003; Nishita et al., 2000; Zaffran et al., 2002).

Protein-protein interactions between TFs can often introduce sequence constraints within enhancers, where the relative spacing and orientation of motifs must maintain a certain configuration to facilitate protein interaction and binding (Panne, 2008). We searched for this type of motif grammar, examining the relative motif spacing and orientation of Doc, Pnr, pMad, dTCF, and Tin TFBS within All TF CRMs. Surprisingly, we found no evidence of consistent grammar as a characteristic signature of All TF CRMs (“CRM Grammar Analysis” in Extended Experimental Procedures and Figures S4C and S4D). Moreover, the motif content itself is highly diverse, whereby the occurrence of pMad, dTCF, and Tin sites and distance between them varies between each All TF CRM. Despite this motif heterogeneity, however, these enhancers recruit all five TFs and function as heart enhancers in vivo, mirroring the cooperative function of these TFs during heart development.

The Presence of Pnr and Doc Is Essential for Tin-pMad-dTCF-Mediated Enhancer Activation

The collective occupancy and activity of the All TF enhancers suggests that the high level of cooperativity observed between these factors at a genetic level extends to their downstream *cis*-regulatory network (Figures 1C and 1D). To examine this further, we generated a cell culture-based model that expresses all five TFs in their active forms. Although this system lacks the spatial and temporal context of the developing embryo, it provides a more homogenous cell population. Based on extensive RNA-seq data (Cherbas et al., 2011), we found that DmD8 cells (an established *Drosophila* cell line derived from dorsal mesothoracic disc) express *pnr* and *doc*, but not *tin*, which we also confirmed at the protein level (Figure S5A). Although all components of the Wg and Dpp signaling cascades are expressed, the ligands are not; therefore, these signaling pathways are inactive in this cell line. To obtain activated dTCF and pMad, we generated conditioned DmD8 medium containing secreted Wg and Dpp. Applying this conditioned medium to fresh DmD8 cells resulted in the phosphorylation of Mad and the activation of the Wg signaling pathway (Figure S5B). Therefore, upon *tin* transfection, all five TFs were active in this cell culture system.

We used this cell culture system to examine: (1) the co-occupancy of all TFs by ChIP followed by quantitative PCR and (2) the requirement of each TF for enhancer activity by luciferase assay. The results for one enhancer (CRM 3436) are highlighted in Figure 5. CRM 3436 is bound by all five TFs in vivo (Figure 5A) and is sufficient to regulate expression in a segmentally repeated pattern encompassing part of the cardiogenic mesoderm (Figure 5B). Performing ChIP for all five factors in cell culture revealed significant occupancy of each TF on the endogenous enhancer locus compared to an unbound negative region (Figure 5C). A similar significant enrichment in the occupancy of all TFs was observed for all six enhancers analyzed (Figures S5C and S5D), confirming the collective occupancy observed in vivo.

To examine the regulatory input of these TFs, three All TF CRMs were placed upstream of a minimal promoter driving a luciferase reporter and transfected into DmD8 cells where both Pnr and Doc were depleted using RNAi to obtain a basal level of the enhancer’s activity. The presence of either Pnr or Doc alone had no significant effect on enhancer activity, whereas both together caused a marginal increase (Figures 5D, S5E, and S5F). Addition of Tin in the presence of Pnr and Doc, however, caused a significant increase in activity, whereas the presence of all five activated TFs had the most dramatic effect, leading to a 15-fold increase over the basal level (Figure 5D). These results demonstrate that all five TFs contribute to the enhancers’ activity and are required for maximal enhancer activation (Figures 5D, S5E, and S5F).

The clear differences in the enrichment and conservation of Pnr and Doc motifs compared to those of Tin, dTCF, and pMad suggest that these two TFs may preferentially serve as anchors for the collective TF binding. Taking advantage of this cell system, we systematically tested this hypothesis by removing Doc alone, Pnr alone, or both in the presence of the other three TFs. As shown in Figure 5D (red asterisk), removal of Doc had a significant effect, whereas the removal of Pnr alone reduced the enhancers activity back to its basal level, despite the presence of Tin, pMad, and dTCF. Therefore, the presence of Pnr and Doc is required for the ability of Tin, dTCF, and pMad to activate the enhancer. The fact that Pnr alone or in combination with Doc is not sufficient for significant enhancer activation suggests that these TFs are essential for the collective recruitment of all five TFs, consistent with the motif content of these CRMs.

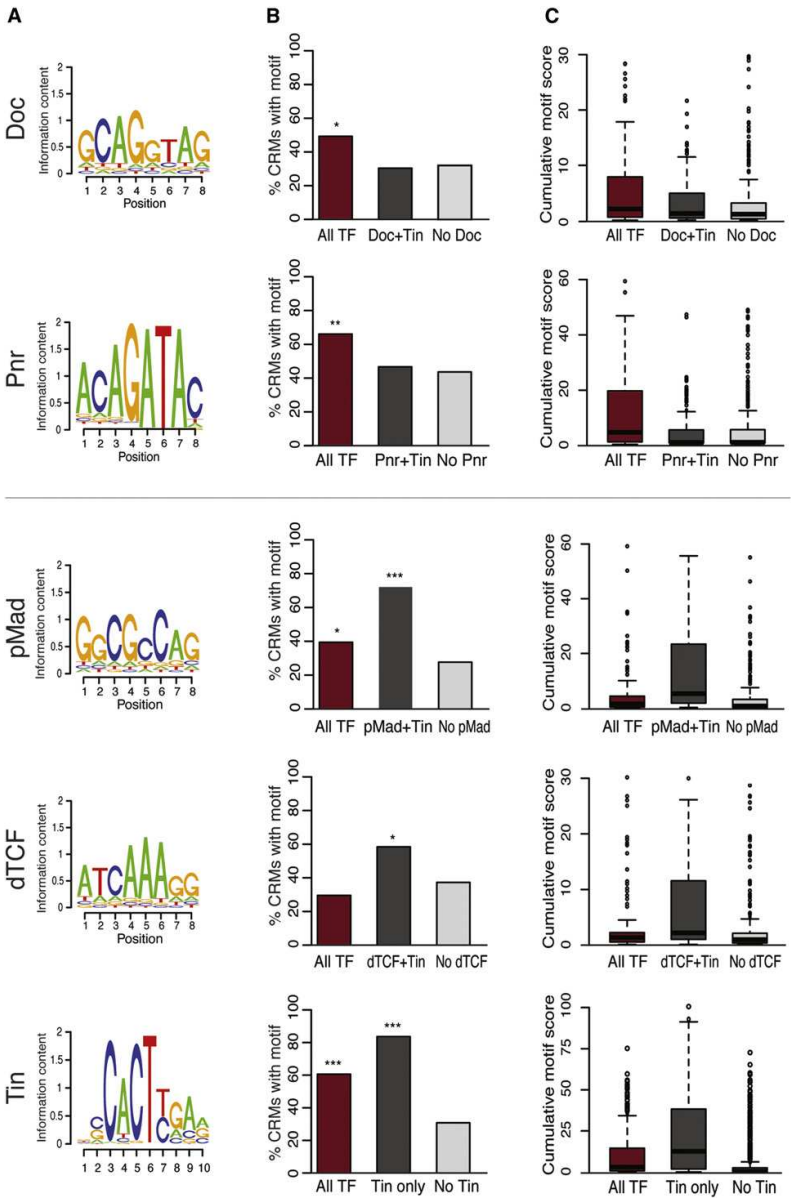


Figure 4. Sequence Properties of All TF CRMs versus Two TF CRMs

(A) Motifs discovered de novo for Doc, Pnr, pMad, dTCF, and Tin in all regions bound by the respective TF are similar to those reported previously (Figure S4A).

(B) Enrichment of TF-binding sites in different CRM classes. Doc and Pnr motifs are more frequently found in All TF CRMs compared to their two TF classes, whereas pMad, dTCF, and Tin motifs are more frequently found in their respective two-TF CRMs, compared to All TF CRMs.

(C) Cumulative motif enrichment scores (computed without score thresholds; TRAP) confirm the differential motif enrichment: Doc and Pnr motifs have elevated cumulative scores in All TF CRMs compared to their respective two TF CRMs (Wilcoxon test $p = 0.02$ and $p = 3.8 \times 10^{-12}$, respectively) and those not bound by the analyzed TF ($p = 2.5 \times 10^{-6}$ and $p = 6.7 \times 10^{-14}$). In contrast, pMad, dTCF, and Tin have lower cumulative motif scores in All TF CRMs compared to their respective two-TF CRMs (pMad $p = 2.3 \times 10^{-8}$, dTCF $p = 6.7 \times 10^{-6}$, Tin $p = 8.8 \times 10^{-11}$). Cumulative motif scores (computed using TRAP) are normalized to the median value for each TF.

See also Figure S4.

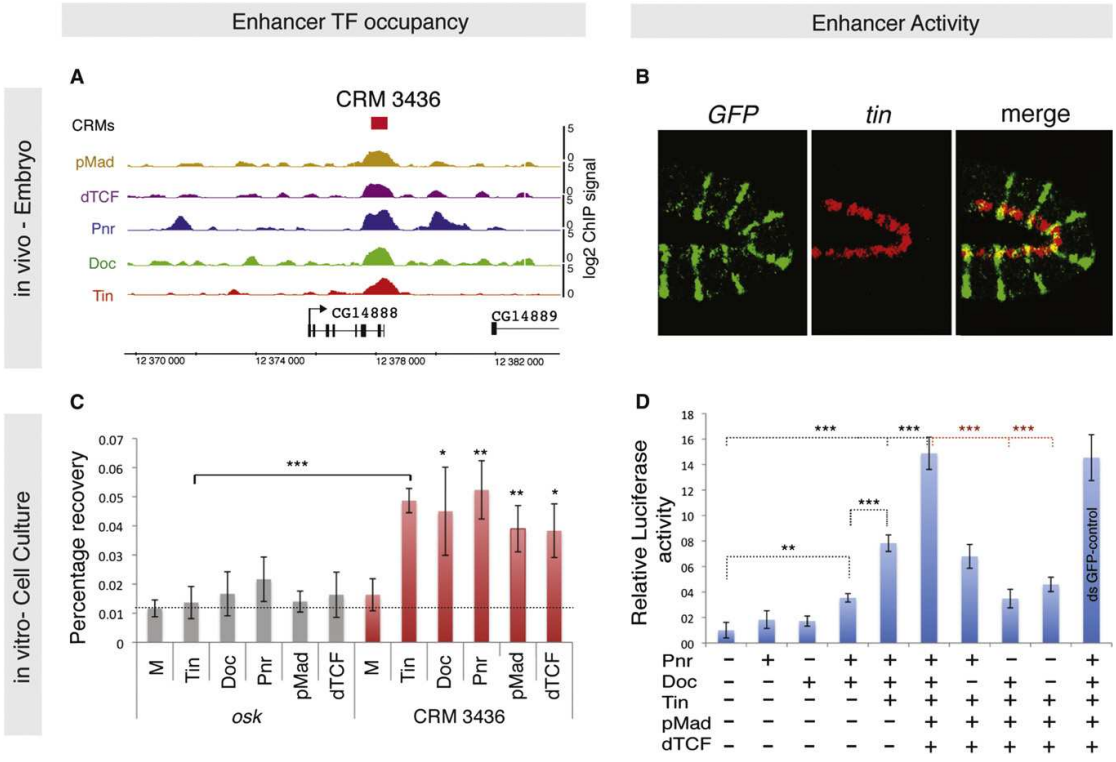


Figure 5. The Presence of Pnr and Doc Is Essential for the Ability of Tin, pMad, and dTCF to Activate Heart Enhancers

(A) CRM 3436 is bound by all TFs in vivo. Shown is $\log_2$ ChIP signal for each TF at embryonic stages 9–11 (merged 4–8 hr data).

(B) CRM 3436 spatiotemporal activity assayed by in situ hybridization of transgenic embryos containing a stable insertion of the ChIP-bound region (red rectangle in A) regulating a GFP reporter. Antisense RNA probes directed against *GFP* (green) and *tin* (red) as a marker of dorsal mesoderm and its derivatives reveal overlapping expression in cardiac mesoderm, indicated by the yellow area of coexpression (merge panel).

(C) CRM 3436 is occupied by all TFs in DmD8 cells. DmD8 cells containing activated forms of all five TFs were used for ChIP experiments followed by real-time PCR of the endogenous enhancer. Gray histograms represent percentage recovery of input for a negative unbound region (maternal gene *oskar*, *osk*); red histograms represent occupancy on CRM 3436. The binding of each TF is significantly enriched on CRM 3436 compared to that TF's enrichment on the negative region (indicated by solid line for Tin). *M*, mock reaction. Error bars show the standard deviations of triplicate experiments. p values (one-tailed type 2 *t* test): *p = < 0.05; **p = < 0.01; ***p = < 0.001.

(D) Luciferase assay of CRM 3436 activity in DmD8 cells. The first column indicates the basal level of the enhancer's activity, using dsRNAi to remove Pnr and Doc (Figure S5A). Error bars show the standard deviations of two biological replicates, each conducted in triplicate. p values (two-tailed type 3 *t* test): *p = < 0.05; **p = < 0.01; ***p = < 0.001. Results from all CRMs tested are shown in Figure S5.

See also Figure S5.

Biniou Binding Is Predictive of VM Activity for CRMs Co-Occupied by Cardiogenic TFs

Examining the activity of the All TF CRMs revealed that, while 50% regulate expression in the cardiogenic mesoderm (Figure 3), an additional 25% have specific activity in the visceral mesoderm (Figure 6A). This VM activity was unexpected given the collective binding of all five cardiogenic TFs (that are not coexpressed in the VM), which we further confirmed for three CRMs in our cell culture-based system (Figure S5D). Of note, the CM and VM activity were mutually exclusive, suggesting a "CM-VM" regulatory switch. To dissect the mechanism of this bimodality, we first assessed whether a central VM-specific regulator, Biniou, is bound to these enhancers based on our previously published data (Zinzen et al., 2009). Biniou is a FoxF TF that is specifically expressed in VM, where it is essential for its specification and subsequent differentiation (Jakobsen et al., 2007; Zaffran et al., 2001). Consistent with our expectation, Biniou ChIP signal is significantly higher at characterized enhancers with VM-specific activity compared to those active in CM (Figure 6B, top; Wilcoxon test p = 0.01). Biniou occupies these CRMs only at the early stages of dorsal mesoderm specification into VM and CM (6–8 hr) and not at later development stages (8–10 hr), mirroring their transient activity (Figure 6B, bottom). Extending this analysis to the entire All TF class

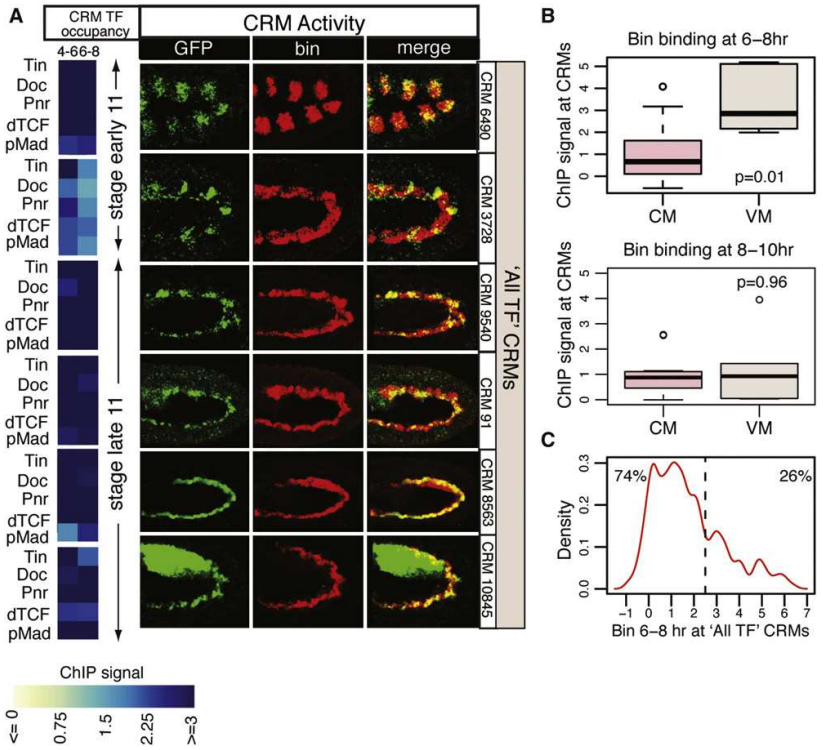


Figure 6. Biniou Occupancy Predicts Visceral Muscle Activity for Enhancers Collectively Bound by Cardiogenic TFs

(A) Unanticipated CRM activity in visceral mesoderm and not cardiac mesoderm for 25% of All TF CRMs tested. (Left) TF binding signals (mean moving average ChIP signal per CRM, wherein blue represents high enrichment). (Right) CRM activity by in situ hybridization using antisense RNA probes directed against the *GFP* reporter gene (green) and *biniou* (red) as a specific marker for VM. The CRMs drive expression in trunk VM (CRMs 6490, 91, 8563, 9540, and 10845) or in restricted populations of VM cells (CRM 3728).

(B) Box plots showing significantly higher levels of Biniou (Bin) occupancy at All TF CRMs driving VM (visceral mesoderm) expression compared to CM (cardiac mesoderm) (Wilcoxon test p = 0.01). This enrichment is only present at 6–8 hr, the stages of dorsal mesoderm specification (stages 10 and 11, top) and not at later stages (bottom).

(C) Density of all All TF CRMs with high or low Biniou (Bin) occupancy at 6–8 hr (x axis). Twenty-six percent of All TF CRMs have high levels of Bin binding, consistent with the proportion of tested CRMs with VM activity (25%).

revealed high levels of Biniou binding at ~25% of enhancers (Figure 6C), consistent with the proportion of tested CRMs showing VM activity. Therefore, a high level of Biniou binding at 6–8 hr (stages 10 and 11) is highly predictive of VM-specific activity, as indicated by the largely nonoverlapping distributions in ChIP signals (Figure 6B, top), and is consistent with the model of Biniou as an instructive regulator of VM cell fate (Jakobsen et al., 2007; Zaffran et al., 2001).

A Lineage Switch Motif Occupied by Two Fox Transcription Factors

Based on our current knowledge of how VM enhancers function, the binding signatures of Biniou, Tin, pMad, and another regulator Bagpipe (Azpiazu and Frasch, 1993) fully explain enhancer activity in the trunk visceral mesoderm at stage 10 (Lee and Frasch, 2005; Lee et al., 2006). However, this model does not explain the observed collective occupancy of cardiogenic TFs on these enhancers in the juxtaposed heart field or the fact that this binding signature is not sufficient to induce CM transcription, whereas other enhancers with similar binding signatures exhibit CM activity (compare Figure 6A to 3B). We reasoned that a transcriptional repressor likely binds to these "complex-VM" enhancers in cardioblasts and blocks the collective activity of pMad, dTCF, Tin, Pnr, and Doc. Sloppy paired (Slp) is a good candidate, as it is expressed in the cardiogenic mesoderm at these stages (Lee and Frasch, 2000) and is required to repress the activity of a VM enhancer in the *bagpipe* locus (*bap3*) in the cardiogenic domain (Lee and Frasch, 2005).

To investigate a potential role of Slp, we performed genome-wide ChIP-on-chip experiments against Slp at the same stages of development as the other TFs and then examined Slp recruitment to the 4,041 Tin-bound CRMs (Table S6). In

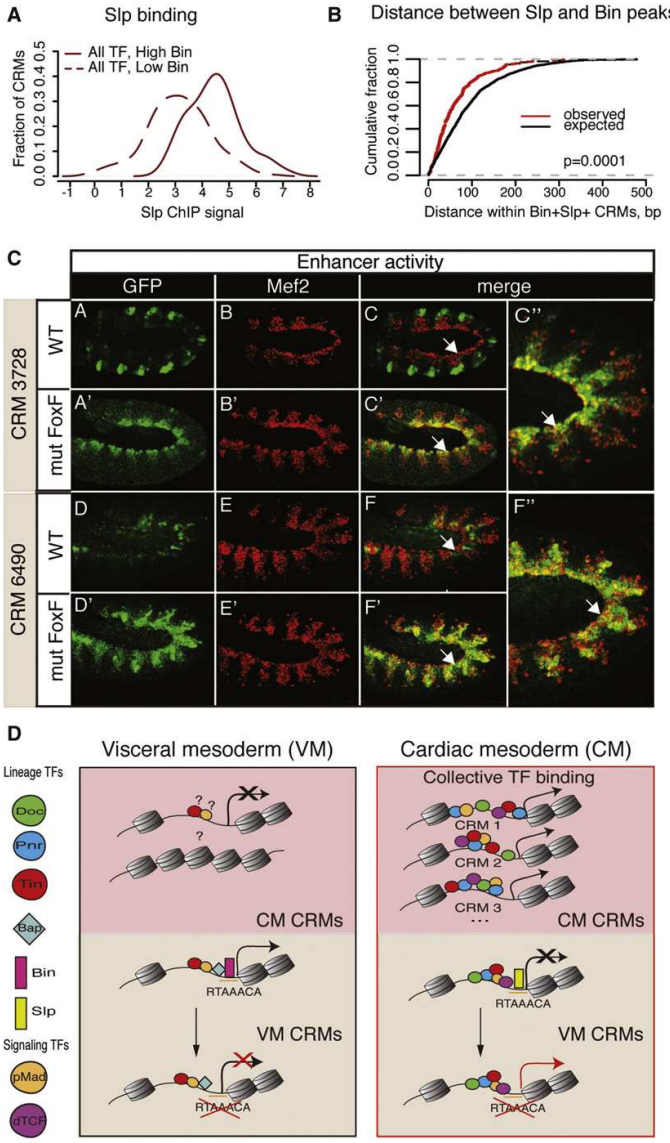


Figure 7. Sloppy Paired Represses the Activity of Dormant TF Binding in Cardiac Cells

(A) The distribution of Sloppy paired (Slp) binding signal at All TF CRMs depending on the levels of Biniou binding. Highest Slp signals were observed at All TF CRMs with high Biniou levels (visceral muscle enhancers) compared to low-Bin All TF CRMs (Wilcoxon test $p = 5.7 \times 10^{-12}$).

(B) Distance between Slp and Bin ChIP peaks within Bin-Slp-Tin-positive CRMs. Cumulative density distributions of observed distances (red) are shifted to the left compared to those expected at random (black), indicating that these peaks nonrandomly localize in proximity to each other. Wilcoxon test p values, p = 0.0001 (observed versus expected).

(C) Mutation of Slp-FoxF motif facilitates enhancer activity in heart and dorsal mesoderm. Immunostaining of transgenic embryos containing the wild-type (WT) and mutant (mut FoxF) enhancers using anti-GFP (enhancer reporter, green) and anti-Mef2 (a mesodermal marker, red) antibodies. Mutated Slp-FoxF sites are shown in Figure S6E. CRM 3728 and 6490 are active in the VM (Figure 6A), but not in CM (A–C and D–F). Mutation of the Slp FoxF sites leads to new activity in CM (A′–C′ and D′–F′, arrow). C″ and F″ are higher magnification images of C′ and F′, respectively.

(D) Proposed model for the regulation of cardiac and visceral mesoderm CRMs in both cell types. VM enhancers (left) contain FoxF motifs that recruit Biniou (Bin) in VM and Slp in cardiac cells, whereas all five heart TFs occupy these enhancers in cardiac cells. Slp counteracts the activity of the cardiogenic TF collective by repressing transcription. In contrast, enhancers that recruit the five heart TFs but lack FoxF motifs drive expression in cardiac cells (right).

See also Figure S6.

addition to the previously described binding on the *bap3* enhancer (Figure S1H), Slp binding is enriched at all enhancers within the All TF class with characterized VM activity (Figure S6A), as well as at those with predicted VM activity based on high levels of Biniou occupancy (VM CRMs) (Figure 7A). Moreover, Slp and Biniou binding peaks nonrandomly localize in close proximity to each other (Figures 7B and S6D) and to the Biniou-FoxF motifs (Figure S6C). Both results suggest that Biniou and Slp are recruited to enhancers via the same motif, globally extending the model of the *bap3* enhancer (Lee and Frasch, 2005). However, in contrast to the *bap3* enhancer, the early VM enhancers (Biniou-high CRMs) identified here are collectively bound by the five cardiogenic TFs, in addition to Slp. This complex binding signature promoted us to ask whether the cardiogenic TFs are capable of activating these enhancers once the repressive influence of Slp is removed. To test this, we mutated the Slp-Biniou FoxF motifs in three of the All TF CRMs that regulate expression in VM (Figure 6A, top three enhancers). In two out of three cases examined, mutation of this site was sufficient to facilitate expression in CM and, interestingly, also in the somatic muscle while attenuating activity in VM (Figures 7C and S6F). These results demonstrate that the "dormant" TF occupancy of cardiac factors has the capacity to direct CM activity. FoxF motifs within these enhancers are therefore used to activate transcription within the VM (mediated by Biniou; Figure S6F) and repress CM activity in the cardiogenic mesoderm (mediated by Slp; Figure 7C). These motifs thereby serve as a "lineage switch," ensuring exclusive enhancer activity in one of the two tissues derived from the dorsal mesoderm (Figure 7D).

DISCUSSION

Dissecting transcriptional networks in the context of embryonic development is inherently difficult due to the multicellularity of the system and the fact that most essential developmental regulators have pleiotropic effects, acting in separate and sometimes interconnected networks. Here, we present a comprehensive systematic dissection of the *cis*-regulatory properties leading to cardiac specification within the context of a developing embryo. The resulting compendium of TF binding signatures, in addition to our extensive in vivo and in vitro analysis of enhancer activity, revealed a number of insights into the regulatory complexity of developmental programs.

Cardiogenic TFs Form a Coherent Functional Module during Cardiac Specification

Nkx, GATA, and T box factors regulate each other's expression in both flies and mice (Lien et al., 1999; Molkentin et al., 2000; Reim and Frasch, 2005; Sun et al., 2004), where they form a recursively wired transcriptional circuit (Figure 1D) that acts cooperatively at a genetic level to regulate heart development across a broad range of organisms. Our data demonstrate that this cooperative regulation extends beyond the ability of these TFs to regulate each other's expression. All five cardiogenic TFs (including dTCF and pMad) converge as a collective unit on a very extensive set of mesodermal enhancer elements in vivo (Tin-bound regions) and also in vitro (in DmD8 cells). Importantly, this TF co-occupancy occurs in *cis*, rather than being mediated via crosslinking of DNA-looping interactions bringing together distant sites. Examining enhancer activity out of context, for example, in transgenic experiments and luciferase assays, revealed that the TF collective activity is preserved in situations in which these regions are removed from their native genomic "looping" context.

In keeping with the conserved essential role of these factors for heart development, the integration of their activity at shared enhancer elements may also be conserved. Recent analyses of the mouse homologs of these TFs (with the exception of the inductive signals from Wg and Dpp signaling) in a cardiomyocyte cell line support this, revealing a signifcant overlap in their binding signatures (He et al., 2011; Schlesinger et al., 2011), although interestingly not in the collective "all-or-none" fashion observed in *Drosophila* embryos. This difference may result from the partial overlap of the TFs examined, interspecies differences, or the inherent differences between the in vivo versus in vitro models. Examining enhancer output for a large number of regions indicates that this collective TF occupancy signature is generally predictive of enhancer activity in cardiac mesoderm or its neighboring cell population, the visceral mesoderm—expression patterns that cannot be obtained from any one of these TFs alone.

TF Collective: Cooperative Enhancer Regulation Using Flexible Sequence Context

There are currently two prevailing models of how enhancers function. The enhanceosome model suggests that TFs bind to enhancers in a cooperative manner directed by a specific arrangement of motifs, often having a very rigid motif grammar (Panne, 2008). An alternative, the billboard model, suggests that each TF (or submodule) is recruited independently via its own sequence motif, and therefore the motif spacing and relative orientation have little importance (Kulkarni and Arnosti, 2003). Our results indicate that cardiogenic TFs are corecruited and activate enhancers in a cooperative manner, but this cooperativity occurs with little or no apparent motif grammar to such an extent that the motifs for some factors do not always need to be present. This is at odds with either the enhanceosome (cooperative binding; rigid grammar) or billboard (independent binding; little grammar) models and represents an alternative mode of enhancer activity, which we term a "TF collective" (cooperative binding; no grammar), and likely constitutes a common principle in other systems.

Our data suggest that the TF collective operates via the cooperative recruitment of a large number of TFs (in this case, at least five), which is mediated by the presence of high-affinity TF motifs for a subset of factors initiating the recruitment of all TFs. The occupancy of any remaining factor(s) is most likely facilitated via protein-protein interactions or cooperativity at a higher level such as, for example, via the chromatin activators CBP/p300, which interact with mammalian GATA and Mad homologs (Dai and Markham, 2001; Feng et al., 1998). This model allows for extensive motif turnover without any obvious effect on enhancer activity, consistent with what has been observed in vivo for the *Drosophila spa* enhancer (Swanson et al., 2010) and mouse heart enhancers (Blow et al., 2010).

Dormant TF Occupancy Reflects the Developmental History of a Cell's Lineage

Integrating the TF occupancy data for all seven major TFs involved in dorsal mesoderm specification (the five cardiogenic factors together with Biniou and Slp) revealed a very striking observation: the developmental history of cardiac cells is reflected in their TF occupancy patterns. VM and CM are both derived from precursor cells within the dorsal mesoderm. Once specified, these cell types express divergent sets of TFs: Slp, activated dTCF, Doc, and Pnr function in cardiac cells, whereas Biniou and Bagpipe are active in the VM (Figures 1A and 7D). Despite these mutually exclusive expression patterns, the cardiogenic TFs are recruited to the same enhancers as VM TFs in the juxtaposed cardiac mesoderm (Figure 7D). Moreover, dependent on the removal of a transcriptional repressor, these combined binding signatures have the capacity to drive expression in either cell type. This finding provides the exciting possibility that dormant TF occupancy could be used to trace the developmental origins of a cell lineage. It also explains why active repression in *cis* is required for correct lineage specification, which is a frequent observation from genetic studies.

At the molecular level, it remains an open question why the VM-specific enhancers are occupied by the cardiac TF collective. We hypothesize that this may occur through chromatin remodeling in the precursor cell population. An "open" (accessible) chromatin state at these loci in dorsal mesoderm cells, which is most likely mediated or maintained by Tin binding prior to specification, could facilitate the occupancy of cell type-specific TFs in both CM and VM cells. Such early "chromatin priming" of regulatory regions active at later stages has been

observed during ES cell differentiation (Liber et al., 2010; Walter et al., 2008). Our data provide evidence that this also holds true for TF occupancy and not just chromatin marks. On a more speculative level, this developmental footprint of TF occupancy may reflect the evolutionary ancestry of these two organs (Pérez-Pomares et al., 2009). Visceral and cardiogenic tissues are derived from the splanchnic mesoderm in both flies and vertebrates. These complex VM-heart enhancers may represent evolutionary relics containing functional binding sites that reflect enhancer activity in an ancestral cell type.

Taken together, the collective TF occupancy on enhancers during dorsal mesoderm specification illustrates how the regulatory input of cooperative TFs is integrated in *cis*, in the absence of any strict motif grammar. We expect this more flexible mode of cooperative *cis* regulation to be present in many other complex developmental systems.

EXPERIMENTAL PROCEDURES

Chromatin Immunoprecipitation

Chromatin immunoprecipitations (ChIPs) were performed as described previously (Sandmann et al., 2006). The following antibodies were used here: rabbit anti-dTCF (M. Bienz), rabbit anti-pMad (C.-H. Heldin), rabbit anti-Doc2 (M. Frasch), and guinea-pig anti-Slp (H. Jackle). The rabbit anti-Pannier serum was generated in this study and raised against amino acids 125–294 and 206–336. The quality of each antibody was assessed by immunostains (data not shown) and western blot (Figure S5), and all ChIP data was integrated with our previously published Tin data, which was based on two independent anti-Tin antibodies. *Doc2* and *Doc3* have almost identical expression patterns and are functionally redundant and are therefore expected to occupy the same sites. Although we used an antibody directed against Doc2, we refer to the data as Doc binding to reflect the redundancy between these TFs. ChIP DNA was amplified and hybridized to Affymetrix GeneChip *Drosophila* high-density Tiling array1.0R. ChIP of endogenous loci in DmD8 cells was performed using a similar protocol 4 days posttransfection of pRM-Tin and 1 day postincubation with Wg+Dpp-conditioned medium (Figure S5B); signal was detected by real-time PCR. See Extended Experimental Procedures for more details.

Defining TF Binding Events and ChIP-Defined CRMs

Quantile normalization (Bolstad et al., 2003) was applied to the four data sets for each TF (two ChIP experiments and two mock controls) for each of the 14 conditions (seven TFs at two time points). High-confidence binding events (shown in Tables S1 and S7) were defined using TileMap (Ji and Wong, 2005). CRMs (listed in Table S2) were defined as neighboring clusters of high-confidence TF binding peaks, as described previously (Zinzen et al., 2009). Slp and Bin signals at CRMs are shown in Table S6. All ChIP data are available in ArrayExpress with accession number E-TABM-1184 and on the Furlong lab web page. See Extended Experimental Procedures for greater detail.

Autoclass Clustering of TF Binding Signals

Clustering was performed using Autoclass-C (Cheeseman, 1996) based on maximum moving average probe-wise ChIP signals (Wilczyński and Furlong, 2010) for each TF/time per CRM (window size = 200 bp). The results were filtered to exclude CRMs with maximum posterior probabilities of cluster assignment less than 0.5 and/or probabilities of best and second-best cluster assignment differing by less than 2-fold. See Table S4 for the list of classified CRMs. More details in Extended Experimental Procedures.

Transgenic Reporter Assays

CRM activity was assayed using transgenic reporter assays by placing the ChIP-defined genomic region upstream of a minimal promoter driving a *GFP* reporter gene in a modified version of pDuo2n-attB (Zinzen et al., 2009); see Extended Experimental Procedures. All constructs were targeted to chromosomal arm 3L via attB/phiC31-mediated integration (Bischof et al., 2007). Transgenic lines were balanced, homozygosed, and tested by double-fluorescent in situ hybridization using probes directed against the *GFP* reporter gene (green) and *tin* (red). CRM activity in dorsal mesoderm, cardiac mesoderm, or visceral mesoderm is readily apparent via the coexpression of *GFP* and *tin* at specific developmental stages. Images were taken using a Zeiss LSM510meta confocal microscope and were processed in Adobe Photoshop. Results are listed in Table S5 (results of double-fluorescent in situ hybridizations for selected endogenous genes are summarized in Table S3).

