universität
wien

DISSERTATION

Titel der Dissertation

Computational analyses of transcription factor binding across species and tissues

Verfasserin

Anaïs Flore Bardet

angestrebter akademischer Grad

Doktorin der Naturwissenschaften (Dr. rer.nat.)

Wien, November 2012

Studienkennzahl:	A 091490
Dissertationsgebiet:	Dr. Studium der Naturwissenschaften Molekulare Biologie
Betreuer:	Dr. Alexander Stark

To Elsa

Acknowledgements

First I would like to thank Alex for sharing his passion for science, allowing me to work on great challenging projects, training me in bioinformatics but also helping me think critically and develop as a scientist.

Thanks to Julius Brennecke and his group for constructive feedbacks during our shared group meetings, my PhD committee members Julius Brennecke, Florian Raible and Magnus Nordborg for helpful discussions as well as Eileen Furlong and Florian Raible for reviewing this thesis.

Thanks to my collaborators who have been crucial to this work: Julia Zeitlinger, Qiye He, Antonio Meireles-Filho, Jonas Steinmann and Denes Hnisz, Wolfgang Lugmayr whose technical support makes our research much easier and Christopher Robinson for taking such a good care of all PhD students.

Thanks to all the current, past and adopted members of the Stark lab for creating such an amazing atmosphere that makes me happy to go to work everyday. Very special thanks go to Omar for being the best colleague ever, helping me with my endless scientific and personal interrogations, Bori and Dasha for their essential girly support and friendship, Gerald for always being excited about our projects, as well as Christian, Antonio, Evgeny, Christoph and Daniel.

Thanks to other people on campus with whom I have shared some relaxing times: Martina, Derya, Flavia as well as the badminton, squash & poker teams.

Thanks to the people who have first welcomed me in my Viennese scientific adventure and have become great friends: Alexandra, Pawel, German, Brian and Andi. Special thanks go to Alexandra for translating the abstract in German and to Brian for the English proofreading of this thesis.

Thanks to my awesome friends with whom I’ve shared fantastic travel adventures and who have watched and helped me evolve since I became interested in gene regulation during our master degree: Sarah, Elodie, Nadjia, PA, Romain, Lieng, Guillaume, Thomas and Clement.

Finally, I want to thank my parents for shaping me into the person I am today and supporting me in all my choices even from far away and in difficult times, but above all, I would like to thank Elsa for giving me all the strength I need to go through a lifetime of happiness.

Summary

One of the main determinants of development is the complex and specific regulation of gene expression. This is mediated through the binding of transcription factors to genomic DNA at specific regulatory sequences called enhancers. Combinations of transcription factors binding to enhancers can activate transcription in a condition specific manner such as in different cell-types or developmental time points. Comparative genomics has been widely used to identify divergence and conservation of enhancers using sequence conservation and transcription factor motif occurrences.

During my PhD, I have assessed the conservation of experimentally determined transcription factor binding sites across *Drosophila* species for the highly conserved transcription factor Twist, which is involved in mesoderm development in the early embryo. For this purpose, I have developed a computational pipeline for the comparison of transcription factor binding sites across conditions as well as a method to identify transcription factor binding sites at high resolution. We found that the binding landscape of Twist is highly conserved across six *Drosophila* species, which was surprising given recent studies about the variation of transcription factor binding events among vertebrates and among yeasts. Furthermore, we observed that Twist binding is dependent on sequence motifs for Twist and partner transcription factors. To determine if the dependence on other transcription factors holds true in different contexts, I have investigated the sequence determinants of tissue-specific binding of the transcription factors Clock and Cycle responsible for circadian rhythms in animal bodies. We found that Clock and Cycle bind together in heads and bodies of adult flies and that their binding at body-specific sites are defined by the sequence motif of the transcription factor Serpent.

Understanding how combinations of TFs bind to DNA in a condition-specific manner to regulate gene expression is a stepping-stone towards the elucidation of a regulatory code.

Zusammenfassung

Ein einflussreicher Faktor für die Entwicklung von Organismen ist die komplexe und spezifische Regulation der Genexpression. Transkriptionsfaktoren (TF) steuern Genregulation indem sie an Enhancer, regulative Sequenzen in der DNA, binden können und dort in flexiblen Kombinationen entwicklungs- und gewebsspezifische Transkription auslösen. In der vergleichenden Genomik, können mithilfe von Sequenzkonservierung und TF Motifsuche, Unterschiede und Gemeinsamkeiten in funktionellen Elementen gefunden.

Während meines Doktorats habe ich das Ausmaß der Konservierung experimentell ermittelter TF-Bindungsstellen für den TF Twist im Genom sechs *Drosophila* Spezies untersucht. Der untersuchte TF ist hochkonserviert und mitverantwortlich für die Entwicklung des Mesoderms in frühen Embryos. Um TF-Bindungsstellen in unterschiedlichen Bedingung zu vergleichen habe ich eine computergestützte Methode entwickelt, welche TF-Bindungsstellen mit hoher Genauigkeit identifizieren kann. Die Analyse zeigt, dass die Bindungslandschaft von Twist zwischen den *Drosophila* Spezies hochkonserviert ist. Dieses Ergebnis ist überraschend da es im Kontrast zu publizierten Studien über die Variabilität von TF-Bindungsstellen in Vertebraten und in Hefen steht. Des Weiteren haben wir festgestellt, dass das Binden von Twist von den Sequenzmotiven für Twist und seinen Partner TF abhängt. Um festzustellen ob Abhängigkeiten zwischen TF zu unterschiedlichen Bedingungen bestehen, habe ich die gewebespezifischen TF Clock und Cycle, welche für den circadianen Rhythmus in Tieren verantwortlich sind, untersucht. Das Ergebnis zeigt, dass in erwachsenen Fliegen Clock und Cycle gemeinsam sowohl in Köpfen als auch Körpern binden und dass die Bindung an „körperspezifischen Stellen“ durch eine Kombination der beiden TF, Clock und Cycle, als auch dem TF Serpent entsteht.

Das Verständnis wie Kombinationen von TF abhängig von Konditionen an die DNA binden und dadurch die Genexpression regulieren, ist ein wichtiger Schritt in der Aufklärung des regulatorischen Codes.

Contents

1. Introduction 1

1.1. Transcriptional regulation of gene expression 2

1.1.1. Transcription initiation by RNA polymerase II 2

1.1.2. Activation and repression by transcription factors 2

1.1.3. Combinatorial regulation at enhancers 4

1.1.4. Influence of chromatin states 4

1.2. Functional characterization of enhancers 6

1.2.1. Identification of transcription factor binding sites 6

1.2.2. Conservation of transcription factor binding sites 9

1.2.3. Identification of enhancers 11

1.2.4. Functionality of enhancers 12

1.2.5. Context-specific activity of enhancers 13

2. Aims of the thesis 14

3. Methods 15

3.1. Comparative analysis of ChIP-seq data 15

3.2. High resolution peak finding in ChIP-seq data 15

4. Results & Discussion 16

4.1. Conservation of transcription factor binding during early *Drosophila* development 16

4.2. Tissue-specific sequence features of the *Drosophila* circadian clock transcription factors 20

References 22

Publications & Manuscripts 33

A. High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species 33

B. Cis-regulatory requirements for tissue-specific programs of the circadian clock 93

C. A computational pipeline for comparative ChIP-seq analyses 155

D. Identification of transcription factor binding sites from ChIP-seq data at high-resolution 179

List of figures

Figure 1: Transcription factor DNA binding domain 3

Figure 2: Transcriptional regulation by enhancers 5

Figure 3: Chromatin Immunoprecipitation (ChIP) 7

Figure 4: Improved resolution of the ChIP techniques 8

Figure 5: Conservation of transcription factor binding sites 10

Figure 6: Context specific binding of transcription factors 13

Figure 7: Targets of the dorso-ventral gene regulatory network 16

Figure 8: High conservation of Twist transcription factor binding sites across six *Drosophila* species 17

Figure 9: The *Drosophila* circadian clock 20

1. Introduction

Each one of us has a unique genome made of DNA encoding our genetic information. Although most of our cells in our bodies contain the exact same copy of our genome, they exhibit a wide variety of functions. The mechanism by which one fertilized egg gives rise to a diverse set of cells with specialized functions is called cell differentiation and is fundamental to the development of multicellular organisms. The identity of a cell is defined by the specific set of genes that it activates by reading the DNA: genes are transcribed into messenger RNA (mRNA) and then translated into proteins following the genetic code. The human genome contains about 20000 protein-coding genes (International Human Genome Sequencing Consortium 2004) and other eukaryotic genomes contain between 6000 to 60000 but this number as well as their genome size does not correlate with phenotypic complexity (Pray 2008). In fact, only 1.5% of the human genome consists of genes and the rest of the genome, which was initially thought to be 'junk' DNA', determines when and where mRNAs are produced leading to a specific set of proteins. This tight control of mRNA production is part of the regulation of gene expression and its specific regulation at the transcriptional level is the subject of this thesis.

1.1. Transcriptional regulation of gene expression

1.1.1. Transcription initiation by RNA polymerase II

Transcription of DNA into RNA is performed by RNA polymerase II (Pol II) and is initiated by its recruitment to gene transcriptional start sites (TSSs). Regulatory proteins able to recognize specific DNA sequence elements, called general transcription factors (TFs), bind to core promoter regions, encompassing gene TSSs, and assemble into the preinitiation complex to recruit Pol II (Smale and Kadonaga 2003). This basal transcriptional machinery already enables low levels of transcription but the additional interaction of specific TFs is required to regulate Pol II activity at specific genes.

1.1.2. Activation and repression by transcription factors

Two main structural domains characterize TFs: The DNA binding domain enables TFs to bind to DNA through the recognition of specific DNA sequences, and the regulatory domain enables protein-protein interactions to control Pol II activity.

The DNA binding domains of TFs specifically recognize 6 to 12 base pairs (bp) long DNA sequence motifs (**Figure 1**) (Spitz and Furlong 2012). Consensus motifs are typically represented as position weight matrices (PWMs) integrating the TF affinity for each position (assuming independence between the positions) (Stormo and Zhao 2010). Families of TFs are defined according to the structure of their DNA binding domain and members of the same family often recognize similar DNA sequence motifs (Wei et al. 2010).

Regulatory domains of TFs confer both activating and repressing roles by promoting protein-protein interactions. Activators stimulate Pol II activity by binding directly to the general TFs of the basal transcriptional machinery

(Ptashne and Gann 1997) or to intermediate players called co-activators such as the mediator complex (Malik and Roeder 2010). Repressors can either interact with an activator therefore repressing its function, recruit corepressors, or compete with an activator for the same DNA binding site, therefore blocking its function.

The effect of individual TFs on transcription can be regulated by their own activation (e.g. by binding of a ligand or by biochemical modifications), by the TF concentration in the nucleus and by variation in the DNA sequence motif to modulate the affinity of the corresponding TF. But many TFs also interact with each other, forming homo- or hetero-dimers. This feature is particularly common for some TF families, such as helix-loop-helix or leucine zipper TFs. The vast number of possible combinations of transcriptional activators and repressors enable a complex regulation of gene expression.

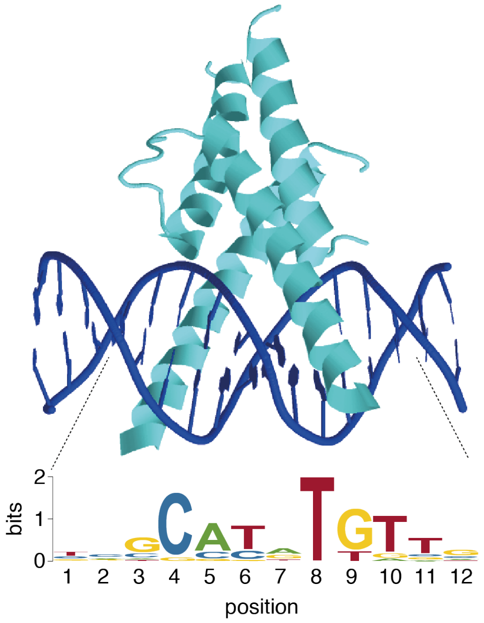


Figure 1: Transcription factor DNA binding domain. Crystal structure of a helix-loop-helix domain bound to DNA (from the protein data bank accession 2QL2) and motif logo (Schneider and Stephens 1990) representing the *Drosophila* Twist PWM (from Jaspar accession MA0249.1)

1.1.3. Combinatorial regulation at enhancers

Sequence-specific TFs that are required to modulate the basal activity of the transcriptional machinery bind to regulatory elements called enhancers (**Figure 2**). Enhancers are defined as being sufficient to activate the transcription of their target genes from a minimal promoter (Banerji et al. 1981) and can function independently of their location and orientation relative to the gene TSS. Enhancers, also called cis-regulatory modules (CRMs), are composed of clusters of different transcription factor binding sites (TFBSs). Importantly, it is the combinatorial binding of TFs to enhancers that regulates gene expression in a spatio-temporal manner (detailed in section 1.2). Additionally, to ensure robust and precise patterns of gene expression, most eukaryotic genes are regulated by multiple enhancers, each contributing to its gene's activity in an redundant (e.g. shadow enhancers) or additive spatiotemporal manner (Hong et al. 2008; Barolo 2012; Frankel et al. 2010). Whereas proximal enhancers can recruit TFs directly in the vicinity of the transcriptional machinery, distal enhancers are believed to be brought to their specific target genes by DNA looping (Blackwood and Kadonaga 1998; Bulger and Groudine 1999; Kagey et al. 2010). Additional regulatory elements are responsible for shaping the three-dimensional organization of the genome by promoting or blocking enhancer-promoter interactions, such as tethering elements (Calhoun et al. 2002) or insulators (e.g. CTCF) (Valenzuela and Kamakaka 2006; Gaszner and Felsenfeld 2006; Phillips and Corces 2009).

1.1.4. Influence of chromatin states

Another important feature of transcriptional regulation is the physical access of TFs to promoters and enhancers (Ong and Corces 2011). Eukaryotic DNA sequences are packaged into a higher order structure called chromatin, consisting of nucleosomes, which are made of DNA wrapped around histone protein octamers. Typically, active regulatory regions are depleted from nucleosomes ("open") while inactive regions are nucleosome dense ("closed") (Lee et al. 2004). 'Pioneer' TFs are able to bind non-accessible chromatin and recruit remodelling complexes, such as the SWI/SNF complex (Clapier and Cairns 2009), that can displace or remove nucleosomes at promoters and enhancers to facilitate binding of TFs and activate transcription (Lupien et al. 2008; Harrison et al. 2011; Nien et al. 2011; Zaret and Carroll 2011). TFs can also recruit histone-modifying enzymes, such as the histone acetyltransferase CBP/p300 (Visel et al. 2009), to biochemically modify the histone tails of nucleosomes. Histone modifications can then be used as markers for genomic regions of certain functions (Heintzman et al. 2007; Rando 2007): for example, histone H3 tri-methylation on lysine 4 (H3K4me3) and histone H3 acetylation on lysine 9 (H3K9ac) mark active promoters, H3K36me3 marks the body of transcribed genes, H3K4me1 and H3K27ac mark enhancers and H3K27me3 and H3K9me2/3 are repressive marks.

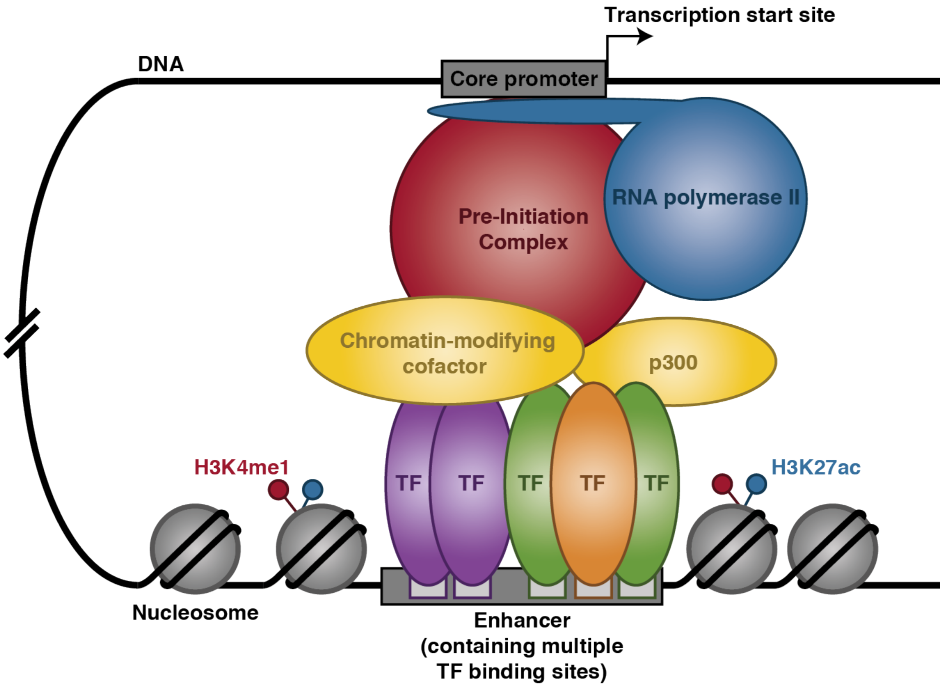


Figure 2: Transcriptional regulation by enhancers. Enhancers are DNA sequence elements that can recruit specific transcription factors to activate gene expression. (Adapted from Maston et al. 2012)

1.2. Functional characterization of enhancers

Over the last three decades, the functional characterization of enhancers has been a major challenge. How are those sequence elements able to drive specific patterns of gene expression underlying the development of multicellular organisms? One way to start answering this question is to identify TFBSs that constitute functional enhancers.

1.2.1. Identification of transcription factor binding sites

TF binding preferences can be determined experimentally *in vitro* using SELEX (Klug and Famulok 1994) or protein-binding microarrays (PBM) (Berger et al. 2006). The resulting sequence motifs are available is specific databases (Sandelin et al. 2004; Matys et al. 2003) but their number is still limited. They can be used to scan genomes for putative TFBSs. However, out of thousands of motif occurrences found, only a small fraction will be bound by their respective TF in a context-specific manner leading to a high number of false positives (Yáñez-Cuna et al. 2012a; Harrison et al. 2011).

Chromatin immunoprecipitation (ChIP) coupled to DNA microarrays ChIP-chip) (Ren et al. 2000; Iyer et al. 2001) or deep sequencing (ChIP-seq) (Johnson et al. 2007; Robertson et al. 2007) has now become the method of choice to identify *in vivo* TFBSs genome-wide. The ChIP method relies on the selection of DNA fragments bound by the TF of interest using a specific antibody directed against the TF (**Figure 3**). The resulting DNA fragments can then be hybridized onto a DNA chip or sequenced to recover the genomic positions that were bound by the TF of interest. An alternative method is to fuse the protein of interest with an epitope for which a good antibody is available, while making sure that it does not affect the endogenous TF activity.

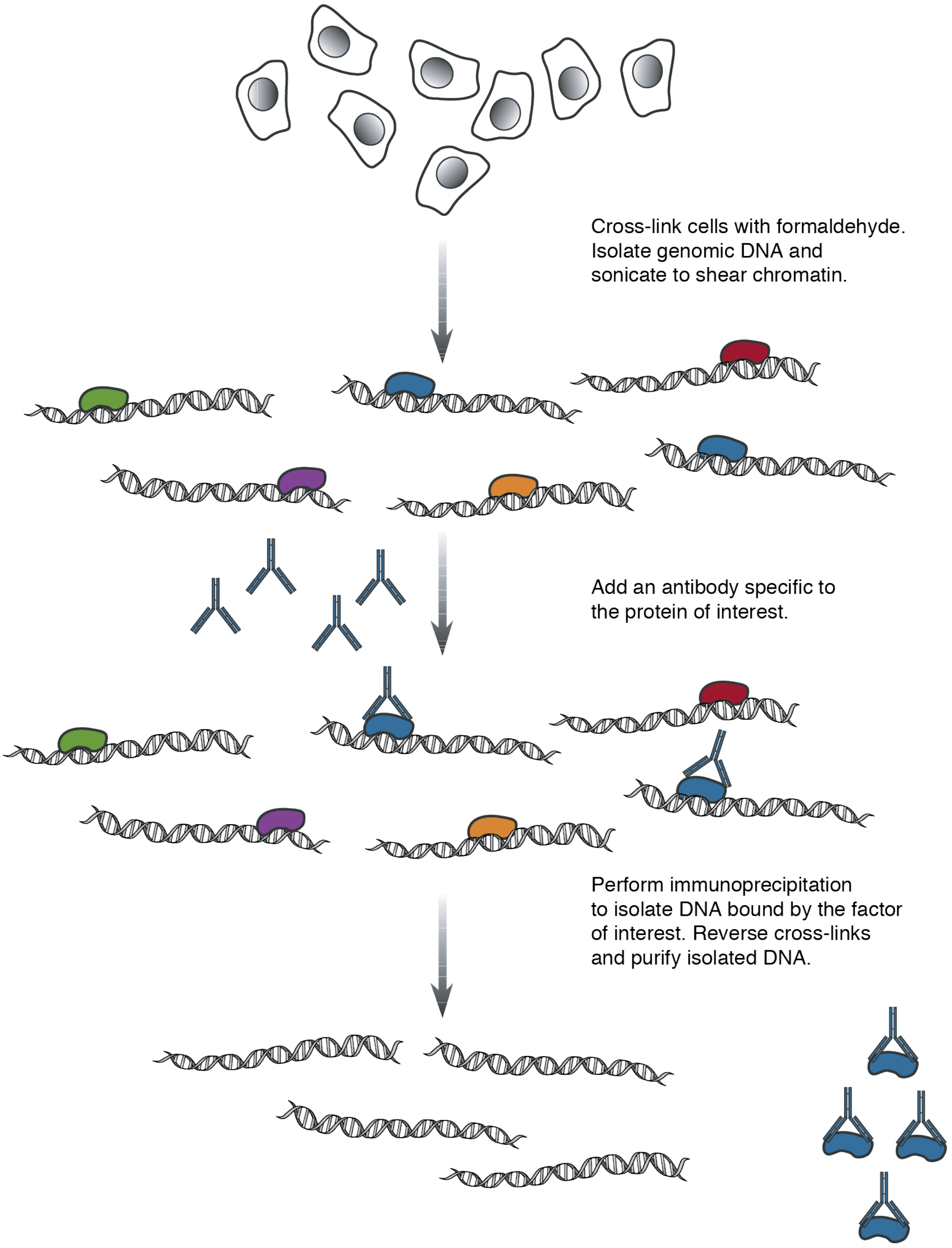

Cross-link cells with formaldehyde.
Isolate genomic DNA and
sonicate to shear chromatin.
Add an antibody specific to
the protein of interest.
Perform immunoprecipitation
to isolate DNA bound by the factor
of interest. Reverse cross-links
and purify isolated DNA.

Figure 3: Chromatin Immunoprecipitation (ChIP). In a population of cells, proteins bound to DNA are cross-linked, chromatin is isolated and fragmented, a specific antibody is used to only select the fragments bound by the protein of interest and proteins are removed from the DNA fragments. (Adapted from Hoffman and Jones 2009)

ChIP-seq commonly determines thousands of regions bound by a single TF, many of which might in fact not be functional (Li et al. 2008). Since ChIP-seq is a quantitative method, the enrichment of sequences within a region reflects the occupancy of the TF at this genomic position. This measure can then be used to select the best TFBSs that are less likely to be false positives since low occupancy TFBSs could in fact represent non-specific binding. TFBSs can also be searched for the specific sequence motif recognized by the TF. Although some TFs bind to very specific motifs always found in their TFBSs, in most of the cases, a substantial fraction of binding sites, even high occupancy ones, do not have the corresponding motif (e.g. E2F TF family) (Rabinovich et al. 2008). This could be explained by an artefact of the ChIP-seq method, where sequence-specific co-factors of the TF of interest and the corresponding DNA fragments could be retained during the cross-linking step. Furthermore, motif PWMs might not capture correctly all TF binding preferences (Badis et al. 2009).

Progress in DNA sequencing techniques (Metzker 2010) has revolutionized the use of the ChIP technology. ChIP-seq has considerably improved the resolution of ChIP-chip to identify TFBS genomic positions and more recently the ChIP-exo method has pushed the resolution from hundreds to tens of nucleotides (Rhee and Pugh 2011) (**Figure 4**). Since enhancers are composed of clusters of TFBSs, the experimental gain in resolution is crucial to

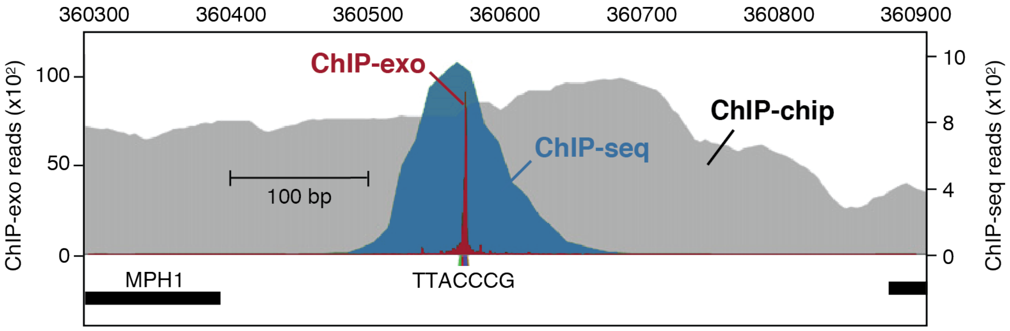


Figure 4: Improved resolution of the ChIP techniques. Comparison of the ChIP-chip, ChIP-seq and ChIP-exo resolution for the Reb1 yeast transcription factor. (Adapted from Rhee and Pugh 2011)

their identification. However, computational methods that identify TFBSs in ChIP-seq data (Wilbanks and Facciotti 2010) do not always take advantage of the experimental resolution. I address this limitation in the section 3.2 of this thesis.

1.2.2. Conservation of transcription factor binding sites

The increasing number of genomes being sequenced enables the study of natural selection underlying evolutionary processes. Pairwise as well as multiple genome comparisons have revealed a high degree of sequence conservation of protein-coding genes corresponding to the conservation of their function. When assuming that functional sequences are under negative selection and less likely to accumulate mutations, sequence conservation should also help to identify functional non-coding regulatory elements composed of TFBSs. Indeed, multiple genome alignments for mammalian (Lindblad-Toh et al. 2011) or *Drosophila* species (Drosophila 12 Genomes Consortium et al. 2007) have been generated and sequence conservation alone has been successfully used to identify conserved TF motifs that constitute functional enhancers (Kheradpour et al. 2007). Motif conservation within experimentally determined TFBSs (e.g. by ChIP) has become a powerful approach for the identification of functional TFBSs. However, it remains limited since regulatory function has also been found to be conserved despite low sequence similarity (Blow et al. 2010; Meireles-Filho and Stark 2009).

More recently, ChIP-seq datasets assessing TF binding across different species have been generated independently of sequence conservation (**Figure 5**). A surprisingly low conservation of TFBSs was reported in vertebrate species in differentiating tissues (Schmidt et al. 2010), embryonic stem cells (Kunarso et al. 2010) or differentiated cell lines (Mikkelsen et al. 2010). Whereas slightly higher conservation results have been found in yeast species (Borneman et al. 2007) those studies suggested an extensive turnover of TFBSs. However, the conservation of TFBSs might not be a direct readout of the conservation of function. As mentioned above, out of thousands of identified TFBSs, some might not be functional and would therefore not be conserved. Furthermore, the effect of the loss of a TFBS could be compensated by the appearance of a new TFBS nearby targeting the same gene (Dowell 2010). This low conservation of TFBSs might also be due to the choice of transcription factor and developmental context as the constraints on regulation within differentiated tissues and during early development are expected to be different. In the section 4.1 of this thesis, I assess the conservation of TFBSs in a developmental context.

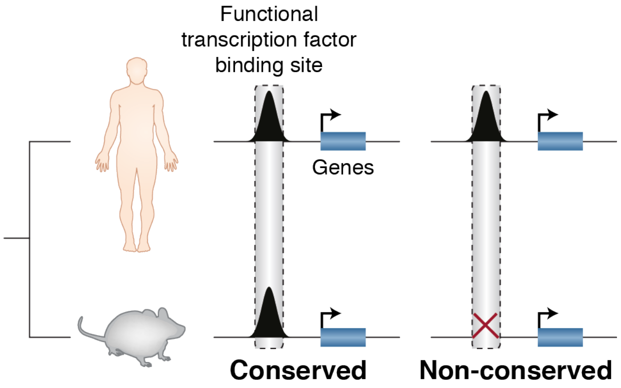


Figure 5: Conservation of transcription factor binding sites. Using genome alignments anchored on homologous genes, conservation of transcription factor binding sites, defined by ChIP experiments, can be assessed independently of sequence conservation. (Adapted from Pennacchio and Visel 2010)

1.2.3. Identification of enhancers

Co-occurrence of different TFBSs in clusters characterizes functional enhancers (Guss et al. 2001; Spitz and Furlong 2012). Several computational methods have been first successfully applied in the *Drosophila* embryo (Berman et al. 2002; Markstein et al. 2002; Schroeder et al. 2004). They identify enhancers by searching for clusters of TF motifs (Su et al. 2010; Hardison and Taylor 2012). Later, studying ChIP datasets, TF binding to enhancers has appeared to be highly cooperative and clustering of different *in vivo* TFBSs has become a common approach to identify enhancers (Zinzen et al. 2009). Extreme cases of regions bound by many TFs (i.e. highly occupied target or HOT regions) (MacArthur et al. 2009; Gerstein et al. 2010; modENCODE Consortium et al. 2010; Nègre et al. 2011) have even been shown to drive diverse spatio-temporal patterns of gene expression (Kvon et al. 2012).

By experimentally dissecting functional enhancers to study the contribution of specific TF motifs (Swanson et al. 2010) and examining the underlying sequence conservation, different models of enhancer structure or ‘grammar’ have been proposed so far: the ‘enhanceosome’ model defines a very constrained organization with fixed motif composition and positioning found in the interferon-beta enhancer in human (Thanos and Maniatis 1995; Panne 2008), whereas the ‘billboard’ model allows for a more flexible motif organization well characterized in the even skipped stripe 2 enhancer in *Drosophila* (Stanojevic et al. 1991; Arnosti and Kulkarni 2005; Meireles-Filho and Stark 2009).

Furthermore, complementary information about chromatin marks or chromatin accessibility can also be used to identify enhancers. ChIP-seq experiments can be performed for specific histone modifications that specifically mark enhancers such as H3K4me1 and H3K27ac. Other methods, involving sequencing, such as DNase I hypersensitivity (Boyle et al. 2008) or FAIRE (Giresi et al. 2007) enable the determination of accessible regions as regions depleted from nucleosomes and more likely to be active enhancers (Bell et al. 2011).

1.2.4. Functionality of enhancers

A direct approach to test the regulatory potential of an enhancer sequence is the use a reporter gene assay. *In vitro*, the candidate DNA sequence and a minimal promoter driving basal activity is placed in front of a luciferase reporter gene and transfected into cultured cells resulting in a quantitative estimate of enhancer activity (i.e. luminescence). *In vivo*, the candidate DNA sequence and a minimal promoter is placed in front of a reporter gene, injected into an embryo (e.g. mouse or *Drosophila*) and integrated in the genome resulting in a spatiotemporal pattern of gene expression (Hardison and Taylor 2012). Those assays are especially powerful to study the contribution of specific sequence motifs within enhancer by mutagenesis.

Functionality of TFBSs can also be inferred by considering the function of their target genes. When functional annotations are available, the enrichment of the target genes for particular functions can be compared to the function of the regulatory protein of interest (e.g. known target genes, gene ontology analyses). Occupancy and even dynamics of TF binding across conditions can also be correlated with the level of expression of their target genes (e.g. determined by RNA-seq). However, such analyses related to target genes rely on the accurate assignment of TFBSs to genes. This task is commonly performed computationally by assigning each TFBS to their closest gene TSS, which is not always appropriate, considering the ability of enhancers to drive expression from a distance. The mouse Sonic hedgehog enhancer is a prominent example of an enhancer acting from the intron of another gene more than 1Mb away from its gene promoter (Lettice et al. 2003). Experimental techniques involving sequencing have been developed to determine the chromosomal spatial contacts of the genome such as chromosome conformation capture (3C (Dekker et al. 2002) or more recently Hi-C (Lieberman-Aiden et al. 2009) and ChIP-PET (Fullwood et al. 2009)). They aim to overcome this problem by determining enhancer-promoter interactions but are still limited by their resolution (de Wit and de Laat 2012).

1.2.5. Context-specific activity of enhancers

Combinatorial binding of TFs to enhancers is the foundation of gene regulatory networks that control animal development by means of context-specific gene expression (Erwin and Davidson 2009; Davidson 2010). The profiling of multiple TFs using ChIP techniques have been performed in several conditions (e.g. cell types, tissues, developmental time-points) in specific studies (Zeitlinger et al. 2003; Harbison et al. 2004; Boyer et al. 2005; Sandmann et al. 2006), and more recently by large consortiums that aim to determine all functional elements in the genomes of model organisms and the human (e.g. ENCODE (ENCODE Project Consortium 2004), modENCODE (Celniker et al. 2009), Mouse ENCODE (Mouse ENCODE Consortium et al. 2012)). TF binding appears to be largely context-specific and several examples of such a dynamic and combinatorial TF binding have been demonstrated in different cell types, tissues or developmental time-points (Chen et al. 2008; Zinzen et al. 2009; Wilczyński and Furlong 2010; Lin et al. 2010; Junion et al. 2012). More recently, the concept of a master TF that can recruit TFs to cell-type specific enhancers has emerged (**Figure 6**) (Mullen et al. 2011; Palii et al. 2011; Trompouki et al. 2011). This dynamic organization of TFBSs highlights the existence of a regulatory code that could explain the diversity of transcriptional outputs (Yáñez-Cuna et al. 2012b). Using context-specific datasets, one can then predict the sequence features underlying context-specific binding (Narlikar et al. 2010; Lee et al. 2011; Yáñez-Cuna et al. 2012a; Busser et al. 2012), which I assess in the section 4.2 of this thesis.

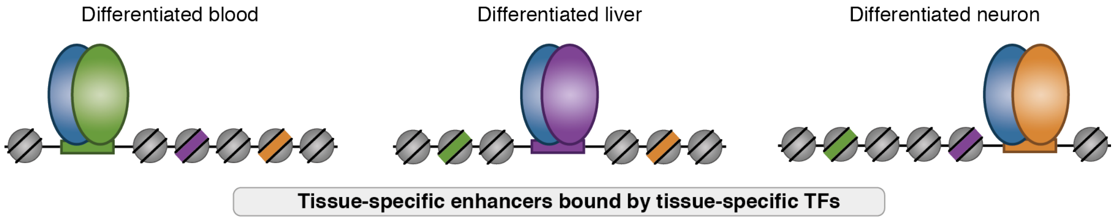


Figure 6: Context specific binding of transcription factors. Master transcription factors binding to enhancers recruit different context-specific transcription factors. (Adapted from Maston et al. 2012)

2. Aims of the thesis

The aims of my projects were to study TF binding across different species and conditions. More specifically, I have asked the following questions:

- Are developmental TFBSs conserved across species and how well does this conservation correlate with the underlying sequence conservation? (**Publication A**)
- What are the sequence determinants of tissue-specific TF binding? (**Manuscript B**)

To answer those questions, I have set up a computational pipeline for the comparative analysis of ChIP-seq data (**Publication C**), including the development of a new method to identify TFBSs in ChIP-seq data at high resolution (**Manuscript D**).

This work has been conducted between February 2009 and August 2012 in the laboratory of Dr. Alexander Stark at the Research Institute of Molecular Pathology (IMP) in Vienna, Austria. This thesis is written as a cumulative dissertation.

3. Methods

3.1. Comparative analysis of ChIP-seq data

The computational challenges for the analysis of ChIP-seq data for TFs have been previously described (Park 2009; Pepke et al. 2009). **Publication C** reports the computational pipeline developed for the comparative analysis of TF binding across species specifically and across different conditions in general. It describes and provides supporting code for data preprocessing, read mapping, translation into common coordinates, peak calling, data visualization, comparative analyses such as global similarity, peak conservation and quantitative changes, and introduces functional and sequence analyses.

3.2. High resolution peak finding in ChIP-seq data

A large collection of computational methods has been developed to identify TFBSs in ChIP-seq data (Wilbanks and Facciotti 2010). However, in our experience, they do not cope well with the recent technical advances that aim to determine TFBSs at high resolution (e.g. ChIP-exo (Rhee and Pugh 2011)). **Manuscript D** presents a new computational approach, peakzilla, which fully exploits the bimodal distribution of sequenced reads, to identify true binding events at high resolution.

4. Results & Discussion

4.1. Conservation of transcription factor binding during early Drosophila development

The transcription factor Twist is a key regulator specifying dorso-ventral patterning and mesoderm specification in the early *Drosophila* embryo (Baylies and Bate 1996; Levine and Davidson 2005). Twist has been the subject of many developmental genetics and genomics studies (Jiang et al. 1991; Leptin 1991; Ip et al. 1992; Sandmann et al. 2007; Zeitlinger et al. 2007; Zinzen et al. 2009) and many functional enhancers have been identified and dissected to determine their motif requirements (**Figure 7**) (Pan et al. 1991; Szymanski and Levine 1995; Markstein et al. 2004; Zinzen et al. 2006; Ozdemir et al. 2011; Reeves and Stathopoulos 2009).

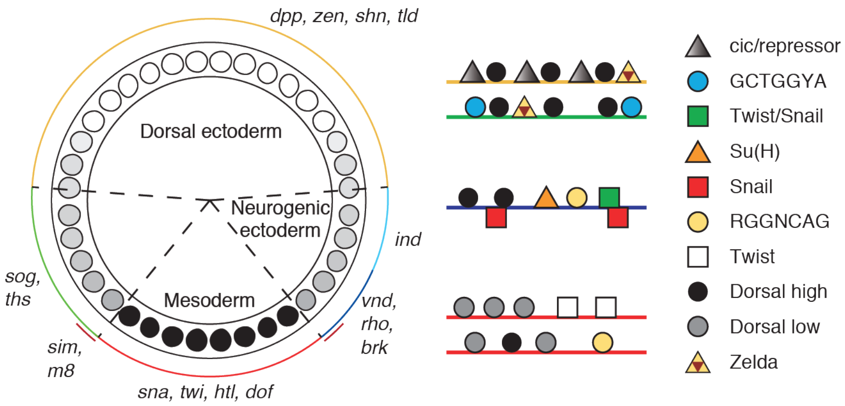


Figure 7: Targets of the dorso-ventral gene regulatory network. The Dorsal gradient divides the embryo into three main tissues: mesoderm, neurogenic ectoderm and dorsal ectoderm, each of which has specific targets genes with different motif organization at their enhancers. (Adapted from Reeves and Stathopoulos 2009)

In **publication A**, our collaborators Qiye He and Julia Zeitlinger (Stowers Institute for Medical Research, Kansas City, Missouri, USA) performed ChIP-seq experiments in the early *Drosophila* embryo for Twist in six *Drosophila* species: *D. melanogaster*, *simulans*, *yakuba*, *erecta*, *ananassae* and *pseudoobscura* (**Figure 8**). Their evolutionary distances, as measured by substitution per neutral site, are comparable up to human-chicken distances (Stark et al. 2007). Importantly, the Twist protein is highly conserved among these *Drosophila* species and an antibody raised against Twist in *melanogaster* cross reacts in the other species as shown in immunostainings (**Figure 8**). In an example at the Tinman locus (**Figure 8**), we see that a Twist binding peak in *melanogaster* is conserved in all other species matching the position of a known enhancer. Generally, we found that Twist binding is highly conserved across the six *Drosophila* species and that the conservation of *melanogaster* peaks in other species correlates with the evolutionary distances of the phylogenetic tree. This conservation holds true for the TF Snail in a closely related species (*D.melanogaster - D.simulans*) and is consistent with the results found for other developmental TFs at similar evolutionary distances (*D.melanogaster - D.yakuba*) (Bradley et al. 2010).

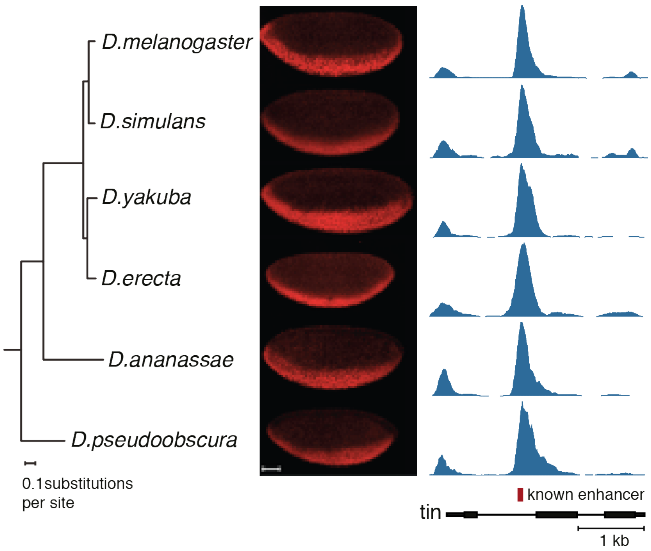


Figure 8: High conservation of Twist transcription factor binding sites across six *Drosophila* species. *Drosophila* phylogenetic tree, antibody staining for Twist in all species and conservation of transcription factor binding site at the Tinman enhancer. (Tree adapted from Stark et al. 2007)

We found 34% of the *melanogaster* Twist peaks to be conserved in all other species, which is 100 fold higher than the conservation found by Schmidt et al. for a differentiated liver TF (CEBPα) in five vertebrate species (0.3% conservation of human peaks in mouse, dog, opossum and chicken) (Schmidt et al. 2010). As discuss in the introduction, this might indicate that the regulation of developmental processes is indeed under stronger evolutionary constraints than within differentiated tissues. Furthermore, we assessed the functionality of peaks and found that conserved peaks are more likely to be functional (using expression data of Twist in mutant embryo, gene ontology categories of target genes or known enhancers). When additionally taking into account the higher number of TFBSs found in vertebrate species, this apparently low conservation of TFBSs in vertebrate species could stem from a lower proportion of vertebrate TFBSs being functional as well as an increased propensity for functional turnover (i.e. compensating loss and gain) of TFBSs enabled by the larger size of their genomes.

The high conservation of TFBSs provides a unique opportunity to identify functional features of TF binding and enhancer organization. We have shown that clustered peaks are preferentially conserved highlighting the importance of homotypic binding of TFs discussed in the **manuscript D** of this thesis (Gotea et al. 2010). We found that Twist binding is dependent on the specific sequences of the Twist motifs and not of the entire peak regions: Twist motifs are preferentially conserved in conserved peaks as opposed to in species-specific peaks. We also found that the quantitative changes in Twist binding can be explained by the quality of their Twist motifs. However, in only 24% of the cases, the conservation pattern of a *melanogaster* peak in other species matches exactly the conservation pattern of the Twist motif. Indeed, we found that partner TF motifs, including Snail and Dorsal, are significantly more conserved in conserved peaks than species-specific ones and can explain the conservation pattern of a twist peak when the Twist motif itself cannot. This emphasizes the importance of cooperative binding of TFs for enhancer functionality.

Finally, we investigated the functionality of low-occupancy peaks. As in any common ChIP-seq analysis, we had initially focused on the best *melanogaster* peaks or high occupancy peaks as we expect them to have less false positives. However, some low-occupancy peaks are still conserved above random and we found them to still contain more Twist motifs than the average genome and to be significantly conserved in all species, indicating that they might be functionally important. Using ChIP-chip datasets of Twist in different time points (Zinzen et al. 2009), we found that low-occupancy peaks at the same time point of our study correspond to peaks that get more strongly bound, and might thus be functional, at later time points during embryonic development.

4.2. Tissue-specific sequence features of the Drosophila circadian clock transcription factors

The circadian clock is a well-known example of a gene regulatory network that modulate the transcriptional outputs of their target genes in a cell-type specific manner (Abruzzi et al. 2011). The *Drosophila* circadian clock is composed of negative feedback loops controlled by the transcription factors Clock (CLK) and Cycle (CYC) that can heterodimerize to bind the E-box motifs (CACGTG) near their target genes (**Figure 9A**) (Hardin 2011). Although Clock and Cycle are nearly ubiquitously expressed in *Drosophila* tissues (Plautz et al. 1997), they are able to control various tissue-specific biological processes (Rey et al. 2011).

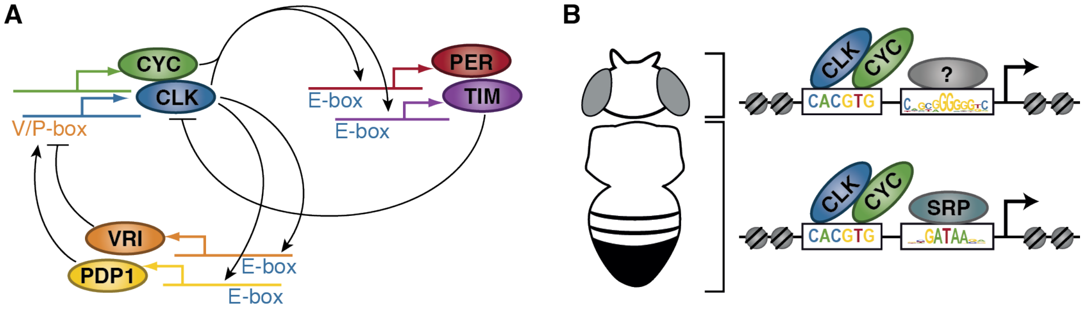


Figure 9: The *Drosophila* circadian clock. A. Transcription factors of the *Drosophila* circadian clock network. (Adapted from Doherty and Kay 2010).
B. Combinatorial binding of clock and cycle together with serpent specifies body specific targets.

In **manuscript B**, my collaborator Antonio C. A. Meireles-Filho (Stark group, Research Institute of Molecular Pathology, Vienna, Austria) performed ChIP-seq experiments for clock and cycle in the heads and bodies of adult flies. We found that as expected, Clock and Cycle share most of their binding sites corresponding to their function as heterodimers. However, whereas Clock and Cycle share a substantial fraction of their binding sites in heads and in bodies, including the core components of the circadian clock, we were able to identify tissue-specific binding sites that target genes with the corresponding tissue-specific functions. When investigated the sequence features of Clock and Cycle binding, we found that the E-box motif (CACGTG) was significantly enriched in binding sites shared in heads and bodies and that among others, a GATA motif was significantly enriched in body-specific binding-sites. Further experimental validations *in vitro* have identified the TF Serpent (SRP) to be involved together with Clock and Cycle to define body-specific targets (**Figure 9B**). Performing ChIP-seq experiments in further refined tissues might help identifying new specific partner TFs of Clock and Cycle.

References

Abruzzi KC, Rodriguez J, Menet JS, Desrochers J, Zadina A, Luo W, Tkachev S, Rosbash M. 2011. Drosophila CLOCK target gene characterization: implications for circadian tissue-specific gene expression. *Genes Dev* **25**: 2374–2386.

Arnosti DN, Kulkarni MM. 2005. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem* **94**: 890–898.

Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, et al. 2009. Diversity and complexity in DNA recognition by transcription factors. *Science* **324**: 1720–1723.

Banerji J, Rusconi S, Schaffner W. 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**: 299–308.

Barolo S. 2012. Shadow enhancers: frequently asked questions about distributed cis-regulatory information and enhancer redundancy. *Bioessays* **34**: 135–141.

Baylies MK, Bate M. 1996. twist: a myogenic switch in Drosophila. *Science* **272**: 1481–1484.

Bell O, Tiwari VK, Thomä NH, Schübeler D. 2011. Determinants and dynamics of genome accessibility. *Nat Rev Genet* **12**: 554–564.

Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, Bulyk ML. 2006. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* **24**: 1429–1435.

Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB. 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci USA* **99**: 757–762.

Blackwood EM, Kadonaga JT. 1998. Going the distance: a current view of enhancer action. *Science* **281**: 60–63.

Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, 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.

Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M. 2007. Divergence of transcription factor binding sites across related yeast species. *Science* **317**: 815–819.

Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, et al. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**: 947–956.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**: 311–322.

Bradley RK, Li X-Y, Trapnell C, Davidson S, Pachter L, Chu HC, Tonkin LA, Biggin MD, Eisen MB. 2010. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. *PLoS Biol* **8**: e1000343.

Bulger M, Groudine M. 1999. Looping versus linking: toward a model for long-distance gene activation. *Genes Dev* **13**: 2465–2477.

Busser BW, Taher L, Kim Y, Tansey T, Bloom MJ, Ovcharenko I, Michelson AM. 2012. A machine learning approach for identifying novel cell type-specific transcriptional regulators of myogenesis. *PLoS Genet* **8**: e1002531.

Calhoun VC, Stathopoulos A, Levine M. 2002. Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex. *Proc Natl Acad Sci USA* **99**: 9243–9247.

Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, et al. 2009. Unlocking the secrets of the genome. *Nature* **459**: 927–930.

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, et al. 2008. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**: 1106–1117.

Clapier CR, Cairns BR. 2009. The biology of chromatin remodeling complexes. *Annu Rev Biochem* **78**: 273–304.

Davidson EH. 2010. Emerging properties of animal gene regulatory networks. *Nature* **468**: 911–920.

de Wit E, de Laat W. 2012. A decade of 3C technologies: insights into nuclear organization. *Genes Dev* **26**: 11–24.

Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. *Science* **295**: 1306–1311.

Doherty CJ, Kay SA. 2010. Circadian control of global gene expression patterns. *Annu Rev Genet* **44**: 419–444.

Dowell RD. 2010. Transcription factor binding variation in the evolution of gene regulation. *Trends in genetics : TIG*.

Drosophila 12 Genomes Consortium, Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**: 203–218.

ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**: 636–640.

Erwin DH, Davidson EH. 2009. The evolution of hierarchical gene regulatory networks. *Nat Rev Genet* **10**: 141–148.

Frankel N, Davis GK, Vargas D, Wang S, Payre F, Stern DL. 2010. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466**: 490–493.

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, et al. 2009. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**: 58–64.

Gaszner M, Felsenfeld G. 2006. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* **7**: 703–713.

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**: 1775–1787.

Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* **17**: 877–885.

Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. 2010. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res* **20**: 565–577.

Guss KA, Nelson CE, Hudson A, Kraus ME, Carroll SB. 2001. Control of a genetic regulatory network by a selector gene. *Science* **292**: 1164–1167.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, MacIsaac KD, Danford TW, Hannett NM, Tagne J-B, Reynolds DB, Yoo J, et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**: 99–104.

Hardin PE. 2011. Molecular genetic analysis of circadian timekeeping in Drosophila. *Adv Genet* **74**: 141–173.

Hardison RC, Taylor J. 2012. Genomic approaches towards finding cis-regulatory modules in animals. *Nat Rev Genet* **13**: 469–483.

Harrison MM, Li X-Y, Kaplan T, Botchan MR, Eisen MB. 2011. Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. *PLoS Genet* **7**: e1002266.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**: 311–318.

Hoffman BG, Jones SJM. 2009. Genome-wide identification of DNA-protein interactions using chromatin immunoprecipitation coupled with flow cell sequencing. *J Endocrinol* **201**: 1–13.

Hong J-W, Hendrix DA, Levine MS. 2008. Shadow enhancers as a source of evolutionary novelty. *Science* **321**: 1314.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931–945.

Ip YT, Park RE, Kosman D, Yazdanbakhsh K, Levine M. 1992. dorsal-twist interactions establish snail expression in the presumptive mesoderm of the Drosophila embryo. *Genes Dev* **6**: 1518–1530.

Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**: 533–538.

Jiang J, Kosman D, Ip YT, Levine M. 1991. The dorsal morphogen gradient regulates the mesoderm determinant twist in early Drosophila embryos. *Genes Dev* **5**: 1881–1891.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**: 1497–1502.

Junion G, Spivakov M, Girardot C, Braun M, Gustafson EH, Birney E, Furlong EEM. 2012. A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell* **148**: 473–486.

Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS, et al. 2010. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**: 430–435.

Kheradpour P, Stark A, Roy S, Kellis M. 2007. Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Res* **17**: 1919–1931.

Klug SJ, Famulok M. 1994. All you wanted to know about SELEX. *Molecular Biology Reports* **20**: 97–107.

Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, Chan Y-S, Ng H-H, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**: 631–634.

Kvon EZ, Stampfel G, Yáñez-Cuna JO, Dickson BJ, Stark A. 2012. HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. *Genes Dev* **26**: 908–913.

Lee C-K, Shibata Y, Rao B, Strahl BD, Lieb JD. 2004. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat Genet* **36**: 900–905.

Lee D, Karchin R, Beer MA. 2011. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res* **21**: 2167–2180.

Leptin M. 1991. twist and snail as positive and negative regulators during Drosophila mesoderm development. *Genes Dev* **5**: 1568–1576.

Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12**: 1725–1735.

Levine M, Davidson EH. 2005. Gene regulatory networks for development. *Proc Natl Acad Sci USA* **102**: 4936–4942.

Li X-Y, MacArthur S, Bourgon R, Nix D, Pollard DA, Iyer VN, Hechmer A, Simirenko L, Stapleton M, Luengo Hendriks CL, et al. 2008. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol* **6**: e27.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289–293.

Lin YC, Jhunjhunwala S, Benner C, Heinz S, Welinder E, Mansson R, Sigvardsson M, Hagman J, Espinoza CA, Dutkowski J, et al. 2010. A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. *Nat Immunol* **11**: 635–643.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**: 476–482.

Lupien M, Eeckhoute J, Meyer CA, Wang Q, Zhang Y, Li W, Carroll JS, Liu XS, Brown M. 2008. FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132**: 958–970.

MacArthur S, Li X-Y, Li J, Brown JB, Chu HC, Zeng L, Grondona BP, Hechmer A, Simirenko L, Keränen SVE, et al. 2009. 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.

Malik S, Roeder RG. 2010. The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation. *Nat Rev Genet* **11**: 761–772.

Markstein M, Markstein P, Markstein V, Levine MS. 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *Proc Natl Acad Sci USA* **99**: 763–768.

Markstein M, Zinzen R, Markstein P, Yee K-P, Erives A, Stathopoulos A, Levine M. 2004. A regulatory code for neurogenic gene expression in the Drosophila embryo. *Development* **131**: 2387–2394.

Maston GA, Landt SG, Snyder M, Green MR. 2012. Characterization of enhancer function from genome-wide analyses. *Annu Rev Genomics Hum Genet* **13**: 29–57.

Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, et al. 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**: 374–378.

Meireles-Filho ACA, Stark A. 2009. Comparative genomics of gene regulation-conservation and divergence of cis-regulatory information. *Curr Opin Genet Dev* **19**: 565–570.

Metzker ML. 2010. Sequencing technologies - the next generation. *Nat Rev Genet* **11**: 31–46.

Mikkelsen TS, Xu Z, Zhang X, Wang L, Gimble JM, Lander ES, Rosen ED. 2010. Comparative Epigenomic Analysis of Murine and Human Adipogenesis. *Cell* **143**: 156–169.

modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Nègre N, Eaton ML, Landolin JM, Bristow CA, Ma L, et al. 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**: 1787–1797.

Mouse ENCODE Consortium, Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, et al. 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol* **13**: 418.

Mullen AC, Orlando DA, Newman JJ, Lovén J, Kumar RM, Bilodeau S, Reddy J, Guenther MG, DeKoter RP, Young RA. 2011. Master transcription factors determine cell-type-specific responses to TGF-β signaling. *Cell* **147**: 565–576.

Narlikar L, Sakabe NJ, Blanski AA, Arimura FE, Westlund JM, Nobrega MA, Ovcharenko I. 2010. Genome-wide discovery of human heart enhancers. *Genome Res* **20**: 381–392.

Nègre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R, et al. 2011. A cis-regulatory map of the Drosophila genome. *Nature* **471**: 527–531.

Nien C-Y, Liang H-L, Butcher S, Sun Y, Fu S, Gocha T, Kirov N, Manak JR, Rushlow C. 2011. Temporal coordination of gene networks by Zelda in the early Drosophila embryo. *PLoS Genet* **7**: e1002339.

Ong C-T, Corces VG. 2011. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet* **12**: 283–293.

Ozdemir A, Fisher-Aylor KI, Pepke S, Samanta M, Dunipace L, McCue K, Zeng L, Ogawa N, Wold BJ, Stathopoulos A. 2011. High resolution mapping of Twist to DNA in Drosophila embryos: Efficient functional analysis and evolutionary conservation. *Genome Res* **21**: 566–577.

Palii CG, Perez-Iratxeta C, Yao Z, Cao Y, Dai F, Davison J, Atkins H, Allan D, Dilworth FJ, Gentleman R, et al. 2011. Differential genomic targeting of the transcription factor TAL1 in alternate haematopoietic lineages. *EMBO J* **30**: 494–509.

Pan DJ, Huang JD, Courey AJ. 1991. Functional analysis of the Drosophila twist promoter reveals a dorsal-binding ventral activator region. *Genes Dev* **5**: 1892–1901.

Panne D. 2008. The enhanceosome. *Curr Opin Struct Biol* **18**: 236–242.

Park P. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*.

Pennacchio LA, Visel A. 2010. Limits of sequence and functional conservation. *Nature genetics*, July.

Pepke S, Wold B, Mortazavi A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* **6**: S22–32.

Phillips JE, Corces VG. 2009. CTCF: master weaver of the genome. *Cell* **137**: 1194–1211.

Plautz JD, Kaneko M, Hall JC, Kay SA. 1997. Independent photoreceptive circadian clocks throughout Drosophila. *Science* **278**: 1632–1635.

Pray LA. 2008. Eukaryotic Genome Complexity. *Nature Education* **1(1)**.

Ptashne M, Gann A. 1997. Transcriptional activation by recruitment. *Nature* **386**: 569–577.

Rabinovich A, Jin VX, Rabinovich R, Xu X, Farnham PJ. 2008. E2F in vivo binding specificity: comparison of consensus versus nonconsensus binding sites. *Genome Res* **18**: 1763–1777.

Rando OJ. 2007. Global patterns of histone modifications. *Curr Opin Genet Dev* **17**: 94–99.

Reeves GT, Stathopoulos A. 2009. Graded dorsal and differential gene regulation in the Drosophila embryo. *Cold Spring Harb Perspect Biol* **1**: a000836.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290**: 2306–2309.

Rey G, Cesbron F, Rougemont J, Reinke H, Brunner M, Naef F. 2011. Genome-wide and phase-specific DNA-binding rhythms of BMAL1 control circadian output functions in mouse liver. *PLoS Biol* **9**: e1000595.