Motif Analysis

De novo motif discovery was performed using Weeder (Pavesi et al., 2004) on 400 bp regions surrounding the positions of the 100 highest-scoring TileMap peaks for each data set (defined as described above) and RSAT (Thomas-Chollier et al., 2008) on CRMs of the All TFs class. Motif scanning was performed using Patser (Hertz and Stormo, 1999), applying thresholds defined on the basis of specificity-sensitivity criteria (data not shown). See Extended Experimental Procedures for details and additional analyses.

ACCESSION NUMBERS

Data have been deposited under ArrayExpress accession number E-MTAB-1184.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, six figures, and seven tables and can be found with this article online at doi:10.1016/j.cell.2012.01.030.

ACKNOWLEDGMENTS

We are extremely grateful to M. Bienz, C.-H. Heldin, M. Frasch, and H. Jackle for antibodies. This work was technically supported by the EMBL Genomics Core facility, with specific thanks to Jos de Graaf for array hybridizations. We thank all members of the Furlong lab for discussions and comments on the manuscript, Stijn Van Dongen for help with assessing the robustness of clustering methods, and Thomas Sandmann for designing dsRNA probes. This work was supported by a Deutsche Forschungsgemeinschaft (DFG FU 750/1) grant and Human Frontier Science Program (HFSP) grant to E.E.M.F. and postdoctoral fellowships to G.J. from EMBO and to M.S. from the EMBL EIPOD program.

Received: August 17, 2010
Revised: August 16, 2011
Accepted: January 17, 2012
Published: February 2, 2012

REFERENCES

Azpiazu, N., and Frasch, M. (1993). tinman and bagpipe: two homeo box genes that determine cell fates in the dorsal mesoderm of Drosophila. Genes Dev. *7*(7B), 1325–1340.

Bischof, J., Maeda, R.K., Hediger, M., Karch, F., and Basler, K. (2007). An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases. Proc. Natl. Acad. Sci. USA *104*, 3312–3317.

Blow, M.J., McCulley, D.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2010). ChIP-Seq identification of weakly conserved heart enhancers. Nat. Genet. *42*, 806–810.

Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185–193.

Brown, C.O., III, Chi, X., Garcia-Gras, E., Shirai, M., Feng, X.H., and Schwartz, R.J. (2004). The cardiac determination factor, Nkx2-5, is activated by mutual

cofactors GATA-4 and Smad1/4 via a novel upstream enhancer. J. Biol. Chem. *279*, 10659–10669.

Brown, C.D., Johnson, D.S., and Sidow, A. (2007). Functional architecture and evolution of transcriptional elements that drive gene coexpression. Science *317*, 1557–1560.

Bruneau, B.G., Nemer, G., Schmitt, J.P., Charron, F., Robitaille, L., Caron, S., Conner, D.A., Gessler, M., Nemer, M., Seidman, C.E., and Seidman, J.G. (2001). A murine model of Holt-Oram syndrome defines roles of the T-box transcription factor Tbx5 in cardiogenesis and disease. Cell *106*, 709–721.

Campos-Ortega, J.A. (1997). The Embryonic Development of Drosophila Melanogaster (New York: Springer-Verlag).

Cheeseman, P. (1996). Bayesian Classification (AutoClass): Theory and Results. In Advances in Knowledge Discovery and Data Mining, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds. (Cambridge, MA: AAAI Press/MIT Press).

Cherbas, L., Willingham, A., Zhang, D., Yang, L., Zou, Y., Eads, B.D., Carlson, J.W., Landolin, J.M., Kapranov, P., Dumais, J., et al. (2011). The transcriptional diversity of 25 Drosophila cell lines. Genome Res. *21*, 301–314.

Cripps, R.M., and Olson, E.N. (2002). Control of cardiac development by an evolutionarily conserved transcriptional network. Dev. Biol. *246*, 14–28.

Dai, Y.S., and Markham, B.E. (2001). p300 Functions as a coactivator of transcription factor GATA-4. J. Biol. Chem. *276*, 37178–37185.

Durocher, D., Charron, F., Warren, R., Schwartz, R.J., and Nemer, M. (1997). The cardiac transcription factors Nkx2-5 and GATA-4 are mutual cofactors. EMBO J. *16*, 5687–5696.

Feng, X.H., Zhang, Y., Wu, R.Y., and Derynck, R. (1998). The tumor suppressor Smad4/DPC4 and transcriptional adaptor CBP/p300 are coactivators for smad3 in TGF-beta-induced transcriptional activation. Genes Dev. *12*, 2153–2163.

Frasch, M. (1999). Intersecting signalling and transcriptional pathways in Drosophila heart specification. Semin. Cell Dev. Biol. *10*, 61–71.

Gajewski, K., Zhang, Q., Choi, C.Y., Fossett, N., Dang, A., Kim, Y.H., Kim, Y., and Schulz, R.A. (2001). Pannier is a transcriptional target and partner of Tinman during Drosophila cardiogenesis. Dev. Biol. *233*, 425–436.

Garg, V., Kathiriya, I.S., Barnes, R., Schluterman, M.K., King, I.N., Butler, C.A., Rothrock, C.R., Eapen, R.S., Hirayama-Yamada, K., Joo, K., et al. (2003). GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. Nature *424*, 443–447.

Halfon, M.S., Carmena, A., Gisselbrecht, S., Sackerson, C.M., Jiménez, F., Baylies, M.K., and Michelson, A.M. (2000). Ras pathway specificity is determined by the integration of multiple signal-activated and tissue-restricted transcription factors. Cell *103*, 63–74.

He, A., Kong, S.W., Ma, Q., and Pu, W.T. (2011). Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. Proc. Natl. Acad. Sci. USA *108*, 5632–5637.

Hertz, G.Z., and Stormo, G.D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics *15*, 563–577.

Ieda, M., Fu, J.D., Delgado-Olguin, P., Vedantham, V., Hayashi, Y., Bruneau, B.G., and Srivastava, D. (2010). Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. Cell *142*, 375–386.

Jakobsen, J.S., Braun, M., Astorga, J., Gustafson, E.H., Sandmann, T., Karzynski, M., Carlsson, P., and Furlong, E.E. (2007). Temporal ChIP-on-chip reveals Biniou as a universal regulator of the visceral muscle transcriptional network. Genes Dev. *21*, 2448–2460.

Ji, H., and Wong, W.H. (2005). TileMap: create chromosomal map of tiling array hybridizations. Bioinformatics *21*, 3629–3636.

Kelly, R.G., and Buckingham, M.E. (2002). The anterior heart-forming field: voyage to the arterial pole of the heart. Trends Genet. *18*, 210–216.

Kulkarni, M.M., and Arnosti, D.N. (2003). Information display by transcriptional enhancers. Development *130*, 6569–6575.

Lee, H.H., and Frasch, M. (2000). Wingless effects mesoderm patterning and ectoderm segmentation events via induction of its downstream target sloppy paired. Development *127*, 5497–5508.

Lee, H.H., and Frasch, M. (2005). Nuclear integration of positive Dpp signals, antagonistic Wg inputs and mesodermal competence factors during Drosophila visceral mesoderm induction. Development *132*, 1429–1442.

Lee, H.H., Zaffran, S., and Frasch, M. (2006). In Development of the Larval Visceral Musculature, H. Sink, ed. (New York: Springer).

Liber, D., Domaschenz, R., Holmqvist, P.H., Mazzarella, L., Georgiou, A., Leleu, M., Fisher, A.G., Labosky, P.A., and Dillon, N. (2010). Epigenetic priming of a pre-B cell-specific enhancer through binding of Sox2 and Foxd3 at the ESC stage. Cell Stem Cell *7*, 114–126.

Lien, C.L., Wu, C., Mercer, B., Webb, R., Richardson, J.A., and Olson, E.N. (1999). Control of early cardiac-specific transcription of Nkx2-5 by a GATA-dependent enhancer. Development *126*, 75–84.

Liu, Y.H., Jakobsen, J.S., Valentin, G., Amarantos, I., Gilmour, D.T., and Furlong, E.E. (2009). A systematic analysis of Tinman function reveals Eya and JAK-STAT signaling as essential regulators of muscle development. Dev. Cell *16*, 280–291.

Lockwood, W.K., and Bodmer, R. (2002). The patterns of wingless, decapentaplegic, and tinman position the Drosophila heart. Mech. Dev. *114*, 13–26.

Molkentin, J.D., Antos, C., Mercer, B., Taigen, T., Miano, J.M., and Olson, E.N. (2000). Direct activation of a GATA6 cardiac enhancer by Nkx2.5: evidence for a reinforcing regulatory network of Nkx2.5 and GATA transcription factors in the developing heart. Dev. Biol. *217*, 301–309.

Nishita, M., Hashimoto, M.K., Ogata, S., Laurent, M.N., Ueno, N., Shibuya, H., and Cho, K.W. (2000). Interaction between Wnt and TGF-beta signalling pathways during formation of Spemann's organizer. Nature *403*, 781–785.

Olson, E.N. (2006). Gene regulatory networks in the evolution and development of the heart. Science *313*, 1922–1927.

Panne, D. (2008). The enhanceosome. Curr. Opin. Struct. Biol. *18*, 236–242.

Pavesi, G., Mereghetti, P., Mauri, G., and Pesole, G. (2004). Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res. *32*(Web Server issue), W199–W203.

Pérez-Pomares, J.M., González-Rosa, J.M., and Muñoz-Chápuli, R. (2009). Building the vertebrate heart - an evolutionary approach to cardiac development. Int. J. Dev. Biol. *53*, 1427–1443.

Reim, I., and Frasch, M. (2005). The Dorsocross T-box genes are key components of the regulatory network controlling early cardiogenesis in Drosophila. Development *132*, 4911–4925.

Roider, H.G., Kanhere, A., Manke, T., and Vingron, M. (2007). Predicting transcription factor affinities to DNA from a biophysical model. Bioinformatics *23*, 134–141.

Sandmann, T., Jakobsen, J.S., and Furlong, E.E. (2006). ChIP-on-chip protocol for genome-wide analysis of transcription factor binding in Drosophila melanogaster embryos. Nat. Protoc. *1*, 2839–2855.

Schlesinger, J., Schueler, M., Grunert, M., Fischer, J.J., Zhang, Q., Krueger, T., Lange, M., Tönjes, M., Dunkel, I., and Sperling, S.R. (2011). The cardiac transcription network modulated by Gata4, Mef2a, Nkx2.5, Srf, histone modifications, and microRNAs. PLoS Genet. *7*, e1001313.

Sepulveda, J.L., Belaguli, N., Nigam, V., Chen, C.Y., Nemer, M., and Schwartz, R.J. (1998). GATA-4 and Nkx-2.5 coactivate Nkx-2 DNA binding targets: role for regulating early cardiac gene expression. Mol. Cell. Biol. *18*, 3405–3415.

Senger, K., Armstrong, G.W., Rowell, W.J., Kwan, J.M., Markstein, M., and Levine, M. (2004). Immunity regulatory DNAs share common organizational features in Drosophila. Mol. Cell *13*, 19–32.

Sun, G., Lewis, L.E., Huang, X., Nguyen, Q., Price, C., and Huang, T. (2004). TBX5, a gene mutated in Holt-Oram syndrome, is regulated through a GC box and T-box binding elements (TBEs). J. Cell. Biochem. *92*, 189–199.

Swanson, C.I., Evans, N.C., and Barolo, S. (2010). Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. Dev. Cell *18*, 359–370.

Takeuchi, J.K., and Bruneau, B.G. (2009). Directed transdifferentiation of mouse mesoderm to heart tissue by defined factors. Nature *459*, 708–711.

Thomas-Chollier, M., Sand, O., Turatsinze, J.V., Janky, R., Defrance, M., Vervisch, E., Brohee, S., and van Helden, J. (2008). RSAT: Regulatory sequence analysis tools. Nucleic Acids Res *36*, W119–W127.

Walter, K., Bonifer, C., and Tagoh, H. (2008). Stem cell-specific epigenetic priming and B cell-specific transcriptional activation at the mouse Cd19 locus. Blood *112*, 1673–1682.

Wilczyński, B., and Furlong, E.E. (2010). Dynamic CRM occupancy reflects a temporal map of developmental progression. Mol. Syst. Biol. *6*, 383.

Xu, X., Yin, Z., Hudson, J.B., Ferguson, E.L., and Frasch, M. (1998). Smad proteins act in combination with synergistic and antagonistic regulators to target Dpp responses to the Drosophila mesoderm. Genes Dev. *12*, 2354–2370.

Zaffran, S., and Frasch, M. (2002). Early signals in cardiac development. Circ. Res. *91*, 457–469.

Zaffran, S., Küchler, A., Lee, H.H., and Frasch, M. (2001). biniou (FoxF), a central component in a regulatory network controlling visceral mesoderm development and midgut morphogenesis in Drosophila. Genes Dev. *15*, 2900–2915.

Zaffran, S., Xu, X., Lo, P.C., Lee, H.H., and Frasch, M. (2002). Cardiogenesis in the Drosophila model: control mechanisms during early induction and diversification of cardiac progenitors. Cold Spring Harb. Symp. Quant. Biol. *67*, 1–12.

Zinzen, R.P., Girardot, C., Gagneur, J., Braun, M., and Furlong, E.E. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. Nature *462*, 65–70.

3.2.4 Discussion

A very common analysis performed on ChIP-chip (and ChIP-seq) data is *de novo* motif discovery to identify the binding characteristics of the ChIPed TF and predict potential co-factors. When a binding model exists for the ChIPed TF, motif discovery can be used to assess the quality of called peaks, and therefore of the ChIP assay itself, by comparing the identified motif(s) with the known one(s). The results of such analyses are not always straightforward to interpret and might even uncover unexpected features.

The sequence analysis of the Pnr binding peaks performed in this study is a very good illustration of this point. Independently of the algorithm used (Weeder, RSAT, MEME), the motif TATCGATA (named Pnr* in supplementary Figure S4A of the enclosed *Cell* paper, annexe 3) was consistently reported, while the expected Pnr signature (GATAag) was not. Although this Pnr* motif contains a GATA subsequence, it only partially matches the expected GATAag Pnr motif, which was only uncovered by comparing the "All TF" CRMs with the "Pnr+Tin" CRMs using the RSAT oligo-diff tool. Interestingly, TATCGATA perfectly matches the signature of both the insulator protein BEAF-32 and the core promoter motif DRE binding factor (DREF) [1]. Of note, BEAF-32 only binds the CGATA core[177], the TATCGATA palindrome therefore potentially holds two overlapping BEAF-32 TFBSs. Guillaume Junion in the Furlong Lab performed whole embryo ChIP-chip against BEAF-32 at 4-6h and 6-8h AEL (unpublished data). Using BEAF-32 and Pnr peaks (400 bp regions centred on the peak summit), Zhen Xuan Yeo and I could verify that these TFs extensively bind to the same regions, with 60% of the Pnr peaks and 75% of BEAF-32 peaks co-localizing (z-score>123 using the genome correction structure statistics[97]). Moreover, 80% of the regions bound by BEAF-32 and Pnr are found within 200 bp of a TSS, suggesting that BEAF-32 and Pnr co-localize at DRE core promoters, thereby raising questions about the functional role of Pnr in this particular context. Although this should be confirmed, the presence of the GATA core in TATCGATA suggests that Pnr most likely directly binds these sites. It is therefore important to determine whether these factors co-localize at these regions in the same cells (BEAF-32 is ubiquitous and Pnr is broadly expressed in the embryo), and, if so, whether they both contact the DNA (potentially in the form of a protein complex). Alternatively, these factors may have antagonistic binding effects at these TSSs and the

observed ChIP signal would reflect binding occurring in distinct cells. Altogether, these results are very exciting and potentially point to new roles of both Pnr and BEAF-32.

The Pnr example clearly shows that the known functional motif is not necessarily the most significant one returned by motif discovery algorithms. In addition, it is not infrequent (at least in my experience) that motifs discovered in ChIP peaks only partially match known motifs, with cases where the newly discovered motif display either higher or lower information content compared to the known motif (for example pMad and Tin, respectively, as shown on Figure S4 of the enclosed *Cell* paper, annexe 3). In such situations, the recurring question is whether the observed difference is 'acceptable'.

Historically, motifs were build with only a handful of experimentally determined footprints typically generated from *in vitro* experiments using purified protein and naked DNA (for example the Bap PWM in FlyReg – now REDfly[25]-, is based on only three sites) and might therefore be a biased representation of the TF binding specificity. Even when more footprints are available, these footprints were often discovered in a particular biological context or using an *in vitro* assay, and the resulting PWM might still be biased. New experimental methods such as bacterial-1-hybrid[148], protein-binding microarrays[149], SELEX-seq[178,179], MITOMI[151] are currently used to determine TF binding preferences at much larger scales. Importantly, these are *in vitro* approaches and the resulting binding preferences might therefore be biased for multiple reasons[180], such as the lack of post-translational modifications (eukaryotic proteins produced in bacteria), the use of the DBD only instead of the full-length protein, or yet the fact that these assays are conducted out of relevant biological context (absence of co-factors and proper chromatin environment). Binding models obtained using these techniques are nevertheless very useful and are in general in good agreement with existing data[179]. In contrast, ChIP studies provide hundreds to thousands of sites from *in vivo* binding and can therefore much more accurately reflect TF binding preferences. It is therefore not surprising to observe differences between discovered and known models, and determining whether the observed difference is 'acceptable' is a case-by-case decision.

Altogether, these observations have several technical and methodological implications. First, assessing ChIP assay quality using match enrichment of a known PWM in called peaks might underestimate the actual assay quality, as the used PWM might be accurate for a fraction of the ChIP peaks only. In addition to the necessary technical quality assessment of

the experiment in the form of diagnostic plots (for example using the R package arrayQualityMetrics[181] for microarrays), it is advisable, in my opinion, to evaluate the ChIP assay quality and specificity by extensive visualization (to acquire a human opinion of the signal-to-noise ratio and of the called peaks) and by looking, for example, at the recall of known enhancers/TFBSs, or at the enrichment for biologically relevant genes. Second, using the PWM enrichment metric to determine the threshold for calling peaks or post filtering called peaks (considering peaks lacking a good PWM match as false positives) is, in my opinion, not recommended and might even hide important aspects of the studied TF biology. This study actually provides a concrete example where TF presence at enhancers occurs in the absence of *bona fide* TFBSs for all TFs. Third, in the case where the motif of the ChIP'ed TF is unknown, results of *de novo* discovery should be interpreted cautiously and should be further experimentally validated, for example by gel retardation assays or SELEX. Finally, TFs may have different binding specificities depending on the functional context, and this might be reflected in the motif discovery results[13,182]. It might then be advisable to consider multiple PWMs in subsequent analyses. Discovering different PWMs might further reflect the presence of collaborating factors.

Recent technological developments at both the experimental and computational levels may help to disentangle such situations. On the experimental side, ChIP-seq has proven to be more sensitive (reduction of noise) and of higher spatial resolution (narrower peaks) than ChIP-chip[161,183,184]. More precise peak identification greatly facilitates subsequent computational analyses by reducing the search space (and thus the noise). Recently, Rhee *et al.* claimed single bp accuracy with their new ChIP-exo method, a modified ChIP-seq protocol that includes exonuclease digestion mediated trimming of ChIP'ed DNA to the cross-linked protein of interest[4].

On the computational side, novel *de novo* discovery tools have been developed to specifically deal with large sequence sets produced by ChIP-seq (and ChIP-chip) [152-154]. In particular RSAT peak-motifs[154] identifies k-mers with significant positional bias (on top of usual overrepresentation analysis) and systematically builds motif position profiles anchored on the peak summits allowing to readily distinguish between different affinities of the ChIPed TF (motif enrichment profile centred on peak summits) and signatures of collaborating factors (motif enrichment profile uniformly or symmetrically distributed around peak summits).

4 Conclusion and perspectives

We previously reported the integration of ChIP-chip datasets for five key mesoderm-specific TFs and showed that their combination is sufficient to predict spatio-temporal activity of the enhancers[36]. Here, we investigated how five TFs essential for cardiac development operate in *cis* in the dorsal mesoderm. Although only one of these TFs is specifically expressed in this tissue, we could demonstrate that these TFs are recruited as a "TF collective" at cardiac enhancers, and that this occurs in absence of strong sequence requirements, suggesting a novel model for enhancer activity, alternative to the billboard and enhanceosome models.

We further characterised a novel property of developmental enhancers, whereby dormant TF binding signatures can reflect a developmental footprint of a cell lineage: ''Cardiac'' TFs occupy enhancer elements that are active in the neighbouring VM, echoing the fact that both cell populations are derived from the dorsal mesoderm.

Finally, we demonstrated the power of the BiTS-ChIP-seq assay, a tissue-specific ChIP protocol relying on FACS sorting of nuclei followed by deep sequencing. We applied this protocol to map histone PTMs and Pol II occupancy in the mesoderm of the developing *Drosophila* embryo, and subsequently characterised *in vivo* epigenetic marking of chromatin at active enhancers. We showed that active enhancers are enriched for H3K27Ac, H3K79me3 and Pol II, and that the presence and shape of these marks dynamically correlates with enhancer activity timing and nucleosome positioning. Finally, using a machine learning approach, we predicted novel enhancers presumably active in the mesoderm at a specific developmental stage based on histone PTMs and Pol II occupancy, and successfully validate 89% of them.

BiTS-ChIP-seq opens new avenues in genome biology research within developing embryos with the possibility to probe genome-wide occupancy of ubiquitous factors in specific tissues. The acquisition of cell-specific occupancy maps for histone PTMs, Pol II and general factors, e.g. the activator p300 (CBP) or the repressor CTBP, as well as for nucleosome depleted regions (DHS and FAIRE), will facilitate novel biological questions being addressed. Doing this over the course of development will shed light on the genomic

regulatory mechanisms and organisational principles underlying cell specification and tissue differentiation.

In this respect, we showed that chromatin state can be used to find active regions that are frequently large enough to contain multiple CRMs. The integration of features derived from these different maps (signal level, signal shape, signal combination) will lead to more accurate location of active enhancers, including within genes. In addition, subsequent sequence analysis of these enhancers using new analysis pipelines, such as RSAT peak-motifs[154], will help uncover the signature of novel TFs. This approach is especially interesting to investigate developmental networks for which little is known, in particular when the underlying major TFs or important co-factors have not yet been identified. Furthermore, the combination of maps acquired at successive developmental stages will allow the deciphering of the different sequential steps of enhancer activation, such as which TF or histone marks prime enhancers for activity, which are hallmarks of active enhancers, how long do these marks remain once enhancers become inactive, and which are indicative of regulatory features in a repressed state. Similar questions apply to genes.

A long-standing problem associated with genome-wide enhancer mapping is the determination of enhancers' target genes. The most common strategy is to consider the closest gene (i.e., the closest proximal TSS). This approach has a number of limitations: enhancers can be located hundreds of thousands of bases from their target genes, are able to 'jump' over intervening genes, and can be located within other genes. Moreover, in small genomes such as *Drosophila*, the gene density is high, and associating intergenic enhancers to the closest TSS might result in incorrect assignments. Importantly, it is also unclear how many genes an enhancer may actually target (simultaneously or not). More advanced methods to estimate the likelihood of an enhancer to target a gene (or vice versa) thus need to be developed. Here again, the determination of tissue specific time series of maps mentioned above provides exciting possibilities to tackle this issue, for example by correlating enhancer activity profiles with putative target gene activity profiles.

PWMs are commonly used to represent TF binding specificities. Although this representation has a number of advantages (e.g. compact representation, easy to compute and further use to predict TFBSs in a probabilistic way), it does not incorporate potential position interdependencies, a phenomenon that is far from anecdotal[182]. Genome-wide ChIP assays

against sequence-specific TFs commonly provide thousands of bound regions, thereby enabling the creation of more complex probabilistic models to represent TF binding specificity (for example HMMs), potentially accounting for such positional interdependencies. With the recent development of SELEX-seq[179] (together with other existing *in vitro* methods like protein-binding microarrays[149]) and the increase in throughput of sequencing technologies (see below), the set of available TF models will grow rapidly. Together with tissue-specific chromatin and nucleosomes density maps, these new TF binding models should provide much more accurate TFBS predictions (as already shown by combining PWMs with different chromatin data and sequence conservation[185,186]) and thereby facilitate the deciphering of enhancer organisation.

In only five years, the throughput of 'next generation' sequencers has increased from a few million reads (first Illumina Genome Analyzer®) to ~150 millions reads (Illumina HiSeq® 2000 platform) per lane, and it is now common to sequence multiple samples in the same lane, making the technology much more affordable. With the development of cheaper bench sequencers, such as 454 GS Junior (Roche), MiSeq (Illumina) and Ion Torrent PGM (Life Technologies), high throughput sequencing will become routine[187]. Over the coming years, ChIP-seq and RNA-seq assays will generate data at an ever increasing pace. Although exciting, this prospect raises a number of issues. First, data storage has become a bottleneck, and storage cost will soon exceed the cost of sample sequencing. It is therefore important to develop robust data management strategies, where samples can be properly stored and described together with the raw sequencing results (this is particularly important in case of multiplexing, where a result file contains sequences of several samples). Next, accessing raw and processed data using, for example, a genome browser necessitates efficient strategies for mixing random file access (of indexed files) and traditional relational database storage[188]. Finally, adequate and friendly analysis pipelines need to be developed to enable experimentalists to analyse their results themselves. In this respect, tools such as Galaxy[189] are promising as they ease the development of workflows by bioinformaticians, which can then be used by experimentalists to run analyses on computer farms or using "computer clouds".

References

1. Ohler, U., Liao, G.-C., Niemann, H. & Rubin, G. M. Computational analysis of core promoters in the Drosophila genome. *Genome Biol.* **3**, RESEARCH0087 (2002).
2. Wray, G. A. *et al.* The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20**, 1377–1419 (2003).
3. Bulger, M. & Groudine, M. Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev. Biol.* **339**, 250–257 (2010).
4. Rhee, H. S. & Pugh, B. F. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell* **147**, 1408–1419 (2011).
5. Bryne, J. C. *et al.* JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* **36**, D102–6 (2008).
6. Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23 (2000).
7. D'haeseleer, P. What are DNA sequence motifs? *Nat Biotechnol* **24**, 423–425 (2006).
8. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
9. Zaffran, S., Xu, X., Lo, P. C., Lee, H. H. & Frasch, M. Cardiogenesis in the Drosophila model: control mechanisms during early induction and diversification of cardiac progenitors. *Cold Spring Harb. Symp. Quant. Biol.* **67**, 1–12 (2002).
10. Gajewski, K. *et al.* Pannier is a transcriptional target and partner of Tinman during Drosophila cardiogenesis. *Dev. Biol.* **233**, 425–436 (2001).
11. Garg, V. *et al.* GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. *Nature* **424**, 443–447 (2003).
12. Bruneau, B. G. *et al.* A murine model of Holt-Oram syndrome defines roles of the T-box transcription factor Tbx5 in cardiogenesis and disease. *Cell* **106**, 709–721 (2001).
13. Slattery, M. *et al.* Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox Proteins. *Cell* **147**, 1270–1282 (2011).
14. Castanon, I., Stetina, Von, S., Kass, J. & Baylies, M. K. Dimerization partners determine the activity of the Twist bHLH protein during Drosophila mesoderm development. *Development* **128**, 3145–3159 (2001).
15. Sandmann, T. *et al.* A core transcriptional network for early mesoderm development in Drosophila melanogaster. *Genes Dev.* **21**, 436–449 (2007).
16. Lupien, M. *et al.* FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-Specific Transcription. *Cell* **132**, 958–970 (2008).
17. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
18. Arnone, M. I. & Davidson, E. H. The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124**, 1851–1864 (1997).
19. Atchison, M. L. Enhancers: mechanisms of action and cell specificity. *Annu. Rev. Cell Biol.* **4**, 127–153 (1988).
20. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is

enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).
21. Dynan, W. S. Modularity in promoters and enhancers. *Cell* **58**, 1–4 (1989).
22. Ip, Y. T., Kraut, R., Levine, M. & Rushlow, C. A. The dorsal morphogen is a sequence-specific DNA-binding protein that interacts with a long-range repression element in Drosophila. *Cell* **64**, 439–446 (1991).
23. Small, S., Blair, A. & Levine, M. Regulation of two pair-rule stripes by a single enhancer in the Drosophila embryo. *Dev. Biol.* **175**, 314–324 (1996).
24. Wilczyński, B. & Furlong, E. E. M. Dynamic CRM occupancy reflects a temporal map of developmental progression. *Mol Syst Biol* **6**, – (2010).
25. Gallo, S. M. *et al.* REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. *Nucleic Acids Res.* **39**, D118–23 (2011).
26. Yin, Z., Xu, X. L. & Frasch, M. Regulation of the twist target gene tinman by modular cis-regulatory elements during early mesoderm development. *Development* **124**, 4971–4982 (1997).
27. Panne, D. The enhanceosome. *Curr. Opin. Struct. Biol.* **18**, 236–242 (2008).
28. Maniatis, T. *et al.* Structure and function of the interferon-beta enhanceosome. *Cold Spring Harb. Symp. Quant. Biol.* **63**, 609–620 (1998).
29. Honda, K. *et al.* IRF-7 is the master regulator of type-I interferon-dependent immune responses. *Nature* **434**, 772–777 (2005).
30. Escalante, C. R., Nistal-Villán, E., Shen, L., García-Sastre, A. & Aggarwal, A. K. Structure of IRF-3 Bound to the PRDIII-I Regulatory Element of the Human Interferon-β Enhancer. *Mol Cell* **26**, 703–716 (2007).
31. Panne, D., Maniatis, T. & Harrison, S. C. Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon-beta enhancer. *The EMBO Journal* **23**, 4384–4393 (2004).
32. Panne, D., Maniatis, T. & Harrison, S. C. An atomic model of the interferon-beta enhanceosome. *Cell* **129**, 1111–1123 (2007).
33. Senger, K. *et al.* Immunity regulatory DNAs share common organizational features in Drosophila. *Mol Cell* **13**, 19–32 (2004).
34. Kulkarni, M. M. & Arnosti, D. N. Information display by transcriptional enhancers. *Development* **130**, 6569–6575 (2003).
35. Brown, C. D., Johnson, D. S. & Sidow, A. Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science* **317**, 1557–1560 (2007).
36. Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M. & Furlong, E. E. M. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* **462**, 65–70 (2009).
37. Davidson, E. H. & Erwin, D. H. Gene regulatory networks and the evolution of animal body plans. *Science* **311**, 796–800 (2006).
38. Bonn, S. & Furlong, E. E. M. cis-Regulatory networks during development: a view of Drosophila. *Curr. Opin. Genet. Dev.* **18**, 513–520 (2008).
39. Borok, M. J., Tran, D. A., Ho, M. C. W. & Drewell, R. A. Dissecting the regulatory switches of development: lessons from enhancer evolution in Drosophila. *Development* **137**, 5–13 (2010).
40. Visel, A. *et al.* Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.* **40**, 158–160 (2008).
41. Blow, M. J. *et al.* ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.* **42**, 806–810 (2010).