Rhee HS, Pugh BF. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**: 1408–1419.

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**: 651–657.

Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**: D91–4.

Sandmann T, Girardot C, Brehme M, Tongprasit W, Stolc V, Furlong EEM. 2007. A core transcriptional network for early mesoderm development in Drosophila melanogaster. *Genes Dev* **21**: 436–449.

Sandmann T, Jensen LJ, Jakobsen JS, Karzynski MM, Eichenlaub MP, Bork P, Furlong EEM. 2006. A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Dev Cell* **10**: 797–807.

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**: 1036–1040.

Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18**: 6097–6100.

Schroeder MD, Pearce M, Fak J, Fan H, Unnerstall U, Emberly E, Rajewsky N, Siggia ED, Gaul U. 2004. Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biol* **2**: E271.

Smale ST, Kadonaga JT. 2003. The RNA polymerase II core promoter. *Annu Rev Biochem* **72**: 449–479.

Spitz F, Furlong EEM. 2012. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**: 613–626.

Stanojevic D, Small S, Levine M. 1991. Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo. *Science* **254**: 1385–1387.

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, et al. 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* **450**: 219–232.

Stormo GD, Zhao Y. 2010. Determining the specificity of protein-DNA interactions. *Nat Rev Genet* **11**: 751–760.

Su J, Teichmann SA, Down TA. 2010. Assessing computational methods of cis-regulatory module prediction. *PLoS Comput Biol* **6**: e1001020.

Swanson CI, Evans NC, Barolo S. 2010. Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. *Dev Cell* **18**: 359–370.

Szymanski P, Levine M. 1995. Multiple modes of dorsal-bHLH transcriptional synergy in the Drosophila embryo. *EMBO J* **14**: 2229–2238.

Thanos D, Maniatis T. 1995. Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell* **83**: 1091–1100.

Trompouki E, Bowman TV, Lawton LN, Fan ZP, Wu D-C, DiBiase A, Martin CS, Cech JN, Sessa AK, Leblanc JL, et al. 2011. Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration. *Cell* **147**: 577–589.

Valenzuela L, Kamakaka RT. 2006. Chromatin insulators. *Annu Rev Genet* **40**: 107–138.

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**: 854–858.

Wei G-H, Badis G, Berger MF, Kivioja T, Palin K, Enge M, Bonke M, Jolma A, Varjosalo M, Gehrke AR, et al. 2010. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J* **29**: 2147–2160.

Wilbanks EG, Facciotti MT. 2010. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE* **5**: e11471.

Wilczyński B, Furlong EEM. 2010. Dynamic CRM occupancy reflects a temporal map of developmental progression. *Mol Syst Biol* **6**: 383.

Yáñez-Cuna JO, Dinh HQ, Kvon EZ, Shlyueva D, Stark A. 2012a. Uncovering cis-regulatory sequence requirements for context-specific transcription factor binding. *Genome Res*.

Yáñez-Cuna JO, Kvon EZ, Stark A. 2012b. Deciphering the transcriptional cis-regulatory code. *Trends in genetics : TIG*.

Zaret KS, Carroll JS. 2011. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev* **25**: 2227–2241.

Zeitlinger J, Simon I, Harbison CT, Hannett NM, Volkert TL, Fink GR, Young RA. 2003. Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell* **113**: 395–404.

Zeitlinger J, Zinzen RP, Stark A, Kellis M, Zhang H, Young RA, Levine M. 2007. 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.

Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EEM. 2009. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* **462**: 65–70.

Zinzen RP, Senger K, Levine M, Papatsenko D. 2006. Computational models for neurogenic gene expression in the Drosophila embryo. *Curr Biol* **16**: 1358–1365.

Publications & Manuscripts

Publication A

High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species

He Q*, **Bardet AF***, Patton B, Purvis J, Johnston J, Paulson A, Gogol M, Stark A, Zeitlinger J. **Nat Genet.** 2011 May;43(5):414-20.

* contributed equally

Q.H. performed the ChIP experiments and library preparation, and J.J., A.P., M.G. and J.Z. established the ChIP-Seq pipeline. B.P. and J.P. raised the different *Drosophila* species, harvested the embryos and staged them, **A.F.B.** and A.S. analyzed the data, and Q.H., **A.F.B.**, A.S. and J.Z. wrote the manuscript.

nature genetics

High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species

Qiye He[1,4], Anaïs F Bardet[2,4], Brianne Patton[1], Jennifer Purvis[1], Jeff Johnston[1], Ariel Paulson[1], Madelaine Gogol[1], Alexander Stark[2] & Julia Zeitlinger[1,3]

© 2011 Nature America, Inc. All rights reserved.

The binding of some transcription factors has been shown to diverge substantially between closely related species. Here we show that the binding of the developmental transcription factor Twist is highly conserved across six *Drosophila* species, revealing strong functional constraints at its enhancers. Conserved binding correlates with sequence motifs for Twist and its partners, permitting the *de novo* discovery of their combinatorial binding. It also includes over 10,000 low-occupancy sites near the detection limit, which tend to mark enhancers of later developmental stages. These results suggest that developmental enhancers can be highly evolutionarily constrained, presumably because of their complex combinatorial nature.

Conservation of functional genomic elements during evolution by selection against fitness-impairing mutations is a fundamental concept in biology. However, the conservation of *cis*-regulatory elements that drive developmental gene expression has remained puzzling. On one hand, transcription factor binding patterns can differ substantially between closely related species[1,2], suggesting high turnover of *cis*-regulatory elements and regulatory rewiring[3]. On the other hand, regulatory relationships that specify certain cell types and organs can be maintained over large evolutionary distances[4]. Furthermore, *cis*-regulatory elements that control development are often complex, making it unlikely that they frequently arise *de novo* from nonfunctional sequence by random mutations. In this study, we investigated the binding pattern of a developmental transcription factor during embryogenesis across six *Drosophila* species and found that it is highly conserved. This not only indicates that developmental gene regulation can be highly constrained during evolution but also provides a unique opportunity to analyze where such constraints occur at the level of gene structure and *cis*-regulatory sequence composition.

We systematically compared the binding landscapes of the basic helix-loop-helix transcription factor Twist during mesoderm formation across six *Drosophila* species. The evolutionary distances between these species, as measured by substitutions per neutral site, are comparable to the distances between human and primates, human and mouse, and human and chicken (*Drosophila melanogaster*, *Drosophila simulans*, *Drosophila yakuba*, *Drosophila erecta*, *Drosophila ananassae* and *Drosophila pseudoobscura*)[5,6]. Twist is not only a master regulator for mesoderm development[7] that has been well characterized by developmental genetics and genomics studies[8–12], but it is also structurally and functionally conserved[13,14] (**Supplementary Fig. 1**), and polyclonal antibodies raised against *D. melanogaster* Twist[10,15] cross react with Twist orthologs of the other *Drosophila* species and reveal conserved mesodermal expression (**Supplementary Fig. 2**). Because transcription factor binding can differ between different developmental stages[16,17], we used stage-matched embryos that encompassed mesoderm formation (2–4 h after egg laying in *D. melanogaster*) for each species (**Supplementary Table 1**) and performed chromatin immunoprecipitation (ChiP) followed by deep sequencing (ChIP-Seq) on two independent biological replicates per species with an Illumina Genome Analyzer 2 using genomic input (whole cell extract (WCE)) as a control (**Fig. 1a**, **Supplementary Table 2** and **Supplementary Fig. 3**). Because of the high quality of the genomic sequence and annotation, we performed a *D. melanogaster*–centric analysis by mapping the sequence reads to each species' reference genome and translating them directly to the genome coordinates of *D. melanogaster* for further analysis (**Fig. 1a** and **Supplementary Tables 3**–**6**). Using a false discovery rate (FDR) of 0.1%, we obtained 3,488 peaks in *D. melanogaster* (**Supplementary Table 7**) which are in good agreement with Twist binding sites from previous ChIP-chip studies (**Supplementary Fig. 4**).

RESULTS

Twist binding is highly conserved across species

Our results show that the binding landscape of Twist is very similar across all six *Drosophila* species. For example, the Twist binding peaks at the known Twist-dependent enhancer of the *tin* locus are nearly identical in each species (see **Fig. 1b** and **Supplementary Fig. 5**

---

[1]Stowers Institute for Medical Research, Kansas City, Missouri, USA. [2]Research Institute of Molecular Pathology (IMP), Vienna, Austria. [3]Department of Pathology, University of Kansas Medical School, Kansas City, Missouri, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to A.S. (stark@imp.ac.at) or J.Z. (jbz@stowers.org).

Received 20 September 2010; accepted 21 March 2011; published online 10 April 2011; doi:10.1038/ng.808

© 2011 Nature America, Inc. All rights reserved.

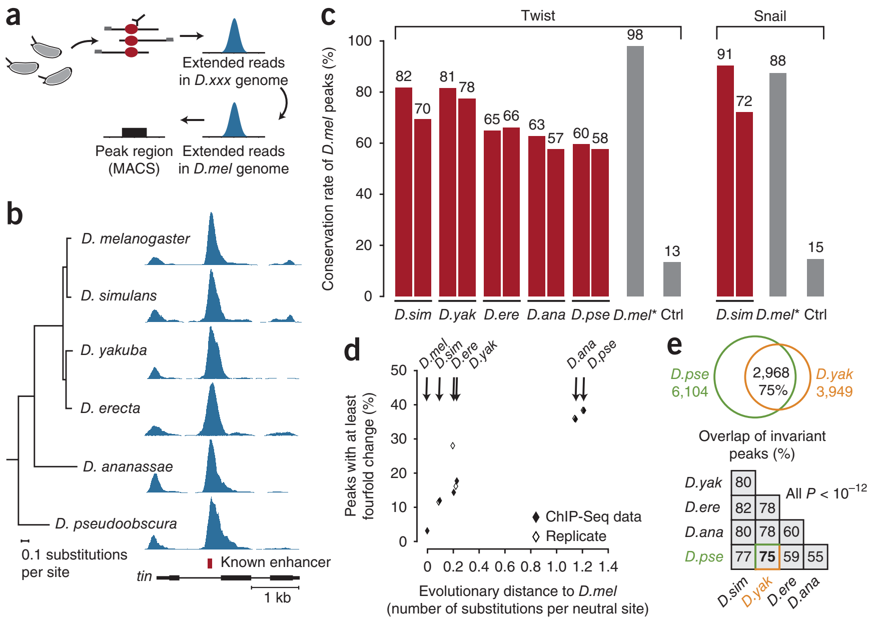


Figure 1 Evolutionary constraints on Twist binding across six *Drosophila* species. (**a**) Overview of the comparative ChIP-Seq pipeline. We directly translated the genomic coordinates of matched reads to *D. melanogaster* for peak calling and analysis (see **Supplementary Tables 3–6** for alternatives). (**b**) Twist binding at the *tin* enhancer[52] is highly similar across six *Drosophila* species. (**c**) Conservation of *D. melanogaster* Twist (left) and Snail (right) binding sites across *Drosophila* species (red; two independent biological replicates per species) compared to a biological replicate in *D. melanogaster* (*) and a control that assessed the background conservation rate by offsetting all *D. melanogaster* peaks by 20 kb (gray). Note that conservation levels varied with the ChIP enrichments; for example, conservation levels are lower than expected for *D. erecta*. (**d**,**e**) Quantitative changes of Twist binding increase with the evolutionary distance. (**d**) The number of Twist binding peaks with ≥fourfold changes in height (normalized read density) increased approximately linearly with the phylogenetic distance ($y = 0.24x + 0.09$; $R^2 = 0.86$). Percentages are based on 8,796 peaks called independently in at least one ChIP experiment. Note that one *D. erecta* replicate is an outlier because of lower ChIP enrichments. (**e**) Invariant peaks are consistent between species comparisons. Seventy-five percent (2,968 of 3,949) of the invariant peaks (≤twofold change) between *D. melanogaster* and *D. pseudoobscura* are also invariant between *D. melanogaster* and *D. yakuba*, which corresponds to a highly significant overlap ($P = 10^{-26}$). The overlaps of invariant peaks were also highly significant between all other species pairs; numbers indicate percentage of overlap (with binomial P values all $\leq 4 \times 10^{-13}$). *D.xxx,* any non-*melanogaster Drosophila* species: *D.mel*, *D. melanogaster*; *D.sim*, *D. simulans*; *D.yak*, *D. yakuba*; *D.ere*, *D. erecta*; *D.ana*, *D. ananassae*; *D.pse*, *D. pseudoobscura.*

for an extended view). We will refer to binding that is shared across species as binding conservation, independent of sequence conservation. At the genome-wide level, we found that the majority of the 3,488 binding peaks in *D. melanogaster* are conserved: more than 80% were bound in *D. simulans* and *D. yakuba* and more than 60% were bound in the other species, including *D. pseudoobscura*, at an evolutionary distance comparable to human with chicken[6] (**Fig. 1c**). Peaks called in the other species showed a similar conservation in *D. melanogaster* (inverse analysis; **Supplementary Fig. 6**), and clustering of the binding data across species recapitulated the established phylogenetic tree, suggesting that the ChIP-Seq data reflect evolutionary events (**Supplementary Fig. 7**). As conservation estimates are threshold dependent, we confirmed that they remain high with different threshold values and using a threshold-independent comparison of the entire Twist binding landscape (**Supplementary Tables 8**,**9**, **Supplementary Fig. 8** and see below). We also show that they are in agreement with the range of conservation estimates derived from the presence of Twist motifs across species[5,18] (**Supplementary Fig. 9** and below). We also confirmed our conservation estimates between *D. melanogaster* and *D. simulans* by performing ChIP-Seq experiments for an additional factor, Snail, which binds to almost identical genomic regions as Twist[11] (**Fig. 1c** and **Supplementary Fig. 10**). Furthermore, our findings are consistent with the high conservation reported for six developmental transcription factors between *D. melanogaster* and *D. yakuba*[19] (**Supplementary Table 10**). In summary, our results show high conservation rates for Twist, with at least ~50% conservation between *D. melanogaster* and *D. pseudoobscura.*

Finally, we assessed whether binding peaks are also evolutionarily constrained at the quantitative level. For this, we identified peaks in each species independently and compared the height of each peak with that of the corresponding peak in *D. melanogaster* (similar to a previous study[19]; **Supplementary Table 11** and **Supplementary Figs. 11**,**12**). The number of peaks that changed at least fourfold increased approximately linearly with the evolutionary distance to *D. melanogaster*, with approximately 2.4% per 0.1 substitutions per neutral site (coefficient of determination, $R^2 = 0.86$; **Fig. 1d**). This suggests that binding divergence may follow a molecular clock, with ~6% of binding sites changing in occupancy levels by more than fourfold every ten million years[20]. Peaks that are invariant (≤twofold change) strongly overlapped between all species comparisons (**Fig. 1e**). These peaks are predominantly located near regulatory genes such as transcription factors ($P = 2 \times 10^{-30}$), whereas variable peaks are not ($P = 0.24$) (**Supplementary Table 12**). This not only argues that binding peaks are highly conserved but also that their level of occupancy is evolutionarily constrained.

Functional binding sites are preferentially conserved

Thirty-four percent of all *D. melanogaster* binding peaks are shared among all six species and thus form a core set of Twist developmental enhancers in *Drosophila*. To assess the functional importance of these deeply conserved binding events, we assigned the peaks to neighboring genes, taking into account the genomic location of regulatory insulators[21]. Conserved peaks showed a clear enrichment near genes that are downregulated in *twist* mutant embryos[10]: for example, ~50% of all peaks that are assigned to genes downregulated in *twist* mutant embryos[10] are deeply conserved, whereas peaks assigned to genes that do not change in the mutant are conserved below average (**Fig. 2a**). Conservation of binding is even higher near genes in Gene Ontology categories related to the developmental role of Twist (up to 71%; **Fig. 2b** and **Supplementary Table 13**) and at known Twist-regulated developmental enhancers (73%; **Supplementary Table 14**), as well as for the highest binding peaks (70%; **Supplementary Fig. 13**), which are thought to be functionally more important[22]. These results show that important binding sites of Twist are maintained over large evolutionary distances.

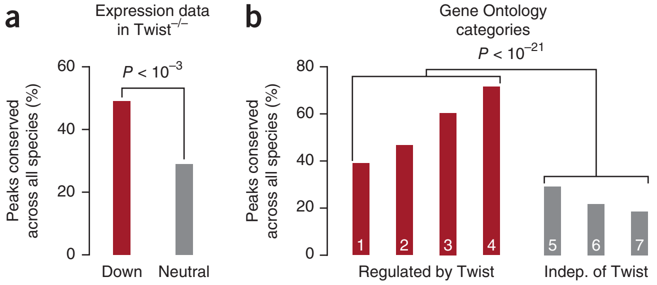


Figure 2 High conservation of functional Twist binding across six *Drosophila* species. (**a**) Preferential conservation of peaks near genes that are downregulated at least twofold in *twist* mutant embryos (red) compared to control genes that do not change (gray; data from a previous study[10]); the fraction of *D. melanogaster* peaks that are conserved across all six species was significantly different, with binomial $P < 10^{-3}$. (**b**) Preferential conservation of peaks near genes in Gene Ontology categories associated with Twist function (red; (1) dorsoventral axis specification, (2) gastrulation, (3) mesodermal cell fate determination, (4) muscle fiber development) or Gene Ontology categories not related to Twist function (gray; (5) carbohydrate metabolic process, (6) amino acid metabolic process, (7) mRNA metabolic process). The difference between all genes in the combined functional versus Twist-independent categories was significant, with a binomial $P < 10^{-21}$. For an overview of all Gene Ontology categories, see **Supplementary Table 13**.

© 2011 Nature America, Inc. All rights reserved.

Enhancers have been reported to lie upstream or downstream of genes, in introns or even overlapping with coding exons[23,24], and, indeed, Twist binds to different genomic regions (**Supplementary Fig. 14**). However, despite the high overall sequence conservation of protein coding regions, Twist binding in coding exons is poorly conserved (**Fig. 3a**). We also observed low levels of conservation of binding in 3′ untranslated regions (UTRs), wheras conservation rates of peaks in promoters, 5′ UTRs, intronic regions and intergenic regions were uniformly high (**Fig. 3a**). The deep conservation of binding peaks is independent of the distance to the nearest transcription start sites, even at distances of over 20 kb (**Fig. 3b**), suggesting evolutionary selection of distant enhancers, which are commonly found in flies and vertebrates[23,24].

Clustered binding sites are preferentially conserved

Specific developmental expression patterns of genes are often regulated by multiple enhancers, which can act redundantly[25] or can each be essential for fitness[26–28]. Twist target genes frequently have multiple Twist binding peaks[10,11], and some of the enhancers at these peaks can direct similar expression patterns[11,29]. Whether regulation by multiple enhancers is generally more likely to be redundant or essential has remained unclear.

Clustered peaks that were assigned to the same gene are significantly more often deeply conserved than isolated peaks that are uniquely assigned to a gene (54% compared to 34%, $P < 3 \times 10^{-4}$). The preferential conservation of clustered peaks was also apparent when we classified peaks based on the distance to their closest neighbor, independent of their gene assignments. The conservation rate was highest for peak-to-peak distances less than 5 kb and decreased gradually with greater distances (**Fig. 3c**). This suggests that clustered binding sites and 'shadow enhancers'[29] (**Supplementary Table 14** and **Supplementary Fig. 15**) may be functionally important, perhaps because the enhancers' activities are not fully redundant due to different input factors[30], or to ensure robustness and precision of expression patterns[26,27].

Twist binding correlates with transcription factor motifs

Comprehensive comparative ChIP-Seq data provide a unique opportunity to study the sequence basis of conserved binding. We found that Twist binding peaks that are shared across all species have similar average sequence conservation compared to binding peaks that are specific to *D. melanogaster* as assessed by phastCons scores or by the number of fully conserved nucleotides (**Fig. 4a**). In contrast, 37% of all Twist sequence motifs found in shared peaks, but only 9% in *D. melanogaster*–specific peaks, are present in all species ($P < 10^{-17}$; **Fig. 4a**). The correlation between peak and motif presence was similar when motif movements were allowed (46% compared to 13%; $P = 4 \times 10^{-21}$), held for pairwise comparisons between species and species-specific losses of peaks (**Supplementary Fig. 16**) and allowed for the *de novo* discovery of the Twist motif (**Supplementary Table 15**). Overall, ~24% of Twist peaks had a binary (presence or absence) binding pattern across the six species that exactly matched that of the Twist sequence motifs (eightfold more than expected if peaks and motifs occurred independently, $P < 2 \times 10^{-58}$). For all divergent peaks, we determined the types of mutations that caused the species-specific Twist motif loss and found that the majority of motif losses were caused by point mutations, followed by deletions and insertions (**Fig. 4b** and **Supplementary Fig. 17**). Finally, changes in the quality of the Twist motif across species are also significantly correlated with quantitative changes in Twist binding (**Fig. 4c** and **Supplementary Fig. 18**). In summary, the conservation of binding peaks correlates with the conservation of motifs, rather than overall enhancer sequence, suggesting specific selection against motif-disrupting point mutations and insertions or deletions.

However, a substantial fraction of Twist binding losses cannot be attributed to the loss of the Twist motif. For example, 14% of the Twist

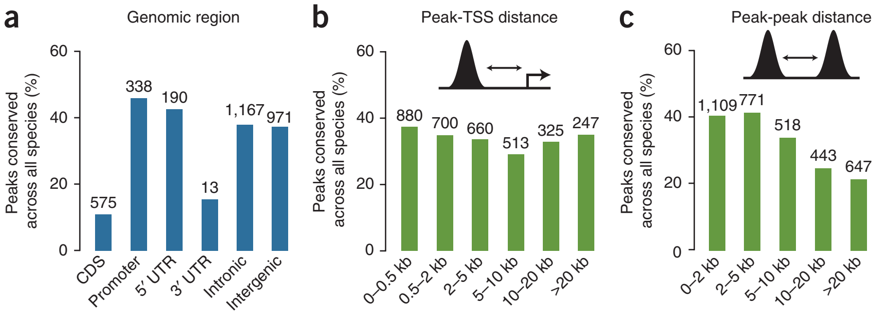


Figure 3 Preferential conservation of clustered binding peaks. (**a**) Conservation rates (percent of *D. melanogaster* peaks that are conserved across all six species) for peaks in different genomic regions. CDS, coding-sequence; UTR, untranslated region. The number of *D. melanogaster* peaks in each region is shown on top. (**b**) Conservation rates are as in **a** but are dependent on the distances of the peak summits to the nearest gene transcription start sites (TSS). (**c**) Conservation rates are as in **a** but dependent on the distances between two neighboring peak summits (independent of the conservation of either peak). Isolated peaks are significantly less highly conserved ($P < 10^{-45}$ compared to the leftmost bin). Note that the 0–0.5-kb bin is not populated because of the width of the peaks.

© 2011 Nature America, Inc. All rights reserved.

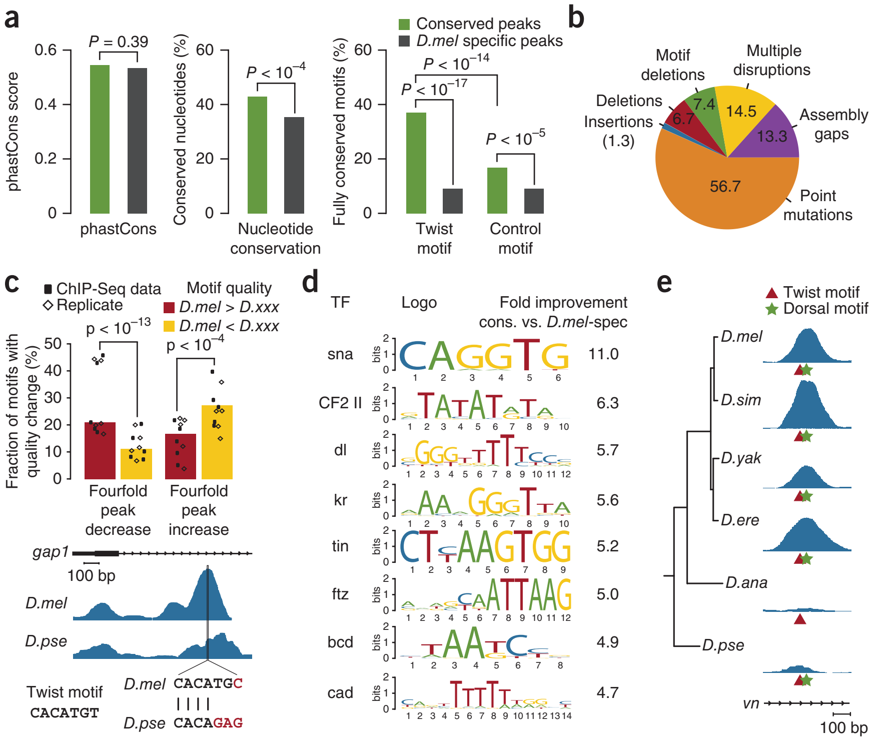


Figure 4 Twist binding depends on the sequence motifs of Twist and its partner transcription factors. (**a**) Twist binding peaks shared across all species (conserved) or *D. melanogaster*–specific (*D.mel*-spec) peaks have similar overall phastCons scores (left; Wilcoxon $P = 0.39$) and nucleotide conservation (middle; Wilcoxon $P < 10^{-4}$) but different conservation rates for the Twist motif (hypergeometric $P < 10^{-17}$). (**b**) Sequence changes (in percent) that cause motif and peak loss (**Supplementary Fig. 17**). (**c**) At top, quantitative changes of peak height correlate with Twist motif quality (MAST score). Peaks that are ≥fourfold lower in a second species compared to *D. melanogaster* (left) contain more motifs with lower scores in that species than in *D. melanogaster* ($P < 10^{-13}$ for all). The reverse is true for peaks that are ≥fourfold higher (right; $P < 10^{-4}$ for all except the *D. erecta* 2 replicate, which had $P = 0.43$). Circles and diamonds represent the fraction of changed motifs in each pairwise comparison, and bar heights indicate the median values. At bottom, an example of a quantitative change of Twist binding at the *gap1* gene locus that correlates with Twist motif quality (red, mismatches to the consensus motif). (**d**) Motifs of Twist partner transcription factors correlate with Twist binding. Shown are the top non-Twist motifs[6] that are conserved in fully conserved Twist peaks but not *D. melanogaster*–specific peaks (fold improvement between motif conservation rates). (**e**) Loss of Twist binding in *D. ananassae* despite a conserved Twist motif correlates with the loss of a Dorsal motif in the *vn* (*vein*) intron (read density scales are identical across species).

peaks that were lost in at least one species nevertheless contained a conserved Twist motif. We therefore explored whether Twist binding could be disrupted through the loss of the motif for a partner transcription factor. We identified several motifs for transcription factors other than Twist that are significantly more highly conserved in conserved Twist peaks than in species-specific Twist peaks or the average genome (**Fig. 4d** and **Supplementary Table 16**). These factors include Snail (11.1-fold increased conservation) and Dorsal (5.7-fold increased conservation), both of which are known to function together with Twist[11,12,31]. As shown in **Figure 4e**, a *D. ananassae*–specific disruption of a Dorsal motif at the *vn* (*vein*) enhancer, a known Twist enhancer in *D. melanogaster*[32], might explain the divergence of Twist binding despite a conserved Twist motif. Indeed, genome-wide, Snail and Dorsal motifs are able to explain 19% of the losses of Twist binding that occur despite a conserved Twist motif, and the top ten identified motifs explain 49% of the losses. Transcription factors for these motifs[6] include factors involved in mesoderm development (tinman and CF2II), segmentation (bicoid and caudal) or both (Kruppel and fushi tarazu). Both muscle and segmentation transcription factors frequently co-occupy Twist enhancers and may cooperate with Twist in gene regulation[11,33,34]. These results suggest that cross-species ChIP-Seq analysis can be used to identify combinatorial relationships between transcription factors, similar to ChIP-Seq analyses in yeast and human haplotypes[35,36].

Twist has widespread access to inactive enhancers

Interestingly, we noticed that sites with low Twist occupancy also tend to be conserved across species: the Twist binding landscape is very similar overall across the entire genome (with Pearson correlation coefficients above 0.45 between *D. melanogaster* and all five comparative species) (**Fig. 5a** and **Supplementary Table 2**). This similarity persists when we excluded the 3,488 identified peaks (corresponding to 2.1% of the genome; **Fig. 5a**), indicating that the similarity in binding extends to low-occupancy peaks near the detection limit. To test this directly, we identified a large number of putative binding sites in the intergenic and intronic regions of *D. melanogaster* by lowering the threshold for peak identification ($P < 10^{-5}$, FDR = 22%; note that this will include many false positives). Many thousand *D. melanogaster* peaks near the detection limit, but not their randomly placed counterparts, had enrichments of ChIP over WCE in the other species (**Fig. 5b**). Finally, these low-occupancy sites are enriched for Twist motifs, which are specifically conserved above background (**Fig. 5c**), suggesting that many of these motifs have been selectively maintained throughout evolution and are likely functionally important.

One possibility is that the function of low-occupancy peaks is to increase the local concentration of the transcription factor near high-occupancy binding peaks. Indeed, low-occupancy sites are more often conserved if they occur near high-occupancy sites (**Fig. 5d**). For example, the *E(spl)* (*Enhancer of split*) cluster on chromosome 3R has several conserved low-occupancy peaks in the vicinity of high-occupancy peaks, and they are shared between the species (**Supplementary Fig. 19**). This finding is also consistent with the preferential conservation of clustered enhancers.