42. Fisher, S. Conservation of RET Regulatory Function from Human to Zebrafish Without Sequence Similarity. *Science* **312**, 276–279 (2006).
43. Hare, E. E., Peterson, B. K., Iyer, V. N., Meier, R. & Eisen, M. B. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS Genet* **4**, e1000106 (2008).
44. Ho, M. C. W. *et al.* Functional Evolution of cis-Regulatory Modules at a Homeotic Gene in Drosophila. *PLoS Genet* **5**, e1000709 (2009).
45. Parker, S. C. J., Hansen, L., Abaan, H. O., Tullius, T. D. & Margulies, E. H. Local DNA topography correlates with functional noncoding regions of the human genome. *Science* **324**, 389–392 (2009).
46. Greenbaum, J. A., Pang, B. & Tullius, T. D. Construction of a genome-scale structural map at single-nucleotide resolution. *Genome Res.* **17**, 947–953 (2007).
47. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251–260 (1997).
48. Biswas, M., Voltz, K., Smith, J. C. & Langowski, J. Role of Histone Tails in Structural Stability of the Nucleosome. *PLoS Computational Biology* **7**, e1002279 (2011).
49. Turner, B. M. Histone acetylation and an epigenetic code. *Bioessays* **22**, 836–845 (2000).
50. Li, G. & Reinberg, D. Chromatin higher-order structures and gene regulation. *Curr. Opin. Genet. Dev.* **21**, 175–186 (2011).
51. Campos, E. I. & Reinberg, D. New chaps in the histone chaperone arena. *Genes Dev.* **24**, 1334–1338 (2010).
52. Mito, Y., Henikoff, J. G. & Henikoff, S. Genome-scale profiling of histone H3.3 replacement patterns. *Nat. Genet.* **37**, 1090–1097 (2005).
53. Ni, T. *et al.* A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat. Methods* **7**, 521–527 (2010).
54. Rach, E. A. *et al.* Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS Genet* **7**, e1001274 (2011).
55. Schoenfelder, S. *et al.* Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat. Genet.* **42**, 53–61 (2010).
56. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948–951 (2008).
57. Wen, B., Wu, H., Shinkai, Y., Irizarry, R. A. & Feinberg, A. P. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nat. Genet.* **41**, 246–250 (2009).
58. Finlan, L. E. *et al.* Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS Genet* **4**, e1000039 (2008).
59. Reddy, K. L., Zullo, J. M., Bertolino, E. & Singh, H. Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature* **452**, 243–247 (2008).
60. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).
61. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074–1080 (2001).
62. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human

Genome. *Cell* **129**, 823–837 (2007).
63. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.* **40**, 897–903 (2008).
64. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
65. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
66. Zhou, V. W., Goren, A. & Bernstein, B. E. Charting histone modifications and the functional organization of mammalian genomes. *Nature Publishing Group* **12**, 7–18 (2010).
67. Cheutin, T. & Cavalli, G. Progressive Polycomb Assembly on H3K27me3 Compartments Generates Polycomb Bodies with Developmentally Regulated Motion. *PLoS Genet* **8**, e1002465 (2012).
68. Hawkins, R. D. *et al.* Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**, 479–491 (2010).
69. Lafos, M. *et al.* Dynamic Regulation of H3K27 Trimethylation during Arabidopsis Differentiation. *PLoS Genet* **7**, e1002040 (2011).
70. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
71. Young, M. D. *et al.* ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Res.* (2011).doi:10.1093/nar/gkr416
72. Mendenhall, E. M. & Bernstein, B. E. Chromatin state maps: new technologies, new insights. *Curr. Opin. Genet. Dev.* **18**, 109–115 (2008).
73. Santos-Rosa, H. *et al.* Active genes are tri-methylated at K4 of histone H3. *Nature* **419**, 407–411 (2002).
74. Bernstein, B. E. *et al.* Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 8695–8700 (2002).
75. Schübeler, D. *et al.* The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev.* **18**, 1263–1271 (2004).
76. Schneider, R. *et al.* Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. *Nature Cell Biology* **6**, 73–77 (2004).
77. Ng, H.-H., Robert, F., Young, R. A. & Struhl, K. Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Mol Cell* **11**, 709–719 (2003).
78. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**, 77–88 (2007).
79. Pekowska, A. *et al.* H3K4 tri-methylation provides an epigenetic signature of active enhancers. *The EMBO Journal* (2011).doi:10.1038/emboj.2011.295
80. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
81. Pokholok, D. K. *et al.* Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122**, 517–527 (2005).
82. Bernstein, B. E. *et al.* Genomic maps and comparative analysis of histone

modifications in human and mouse. *Cell* **120**, 169–181 (2005).
83. Kolasinska-Zwierz, P. *et al.* Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat. Genet.* **41**, 376–381 (2009).
84. Barrand, S. & Andersen, I. ScienceDirect.com - Biochemical and Biophysical Research Communications - Promoter-exon relationship of H3 lysine 9, 27, 36 and 79 methylation on pluripotency-associated genes. *Biochemical and biophysical research ...* (2010).
85. Kharchenko, P. V. *et al.* Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471**, 480–485 (2010).
86. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**, 817–825 (2010).
87. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 1–9 (2011).doi:10.1038/nature09906
88. The modENCODE Consortium *et al.* Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. *Science* **330**, 1787–1797 (2010).
89. Filion, G. J. *et al.* Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* **143**, 212–224 (2010).
90. van Steensel, B. Chromatin: constructing the big picture. *The EMBO Journal* **30**, 1885–1895 (2011).
91. van Steensel, B., Delrow, J. & Henikoff, S. Chromatin profiling using targeted DNA adenine methyltransferase. *Nat. Genet.* **27**, 304–308 (2001).
92. Roh, T. Y. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev.* **19**, 542–552 (2005).
93. Roh, T. Y., Wei, G., Farrell, C. M. & Zhao, K. Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns. *Genome Res.* **17**, 74–81 (2006).
94. Crawford, G. E. *et al.* Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* **16**, 123–131 (2006).
95. Gross, D. S. & Garrard, W. T. Nuclease Hypersensitive Sites in Chromatin. *Annu. Rev. Biochem.* **57**, 159–197 (1988).
96. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 21931–21936 (2010).
97. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
98. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
99. Ong, C.-T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Publishing Group* **12**, 283–293 (2011).
100. Jin, C. *et al.* H3.3/H2A.Z double variant|[ndash]|containing nucleosomes mark “nucleosome-free regions” of active promoters and other regulatory regions. *Nat. Genet.* **41**, 941–945 (2009).
101. Lin, Y. C. *et al.* A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. *Nat. Immunol.* **11**, 635–643 (2010).
102. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
103. De Santa, F. *et al.* A Large Fraction of Extragenic RNA Pol II Transcription Sites

Overlap Enhancers. *PLoS Biol.* **8**, e1000384 (2010).
104. Koch, F. *et al.* Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat. Struct. Mol. Biol.* **18**, 956–963 (2011).
105. Gilbert, S. F. *Developmental Biology, Ninth Edition (Developmental Biology Developmental Biology)*. 711 (Sinauer Associates, Inc.: 2010).
106. Hartenstein, V. *Atlas of Drosophila Development*. 58 (Cold Spring Harbor Laboratory Press: 1995).
107. Sanson, B. Generating patterns from fields of cells: Examples from Drosophila segmentation. *EMBO Reports* **2**, 1083–1088 (2001).
108. Beer, J. & Technau, G. Lineage analysis of transplanted individual cells in embryos of Drosophila melanogaster. *Development Genes and …* (1987).
109. Farrell, E. R. & Keshishian, H. Laser ablation of persistent twist cells in Drosophila: muscle precursor fate is not segmentally restricted. *Development* **126**, 273–280 (1999).
110. Sink, H. *Muscle Development in Drosophila*. (Landes Bioscience: 2006).
111. Furlong, E. E. Integrating transcriptional and signalling networks during muscle development. *Curr. Opin. Genet. Dev.* **14**, 343–350 (2004).
112. Baylies, M. K. & Bate, M. twist: a myogenic switch in Drosophila. *Science* **272**, 1481–1484 (1996).
113. Bodmer, R. The gene tinman is required for specification of the heart and visceral muscles in Drosophila. *Development* **118**, 719–729 (1993).
114. Maggert, K., Levine, M. & Frasch, M. The somatic-visceral subdivision of the embryonic mesoderm is initiated by dorsal gradient thresholds in Drosophila. *Development* **121**, 2107–2116 (1995).
115. Azpiazu, N., Lawrence, P. A., Vincent, J. P. & Frasch, M. Segmentation and specification of the Drosophila mesoderm. *Genes Dev.* **10**, 3183–3194 (1996).
116. Reim, I. & Frasch, M. The Dorsocross T-box genes are key components of the regulatory network controlling early cardiogenesis in Drosophila. *Development* **132**, 4911–4925 (2005).
117. Jakobsen, J. S. *et al.* Temporal ChIP-on-chip reveals Biniou as a universal regulator of the visceral muscle transcriptional network. *Genes Dev.* **21**, 2448–2460 (2007).
118. Tapanes-Castillo, A. & Baylies, M. K. Notch signaling patterns Drosophila mesodermal segments by regulating the bHLH transcription factor twist. *Development* **131**, 2359–2372 (2004).
119. Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nature Publishing Group* **5**, 276–287 (2004).
120. Van Loo, P. & Marynen, P. Computational methods for the detection of cis-regulatory modules. *Briefings in Bioinformatics* **10**, 509–524 (2009).
121. Aerts, S. Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. *Curr. Top. Dev. Biol.* **98**, 121–145 (2012).
122. Hertz, G. Z. & Stormo, G. D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563–577 (1999).
123. Thomas-Chollier, M. *et al.* RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.* **36**, W119–27 (2008).
124. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).

125. Katzman, S. *et al.* Human genome ultraconserved elements are ultraselected. *Science* **317**, 915 (2007).
126. Pennacchio, L. A. *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).
127. Pennacchio, L. A. & Visel, A. Limits of sequence and functional conservation. *Nat. Genet.* **42**, 557–558 (2010).
128. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
129. Davidson, E. *Genomic Regulatory Systems: Development and Evolution*. (Academic Press, San Diego, USA: 2001).
130. Gotea, V. *et al.* Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res.* **20**, 565–577 (2010).
131. Lifanov, A. P., Makeev, V. J., Nazina, A. G. & Papatsenko, D. A. Homotypic regulatory clusters in Drosophila. *Genome Res.* **13**, 579–588 (2003).
132. Markstein, M., Markstein, P., Markstein, V. & Levine, M. S. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 763–768 (2002).
133. Rajewsky, N., Vergassola, M., Gaul, U. & Siggia, E. D. Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformatics* **3**, 30 (2002).
134. Frith, M. C., Li, M. C. & Weng, Z. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.* **31**, 3666–3668 (2003).
135. Kim, J. *et al.* Functional Characterization of Transcription Factor Motifs Using Cross-species Comparison across Large Evolutionary Distances. *PLoS Computational Biology* **6**, e1000652 (2010).
136. Halfon, M. S., Grad, Y., Church, G. M. & Michelson, A. M. Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res.* **12**, 1019–1028 (2002).
137. Berman, B., Nibu, Y. & Pfeiffer, B. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. (2002).
138. Schroeder, M. D. *et al.* Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biol.* **2**, E271 (2004).
139. Ferretti, V. *et al.* PReMod: a database of genome-wide mammalian cis-regulatory module predictions. *Nucleic Acids Res.* **35**, D122–6 (2007).
140. Blanchette, M. *et al.* Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.* **16**, 656–668 (2006).
141. Jensen, L. J. & Bateman, A. The rise and fall of supervised machine learning techniques. *Bioinformatics* **27**, 3331–3332 (2011).
142. Wasserman, W. W. & Fickett, J. W. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **278**, 167–181 (1998).
143. Krivan, W. & Wasserman, W. W. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.* **11**, 1559–1566 (2001).
144. Portales-Casamar, E. *et al.* JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **38**, D105–10 (2010).

145. Zhu, L. J. *et al.* FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res.* **39**, D111–7 (2011).
146. Narlikar, L. *et al.* Genome-wide discovery of human heart enhancers. *Genome Res.* **20**, 381–392 (2010).
147. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88–D92 (2007).
148. Noyes, M. B. *et al.* A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. *Nucleic Acids Res.* **36**, 2547–2560 (2008).
149. Berger, M. F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* **24**, 1429–1435 (2006).
150. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505–510 (1990).
151. Maerkl, S. J. & Quake, S. R. A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. *Science* **315**, 233–237 (2007).
152. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).
153. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653–1659 (2011).
154. Thomas-Chollier, M. *et al.* RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res.* **40**, e31–e31 (2012).
155. Nègre, N. *et al.* A cis-regulatory map of the Drosophila genome. *Nature* **471**, 527–531 (2011).
156. Vokes, S. A., Ji, H., Wong, W. H. & McMahon, A. P. A genome-scale analysis of the cis-regulatory circuitry underlying sonic hedgehog-mediated patterning of the mammalian limb. *Genes Dev.* **22**, 2651–2663 (2008).
157. Sandmann, T. *et al.* A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Dev. Cell* **10**, 797–807 (2006).
158. Zeitlinger, J. *et al.* Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. *Genes Dev.* **21**, 385–390 (2007).
159. MacArthur, S. *et al.* Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol.* **10**, R80 (2009).
160. Ji, H. & Wong, W. H. TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics* **21**, 3629–3636 (2005).
161. Zhang, Y., Liu, T., Meyer, C. & Eeckhoute, J. Model-based analysis of ChIP-Seq (MACS). *Genome ...* (2008).
162. Li, X.-Y. *et al.* Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol.* **6**, e27 (2008).
163. Kvon, E. Z., Stampfel, G., Yanez-Cuna, O. J., Dickson, B. J. & Stark, A. HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. *Genes Dev.*

164. Wu, C., Bingham, P. M., Livak, K. J., Holmgren, R. & Elgin, S. C. The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. *Cell* **16**, 797–806 (1979).
165. Nagy, P. L., Cleary, M. L., Brown, P. O. & Lieb, J. D. Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 6364–6369 (2003).
166. Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. & Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* **17**, 877–885 (2007).
167. Sabo, P. J. *et al.* Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat. Methods* **3**, 511–518 (2006).
168. Crawford, G. E. *et al.* DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat. Methods* **3**, 503–509 (2006).
169. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
170. Giresi, P. G. & Lieb, J. D. Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). *Methods* **48**, 233–239 (2009).
171. Song, L. *et al.* Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.* (2011).doi:10.1101/gr.121541.111
172. Dojer, N., Gambin, A., Mizera, A., Wilczyński, B. & Tiuryn, J. Applying dynamic Bayesian networks to perturbed gene expression data. *BMC Bioinformatics* **7**, 249 (2006).
173. Wilczyński, B. & Dojer, N. BNFinder: exact and efficient method for learning Bayesian networks. *Bioinformatics* **25**, 286–287 (2009).
174. He, H. H. *et al.* Nucleosome dynamics define transcriptional enhancers. *Nat. Genet.* **42**, 343–347 (2010).
175. Kowalczyk, M. S. *et al.* Intragenic enhancers act as alternative promoters. *Mol Cell* **45**, 447–458 (2012).
176. May, D. *et al.* Large-scale discovery of enhancers from human heart tissue. *Nat. Genet.* **44**, 89–93 (2011).
177. Emberly, E. *et al.* BEAF regulates cell-cycle genes through the controlled deposition of H3K9 methylation marks into its conserved dual-core binding sites. *PLoS Biol.* **6**, 2896–2910 (2008).
178. Zykovich, A., Korf, I. & Segal, D. J. Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. *Nucleic Acids Res.* **37**, e151 (2009).
179. Jolma, A. *et al.* Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* **20**, 861–873 (2010).
180. Wang, J., Lu, J., Gu, G. & Liu, Y. In vitro DNA-binding profile of transcription factors: methods and new insights. *J. Endocrinol.* **210**, 15–27 (2011).
181. Kauffmann, A., Gentleman, R. & Huber, W. arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. *Bioinformatics* **25**, 415–416 (2009).
182. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).

183. Ho, J. W. K. *et al.* ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics* **12**, 134 (2011).
184. Chen, Y. *et al.* Systematic evaluation of factors influencing ChIP-seq fidelity. *Nat. Methods* (2012).doi:10.1038/nmeth.1985
185. Cuellar-Partida, G. *et al.* Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics* **28**, 56–62 (2011).
186. Ernst, J., Plasterer, H. L., Simon, I. & Bar-Joseph, Z. Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res.* **20**, 526–536 (2010).
187. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* (2012).doi:10.1038/nbt.2198
188. Clarke, L. *et al.* The 1000 Genomes Project: data management and community access. *Nat. Methods* **9**, 459–462 (2012).
189. Goecks, J., Nekrutenko, A., Taylor, J.Galaxy Team Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**, R86 (2010).

Annexes

Annexe 1: The IUPAC code for nucleotide symbols

Nucleotide Symbols IUPAC notions		
A	A	**A**denine
C	C	**C**ytosine
G	G	**G**uanine
T	T	**T**hymine
U	U	**U**racil
R	A or G	pu**R**ine
Y	C or T (U)	p**Y**rimidine
M	A or C	a**M**ino
K	G or T (U)	**K**eto
S	C or G	**S**trong (triple '3 H' bonds)
W	A or T (U)	**W**eak (double '2 H' bonds)
B	C or G or T (U)	not A
D	A or G or T (U)	not C
H	A or C or T (U)	not G
V	A or C or G	not T (U)
N	A or C or G or T (U)	a**N**y nucleotide

Annexe 2: Supplementary Information for "Tissue specific analysis of chromatin state reveals temporal signatures of enhancer activity during embryonic development"

Supplementary Information

Tissue specific analysis of chromatin state reveals temporal signatures of enhancer activity during embryonic development

Stefan Bonn*, Robert P. Zinzen*, Charles Girardot*, E. Hilary Gustafson, Alexis P. Gonzalez, Nicolas Delhomme, Yad Ghavi-Helm, Bartek Wilczynski, Andy Riddell and Eileen E.M. Furlong

Table of Contents

Supplementary Note

I. Experimental procedures

I.1 Antibodies for immunoprecipitation

The commercially available antibodies detecting H3 (ab1791), H3K4me3 (ab71998), H3K4me1 (ab8895), H3K27ac (ab4729), H3K36me3 (ab9050), and H3K79me3 (ab2621) were purchased from abcam® and anti-H3K27me3 from Active Motif (39155). The Rpb3 (RNA Polymerase II) antibody was a generous gift from John Lis[1] and the anti-Mef2 antibody was generated in the Furlong lab[2]. The specificity of the majority of the commercial antibodies used were recently assessed[3] and shown to be of high specificity for their antigen and of ChIP quality (H3K4me1, H3K27ac, H3K36me3, H3K79me3, H3K27me3). The H3 antibody performed well in Western blots, but apparently failed in *Drosophila* chromatin IPs[3], yet it performed well in our hands. The IP conditions for the individual antibodies were optimized for recovery and enrichment using small amounts of chromatin (2-10µg) to yield enough material for subsequent library generation (3-40ng). The quality of each IP was assessed by real-time PCR (for primer info see Supplementary Table 6). Real-time PCR primer combinations were as follows: H3 – *osk*/*twi*-promoter; H3K4me1 – *Rpl32-5'*/*osk*; input, H3K27ac, H3K79me3 and H3K4me3 – *Rpl32*-promoter/*Rpl32-5*; H3K36me3 – *Rpl32-5'*/*Rpl32*-promoter; H3K27me3 – *tup*-promoter/*Rpl32*-promoter; Mef2 – *act*/*osk*; Rpb3 – *twi*-promoter/*Rpl32-5'*.

I.2 FACS sorting conditions for fixed nuclei sorting

Nuclear samples were run on a Beckman Coulter MoFlo cell sorter using Summit software version 4.3. A Coherent Innova 90C Argon ion laser (Coherent Inc.), tuned to 488nm TEM_{00} mode (200mW), was used as primary laser. Small width obscuration bars were placed in front of the Forward Scatter and wide-angle 90° light collection lenses. Laser illumination, Moflo's L-configuration optical layout and sorting were optimized using Flow-Check™ Fluorospheres (Beckman Coulter Inc.).

A BD FACSFlow™ sheath (Becton Dickinson GmbH), filtered in-line through a PALL Fluorodyne II filter 0.2µm (Pall GmbH), was used in the acquisition and sorting of Alexa488 stained nuclei. The differential pressure was kept low to limit illumination variations. Each sample was collected using FSC and Alexa488 correlated data parameters while thresholding on FSC. Sample acquisition rates averaged 30000 events per second. Alexa488 fluorescence

intensity from the immunostained tagged histone was measured after passing collected light through a 530/40 nm bandpass filter. A second detector collected fluorescence through a 670/40 nm bandpass filter. Temperature on both sample and collection tubes was kept at 4°C during sorting. The sort decision gate was based on a combination of scatter, pulse width and fluorescence parameters. Post-acquisition analysis was performed using FlowJo version 9.2 for Macintosh (Tree Star).

To independently assess sorting purity small aliquots of sorted samples were DAPI stained and assayed under an epifluorescent microscope. Samples with >5% DAPI-positive, Alexa488-negative events were discarded or resorted (which usually resulted in >99% purity).

I.3 Chromatin preparation and immunoprecipitation

Immediately after sorting, the nuclei were centrifuged at 3500 g for 10 min and the pellet resuspended in 300 µl RIPA buffer and transferred into a 1.5 ml tube (RIPA: 140m*M* NaCl, 10 m*M* Tris-HCl pH 8.0, 1 m*M* EDTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate, Roche proteinase inhibitors). After incubation for 10 min on ice, the chromatin was sheared into 200 bp fragments using a Diagenode BioRuptor (18 cycles, high intensity, 30s on/off intervals, ice-cold water changed every 6 cycles to maintain the sample at ~4°C). The sample was centrifuged for 2 min at 18000 g and the supernatant was transferred to a low-binding tube (Biozym Scientific GmbH, 710176). While the majority of sheared chromatin was subsequently snap-frozen in liquid nitrogen, a small aliquot was used to measure DNA concentration and fragment size after reverse cross-linking[4].

The chromatin IP was performed as described previously[4] with the following modifications. The chromatin was precleared to extract any antibodies left from the staining procedure. For this, 20 µl of 50% ProtG suspension (Protein G Sepharose, Sigma, P3296) was washed twice with 1 ml RIPA buffer and the beads were added to the chromatin using 700 µl RIPA. The suspension was incubated for 1h on a rotating wheel, centrifuged for 2 min at 1000 g and the cleared supernatant was used for chromatin IP as described. The ProtG Sepharose-bound material was used for chromatin IP in two replicates and sequenced. Comparing these SBP libraries to H3 identified only 50 enriched regions genome-wide, demonstrating that SBP preclearing does not introduce any bias.

I.4 Solexa library preparation and sequencing

Solexa libraries were prepared according to manufacturers recommendations with small modifications. In short, 3–10ng of IP-purified, RNase treated, and reverse cross-linked

genomic DNA was end-repaired and terminal adenosine residues were added using the NEBNext® reagents. PE adapters (Illumina) were ligated, after which the material was size selected at ~230-250 bp (size equals sheered chromatin fragments plus adapters) on a 2% GelGreen™ (Biotium) stained agarose gel using SafeXtractor™-100 gel extractors (5Prime) under blue light. PCR amplification was performed using PE1.0 and PE2.0 primers (Illumina) for 18 cycles according to manufacturer's recommendation using the Phusion® High-Fidelity PCR Kit (Finnzyme). The PCR-amplified library was purified on a 2% agarose gel, avoiding unincorporated primers and primer-self-ligation products. Library quality was assessed on a Bioanalyzer2100 system (Agilent). To increase coverage and ensure detection saturation and reproducibility, two biological replicates of every mark were single-end sequenced with 36 bp reads using an Illumina Genome Analyzer IIx by the EMBL Genomics Core facility. Sequencing information for each mark is shown in Supplementary Table 3. The sequencing data is accessible at the Sequence Read Archive (SRA) under the study accession number ERP000560 (http://www.ebi.ac.uk/ena/data/view/ERP000560).

II. ChIP-Seq data processing

II.1 ChIP-Seq data quality assurance

The quality assurance and pre-processing workflow was implemented in Galaxy[5]. First, quality assurance was performed using the Bioconductor package ShortRead[6,7]. Every sequencing lane passed the quality threshold; see Supplementary Table 3 for the number of reads. Second, biological replicates were compared to evaluate their reproducibility. Briefly, the data from every experiment was corrected for library size in a manner similar to the RPKM correction used for RNA-Seq data[8] using an in house R package (HistoneChIPseq). As we are examining chromatin marks, the mappable genome size (135 Mb for 36bp long reads) was used as the reference to apply the correction, which has the advantage of direct genomic coverage readout. The coverage enrichment **E** of a library is $\mathbf{E = R * l / M}$ (1), where **M** is the mappable genome size, **R** the number of reads in the library and **l** the length of the reads (or of the fragments, whether one shifts or extends the reads, respectively). Thus **E** represents the expected coverage of the library if the reads were uniformly distributed. For a given base pair in the genome overlapped by **r** reads, its corrected score **s** is: $\mathbf{s = r/E}$ (2). We call this corrected value "Read Per Genomic Coverage" (RPGC). The obtained scores were summarized (median value) using non-overlapping, adjacent windows of size 200bp or 2 kb

for input (BiTS-isolated chromatin) and H3K27me3. The obtained values were compared between biological replicates using Pearson correlation (Supplementary Table 3).

II.2 Read Alignment

The Illumina export files were converted into FASTQ[9] formatted data, which were aligned against the *D. melanogaster* dm3 genome obtained from FlyBase[10] using bowtie[11] with the following parameters: -n 2 -e 70 -l 28 --maxbts 800 -y -m 1 --best --strata -S -q --phred64-quals. All parameters are default values with the exception of -m 1, which ensures that reads aligning to more than one locus are omitted to avoid alignment bias. The obtained Sequence Alignment/Map (SAM) files were converted into sorted Binary Alignment/Map (BAM) files using the SAMtools suite[12].

II.3 ChIP-seq read summarization and binning

Samples were corrected to the mappable genome size (135 Mb) using the HistoneChIPseq R package, generating Read Per Genomic Coverage (RPGC) scores that were summarized (median value) into adjacent non-overlapping bins of a defined size (bin sizes of 25 or 50 bp were used for all analyses using binned values). Two described approaches (NormDiff and Background Subtracted)[13] were used to independently perform the background correction for chromatin modifications on histones (using H3 data as the background model) and for H3, RNA Polymerase II (Pol II) and Mef2 (using input data as the background model. We used Background Subtracted normalized data for the visualization of genomic loci in IGB[14] and for 'gene and CRM intensity profiles' and NormDiff[13] normalized data if not otherwise indicated.

II.4 Peak Finding

BAM files were converted into Bed files using the bamToBed tool from the BEDTools suite[15]. Peaks were defined with MACS (v1.3.7.1)[16] using the following parameters: --tsize=36 --nomodel --pvalue=0.00001 --format=BED --shiftsize=90 (in agreement with the fragment length estimated physically using the Agilent Bioanalyzer i.e. ~180 bp) --bw=100 --gsize=135000000. Histone marks were analyzed against H3 control data, while for Pol II and Mef2 input was used.

II.5 Analysis of data saturation

To assess if the sequencing libraries contained enough reads to reach saturation in peak calling using MACS, we performed a sub-sampling analysis. For this, we sampled increasing

amounts of duplicate-filtered data (from 10% to 100%, with 10% steps) for each biological condition and called peaks using MACS (the background library was always identical and composed of 100% of available reads). Saturation was considered reached when including 10% more reads consistently increased the coverage of peaks recalled by less than 5%. This measure was preferred over peak number because of the observed effect of frequent peak merging (see Supplementary Fig. 9). We reached saturation for all chromatin marks examined, as shown (Supplementary Table 3, Supplementary Fig. 9).

III. The chromatin marks studied here represent four of the five major chromatin types

A comprehensive study by Filion *et al.* used the occupancy of 53 chromatin binding proteins in Kc167 cells to segment the *Drosophila* genome into 5 major chromatin types[17]. A subset of only five of these proteins (which collectively occupy 97.6% of the genome) recapitulates the five-chromatin states with 85.5% accuracy (marker proteins). The six chromatin marks that we examined, in addition to histone H3, represent four of the five-chromatin states as indicated in the table below. The only state that we have not examined is heterochromatin. Since not a single known *Drosophila* enhancer (in CAD2) maps to regions of the *Drosophila* genome annotated as heterochromatin, the absence of heterochromatin marks (e.g. H4K9) should not impact on any of the conclusions made in this study, or on our ability to directly learn what combinations of chromatin marks are predictive of enhancer activity.

Filion *et al* (marker proteins)	**Purpose**	**Our study**
Histone H1	Nuclesosome density	Histone H3
PC	Polycomb repressed regions	H3K27me3
HP1	Heterochromatin	—
MRG15	Active chromatin with K36me3 (yellow)	H3K36me3
BRM	Active chromatin, no K36me3 (red)	H3 K79me3, K4me3, K27ac

'Blue' chromatin (repressed) is represented by H3K27me3, which is placed by the Polycomb complex, while 'black' chromatin is represented by the presence of Histone H3 and a general absence of other chromatin modifications (as shown in Fig 3B and S1E, Filion *et al*). Active chromatin is detected using H3K4me3, H3K27ac, H3K79me3 and H3K36me3. 'yellow' and 'red' active chromatin can be readily distinguished from each other by splitting active regions (denoted by the active chromatin marks H3 K4me3, K27ac and K79me3) into those that have

H3K36me3 ('yellow') and those that do not ('red'). Interestingly, comparing 'yellow' and 'red' chromatin, Filion *et al.* describe that (1) the nucleosome-remodeling ATPase Brahma (BRM) and the Mediator subunit MED31 are exclusively found in 'red' chromatin, (2) that 'red' chromatin is characterized by the presence of H3K79me3 and a lack of H3K36me3 and (3) that 'red' chromatin contains genes with restricted expression domains and that are linked to more specific processes than genes found in the 'yellow' chromatin. Based on this, the authors suggest that the intergenic 'red' chromatin may contain more regulatory chromatin complexes. Our data provides direct evidence for this – H3K79me3 is a mark of active enhancers, while H3K36me3 is depleted on active enhancers (Supplementary Fig. 5 and Supplementary Fig. 10).

IV. modENCODE data

IV.1 Processing of whole-embryo ChIP-seq data (modENCODE)

(Used in the analysis shown in Supplementary Figure 4)

Whole-embryo ChIP-seq data produced by the modENCODE consortium for H3K27ac (SRR030292 i.e. E4-8_H3K27Ac_ChIPSeq_1), H3K4me3 (SRR030287 i.e. E4-8_H3K4Me3_ChIPSeq_1), H3K4me1 (SRR030294 i.e. E4-8_H3K4Me1_ChIPSeq_1) and Pol II (SRR030327 i.e. E4-8_PolII_ChIPSeq_1) at 4-8 hrs together with their respective input controls (SRR030288 i.e. E4-8_INPUT for chromatin marks and SRR030345 i.e. E4-8_INPUT_PolII for Pol II) were obtained from the Short Read Archive. Reads were aligned and binned following the same procedure as for BiTS-ChIP-seq data, which is fully described in the II.2 'Read Alignment' and II.3 'ChIP-seq read summarization and binning' sections.

IV.2 Non-mesodermal TFs

(Used in the analysis shown in Figure 3, light grey columns)

To create a list of binding sites occupied by non-mesodermal TFs, we utilized modENCODE ChIP-chip data for TFs that are not expressed in mesoderm at 6-8 hrs but are expressed at this stage in other tissues[18]. The following criteria were applied to select the TFs used: (1) The TFs must have no annotated expression in the mesoderm at 6-8 hrs, using the BDGP *in situ* hybridization database[19], (2) have no mesodermal annotation at 6-8 hrs in Flybase[10], (3) to identify potentially unannotated expression in mesoderm, we visually inspected TF loci using BiTS-ChIP data and excluded any genes that had activity marks present in the gene promoter (H3K27ac, H3K4me3) or gene body (Pol II, H3K79me3, H3K36me3). The resulting seven

non-mesodermal TFs used were GATAe, cnc, D, disco, dll, hkb, and sens. modENCODE accession numbers were 2573 (E0-8h_GATAe), 627 (E0-12h-CNC), 2571 (E0-8_D), 2572 (E0-8h_disco), 606 (E-0-12h-dll), 2575 (E0-8h_hkb), 2577 (E4-8h_sens).

V. Gene lists using the BDGP *in situ* hybridization database

(Used in the analysis shown in Figure 2c-f, Supplementary Figures 3 and 4)

The following gene lists were assembled using the Berkeley *Drosophila* Genome Project *in situ* database BDGP[19] (downloaded on July 2010):

‘active 6-8h’ *Genes expressed ubiquitously and in the mesoderm at 6-8 hrs*

List is composed of genes that are expressed throughout most of the embryo, i.e. ubiquitously expressed or annotated with one of the anatomical terms and one of the stage terms listed in Supplementary Table 8. Of the 572 ‘active 6-8h’ genes, 267 are expressed ubiquitously, 229 are expressed in several tissues including the mesoderm, and 38 are expressed only in mesoderm.

‘meso 6-8h’ *Genes expressed in mesoderm but not ubiquitously at 6-8 hrs*

List is composed of genes expressed in mesodermal cells at 6-8 hrs. Genes have to be annotated with one of the anatomical terms and one of the stage terms listed in Supplementary Table 8. This includes genes expressed in mesoderm and one or more other tissues, but excludes ubiquitously expressed genes.

’only meso 6-8h’ *Genes expressed exclusively in the mesoderm at 6-8 hrs*

List is composed of mesoderm-specific genes that (1) are expressed only in the mesoderm at 6-8 hrs and (2) are not ubiquitously expressed at 6-8 hrs. Unlike the ‘meso 6-8h’ gene list, the ‘only meso 6-8h’ genes are not expressed anywhere outside of the mesoderm at 6-8 hrs of development.

’no meso 6-8h’ *Genes expressed only outside of the mesoderm at 6-8 hrs*

List is composed of genes that have (1) no mesodermal annotation at any stage of development and (2) must be expressed in a tissue outside of the mesoderm at stages 9-10 or 11-12 to be selected.

‘inactive 6-8h’ *Genes inactive at 6-8 hrs but active later in development*

List is composed of genes that (1) have “no staining” annotation term at stages ‘stage 1-3’, ‘stage 4-6’, ‘stage 7-8’, ‘stage 9-10’ and ‘stage 11-12’ and (2) are

expressed at ‘stage 13-16’ anywhere but in a mesodermal tissue. This second criteria ensures that it is possible to detect these genes’ expression by *in situ* hybridization, thereby eliminating genes annotated as having ‘no staining’ simply due to problematic *in situ* probes.

To avoid confounding overlapping signatures from multiple TSSs, we selected genes that contain a single annotated TSS and are further than 1 kb from another TSS. The assembled lists (provided in Supplementary Table 7) contain 572 (active 6-8h), 267 (meso 6-8h), 38 (only meso 6-8h), 275 (no meso 6-8h) and 78 (inactive 6-8h) genes. Note, as the data shown in Figures 2c-f and Supplementary Figures 3 and 4 was summarized using a trimmed mean and filtered for genes larger than 850 bp (see section VII ‘Gene and CRM intensity profiles’), the total number of genes plotted is smaller (427 ‘active 6-8h’ ; 201 ‘meso 6-8h’; 30 ‘only meso 6-8h’; 209 ‘no meso 6-8h’; 56 ‘inactive 6-8h’). The following mesodermal anatomical terms were ignored when assembling the gene lists with mesodermal activity as these cell types were excluded from the FACS sorting due to the restricted specific expression of the *twist* enhancer in the trunk mesoderm: ‘embryonic/larval circulatory system’, ‘embryonic/larval fat body’, ’fat body specific anlage’, ’fat body/gonad primordium’, ’longitudinal visceral mesoderm primordium’, ’longitudinal visceral muscle fibers’.

VI. Definition and processing of CAD2 and TF-Meso-CRMs databases

VI.1 CAD2- and TF-Meso-CRMs

(Basis for the analysis shown in Figure 3, 4 and 5)

The CRM Activity Database version 2 (CAD2, Supplementary Table 1) is based on CAD[20], but was updated to include new *Drosophila* enhancers reported since 2009 in REDfly[21] and elsewhere (PMIDs in Supplementary Table 1). Entries were filtered for size (entries are ≤ 2 kb) and remapped to dm3 using the UCSC LiftOver tool[22]. CAD2 contains literature-based annotation of the *in vivo* activity of *Drosophila* enhancers reported using transgenic embryos. Each enhancer was therefore individually annotated for its specific spatio-temporal activity according to published expression annotation and images reported in the literature, where we manually confirmed the activity of all enhancers using images from the original publications. Stage-specific annotation encompasses enhancer activity from embryonic stage 5 or prior, st.6, st.7, st.8, st.9, st.10, st.11, st.12, st.13, st.14, and st.15 or later; annotations distinguish between expression in the mesoderm (and its derivatives,

column headers M5-15 in Supplementary Table 1) and/or expression in non-mesodermal tissues (column headers O5-15 in Supplementary Table 1). We manually annotated the activity of 'mesoderm' enhancers into specific domains where applicable (e.g. somatic mesoderm, cardiac mesoderm, visceral mesoderm, dorsal mesoderm and mesoderm flagged with 'S', 'C', 'V', 'dM' and 'M' respectively in M5-15 columns in Supplementary Table 1), while all enhancers active outside of mesoderm were annotated as active in 'non-mesodermal tissues' (flagged with a '1' in O5-15 columns in Supplementary Table 1). Enhancers described as inactive are flagged with a '0' while a 'ni' flag is used when activity information was not reported at that stage in the original publication (Supplementary Table 1).

The Mesodermal ChIP-CRM atlas (TF-Meso-CRMs) was obtained from Zinzen *et al.*[20] and remapped to dm3 using the UCSC LiftOver tool[22].