Another possibility is that low-occupancy peaks correspond to sites that are more strongly bound at different developmental stages. ChIP-chip studies at different time points during embryonic mesoderm and muscle development have shown that Twist and other transcription factors change their binding sites over time[10,30,37]. Indeed, low-occupancy peaks from our ChIP-Seq study at 2–4 h after egg laying strongly overlap with regions determined to be bound by Twist at later time points[10,30] (**Fig. 5e**). In contrast, sites with high Twist occupancy showed a decreasing overlap with sites bound at later time points (**Fig. 5e**). These opposing trends argue that sites with low occupancy are likely to be bound in different developmental

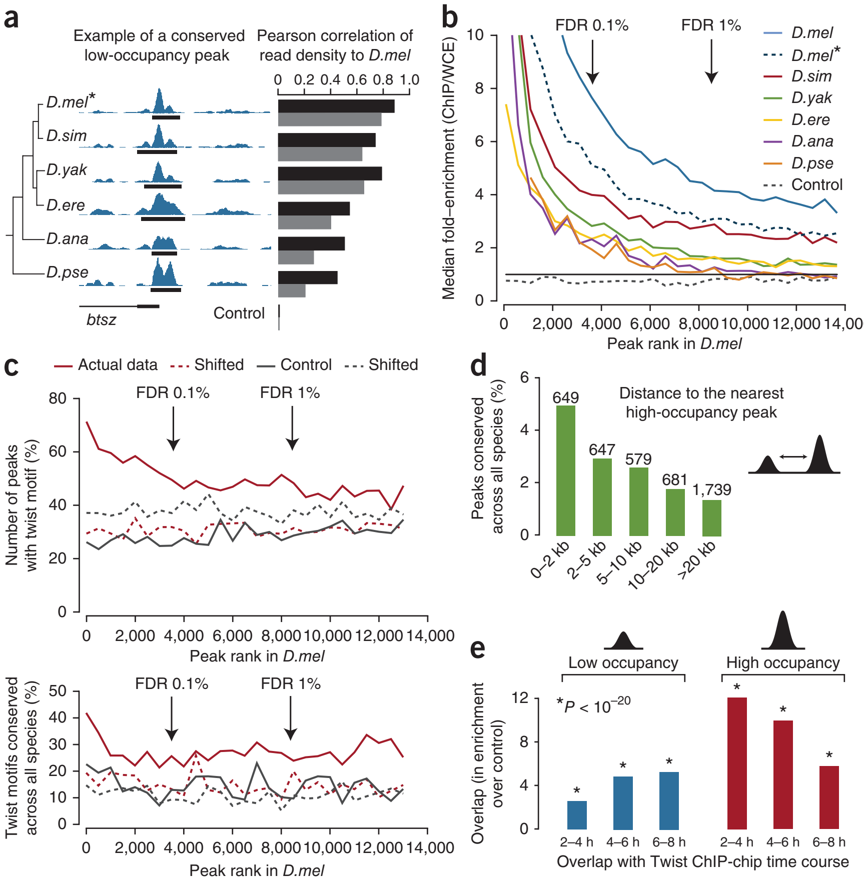


Figure 5 Conservation of low-occupancy peaks. (**a**) The similarity of Twist binding (blue) extends beyond the peak (black bar) at the *btsz* (*bitesize*) locus (left). Read densities are similar across species (black) even when excluding peak regions (gray). *D. mel**, biological replicate; control, independence is simulated by reverting the read density. (**b**) Several thousand peaks are detectably bound across species. Shown is the fold enrichment (ChIP/WCE) at the position aligned to the *D. melanogaster* peak summit (median of 500 peaks per bin; *D.mel**, biological replicate; control, *D. melanogaster* peaks shifted by 20 kb). (**c**) Several thousand peaks contain Twist motifs that are specifically conserved. Top, at any rank, peaks (solid red) contained more Twist motifs than expected given shifted peaks (dashed red), randomized motifs (solid gray; all $P < 10^{-144}$ for high-occupancy peaks and $P < 10^{-57}$ for low-occupancy peaks). Bottom, Twist motifs in peaks at any rank (bins of 500) were more often conserved across all species than expected given the average conservation of the peak region (randomized motifs) or the genome-wide conservation of the Twist motif (shifted; all $P < 10^{-3}$ for high-occupancy peaks and $P < 10^{-5}$ for low-occupancy peaks). (**d**) The conservation rate of low-occupancy peaks dropped with increasing distance to the nearest high-occupancy peak ($P < 10^{-8}$ between the outermost bins). (**e**) Low-occupancy peaks overlapped increasingly with ChIP-chip data[30] from later time points (top), whereas high-occupancy peaks showed the opposite trend. To account for different numbers of ChIP-chip peaks at different time points, we calculated the enrichments against shifted peak locations (all $P < 10^{-20}$).

© 2011 Nature America, Inc. All rights reserved.

contexts when different partner transcription factors are present or when changes in chromatin allow increased access. This implies that many low-occupancy binding sites observed in ChIP experiments might constitute functional sites under different conditions rather than promiscuous nonfunctional binding.

DISCUSSION

We find that the binding landscape of Twist is highly conserved across six *Drosophila* species, with preferential conservation of peaks near relevant Twist target genes. This is consistent with the high binding conservation for six transcription factors between *D. melanogaster* and *D. yakuba*[19]. However, it stands in contrast to recent reports in yeast[2], in adult vertebrate liver[1,3,38], in human and mouse embryonic stem cells[39], and during human and mouse adipogenesis[40], in which the binding of transcription factors has diverged substantially. Thus, there appears to be a wide range by which transcription factor binding is conserved, presumably reflecting different evolutionary dynamics.

On one hand, *cis*-regulatory changes and binding divergence may be an important driving force for adaptive evolution (for example, see ref. 41). Indeed, rapid evolutionary adaptation to different ecological niches has been suggested to be the primary reason for the high turnover of binding in yeast[2]. In flies and vertebrates, species-specific binding might also alter gene expression and contribute to adaptation and speciation. In vertebrates, for example, transposable elements seem to substantially contribute to species-specific binding[39,40], consistent with the hypothesis that transposons could effectively contribute to regulatory changes during evolution[42].

On the other hand, strong evolutionary constraints are expected for deeply conserved developmental processes. For example, the mesoderm formation studied here is thought to be shared between all bilateria, with transcription factors such as Twist being ancestrally involved in mesoderm development[13,14,43]. Furthermore, individual Twist-dependent enhancers can be conserved from *Drosophila* to insects as distant as *Tribolium*[44], presumably because complex developmental enhancers with specific combinations of transcription factor binding sites cannot easily evolve *de novo*. In contrast, transcriptional regulation in differentiated cell types and organs may work through enhancers with simpler inputs[45] and may even be maintained independent of enhancers by switching components of the core transcription machinery[46]. This might allow binding sites to evolve more easily *de novo* and reduce the evolutionary constraints on enhancers of differentiated tissues.

Some of the differences in conservation of binding between flies and vertebrates might also be due to the smaller population size of vertebrates, which could increase evolutionary drift. Furthermore, vertebrates have much larger genomes, which may allow for more nonfunctional or selectively neutral binding[47] as well as binding site movements. Consistent with this hypothesis, comparative ChIP-Seq studies in vertebrates reported an order of magnitude higher in the numbers of binding sites[1,39,40] compared to *Drosophila*. Notably, the absolute number of conserved sites appears to be roughly constant, perhaps indicating a similar number of core regulatory connections that need to be maintained.

Taken together, our results suggest that the high conservation of binding that we found for Twist will apply to complex developmental enhancers in all metazoans, including vertebrates. Vertebrate developmental enhancers are among the most highly conserved sequences[48], and many vertebrate *cis*-regulatory motifs and their target genes can be identified based on conservation[49–51]. In addition, the liver transcription factor binding sites that are deeply conserved are near genes involved in liver organogenesis[1], and binding sites in embryonic stem cells and adipocytes are substantially more highly conserved near functional targets[39,40].

The high conservation of Twist binding also provides a unique opportunity to globally identify functionally important features of transcription factor binding and enhancer organization. Specifically, we have shown that clustered peaks or 'shadow enhancers'[29] tend to be more conserved than isolated peaks, suggesting that gene regulation by multiple enhancers may be essential for fitness rather than being redundant. Furthermore, Twist binding correlates with sequence motifs for Twist and partner transcription factors, which suggests widespread cooperative binding and may explain why developmental transcription factors can bind and regulate different developmental programs[16,17,30]. This notion is consistent with thousands of low-occupancy Twist sites that we identified and for which we provided evidence that many are functional in different developmental conditions. This suggests that transcription factors such as Twist can access and bind to inactive enhancers at low levels. Whether low-occupancy binding is because of the lack of partner transcription factors at this condition, properties of chromatin or both remains to be shown. We predict that low-occupancy binding and strong evolutionary conservation will be relevant to developmental gene regulation in complex multicellular organisms in general.

METHODS

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

Accession code. The data from this study are deposited in ArrayExpress under the accession code E-MTAB-376.

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

ACKNOWLEDGMENTS

We would like to thank E. Furlong (European Molecular Biology Laboratory (EMBL)) and M. Levine (University of California Berkeley) for generously providing Twist antibodies; B. Dickson, W. Lugmayr, R. Revilla (IMP), C. Schlötterer (University of Veterinary Medicine Vienna), M. Levine (University of California Berkeley) and R. Krumlauf (Stowers Institute for Medical Research) for discussions, help and advice. A.F.B. was supported by the Austrian Ministry for Science and Research through the Genome Research in Austria (GEN-AU) Bioinformatics Integration Network III. J.Z. is a Pew scholar. A.S. is supported by an European Research Council (ERC) Starting Grant from the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 242922.

AUTHOR CONTRIBUTIONS

Q.H. performed the ChIP experiments and library preparation, and J.J., A.P., M.G. and J.Z. established the ChIP-Seq pipeline. B.P. and J.P. raised the different *Drosophila* species, harvested the embryos and staged them, A.F.B. and A.S. analyzed the data, and Q.H., A.F.B., A.S. and J.Z. 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://npg.nature.com/reprintsandpermissions/.

1. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036–1040 (2010).
2. Borneman, A.R. *et al.* Divergence of transcription factor binding sites across related yeast species. *Science* **317**, 815–819 (2007).
3. Odom, D.T. *et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet.* **39**, 730–732 (2007).
4. Davidson, E.H. & Erwin, D.H. Gene regulatory networks and the evolution of animal body plans. *Science* **311**, 796–800 (2006).
5. Clark, A.G. *et al.* Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**, 203–218 (2007).
6. Stark, A. *et al.* Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450**, 219–232 (2007).
7. Baylies, M.K. & Bate, M. Twist: a myogenic switch in *Drosophila. Science* **272**, 1481–1484 (1996).
8. 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).
9. Leptin, M. Twist and snail as positive and negative regulators during *Drosophila* mesoderm development. *Genes Dev.* **5**, 1568–1576 (1991).
10. Sandmann, T. *et al.* A core transcriptional network for early mesoderm development in *Drosophila melanogaster. Genes Dev.* **21**, 436–449 (2007).
11. 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).
12. Ip, Y.T., Park, R.E., Kosman, D., Yazdanbakhsh, K. & Levine, M. Dorsal-Twist interactions establish snail expression in the presumptive mesoderm of the *Drosophila* embryo. *Genes Dev.* **6**, 1518–1530 (1992).
13. Castanon, I. & Baylies, M.K. A Twist in fate: evolutionary comparison of Twist structure and function. *Gene* **287**, 11–22 (2002).
14. Technau, U. & Scholz, C.B. Origin and evolution of endoderm and mesoderm. *Int. J. Dev. Biol.* **47**, 531–539 (2003).
15. Zinzen, R.P., Senger, K., Levine, M. & Papatsenko, D. Computational models for neurogenic gene expression in the *Drosophila* embryo. *Curr. Biol.* **16**, 1358–1365 (2006).
16. Zeitlinger, J. *et al.* Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell* **113**, 395–404 (2003).
17. 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).
18. Richards, S. *et al.* Comparative genome sequencing of *Drosophila* pseudoobscura: chromosomal, gene, and *cis*-element evolution. *Genome Res.* **15**, 1–18 (2005).
19. Bradley, R.K. *et al.* Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related *Drosophila* species. *PLoS Biol.* **8**, e1000343 (2010).
20. Tamura, K., Subramanian, S. & Kumar, S. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol. Biol. Evol.* **21**, 36–44 (2004).
21. Nègre, N. *et al.* A comprehensive map of insulator elements for the *Drosophila* genome. *PLoS Genet.* **6**, e1000814 (2010).
22. 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).
23. Visel, A., Rubin, E.M. & Pennacchio, L.A. Genomic views of distant-acting enhancers. *Nature* **461**, 199–205 (2009).
24. Bulger, M. & Groudine, M. Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev. Biol.* **339**, 250–257 (2010).
25. Degenhardt, K.R. *et al.* Distinct enhancers at the *Pax3* locus can function redundantly to regulate neural tube and neural crest expressions. *Dev. Biol.* **339**, 519–527 (2010).
26. Perry, M.W., Boettiger, A.N., Bothma, J.P. & Levine, M. Shadow enhancers foster robustness of *Drosophila* gastrulation. *Curr. Biol.* **20**, 1562–1567 (2010).
27. Frankel, N. *et al.* Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466**, 490–493 (2010).
28. O'Meara, M.M. *et al. Cis*-regulatory mutations in the *Caenorhabditis elegans* homeobox gene locus cog-1 affect neuronal development. *Genetics* **181**, 1679–1686 (2009).
29. Hong, J.W., Hendrix, D.A. & Levine, M.S. Shadow enhancers as a source of evolutionary novelty. *Science* **321**, 1314 (2008).
30. 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).
31. García-Zaragoza, E., Mas, J.A., Vivar, J., Arredondo, J.J. & Cervera, M. CF2 activity and enhancer integration are required for proper muscle gene expression in *Drosophila. Mech. Dev.* **125**, 617–630 (2008).
32. Markstein, M. *et al.* A regulatory code for neurogenic gene expression in the *Drosophila* embryo. *Development* **131**, 2387–2394 (2004).
33. Li, X.Y. *et al.* Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol.* **6**, e27 (2008).
34. Qian, S., Capovilla, M. & Pirrotta, V. Molecular mechanisms of pattern formation by the BRE enhancer of the *Ubx* gene. *EMBO J.* **12**, 3865–3877 (1993).

© 2011 Nature America, Inc. All rights reserved.

35. Kasowski, M. *et al.* Variation in transcription factor binding among humans. *Science* **328**, 232–235 (2010).
36. Zheng, W., Zhao, H., Mancera, E., Steinmetz, L.M. & Snyder, M. Genetic analysis of variation in transcription factor binding in yeast. *Nature* **464**, 1187–1191 (2010).
37. Wilczynski, B. & Furlong, E.E. Dynamic CRM occupancy reflects a temporal map of developmental progression. *Mol. Syst. Biol.* **6**, 383 (2010).
38. Conboy, C.M. *et al.* Cell cycle genes are the evolutionarily conserved targets of the E2F4 transcription factor. *PLoS ONE* **2**, e1061 (2007).
39. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42**, 631–634 (2010).
40. Mikkelsen, T.S. *et al.* Comparative epigenomic analysis of murine and human adipogenesis. *Cell* **143**, 156–169 (2010).
41. Carroll, S.B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008).
42. Davidson, E.H. & Britten, R.J. Regulation of gene expression: possible role of repetitive sequences. *Science* **204**, 1052–1059 (1979).
43. Arendt, D. The evolution of cell types in animals: emerging principles from molecular studies. *Nat. Rev. Genet.* **9**, 868–882 (2008).
44. Cande, J., Goltsev, Y. & Levine, M.S. Conservation of enhancer location in divergent insects. *Proc. Natl. Acad. Sci. USA* **106**, 14414–14419 (2009).
45. Flames, N. & Hobert, O. Gene regulatory logic of dopamine neuron differentiation. *Nature* **458**, 885–889 (2009).
46. Deato, M.D. & Tjian, R. Switching of the core transcription machinery during myogenesis. *Genes Dev.* **21**, 2137–2149 (2007).
47. 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).
48. Pennacchio, L.A. *et al. In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).
49. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).
50. Ettwiller, L. *et al.* The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. *Genome Biol.* **6**, R104 (2005).
51. Del Bene, F. *et al. In vivo* validation of a computationally predicted conserved Ath5 target gene set. *PLoS Genet.* **3**, 1661–1671 (2007).
52. 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).

© 2011 Nature America, Inc. All rights reserved.

© 2011 Nature America, Inc. All rights reserved.

ONLINE METHODS

Stock maintenance and embryo collection. All six *Drosophila* species were raised at 25 °C and 60% humidity. The collection window for the different *Drosophila* species was 2–4 h after egg laying (AEL) except for *D. simulans,* which was 1–3 h AEL, and for *D. pseudoobscura*, which was 3–5 h AEL. The time windows were derived from literature[53] and empirical optimization to obtain high signal-to-noise ratios in the ChIP-Seq experiments and to obtain the majority of embryos within Bownes stage 5–8 (**Supplementary Table 1**). For staging, formaldehyde-fixed embryos were rehydrated, stained with DAPI and imaged using the MosaiX tool from Zeiss. Data from independent collections were pooled, analyzed and compared to the Bownes stages of *D. melanogaster*.

Embryo immunostaings. Embryos from the six species were collected and fixed according to published protocols[54]. Embryos were incubated with Twist antibodies[15] (1:200 dilution) at 4 °C overnight. Incubation with the secondary antibody (AlexaFluor555-conjugated guinea pig antibody from Invitrogen A-21435 at 1:500 dilution) was performed for ~3 h at room temperature (22 °C). To visualize nuclei, embryos were stained with DAPI (1 µg/ml) for 15 min. Embryos were mounted in 70% glycerol and observed with an LSM 5 Pascal confocal microscope (Zeiss).

Chromatin immunoprecipitation (ChIP) and library preparation for Solexa sequencing. ChIP was performed using modified protocols from the Zeitlinger lab[11] and the Furlong lab[17]. Briefly, embryos were cross linked in 1.8% formaldehyde, chromatin was sonicated to an average size of ~500 bp (whole cell extract (WCE)), and 300 µl WCE from ~120 mg embryos was incubated with protein A–conjugated Dynabeads (Invitrogen 100-02D), coated with antibodies against *D. melanogaster* Snail (Millipore MAB5494) or Twist; the experiments in *D. melanogaster* were performed with antibodies against full-length Twist (a generous gift from M. Levine[15]), and antibodies raised against the C terminus of Twist (a generous gift from E. Furlong[10]) were used for the other species because they yielded higher enrichment ratios. For controls, 50 µl WCE was used. The level of Twist enrichment was monitored by real-time PCR (StepOnePlus, Applied Biosystems) using primers for *brk* and *rho* enhancers in *D. melanogaster,* for *tup* and *Dscam* in the non–*D. melanogaster* species and primers for a non-genic region (NonG) as negative control[11].

Preparations of DNA libraries for single-end sequencing were done according to instructions from Illumina with 36 cycles of extension. Up to 20 ng ChIP DNA, or 100 ng WCE DNA, was used in each preparation.

Reads processing. We mapped the reads to each genome reference (dm3 (not chrU, chrUextra), droSim1, droYak2, droEre2, droAna3 and dp4) from UCSC[55] using Eland from the Illumina Solexa data processing pipeline with default parameters. We translated all non *D. melanogaster* reads into *D. melanogaster* coordinates using the liftOver program[55] (using default parameters, except *minmatch* = 0.7). We extended each read to the average length of the genomic fragments for each experiment and calculated a normalized read count and fold enrichment (ChIP versus WCE) for each genomic position.

Peak calling and conservation. We defined peak regions in each experiment from the ChIP reads and the corresponding WCE reads using MACS v1.3.2 (ref. 56) with the maximum possible *mfold* parameter. For *D. melanogaster,* we focused on the 3,488 high-occupancy peaks with an FDR below 0.1% and defined those with an FDR greater than 1% as low-occupancy peaks. For each peak, we determined a summit as the position with the highest read count and calculated its fold enrichment. We called a *D. melanogaster* peak conserved if its region overlapped with the peak summit in another experiment. We controlled for the background binding conservation by determining the conservation of *D. melanogaster* peaks against themselves offset by 20 kb. Independently, we calculated the Pearson correlation coefficient between the read counts of two experiments (excluding runs of zeros, which would artificially increase the correlation).

Quantitative changes. We called peaks independently in each ChIP experiment (MACS $P = 10^{-22}$, which corresponds to an FDR = 0.1% in *D. melanogaster*) and combined overlapping peaks. We scored each peak in each ChIP experiment by the highest read count in a 151-bp window around the summit. We excluded peaks with a read count of zero in any experiment and normalized the remaining 8,796 peaks using quantile normalization. We defined peaks as invariant ('no change') if their heights changed less than twofold and variable (decreasing or increasing) if their heights changed more then fourfold.

Functional analyses. We assigned each peak to its closest gene transcription start site (FlyBase r5.11) but not across insulators[21] (CTCF peaks and the intersection of CP190 and BEAF peaks). We calculated the conservation rate of peaks assigned to genes in different genomic regions (FlyBase r5.11), functional categories from Gene Ontology[57] (GO:0009950; GO:0007369; GO:0007500; GO:0048747 (Twist-related) versus GO:0005975; GO:0006520; GO:0016071 (unrelated to Twist)) and from expression data in Twist mutants[10] as Twist targets (twofold downregulated versus neutral (less than 0.00098-fold change)).

Motif and sequence analysis. We searched for motif occurrences of known motifs[6] including the Twist motif CACATGT[15] in an area 151 bp (average genomic fragment length) around each peak summit. We used a Position Weight Matrix cutoff of 4×10^{-3}, corresponding to one allowed mismatch for the Twist motif such that 59% of peaks have at least one motif. As controls, we used shuffled columns of PWMs as done previously[58] and peak coordinates shifted by 20 kb.

For each identified motif occurrence or peak region, we extracted the orthologously aligned sequence for each of the five species from multiple genome alignment[6] and evaluated the sequence conservation of motif occurrences and peak regions by perfect conservation, point mutations, deletions (gap in *D. melanogaster*), insertions (gaps in the other species), deletions of entire motifs and alignment gaps (absence of nucleotides in a ± 20-bp window around the motif). All changes were summed across all five species and normalized to the region length in *D. melanogaster*. When assessing whether a motif fully explains the phylogenetic distribution of a peak, we considered only the motif occurrence closest to the peak summit and required that the presence or absence patterns of peak and motif across species matched exactly.

For the analysis of motif quality in quantitative changes in Twist binding, an unbiased pairwise symmetrical comparison between species was performed. For each peak, we searched for motif matches independently in both species, scored each match and the aligned sequence by MAST and counted how often each species' sequence scored more highly for peaks that decreased or increased.

Overlap with peaks at later stages. We counted the overlap of high- and low-occupancy peaks with ChIP-chip Twist binding regions identified during only one time point (2–4 h, 4–6 h or 6–8 h; excluding peaks in CDS regions) from a previous study[30] and calculated the enrichment over controls shifted by 20 kb.

53. Kim, J., Kerr, J.Q. & Min, G.S. Molecular heterochrony in the early development of *Drosophila*. *Proc. Natl. Acad. Sci. USA* **97**, 212–216 (2000).
54. Rothwell, W.F. & Sullivan, W. *Drosophila Protocols.* **141** (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 2000).
55. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
56. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
57. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
58. Kheradpour, P., Stark, A., Roy, S. & Kellis, M. Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Res.* **17**, 1919–1931 (2007).

doi:10.1038/ng.808

High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species

Qiye He[*1], Anaïs F. Bardet[*2], Brianne Patton[1], Jennifer Purvis[1], Jeff Johnston[1], Ariel Paulson[1], Madelaine Gogol[1], Alexander Stark[2]#, Julia Zeitlinger[1,3]#

[1]Stowers Institute for Medical Research, Kansas City;
[2]Research Institute of Molecular Pathology (IMP), Vienna
[3]Department of Pathology, University of Kansas Medical School, Kansas City

[*]These authors contributed equally
[#]Corresponding authors (stark@imp.ac.at & jbz@stowers.org)

Nature Genetics: doi.10.1038/ng.808

Supplementary Tables

Table S1: Staging of the embryos from the different *Drosophila* species 1
Table S2: Pearson correlation of biological replicates 2
Table S3: Sensitivity of translating reads between genomes 3
Table S4: Number of reads for each sample 5
Table S5: Sensitivity of translating peaks between genomes 7
Table S6: Comparison of peak-calling protocols for comparative ChIP 8
Table S7: Conservation of *D. melanosgaster* Twist binding peaks in six *Drosophila* species 9
Table S8: MACS independent assessment of peak conservation 10
Table S9: Conservation estimates using different cutoffs and measures 11
Table S10: Re-analysis of the binding conservation rates from Bradley et al.[7] 13
Table S11: Quantitative changes of Twist binding peaks in all six *Drosophila* species .. 14
Table S12: GO analysis of invariant vs. variant peaks 14
Table S13: GO analysis of genes near Twist binding peaks 14
Table S14: Conservation at known Twist enhancers 15
Table S15: Most conserved 7mers 16
Table S16: Conservation of other motifs 17

Supplementary Figures

Figure S1: Protein sequence alignment of Twist and Snail 19
Figure S2: Cross-reactivity of the Twist antibody in immunostainings 22
Figure S3: Venn diagrams of peak overlap between biological replicates 24
Figure S4: Twist binding overlap with other studies 25
Figure S5: Twist binding at the *tinman* locus (extended view) 26
Figure S6: Conservation of peaks from a species-specific view 28
Figure S7: Clustering of Twist binding data recapitulates the phylogenetic tree 29
Figure S8: Choice of a sensitive cutoff to adjust for the winner's curse 30
Figure S9: Expected conservation rates based on sequence conservation....................... 32
Figure S10: Snail binding conservation 34
Figure S11: Raw data of quantitative changes 36
Figure S12: Quantitative changes 37
Figure S13: Conservation rate of *D. melanogaster* peaks at different ranks 38
Figure S14: Distribution of Twist binding peaks relative to annotated genes 39
Figure S15: Example of a conserved shadow enhancer.. 40
Figure S16: Pairwise conservation rate of Twist motifs 41
Figure S17: Sequence basis for binding peak losses 42
Figure S18: Correlation between Twist binding levels and motif quality 44
Figure S19: Conserved low-occupancy peaks at the *Enhancer-of-Split* locus 45

References.. 46

Table S1: Staging of the embryos from the different *Drosophila* species

Species	Collection window (hours post egg laying)	Percentage of embryos in Bownes stage 5-8
D. melanogaster	2-4	88%
D. simulans	1-3	43%*
D. yakuba	2-4	95%
D. erecta	2-4	60%*
D. pseudoobscura	3-5	97%
D. ananassae	2-4	92%

Developmental stages represented in embryos used for Twist ChIP. We staged the embryos and determined the percentage of embryos that were within Brownes stage 5-8, which encompass mesoderm formation.

* The 1-3 hour collections of *D. simulans* and *D. erecta* embryos have ~40% embryos in Bownes stage 2 and 3. However, the level of Twist message is very low at these stages and should not give rise to a false signal.

Table S2: Pearson correlation of biological replicates

Species	Correlation between biol. replicates
D. melanogaster	0.88
D. simulans	0.79
D. yakuba	0.92
D. erecta	0.54
D. ananassae	0.92
D. pseudoobscura	0.93

Pearson correlation coefficient between the genome-wide read densities of biological replicates in each species. To confirm the high conservation estimates with biological replicates, we performed ChIP experiments from independent embryo collections. The resulting genome-wide profiles of read densities are highly similar based on Pearson correlation coefficients. Note that the lower ChIP enrichments for *D. simulans* 2 and *D. erecta* 2 explain the lower correlation with the second replicate.

Table S3: Sensitivity of translating reads between genomes

		D. melanogaster	*D .simulans*	*D. yakuba*	*D. erecta*	*D. ananassae*	*D. pseudoobscura*
Eland	Reads mapped (% of genome)	90.57	78.64	81.94	83.25	68.36	94.24
LiftOver	Reads translated to D.mel (% of mapped reads)	NA	97.94	95.52	92.64	75.66	73.85
	D.mel genome covered by translated reads (% of entire genome)	NA	73.85	80.64	83.80	72.88	61.32

Assessment of the sensitivity of mapping and translating hypothetical ChIP-Seq reads to *D. melanogaster* using liftOver. Comparative analyses require that the data from different species are compared in a common reference. We used *Drosophila melanogaster* as the common reference genome in our analysis because the genome annotation and assembly quality is the best of all analyzed species (e.g. *D. erecta, D. ananassae, D. pseudoobscura* are assembled as contigs only). To determine the maximal fraction of sequencing reads that could be lost at the liftOver step, we first determined the number of all possible 36 nucleotide long reads (created from the reference genomes in one-nucleotide steps) that could be mapped back uniquely to the respective genome using Eland (first row). The results are less than 100% because Eland maps only uniquely occurring reads. The lowest percentage of mappable reads is found for *D. ananassae*, which has the largest genome size (231 Mb). We then measured the fraction of mapped reads of these species that could be unambiguously translated to the *D. melanogaster* genome using liftOver (second row) and calculated the fraction of the *D. melanogaster* genome that is covered by these reads (third row). The numbers are very high, e.g. the translated reads of *D. erecta* cover up to 83.8% of the *D. melanogaster* genome, 90.57%

of which is mappable using Eland (see first row). Even in the most distant *D. pseudoobscura* genome, at least 73.85% of the mappable reads can be translated to *D. melanogaster*, thus at most 26.15% of reads may be lost at the liftOver step. While we found that the translation of genome coordinates between species works generally well, the fraction of reads that can be translated between genomes is lower between more distant species. Since we are using translated reads to assess the conservation of *D. melanogaster* peaks called in the *D. melanogaster* genome, the loss caused by translating reads argues that our conservation estimates are conservative, i.e. if read translation was perfect and without loss, our conservation estimates would by definition be higher. In summary, we believe that errors in our estimates due to mapping/translation problems are minimal and would lead to an underestimation of the conservation rates.