As several chromatin modifications and Pol II have known gene specific signatures, we excluded all gene proximal enhancers to avoid confounding chromatin signatures coming from active promoters and gene transcription. We therefore removed enhancers (CAD2 and TF-Meso-CRMs) residing within 1 kb of gene annotations, both 5' and 3' (note that the *Drosophila* genome is very dense, with one TSS every ~5.6 kb). As gene specific signals for H3 K79me3, K36me3 , K4me1, K4me3, K27ac, and Pol II are well confined within gene boundaries (see Fig. 2c-f and Supplementary Fig. 3), we are very confident that filtering for gene regions extended by1 kb removes any chromatin signal for known annotated genes. In addition, enhancers that overlapped by at least one base with an H3K4me3 enriched region (as reported by the MACS peak caller) were excluded to avoid including unannotated TSSs. After filtering, CAD2 contains 144 intergenic enhancers and the Mesodermal ChIP-CRM atlas has 1717 TF-Meso-CRMs (of which 844 are bound at 6-8 hours, used in Fig. 5a and Supplementary Fig. 8). All analysis performed in this study used these subsets as a general starting point. Any further filtering for spatial and/or temporal aspects of CAD2 and TF-Meso-CRM activity (binding) is described in the following section.

VI.2 Defining enhancer classes with specific spatio-temporal activity
(Used in the analysis shown in Figure 3-5 and Supplementary Fig. 7 and 10)

Two enhancer databases (CAD2 and TF-Meso-CRMs) were used to define groups of enhancers with specific spatio-temporal activity:

CAD2: To obtain groups of active and inactive enhancers (used for the analysis in Fig. 3), we used CAD2 literature-based information on the spatial activity of enhancers in the

mesoderm and outside of the mesoderm. To obtain CAD2 enhancers that are active in mesoderm (class 'A'), we searched for the terms "S" = somatic mesoderm, "C" = cardiac mesoderm, "V" = visceral mesoderm, "M" = mesoderm, and "dM" = dorsal mesoderm at 6-8 hrs of development (i.e. stages 10 and 11). All enhancers with a positive annotation were assigned to the active class at that time. Within the 144 intergenic CAD2 enhancers, 22 are active in mesoderm at 6-8 hrs.

To obtain enhancers that are strictly inactive in mesoderm at 6-8 hrs (class 'I'), we selected CAD2 enhancers that have annotations for inactivity in mesoderm for the stages of interest (stages 10 and 11) and the directly adjacent stages (stages 9 and 12). We chose this rather stringent filter because assessing inactivity is inherently more difficult and thus more error prone than activity. Within the 144 intergenic CAD2 enhancers, 21 are strictly inactive in mesoderm at 6-8 hrs.

To obtain enhancers that are active (exclusively) outside of the mesoderm at 6-8 hrs (class 'O'), we selected CAD2 enhancers that had any activity information outside mesoderm and no activity annotation in mesoderm at stages 10 and 11 (i.e. 6-8 hrs). This reduced the set of 144 CAD2 enhancers to 140 of which 31 are active only outside of mesoderm at 6-8 hrs.

To obtain temporal classes of CAD2 enhancers (used for Fig. 4a) that are active in the mesoderm before ('E'arly), at ('A't), or after ('L'ate) 6-8 hrs, we searched CAD2 for the terms "S" = somatic mesoderm, "C" = cardiac mesoderm, "V" = visceral mesoderm, "M" = mesoderm, and "dM" = dorsal mesoderm at the time of interest (stages 10 and 11). Enhancers active before 6-8 hrs ('E'arly class) must have at least one active annotation at stages 5-9 (2-6 hrs) of development but be devoid of activity annotation at stages 10-15 (6-8 hrs or later). Similarly, enhancers active after 6-8 hrs ('L'ate class) must have at least one active annotation at stages 12-15 of development but be devoid of activity annotation at stages 5-11 (2-8 hrs) of development.

TF-Meso-CRMs: Three temporal enhancer lists (used for the analysis shown in Fig. 4b-f) were assembled using existing TF-binding data for five mesoderm specific transcription factors (Mef2, Twist, Tin, Bin and Bap) for 5 different stages of development[20]. Only TF-Meso-CRMs that were bound by at least two TFs (Mef2, Twist, Tin, Bap or Bin) at the same developmental stage (either 2-4, 4-6, 6-8, 8-10, or 10-12 hrs)) were considered to obtain a stringent list of high-confidence enhancers. (1) The temporal class '6-8h' is composed of 297 CRMs that are co-bound at 6-8 hrs. (2) The class '8-12h' contains 10 CRMs that are bound

only after 6-8 hrs (co-bound either at 8-10 or 10-12 hrs) and have no binding events before or at 6-8 hrs. (3) The class '2-6h' is composed of 88 CRMs that are bound only before 6-8 hrs (co-bound either at 2-4 or 4-6 hrs) and have no binding events at or after 6-8 hrs. Notably, since the data in figures 4 b-f was summarized using a trimmed mean (see section VII 'Gene and CRM intensity profiles'), the total number of CRMs plotted is smaller (239, 8, and 72 CRMs for the '6-8h', '8-12h' and '2-6h' classes, respectively).

VI.3 Enrichment, precision and recall of enhancer activity

(Analysis shown in Figure 3)

Enrichment of active enhancers (Fig. 3b and c) is defined as the fraction of active enhancers containing a specific modification compared to the fraction of active enhancers in the full enhancer dataset. The full dataset contained 22 active mesoderm enhancers at 6-8 hrs and 144 enhancers total, giving a baseline of 15% active enhancers. The enrichment of active enhancers for H3K27ac is for example: H3K27ac marks 23 out of the 144 enhancers, 13 of these have activity annotation in mesoderm at 6-8 hrs. The enrichment of active enhancers over background is thus (13/23)/(22/144) = 3.7 fold. We note that a number of the 10 H3K27ac marked enhancers that are not annotated as being active have actually no known activity information at 6-8 hrs of development (i.e. flagged with 'ni'). The significance of enrichment was calculated using a two-sided Fisher's exact test. For the enrichment of active enhancers outside mesoderm we excluded enhancers that have activity both in the mesoderm and outside of the mesoderm at 6-8 hrs, reducing the full enhancer dataset to 140 enhancers (described in section VI.2 'Defining enhancer classes with specific spatio-temporal activity').

Precision and Recall (Fig. 3e) were calculated according to their usual definitions: Precision = TP/(TP+FP) while Recall = TP/(TP+FN), where TP is the number of true positives, FP the number of false positives and FN the number of false negatives. Given a set of enriched regions 'S' (histone mark and Pol II peaks as defined by MACS or mesodermal TFs as obtained from Zinzen *et al.*), true positives were defined as the number of active mesodermal enhancers at 6-8 hrs that overlap with a peak of S, false positives as the number of inactive mesodermal enhancers at 6-8 hrs that overlap with a peak of S and false negatives as the number of active mesodermal enhancers at 6-8 hrs that do not overlap with a peak of S. Note, only CAD2 enhancers that are either active (class 'A', n = 22) or fully inactive (class 'I', n = 21) in mesoderm at 6-8 hrs of development were considered (see section VI.2

'Defining enhancer classes with specific spatio-temporal activity'). Thus the list of enhancers for this analysis consisted of 43 entries.

VI.4 Definition of region overlap

Enhancers were considered marked by a histone modification or Pol II if MACS regions were overlapping with the enhancer by at least one base. For TFs, the center of the TF binding region[18,20] had to be inside the enhancer to be considered 'bound'.

VII Gene and CRM intensity profiles

(Quantitative meta-gene analysis shown in Figure 2c-f, Supplementary Figures 3 and 4 and TF-Meso-CRMs intensity profiles shown in Figure 4b-f)

VII.1 Gene intensity profiles

For gene-centered intensity profiles, background subtracted counts per 25-base bins were taken from -500 to +300 bp around the TSS, and -300 to +500 bp around the transcriptional termination site (end) of the genes. The gene signal between +300 bp downstream of the TSS to -300 bp upstream of the gene end was calculated by cubic spline interpolation of the values into 100 bins using the R stats package. For this we required the genes to span ≥ 850 bp to acquire at least 10 independent values for the interpolation. For signal summarization, we used a trimmed mean (10% - 90% of the intensity ranked data) to avoid the impact of outlier values that arise from mis-, or un-annotated genes from BDGP. We chose this rather stringent filter because assessing inactivity is inherently more difficult (and thus error prone) than activity (e.g. genes that have no mesodermal activity annotation by BDGP but seem to be active when looking at the BDGP *in situ* images at 6-8 hrs and at chromatin modifications). The resulting values were smoothed using the mean intensity for every 5 bins. The 95% confidence intervals of the mean were calculated using the trimmed values assuming a Normal distribution of the data. Genes were grouped into four classes according to their annotated expression by *in situ* hybridization by the Berkeley *Drosophila* Genome Project (BDGP) (see section V 'Gene lists using the BDGP *in situ* hybridization database').

VII.2 CRM intensity profiles

For TF-Meso-CRM-centered intensity profiles (Fig. 4b-f), trimmed mean background subtracted counts per 25-base bin were calculated from -1.6 kb to +1.6 kb of the TF-Meso-

CRM center. We used a trimmed mean (10% - 90% of the data) to avoid the impact of outlier values that arise from un-annotated/unfiltered genes (e.g. Pol II peaks on active TSSs are roughly ~10x stronger than on bound CRMs). The resulting values were smoothed using a mean intensity for every 5 bins. The 95% confidence intervals of the mean were calculated using the trimmed values assuming a Normal distribution of the data. We did not plot the 95% confidence interval for the bound '8-12h' class of TF-Meso-CRMs since it only contains 8 CRMs (10 before trimming of the values). For such low sample numbers the Normal distribution does not result in an adequate estimation of the confidence interval of the mean. To obtain scaled values between 1 and 0 (Fig. 4b), the values were shifted by the difference between 0 and the minimum value and then divided by the maximum intensity for a given mark within the region -1.6 to +1.6 kb around the TF-Meso-CRMs.

VIII. Assessment of overlap significance between two region sets

(Significance analysis of overlap between H3K27ac and H3K4me1, between BNFinder enhancer predictions and TF-Meso-CRMs and analysis in Supple. Fig. 5)

Significance (p-value) of overlap between a region set RS_A and a reference region set RS_{Ref} was estimated by bootstrap following recommendations from[23]. 999 sequence sets were randomly generated, each of which contained non-overlapping sequences mimicking sequences of RS_A: for each original region of RS_A, a region of the same size is randomly sampled from the chromosome of the original RS_A region restricted to the allowed genome space. The allowed genome space considered for sampling random regions is matched to RS_A properties and was either the whole genome or the intergenic genome (i.e. whole genome excluding genes±1kb and H3K4me3 MACS peaks). For each random sequence set, we determined the percentage of random regions that overlapped with RS_{Ref}. The observed percentage of overlap between RS_A and RS_{Ref} (P_{obs}) is ranked within the 999 random percentages and the p-value is estimated given the formula: $rank_{obs}/(N_{rdm}+1)$, where $rank_{obs}$ is the rank of P_{obs} and N_{rdm} is the number of random dataset used in the simulation (always 999 in this study). Note that with 999 random sets, 0.001 is the best p-value one can obtain.

This procedure was applied to estimate the p-value of the overlap:

(a) between H3K27ac and H3K4me1 MACS peaks (RS_{Ref} = H3K4me1 peaks, RS_A = H3K27ac peaks, N_{rdm} = 999, random regions sampled from the whole genome),
(b) between BNFinder enhancer predictions and TF-Meso-CRMs (RS_{Ref} = TF-Meso-CRMs,

RS_A = enhancer predictions, N_{rdm} = 999, random regions sampled from the intergenic genome as defined earlier) as presented in section IX.6 '*De novo* prediction of regulatory regions active in mesoderm at 6-8 hrs',
(c) between each of the 7 peak sets called by MACS (H3K4me3, H3K4me1, H3K27me3, H3K27ac, H3K36me3, H3K79me3 and Pol II) and the 144 intergenic CAD entries (RS_{Ref} = each of the MACS peak sets, RS_A = 144 intergenic CAD entries, N_{rdm} = 999, random regions sampled from the intergenic genome as defined earlier) as presented in Supplementary Fig. 5.

IX. Clustering and Bayesian modeling

IX.1 Intensity summarization of H3 modifications and Pol II signals covering CAD2 enhancers

(Analysis linked to Figure 5b, c (clustering and training data for the Bayesian Network)

For quantitative analyses (used for clustering, BNFinder) the level of signal (H3 modifications or Pol II) covering CAD2 enhancers was summarized into a unique value: NormDiff intensity values (using 50bp bins) were summarized into a unique intensity value for each CAD2 enhancer using a moving average approach. A window of defined width (e.g. 1 kb) was moved over each enhancer using a single-bin step size; the maximum observed average was used as the summarized enhancer intensity. For enhancers smaller than the used window size, a single window of that width (e.g. 1 kb) was centered on the enhancer.

IX.2 Clustering

(Analysis linked to Figure 5b, c)

Hierarchical clustering of the CAD2 training set presented in Fig. 5b (Manhattan distance, complete linkage) was performed in the MeV application[24] using NormDiff summarized intensities (1 kb bandwidth) as described above (provided in Supplementary Table 9). Hierarchical clustering of BNFinder predictions presented in Fig. 5c (Euclidean distance, average linkage) was also performed in the MeV application[24] using the NormDiff summarized intensities as used by the BNFinder during the prediction phase (see section IX.4 "Application of the Bayesian Network").

IX.3 Bayesian networks as a predictive model of enhancer activity

Bayesian network (BN) inference is a popular tool for describing multivariate probabilistic models with complex structure of dependencies between variables. The model is

represented as a graph with nodes representing variables and edges representing conditional dependency of probability distributions between them (as presented in Supplementary Fig. 11a). Given the dependency structure, conditional probability distributions can be estimated from observations, however the problem of reconstructing the most likely dependency structure is generally computationally difficult[25]. It was recently shown [26] that in special cases, when the acyclicity of the graph can be ensured, it is possible to find the optimal dependency structure effectively.

In the context of enhancer activity prediction, the model consists of two types of variables: observed histone modification and Pol II occupancy quantitative levels are used as "input" variables and the binary enhancer activity classification variables are used as output. Since we are only interested in recovering connections linking "input" variables to "output" variables, there is no possibility of obtaining a graph with cycles allowing us to use the efficient algorithm as implemented in the BNFinder package[27].

While Bayesian Networks typically use variables with discrete values, BNFinder allows the user to use a mixture model for quantitative observations instead of a threshold-based discretization. In this setting each variable with continuous observations is assumed to have two possible states: "high" and "low" each of which gives rise to experimental observations following a specific Normal distribution. Data for such variables is supplied for training without discretization and the actual value of observed signal is converted to the probability of this observation coming from "high" signal based on estimated mixture model for this variable. BN models can naturally handle probabilities instead of discrete observations facilitating completely automatic treatment of quantitative variables without the negative effects of discretization for classification.

For each member of the training set (each enhancer), we provide a vector of observed continuous signals (for all chromatin modifications and Pol II) describing the state of chromatin at the locus and the binary annotations for the activity. Using a dataset combining a number of such observations, BNFinder recovers the best network structure according to the BDE (Bayesian Dirichlet Equivalence) criteria, which corresponds to the network with maximal posterior probability given the data. This also allows construction of the conditional probability distributions of observing any activity class depending on the measured state of each of the input (histone modifications and Pol II occupancy) included in the network (i.e. conditional dependencies between inputs and outputs or network edges as presented in Supplementary Fig. 11a). Such models can then be used to score unseen examples of

genomic loci and assign posterior probabilities to them of belonging to any activity class (prediction or 'test' phase).

As in any other machine learning procedure, the use of an independent testing set is required to faithfully assess the accuracy of the method. To this end we employed a cross-validation scheme for BN training. This is achieved by dividing the training set into **n** cross-validation "folds" and subsequently using each of them as a testing set for the BN trained on the remaining **n-1** folds (in this study, n=4). This procedure ensures that all the examples were scored against a model trained on an independent dataset. In addition, we can use a similar strategy to assess the robustness of inferred relationships between histone marks and Pol II and the activity classes by counting how often we recover each relationship from sub-samples of the original dataset. If a relationship is robust, we expect to see it in a majority of networks obtained via such a sub-sampling strategy.

Our approach is a supervised learning strategy, which suits our purpose of finding histone modifications predictive of independently assayed activity of known enhancers very well. It is in contrast with recently published unsupervised methods used for genome-wide *de novo* annotations of human[28] or Drosophila[18] genomes based on chromatin marks integrated in a Hidden Markov Model. While such unsupervised strategies discover the most frequently occurring constellations of histone marks and correlate them post hoc with known annotations, they face great difficulty in discovering rare combinations of marks and there is no possibility of ensuring that the resulting model will discern between similarly marked regions (e.g. enhancers vs promoters) or discriminate between the features of interest (e.g. active versus inactive enhancers). Since our model is recovering relationships between combinations of marks and Pol II and activity state, it is natural to ask whether a simpler model based on linear or logistic regression (e.g. similar to the one used by Karlic *et al.*[29] in their recent study linking histone marks with gene expression levels) could be used. While it might be argued that a regression model estimated with the LASSO method could produce similar results, the Bayesian method is not limited to additive functions and indeed recovers some non-additive interactions (e.g. Pol II and H3K27ac, probability table in Fig 5b). In addition to this advantage, it is not susceptible to issues related to differences in dynamic ranges between Pol II signals (normalized to input) and histone marks (normalized to H3), which could lead to an artificial dominance of one of the inputs. Finally, the ability of the Bayesian model to produce scores that can be interpreted as posterior probability of activity makes it easier to interpret the significance of choosing different thresholds.

IX.4 Application of the Bayesian Network

(Analysis linked to Figure 5)

BNFinder (version 3.3, available at http://bioputer.mimuw.edu.pl/software/bnf/)[27] was used to train a Bayesian Network to understand the relationship between H3 histone modifications and Pol II occupancy (the regulators or inputs) and enhancer activity states (the outputs). We chose two different activity states: (1) expression in mesoderm (no time constraint) and (2) expression in mesoderm at the examined time (6-8 hrs), taking advantage of the temporal enhancer activity annotation in CAD2. CAD2 entry expression was first summarized as "expressed in the mesoderm" and "expressed outside the mesoderm" using four time windows ('2-4 hrs' using stages 5-7, '4-6 hrs' using stages 8-9, '6-8 hrs' using stages 10-11 and '8 hrs+' using stages 12 and onward, i.e. after 8 hrs) and three expression values: '1' for expressed, '0' for not expressed and 'NA' for missing value. For example, the "*sug*" CRM (chr2R:8813219-8814579, dm3) has the following summarized expression vector {NA,1,1,1,NA,0,0,0} in which the first four values represent mesodermal expression for '2-4 hrs', '4-6 hrs', '6-8 hrs' and '8 hrs+' respectively and the last four the expression outside the mesoderm (in the same order). The training set contained CAD2 enhancers that (1) are located at least 1 kb from any gene boundaries (to avoid interference with gene-related signals), (2) have a maximum of one missing value ('NA') at 6-8 hrs (see above) and (3) contain no H3K4me3. This last filter removed TF-Meso-CRM-633, chr2R:9224752-9225751 (dm3) expression profile {1,1,1,0,0,0,0,0}, as careful inspection revealed that it sits at the promoter of an un-annotated gene, revealed by the high H3K4me3 signal level (Hierarchical clustering in Fig. 5b, first row). In total, 65 CAD enhancers match these conditions: 22 are active in the mesoderm at 6-8 hrs, 36 are active outside the mesoderm at 6-8 hrs, 5 have activity both inside and outside the mesoderm at 6-8 hrs, and 12 enhancers are inactive both inside and outside the mesoderm at 6-8 hrs. Summarized intensity values were obtained as described earlier using a window size of 200bp for marks exhibiting a peak-like shape (Pol II and H3K4me3) and a 1 kb window for marks that spread over larger regions (H3K4me1, H3K27ac, H3K27me3, H3K79me3, H3K36me3).

Missing values for enhancer activity annotation (i.e. 'NA' values) were considered as '0', i.e. as not expressed – for example, an enhancer with the expression profile {1,1,NA,0,0,0,0,0} is considered 'expressed in the mesoderm' and 'not expressed in the mesoderm at 6-8 hrs' due to the missing value for mesodermal expression at 6-8 hrs. A file with regulators' intensities and enhancer activity states extracted from expression profiles was

assembled for the 'bnf' tool (provided in Supplementary Table 10)[27] (--data-factor option was set to 1000000, the default BDE scoring function was used). The trained network is shown in Fig. 5b and Supplementary Fig. 11a.

The accuracy of the trained Bayesian Networks was assessed using a 4-fold cross-validation scheme. Cross-validation results are presented in Supplementary Fig. 11a as TPR/FDR ROC curves and were obtained using the 'bnf-cv' tool of the BNFinder package (options common to 'bnf' were identical, k was set to 4). The robustness of inferred relationships was estimated by running the 4-fold cross-validation 250 times (i.e. summing up to 1000 training runs) and counting how many times a particular edge was reported (network in Supplementary Fig. 11a). The final Bayesian Network (learned with all 65 training set enhancers) is shown in Supplementary Fig. 11a and its performance is presented in Supplementary Fig. 11b (TPR/FDR ROC and Precision/Recall curves; these were obtained using bnf-cv with k set to 1). Figure 5b shows the conditional probability table (extracted from BNFinder output presented in next section) listing the posterior probabilities of an enhancer being active and inactive in the mesoderm at 6-8h given H3K27ac, H3K79me3 and Pol II present/absent states. Globally, validation results show that both states can be reliably estimated based on the data. The best classification performance is obtained for the "Expression in mesoderm at the examined time" state (area under the ROC curve: 0.82 with a precision of 100% up to a recall of ~40%), which uses various levels of H3K27ac, H3K79me3 and Pol II as regulators. H3K27me3 is reported to be negatively associated with mesodermal expression.

Lastly, we verified that the Bayesian classifier uniformly scores proximal and distal enhancers active in mesoderm at 6-8 hrs. The posterior probability of an enhancer to be active in the mesoderm at 6-8 hrs was computed for each 65 CAD2 enhancers of the training set using the model trained to predict enhancers active in the mesoderm at 6-8 hrs. The mean posterior probabilities as a function of the distance between these 65 CAD2 enhancers (using enhancers' centers) and the TSS of their target genes (or of the closest TSS when the target gene was unknown) was plotted (Supplementary Fig. 12) using the following procedure: the mean posterior probabilities for active and inactive enhancers were computed using a running window of 4000bp by steps of 1000 bp (only windows with at least 5 enhancers were considered). As shown on Supplementary Fig. 12, no bias due to TSS proximity was observed, confirming that the trained model can be applied to the whole intergenic genome.

IX.5 BNFinder output: Conditional probabilities from the trained Bayesian network


```
     {
       'K27Ac_6-8h' :   {
         'cpds' :     {
           None : 0.5,
           () :       {
             None : 0.014925373134328358,
             0 : 0.79358467163638391,
             1 : 0.20641532836361609,
           },
         },
         'floatParams' : '(0.00066371522000000405, 1.623554267533333,
0.44891250780649417, 0.76923076923076927, 0.23076923076923078)',
         'pars' : [],
         'vals' : [],
       },
       'K27me3_6-8h' :   {
         'cpds' :     {
           None : 0.5,
           () :       {
             None : 0.014925373134328358,
             0 : 0.52877506667568142,
             1 : 0.47122493332431853,
           },
         },
         'floatParams' : '(0.41917207761764697, 1.6956007453225803,
0.3569109716479309, 0.52307692307692311, 0.47692307692307695)',
         'pars' : [],
         'vals' : [],
       },
       'K36me3_6-8h' :   {
         'cpds' :     {
           None : 0.5,
           () :       {
             None : 0.014925373134328358,
             0 : 0.52913633928092096,
             1 : 0.47086366071907931,
           },
         },
         'floatParams' : '(-0.97368754708823557, -0.67679223716129044,
0.11636108917651469, 0.52307692307692311, 0.47692307692307695)',
         'pars' : [],
         'vals' : [],
       },
       'K4me1_6-8h' :   {
         'cpds' :     {
           None : 0.5,
           () :       {
             None : 0.014925373134328358,
             0 : 0.44082866252944897,
             1 : 0.55917133747055092,
           },
         },
         'floatParams' : '(0.35367498679310344, 2.2124551275555557,
0.59777506658900637, 0.44615384615384618, 0.55384615384615388)',
         'pars' : [],
         'vals' : [],
       },
       'K4me3_6-8h' :   {
         'cpds' :     {
           None : 0.5,
           () :       {
             None : 0.014925373134328358,
             0 : 0.47182857060476407,
             1 : 0.5281714293952362,
           },
         },
```

```
        'floatParams' : '(0.1136699421935484, 0.65945637017647063,
0.1956842581222748, 0.47692307692307695, 0.52307692307692311)',
        'pars' : [],
        'vals' : [],
      },
      'K79me3_6-8h' :   {
        'cpds' :     {
          None : 0.5,
          () :       {
            None : 0.014925373134328358,
            0 : 0.73814824281312486,
            1 : 0.26185175718687559,
          },
        },
        'floatParams' : '(-0.3179687559148936, 0.78704234205555534,
0.31981414490555626, 0.72307692307692306, 0.27692307692307694)',
        'pars' : [],
        'vals' : [],
      },
      'PolII_6-8h' :   {
        'cpds' :     {
          None : 0.5,
          () :       {
            None : 0.014925373134328358,
            0 : 0.8085037277517434,
            1 : 0.19149627224825641,
          },
        },
        'floatParams' : '(0.0049082834901960845, 1.0598374132142856,
0.27928208861427284, 0.7846153846153846, 0.2153846153846154)',
        'pars' : [],
        'vals' : [],
      },
      'expInMeso' :   {
        'cpds' :     {
          None : 0.5,
          (0, 0, 0) :       {
            None : 0.050474507877100318,
            0 : 0.49120413112282157,
            1 : 0.50879586887717887,
          },
          (0, 0, 1) :       {
            None : 0.035006964753929018,
            0 : 0.73609081840911161,
            1 : 0.26390918159088794,
          },
          (0, 1, 0) :       {
            None : 0.17389366118532748,
            0 : 0.55854485789317521,
            1 : 0.44145514210682502,
          },
          (0, 1, 1) :       {
            None : 0.42963335885316389,
            0 : 0.4556377722679007,
            1 : 0.5443622277320993,
          },
          (1, 0, 0) :       {
            None : 0.13910763034904289,
            0 : 0.40392563268850484,
            1 : 0.59607436731149521,
          },
          (1, 0, 1) :       {
            None : 0.21721331412329556,
            0 : 0.62245333922181312,
            1 : 0.3775466607781871,
          },
          (1, 1, 0) :       {
            None : 0.10334177199952997,
```

```
      0 : 0.11704892410268979,
      1 : 0.88295107589731026,
    },
    (1, 1, 1) :       {
      None : 0.3252042058790629,
      0 : 0.34693143429624779,
      1 : 0.65306856570375227,
    },
  },
  'floatParams' : 'None',
  'pars' : ['K79me3_6-8h', 'K27Ac_6-8h', 'K27me3_6-8h'],
  'vals' : ['0', '1'],
},
'expInMesoAtTime.no.NA' :   {
  'cpds' :     {
    None : 0.5,
    (0, 0, 0) :       {
      None : 0.023159468772417237,
      0 : 0.80537106910656964,
      1 : 0.19462893089343028,
    },
    (0, 0, 1) :       {
      None : 0.22009915794498627,
      0 : 0.48668668407992621,
      1 : 0.51331331592007379,
    },
    (0, 1, 0) :       {
      None : 0.10553840214126624,
      0 : 0.66717056817406173,
      1 : 0.33282943182593855,
    },
    (0, 1, 1) :       {
      None : 0.25174739807774354,
      0 : 0.33666824274259244,
      1 : 0.66333175725740756,
    },
    (1, 0, 0) :       {
      None : 0.19235052576311498,
      0 : 0.41842257647650166,
      1 : 0.58157742352349828,
    },
    (1, 0, 1) :       {
      None : 0.28290142502690047,
      0 : 0.59558475154872181,
      1 : 0.4044152484512783,
    },
    (1, 1, 0) :       {
      None : 0.43155092244474735,
      0 : 0.45529602792260149,
      1 : 0.54470397207739851,
    },
    (1, 1, 1) :       {
      None : 0.11390320251367368,
      0 : 0.11609426184485547,
      1 : 0.8839057381551445,
    },
  },
  'floatParams' : 'None',
  'pars' : ['PolII_6-8h', 'K79me3_6-8h', 'K27Ac_6-8h'],
  'vals' : ['0', '1'],
},
}
```

IX.6 *De novo* prediction of regulatory regions active in mesoderm at 6-8 hrs

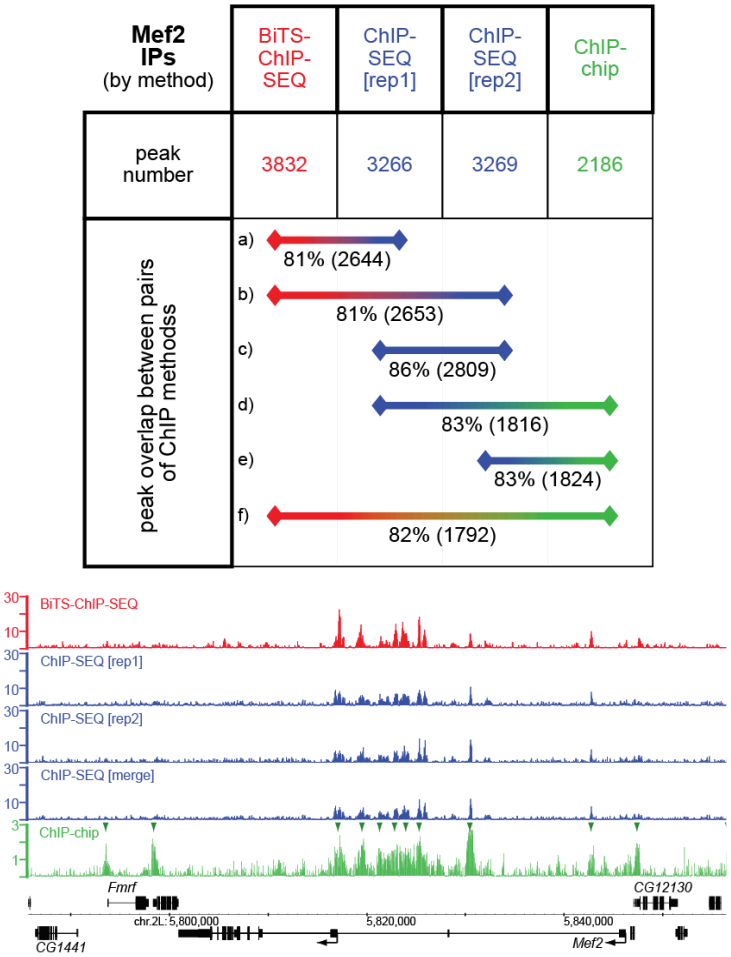
(Analysis linked to Figure 5c)

De novo prediction of regions active in mesoderm at 6-8 hrs was performed as follows: every 1 kb window spanning the intergenic genome was covered using a 50bp step (the used bin size). Windows located at least 1 kb away from gene boundaries and not overlapping with H3K4me3 enriched regions (as reported by MACS) were considered further. Each 1 kb window was considered as a potential regulatory sequence and the average NormDiff over the whole window was computed for H3K27ac and H3K79me3. Pol II summarized intensity was computed as previously explained (best NormDiff moving average found within the 1 kb window using a bandwidth of 200bp). Note that the "Expression in mesoderm at the examined time (6-8 hrs)" state only depends on these 3 regulators (H3K27ac, Pol II and H3K79me3). Windows were scored using the 'bnc' tool of the BNFinder package using conditional probabilities from the Bayesian network trained on the whole dataset (section IX.5). Windows with a posterior probability greater than 0.582 (i.e. corresponding to a precision of 100% and a recall of 36% as recomputed by overlapping CAD2 with the set of final predictions) were further selected and overlapping windows merged. Resulting unique regions retain the best score of merged overlapping 1 kb windows. A total of 121 regions were defined using this procedure. As the learned BN model does not consider H3K4me1 presence (as both active and inactive enhancers have significant levels of H3K4me1), it identifies putative regulatory regions of two classes: 1) those with high levels of Pol II, which most likely represent an unannotated group of mesoderm specific stalled TSSs and 2) those with detectable level of H3K4me1, which we consider as good candidates for active enhancers. To obtain a list of predicted active enhancers, we therefore post-filtered the BN predictions for a minimum NormDiff level of H3K4me1 of 0.5 and a maximum NormDiff level of Pol II of 2. The final 112 predicted active enhancer regions are reported in Supplementary Table 4. We note that as these regions have an average size of 2.7 kb (median size of 2.2 kb), they most likely contain multiple *cis*-regulatory modules. A comparison of these regions to our TF-Meso-CRMs[20] confirms this – 87 of the 112 (78%) predicted regions contain 146 TF-Meso-CRMs bound by at least one TF at 6-8h. The significance of this overlap was assessed using the procedure described in the section VIII "Assessment of overlap significance between two region sets". The maximum observed random overlap was 12.5% corresponding to an estimated p-value of 0.001 (the best p-value that could be obtained given the number of randomization).

IX.7 Testing of predicted regulatory regions

Candidate regulatory regions were select from BNFinder predictions covering the range of posterior probabilities (PP) above the cut-off (PP=0.582, corresponding to a specificity of 100%) to the maximum of 0.884. Tested regions were chosen to represent a variety of histone and Pol II signatures. As predicted regions are rather large and may encompass several CRMs, tested regions were chosen to include distinctive chromatin signatures; in some cases, more than one region was cloned within a specific region to assay chromatin/PolII signature dependencies.

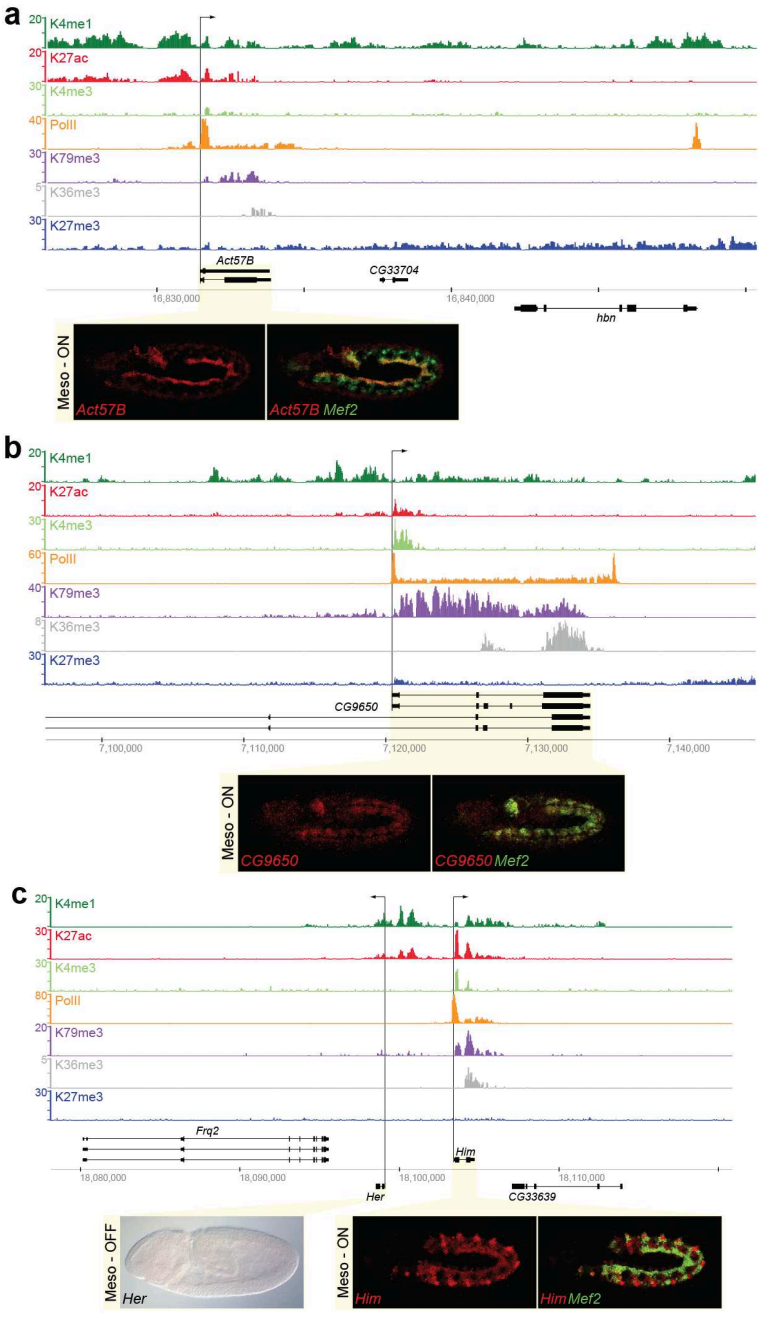
Supplementary Figures



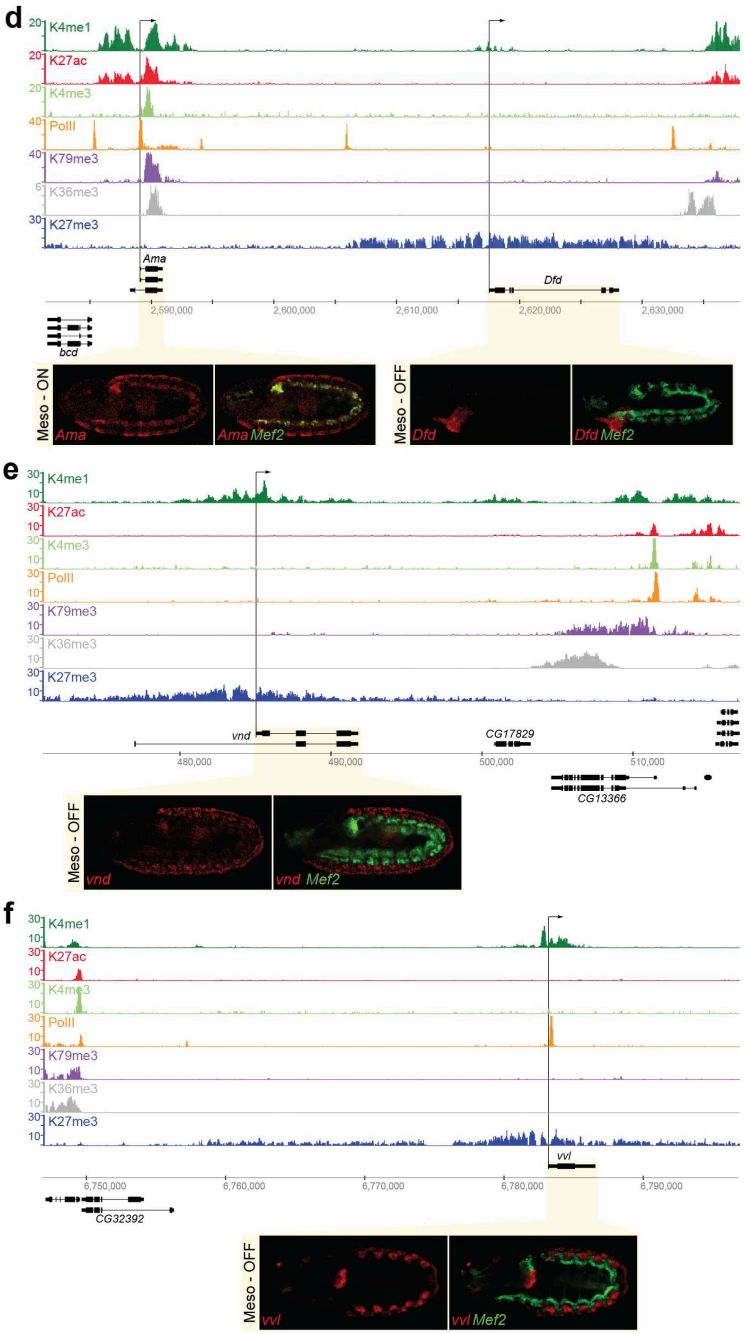
Supplemental Fig. 1: Comparison of Mef2-ChIP data generated by three methods

Top: *Overlap of Mef2 bound regions as identified by three methods.*
Genome-wide binding of Mef2, a mesoderm-specific TF, was assessed via immunoprecipitation from fixed sorted nuclei followed by Solexa sequencing (BiTS-ChIP-SEQ, red), regular ChIP-SEQ (two replicates are shown in blue), and ChIP followed by microarray analysis (ChIP-chip, green). The percentage overlap of identified peaks (by MACS for deep sequencing, by TileMap for chip) between dataset pairs is indicated (a-f). Note that the percentage overlap between BiTS-ChIP-SEQ and ChIP-chip (f; 82%) is similar to the overlap between normal ChIP-SEQ and ChIP-chip (d, e; 83%)

Bottom: *Mef2 locus showing high concordance between the three data sets*. Shown is Mef2 signal enrichment generated by BiTS-ChIP-SEQ (red), by regular ChIP-SEQ (blue, 2 replicates are shown individually, as well as merged), and ChIP-chip data (green, significant binding peaks reported are indicated by arrowheads); genomic model below. Additional regions are shown in Fig. 1d.



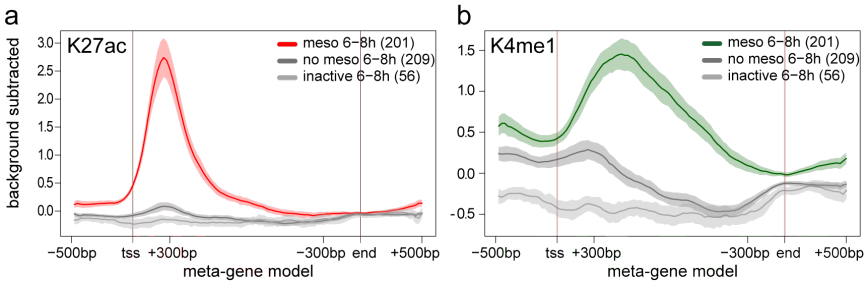
Supplemental Fig. 2 (part 1): Chromatin and Pol II signatures in mesodermally active and inactive gene loci.



Supplemental Fig. 2 (part 2): Chromatin and Pol II signatures in mesodermally active and inactive gene loci.

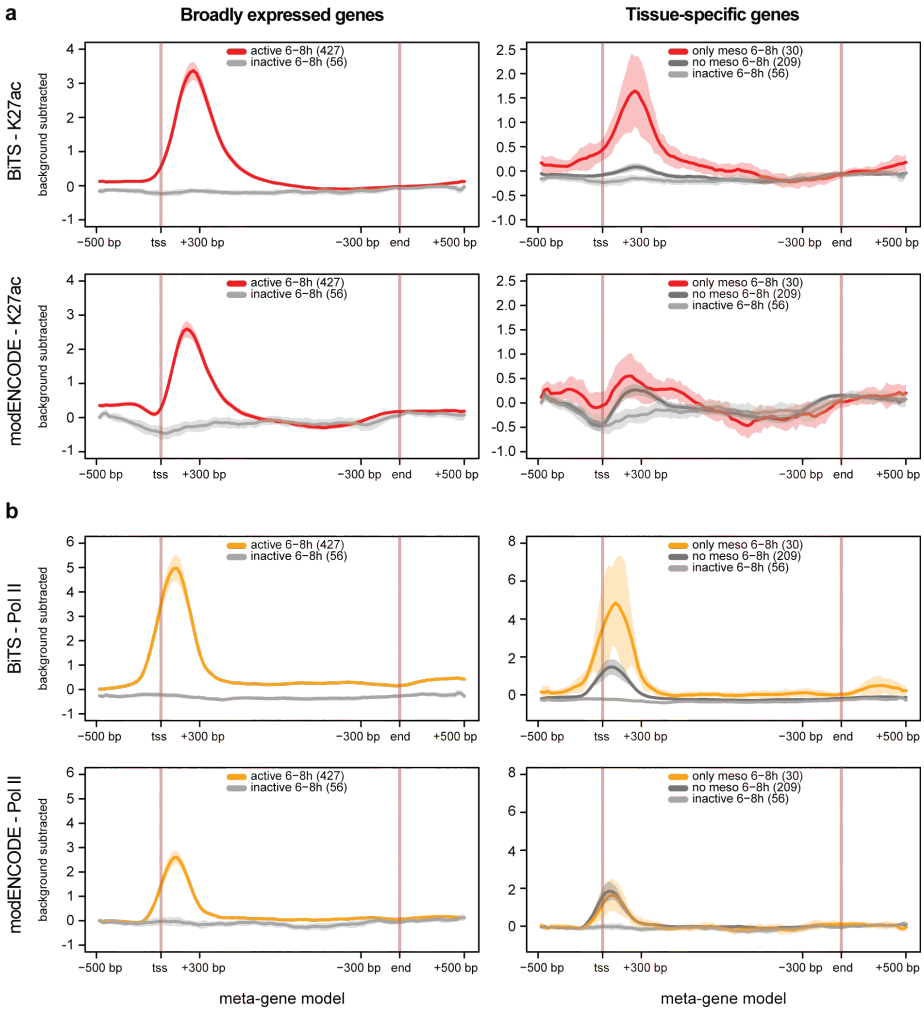
Supplemental Fig. 2 (legend): Chromatin and Pol II signatures in mesodermally active and inactive gene loci.

Shown are the chromatin mark and Pol II enrichments (background subtracted signal) for active **(a-d)** and mesodermally inactive (**c-f**) genes. Arrows indicate the promoter and the direction of transcription of the genes for which RNA *in situ* images are shown. *In situ* hybridizations of mesodermally active genes (*Act57B* (**a**); *CG9650* (**b**); *him* (**c**); *Ama* (**d**)) and mesodermally inactive genes (*Her* (**c**); *Dfd* (**d**); *vnd* (**e**); *vvl* (**f**)) are shown below the enrichment and gene model tracks in each panel (the *Her* in situ was taken from the BDGP gene expression database, note the complete absence of expression).

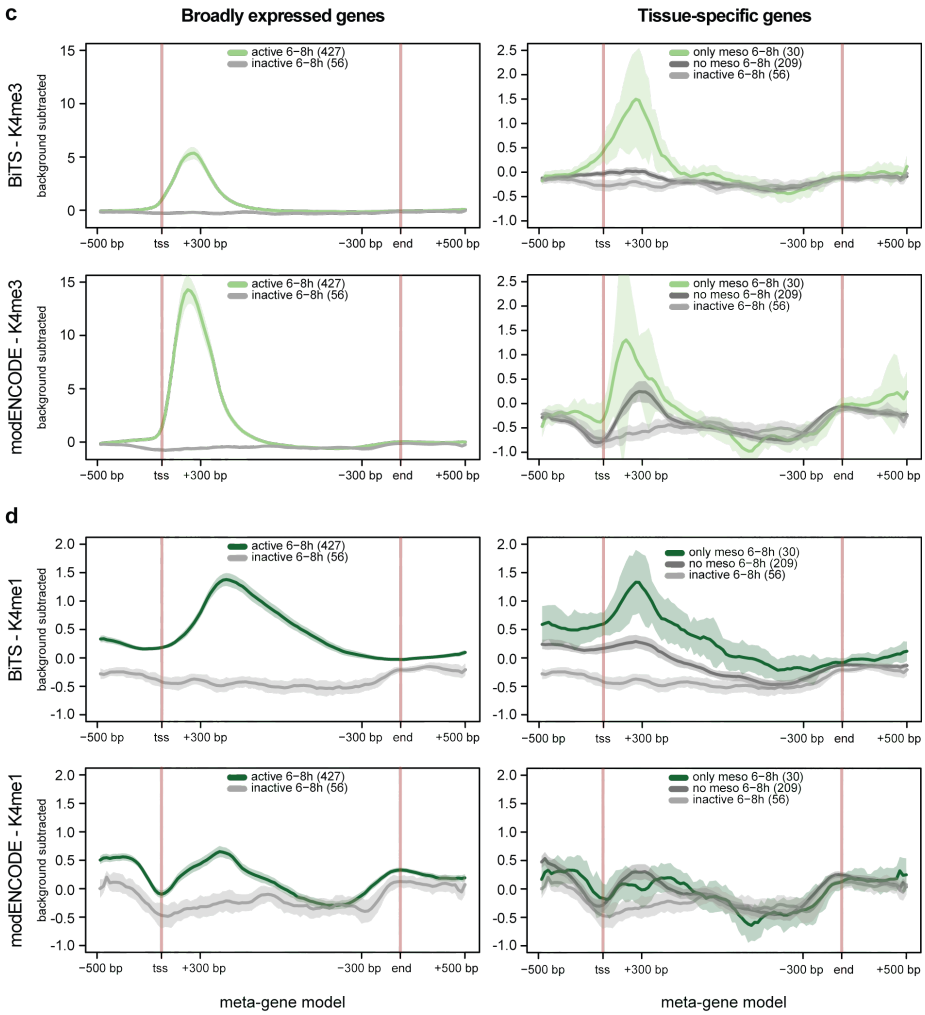


Supplemental Fig. 3: Global assessment of tissue-specificity

Trimmed mean background subtracted signal is shown for H3K27ac (**a**), and H3K4me1 (**b**) with shading indicating the 95% confidence intervals of the signal (see Supplementary Note section VII 'Gene and CRM intensity profiles'). Gene activity information was obtained from large scale *in situ* hybridization data (BDGP) and signal was visualized for genes active in mesoderm at 6-8 hrs ('meso 6-8h' excluding ubiquitously expressed genes, coloured lines, 201 genes), genes active only in other cell types at 6-8 hrs ('no meso 6-8h', dark grey lines, 209 genes) or genes active in other cell types at later stages ('inactive 6-8h', light grey lines, 56 genes). Mesodermal gene activity is accompanied by an increase in K27ac peak ~250 bp downstream of the transcriptional start site (TSS) (a), and an increase in H3K4 monomethylation peaking further downstream of the TSS (b). Genes expressed outside of mesoderm at 6-8 hrs or inactive genes show strongly reduced to no signal, in comparison. (Similar analyses of H3K4me3, H3K79me3, H3K36me3 and Pol II are shown in Figure 2c-f).



Supplemental Fig. 4 (part 1): Comparison of tissue-specific (BiTS) vs. whole-embryo (modENCODE) ChIP-Seq data



Supplemental Fig. 4 (part 2): Comparison of tissue-specific (BiTS) vs. whole-embryo (modENCODE) ChIP-Seq data

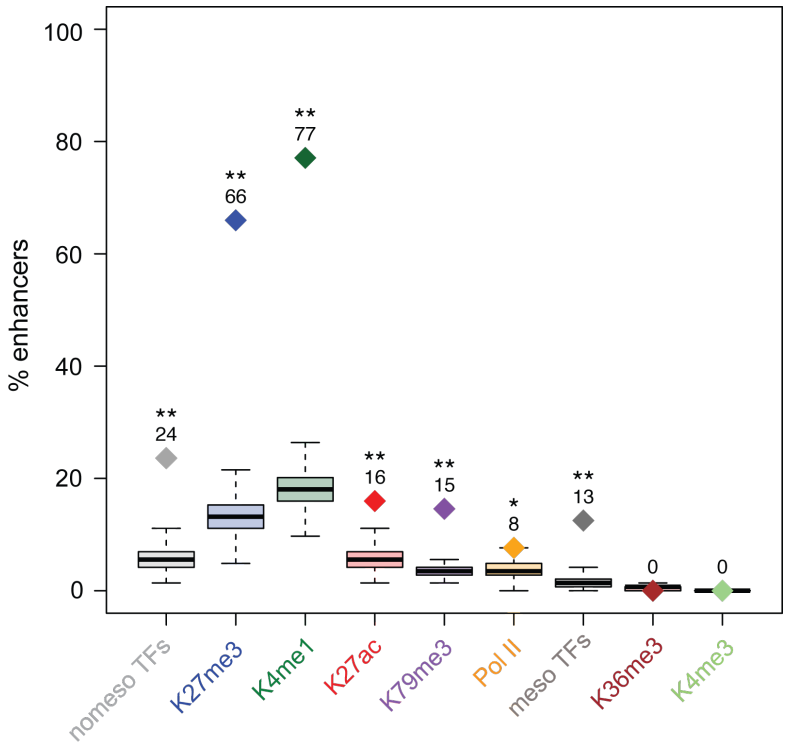
Comparison of mean signal intensities for tissue-specific data (BiTS) or whole-embryo data (modENCODE) for genes expressed in most parts of the embryo (Broadly expressed genes) or genes expressed in small tissues (Tissue-specific genes). The gene sets, calculation of mean values and 95% confidence intervals of the mean (shaded areas) are described in the Supplementary Note sections V 'Gene lists using the BDGP *in situ* hybridization database' and VII 'Gene and CRM intensity profiles', respectively.

Before comparing the quantitative signal (mean background subtracted) for BiTS and modENCODE for mesoderm-specific genes ('Tissue-specific genes', second column) we assessed the quality of the two datasets to determine if they are generally comparable ('Broadly expressed genes', first column). To this end we compared signal intensities for 427 genes expressed throughout most of the embryo (active 6-8h, colored line, first column) or 56 genes expressed only later during development (inactive

6-8h, grey line, first column) for **(a)** H3K27ac, **(b)** Pol II, **(c)** H3K4me3, and **(d)** H3K4me1. As the 'active' and 'inactive' genes are common to both, mesoderm-specific BiTS and whole-embryo modENCODE data, they allow for a direct comparison of the general quality of both datasets. Comparing the signals of the two datasets at these widely expressed genes (left hand panels) reveals two important points: 1) Both tissue-specific data and whole-embryo data can nicely distinguish between the mean signal of expressed versus unexpressed genes at 6-8 hrs. The non-overlap between the 95% confidence intervals of the active (colored) and inactive (grey) lines shows that this is significant. 2) There is no major difference between the general quality of the two datasets when looking at widely expressed genes, which is important for the tissue-specific comparison described below. More specifically, the Pol II **(b)** and H3K4me1 **(d)** BiTS data (left hand panels) shows higher enrichment for active genes compared to modENCODE, while modENCODE H3K4me3 **(c)** data performs better than BiTS, and both datasets seem to perform equally well for H3K27ac **(a)**. These differences in performance for different marks most probably reflect the general quality of the IP and/or the antibody used, library preparation or sequencing. As both tissue-specific and whole-embryo data perform equally well for the chromatin modification H3K27ac **(a)**, this mark is used as the main point of reference for the comparison of sensitivity and specificity of tissue-specific versus whole-embryo data.

We next compared the sensitivity and specificity of tissue-specific BiTS data and whole-embryo modENCODE data at 'Tissue-specific genes' (right hand panels). The mean background subtracted signal intensities were compared for all 30 genes annotated by BDGP to be exclusively expressed in mesoderm (only meso 6-8h; colored line, second column) or 209 genes expressed only outside of mesoderm (no meso 6-8h; dark grey line, second column) at 6-8 hrs during development. The mean value for genes not expressed at 6-8h but only later in development, was plotted as a negative control for genes that are clearly inactive at this stage of development (inactive 6-8h; light grey line). This analysis highlights two key points: (1) BiTS data has the tissue specificity to distinguish between mesoderm-specific genes and non-mesoderm genes, while modENCODE whole-embryo data cannot: For all genes examined, BiTS data shows significant differences (as estimated from the confidence intervals) between genes expressed exclusively in the mesoderm (only meso 6-8h, colored lines, second column) and genes expressed only outside of the mesoderm (no meso 6-8h, dark grey lines, second column) at 6-8 hrs. In contrast the signals for mesoderm specific genes and non-mesoderm genes are indistinguishable with modENCODE data (seen by the overlapping colored and dark lines). Although expected, this confirms the specificity of the BiTS method. (2) BiTS data has increased sensitivity to detect tissue-specific genes, including those expressed in small populations of cells. Genes with broad expression in multiple tissues of the embryo have roughly equal quantitative levels of H3K27ac signal in whole-embryo versus tissue-specific data (shown in panel **a**, left hand column), having a peak height of ~ 3 RPGC (H3 background subtracted). However, the level of signal is much lower on mesoderm-specific genes using modENCODE data compared to BiTS data (shown in panel **a**, right hand column). Tissue-specific BiTS data has almost no signal loss comparing H3K27ac data for genes expressed throughout the embryo versus genes only expressed in small populations of cells (panel **a**, first row). In contrast, the quantitative levels of signal for whole-embryo modENCODE data is strongly reduced when genes are expressed only in specific tissues (panel **a**, second row), for H3K27ac and H3K4me1 the signal is barely above background (panel a and d, second row, grey lines).

In summary, BiTS-ChIP-Seq provides information with high tissue-specificity and is conservatively ~3 times as sensitive as whole embryo ChIP-Seq in detecting tissue-specific (non-ubiquitous) signals, thereby facilitating analysis of genes with very restricted expression.

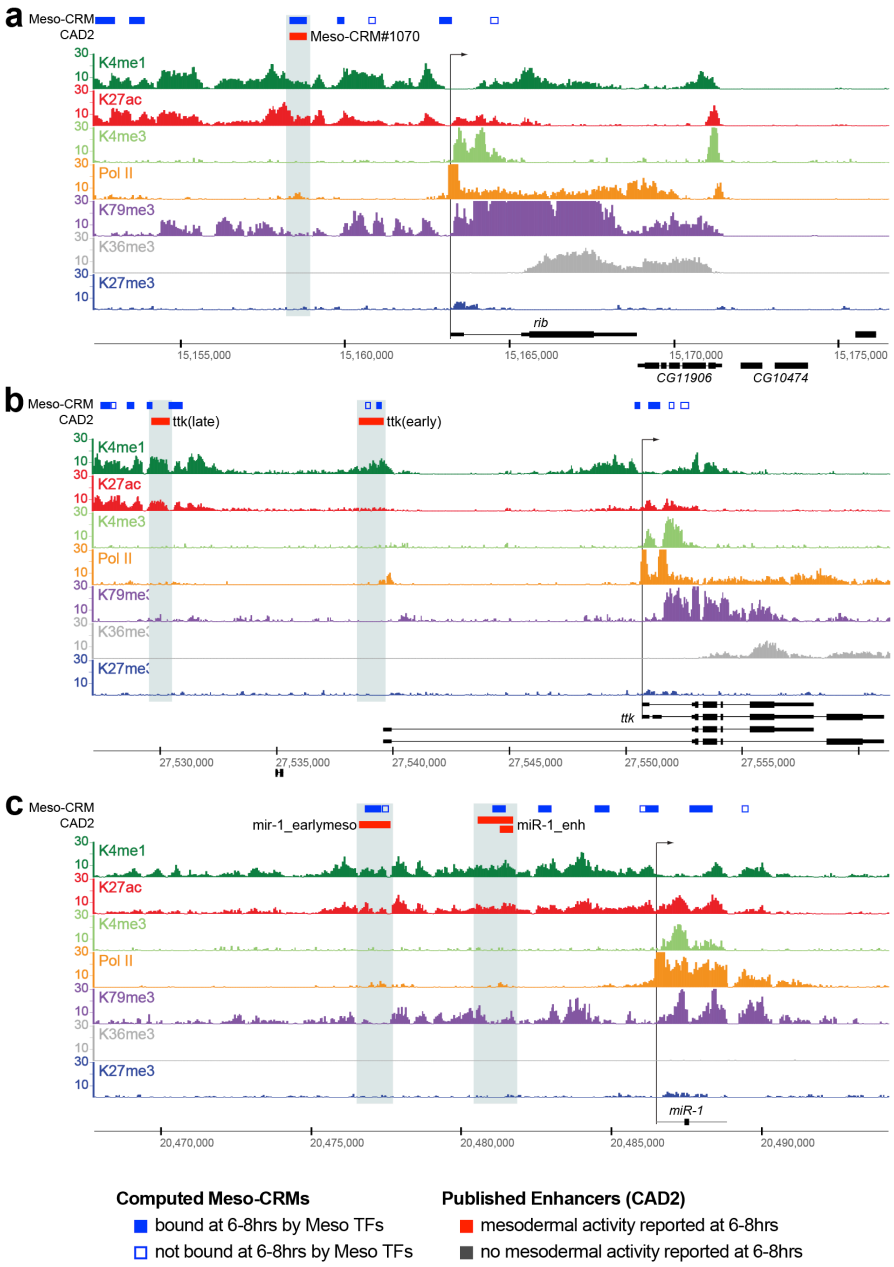


Supplementary Fig. 5: The distribution of chromatin modifications, Pol II and TF occupancy on developmental enhancers

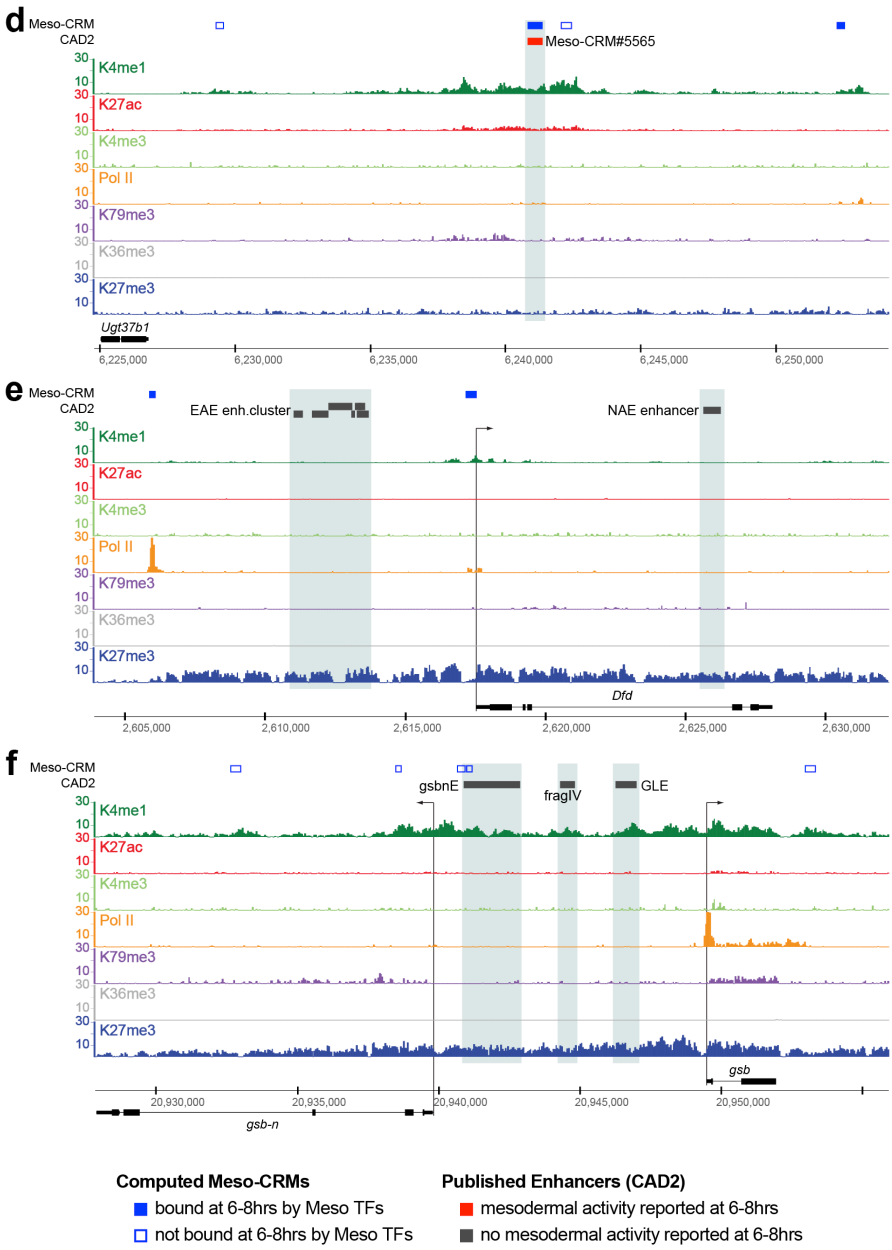
Solid diamonds show the observed percentage of developmental enhancers in the CAD2 database containing indicated chromatin modifications, Pol II or TF occupancy, regardless of their activity state (the number is indicated above the diamond). As for all other analyses, intragenic enhancers (within known genes or within +/- 1 kb) and promoter proximal enhancers (i.e., H3K4me3-positive) were excluded to avoid confounding enhancer with gene body signals, leaving 144 CAD2 enhancers. Interestingly, although two chromatin marks examined are associated with Poll II elongation (H3K36me3 and H3K79me3), only K79me3 is found on enhancers.

The boxplots show the distribution of percentages obtained using 999 random sets of 144 size-matched intergenic regions (excluding known genes +/- 1 kb and H3K4me3 positive regions).

Significant differences of the observed values (diamonds) from the random distribution (boxplots) were estimated by bootstrapping (see Supplementary Note section VIII 'Assessment of overlap significance between two region sets'). Estimated p-values: * p <= 0.05; ** p = 0.001, where 0.001 is the lowest possible p-value that can be obtained.



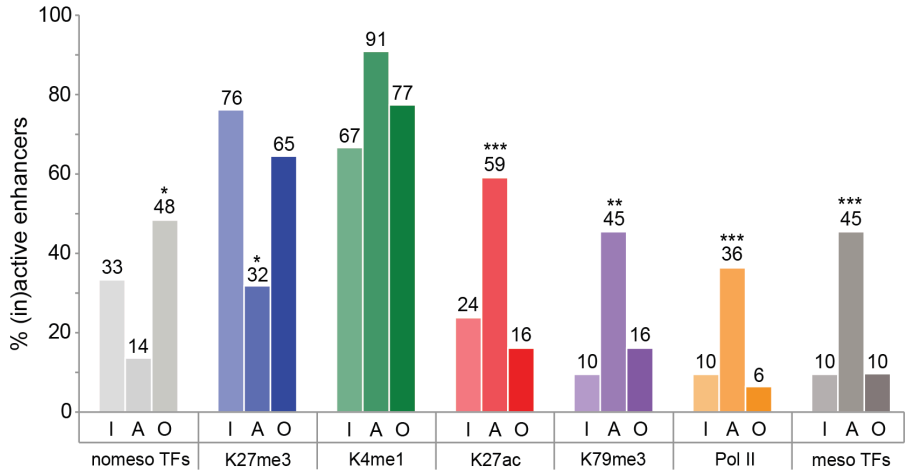
Supplemental Fig. 6 (part 1): Histone H3 modifications and Pol II occupancy at active and inactive developmental enhancers



Supplemental Fig. 6 (part 2): Histone H3 modifications and Pol II occupancy at active and inactive developmental enhancers (legend on following page)

Supplemental Fig. 6 (legend): Histone H3 modifications and Pol II occupancy at active and inactive developmental enhancers

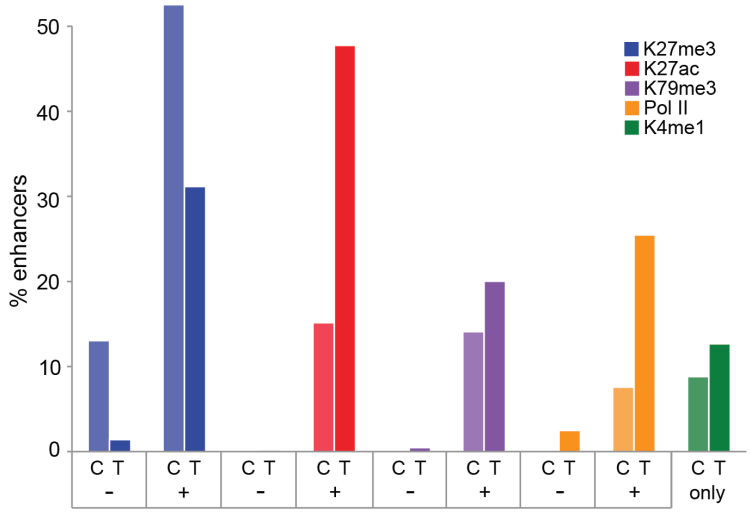
Genomic loci containing active **(a-d)** and inactive **(e-f)** enhancer elements (indicated by the grey shading) showing BiTS-ChIP enrichment for histone modifications and Pol II. Tracks for all chromatin marks and Pol II show background subtracted signal (see Supplementary Note section II.3 'ChIP-seq read summarization and binning'). Blue boxes indicate CRMs defined by mesoderm transcription factor occupancy (TF-Meso-CRMs); solid box = bound at 6-8h, outlined = not bound at 6-8h. Red boxes indicate known mesoderm enhancers active at 6-8hrs, grey boxes indicate enhancers inactive in mesoderm (from CAD2). Promoter arrows indicated the enhancers' target genes.