Table S4: Number of reads for each sample

Species		Raw	Not linker	Mapped	D.mel coords	Not U	Kept %
D. melanogaster 1	TWI	6924965	6924031	5543574	NA	5434925	79.33
	WCE	8594417	8581024	6747496	NA	6521548	77.36
	SNA	4291074	4285633	2737771	NA	2688733	62.66
	WCE	4562107	4556454	3364128	NA	3298782	72.31
D. melanogaster 2	TWI	14200801	14129968	10716228	NA	10581462	74.51
	WCE	11221132	11188236	7848693	NA	7480236	68.51
	SNA	3953419	3945886	2719433	NA	2674348	67.65
	WCE	3522437	3518412	2575068	NA	2522295	71.61
D. simulans 1	TWI	10160113	8001743	5617543	5326639	5243769	51.85
	SNA	27934713	27934187	15511276	14630075	14412159	51.59
	WCE	6767252	6733014	4438638	4061284	3920495	58.50
D. simulans 2	TWI	9610826	9607567	6880848	6525575	6474027	67.36
	SNA	8816467	8798729	5425635	5075268	4972998	56.41
	WCE	5145042	5143262	3839420	3610384	3579424	69.57
D. yakuba 1	TWI	8784288	8708042	6494404	5909963	5811634	66.81
	WCE	12567553	12410303	9824301	8774190	8549213	69.06
D. yakuba 2	TWI	6137729	5685916	4042366	3654447	3623654	59.04
	WCE	16354071	16310705	12647066	11193723	11059560	67.63
D. erecta 1	TWI	10100897	9970701	8150334	7406680	7273869	72.73
	WCE	13173366	13172855	11698663	10438238	10158082	78.49
D. erecta 2	TWI	11146650	6476678	4275147	3913429	3879556	34.80
	WCE	15603994	15501342	13101697	11955738	11879663	76.13
D. ananassae 1	TWI	12888726	12820719	9090403	6526596	6325165	50.10
	WCE	14973513	14968247	11058518	8084193	7681642	53.25
D. ananassae 2	TWI	15181955	14954445	9805781	6880465	6797084	44.77
	WCE	15414157	15405212	10703889	7526340	7391218	47.95
D. pseudoobscura 1	TWI	14574340	14259587	8698705	5187271	5093405	35.38
	WCE	13970438	13706932	9553915	5824033	5642428	41.27
D. pseudoobscura 2	TWI	14627544	13085530	9431110	5893564	5863416	40.08
	WCE	14168724	14011345	11164353	7312950	7271167	51.32

Overview of analysis steps for each sample and the corresponding read counts. TWI: Twist ChIP; SNA: Snail ChIP; WCE: Whole Cell Extract (used as control for corresponding ChIP sample). The raw reads from sequencing are screened against the Solexa linker sequences, mapped to the respective genome and translated to *D.melanogaster* coordinates. All reads that are not located on unassembled sequences (chrU, chrUextra) are used for the analysis. Their total number and their percentage to the original read numbers are shown in the last two columns.

Table S5: Sensitivity of translating peaks between genomes

Species	Number of peaks in the original species	Number of peaks translated to D.mel coordinates	Peaks translated (%)
D. melanogaster 1	3488	NA	NA
D. melanogaster 2	3583	NA	NA
D. simulans 1	2174	2136	98.25
D. simulans 2	773	760	98.32
D. yakuba 1	3158	3103	98.26
D. yakuba 2	2468	2425	98.26
D. erecta 1	2848	2789	97.93
D. erecta 2	721	710	98.47
D. ananassae 1	4427	3621	81.79
D. ananassae 2	3352	2731	81.47
D.pseudoobscura 1	4709	2970	63.07
D.pseudoobscura 2	3797	2420	63.73

Number of peaks called in each reference genome (first column) that can be translated to *D. melanogaster* coordinates using liftOver (third column) and what fraction they represent (fourth column). As an alternative approach to the one used in **Table S4**, peaks were here called in individual species' genomes first (with a p-value cutoff of $10^{-21.8}$ corresponding to an FDR of 0.1% in *D. melanogaster*) before being translated into *D. melanogaster* genome coordinates. The number of peaks that could be translated to the *D. melanogaster* genome was very high for species closely related to *D. melanogaster* but dropped to 63% for remote species, consistent with the phylogeny and our results above (**Table S4**). Note that the lower ChIP enrichments for *D. simulans* 2 and *D. erecta* 2 mean that fewer peaks can be called at the stringent cutoff. Although we did not use this approach (see **Table S6** for further analysis and explanations), this table shows that our estimates of the fraction of *D. melanogaster* peaks that are conserved in the other species are conservative, especially for remote species where some peak loss occurs during peak translation.

Table S6: Comparison of peak-calling protocols for comparative ChIP

Species	Number of peaks from translated peaks (a)	Number of peaks from translated reads (b)	Overlap (%)
D. simulans 1	2136	2164	92.84
D. simulans 2	760	785	93.42
D. yakuba 1	3103	2911	88.72
D. yakuba 2	2425	2335	90.72
D. erecta 1	2789	2629	88.10
D. erecta 2	710	663	84.93
D. ananassae 1	3621	3723	79.84
D. ananassae 2	2731	2840	78.94
D. pseudoobscura 1	2970	3652	81.04
D. pseudoobscura 2	2420	2812	77.07

A comparison of the number of peaks obtained as *D. melanogaster* coordinates from different species using two different protocols. (a) In the first protocol, peaks are called in the original species and the peak coordinates are translated to *D. melanogaster* coordinates (as in **Table S5**). **(b)** In the second protocol, the reads from the ChIP-seq experiment are directly translated to *D. melanogaster* coordinates and peaks are then called in *D. melanogaster.* In both protocols, peaks were called with a fixed p-value cutoff of $10^{-21.8}$ corresponding to an FDR of 0.1% in *D .melanogaster* 1. Both protocols give a very similar number of peaks with strong overlap between 77% and 93%. Neither of the two approaches yields systematically more (or fewer) peaks, arguing that the protocols are equivalent and that the observed differences are mainly due to small variations that make peaks fall above or below the cutoff. We used strategy **b** because first translating the reads to a common reference has the advantage that it allows the comparative analysis of the raw reads (e.g. for Pearson correlations).

Table S7: Conservation of *D. melanosgaster* Twist binding peaks in six *Drosophila* species

Table_S7 provides all 3488 Twist binding peaks identified in *D. melanogaster*, their conservation in other species, and genes that are associated to these peaks

Table S8: MACS independent assessment of peak conservation

Species	IP/WCE > 2		SUM/FLK > 2	
	All	Offset	All	Offset
D. melanogaster 1	100	19	96	20
D. melanogaster 2	99	15	88	20
D. simulans 1	95	15	85	18
D. simulans 2	96	26	78	17
D. yakuba 1	91	8	84	19
D. yakuba 2	94	7	81	17
D. erecta 1	92	8	62	15
D. erecta 2	77	7	81	20
D. ananassae 1	86	13	66	18
D. ananassae 2	89	12	59	18
D. pseudoobscura 1	87	10	68	19
D. pseudoobscura 2	71	5	68	19

Results of a MACS-independent analysis of the conserved Twist binding peaks in the other species. To ensure that our MACS p-value cutoffs do not call peaks conserved without substantial ChIP enrichment, we tested the peak height and shape of all conserved peaks by an alternative method. Genomic regions that orthologously align to the summits of conserved *D. melanogaster* peaks are directly tested for enrichment of IP vs. WCE (% of conserved peaks with enrichment >2 fold) and for their peaky shape (% of conserved peaks with an enrichment >2 fold of peak summit over the inferred flanking regions, which are 75bp long windows located 150bp offset from the peak summit on each side). The majority of all conserved peaks pass both measures in the respective comparative species, and the fraction of peaks is much above the expected background given a control with offset peaks, providing independent evidence that our sensitive cutoffs used to determine peak conservation do not select peaks without substantial ChIP enrichment.

Table S9: Conservation estimates using different cutoffs and measures

Species	P-value 10-5	P-value 10-10	P-value 10-15	Best 3488	Best 5000	Best 6000	4 fold	2 fold	Correlation	Correlation w/o peaks*
D. melanogaster 2	98	94	87	74	84	88	100	92	0.88	0.78
D. simulans 1	82	71	59	58	66	70	94	77	0.74	0.64
D. simulans 2	70	48	32	54	61	64	93	73	0.62	0.55
D. yakuba 1	81	72	64	60	69	73	95	79	0.79	0.65
D. yakuba 2	78	66	57	59	68	72	95	80	0.75	0.57
D. erecta 1	65	56	50	49	56	59	93	70	0.54	0.40
D. erecta 2	66	42	28	50	57	60	95	69	0.59	0.61
D. ananassae 1	63	54	49	42	49	53	91	70	0.51	0.27
D. ananassae 2	57	49	43	42	50	54	91	69	0.48	0.24
D. pseudoobscura 1	60	52	47	41	48	52	91	67	0.45	0.21
D. pseudoobscura 2	58	47	41	40	46	48	90	65	0.46	0.30
All species (w/o replicates)	34	26	20	17	23	26	-	-	-	-
Offset 20 kb	13	7	5	4	5	6	24	8	0.01	0.01

Conservation rates of Twist binding (in percent) when six different cutoffs for identifying peaks in the non-*melanogaster* data are used (columns 2-7). These conservation rates all refer to the fixed number of 3844 peaks in *D. melanogaster,* which were identified by MACS using an FDR of 0.1% (p-value $10^{-21.8}$). In the other species, the conservation rates for p-value cutoffs of 10^{-5}, 10^{-10} and 10^{-15} are shown (columns 2-4), which represent increasingly stringent cutoffs for the ChIP enrichment in the comparative genome. We used a p-value cutoff of 10^{-5} in our main analysis (column 1). We also applied rank cutoffs in the other species (columns 5-7), using the top 3844 peaks (to match the number in *D. melanogaster*; similar to Schmidt et al.[6]), top 5000 or top 6000 peaks. Similar to Bradley et al.[7] who defined divergent peaks by a change of more than 10-fold, we also defined peaks as conserved, if their heights changed less than 4- or 2-

fold (columns 8-9). Note that we normalized the peak heights between species using pairwise quantile normalization. To control for random conservation at different cutoffs, the average conservation rate is shown for peaks that have been offset by 20 kb (last row). In summary, we found a high conservation rate above 40% for all cutoffs in all species.

To compare the similarity of the Twist binding landscapes in a cutoff-free manner, we calculated the Pearson correlation coefficient of the normalized read counts between *D. melanogaster* and each of the other species/experiments, either genome-wide (column 10) or without the 3488 identified peak regions (*, column 11). When doing so, we excluded positions at which the compared experiments had both zero reads, since these would increase the correlation coefficient (not shown). The high correlation confirms the conservation rates above. The correlation drops but remains high without the peak regions, indicating that the peaks contribute positively to the high similarity but that the similarity extends beyond the peaks.

Table S10: Re-analysis of the binding conservation rates from Bradley et al.[7]

ChIP (TF)	Number of peaks in D.mel	Total number of regions with sign. enrichment in D.yak	Conservation rate (%)
BCD	456	484	56
HB1	2441	2837	56
HB2	1952	1562	47
KR1	2754	3155	63
KR2	2647	3610	70
GT	916	2738	85
KNI	104	388	81
CAD	1773	1744	56

Conservation rates between *D. melanogaster* and *D. yakuba* for transcription factors involved in AP-patterning (segmentation) at a similar stage of development (~2h AEL). We re-analyzed the data from Bradley et al.[7] (BCD: Bicoid, HB: Hunchback, KR: Krüppel, GT: Giant, KNI: Knirps and CAD: Caudal) using our approach. Peaks were called in *D. melanogaster* at a p-value 10^{-22} and required a significant enrichment of the corresponding regions in *D. yakuba* (p-value 10^{-5}). Note that the conservation rates (column 3) drop, if the total number of regions with significant enrichment in *D. yakuba* is lower than in *D. melanogaster*. Considering this, the conservation rates for these factors are similarly high to ours, suggesting that the binding of developmental transcription factors is generally highly conserved.

Table S11: Quantitative changes of Twist binding peaks in all six *Drosophila* species

Table_S11 shows all the quantitative changes, as measured by the fold enrichment (IP/control) at the summit of Twist peaks, of the 8796 Twist peaks that were identified in at least one ChIP experiment.

Table S12: GO analysis of invariant vs. variant peaks

In **Table_S12**: m: number of genes near invariant/variable peaks that belong to a given GO category. M: number of genes in a given GO category. n: total number of genes near invariant/variable peaks. N: total number of genes in all GO categories.

Table S13: GO analysis of genes near Twist binding peaks

In **Table_S13**: m: number of genes near Twist peaks that belong to a given GO category. M: number of genes in a given GO category. n: total number of genes near Twist peaks. N: total number of genes in all GO categories.

Table S14: Conservation at known Twist enhancers

Genomic Position			**Gene**	**Peak conserved in…**
chr2L	2456194	2457304	dpp	dsim,dyak,dere,dpse
chr2L	15479779	15480368	sna	dsim,dyak,dere,dpse
chr2L	15485442	15486919	sna-S	dsim,dyak,dere,dana,dpse
chr2L	18874869	18876181	tup	dsim,dyak,dere,dana,dpse
chr2L	20475806	20477901	mir1-1	dsim,dyak,dere,dana
chr2L	20480578	20481741	mir1-2	dsim,dyak,dere,dana,dpse
chr2L	21851517	21853665	tsh	dsim,dyak,dere,dana,dpse
chr2R	3133869	3134085	ady	not detected at FDR 0.1%
chr2R	7681727	7682234	ths	dsim,dyak,dere,dana,dpse
chr2R	8833934	8834294	mdr49	dsim,dyak,dere
chr2R	11775659	11776615	sli	not detected at FDR 0.1%
chr2R	18932428	18933842	twi	dsim,dyak,dere,dana,dpse
chr2R	19875348	19875790	phm	dsim,dere,dpse
chr3L	1461823	1462121	rho	dsim,dyak,dere,dana,dpse
chr3L	5828771	5829267	vn	dsim,dyak,dere,dpse
chr3L	9032265	9033283	doc	dsim,dyak,dere,dana,dpse
chr3L	9797449	9797743	DyakIlp4	not detected at FDR 0.1%
chr3L	15032420	15033848	ind	not detected at FDR 0.1%
chr3R	2580903	2581527	zen	dsim,dyak,dere,dpse
chr3R	8895836	8896466	sim	dsim,dyak,dere,dana,dpse
chr3R	9118955	9119462	wntD	dsim,dyak,dere,dpse
chr3R	10423060	10424098	stumps	not detected at FDR 0.1%
chr3R	11854390	11855212	pnr	dsim,dyak,dere,dana,dpse
chr3R	13875600	13876391	htl	dsim,dyak,dere,dana,dpse
chr3R	17205818	17205999	tin	dsim,dyak,dere,dana,dpse
chr3R	20574722	20575521	tld	dsim,dyak,dere,dana,dpse
chr3R	21861313	21862676	E(spl)	dsim,dyak,dere,dana,dpse
chrX	477514	479251	vnd-V	dsim,dyak,dere,dana,dpse
chrX	479413	481119	vnd-M	dsim,dyak,dere,dana,dpse
chrX	486761	487503	vnd	dsim,dyak,dere,dana,dpse
chrX	7190967	7191464	brk	dsim,dyak,dere,dana,dpse
chrX	7214537	7215702	brk-S	dsim,dyak,dere,dana,dpse
chrX	13574341	13574673	cg12177	dsim,dyak,dere,dana,dpse
chrX	15518731	15519122	sog	dsim,dyak,dere,dana,dpse
chrX	15540627	15541510	sog-S	dsim,dyak,dere,dana,dpse

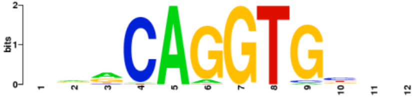
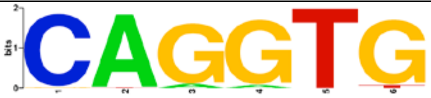
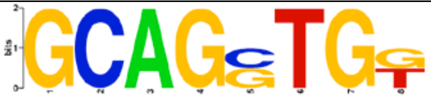
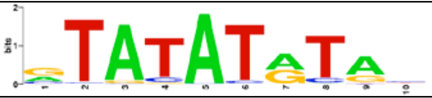
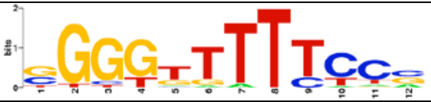
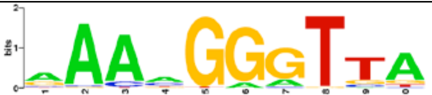
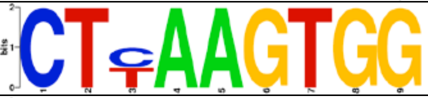
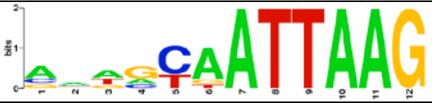
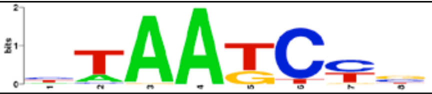
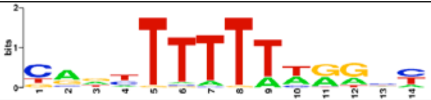
Known Twist enhancers have been manually assembled based on Zeitlinger et al.[3] and subsequent additions based on data from Mike Levine's lab. The majority of known enhancers, including the 'shadow enhancers' identified by Hong et al.[11] (marked with "–S"), are conserved across all species.

Table S15: Most conserved 7mers

7mer	Conservation rate in conserved peaks	Improvement over shifted peaks	Similarity to known motifs	Motif
ACATCTG	50.0	3.8	0.8	Twist
CATATGG	49.3	Inf	0.7	Twist
ACACGTG	45.0	1.8	0.8	Twist
AACATGT	44.8	3.0	0.9	Twist
ACATGTG	43.5	2.6	1.0	Twist
ACAGCTG	40.2	1.8	0.7	Snail
CATGCGC	39.5	1.6	0.7	Twist
CATGTGG	39.1	3.1	0.9	Twist
CATATGC	38.9	Inf	0.7	Twist
CACATGG	38.7	4.6	0.8	Twist

Top 10 most conserved 7mers in fully conserved peaks, their rate of conservation across all species (in %) and the similarity to known motifs from Stark et al.[12]. Nine of the top ten 7mers correspond to the Twist motif, while one is a Snail motif. For example, 50% of all occurrences of the 7mer ACATCTG (first row) in fully conserved binding peaks are conserved across all six species, which is 3.8-fold more than expected given the 7mers conservation in shifted control peaks. The 7mer-to-motif similarity was assessed using the Pearson correlation of the corresponding PWMs as in Stark et al.[12].

Table S16: Conservation of other motifs

Transcription factor	Logo	Fold improvement of conserved vs. species-specific peaks	P-value	Fold improvement of conserved vs. genome	P-value	Percent of peak explained	Percent of peak loss explained	Conservation rate (in conserved and in non-conserved peaks)
esg		**11.1**	**7.7×10^{-5}**	**2.2**	**7.1×10^{-14}**	**16.6**	**12.3**	**66.8** **51.1**
sna		**11.0**	**5.1×10^{-5}**	**2.1**	**3.3×10^{-12}**	**19.4**	**7.7**	**73.4** **57.2**
sc		**6.3**	**8.2×10^{-5}**	**3.0**	**1.0×10^{-20}**	**20.7**	**9.9**	**73.0** **48.4**
CF2 II		**6.3**	**2.6×10^{-4}**	**2.2**	**1.6×10^{-18}**	**13.7**	**17.4**	**51.7** **41.8**
dl		**5.7**	**2.1×10^{-2}**	**2.9**	**9.8×10^{-13}**	**12.2**	**20.4**	**51.9** **32.2**
kr		**5.6**	**1.8×10^{-2}**	**3.2**	**2.5×10^{-14}**	**12.1**	**10.7**	**56.1** **41.4**
tin		**5.2**	**2.9×10^{-2}**	**2.6**	**5.7×10^{-7}**	**13.2**	**5.2**	**66.5** **49.7**
ftz		**5.0**	**2.9×10^{-2}**	**2.1**	**2.0×10^{-7}**	**13.6**	**6.9**	**60.7** **49.6**
bcd		**4.9**	**2.6×10^{-2}**	**3.7**	**4.2×10^{-20}**	**17.4**	**7.1**	**63.9** **44.0**
cad		**4.7**	**4.9×10^{-2}**	**2.9**	**8.1×10^{-12}**	**13.2**	**13.2**	**54.0** **39.7**

Top 10 transcription factor motifs (excluding Twist) that are conserved in conserved Twist-binding peaks. Shown are the motif logos for their known motifs[12], the fold improvement of the conservation rates between fully conserved and *D. melanogaster*-specific binding peaks or the average genome, respectively, and the corresponding hypergeometric P-values, the fraction of peaks that have the identical species distribution as the corresponding motif ("explained"), the fraction of peak losses explained by motif losses in each species, and the motifs' conservation rates in regions of conserved and non-conserved *D. melanogaster* peaks (average over all cases in any species). Note that the motifs for several factors (e.g. esg, sc) are similar to the Snail motif, which could explain their high conservation.

Figure S1: Protein sequence alignment of Twist and Snail

a. Twist

```
D.mel            MMSARSVSPKVLLDISYKPTLPNIMELQNNVIKLIQVEQQAYMQSGYQLQH-QQQHLHSH 59
D.sim            MMSTRSVSPKVLLDISYKPTLPNIMELQNNVIKLIQVEQQAYMQSGYQLQH-QQQHLHAH 59
D.yak            MMSARSVSPKVLLDISYKPTLPNIMELQNNVIKLIQVEQQAYMQSGYQLQHQQQQHLHSH 60
D.ere            MMSARSVSPKVLLDISYKPTLPNIMELQNNVIKLIQVEQQAYMQSGYQLQH-QQQHLHSH 59
D.ana            MMSARSVSPKVLLDISYKPTLPNIMELQHNVIKLIQVEQQAYMQSGY------PQHAIMQ 54
D.pse            MMSTRSVSPKVLLDISYKPTLPNIMELQHNVIKLIQVEQQAYVHSSHY-----IHQSPVH 55
                 ***:************************:*************::*.:       ::   :

D.mel            QHHQQHH----QQQHAQYAPLPSEYAAYGITELEDTDYNIPSNEVLSTSS----NQSAQS 111
D.sim            QHHQQHH----QQQHTQYAPLPSEYAAYGITELEDTDYNIPSNEVLSTSS----NQSAQS 111
D.yak            QQHHQQHQQPQQQQHPQYAPLPSEYAAYGITELEDTDYNIPSNEVLSTSS----NQSAQS 116
D.ere            QHHQQHQQ---QQQHTQYAPLPSEYAAYGITELEDTDYNIPSNEVLSTSS----NQSAQS 112
D.ana            QHQQQQQ----QQQQPQYAPLPSEYAAYGITELEDTDYNIPSNEVLSTSS----NQSAQS 106
D.pse            QHQQQQHQ---QQQNPQYAPLPSEYAAYGITELEDTDYNIPSNEILSTSSSTHSNHSAQS 112
                 *:::*::    ***:.****************************:*****    *:****

D.mel            TSLELNNNNTSSNTNSSGNNPSGFDGQ--ASSGSSWNEHGKRARSSGDYDCQTGGSLVMQ 169
D.sim            ASLELNNNNTSSNTTSSGNNSTRQEG-----------------RSSGDYDCQAGGSLVMQ 154
D.yak            ASLELNNNNTSSNNTSSGNNPNGFDGQ--ASSGSSWNEHGKRARSSGDYDCQTGGSLAMQ 174
D.ere            TSLEMNNNNTSSNNTSSGNNPSGFDGQ--ASSGSSWNEHGKRARSSGDYDCQTGGSLVMQ 170
D.ana            ASLELNNNNTSSN-TSSNNNFDPQNGN--GN-GSAWNEHGKRTRSSGDYDCQTGGSLVMQ 162
D.pse            ASLELNNNHTASNSNGSSNNQNVFDQQSVAGSGSSWNEHGKRARSSSDYDCQSGGTLAMQ 172
                 :***:***:*:** ..*.**    :                  ***.*****:**:*.**

D.mel            PEHKKLIHQQQQQ-----QQQ-HQQQIYVDYLPTTVDEVASAQSCPGVQSTCTSPQSHFD 223
D.sim            PEHKKLIHQQQQQ-----QQQQHQQQIYVDYLPTTVDEVASAQSCPGVQSTCTSPQSHFD 209
D.yak            PEHKKLIHQQQQQ------QQQHQQHIYVDYLPTTVDEVASAQSCPGVQSTCTSPQSHFD 228
D.ere            PEHKKLIHQQQQQPQ---QQQQHQQHIYVDYLPTTVDEVAAAQSCPGVQSTCTSPQSHFD 227
D.ana            PEHKKLIHQQHQQQQ---QHQQQQQHIYVDYLPTTVDEVASAQACPGVQSTCTSPHSHFD 219
D.pse            PDHKKLLHQQQHQQQQQHQQQQQQQQIYVDYLPTTVDEVASAQTCAGPQSTCTSPHSHFE 232
                 *:****:***::*      :* :**:**************:**:*.* *******:***:

D.mel            FPDEELPEHKAQVFLPLYNNQQQQSQQLQQQQP----HQQSHAQMHFQNAYRQSFEGYEP 279
D.sim            FPDEELPEHKAQVFLPLYNNQQQQSQQQQQQQP----HQQSHAQMHFQNAYRQSFDGYEP 265
D.yak            FPDEELPEHKTQVFLPLYSNQQQS----QQQQS----HQQNHAQMHFQNAYRQSFEGYEP 280
D.ere            FPDEELPEHKAQVFLPLYNNQQQ-----SQQQP----HQQNHAQMHFQNAYRQSFESYEP 278
D.ana            FPDEELPEHKTQVFLPLYTNQQQQ---QQQQQPQHQLHQQSQAQMHFQAAYRQSFEGYEP 276
D.pse            FPDEELSEHKAQVFLPLYTNQHQPQQQATHQQQQ---QQPQNPQLHFQNSYRQSFDGYEP 289
                 ******.***:*******.**:*      :**     :* .:.*:*** :*****:.***

D.mel            ANSLNGSAYSSSDRDDMEYARHNALSSVSDL------NGGVMSPACLADDGSAGSLLDGS 333
D.sim            ANSLNGSAYSSSDRDDMEYARHNALSSVSDL------NGGVMSPACLADDGSAGSLLDGS 319
D.yak            ANSLNGSAYSSSDRDDMEYARHNGLSSVSDL------NGGVMSPACLADDGSAGSLLDGS 334
D.ere            ANSLNGSAYSSSDRDDMEYARHNALSSVSDL------NGGVMSPACLADDGSAGSLLDGS 332
D.ana            ANSLNGSAYSSSDRDEMEYARHTALSSVNDL------NGG-MSPACLGDDGSAGSLLDGS 329
D.pse            ANSLNGSAYSSSDRDDMEYVRHTALSSVSDLAAGGGVNGGGMSPACLADDGSSGSLLDGV 349
                 ***************:***.**..****.**      *** ******.****:******

D.mel            DAGGKAFRKPRRRLKRKPSKTEETDEFSNQRVMANVRERQRTQSLNDAFKSLQQIIPTLP 393
D.sim            DAGGKAFRKPRRRLKRKPSKTEETDEFSNQRVMANVRERQRTQSLNDAFKSLQQIIPTLP 379
D.yak            DAGGKAFRKPRRRLKRKPSKTEETDEFSNQRVMANVRERQRTQSLNDAFKSLQQIIPTLP 394
D.ere            DAGGKAFRKPRRRLKRKPSKTEETDEFSNQRVMANVRERQRTQSLNDAFKSLQQIIPTLP 392
D.ana            DAGGKAFRKPRRRLKRKPSKTEDTDEFSNQRVMANVRERQRTQSLNDAFKALQQIIPTLP 389
D.pse            DGAGKAFRKPRRRLKRKPSKTEETDEFSNQRVMANVRERQRTQSLNDAFKSLQQIIPTLP 409
                 *..*******************:***************************:*********

D.mel            SDKLSKIQTLKLATRYIDFLCRMLSSSDISLLKALEAQ----GSPSAYGSASSLLSAAAN 449
D.sim            SDKLSKIQTLKLATRYIDFLCRMLSSSDISLLKALEAQ----GSPSAYGSASSLLSAAAN 435
D.yak            SDKLSKIQTLKLATRYIDFLCRMLSSSDISLLKALEAQ----GSPSAYGSASSLLSAAAN 450
D.ere            SDKLSKIQTLKLATRYIDFLCRMLSSSDISLLKALEAQ----GSPSAYGSASSLLSAAAN 448
D.ana            SDKLSKIQTLKLATRYIDFLCRMLSSSDISLLKALEAQ----GSPSSYGSASSLLSAAAN 445
D.pse            SDKLSKIQTLKLATRYIDFLCRMLSSSDISLLKALEAQVSPMGSSSPYGAASTLLSAAAN 469
                 **************************************    **.*.**:**:*******
```

```
D.mel                 GAEADLKCLRKANGAPIIPPEKLSYLFGVWRMEGDAQHQKA 490
D.sim                 GAEADLKCLRKANGAPIIPPEKLSYLFGVWRMEGDAQHQKA 476
D.yak                 GAEADLKCLRKANGAPIIPPEKLSYLFGVWRMEGDAQHQKA 491
D.ere                 GAEADLKCLRKANGAPIIPPEKLSYLFGVWRMEGDAQHQKA 489
D.ana                 GAEADLKCLRKANGAPIIPPEKLSYLFGVWRMEGDAQHQKA 486
D.pse                 GADADLKCLRKANGAPIIPPEKLSYLFGVWRMEGDVQHQKA 510
                      **:********************************.*****
```

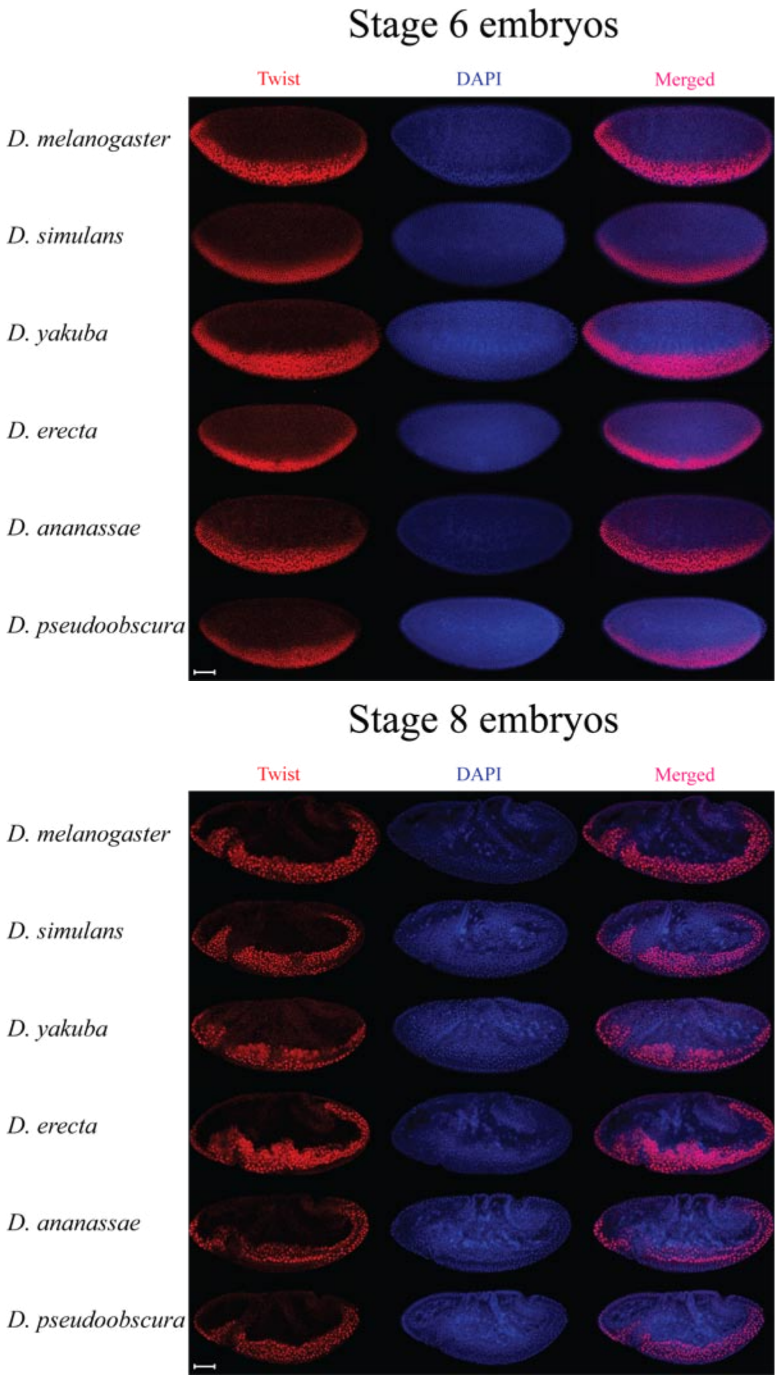
b. Snail

```
D.mel                 MAANYKSCPLKKRPIVFVEERLPQTEALALTKDSQFAQ-----DQPQDLSLKR--GRDEE 53
D.sim                 MAANYKSCPLKKRPFVFVEERLPQTEALALTKDSQFAQ-----DQPQDLSLKR--GRDEE 53
D.yak                 MAANYKSCPLKKRPIVFVEERLPQTEALALTKDLQFAQ-----DQPQDLSLKR--GREEE 53
D.ere                 MAANYKSCPLKKRPIVFVEERLPQTEALALTKDSQFAQ-----DQPQDLSLKR--GREEE 53
D.ana                 MAANYKSCPLKKRPIVFVDQQ---TEALALTKDFQFAVD----DEPQDLSVKR---IKLD 50
D.pse                 MAANYKSCPLKKRPFVFVEEH-PQTEALALTKNSSFAALPAGEDQPQDLSLKRKASREQD 59
                      **************:***:::   ********: .**      *:*****:**    . :

D.mel                 TQDYQQPEPKRDYVLNLSKTPER-NSSSSSNSCLLSPPVEAQDYLP----------TEIH 102
D.sim                 TQDYQQPEPKRDYVLNLSKTPER-ISSSSSNSCLLSPPVEAQDYLP----------TEIH 102
D.yak                 TQDYQQPEPKRDYVLNLSKTPER-ISSSSSNSCLLSPPVEAQDYLP----------TEIH 102
D.ere                 TQDYQQPEPKRDYVLNLSKTPER-ISSSSSNSCLLSPPVEAQDYLP----------TEIH 102
D.ana                 ADHYEQP----DYALNLSKTPER-MPFSGPSSCLLSPPADGEDYQP--------PTSNIH 97
D.pse                 FEDYELP-AKREYVLNLSKTPETPRSASPLCSALLSPIAEHSDYQPESESQAQCQPIDIH 118
                       :.*: *    :*.********   . *   *.**** .: .** *           :**

D.mel                 MRGLTAGTTGYTTATPTTINPFQSAFVMAAGCNPISALWSSYQP----HLAAFPSPASSM 158
D.sim                 MRGLTAGTTGYTTATPTTINPFQSAFVMAAGCNPISALWSSYQP----HLAAFPSPASSM 158
D.yak                 MRGLTAGTTGYTTASPTTINPFQSAFVMAAGCNPISALWSSYQP----HLAAFPSPASSM 158
D.ere                 MRGLTSGTTGYTTATPTTINPFQSAFVMAAGCNPISALWSSYQP----HLAAFPSPASSM 158
D.ana                 LRGLTAGTTGYTTTSP-TANPFQSAFVMAAGCNPISALWSSYQP----HLAAFPSPASSM 152
D.pse                 MRGLTAATAGYTT------NPYQSAFVMAAGCNPISALWSSYQPHIASHLSAFPSPASSM 172
                      :****:.*:****      **:**********************    **:*********

D.mel                 AS----PQSVYS----YQQMTPPSSPGSDLETGSEPEDLSVRNDIPLPALFHLFDEAKSS 210
D.sim                 AS----PQSVYS----YQQMTPPSSPGSDLETGSEPEDLSVRNDIPLPALFHLFDEAKSS 210
D.yak                 AS----PQSAYS----YQQMTPPSSPGSDLETGSEPEDLSVRNDIPLPALFHLFDEAKSS 210
D.ere                 AS----PQSVYS----YQQMTPPSSPGSDLETGSEPEDLSVRNDIPLPALFHLFDEAKSS 210
D.ana                 AS----PQSVYSSGYPQQMMTPPSSPGSEVDSGSEPEDLSVRNDIPLPALFHLFDEARSS 208
D.pse                 ASSMASPHSVYS----YQQMTPPSSPGS--EASSEPEDLSVRNDIPLPALFHLFDEARSS 226
                      **    *:*.**     * *********  ::.************************:**

D.mel                 S----SGASVSSSSGYSYTPAMSASSAS----VAANHAKNYRFKCDECQKMYSTSMGLSK 262
D.sim                 S----SGASVSSSSGYSYTPAMSASSAS----VAANHAKNYRFKCDECQKMYSTSMGLSK 262
D.yak                 S----SGASVSSSSGYSYTPAMSASSAS----VAANHAKNYRFKCDECQKMYSTSMGLSK 262
D.ere                 S----SGASVSSSSGYSYTPAMSASSAS----VAANHAKNYRFKCDECQKMYSTSMGLSK 262
D.ana                 SNSSVTSSSSSTGSSYLYNSSNSSGSVSGSGSAASSAAKNYRFKCDQCQKMYSTSMGLSK 268
D.pse                 S----ASSSSGSVGSYAYLAASSAPNASGAVGSASSAAKNYRFKCDQCQKMYSTSIGLSK 282
                      *    :.:* .: ..* * .: *: ..*     *:. *********:********:****

D.mel                 HRQFHCPAAECNQEKKTHSCEECGKLYTTIGALKMHIRTHTLPCKCPICGKAFSRPWLLQ 322
D.sim                 HRQFHCPAAECNQEKKTHSCEECGKLYTTIGALKMHIRTHTLPCKCPICGKAFSRPWLLQ 322
D.yak                 HRQFHCPAAECNQEKKTHSCEECGKLYTTIGALKMHIRTHTLPCKCPICGKAFSRPWLLQ 322
D.ere                 HRQFHCPAAECNQEKKTHSCEECGKLYTTIGALKMHIRTHTLPCKCPICGKAFSRPWLLQ 322
D.ana                 HQQFHCPAAECNQEKKTHSCEECGKLYTTIGALKMHIRTHTLPCKCPICGKAFSRPWLLQ 328
D.pse                 HRQFHCPAAECNQEKKTHSCEECGKLYTTIGALKMHIRTHTLPCKCPICGKAFSRPWLLQ 342
                      *:**********************************************************

D.mel                 GHIRTHTGEKPFQCPDCPRSFADRSNLRAHQQTHVDVKKYACQVCHKSFSRMSLLNKHSS 382
D.sim                 GHIRTHTGEKPFQCPDCPRSFADRSNLRAHQQTHVDVKKYACQVCHKSFSRMSLLNKHSS 382
D.yak                 GHIRTHTGEKPFQCPDCPRSFADRSNLRAHQQTHVDVKKYACQVCHKSFSRMSLLNKHSS 382
D.ere                 GHIRTHTGEKPFQCPDCPRSFADRSNLRAHQQTHVDVKKYACQVCHKSFSRMSLLNKHSS 382
D.ana                 GHIRTHTGEKPFECPECPRSFADRSNLRAHQQTHVEVKKYACQVCHKSFSRMSLLNKHTS 388
D.pse                 GHIRTHTGEKPFQCPDCPRSFADRSNLRAHQQTHVDVKKYACQVCHKSFSRMSLLNKHSA 402
                      ************:**:*******************:**********************::
```

```
D.mel            SNCTITIA 390
D.sim            SNCTITIA 390
D.yak            SNCTITIA 390
D.ere            SNCTITIA 390
D.ana            SNCSVTVA 396
D.pse            SNCTITIA 410
                 ***::*:*
```

Protein sequence alignment of (a) Twist and (b) Snail and their orthologs across the six *Drosophila* species. Comparing transcription factor binding sites across a variety of species by ChIP-seq requires tight controls to make sure that variations are mostly the result of evolutionary changes rather than differences in the developmental stage, cross-reactivity of the antibodies or other experimental variables. As a first test for the possible cross-reactivity of the antibodies, we examined the evolutionary conservation of the Twist and Snail protein by a ClustalW sequence alignment. Twist shows high sequence conservation, especially in the C-terminal part to which the polyclonal antibody had been raised[1]. Snail also shows a high degree of conservation but antibodies raised against the full-length *D. melanogaster* protein (from Mike Levine[2] or a commercially available antibody (Millipore MAB5494)) did not cross-react well in the other species.

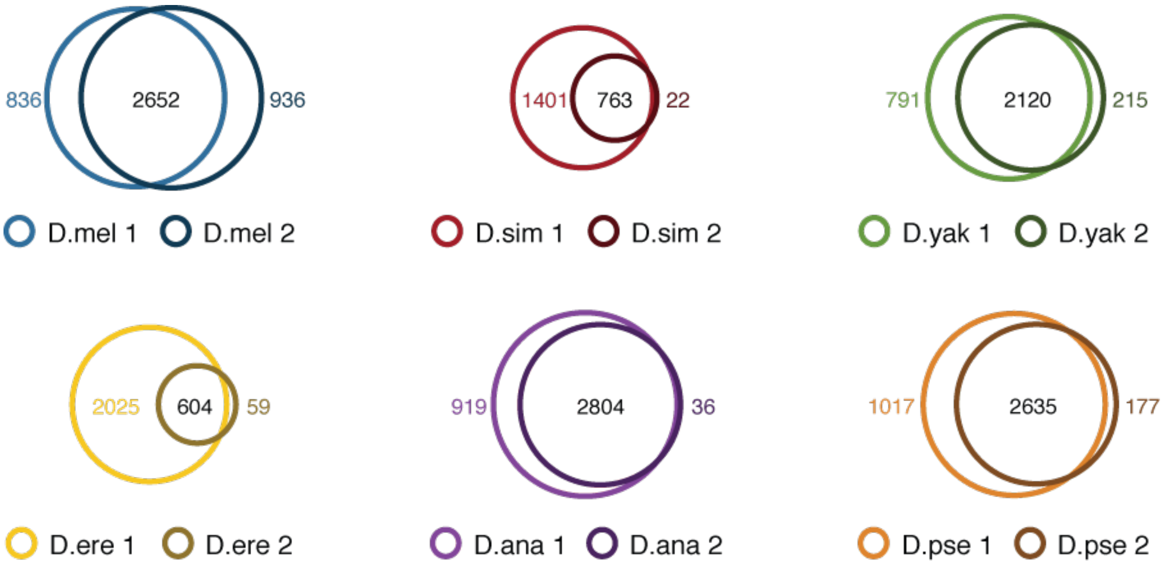
Figure S2: Cross-reactivity of the Twist antibody in immunostainings



Both Twist antibodies[1,2] used in ChIP assays react robustly with Twist homologues in embryos of the six *Drosophila* species used in this study (anterior: left; posterior: right; scale bar: 50 µm). The figure shows stainings with the Twist antibody from Mike

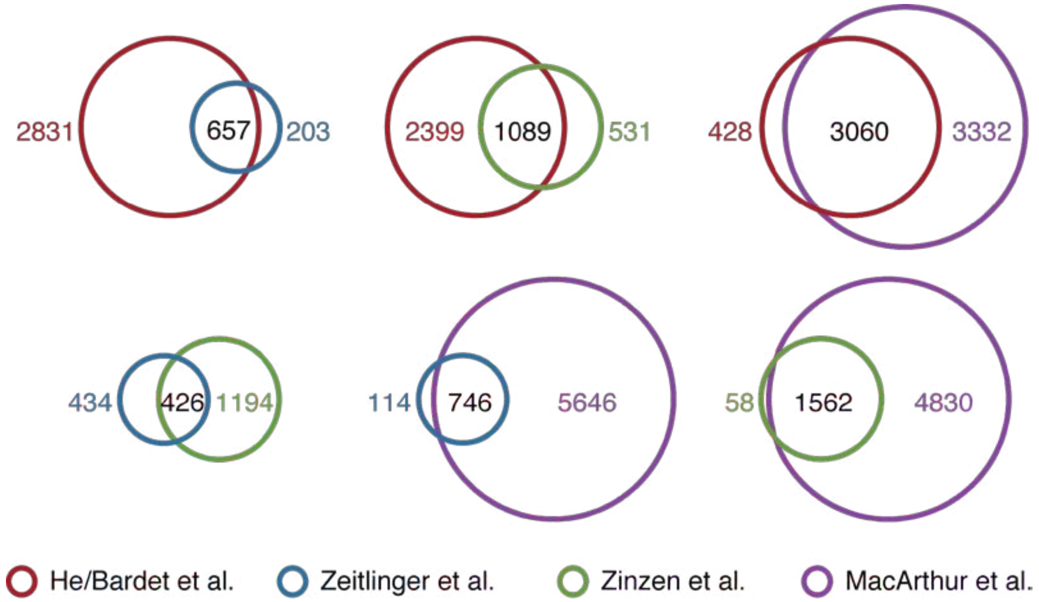
Levine[2]. Even stronger signal is seen in all species with the Twist antibody from Eileen Furlong[1] (not shown). Note that the Twist expression pattern is well conserved in stage 6 and 8 embryos of all species.

Figure S3: Venn diagrams of peak overlap between biological replicates



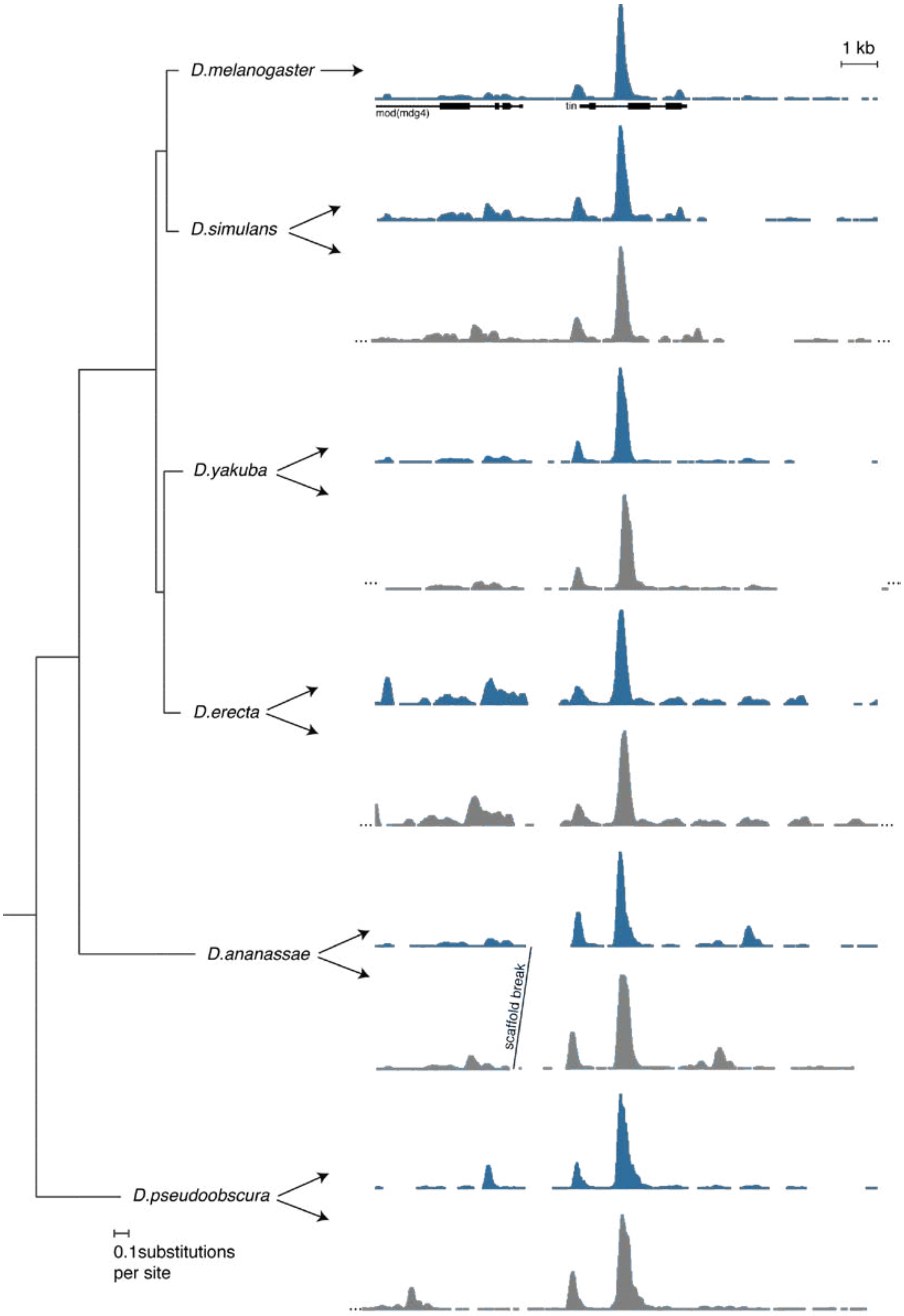
High overlap between peaks from independent biological replicates. We called peaks in each species using a constant p-value cutoff of 10^{-22}, which corresponds to an FDR of 0.1% in *D. melanogaster* (**Table S6**). Note that such analysis underestimates the overlap because of the winner's curse phenomenon (see **Table S9** and **Figure S8** for more information).

Figure S4: Twist binding overlap with other studies



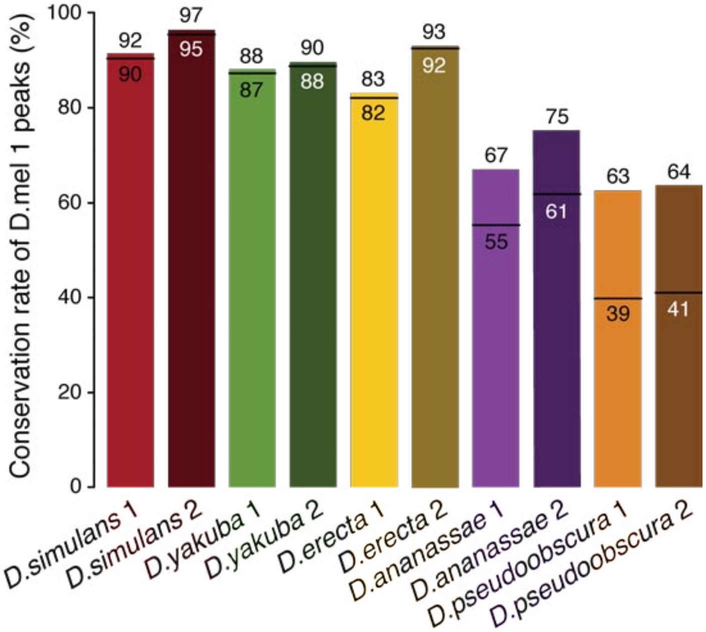
Pairwise comparison (as Venn diagrams) between Twist binding peaks obtained from different studies. Shown are the 3488 ChIP-seq peaks from this study (red), the 860 ChIP-chip Twist Snail (TS) peaks from Zeitlinger et al.[3] (blue), the 1620 ChIP-chip peaks from Zinzen et al.[4](green) and 6392 ChIP-chip peaks (the intersection of two experiments using different antibodies against Twist) from MacArthur et al. [5](purple). All studies performed ChIP on 2-4 hours embryo collections, except for MacArthur et al.[5], who used a 2-3 hours embryo collection. The overlap between the identified binding peaks is large in all cases, especially given that the data were produced in different labs, with different methodologies. The MacArthur et al.[5] and Zinzen et al.[4] data overlap particularly well, presumably because both were performed using the Affymetrix protocol and arrays. Our ChIP-seq data also compare very favorably to previous ChIP-chip data, with overlaps that are in the range of the overlaps of the different ChIP-chip studies, indicating that different detection platforms (microarray or sequencing) produce similar results. Indeed, in our experience, it is the amplification protocol that produces most biases between different technology platforms (not shown).

Figure S5: Twist binding at the *tinman* locus (extended view)



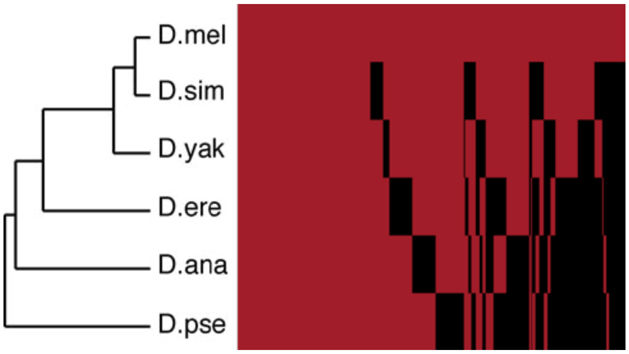
Read densities in an extended 14 kb window of the *tinman* locus in the different species either after translation to *D. melanogaster* coordinates (blue) or in coordinates of the respective original genomes (grey). For the latter, regions orthologous to the *tinman* locus have been defined using liftOver and manually centered on the Twist peak. The regions in the original genomes can differ in length due to species-specific insertions and deletions. In *D. ananassae*, the synteny was not maintained throughout the entire 14 kb window, such that the left part stems from a different scaffold (note that the *D. ananassae* genome consists of 13749 scaffolds and we cannot assess if the synteny break exists in the actual chromosomes).

Figure S6: Conservation of peaks from a species-specific view



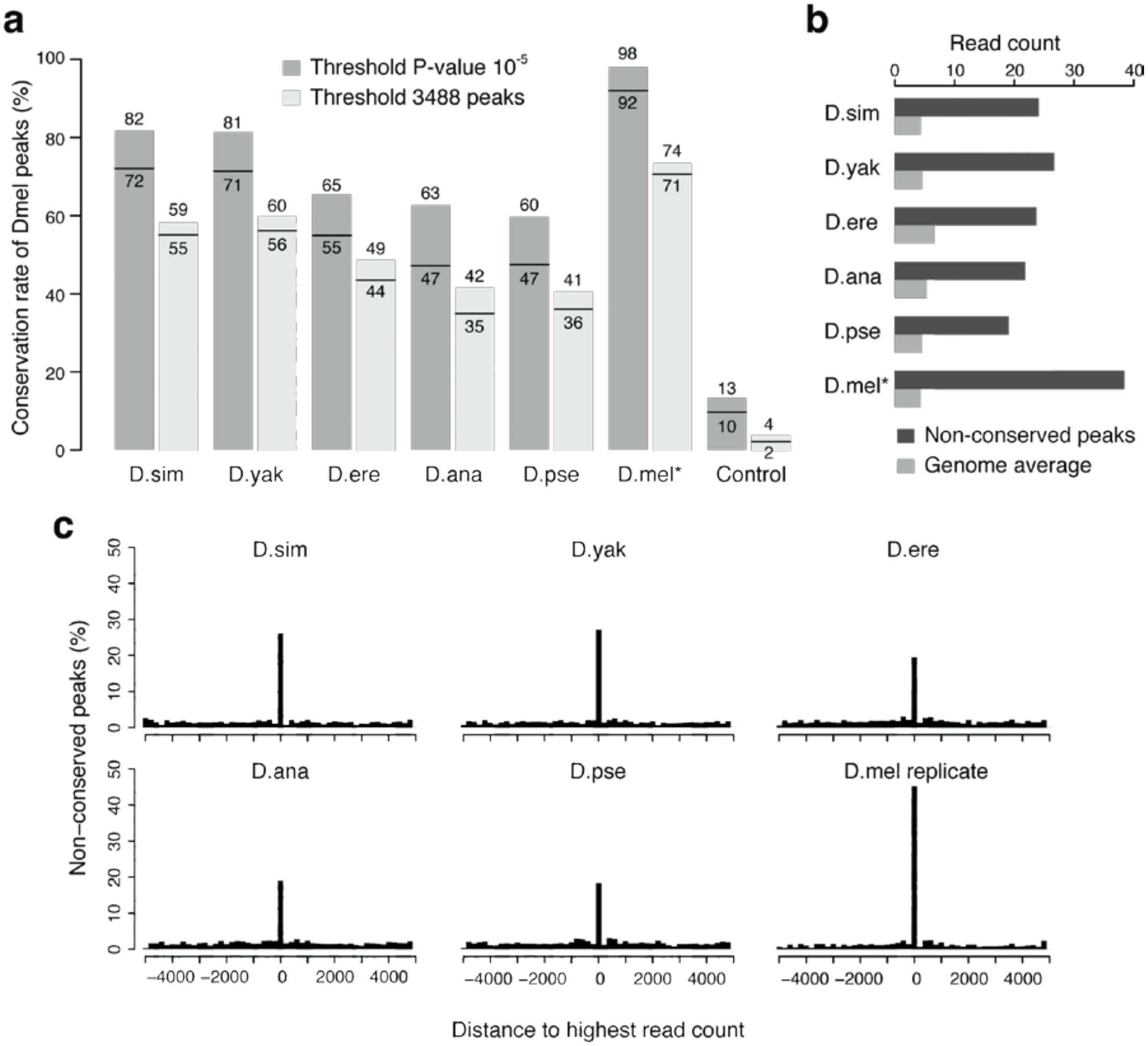
A non-*melanogaster* centric analysis (*inverse analysis*) gives similar results to the *D. melanogaster*-centric analysis. The conservation of peaks directly identified in the original non-*melanogaster* genomes ($p<=10^{-22}$; **Table S5**) compared to *D. melanogaster*, i.e. the reverse analysis to the one presented in the main paper. The bar height indicates conservation relative to all peaks for which an orthologous region exists in *D. melanogaster* (i.e. that can be mapped using liftOver), while the lower black line indicates the conservation relative to all peaks including those for which an orthologous regions is not present in *D. melanogaster*. Interestingly, the conservation rates for *D. simulans*, *D. yakuba*, and *D. erecta* are even higher in the inverse analysis. We noted that this correlates with the quality of the IP in the respective species or replicates: lower enrichments mean that only fewer peaks pass the cutoff (e.g. *D. erecta* 2; **Figure S3**).