Supplementary Fig. 7: Linking enhancer signatures to spatial activity

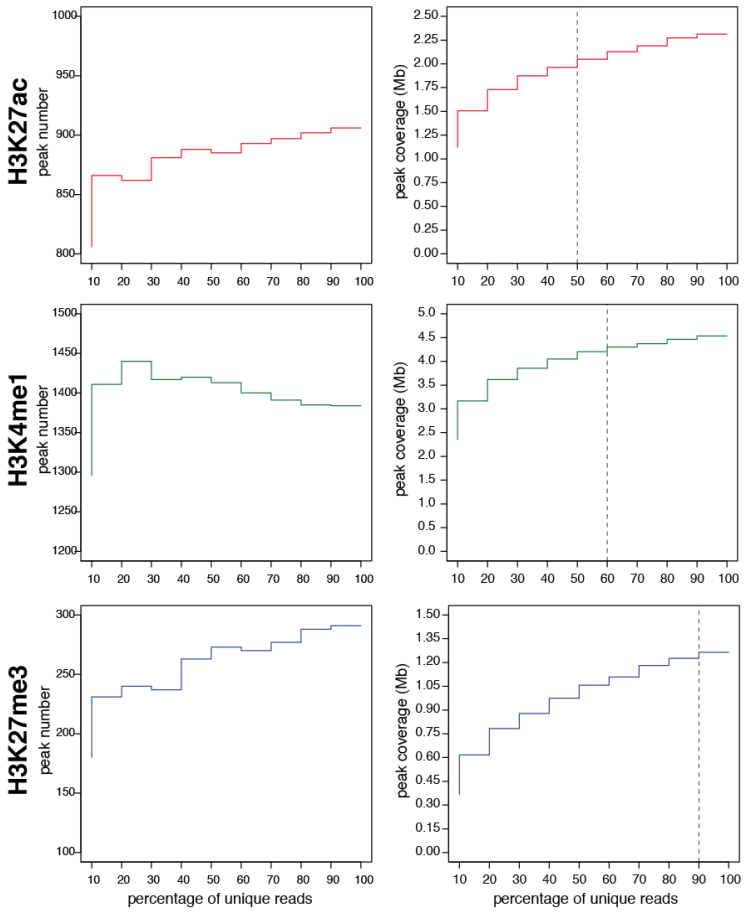
The presence of chromatin marks and Pol II correlates with enhancer activity: CAD-enhancers were divided into enhancers *inactive* at 6-8 hrs (n = 21; '**I**'), enhancers active *outside* of mesoderm at 6-8 hrs (no mesoderm activity annotated at 6-8 hrs; n = 31; '**O**') and enhancers *active in mesoderm* at 6-8 hrs (n = 22; '**A**'). The bar graph shows the percentage of enhancers inactive in mesoderm (I), active in mesoderm (A) or active outside of mesoderm (O), that contain individual chromatin marks, Pol II, or non-mesoderm or mesodermal TFs. H3K27ac, H3K79me3 and Pol II have the highest specific recall for active enhancers, while non-mesoderm TFs (nomeso-TFs) and H3K27me3 having higher recall for inactive enhancers (I and O).

P-values were calculated as in figure 3a, using a two-sided Fisher's exact test. In brief, it was tested if a set of enhancers marked by a specific chromatin modification (TF or Pol II) was significantly enriched for enhancers that are inactive in mesoderm (I), active in mesoderm (A) or active outside of mesoderm (O) at 6-8 hrs of development, as compared to the respective fraction of active enhancers in the total dataset (described in detail in the Supplementary Note section VI.3 'Enrichment, precision and recall of enhancer activity'). For enhancers active in mesoderm at 6-8 hrs (A), the p-values match those of figure 3a. (* p <= 0.05; ** p <= 0.001; *** p <= 0.0001).

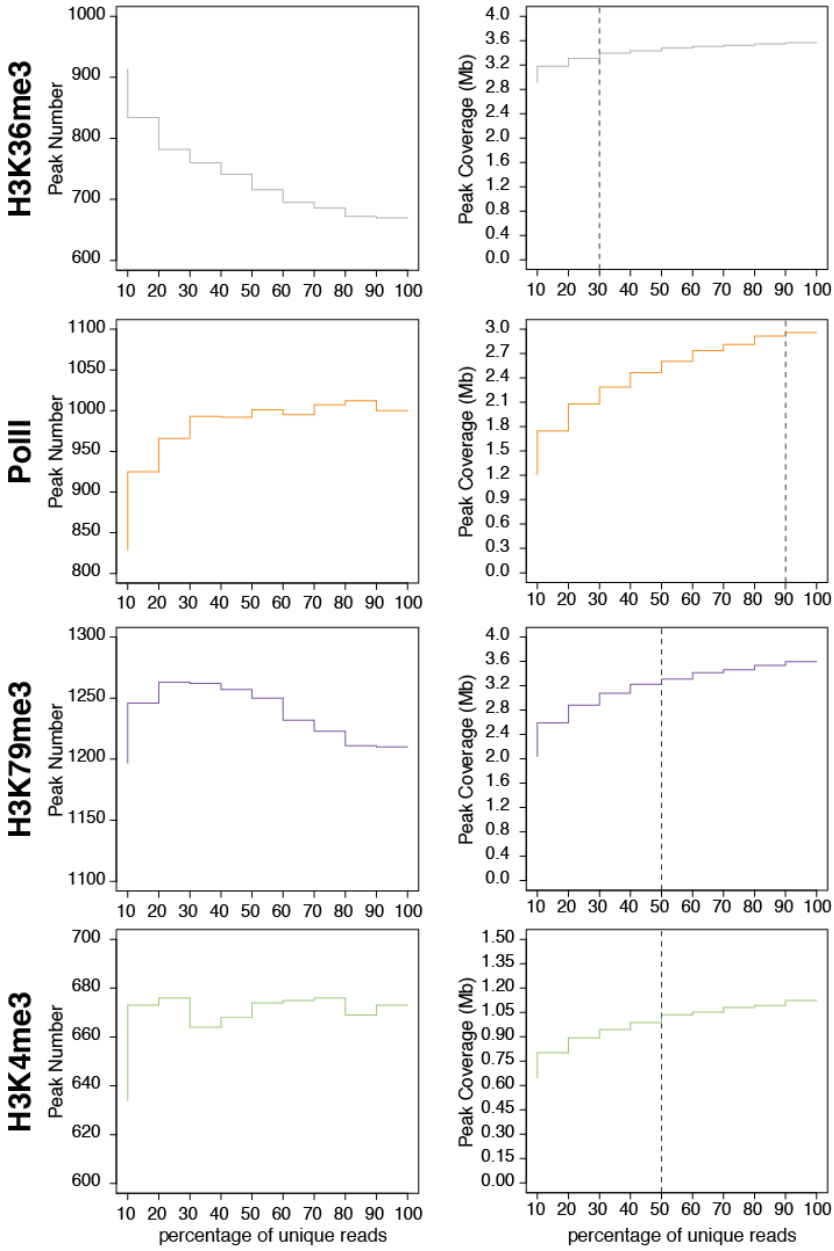


Supplementary Fig. 8: Enhancers positive for K27ac, K79me3 or Pol II are usually co-marked by K4me1

Enhancers marked by each chromatin modification or Pol II were subdivided into those that also contain H3K4me1 (+) and those that do not (-). This analysis was performed for both characterized developmental enhancers from CAD2 (C = 144 enhancers) and for TF-Meso-CRMs (T = 844 enhancers), defined by TF occupancy of mesodermal factors. The presence of H3K27ac, H3K79me3 and Pol II on enhancers is tightly associated with the presence of H3K4me1 (especially in the case of CAD2 enhancers (C)). This is in contrast to the repressive mark, H3K27me3, where many enhancers contain this mark in the absence of H3K4me1. A significant fraction of enhancers contain H3K4me1 only (green bars representing enhancers exclusively carrying K4me1 and no other mark or Pol II). It is interesting to note the high correlation between TF binding to enhancers at 6-8 hrs (T) and the presence of H3K27ac and Pol II at 6-8 hrs, suggesting that TF occupancy is the trigger leading to enhancer activation.



Supplementary Fig. 9 (part 1): Saturation of Histone H3 marks and Pol II data

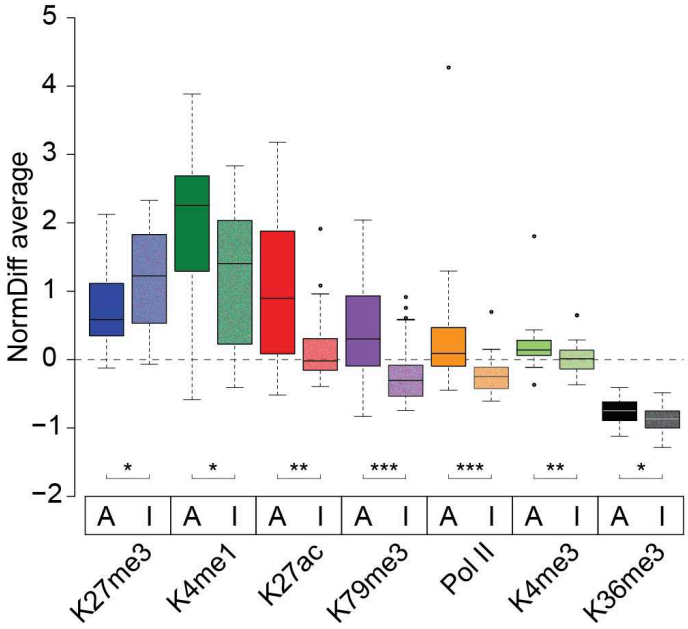


Supplementary Fig. 9 (part 2): Saturation of Histone H3 marks and Pol II data

Supplementary Fig. 9 (legend): Saturation of Histone H3 marks and Pol II data

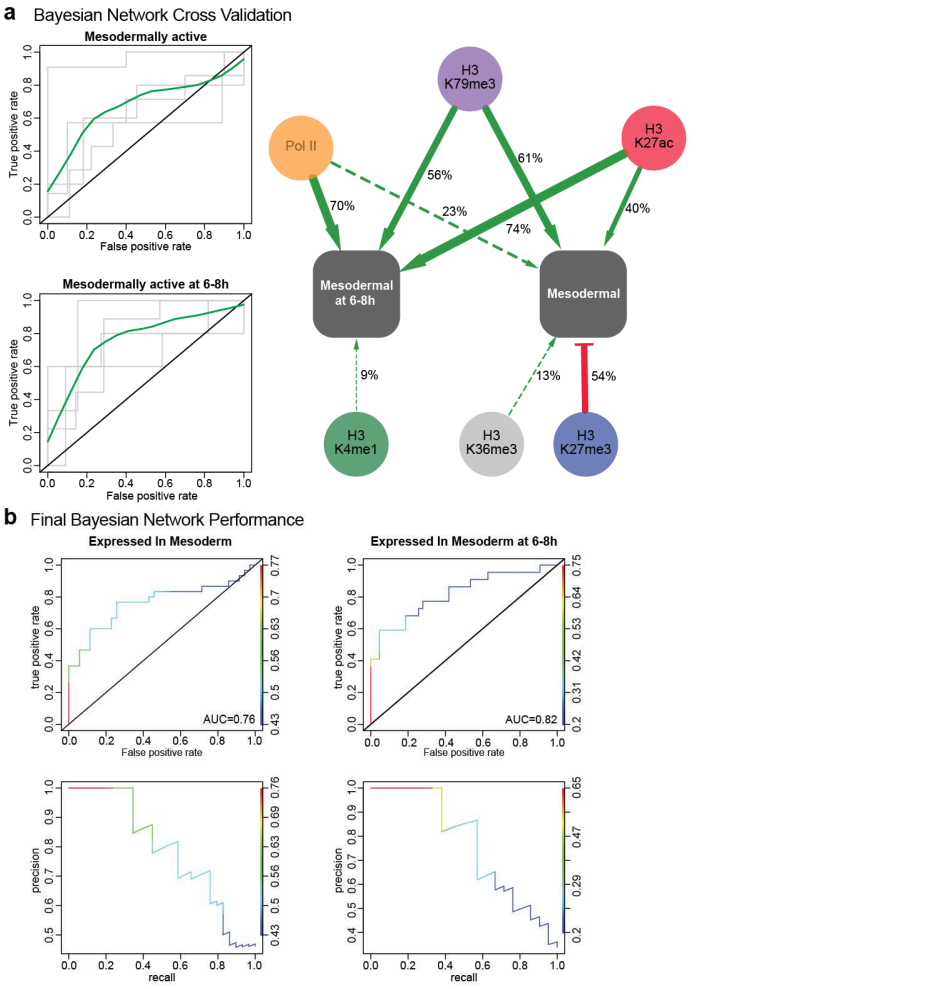
Saturation analysis of the peaks called for H3K27ac, H3K4me1, H3K27me3 (Figure part 1) and H3K36me3, Pol II, H3K79me3, H3K4me3 (Figure part 2) by sub-sampling the data. Reads of biological replicates were merged prior to analysis. Peaks were called by MACS using between 10% to 100% (in steps of 10) of reads mapping to chromosome 2L (duplicate reads were ignored when sub-sampling reads), using H3 as background for all histone marks and input for Pol II.

For each H3 modification and Pol II, the number of called peaks (**left plot**) and the total length covered by called peaks (**right plot**) are presented for each sub-sampled read fraction. While the peak number is a good indicator of saturation in cases of localized signal (H3K4me3 or TFs), the total peak coverage may more accurately assess saturation in cases where marks extend over large genomic regions. Indeed, in undersampling conditions, a unique large region may be split in multiple 'artificial' peaks being called. Hence, when more reads are used, these artificial peaks are merged. This "region splitting" effect is well illustrated on the H3K36me3 plots were the number of called peaks decreases as the proportion of used reads increases while the coverage increases and stabilizes when more than 30% of reads are used. We considered saturation reached when adding 10% more reads constantly resulted in less than 5% of coverage gain (indicated by vertical dashed black lines). All marks reached saturation when using 100% of the reads. Note the scale on the y-axis is different for each mark



Supplemental Fig. 10: Quantitative signal on mesodermally active and inactive enhancers

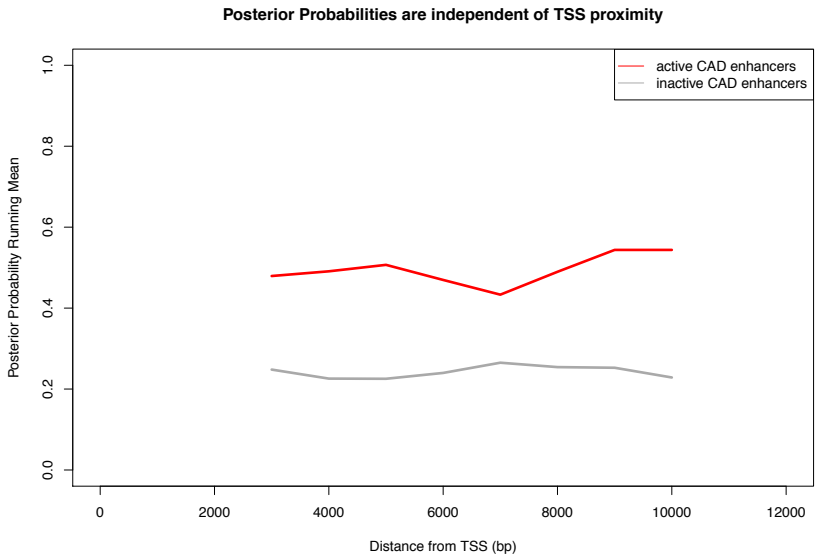
Comparison of the quantitative signal levels for histone H3 modifications and Pol II occupancy between enhancers that are active (A) or inactive (I) in the mesoderm at 6-8 hrs. Active enhancers have clearly significantly higher levels of Pol II, H3K79me3 and H3K27ac. Interestingly, the level of H3K4me3 is also significantly higher at active enhancers while remaining at values considered as noise by peak callers. Also note the clear depletion of H3K36me3 on enhancers.
Two-sided Wilcoxon rank sum test p-values: * p <= 0.05; ** p <= 0.01; *** p <= 0.001.



Supplementary Fig. 11: Bayesian classifiers cross-validation and performance

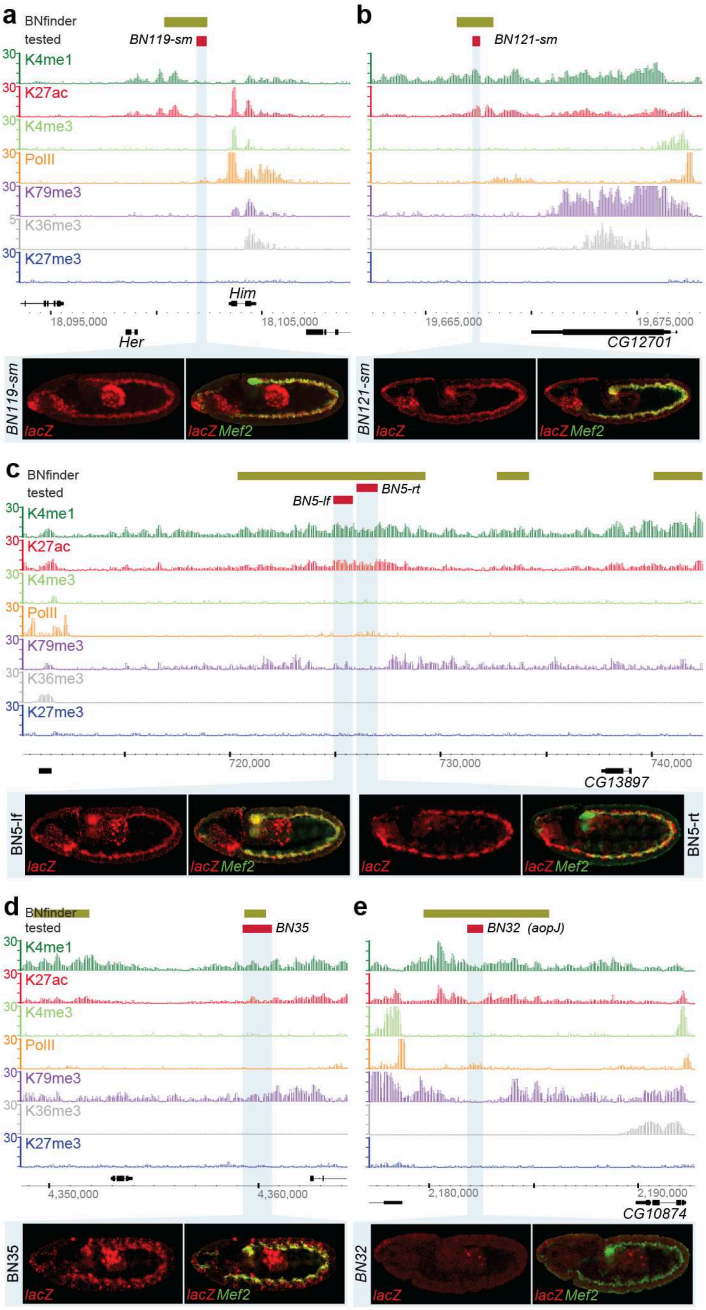
(a) *Cross-validation of the trained Bayesian Network using a 4-fold cross-validation scheme*: **Left**, TPR/FPR ROC curves for each activity state presenting the average performance (green line) and the 4 individual ROC curves (light grey lines) resulting from the 4-fold cross-validation. Both activity states can be accurately predicted using models trained with 75% of the training set. **Right**, network presenting all conditional dependencies reported when performing the 4-fold cross-validation 250 times (edges reported in less than 5% of the 1000 trained networks were omitted for clarity). Positive and negative conditional dependencies are indicated by the green and red arrows, respectively. Solid line arrows indicate conditional dependencies reported when using 100% of the training set, while dashed arrows were not reported when using 100% of the training set. Edge labels indicate the percentage of networks in which the dependency was reported. Edge widths are proportional to edge labels and represent an estimate of the robustness each learned dependency.

(b) Final performance of the learned Bayesian Network for each activity state. Both TPR/FPR and Precision/Recall ROC curves are presented. As expected, predicting activity state at the stage of data collection is more accurate (AUC=0.82).

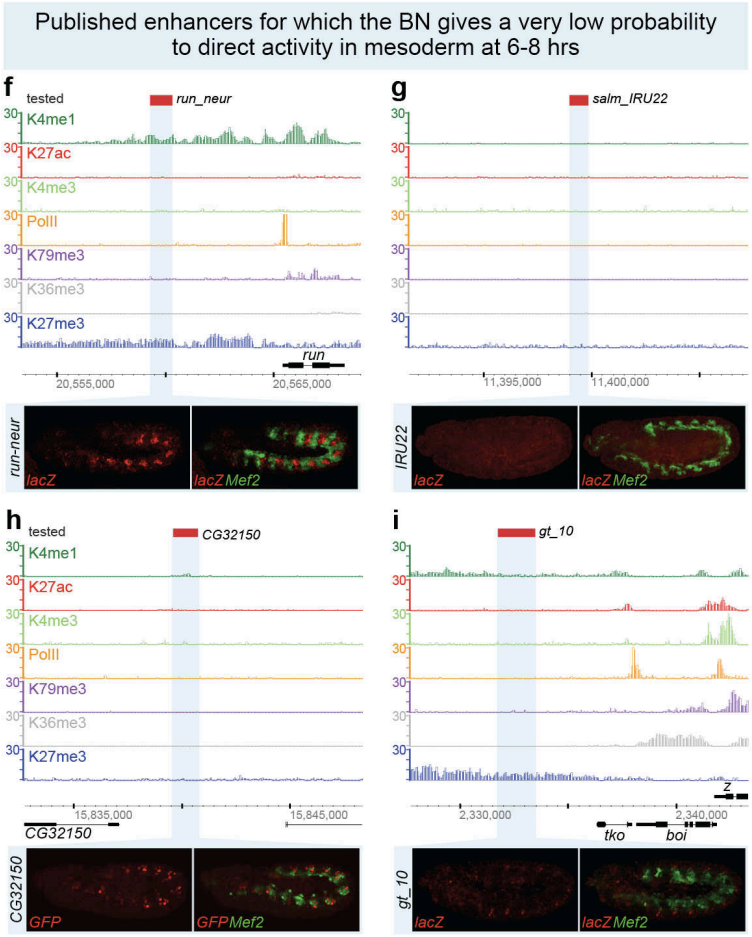


Supplementary Fig. 12: Bayesian classifier uniformly scores proximal and distal enhancers active in the mesoderm at 6-8h

The posterior probability of an enhancer to be active in the mesoderm at 6-8 hrs was computed for each CAD2 enhancer of the training set using the model trained to predict 'mesoderm enhancers active at 6-8 hrs'. The plot shows the mean posterior probabilities as a function of the distance between CAD2 enhancers (using enhancers' centers) and the transcriptional start site (TSS) of their target genes (or of the closest TSS when the target gene was unknown). The mean posterior probabilities for active (red line) and inactive enhancers (grey line) were computed using a running window of 4000bp by steps of 1000 bp (only windows with at least 5 entries were considered). No bias due to TSS proximity was observed, confirming that the trained model can be applied to the whole intergenic genome to predict enhancers active in the mesoderm at 6-8 hrs.



Supplementary Fig. 13 (part 1): Putative enhancers exhibit predicted regulatory activity in vivo.



Supplementary Fig. 13 (part 2): Putative enhancers exhibit predicted regulatory activity in vivo.

Supplementary Fig. 13 (legend): Putative enhancers exhibit predicted regulatory activity *in vivo*.

Upon applying the trained Bayesian network to the intergenic genome, predicted regulatory regions (green bars, 'BNFinder') were cloned and assayed for activity in transgenic reporter assays. **(a-i)** Top: BiTS-ChIP-seq signal enrichment for histone modifications (RPGC, background subtracted using H3) and Pol II (RPGC, background subtracted using input) are shown. Green boxes indicate the boundaries of the predicted regions by the Bayesian network (BN), red boxes indicate the boundaries of cloned regions tested for enhancer activity *in vivo*. Bottom: Embryos showing the expression pattern of the *lacZ* reporter gene driven by the genomic region (red), detected by double *in situ* hybridization with a mesoderm specific marker (*Mef2*, green). All embryos shown are lateral views, st.10-11, anterior left, dorsal up.

(**a-e)** The tested regions have posterior probabilities above the cut-off ($PP_{cut\text{-}off}$ = 0.582, corresponding to an estimated 100% specificity) and are expected to direct expression in the mesoderm at 6-8 hrs (st.10-11) according to the trained Bayesian network. 8 of the 9 regions tested function as mesodermal enhancers *in vivo* at the predicted stages of development (the data for four regions is shown in Figure 6). Note the chromatin signatures within the cloned regions (red bars, 'tested') vary substantially (also see Fig. 6), with BN119 **(a)** having very low levels of all marks, BN35 **(d)** having H3 K4me1, K27ac and K79me3 with no Pol II, while BN5,rt **(c)** has K4me1, K27ac and Pol II will very little K79me3, for example. As the predicted region shown in **(c)** is very large (~9kb), two smaller regions were cloned, which interestingly both give very similar patterns of activity despite having different signatures. The ninth region that did not work is BN32 **(e)**, which encompasses a previously published enhancer 'aopJ'.

(f-i) Tested regions that have a very low posterior probability scores (from the BN) to be active in mesoderm at 6-8 hrs. Here we examined published enhancers for which no expression information was known during embryogenesis ('ni' enhancers). Examining the activity of these predicted 'negative' regions demonstrated that all 4 regions are inactive in the mesoderm at these stages of development.

Information on the activity, location and posterior probabilities (PP) of all regions tested is provided in Supple Table 10.

Supplementary Tables

Supplementary Table 1

CAD2 entries in a tabular format. For each CAD2 entry the following information is provided: the source, name, location (dm3) as well as a staged activity profile - 2 times 11 columns indicating CRM activity by stage (stages early blastoderm st5 or earlier, then stage 6, 7, 8, 9, 10, 11, 12, 13, 14, and finally stage 15 and later) either in the mesoderm and its derivatives (1st 11 columns, 'M') or in non-mesodermal tissues (2nd 11 columns, 'O'). Annotations distinguish between expression in the mesoderm (column headers M5-15) and/or expression in non-mesodermal tissues (column headers O5-15). We manually grouped the annotated activity of 'mesoderm' enhancers into specific domains where applicable (e.g. somatic mesoderm, cardiac mesoderm, visceral mesoderm, dorsal mesoderm and mesoderm flagged with 'S', 'C', 'V', 'dM' and 'M' respectively in M5-15 columns), while all enhancers active outside of mesoderm were annotated as active in 'non-mesodermal tissues' (flagged with a '1' in O5-15 columns). Enhancers described as inactive are flagged with a '0' while a 'ni' flag is used when activity information was not reported at that stage in the original publication. Additional columns indicating target gene IDs and database references where available (mostly transferred from CAD). Supplementary Table 1 is available separately online.

Supplementary Table 2

CAD2 entry list filtered for genes (extended by 1 kb) and potential unannotated TSSs using histone H3K4me3 enriched regions (as reported by MACS). In addition to the names, locations, staged activity profiles (as in Supplementary Table 1) and enhancers' target genes, the table reports the results of overlapping enhancers with regions enriched for H3K4me1, H3K27ac, H3K27me3, H3K4me3, H3K79me3, H3K36me3 and Pol II (as computed by MACS) as well as with TF-Meso-CRMs and binding locations of non-mesodermal TFs. For all these columns: '1' = Overlap, '0' = No Overlap. Supplementary Table 2 is available separately online.

Replicate	Read Number (after quality filtering)	Mapped Read Number	Mapped Read %age	Replicates Correlation (Pearson)	Merged Read Number	Peak Number (MACS)	Saturation Read Number
input R1	28,132,732	27,358,609	97.3	0.73	42,508,317	-	-
input R2	20,779,649	15,149,708	72.9				
H3 R1	22,619,717	16,567,205	73.2	0.87	33,775,626	-	-
H3 R2	23,462,340	17,208,421	73.3				
K27ac R1	25,880,232	21,736,276	84.0	0.96	39,499,392	5154	19,749,696
K27ac R2	18,272,772	17,763,116	97.2				
K4me1 R1	25,793,011	22,708,026	88.0	0.97	43,765,575	7491	26,259,345
K4me1 R2	24,435,944	21,057,549	86.2				
K4me3 R1	24,722,284	18,575,676	75.1	0.88	46,967,572	3847	23,483,786
K4me3 R2	41,247,126	28,391,896	68.8				
K36me3 R1	22,316,923	21,615,422	96.9	0.97	46,531,368	3933	13,959,410
K36me3 R2	25,379,734	24,915,946	98.2				
K79me3 R1	18,359,977	16,642,062	90.6	0.87	46,723,877	6810	23,361,939
K79me3 R2	36,475,879	30,081,815	82.5				
K27me3 R1	25,585,231	19,065,868	74.5	0.67	44,290,975	2047	39,861,878
K27me3 R2	37,004,836	25,225,107	68.2				
PolII R1	27,125,196	20,540,645	75.7	0.85	47,970,735	5328	43,173,662
PolII R2	37,544,693	27,430,090	73.1				
Mef2	26,109,729	19,493,153	74.7	-	10,748,678	3832	-

Supplementary Table 3

This table contains a summary of the **BiTS-ChIP** sequencing results and correlation coefficient between replicates. The 'Saturation read number' is an estimate of the number of (mapped) reads required to reach genome-wide saturation for a given mark using the percentage estimated by sub-sampling simulation (see Supplementary Fig. 9). Comparing this number with the 'Merged Read Number' indicates that all data has reached saturation.

Supplementary Table 4
List of 112 enhancer regions active at 6-8 hrs in the mesoderm as predicted using BNFinder.
Supplementary Table 4 is available separately online.

ID	PP	Location (dm3)			Mesodermal activity (by stage)											Non-mesodermal activity (by stage)										
		chr	start	stop	5	6	7	8	9	10	11	12	13	14	15	5	6	7	8	9	10	11	12	13	14	15
BN predictions with high posterior probabilities (predicted to give mesodermal activity at 6-8hrs)																										
BN5-lf	0.879511562	3L	725026	725949	0	0	0	1	1	1	0	0	0	1	1	0	0	0	0	0	1	1	1	1	1	1
BN5-rt	0.879511562	3L	725929	726951	0	0	0	1	1	1	1	0	0	0	0	0	1	1	0	0	0	1	1	1	1	1
BN32-aopJ	0.877458162	2L	2181840	2182626	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
BN31	0.875260114	2L	1461882	1462726	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1
BN2	0.727943979	3L	612968	615080	0	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
BN119-sm	0.695697135	X	18101918	18102417	0	0	1	1	1	1	1	0	0	0	0	0	1	0	0	0	1	1	1	1	1	1
BN121-sm	0.695367204	X	19667239	19667608	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
BN58	0.695212108	3R	11411695	11413572	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1
BN58-sm	0.695212108	3R	11412688	11413572	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	1
BN106	0.667625779	X	3244689	3247133	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
BN106-sm	0.667625779	X	3245526	3246218	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
BN35	0.583002427	2L	4359310	4360713	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1
BN predictions with low posterior probabilities (not predicted to give mesodermal activity at 6-8hrs)																										
run-neur	0.354533219	X	20559337	20560376	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	1	1	1	1
gt-10	0.199205381	X	2331788	2333533	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0/1	0/1	0/1	0/1	0/1	0/1
CG32150-PE	0.194860392	3L	15839627	15840789	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
salm-IRU22	0.194720601	2L	11399024	11399923	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Supplementary Table 5

BNFinder predictions tested *in vivo* by transgenic reporter assays. The following information is provided: ID, posterior probability (PP), location (chromosome:start-stop in dm3), as well as a staged activity profile – 2 times 11 columns indicating CRM activity by embryonic stage (stages: early blastoderm st5 or earlier, then stage 6, 7, 8, 9, 10, 11, 12, 13, 14, and finally stage 15 and later) either in the mesoderm and its derivatives (1st 11 columns) or in non-mesodermal tissues (2nd 11 columns). For the activity columns, a '1' indicates activity, whereas a '0' indicates "no activity". Note that for enhancer 'gt-10', activity, albeit absent in the mesoderm, was low and variable in the ectoderm between stages 10-15 (denoted by '0/1').

Primer ID	Purpose	Entity	Sequence (5'→3')
FLO#1	real-time PCR	*actin* CRM	CTCGCTCGTTCGCCTTATG
FLO#2	real-time PCR	*actin* CRM	TGCCATCTCGTCCAAGAAGC
FLO#933	real-time PCR	*oskar* 3' region	CACCGTCAAGCAGCGTGTAC
FLO#934	real-time PCR	*oskar* 3' region	TGCGAATGGTCTTCATGGAA
FLO#2110	real-time PCR	*Rpl32* promoter	TTCACGATCTTGGGCCTGTATG
FLO#2111	real-time PCR	*Rpl32* promoter	TTGTTGTGTCCTTCCAGCTTCA
FLO#2112	real-time PCR	*Rpl32* 5' region	GGCACGGCGCCAAAATTAATCA
FLO#2113-1	real-time PCR	*Rpl32* 5' region	CCGATGCCACTGCCTCTTTGGT
FLO#2216	real-time PCR	*twi* promoter	GAGCAGCCGCAAAATGTCAATT
FLO#2217	real-time PCR	*twi* promoter	CGCACTTACGGAACGCAACTGA
FLO#2248	real-time PCR	*tup* promoter	GCGTGCGGATAACGTACGGCAA
FLO#2249	real-time PCR	*tup* promoter	TGCACGGAGACTGCTGAACGAC
FLO#2568	ISH probe	*vnd* genic region	GTACCCAGCAGCCCGCCAGTCCGG
FLO#2569	ISH probe	*vnd* genic region	GGTCTTCTGGCTCATCGCTCCACGGGC
FLO#2570	ISH probe	*tll* genic region	GTCGCATTCTATACCATGTGCCCTGC
FLO#2571	ISH probe	*tll* genic region	GATCTTGCGCTGACTGTACATGTCGG
FLO#2572	ISH probe	*Dfd* genic region	GGACGGTTCGTGTTCGCAGCGCACTCG
FLO#2573	ISH probe	*Dfd* genic region	ATTCCTGCAGTCCCAGCGCCGTGTTCG
FLO#2779	CRM testing	BN2	NNagatctCCAAACTTTCCCATCATATCGC
FLO#2780	CRM testing	BN2	NNggtaccACTGTGCATGTGCAGTAATGG
FLO#2781	CRM testing	BN5-lf	NNggcgcgccAGTTGGGAACCCAGTATTTGTG
FLO#2782	CRM testing	BN5-lf	NNggcgcgccAACGTAGTTTCTCCGAGTGCAT
FLO#2783	CRM testing	BN5-rt	NNggcgcgccTGCACTCGGAGAAACTACGTTA
FLO#2784	CRM testing	BN5-rt	NNggcgcgccACTTGTTCCCAGGGTCTACAAA
FLO#2785	CRM testing	BN31	NNggcgcgccCTTTATTTGAGGTCGTTCCAGG
FLO#2786	CRM testing	BN31	NNagatctGGTTTATGACGTCAGAGGAAGG
FLO#2632	CRM testing	BN32 (aopJ)	NagatctAGCTCGCAGCTGAGGAAGAGAGTGC
FLO#2633	CRM testing	BN32 (aopJ)	NaagcttTCATATCGAATCTCCGTTGCACAATGC
FLO#2766	CRM testing	BN35	NagatctTTCAAGAGCTTGGCAAGGATAG
FLO#2767	CRM testing	BN35	NNggtaccTAGTTGGTTGCAGTGGCATTAC
FLO#2790	CRM testing	BN58	NNggcgcgccATGGTCAGAAAGGACAGGGATA
FLO#2791	CRM testing	BN58	NNggtaccCCAGGACACGCTACTAATCACA
FLO#2789	CRM testing	BN58-sm	NNggcgcgccATCTCTGCATCTTGATGTTGCC
FLO#2791	CRM testing	BN58-sm	NNggtaccCCAGGACACGCTACTAATCACA
FLO#2794	CRM testing	BN106	NNagatctGTCAATCTACTCGCGTTTTTCC
FLO#2795	CRM testing	BN106	NNggtaccGGTCGCAGTTTGTTATCCATTC
FLO#2792	CRM testing	BN106-sm	NNagatctCTGACAGCCAAAAACCGTAAAC
FLO#2793	CRM testing	BN106-sm	NNggtaccGGTAGCGGATCATGCAGTTAAT
FLO#1877	CRM testing	BN119-sm	AGATCTATTATGGCCCATCTTGCATC
FLO#1878	CRM testing	BN119-sm	GGTACCAACACTTTGCAGCGGCTACT
FLO#2015	CRM testing	BN121-sm	AGATCTATTCGGCCAAAAGATGGAGA
FLO#2016	CRM testing	BN121-sm	GGTACCGTGTCTTTGTTTTGTAGAAG

Supplementary Table 6

Primer pairs used in this study. Restriction enzyme sites in lower case where applicable.

Supplementary Table 7

"active 6-8h", "meso 6-8h", "only meso 6-8h", "no meso 6-8h" and "inactive 6-8h" gene lists extracted from Berkeley Drosophila Genome Project in situ database BDGP (see Supplementary Note section V 'Gene lists using the BDGP in situ hybridization database'). Supplementary Table 7 is available separately online.

Anatomical terms used to define genes expressed in the mesoderm (terms are separated by ';')	**Stage terms used for the 6-8h time window**
mesoderm; foregut visceral mesoderm ; embryonic/larval pericardial cell ; embryonic dorsal vessel ; visceral muscle of esophagus ; pericardial cell specific anlage ; somatic mesoderm ; larval visceral muscle ; embryonic heart ; adult muscle system ; hindgut visceral mesoderm ; embryonic/larval somatic muscle ; corpus cardiacum primordium ; larval muscle system ; hindgut visceral mesoderm primordium ; direct flight muscle ; somatic muscle primordium ; embryonic/larval dorsal vessel ; midgut muscle ; trunk mesoderm anlage in statu nascendi ; trunk mesoderm primordium ; mesoderm anlage ; embryonic pericardial cell ; foregut visceral mesoderm primordium ; circular visceral muscle fibers ; cardiac mesoderm primordium ; adult heart ; embryonic/larval muscle system ; larval dorsal vessel ; muscle system primordium ; circular visceral mesoderm primordium ; larval pericardial cell ; trunk mesoderm anlage ; trunk mesoderm primordium P2 ; dorsal vessel specific anlage ; adult dorsal vessel ; embryonic/larval visceral muscle ; embryonic somatic muscle ; visceral muscle primordium ; mesoderm anlage in statu nascendi ; hypodermal muscle of abdomen ; cardioblast ; adult visceral muscle ; larval somatic muscle ; mesothoracic extracoxal depressor muscle 66 ; larval heart ; head visceral muscle primordium ; adult somatic muscle ; adult muscle precursor primordium ; embryonic visceral muscle ; dorsal pharyngeal muscle primordium ; dorsal prothoracic pharyngeal muscle ; head mesoderm anlage ; head mesoderm anlage in statu nascendi ; head mesoderm primordium ; head mesoderm primordium P2 ; head mesoderm primordium P4 ; head mesoderm ; pharyngeal muscle	stage9-10 stage11-12

Supplementary Table 8.

Mapping of BDGP ontological terms.

Supplementary Table 9

List of 66 CAD2 enhancers with summarized intensities (using a 1 kb window, see Supplementary Note section IX.1 'Intensity summarization of H3 modifications and Pol II signals covering CAD2 enhancers') used for hierarchical clustering. Supplementary Table 9 is available separately online.

Supplementary Table 10

BNFinder input file with summarized intensities (200 bp or 1 kb window, see Supplementary Note section IX.1 'Intensity summarization of H3 modifications and Pol II signals covering CAD2 enhancers') and activity states for 65 CAD2 enhancers. Supplementary Table 10 is available separately online.

Supplementary References

1. Adelman, K. et al. Efficient release from promoter-proximal stall sites requires transcript cleavage factor TFIIS. *Mol Cell* **17**, 103-12 (2005).
2. Sandmann, T. et al. A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Dev Cell* **10**, 797-807 (2006).
3. Egelhofer, T.A. et al. An assessment of histone-modification antibody quality. *Nat Struct Mol Biol* (2011).
4. Sandmann, T., Jakobsen, J.S. & Furlong, E.E. ChIP-on-chip protocol for genome-wide analysis of transcription factor binding in Drosophila melanogaster embryos. *Nat Protoc* **1**, 2839-55 (2006).
5. Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**, R86 (2010).
6. Gentleman, R.C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80 (2004).
7. Morgan, M. et al. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* **25**, 2607-8 (2009).
8. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-8 (2008).
9. Cock, P.J., Fields, C.J., Goto, N., Heuer, M.L. & Rice, P.M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* **38**, 1767-71 (2010).
10. Tweedie, S. et al. FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res* **37**, D555-9 (2009).
11. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
12. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
13. Nix, D.A., Courdy, S.J. & Boucher, K.M. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics* **9**, 523 (2008).
14. Nicol, J.W., Helt, G.A., Blanchard, S.G., Jr., Raja, A. & Loraine, A.E. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics* **25**, 2730-1 (2009).
15. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).
16. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
17. Filion, G.J. et al. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* **143**, 212-24 (2010).
18. Roy, S. et al. Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. *Science* (2010).

19. Tomancak, P. et al. Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biol* **8**, R145 (2007).
20. 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).
21. Gallo, S.M. et al. REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. *Nucleic Acids Res* **39**, D118-23 (2010).
22. Kuhn, R.M. et al. The UCSC genome browser database: update 2007. *Nucleic Acids Res* **35**, D668-73 (2007).
23. Phipson, B. & Smyth, G.K. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Stat Appl Genet Mol Biol* **9**, Article39 (2010).
24. Saeed, A.I. et al. TM4 microarray software suite. *Methods Enzymol* **411**, 134-93 (2006).
25. Chickering, D.M., Heckerman, D. & Meek, C. Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research* **5**, 1287-1330 (2004).
26. Dojer, N., Gambin, A., Mizera, A., Wilczynski, B. & Tiuryn, J. Applying dynamic Bayesian networks to perturbed gene expression data. *BMC Bioinformatics* **7**, 249 (2006).
27. Wilczynski, B. & Dojer, N. BNFinder: exact and efficient method for learning Bayesian networks. *Bioinformatics* **25**, 286-7 (2009).
28. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-9 (2011).
29. Karlic, R., Chung, H.R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A* **107**, 2926-31 (2010).

Annexe 3: Supplementary Information for "A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History"

Supplementary Information

A TF collective defines cardioblast cell fate and reflects the developmental history of this cell lineage

Guillaume Junion*, Mikhail Spivakov*, Charles Girardot, Martina Braun,
E. Hilary Gustafson, Ewan Birney and Eileen E.M. Furlong

Extended Experimental Procedures

Generating Wg and Dpp conditioned medium

Although all components of the Wg and Dpp signaling pathways are present in DmD8 cells, the ligands are not. To activate these pathways we generated conditioned Schneider's medium containing secreted Wg and Dpp using a previously described protocol with minor modifications (van Leeuwen et al., 1994). 4ug of *wg* and *dpp* cDNA under an inducible promoter (pRM) were transfected into 750mL ML-DmD8 cells flasks using effectene. Following induction using 0.7μM of CuSO4 24H, the cells were incubated for 4 days to allow the secretion of Wg and Dpp ligands. Cells and debris were removed by centrifugation at 2000 g for 5 min. The conditioned Medium was concentrated 50x using Amicon Ultra 10K (Figure S5A).

ChIP analysis in DmD8 cells

ChIP followed by real time PCR of endogenous loci in DmD8 cells were performed as follows. Four 750mL flasks of cells were transfected, each with 4μg of pRM-Tin using effectene reagent. One day later, cells were induced with 0.7μM of CuSO4 and incubated for 3 days. 24hr prior to chromatin preparation, conditioned medium (Figure S5B) was added (1% final) to allow activation of the Wg and Dpp pathway. The cells were covalently cross-linked by replacing the medium with cold PBS containing 1% formaldehyde and incubating for 10min at RT. After the addition of Glycine (0.125M) for 5 min at RT, the cells were washed twice with PBS and harvested using a cell scraper in 6 ml of SDS buffer (100mM NaCl, 50mM Tris-Cl pH8.1, 5mM EDTA pH8.0, 0.2% NaN3, 0.5% SDS) containing protease inhibitors. Following centrifugation at 600G for 6min, the cells were resuspended in lysis buffer (2 volume SDS buffer + 1 volume triton dilution Buffer (100mM Tris-Cl pH8.6, 100mM NaCl, 5mM EDTA pH8.0, 5% triton X-100) following by incubation for 20 min at RT. The lysate was passaged through needles of 20G, 25G, 27G and distributed in eppendorfs (300-350 ul) for sonication. After 10 min centrifugation at 14000g the chromatin preparations were transferred to low binding tubes and stored at -80 degrees. The ChIP experiments were performed using the same conditions and antibody amounts as used for the embryo ChIP experiments.

Luciferase reporter assays

The three enhancers tested were cloned into the pGL3 luciferase vector (Promega) with an *Hsp70* minimal promoter and the luciferase activity was normalized to a Renilla standard (Promega). Tinman was expressed using a metallothionine promoter (pRM HA3b vector). 100 ng of pRM-Tin, 50ng CRM-pGL3 and 0.5ng of Renilla DNA was transiently transfected into ML-DmD8 cells using Effectene (Qiagen) following manufacturer's instructions. The total amount of transfected DNA was kept constant by supplementing empty pRM vector. Expression was induced 24h after transfection using 0.7μM of CuSO4. Cells were then incubated for 3 days and supplied with Wg+Dpp conditioned medium (1%) 24h before lysis. To remove Pnr and Doc, dsRNA was transfected with the transfection mix containing the CRM-pGL3 and Tin, using 80 ng of dsRNA per well (Figure S5A). Lysed cells were mixed with luciferase substrate and the resulting luminescence was measured using a Mithras LB 940 luminometer. Each experiment was performed as two independent biological replicates, each performed in triplicate in 96-well plates.

ChIP-on-chip data availability

All ChIP-chip hybridization data are available at ArrayExpress (http://www.ebi.ac.uk/arrayexpress) under accession numbers E-TABM-1184 and the array design under A-AFFY-53. The high-confidence transcription factor (TF) binding information, including CRM coordinates and occupancy by different TFs is provided in Tables S1, S2 and S6, and also on the Furlong lab web page at http://furlonglab.embl.de/

Data analysis

Detection of TF bound regions and peaks

All bioinformatics analyses were performed using *D. melanogaster* genome BDGP version 5 (UCSC dm3) (Celniker et al., 2002) and the Flybase 5.9 genome annotation release (Tweedie et al., 2009). Mapping of the Affymetrix GeneChip® *Drosophila* Tiling 1.0R probes to the genome was obtained from the MAT website (http://liulab.dfci.harvard.edu/MAT/). For each of the 10 conditions (5 TFs x 2 time points), TileMap version 2 software (Ji and Wong, 2005) was used to compute TF-specific ChIP signals for each array probe ("probe-level statistic", which can be regarded as adjusted log-ratio signal) based on two biological ChIP replicates and

mock controls (as shown in the table below), and detect continuous regions of signal enrichment ("peak regions") using a Hidden Markov model. A threshold on the probe-wise maximum a posteriori probability of each region returned by TileMap was determined manually for each dataset, as shown in the table below. To limit threshold effect, the top 5% regions below threshold exhibiting more than 75% overlap with one or more above cut-off region (at any of the 10 conditions) were rescued and included in the final 'high-confidence' TF binding profile. For each enriched region, peak position(s) and height were estimated as extrema on a smoothed curve of the log2-ratio signal (Schwartz et al., 2006). The table below shows the datasets and cut-offs used to run the TileMap algorithm and select regions for CRM definition. Cutoffs apply to log transformed TileMap 'max_score' ($-log_{10}(1 - \text{max\_score})$). The table also presents the total number of TileMap regions considered at each condition (the number of sub-cutoff regions rescued as described above is shown in brackets):

TF	Time	IP samples	Mock samples	Cut-off	Number of regions (rescued)
Doc2	4-6h	2xIP at 4-6h	2x4-6h + 2x6-8h	5.5	1696 (279)
Doc2	6-8h	2xIP at 6-8h	2x4-6h + 2x6-8h	5.5	1567 (299)
Tin	4-6h	2xIP at 4-6h	2x4-6h + 2x6-8h	5.0	3252 (124)
Tin	6-8h	2xIP at 6-8h	2x4-6h + 2x6-8h	5.5	1136 (105)
Pnr	4-6h	2xIP at 4-6h	2x4-6h + 2x6-8h	5.0	4653 (78)
Pnr	6-8h	2xIP at 6-8h	2x4-6h + 2x6-8h	5.0	4824 (55)
dTCF	4-6h	2xIP at 4-6h	2x4-6h + 1x6-8h	5.7	779 (259)

Defining CRMs and computing CRM-level quantitative signals

ChIP-peaks across all conditions were merged using a neighbor joining approach with a maximum distance of 200bp between adjacent peaks. Peak cluster boundaries were extended by 100bp past the terminal peak position on each side to account for inaccuracy in peak position precision. The resulting database contained a total of 11287 non-redundant CRMs. Quantitative ChIP signals for each TF per CRM, at each condition, were computed as maximum moving average probe-level statistic per CRM (window size 200bp). Note that the Affymetrix GeneChip® *Drosophila* Tiling 1.0R array is tiled with one probe every 40bp on average.

Defining Tinman-bound and TF-negative CRMs

Tinman-positive CRMs were defined as having high confidence Tinman-specific binding peaks at either time point. TF-negative CRMs (including 1209 'Tinman-negative' CRMs) were defined as having background CRM signal levels for a given TF at both time points. As an estimate of the background signal, we analyzed ChIP signals at 11607 500-bp promoter regions that did not contain strong peaks for any analyzed TF. Based on the signal distribution at these promoters, we chose 0.5 as the maximum CRM-level ChIP signal threshold for TF-negative CRMs, corresponding to the bottom 40-60% of the background distribution (depending on TF and time point).

Unbiased classification of CRMs based on their TF binding signatures

Tinman-positive and negative CRMs (defined as above) were clustered separately. CRM clustering was based on exponentiated scaled CRM-level TF signals for each TF and time point. Tin-4.6 hr signal was not considered for the clustering of Tinman-positive CRMs, as this was used to select these 'Tinman-bound' CRMs at the outset. Similarly, Tin signal at either time point was not considered for the clustering of Tinman-negative CRMs. Autoclass Bayesian clustering was used for classification (Cheesman, 1996; http://ti.arc.nasa.gov/tech/rse/synthesis-projects-applications/autoclass/autoclass-c/ and references thereof). In this method, each class is defined by a multivariate distribution with the number of dimensions equal to the number of attributes - TF/times in our case. Each observation (in our case, CRM) is assigned a probability of belonging to each class. The algorithm then automatically optimizes the class properties and the number of classes so that the observations are best separated and overall have the highest probabilities of class membership. This

"fuzzy" approach provides an easy means of assessment of how well each observation fits the classification and how separable the classes are – properties that are not explicitly available in conventional approaches such as k-means clustering. In this study, we used the original Autoclass-C command line software (Cheeseman, 1996), but a web interface, Autoclass@IJM has recently been developed by others for use with bioinformatics data (Achcar et al., 2009).

Autoclass C software was used with the following principal settings: (1) assumed data model: single_normal_cn (all attributes conform to conditionally independent normal variables); (2) convergence criterion: converge_3 (most stringent of all available); (3) use the following numbers of classes as "starting points" for classification: 5, 8, 10, 15 (note that the final classification, as expected, has departed from these points); (4) assumed relative error of input data: 10% for each attribute.

A classification into 24 clusters was obtained for Tin-positive CRMs using these settings (one cluster included only 3 CRMs characterized by very high Tin_6.8 signals and low signals for other TF/times, and was omitted). As expected, varying the assumed standard error of the input data had an effect on the number of classes and, correspondingly, their tightness. Other parameters did not significantly affect the classification (data not shown). Each CRM was then filtered based on the goodness-of-fit to the classification and the specificity of the cluster assignment using the following criteria: p_best > 0.5, p_best / p_second_best > 2, where p_best and p_second_best are the probabilities of belonging to the best-fitting and second-best-fitting class, respectively. ~75% of Tinman-positive CRMs (3099) passed this filter and were used in further analyses.

Verifying the robustness of CRM classification

To verify the robustness of the classification with respect to potential outliers, we drew one hundred samples from the table of CRM-level TF binding signals, each of which contained 80% of the total number of CRMs. An Autoclass classification was generated for each of the 100 tables. Differences between all pairs of these classifications (100*99/2 = 4950 in total) were assessed using two independent clustering stability metrics: Variation of Information (VI) (Meila, 2005) and Split-Join (SJ) (van Dongen, 2000). The means and standard deviations of these metrics across the 4950 pairs are shown in the table below. Both metrics range from zero (full

consistency) to one (no consistency), and their low values suggest that our classification is generally robust.

Since no clustering stability metric can currently be interpreted definitively in absolute terms (Meila, 2005), we compared the obtained values to two controls. The first control meant to represent a situation in which no consistency between classifications was expected. For this purpose, 100 datasets were generated in which the TF binding signals in each column were randomly reshuffled between CRMs. From each of these tables, the same subset of 80% CRMs was drawn (note that each of them had different TF signals as a result of reshuffling) and classified by Autoclass. VI and SJ were then computed for each pair of classifications. As expected, VI and SJ in this setting were high (see table below), suggesting a much less robust classification than seen with real data. As a second control, we employed one of the most widely used partitioning algorithms, k-means clustering, to classify the same data as Autoclass. K-means classifications (k=25) were generated with the same 100 x 80% samples from the (non-reshuffled) CRM binding table that were used for Autoclass analysis. As can be seen from the below table, the robustness of k-means clustering was significantly lower compared to Autoclass (Wilcox test $p<1e-256$).

	VI (Meila, 2005)	SJ (van Dongen, 2000)
Autoclass	0.12 ± 0.09	0.15 ± 0.09
K-means	0.28 ± 0.08	0.43 ± 0.10
Autoclass: reshuffled	0.69 ± 0.02	0.83 ± 0.02

Grouping and selection of CRM classes for further analyses

Visual inspection of Tinman-positive CRM clusters generated by Autoclass suggested that they could be broadly divided into three large groups based on the signals of the "other" four TFs (Doc, dTCF, pMad and Pnr): 'All TFs', distinguished by generally correlated signals for the four TFs (Doc, dTCF, pMad and Pnr); 'Two TFs', characterized by a clear skew towards one of the four TFs (in addition to Tin); 'Tin only', in which all four TFs showed negligible levels. Numerically, skew was

expressed as the difference between the signal for the strongest-bound TF and the mean signal for the remaining three TFs (other than Tin). Class-wise mean signal across the four TFs was used as a measure of general TF enrichment. (Note: (1) signals from both time points for each TF were combined using mean; (2) we use the notation $skew_4$ and $mean_4$ to emphasize that these parameters do not include Tin). In terms of these two parameters, the 'All TF' group is defined by a low skew and a non-negligible mean. In contrast, the defining property of the 'two-TF' groups is a high skew for the respective TFs. Finally, classes in the 'Tin only' group have low values of both skew and mean. We therefore populated the groups using the following empirical criteria: 'All TF': { $skew_4 \leq 1.2$; $mean_4 > 1$ }; 'two-TF': $skew_4 > 1.5$; 'Tin only': { $skew_4 < 0.6$; $mean_4 < 0.2$ }. To minimize heterogeneity, for further analyses we focused on three 'All TF' classes with the highest mean signals ($mean_4 > 2.7$), while in the larger 'two-TF' groups (Pnr+Tin and dTCF+Tin) we focused on two classes per group that showed the highest skew ($skew_4 > 2.8$) (see Figure S2A).

***De novo* PWM discovery and PWM sources**

De novo position weight matrix (PWM) discovery was performed using Weeder (version 1.3) (Pavesi et al., 2004) on 400bp regions centered on the peak signal positions (as defined earlier) derived from the top 100 ChIP regions (as ranked by TileMap). Weeder was run on each of the 8 conditions (4 TFs excluding Tin, 2 developmental times each) with the following options: -R 50 -O DM -W 6 -e 1 -M -S -T 10. Results and PWMs selected for further analysis are presented in Figure S4A. "Known" PWMs presented in Figure S4A were obtained from published resources: The Tin PWM is from Zinzen et al., 2009, the two Pnr motifs shown are for the *Drosophila* (Haenlin et al., 1997) and vertebrate (MA0035.2 from Jaspar) proteins, respectively. The dTCF and pMad PWMs were obtained from FlyReg (Bergman et al., 2005) and the mouse Tbx6 matrix is from (White and Chapman, 2005); note there is no known motif for Doc and only limited amino acid identify (~59%) in the DNA binding domain to Tbx6.

Motif scanning

Unless otherwise specified, the mapping of TF binding sites (TFBS) at CRMs was performed using Patser (Hertz and Stormo, 1999) with the following thresholds: Tin – 6, dTCF – 7, pMad – 4.5, Pnr – 5.5, Pnr* – 6.7, Doc2 – 5.5. These thresholds were defined on the basis of selectivity – specificity data to ensure a recall of TF-bound

CRMs of 50-70% (in either the 'All TF' or 'two-TF' class, whichever is higher) and the enrichment over TF-negative CRMs of 1.5-2 (data not shown).

Motif analysis on the 'All TF' and 'Two-TF' CRM classes

Differential k-mer composition between 'All TF' CRMs and 'two-TF' CRMs (Figure S4B) was performed using RSAT oligo-diff tool (Thomas-Chollier et al., 2008) with following options: -nopurge -l 6 -2str -noov -lth occ 3 -lth occ_sig 2.

PWM *de novo* discovery on 'All TF' CRMs was performed using RSAT oligo-analysis (Thomas-Chollier et al., 2008). Enriched k-mers of length 6, 7 and 8 were discovered using RSAT oligo-analysis (options -2str -noov -return occ,freq,proba,rank,mseq -sort -lth occ_sig 2) against a background assembled using all 11287 CRMs defined in this study. Enriched k-mers were assembled into patterns (pattern-assembly tool using -2str -sc 9 -subst 0 (or –subst 1) options) and converted to PWMs using the RSAT matrix-from-patterns tool (default parameters).

CRM grammar analysis

For pairwise TF analysis, the expected distances between motifs and/or peaks were computed assuming a uniform distribution of the observed number of motifs/peaks across the length of CRMs. Motif distance analysis results reported in Figure S4D was performed using motif thresholds selected based on motif specificity/sensitivity for TF-bound CRMs (data not shown); repeating this analysis over a range of thresholds did not reveal any consistent associations that would be otherwise overlooked (data not shown). For multiple associations/grammar analysis, the specialized learner software SCRM (Noto and Craven, 2006) was used as follows. First, to specifically address motif positioning rules, 10x 'All TF' CRMs with randomized motif positions, but preserved motif content, were used as the negative set. All rules identified were strongly overfitted and not considered further (for example, the most predictive rule was "a Doc or Pnr or Tin TFBS upstream of Doc or dTCF TFBS" and had the following poor receiver characteristics: #TruePos=52, #FalsePos=733, #TrueNeg=8267 and #FalseNeg=193). Second, we asked whether 'All TF' CRMs could be distinguished from other putative regulatory regions by comparing their motif distribution and content to ~1000 500-bp gene upstream regions with no detected binding of the analyzed TFs. The most predictive rule set learned in this setting was: "three motifs for Doc, pMad, Tin, in any order and any distance apart", with the following, rather poor, receiver characteristics:

#TruePos=166, #FalsePos=246, #TrueNeg=404, and #FalseNeg=74 (it is likely that Pnr motifs were not discovered in this classification due to their similarity to other GATA motifs, which are highly enriched around promoter elements).

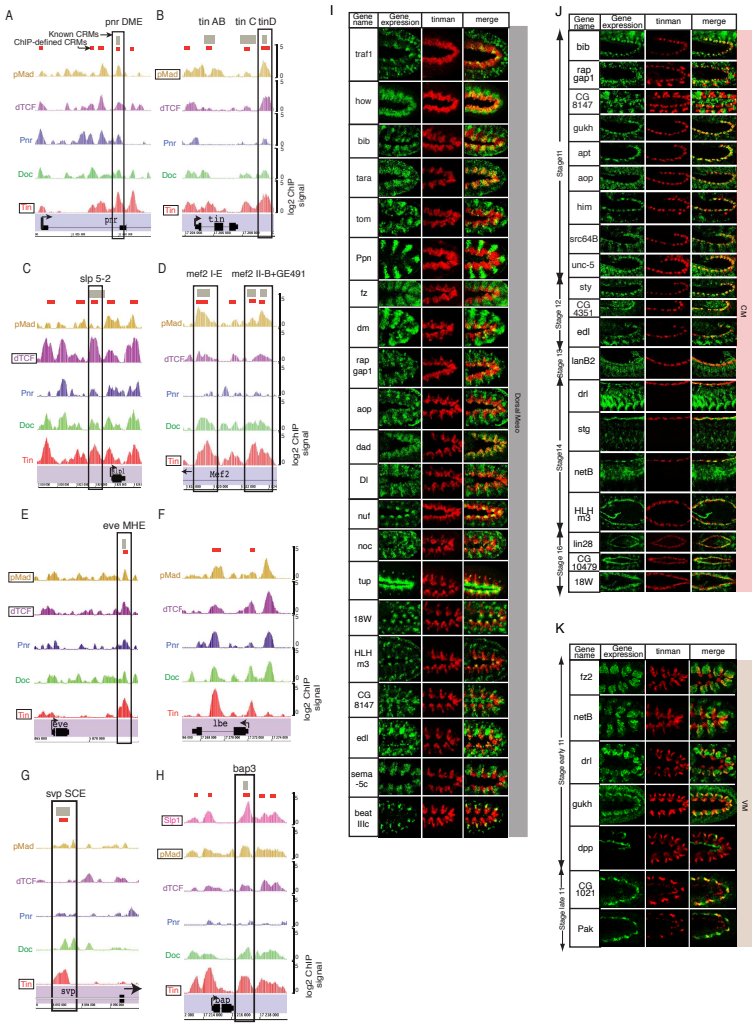


Figure S1 (related to Figure 1)**. TF occupancy on known CRMs and expression of genes proximal to TF binding events**

(A-H) TF occupancy on known dorsal mesoderm and cardiac CRMs active during stages 9-11. Grey boxes at the top of each panel indicate the genomic position of known mesodermal CRMs. Red rectangles represent our ChIP-defined CRMs. Binding signal (log2 ChIP signal for 6-10hrs (mean signal from both time-points for

visualization clarity) is presented for each TF at eight gene loci. The gene model is indicated at the bottom of each panel. The TFs known to regulate each CRM are boxed. **(A)** The *pnr* Dorsal Mesoderm Enhancer is bound by Tin as previously described by *in vitro* and site directed mutagenesis experiments (Gajewski et al., 2001) but also by the other four analyzed factors. **(B)** In the *tinman* locus, the dorsal mesoderm enhancer, *tinD*, has high levels of pMad and Tin binding, confirming mutagenesis analysis (Xu et al., 1998). Our data also revealed a high level of dTCF occupancy suggesting a requirement of the Wg signaling pathway to induce optimal levels of *tin* expression within the dorsal mesoderm, which has not been described previously. Note, the *tinC* enhancer is not bound by Doc and Pnr, suggesting that the maintenance of *tin* expression in cardioblasts occurs via indirect interaction probably via Midline/H15 as previously proposed (Reim et al., 2005) or alternatively via another enhancer. **(C)** slp5-2 is the enhancer in the *slp* locus receiving input from Wg signaling to regulate expression in mesodermal stripes at stage 10 of development. The high level of dTCF binding observed on this enhancer provides *in vivo* validation of previous gel shift assays and site directed mutagenesis showing a requirement of dTCF binding sites for enhancer activity (Lee and Frasch, 2000). The high binding of Tin may provide the competence to drive expression in mesoderm. **(D)** The cardiac enhancers within the intron of *mef2* are bound by Tin, Doc, dTCF and pMad; only Tin was previously reported be a direct activator of cardiac *mef2* expression (Gajewski et al., 1997; Cripps et al. 1999). **(E-G)** Significant binding of multiple TFs was identified in the vicinity of genes involved in the specification of subpopulations of cardioblasts and pericardial cells, which represent very small populations of cells e.g. *eve*, *lbe* and *svp*. **(E)** *eve* Muscle and Heart Enhancer (MHE) is known to be regulated by the co-binding of Tin and the effectors of Dpp and Wg pathways (Knirr and Frasch, 2001). We identified *in vivo* binding of these three TFs to this CRM, demonstrating the sensitivity of our ChIP experiments. **(F)** We identified a number of CRMs within the *ladybird early* (*lbe*) locus, where previously there were no regulatory elements known. **(G)** The *seven-up* (*svp*) SCE enhancer is bound by Tin, matching the known requirement of *tinman* function for its activity (Ryan et al., 2007). Our data revealed that this CRM is also bound by Doc, which is consistent with the coexpression of *doc* and *svp* in two cardioblasts per hemisegment at later stages of development (Zaffran et al., 2006). (**H**) ChIP signals recapitulate known binding of Tin and pMad on the *bap3* CRM (Lee and Frasch, 2005). Interestingly, additional

occupancy of Doc and a low-level occupancy of dTCF are also observed. The Slp1 ChIP experiment recapitulates the *in vitro* binding of Slp1 on the *bap3* CRM (Lee and Frasch, 2005), which acts to repress expression of the VM gene *bagpipe* within the cardiogenic mesoderm.

(I-K) Genes in the vicinity of TF binding events are expressed in the dorsal mesoderm or its derivatives. 50 genes that had binding events for at least Tin, Pnr and Doc were selected for expression analysis. Double fluorescent *in situ* hybridization was performed with an antisense RNA probe directed against the endogenous *tinman* gene (red) and the gene of interest (green). Among the 42 genes giving specific expression patterns, 38 genes are co-expressed with *tin* in the dorsal mesoderm and its derivatives, the cardiac mesoderm (CM) and visceral mesoderm (VM). Twenty-six genes are expressed in the mesoderm at stage 10-11 and are co-expressed with *tin* in the dorsal mesoderm **(I)**. Ten of these genes are also expressed at later developmental stages in cardioblasts **(J)**, as well as ten additional genes that are expressed in the CM at stage 11. Finally, 7 genes with co-expression with *tin* in VM are shown in **(K)**. Note that at least 3 genes, *drl, NetB* and *gukh,* are expressed in both the CM and VM at different stages of development (compare J and K). The expression patterns of the majority of these genes is very complex, with no gene having exclusive expression in CM, which is reflected in the large number of identified CRMs in the vicinity of these genes. 89% of the 38 genes have at least one CRM in their vicinity that is bound by All TFs, confirming our hypothesis that focusing on CRMs co-bound by multiple TFs including Tin will limit our analysis to regulatory elements that activate expression in dorsal mesoderm (and its derivatives) (see Table S4). Among the 38 genes are members of known signaling pathways; *bib, Dl, HLHm3, tom* (Notch pathway)*, fz, fz2* (Wg), *dad* (Dpp), *aop, edl* (EGF), *sty*, *src64B* (FGF); genes with known function in heart development *(tup, apt, lanB2)* and many genes with an unknown role in *Drosophila* heart development; *rapgap1, gukh, unc-5, CG8147, CG10479, 18W, lin28, src64B, netB, drl, him, stg, nuf.* All pictures are lateral view, except the last four rows on panel J, which are oriented in dorsal view.

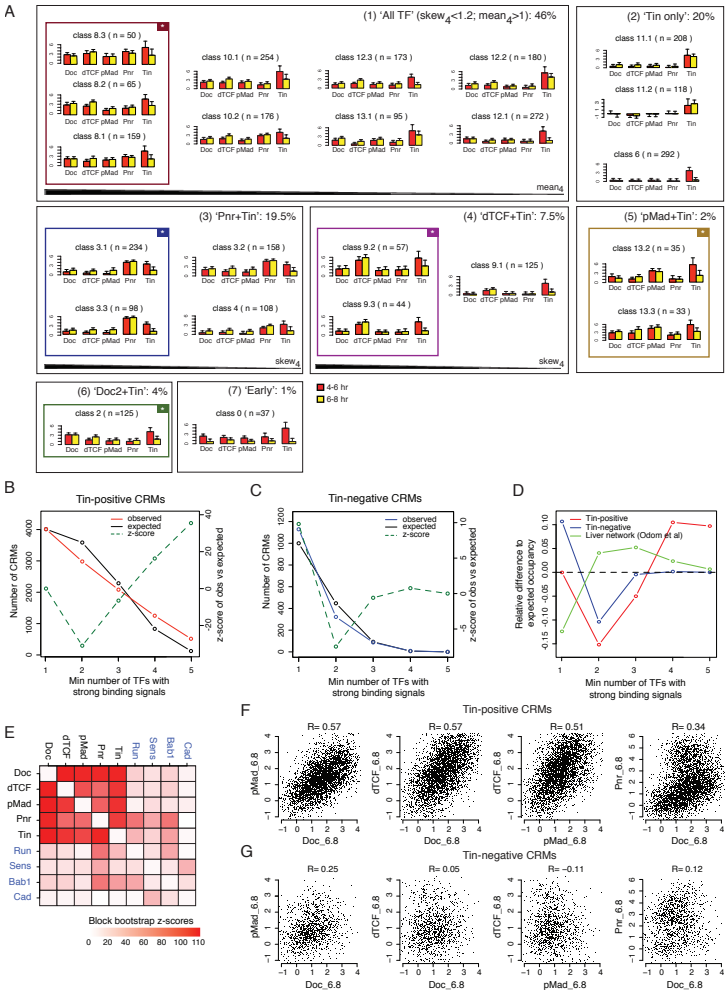


Figure S2 (related to Figure 2). **Collective occupancy of the 'heart' TFs at Tinman-bound CRMs**

(A) Autoclass classification and grouping of Tinman-bound CRMs. Barcharts show mean $\log_2$ ChIP signals in each cluster for each TF at 4-6 hr (red) and 6-8 hr (yellow) of development, while the 'bars' represent the standard deviation from the mean. The number of CRMs per class is given in brackets. Clusters have been combined into

larger classes based on two parameters: (1) the mean binding signals (log_2) of the four TFs other than Tinman (Pnr, dTCF, Doc or pMad) in each cluster, $mean_4$ and (2) the prevalence of the binding signal of either one TF over the three others, $skew_4$., with highly correlated clusters for all TFs having a low skew value. Using these criteria, the following classes were distinguished: (1) "All TFs", including 46% of CRMs, is characterized by generally high correlated signals for all four TFs ($skew_4 < 1.2$) ranging from strong to weak mean TF signal ($1 < mean_4 < 3.6$ [log_2]); (2) Tinman-only CRMs, either at 4-6 hr only or at both times ($skew_4 < 0.2$, $mean_4 < 0.6$; 20% of CRMs); (3)-(6) "two-TF" classes ($skew_4 > 1.5$), with the prevalence of Pnr+Tin (3), dTCF+Tin (4), pMad+Tin (5) and Doc+Tin (6), respectively; (7) enriched signals for all TFs at 4-6 hrs only (excluded from analysis because of the very small number of CRMs in this class). Black arrow below histograms indicates decreasing mean TF binding signal ($mean_4$) in (1) the 'All TF' class, and decreasing correlated signal ($skew_4$) for (2) and (3). Colored rectangles highlight subclasses selected for further analysis (marked with asterisks on the heatmap in Figure 2A and shown in Figure 2C, main manuscript) based on the highest $mean_4$ in the case of "All TFs" and the highest $skew_4$ values in the "two-TF" classes. Cluster numbering is arbitrary. See Extended Supplementary procedures for details on Autoclass analysis and cluster grouping and Table S3 for the list of CRMs in each cluster.