Figure S7: Clustering of Twist binding data recapitulates the phylogenetic tree



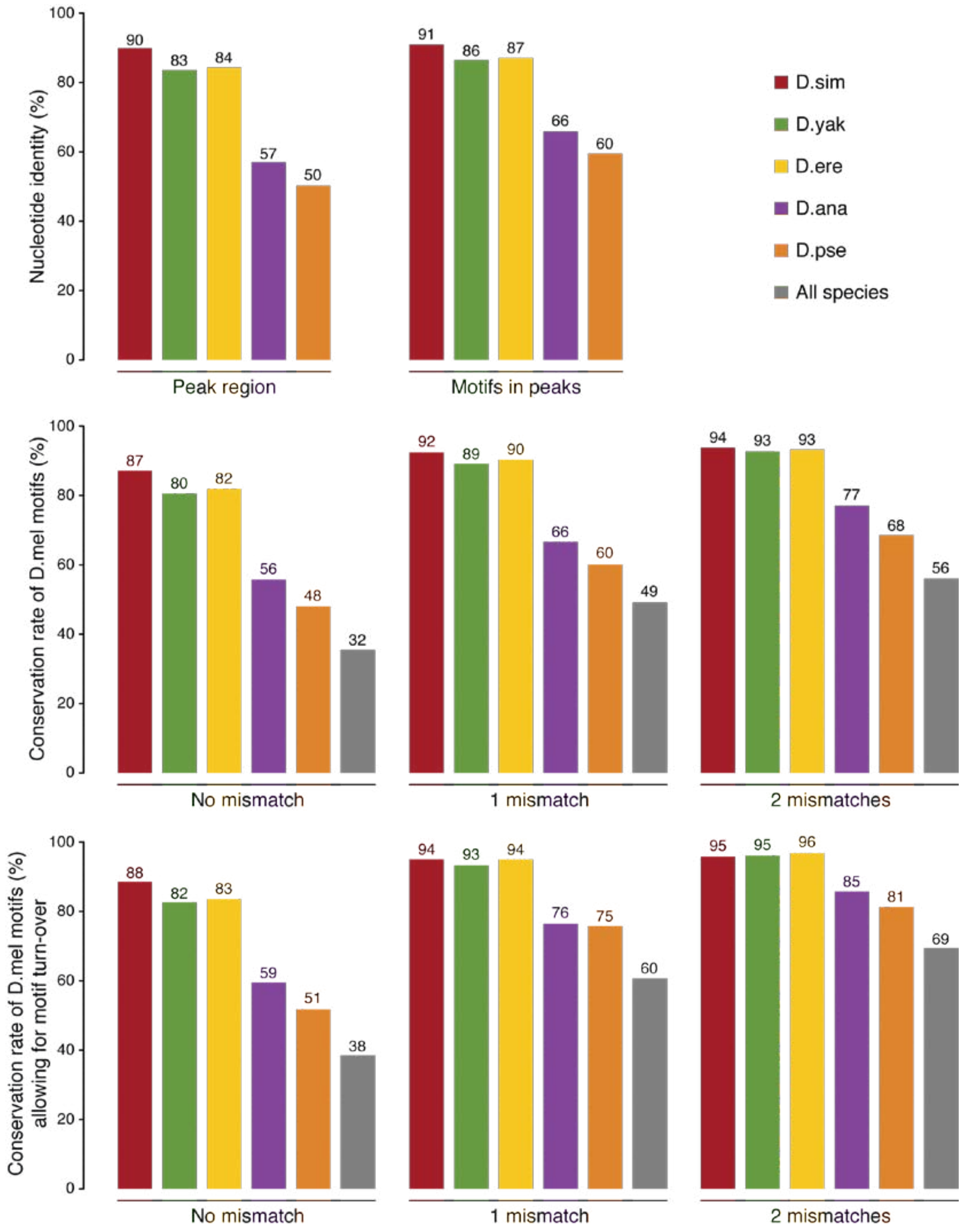
Hierarchical clustering recapitulates the established phylogenetic tree. Twist binding peaks in *D. melanogaster* were clustered based on their conservation in each of the other *Drosophila* species (red: conserved; black: non-conserved). The binary conservation data were clustered by average linkage clustering with centered correlation using Cluster3[10]. The resulting dendrogram on the left recapitulates the known phylogenetic tree. In addition, 54.6% (1905 out of 3488) of the peaks display a conservation pattern in agreement with the branching pattern of the phylogenetic tree, i.e. peaks are strictly progressively lost with increasing phylogenetic distance. This is much higher than the 0.1% that are expected, if the peak conservation patterns were distributed randomly (to all 945 potential bifurcating trees with 6 leaves). Given the influence of species-specific evolutionary changes along these branches, 54.6% is also high. Taken together, the clustering analysis suggests that conservation and divergence in our Twist binding data are mainly evolutionary events.

Figure S8: Choice of a sensitive cutoff to adjust for the winner's curse



Impact of the winner's curse phenomenon on the conservation rates. We define a *D. melanogaster* peak as conserved in another species, if the corresponding genomic position in that species displays a significantly non-random ChIP-Seq enrichment ($p \leq 10^{-5}$). We chose this approach to avoid underestimating conservation due to the so-called "winner's curse": when identifying the best data points (i.e. peaks) in one experiment by applying a cutoff, their values (i.e. peak enrichments) in a replicate experiment are systematically lower– purely due to experimental noise. (This phenomenon has been dubbed the "winner's curse" as winners from auctions tend to pay a higher price than the true value; e.g. see Lohmueller et al.[8]). This figure provides evidence that a stringent cutoff during the assessment of peak conservation underestimates conservation by missing peaks in the comparative species. **(a)** The conservation estimates for each of the species in comparison to *D. melanogaster* are shown, either with adjusting for the winner's curse phenomenon (dark grey columns) or without (light grey columns). Biological replicates are shown for *D. melanogaster* (*D.mel**). The black bar underneath each bar shows the fraction of these conserved peaks that are the highest peak within 10 kb in the other species at the position orthologously aligned to the *D. melanogaster* peak summit. To control for the average conservation in the genome, we offset all *D. melanogaster* peaks by 20 kb and determined their conservation rate (*Control*). **(b** and **c)** If not adjusting for the winner's curse, so-called 'non-conserved' peaks show strong evidence of conservation. **(b)** The average read count for 'non-conserved' peaks (dark grey) is much above the genome average (light grey) for each species, as well as for the biological replicate in *D. melanogaster*. **(c)** The position of 'non-conserved' peaks, when orthologously aligned to the *D. melanogaster* peak summit, is the position of the highest read count within a 10 kb window. Thus, in 19% to 45% of these seemingly divergent cases, there is no doubt that the peak is still present in the other species.

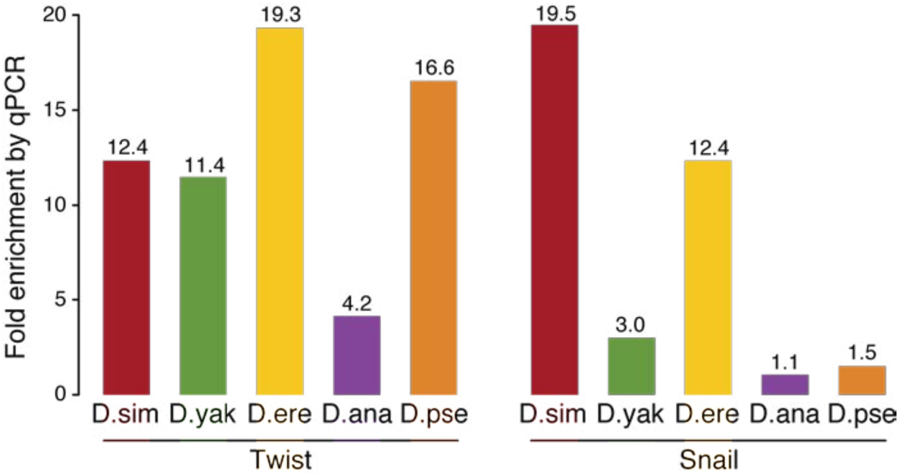
Figure S9: Expected conservation rates based on sequence conservation



Expected Twist binding conservation rates. Rates were calculated based on the sequence identity of the peak region or the Twist motifs in the peak regions (as e.g. in Richards et al.[9]; top) or based on the conservation (i.e. presence) of the Twist motif using different stringencies for motif-matches (as in Kheradpour et al.[10]; middle and bottom). For the latter two, we expect Twist binding to be conserved in the respective species, if the peak region contains at least one conserved Twist motif. The Twist motif is defined by motif-PWM-matches with p-values of 0.015625, 0.00390625, 0.000976562, which corresponds to a maximum allowance of 0, 1, or 2 mismatches. The middle panel shows the results for perfectly aligned motifs, while motif turnover (i.e. movement of 150bp in the alignment) was allowed in the lower panel. The estimates follow the established phylogeny, yet span a wide range from 32% to 69% of peaks that can reasonably be expected to be conserved across all species (even without taking into account motifs of partner transcription factors). Nevertheless, the range of expected binding conservation rates is clearly above those recently reported for vertebrate species.

Figure S10: Snail binding conservation

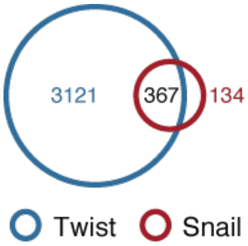
a.



b.

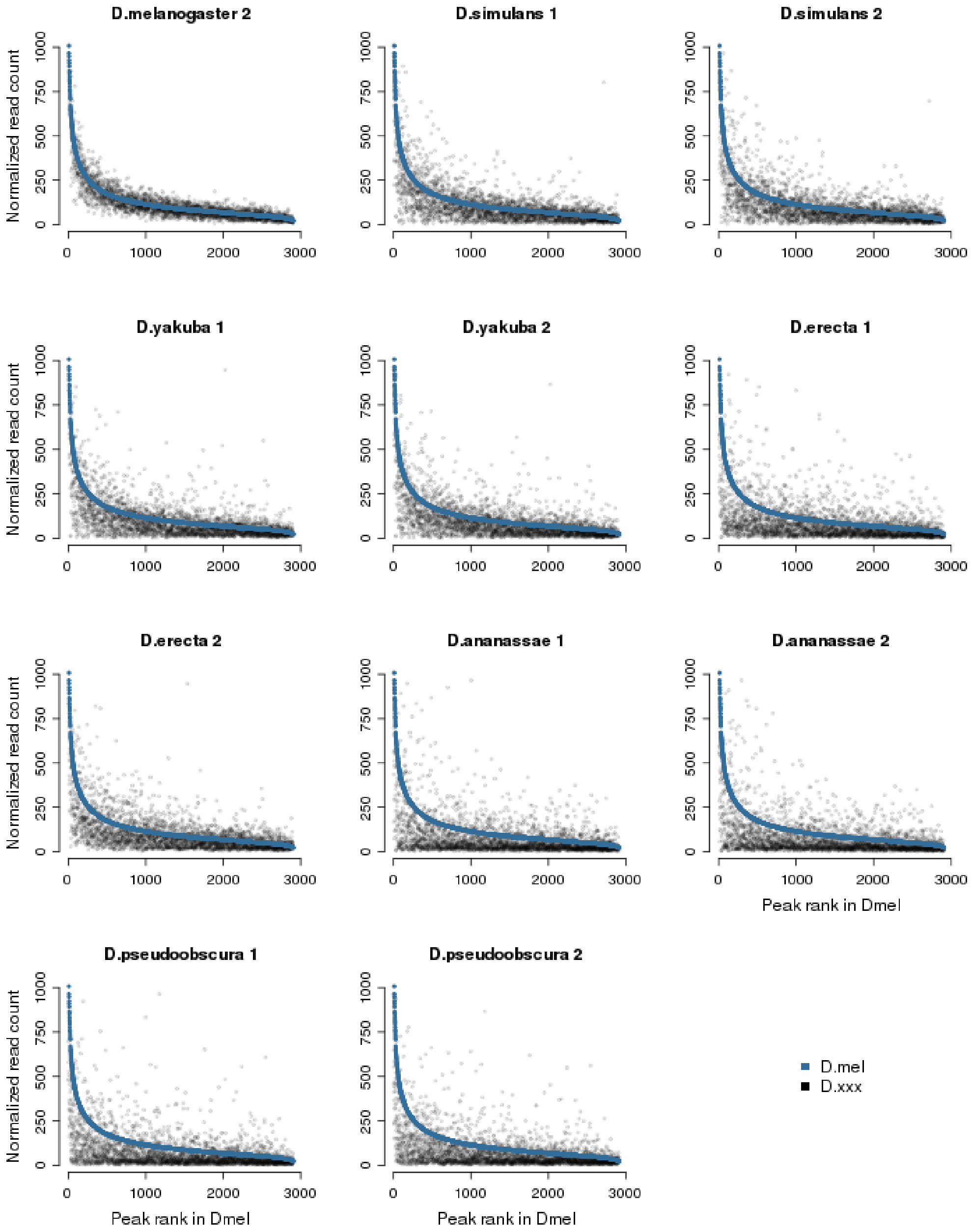
Species	Number of peaks
D. melanogaster 1	501
D. melanogaster 2	765
D. simulans 1	3910
D. simulans 2	179
D. erecta	26

c.



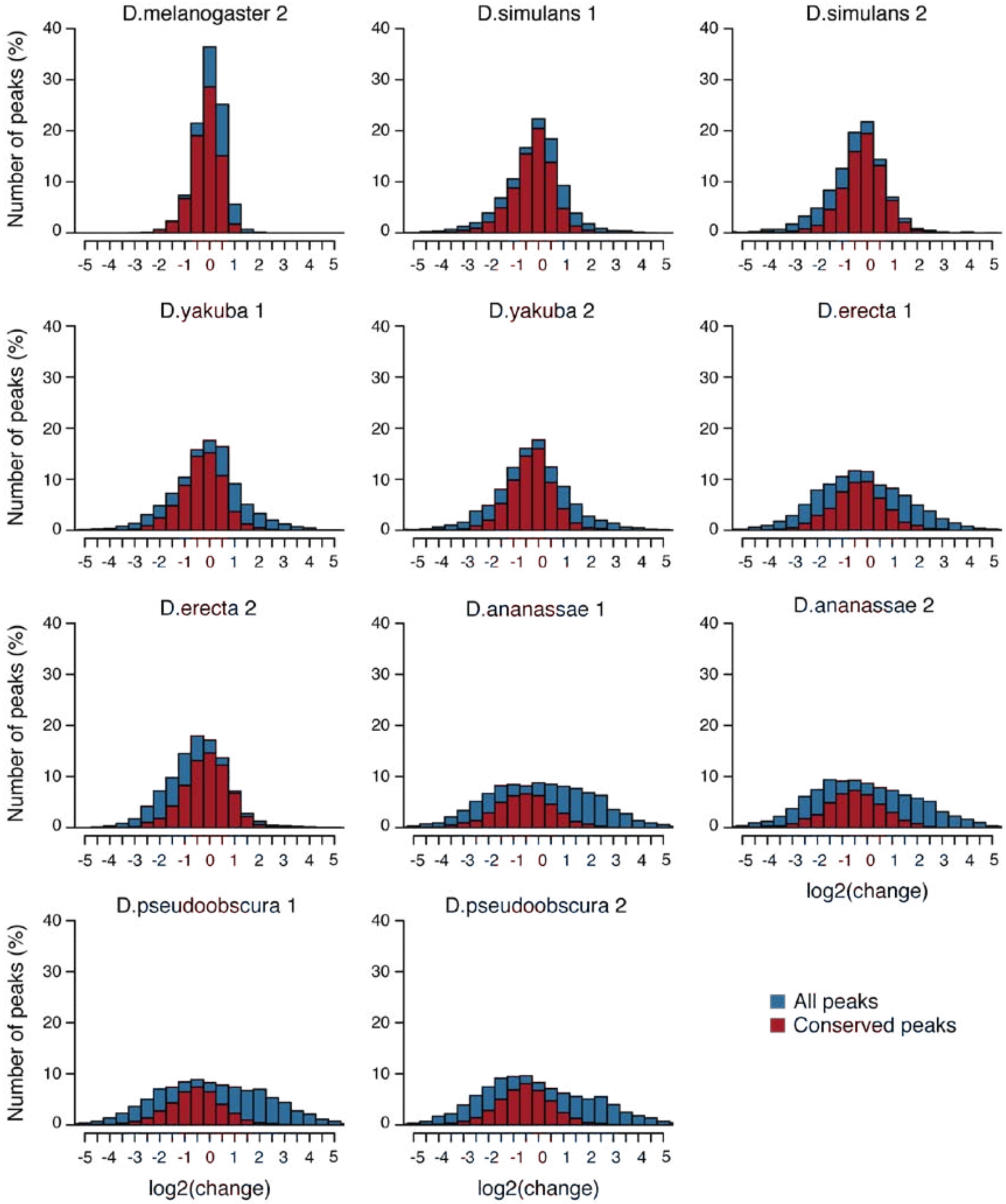
Results of our attempts to obtain comparative Snail ChIP-seq data. (a) To monitor whether individual ChIP samples had worked, we tested the enrichment of Twist and Snail by quantitative PCR (qPCR). The enrichment was tested on an intronic enhancer of the *Dscam* gene (based on Zeitlinger et al.[3]), which has a Twist/Snail motif that is perfectly conserved across all species. While Twist ChIPs show significant enrichment across all five non-*melanogaster* species tested, Snail ChIPs show drastically reduced enrichments at the evolutionarily more distant species (i.e. *D. ananassae* and *D. pseudoobscura*). Indeed, when we sequenced the Snail ChIPs, we found that even the *D. erecta* sample, which showed significant enrichment at *Dscam*, yielded a poor signal-to-noise ratio and low enrichment levels at the genome-wide scale. **(b)** The numbers of binding peaks obtained with Snail ChIPs were very low and not suitable for analysis in the case of *D. erecta*. **(c)** Twist and Snail binding peaks in *D.melanogaster* strongly overlap (75% of the Snail peaks overlap with Twist peaks).

Figure S11: Raw data of quantitative changes



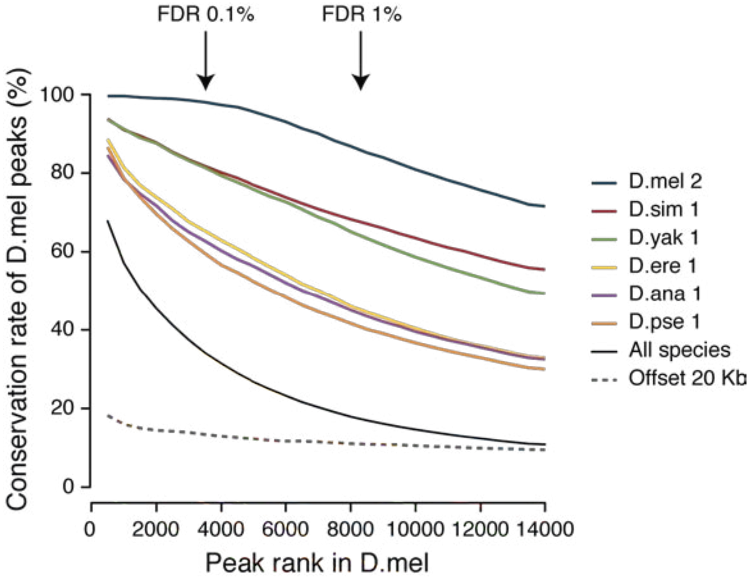
Scatterplots of the read counts in a specific *Drosophila* species vs. their read counts in *D. melanogaster*, ranked by read count in *D. melanogaster* after quantile normalization (similar to Bradley et al.[7]). While there are some quantitative changes between biological replicates, the changes become larger with increasing phylogenetic distances.

Figure S12: Quantitative changes



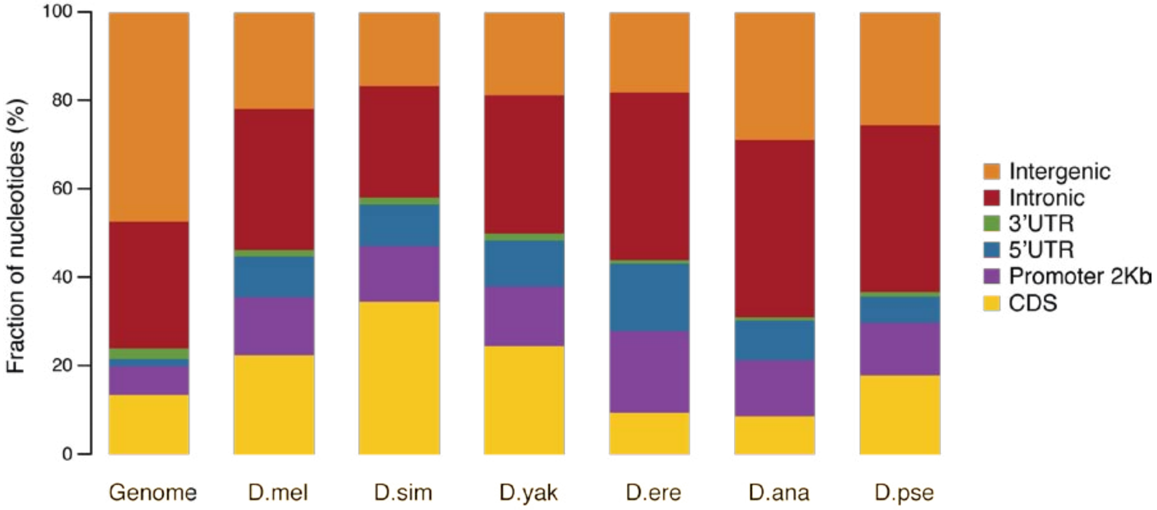
Histogram of the quantitative changes (Dxxx/Dmel) for peaks called in either species (Table S5). The data for all peaks are shown in blue, while those for *D. melanogaster* peaks that we classified as conserved are in red. Since relative rather than absolute changes are shown, peaks that increase or decrease in peak height can also be conserved. The range of quantitative changes that is due to experimental variation can be estimated from the changes in the biological replicate (*D. melanogaster 2*; top left panel).

Figure S13: Conservation rate of *D. melanogaster* peaks at different ranks



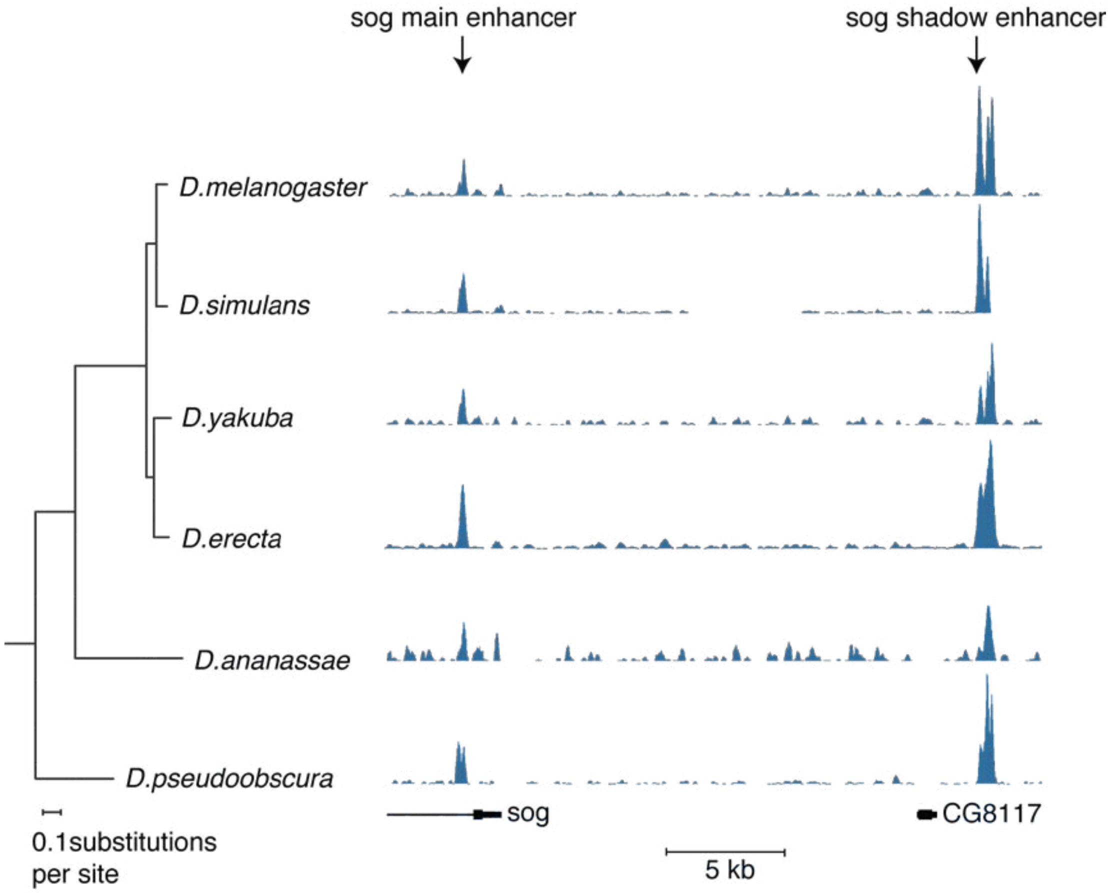
Dependence of conservation rates on the rank of the *D. melanogaster* reference peaks. Up to 70% of the highest ranking peaks are shared across all six species (black line) and conservation rates drop with the peak height, which is shown rank-ordered on the x-axis in bins of 500 peaks. There are likely two reasons for this effect. First, conservation of lower ranking peaks will more likely be missed as they are closer to the cutoff and less reliably distinguishable from noise (the rates drop even in the *D. melanogaster* replicate). Second, bins at lower ranks contain increasingly more false positive peaks that dilute the signal from potentially conserved real peaks. The cutoffs that we chose for our analysis (FDR cutoffs 0.1% to identify peaks and 1% to identify low-occupancy peaks; details see main text) are indicated with arrows at the top of the graph.

Figure S14: Distribution of Twist binding peaks relative to annotated genes



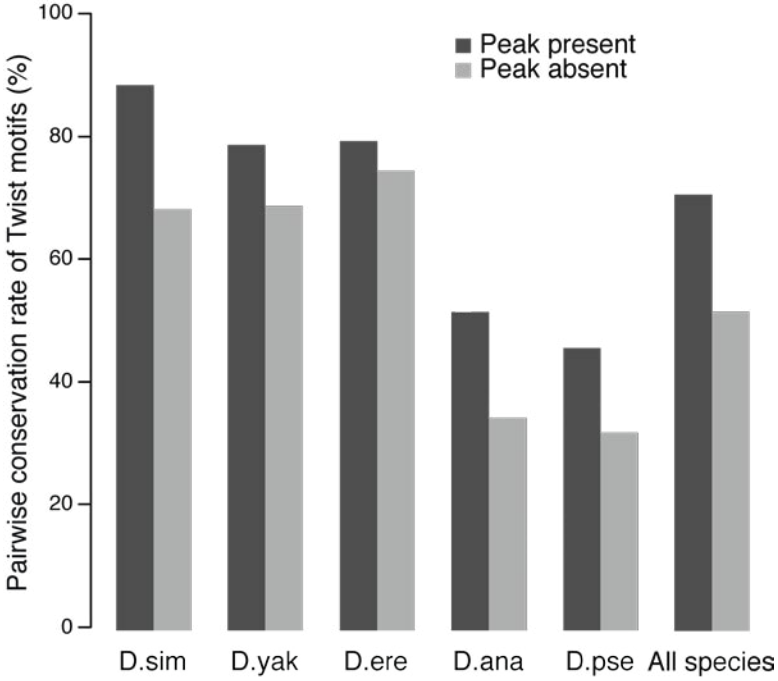
Distribution of the top 3488 peaks identified in each species on different genomic regions (Promoter region = -2 kb to start site, CDS = coding sequence). Each nucleotide in a peak was counted towards its genomic region, i.e. peaks were not assigned to only one region. The first bar represents the genome average for each region type as control. Note that the 5'UTR and the promoter region are enriched for Twist binding peaks in all species.