(B-D) 'Cardiogenic' TFs bind to the same enhancers at a much higher frequency than expected by chance. Combinatorial occupancy of Tin-positive CRMs (**B**, red) and Tin-negative CRMs (**C**, blue) compared to numbers expected assuming independent TF recruitment (black). Dotted lines show the z-scores of the difference between observed and expected values. **(D)** The difference between observed and expected occupancy of Tin-positive and Tin-negative CRMs (within a developing embryo) compared to that reported for the human hepatocyte regulatory network using a tissue culture system (Odom et al., 2006). For TFs other than Tin, a log_2 signals >2, at least at one time point, were considered as 'present', with the rest considered as 'absent'; note this threshold-based approach is more prone to underestimating the numbers of co-bound TFs than other approaches used in this study. Tin-positive and Tin-negative CRMs were defined as described in the Extended Experimental Procedures. Expected occupancies were computed as probabilities of compatible events given the proportion of CRMs occupied by each TF. Z-scores are for the normal approximation of binomial distribution and were preferred over binomial p-values for clarity of

visual presentation. The difference in the shape of the expected occupancy curves between Tin-positive and Tin-negative CRMs is dictated by the fact that all Tin-positive CRMs are *a priori* occupied by at least one TF (Tin), while no Tin-negative CRMs can be occupied by all five TFs. The data demonstrates that Tin-positive CRMs have a very significant skew towards higher TF occupancy, while the occupancy of Tin-negative CRMs is generally similar to what would be expected at random. Note, the level of combinatorial occupancy observed for the 'cardiogenic' TFs (red line) is greater than that of the 'liver' network (green line), even though the liver experiments were performed in primary hepatocytes and therefore represents a more homogeneous population of cells.

(E) Significance of pair-wise overlap between genomic regions bound by the 'cardiogenic' and other developmental TFs. Heatmaps showing the significance of base-pair-wise overlap compared to what would be expected at random. Significance was assessed by the block bootstrap algorithm (Bickel et al., 2010) that corrects for genomic heterogeneity. Block bootstrap z-scores are shown, ranging from zero (white; overlap as expected at random) to 110 (red; extremely highly no-random overlap). As controls for the five 'cardiogenic' TFs profiled in this study (labeled in bold), we used the binding data for four other developmental TFs at comparable developmental times (modENCODE consortium, Roy et al., 2010). Runt (*run*) is expressed in the ectoderm, in stripes overlapping Wg (dTCF) and Doc expression, while senseless (*sens*) is expressed in the dorsal ectoderm in a domain overlapping that of Pnr. Caudal (*Cad*) is expressed in an ectodermal strip at the posterior end of the embryo and later in the hindgut primorida. Bric à brac (*Bab1*) is involved in the proximo-distal patterning of legs and antennae (Couderc et al., 2002). A module consisting of the five 'cardiogenic' TFs is clearly visible in the top left quadrant of the heatmap, exhibiting very high significance of pair-wise overlap with each other. In contrast, Run, Sens or Bab1 have much lower overlap with the 'cardiogenic TFs, with the exception of Pnr which has some overlap with these 'ectodermal' TFs, although at notably much lower significance than with the 'cardiogenic' TFs. This ectodermal signature is expected given Pnr's broad expression in the dorsal ectoderm.

(F, G) Dot plots and Pearson correlation coefficients of ChIP signals for different pairs of TFs at the 6-8 hr time point at Tin-positive **(F)** and Tin-negative CRMs **(G)**.

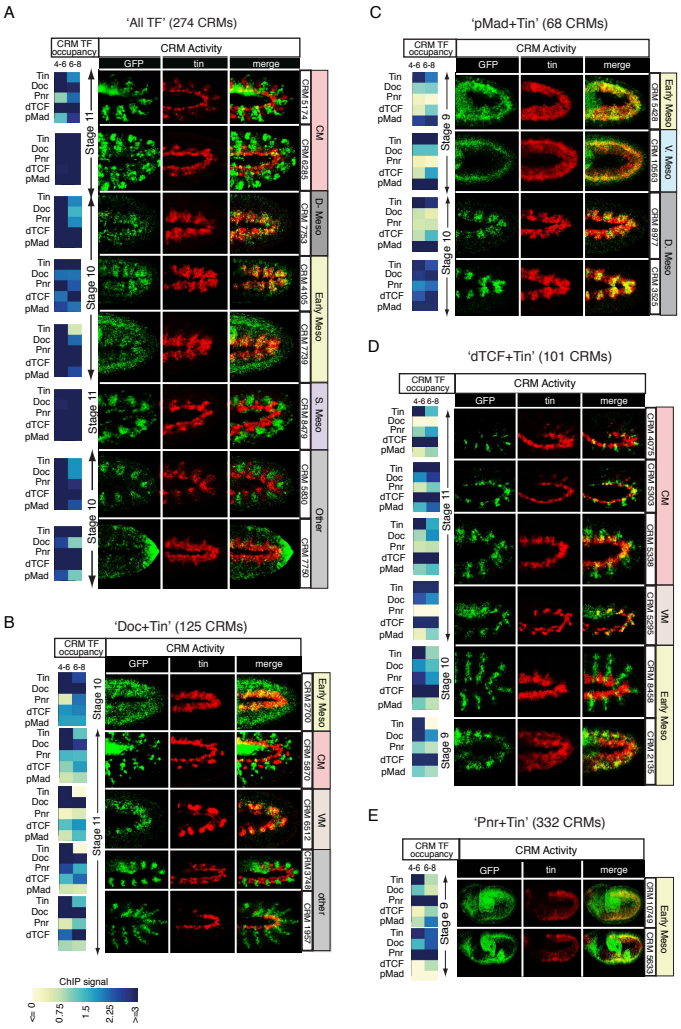


Figure S3 (related to Figure 3). **Activity of CRMs within different binding classes**

(A-E) CRMs from the 'All TF' and 'two-TF' classes in Figure 2C were selected for testing. Left panel: The TF binding signatures of individual tested CRMs shown as mean ChIP signal per CRM for each factor at both time-points (blue represents a high level of binding and yellow no binding). Right Panel: CRM spatio-temporal activity was assayed by *in situ* hybridization of embryos with a transgenic insertion, where the

CRM was placed in front of a minimal promoter and a *GFP* reporter. Panels show *in situ* hybridization using antisense RNA probes directed against a *GFP* reporter gene driven by the CRM (green) and *tinman* (red). Overlapping expression of *GFP* and *tinman* highlights CRM activity in the mesoderm and its derivatives, indicated by the yellow area of coexpression (merge panel). See Table S5 for a complete description.

(A) The binding signatures (left) and spatio-temporal activity (right) of eight individual tested CRMs from the 'All TF' class (note that as expected they generally have high binding signals for all TFs at one or both time points). Two of these CRMs have segmentally repeated activity encompassing the cardiac mesoderm (CRM 5174, 6285), one has specific activity in dorsal mesoderm (CRM 7753), two in early mesoderm (CRM 4105, 7739), one in the somatic mesoderm (CRM 8479) and two regulate expression in other non-mesodermal lineages (CRM 5830, 7750). In summary, 91.6% of active CRMs from this class (22/24) regulate expression in mesodermal lineages, 75% of which are active in cardiac mesoderm (CM) or visceral mesoderm (VM), indicating that the co-binding of All TFs is highly predictive of activity preferentially in these tissues during specification.

(B) The binding signatures (left) and spatio-temporal activity (right) of the tested CRMs within the **'Doc+Tin'** class. Nine CRMs were examined, with four failing to give activity (CRM 4550, 8368, 5377, 3071). The other five CRMs regulate very distinct patterns of expression in the early mesoderm (CRM 2700, yellow rectangle), visceral muscle (CRM 6512), ectoderm (CRM 3748 and 1957) and a subset of cardiac cells and somatic muscle (CRM 5870). These results indicate that the combinatorial binding of Doc with Tin is generally not sufficient to regulate expression in the cardiac mesoderm, but rather may provide a modulating role, restricting activity to very defined spatial and temporal domains.

(C) The binding signatures (left) and spatio-temporal activity (right) of five tested CRMs from the **'pMad+Tin'** class. While one CRM did not give any activity (CRM 8671), the remaining four CRMs regulate expression in the early mesoderm (stage 9), ventral mesoderm and/or dorsal mesoderm at stage 10 (see Table S5). CRM 5428 recapitulates the pan-mesodermal expression of *tin* at stage 9 (yellow rectangle). CRM 10563 activity is restricted to the ventral mesoderm at stage 9 (blue rectangle), indicating that dorso-ventral patterning of the mesoderm already occurs before the restriction of *tin* expression to the dorsal mesoderm at stage 10.. The opposite pattern

was observed for CRM 8977 and 3525, which regulate expression exclusively in the dorsal mesoderm at stage 10 (grey rectangle), indicating positive input from the Dpp pathway and the tissue selector activity of Tin to provide the competence for expression in mesoderm.

(D) The binding signatures (left) and spatio-temporal activity (right) of the eight CRMs tested within the '**dTCF+Tin'** class. While two CRMs did not show any activity (CRMs 5451, 10155), the remaining six CRMs direct expression in mesoderm. Three CRMs (4075, 5303, 5338) regulate activity in the cardiac mesoderm (CM, pink rectangle). CRM 5295 directs exclusive expression in visceral mesoderm (VM, light brown rectangle). Finally, two CRMs drive expression in a striped pattern in the early mesoderm (CRM 2135, 8458, light yellow rectangle). These six CRMs nicely recapitulate the domain of activity for the two TFs occupying them at high levels (dTCF and Tin). Wingless signaling is essential for normal heart development; however, it was not known if this requirement is direct or indirect due to the activation of Sloppy paired (Slp), which in turn blocks VM development within the cardiogenic domain (Lee and Frasch, 2000). Our results demonstrate that dTCF occupies a large number of enhancers ('dTCF+Tin' and 'All TF' CRMs, which are capable of regulating expression in CM (37% and 46% of CRMs respectively), indicating that Wg signaling has direct regulatory input in cardiogenesis in *Drosophila*.

(E) The binding signature (left) and spatio-temporal activity (right) of tested CRMs within the **'Pnr+Tin'** class. Of the 7 tested CRMs, only 2 (CRM 10749, 5633) were capable of functioning as an enhancer *in vivo*. Their activity was specific to the early mesoderm and became inactive from stage 10 of development. This early mesoderm activity does not correlate with *pannier* (*pnr*) expression, which begins in the dorsal mesoderm at stage 10. The 5 inactive CRMs (818, 3828, 219, 2036, 757) have equally high Pnr signal, and it is therefore unclear why these regions could not function as CRMs *in vivo*. We suspect that these regions have a function different from enhancer elements and are currently investigating this in detail.

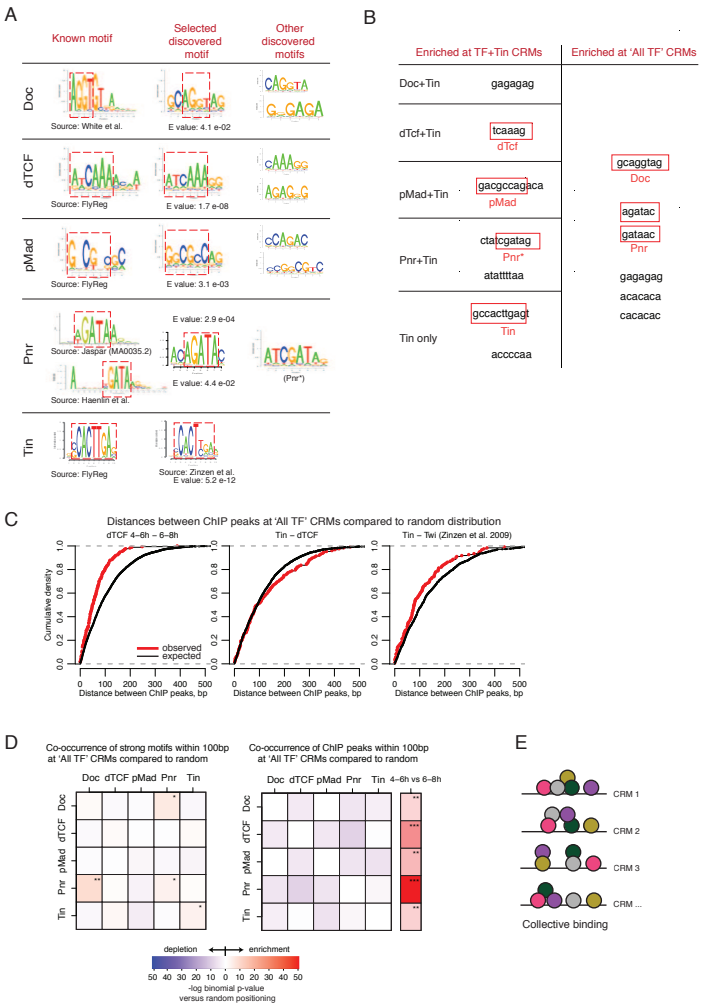


Figure S4 (relates to Figure 4). **Properties and distribution of TF-specific sequence motifs at CRMs**

(A) Motifs discovered at TF-bound regions are similar to the known motifs. The known TF binding sites for dTCF and pMad are from FlyReg (Bergman et al., 2005). As there is no known PWM for the *Drosophila* Doc protein, we compared the similarity to vertebrate Tbx6 from White et al., 2005 (note that as there is only 59%

identity in the proteins DNA binding domains, their DNA binding specificity may have diverged significantly). For Pnr, the two known motifs are for the *Drosophila* (Haenlin et al., 1997) and vertebrate (MA0035.2 from Jaspar) proteins. The Tin PWM used in this study was published previously (Zinzen et al., 2009) and is shown next to an earlier version from FlyReg. *De novo* motif discovery was performed using Weeder on 400bp regions centered on the peaks derived from the 100 highest-scoring ChIP regions. For Doc, dTCF and pMad this yielded motifs similar to known signatures that were selected for further analyses. For Pnr, the motif discovered this way (Pnr*, third column) was enriched in the 'Pnr+Tin', but not the 'All TF' class, and is similar to the DPE promoter element, which contains a GATA sequence. A slightly different Pnr motif (shown in the central column), which is more similar to the known consensus, was discovered in the 'All TF' CRMs using RSAT (panel B) and was used for further analyses. E-values for similarity between the known and discovered motifs computed by STAMP (Mahony and Benos, 2007) are shown under the discovered motif logos. Red dashed squares highlight the part of each motif that is most similar to the previously known motif and its discovered version. Additional discovered signatures are presented in the third column. Note that many of these alternative motifs are shorter versions of the selected ones (redundant motifs were omitted for clarity).

(B) Direct comparison of k-mer enrichment between 'All TF' and 'two-TF' CRMs. Results from RSAT oligo-diff analysis (van Helden, 2003) performed with repeat-masked sequences of 'two-TF' versus 'All TF' CRM classes. The table shows all k-mers that have a significant enrichment in either CRM class (oligo-diff 'occ_sig'>2 corresponding to adjusted p-value <= 0.01; for k-mers enriched in the "Pnr+Tin" class, 'occ_sig'>10 is used instead). Shown are consensus patterns resulting from the alignment of overlapping k-mers using RSAT pattern-assembly tool. Parts of the consensus matching a 'heart' TF's consensus are highlighted by a red box and labeled with the respective TF name. Note for Pnr, two different motifs were observed: the 'tatcgata' palindrome (Pnr*; enriched in the 'Pnr+Tin' class) corresponding to the motif reported by Weeder and the 'agatac' version matching the known Pnr site that is enriched in the 'All TF' class.

(C) Cumulative density plots showing distances between specific TF ChIP peaks at 'All TF' CRMs compared to a random distribution. Left: ChIP peaks for the same TF

at two consecutive time-points are significantly closer to each other than expected at random. As it is expected that these binding events are mediated via the same motifs within a CRM at each time-point, this result indicates that the ChIP-chip data has the resolution to detect TF occupancy in very close proximity (shown for dTCF). Middle and right: ChIP peaks of TF pairs: dTCF–Tin and Tin–Twi (based on data from Zinzen et al., 2009). The cumulative distribution of the distances between Tin and Twi (Twist – a general mesodermal TF) binding peaks are closer then expected at random, indicating that these TFs tend to bind non-randomly at a distance close to each other. In contrast, Tin and dTCF (or any of the other heart TFs) binding peaks are located at a distance further away from each other than expected by random chance, within 'All TF' CRMs.

(D) Heatmaps summarizing the enrichment/depletion of TF-specific motifs (left panel) and TF binding peak pairs (right panel) within 100bp from each other at 'All TF' CRMs compared to random distribution. Enrichment (shades of red) or depletion (shades of blue) scores are represented as –log (binomial p-value) for the proportion of observed versus random peak pairs found within 100bp windows. Pairs significantly enriched in proximity of each other are labeled with * (permutation test p<5e-2), ** (p<5e-3) and *** (p<1e-9). The co-localization of peaks for the same TFs at 4-6hr and 6-8hr is used as positive control to detect peak proximity, as the binding site occupied by a TF within an enhancer is unlikely to change from one time point to the next. Note that these values are not corrected for multiple testing. Taken together, these data support the conclusions of the CRM grammar analysis using SCRM (Noto and Craven, 2006) that also considers more complex spatial relationships. See Extended Supplementary Procedures for details.

(E) The *TF collective* binding model proposed for 'All TF' CRMs based on their collective occupancy and their motif content and TF peak distance analysis. The five TFs (colored circles) are bound in various order and at irregular distances from each other. 'All TF' CRMs also have lower numbers of pMad, dTCF and Tin motifs compared to their 'two-TF' CRMs, suggesting co-operative binding (horizontal lines), rather than independent assembly. It is possible that depending on which strong TF-specific binding sites are present at each CRM, that different TFs act to "anchor" the entire TF collective to DNA (illustrated by 'dipped' circles), while the other factors may bind more loosely to DNA, relying also on protein-protein interactions between these TFs or common cofactors.

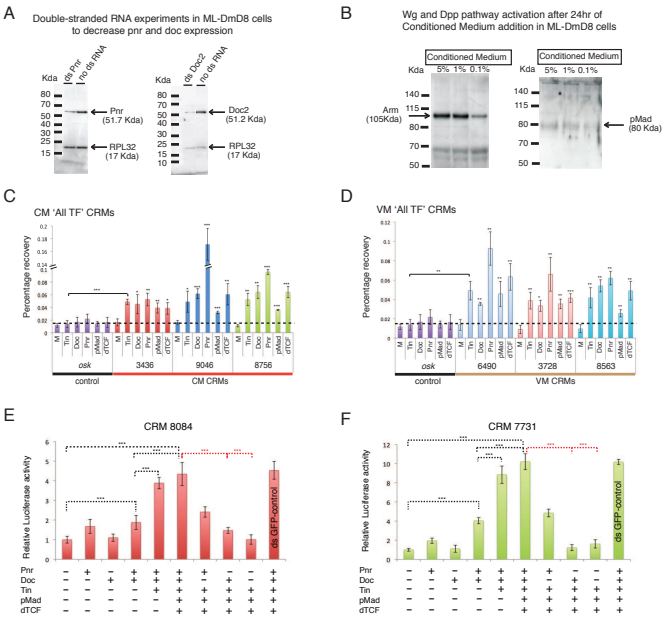


Figure S5 *(related to Figure 5)*: **A cell-based model for cardiac enhancer activity**

(A, B) Establishing a cell-based model using DmD8 cells. **(A)** Western blot of Pnr and Doc showing that both TFs are present in DmD8 cells. dsRNAi was used to knock-down both TFs, to obtain cells that lack all five TFs to assess the basal levels of enhancer activity. RPL32 was used as a loading control. Note, these experiments also clearly demonstrate the specificity of the antibodies used for the ChIP experiments. **(B)** Western blot of conditioned media showing Wg and Dpp signaling

pathway activation. Untransfected DmD8 cells were incubated with various amounts of conditioned medium for 24h, after which the cells were lysed and the supernatant was collected after high-speed centrifugation. Wg and Dpp pathway activation was determined by western blot using antibodies recognizing Armadillo, which becomes stabilized after Wg-mediated receptor activation, and phosopho-Mad, the phosophorlyated form of Mad, which only occurs after Dpp-receptor activation.
(C, D) All five TFs occupy cardiac mesoderm and visceral mesoderm enhancers in a cell culture based model. ChIP experiments in DmD8 cells containing Pnr, Doc, activated dTCF, pMad and Tin. TF occupancy was assessed by real-time PCR as percentage recovery of input DNA. Seven regions were assessed; a negative control region (the *oskar* locus (*osk*) which should not be bound by any of these TFs) and six regions that are bound by all five TFs *in vivo* ('All TF' CRMs). Three of these CRMs drive activity in the cardiac mesoderm (CM) (**C**) while three are active in the visceral mesoderm (VM) (**D**). The CRMs tested are indicated underneath the histogram. M is a mock reaction where the specific antibody was replaced by normal rabbit serum. For each enhancer, the occupancy of each TF is significantly enriched on the 'All TF' CRM compared to that TF's enrichment on the *osk* control region (indicated by the solid line for Tin on CRM 3436). Error bars from three independent biological replicates. P-values (one-tailed Type 2 t-test): *=<0.05, **=<0.01, ***=<0.001.
(E, F) Luciferase assays in DmD8 cells to assess the activity of three 'All TF' CRMs (third CRM shown in Figure 5D, main text). The basal enhancer activity, in the absence of any of the five TFs, was determined by removing Pnr and Doc using dsRNAi (indicated in panels S5A). This level was set to 1 and the luciferase activity for all other experiments are expressed relative to this level (y-axis, relative luciferase activity). While the presence of Pnr and Doc caused a significant increase in enhancer activity, the addition of Tin with Pnr and Doc had an even more dramatic effect. The presence of all five TFs was required for maximal enhancer activation. Note, removal of Pnr or Pnr+Doc reduced all three enhancer's activity back to the basal level (red dashed lines and asterik), even though the other transcription factors are still present. Note, the level of transfected DNA for all experiments was kept constant. Error bars are from two biological replicates, each conducted in triplicate. P-values (two-tailed Type 3 t-test): *=<0.05, **=<0.01, ***=<0.001.

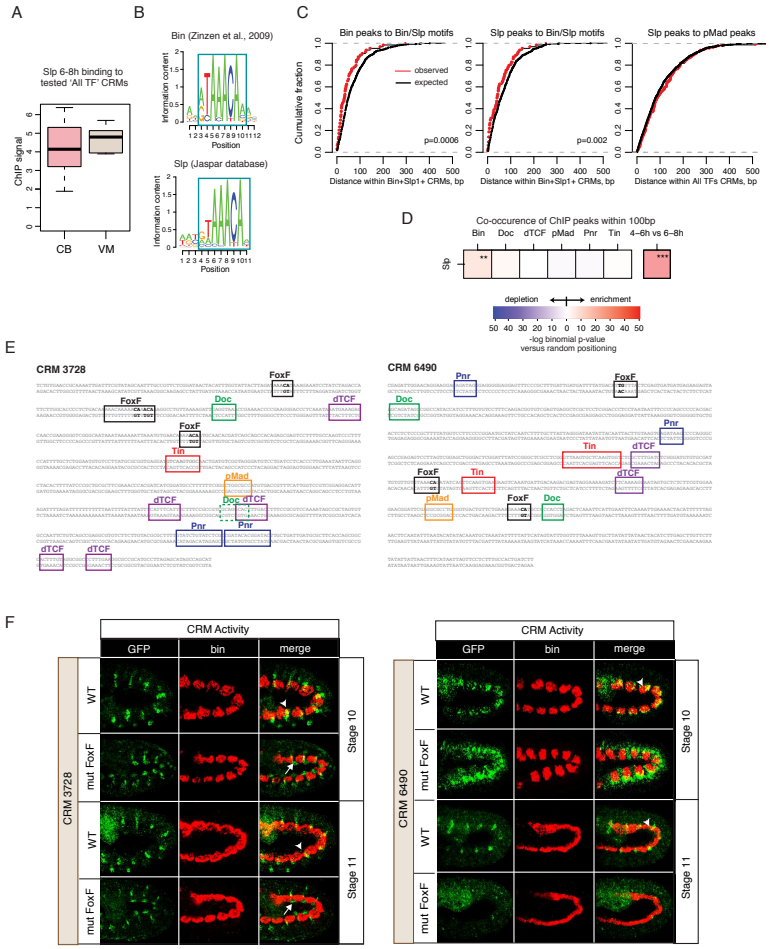


Figure S6 (related to Figure 7)**. Additional properties of Sloppy-paired bound CRMs and FoxF motifs**

(**A**) Sloppy paired (Slp) 6-8hr ChIP signals at the tested 'All TF' CRMs showing cardiac mesoderm (CM) and visceral mesoderm (VM) activity. **(B)** Similarity of known Biniou and Slp motifs, including the source. (**C**) Cumulative density plots showing the distributions of distances between Biniou (Bin) binding peaks and the Bin/Slp motif (left), Slp binding peaks and the Bin/Slp motif (middle), and between

Slp and pMad peaks (right). The TF binding peaks for Bin and Slp are both located in closer proximity to the Bin/Slp motif compared to random. This in line with the closer proximity of the observed Slp and Biniou ChIP-binding peaks (Fig. 7B), which was not observed between Slp and any other TF (shown here for Slp and pMad binding). **(D)** A heatmap summarizing the enrichment/depletion of Slp ChIP peaks in proximity of other TF ChIP peaks at 6-8 hrs of development. Enrichment (shades of red) or depletion (shades of white) scores are represented as –log (binomial p-value) for the proportion of observed versus random peak-pairs found within 100bp windows. Pairs significantly enriched in proximity of each other are labeled with **(p<5e-3) and *** (p<1e-9). The co-localization of Slp peaks at 4-6hr and 6-8hr (right box) is used as positive control for peak proximity, as the binding site occupied by Slp within a CRM is unlikely to change between time-points. **(E)** The sequence of the ChIP-defined regions that were cloned and assayed for enhancer activity *in vivo*. The TF binding sites for Pnr, Doc, Tin, pMad and dTCF are indicated. Both enhancers have three FoxF sites, all three of which were mutated. The base pairs changed are indicated in bold. The nucleotide changes are as follow: on CRM 3728, first FoxF motif: CA is replaced by TC, second FoxF motif: CA is replaced by TT and ACA by TTG, third FoxF motif: ACA is replaced by TTG, on CRM 6490, first FoxF motif: TG is replaced by CA, second motif: CA is replaced by AC, third motif: CA is replaced by AC.

(F) The FoxF motif mediates enhancer activity in visceral mesoderm and enhancer repression in cardioblasts**.** The FoxF motifs within CRM 3728 and CRM 6490 were mutated, as indicated in (E). Double fluorescent *in situ* hybridization of transgenic embryos containing the wild-type (WT) and mutant (mut FoxF) enhancers using an anti-*GFP* (enhancer reporter, green) and anti-*biniou* (a visceral mesodermal marker, red) probes. The wild-type CRM 3728 and 6490 are active in the visceral mesoderm at stages 10-12, indicated by *biniou* and *GFP* colocalization (yellow, arrow-heads). The WT enhancers are not active in the heart (Fig. 7C). When the FoxF sites are mutated, the activity of CRM 3728 in visceral mesoderm is strongly reduced, while the activity of CRM 6490 appears to be completely absent at stage 12. These results, in addition to the *in vivo* occupancy of Biniou on these CRMs, indicate that Biniou is regulating these enhancers activity in visceral mesoderm. In the cardiac mesoderm, this site is occupied by Slp and when mutated both enhancer's activity become derepressed in the heart (shown in Fig. 7C A'-C', D'-F', and arrow above),

demonstrating that the collective occupancy of the heart TFs on these 'VM enhancers' is functional.

Supplementary references

Achcar, F., Camadro, J.M., and Mestivier, D. (2009). AutoClass@IJM: a powerful tool for Bayesian classification of heterogeneous data in biology. Nucleic Acids Res *37*, W63-67.

Bailey, T.L., Williams, N., Misleh, C., and Li, W.W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res *34*, W369-373.

Bergman, C.M., Carlson, J.W., and Celniker, S.E. (2005). Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. Bioinformatics *21*, 1747-1749.

Bickel, P.B., N. Brown, JB. Huang, H , and Zhang, N. (2010). Subsampling methods for genomic inference, Vol 4.

Celniker, S.E., Wheeler, D.A., Kronmiller, B., Carlson, J.W., Halpern, A., Patel, S., Adams, M., Champe, M., Dugan, S.P., Frise, E.*, et al.* (2002). Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. Genome Biol *3*, RESEARCH0079.

Couderc, J.L., Godt, D., Zollman, S., Chen, J., Li, M., Tiong, S., Cramton, S.E., Sahut-Barnola, I., and Laski, F.A. (2002). The bric a brac locus consists of two paralogous genes encoding BTB/POZ domain proteins and acts as a homeotic and morphogenetic regulator of imaginal development in Drosophila. Development *129*, 2419-2433.

Cripps, R.M., Zhao, B., and Olson, E.N. (1999). Transcription of the myogenic regulatory gene Mef2 in cardiac, somatic, and visceral muscle cell lineages is regulated by a Tinman-dependent core enhancer. Dev Biol *215*, 420-430.

Gajewski, K., Kim, Y., Lee, Y.M., Olson, E.N., and Schulz, R.A. (1997). D-mef2 is a target for Tinman activation during Drosophila heart development. EMBO J *16*, 515-522.

Haenlin, M., Cubadda, Y., Blondeau, F., Heitzler, P., Lutz, Y., Simpson, P., and Ramain, P. (1997). Transcriptional activity of pannier is regulated negatively by heterodimerization of the GATA DNA-binding domain with a cofactor encoded by the u-shaped gene of Drosophila. Genes Dev *11*, 3096-3108.

Knirr, S., and Frasch, M. (2001). Molecular integration of inductive and mesoderm-intrinsic inputs governs even-skipped enhancer activity in a subset of pericardial and dorsal muscle progenitors. Dev Biol *238*, 13-26.

Mahony, S., and Benos, P.V. (2007). STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res *35*, W253-258.

Meila, M. (2005). Comparing Clusterings - An Axiomatic View. Proceedings of the 22nd International Conference on Machine Learning. Bonn, Germany, 2005.

Noto, K., and Craven, M. (2006). A specialized learner for inferring structured cis-regulatory modules. BMC Bioinformatics *7*, 528.

Odom, T., Dowell, R.D., Jacobsen E.S., Nekludova, L., Rolfe, P.A., Danford, T.W., Gifford, D.K., Fraenkel, E., Bell, G.T., and Young, R.A. (2006). Core transcriptional regulatory circuitry in human hepatocytes. Mol Syst Biol *2,* 2006.0017.

Pavesi, G., Mereghetti, P., Mauri, G., and Pesole, G. (2004). Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res *32*, W199-203.

Reim, I., Mohler, J.P., and Frasch, M. (2005). Tbx20-related genes, mid and H15, are required for tinman expression, proper patterning, and normal differentiation of cardioblasts in Drosophila. Mech Dev *122*, 1056-1069.

Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., Lin, M.F*., et al.* (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science *330*, 1787-1797.

Ryan, K.M., Hendren, J.D., Helander, L.A., and Cripps, R.M. (2007). The NK homeodomain transcription factor Tinman is a direct activator of seven-up in the Drosophila dorsal vessel. Dev Biol *302*, 694-702.

Schwartz, Y.B., Kahn, T.G., Nix, D.A., Li, X.Y., Bourgon, R., Biggin, M., and Pirrotta, V. (2006). Genome-wide analysis of Polycomb targets in Drosophila melanogaster. Nat Genet *38*, 700-705.

Thomas-Chollier, M., Sand, O., Turatsinze, J.V., Janky, R., Defrance, M., Vervisch, E., Brohee, S., and van Helden, J. (2008). RSAT: regulatory sequence analysis tools. Nucleic Acids Res *36*, W119-127.

Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A., Seal, R*., et al.* (2009). FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Res *37*, D555-559.

van Dongen, S. (2000). Performance criteria for graph clustering and Markov cluster experiments. Technical Report INS-R0012, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, May 2000.

van Leeuwen, F., Harryman Samos, C., Nusse, R. (1994). Biological activity of soluble Wingless protein in cultured Drosophila imaginal disc cells. Nature *368,* 342-344.

White, P.H., and Chapman, D.L. (2005). Dll1 is a downstream target of Tbx6 in the paraxial mesoderm. Genesis *42*, 193-202.

Zaffran, S., Reim, I., Qian, L., Lo, P.C., Bodmer, R., and Frasch, M. (2006). Cardioblast-intrinsic Tinman activity controls proper diversification and differentiation of myocardial cells in Drosophila. Development *133*, 4073-4083.