Figure S15: Example of a conserved shadow enhancer



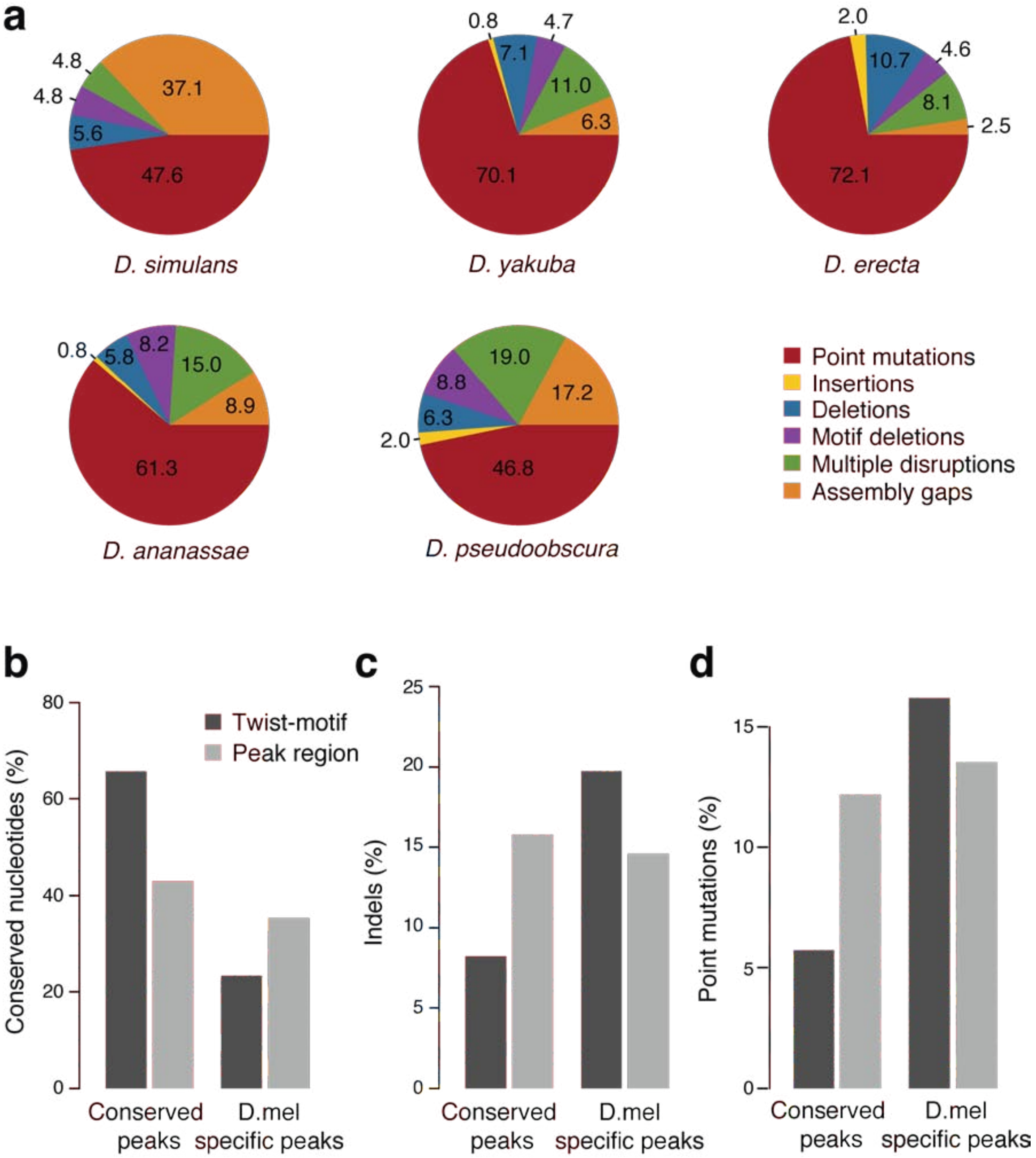
An example of a well-conserved shadow enhancer at the *sog* locus. Note that there can be quantitative changes in Twist binding at shadow enhancers.

Figure S16: Pairwise conservation rate of Twist motifs



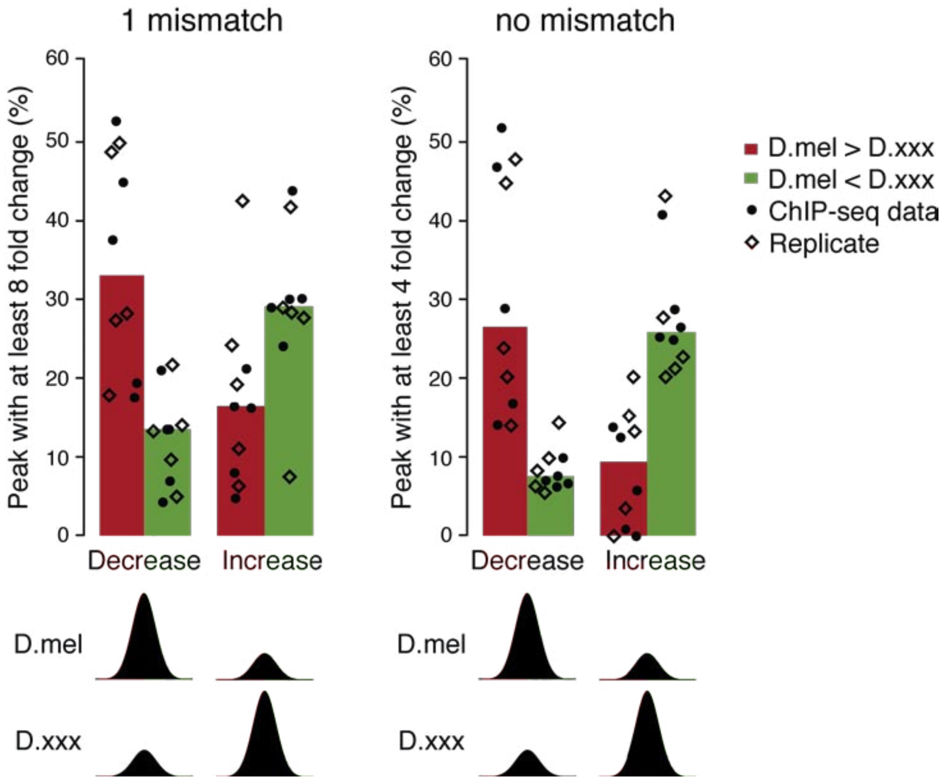
The conservation rate of Twist motifs in all *D. melanogaster* peaks that are present in the indicated species (dark grey) or absent (light grey). Note that this pairwise conservation rate of Twist motifs follows the phylogenetic tree and depends on the presence or absence of experimentally determined Twist binding, i.e. if Twist binding is conserved, the Twist motif is also more often conserved.

Figure S17: Sequence basis for binding peak losses



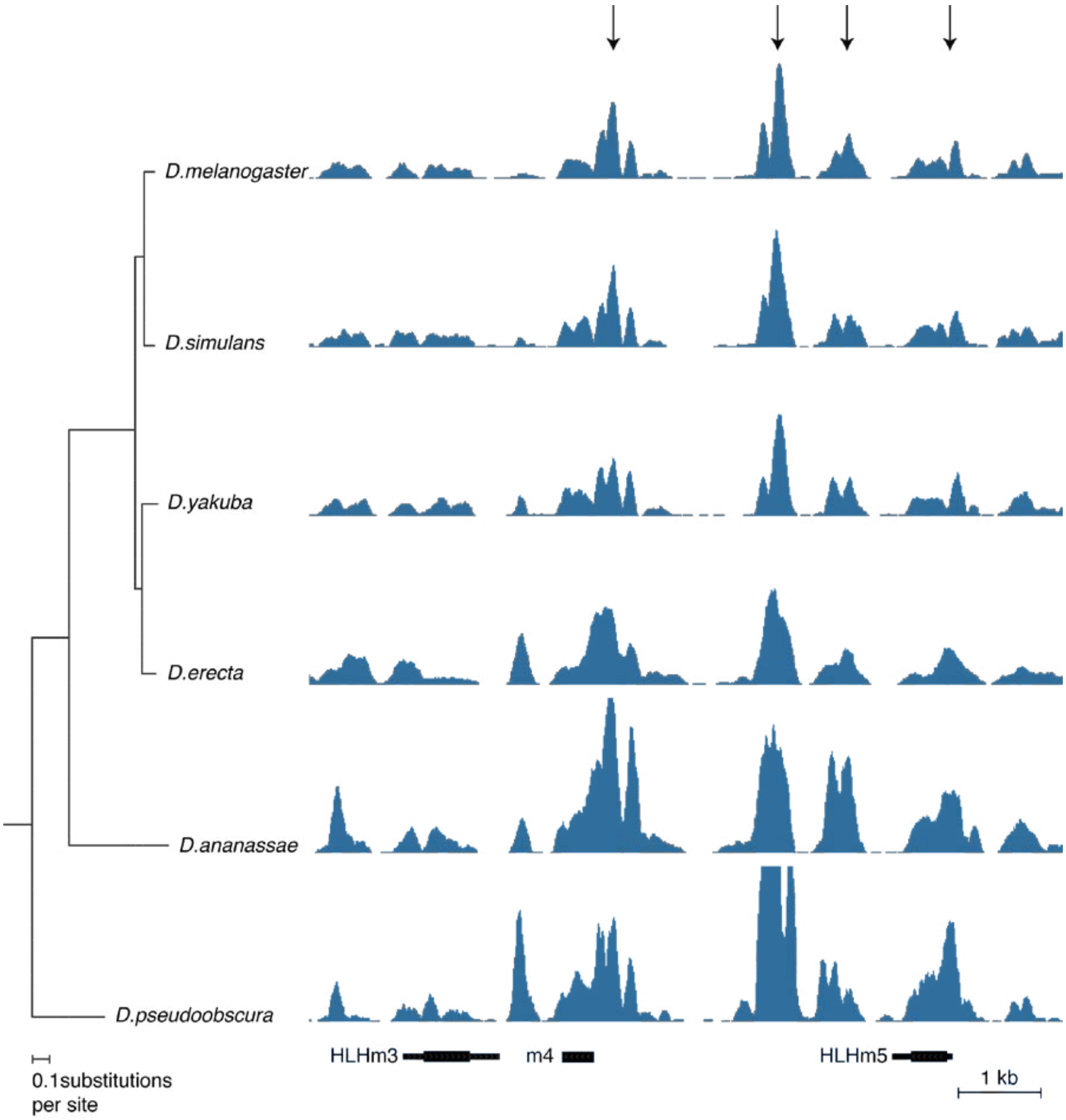
Twist binding peak losses across the *Drosophila* phylogeny are associated with insertions, deletions and point mutations in the Twist motif. (a) Twist motif losses that co-occur with Twist binding losses in any of the species result from nucleotide-to-nucleotide point mutations, deletion of larger regions or assembly gaps, deletions of nucleotides within the motif or of the entire motif, insertions, or multiple disruptions of the above. Note that the fraction of motifs lost due to point mutations decreases with increasing phylogenetic distance from *D. melanogaster* and that the number of alignment/assembly gaps increases in the same order. The *D. simulans* data are an exception since they contain a notably high number of deletions due to assembly gaps (also noted in Stark et al.[12]), despite the short phylogenetic distance to *D. melanogaster*. **(b-d)** In conserved Twist binding peaks, sequence changes are preferentially absent from Twist motifs (PWM-cutoff $P<10^{-3}$). (**b**) Nucleotides in Twist motifs, but not the average peak region, are preferentially conserved across all six species. (**c**) Indels and (**d**) nucleotide-to-nucleotide mutations are reduced in conserved Twist peaks but increased in *D. melanogaster*-specific peaks, while average peak regions display similar numbers of indels and mutations. Indels and point mutations are assessed as their number across all species divided by the length of the region in *D. melanogaster*.

Figure S18: Correlation between Twist binding levels and motif quality



Quantitative changes of peak height correlate with Twist motif quality (MAST score). Results as in main **Figure 4c** but with different parameters (8-fold or 4-fold changes of peak height and for Twist motif matches with 1 or no mismatch allowed)= Peaks that are lower in a second species compared to *D. melanogaster* ("Decrease") contain more motifs with lower scores in that species than in *D. melanogaster*. The reverse is true for peaks that are higher in the second species ("Increase"). Circles and diamonds represent the fraction of changed motifs in each pairwise comparison and bar heights represent the median values.

Figure S19: Conserved low-occupancy peaks at the *Enhancer-of-Split* locus



Example of conserved low-occupancy Twist peaks that are found near high-occupancy peaks (arrows) at the *Enhancer-of-Split* locus. The fact that low-occupancy peaks are preferentially found near high-occupancy peaks raises the possibility that they increase the local concentration of transcription factor at important binding regions.

References

1 Sandmann, T. *et al.* A core transcriptional network for early mesoderm development in Drosophila melanogaster. *Genes Dev* **21**, 436-449 (2007).

2 Zinzen, R. P., Senger, K., Levine, M. & Papatsenko, D. Computational models for neurogenic gene expression in the Drosophila embryo. *Curr Biol* **16**, 1358-1365 (2006).

3 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).

4 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).

5 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, doi:10.1186/gb-2009-10-7-r80 (2009).

6 Schmidt, D. *et al.* Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science*, doi:10.1126/science.1186176 (2010).

7 Bradley, R. K. *et al.* Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. *PLoS Biol* **8**, e1000343, doi:10.1371/journal.pbio.1000343 (2010).

8 Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* **33**, 177-182, doi:10.1038/ng1071 (2003).

9 Richards, S. *et al.* Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome Res* **15**, 1-18, doi:10.1101/gr.3059305 (2005).

10 Kheradpour, P., Stark, A., Roy, S. & Kellis, M. Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Res* **17**, 1919-1931, doi:10.1101/gr.7090407 (2007).

11 Hong, J. W., Hendrix, D. A. & Levine, M. S. Shadow enhancers as a source of evolutionary novelty. *Science* **321**, 1314, doi:10.1126/science.1160631 (2008).

12 Stark, A. *et al.* Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* **450**, 219-232 (2007).

Manuscript B

Cis-regulatory requirements for tissue-specific programs of the circadian clock

Meireles-Filho ACA, **Bardet AF***, Yáñez-Cuna JO*, Stampfel G, Stark A.
(submitted)

* contributed equally

A.C.A.M.F. and A.S. conceived and designed the experiments. A.C.A.M.F. and G.S. performed the experiments. **A.F.B.** analysed the ChIP-seq datasets and J.O.Y.C. performed the predictions. A.C.A.M.F. and A.S. analyzed the data and wrote the paper.

Cis-regulatory Requirements for Tissue-specific Programs of the Circadian Clock

Antonio C. A. Meireles-Filho, Anaïs F. Bardet*, J. Omar Yáñez-Cuna*, Gerald Stampfel, and Alexander Stark#

Research Institute of Molecular Pathology (IMP), 1030 Vienna, Austria

* These authors contributed equally to this work.

# To whom correspondence should be addressed (stark@starklab.org).

Running title: Tissue-specific Programs of the Circadian Clock

Keywords: circadian clock, *Drosophila*, transcription factor, enhancers, ChIP-seq, machine learning; cis-regulatory motifs

ABSTRACT

Broadly expressed transcriptions factors (TFs) control tissue-specific programs of gene expression through interactions with local TF networks. A prime example for a TF network that functions broadly in animals but is modulated in a cell-type specific fashion is the circadian clock. Although the conserved TFs CLOCK (CLK) and CYCLE (CYC or BMAL1) control a core transcriptional circuit throughout animal bodies, downstream clock target genes that give rise to rhythms in behavior and physiology are generated tissue-specifically. Yet, how CLK and CYC determine tissue-specific clock programs have remained unclear. Here, we use a functional genomics approach to determine the *cis*-regulatory requirements for clock specificity. We epitope-tag *Clk* and *cyc* genes in *Drosophila* by homologous recombination and determine their genome-wide binding targets in heads and bodies by ChIP-seq. Both TFs have distinct DNA targets in the two tissue-contexts, suggesting that, in addition to regulating downstream targets indirectly via a hierarchy of TFs, they also bind directly to effector genes in a cell-type specific manner to drive tissue-specific programs. Computational dissection of CLK/CYC context-specific binding sites reveals sequence motifs for putative clock partner factors, which are predictive for individual binding sites. This includes the GATA factor SERPENT (SRP), which is important for correct prediction of body-specific targets and is able to function synergistically with CLK/CYC to activate a body-specific *cis*-regulatory region. This suggests that SRP helps determining direct CLK/CYC targets and is responsible for orchestrating tissue-specific clock outputs. Our results reveal a critical role of SRP in modulating tissue-specific CLK/CYC binding and function, and reveal how universal clock circuits can regulate tissue-specific rhythms. Overall, our study provide insights into the mechanism by which universal TFs can be modulated to drive tissue-specific programs and demonstrates an approach to dissect regulatory interactions more generally.

INTRODUCTION

The basis for multicellular life is the ability to create functional specialization and distinct cell-types through differentially controlled gene expression. This is achieved by intricate gene regulatory networks controlled by transcription factors (TFs) [1]. TFs are regulatory proteins that bind to specific DNA *cis*-regulatory sequences (motifs) within enhancers of target genes, and specify cell-type definite regulatory networks by activating and repressing gene expression. Typically, enhancer activity is determined by defined sets of TFs, such that different TF combinations establish or modulate the activity of individual factors, which are often expressed more broadly and also function in other contexts [2-11].

The circadian clock is an important example of a transcriptional circuit that functions broadly in animals but is modulated in a cell-type specific fashion [12]. Eukaryotic circadian clocks (from the Latin circa diem, meaning 'about a day') are governed by transcriptional negative feedback loops that control daily rhythms in gene expression, ultimately leading to cyclic output rhythms in behavior, metabolism, and physiology [13-15].

In *Drosophila melanogaster*, circadian rhythms are controlled by a network of transcriptional negative feedback loops that are interconnected by the TFs CLOCK (CLK) and CYCLE (CYC) [13,16-18]. In the main loop, they form a heterodimer (CLK/CYC) that binds to E-box sequences (CACGTG) upstream of *period* (*per*) and *timeless* (*tim*) genes. PER and TIM dimerize in the cytoplasm, translocate into the nucleus and feed back on their own regulation by inhibiting CLK/CYC activity [18,19]. Besides being fundamental for the generation and maintenance of the pacemaker mechanism, the core clock mechanism is linked to downstream outputs in part by the CLK/CYC–mediated induction of downstream gene transcription [20-24].

Functional clock circuits are widespread throughout *Drosophila* tissues and specify tissue-specific physiological rhythms by locally generated circadian transcriptional profiles. But, although CLK and CYC are present almost ubiquitously in *Drosophila* [25,26], they produce cell-type specific molecular and physiological outputs [27]. For example, the clock controls rhythms in locomotor activity behavior in the brain [28-30], odor receptivity in antennae [31,32] and expression of metabolic enzymes in the fat body [33,34]. In a broader context, the circadian expression pattern of fly heads vs. bodies was found to be largely non-overlapping: of the 120 and 177 genes that cycled in heads and bodies respectively, only 12 (<10%) cycled in both [22]. Importantly, the existence of tissue-specific differences is widely conserved and has also been noted in mammals, with only a small overlap between genes which cycled in the Suprachiasmatic Nucleus (SCN – the central clock in the mammalian brain), liver, and heart [14,35-37].

Since circadian physiology is tissue-specific and in part controlled by CLK/CYC–mediated gene expression, the cycling genes and functional outcomes in each cell type must be determined by local TFs modulating CLK/CYC and their target spectrum in a context-specific manner. However, such putative partner TFs involved in the circadian clock tissue-specification have remain unclear.

To identify *cis*-regulatory sequence motifs (and the corresponding TFs) that define CLK/CYC context-specific binding and function, we first identified genome-wide CLK and CYC targets in fly heads and bodies by chromatin immunoprecipitation followed by massive parallel sequencing (ChIP-seq). We observed that CLK and CYC shared most of their binding sites in heads and bodies. In addition we observed many different CLK and CYC targets between heads and bodies (i.e. context-specific), and that context-specific binding correlated with tissue-specific functions. We then computationally dissected head- and body-specific binding sites for combinations of sequence motifs of putative partner TFs and identified the motif GATAA to be predictive of context-specific binding in the body. We found that the TF SERPENT (SRP), a key regulator of *Drosophila* endoderm development, as the factor that was able to synergistically activate, along with CLK/CYC, a body specific enhancer. Our results suggest that SRP is an integral component of the downstream clock machinery that shapes tissue-specific expression.

RESULTS

***Clk*^{tag} and *cyc*^{tag} flies**

To enable efficient, stringent, and comparable ChIP for the identification of CLK and CYC binding sites genome-wide, we added a peptide tag including a V5 epitope tag to the 3′ terminus of the *Clk*- and *cyc*-coding sequences at their endogenous genomic *loci* by ends-out homologous recombination [38-40] (see Suppl Fig 1,2 and Methods for details). As increased gene doses of Clk were shown to shorten the locomotor circadian period (probably by increased CLK/CYC-mediated transcription [41]), placing tagged versions of both TFs under the control of their own regulatory regions ensures that they are expressed at physiological levels. In addition, using the same tag for both factors allows ChIPs to be carried out under identical conditions, important for direct quantitative comparisons. We isolated one homozygous viable knock-in line for each *locus* (*Clk*^{tag} and *cyc*^{tag}), for which we confirmed the correct integration by Southern blot and PCR analyses (Suppl Fig 1,2).

Clk^{tag} and cyc^{tag} flies are homozygous viable and do not display any obvious phenotype. To confirm that tagging the endogenous *loci* did also not affect the circadian clock machinery, we analyzed the locomotor activity rhythms of homozygous Clk^{tag} and cyc^{tag} flies, which is the most well characterized *Drosophila* clock output. Behavioral monitoring showed that tagged lines have regular locomotor rhythms with peaks of activity around dawn and dusk and anticipated to lights-on and -off, similar to w^{-} flies that have the same genetic background. To prove that the anticipatory behaviors were not related to the lights-on and -off transitions, we quantified anticipation at the first two days of constant darkness, and confirmed that both tagged lines showed similar morning and evening behavior compared to w^{-} flies (Suppl Fig 3). Clk^{tag} flies showed few arrhythmic flies (13%), and period slightly longer than w^{-} controls (26.63hs compared to 23.23hs). All cyc^{tag} flies were rhythmic and free-run in constant darkness with a period of 24.04hs (Suppl Table 1). These results suggest that the addition of the tag did not significantly affect CLK and CYC function, such that

Clk^{tag} and cyc^{tag} lines were suitable to investigate genome-wide CLK and CYC targets.

Genome-wide DNA-binding profile of CLK and CYC

To identify CLK and CYC DNA binding sites genome-wide, we performed ChIP-seq using a V5 antibody from heads and bodies of homozygous Clk^{tag} and cyc^{tag} decapitated flies at Zeitgeber Time (ZT) 13, at which CLK binding is close to its maximum [42] (ZT0 = lights on; ZT12 = lights off). We sequenced ChIP and input samples, as well as "mock" samples for which we performed ChIP with the V5 antibody from wildtype w^- flies (not carrying the V5 tag) for two independent biological replicates each, summing to 24 libraries in total (Suppl Table 2). The biological replicates from independent fly collections were highly similar with genome-wide Pearson correlation coefficients (PCC) of the normalized read densities between 0.83 and 0.88 (Suppl Fig 4 and Suppl Table 3). We compared each ChIP to the corresponding input sample to identify binding sites (*peaks*) using peakzilla at stringent thresholds and additionally required that both biological replicates concurred (as in [43]; see Methods). This yielded 2059 and 436 candidate peaks for CLK and CYC in heads versus 431 and 484 in bodies, respectively. In contrast, the ChIP from wildtype w^- flies (mock) resulted in only 62 versus 7 candidates, respectively, demonstrating the specificity of the V5 antibody and indicating that the identified candidate peaks had low false discovery rates (FDRs) between 1 and 14 percent. To exclude that antibody cross-reactivity with DNA binding proteins influenced our results, we corrected candidate peaks based on the mock results, yielding 1959 final binding sites for CLK in heads, 369 for CYC in heads, 425 for CLK in bodies, and 481 for CYC in bodies (Suppl Table 4).

Among the CLK binding sites reported previously [42], 58.2% and 59.1% overlapped with CLK heads and CYC heads peaks respectively. The peaks from all four samples agreed well with our expectations: we found multiple peaks in the vicinity of all known core clock components *per*, *tim*, *vri*, *cwo* and *Pdp1*, where they often coincided with previously described E-boxes [44] (Fig 1a and Suppl Fig 5). Within these regions, CLK and CYC have strong overlapping signals in both heads and bodies, which is in agreement with their role as heterodimers in the clock pacemaker [18,45,46]. As expected, CLK and CYC peaks for head and body samples were similarly enriched in 5'UTRs, promoters and intronic regions, while they were depleted in coding DNA sequences (CDS, Fig 1b and Suppl Fig 6).

Besides binding to the core clock components, CLK and CYC are known to bind to a number of regulatory sequences of downstream targets genes [42]. Indeed, the most enriched 6mer motifs in CLK and CYC binding sites in head and body samples recovered the established canonical CLK/CYC E-box motif as expected, but interestingly, also two additional E-box motif versions that were as frequent and some others less abundant (Fig 1c, Suppl Fig 7). Collectively, these E-boxes are found in more than 88% of binding sites in all samples, and show distinct enrichment at the peak summit (Fig 1d). Taken together, these results suggest that the CLK and CYC ChIP identified CLK and CYC binding sites at potential regulatory regions in head and body samples genome-wide.

CLK and CYC exhibit context-specific binding sites

To better understand the specific roles of the circadian clock in different tissues, we analyzed the overlap between CLK and CYC binding sites in heads vs. bodies, and defined shared and differential peaks across the different conditions (see Methods).

Consistent with their known role in the circadian pacemaker, CLK and CYC binding profiles were highly similar (Fig 2 a,b), showing genome-wide PCC above 0.88 (Suppl Table 3). This suggests that CLK and CYC work primarily as heterodimers in activating gene expression of not only the core clock components, but also of downstream targets. In contrast, the binding profiles between heads and bodies differed substantially for both factors (Fig 2c,d,e). We observed that more than 30% of the peaks found in heads were specific to heads, and more than 20% of peaks found in bodies were specific to bodies (e.g. Fig 2e). Genome-wide PCCs confirmed this evident disparity (PCCs between 0.38 and 0.72, below the PCC of the corresponding inputs [0.89]), a striking difference given the concordance of biological replicates within head and body samples (Suppl Table 3). The discrepancy in binding observed here is consistent with previous studies which showed distinctive head and body clock gene expression [22], suggesting that tissue-specific differences in cycling genes might be at least partly caused by direct differential CLK and CYC binding.

To test this hypothesis, we defined stringent classes of CLK and CYC binding sites that are shared across all conditions, head-specific, or body-specific (Fig 3a; leaving a large “gray zone” of sites not assigned to any class, see Methods), assigned each peak to the closest gene transcriptional start site (TSS) and assessed the genes’ functions according to gene ontology (GO; [47]). Shared peaks showed high enrichments for terms related to gene regulation, such as “regulation of transcription” and “transcription factor activity”, suggesting that CLK/CYC are at the top of a gene regulatory hierarchy across tissues. Indeed, 19.8% of the peaks in the shared sample lie close to TFs, a 5-fold enrichment compared to all genes (*P*-value = $2.36\text{x}10^{-09}$). In contrast, we observed striking differences between the other sets: CLK/CYC head-specific peaks were enriched in gene functions related to behavior and vision such as “regulation of synaptic transmission” and “detection of light stimulus”, while CLK/CYC body-specific peaks were consistent with functions in metabolism (e.g. “nitrogenase activity” and “xanthine dehydrogenase activity”, Fig 3b and Suppl Fig 8 for the full list). To test if these differences were also reflected at the expression level, we compared CLK/CYC head- and body-specific binding sites with previously described datasets of cycling mRNAs [22]. We found genes close to head-specific peaks to be 3.39 fold enriched in genes cycling in heads (*P*-value < 0.05) while depleted in genes cycling in the body (0.69 fold). The opposite is also true: genes close to body-specific peaks were 5.16 fold enriched in genes cycling in bodies (*P*-value = 0.057) while we did not observe any body peaks close in genes cycling in the heads, suggesting that tissue-specific differences in cycling genes is at least partly caused by direct CLK and CYC binding. Taken together these results show that, in addition to their role as master regulators in a TF hierarchy, CLK/CYC also binds directly to downstream targets in a cell-type specific manner to drive tissue-specific programs.

Differential motif-content is predictive of context-specific CLK and CYC binding sites

Although CLK and CYC binding sites are highly enriched for E-boxes sequences (Fig 1d, Suppl Fig 7), the presence of this motif alone cannot explain how in each cell type (and its distinct *trans*-regulatory TF environment) CLK and CYC are recruited to tissue-specific *cis*-regulatory targets. Indeed, compelling evidence suggests that TF context-specific binding is achieved by partner TFs that help to define binding targets [3,4,8,10,48]. Therefore, to understand the TF combinations that define tissue-specific CLK and CYC targets, we searched for TF motifs that were differentially enriched between head and body binding sites. We found the "orphan" motif ME50 (predicted by conservation analyses [49]), opa, ME134/odd and Adf1 to be enriched in head-, while depleted in body-specific binding sites and average intergenic genomic sequences. In contrast, the motifs bab1, TATA/Mef2, ME3, GATA/srp and Hox were enriched in body- over head-specific binding and average intergenic genomic sequences, showing that CLK/CYC head and body binding regions have different motif contents (Fig 4a).

To test if combinations of differentially distributed motifs allow the discrimination of head vs. body binding sites, we compared these sequences using a predictive binary classification framework [48]. Using sequence motif content alone, head and body peaks could be accurately distinguished using leave-one-out cross-validation (82% of peaks correctly classified; area under the receiver operating characteristic (ROC) curve (AUC) = 0.909; Fig 4b,c). This indicates that partner TF motifs surrounding CLK/CYC binding sites carry information indicative of binding and have the potential to determine context-specific clock target genes and function. It further suggests that these motifs and their corresponding TFs could have a role in CLK/CYC head vs. body binding sites discrimination and therefore constitute novel clock partner factors.

Identifying the cis-regulatory requirements of individual binding sites

Strikingly, only 5 features (i.e. motif types) were required for the correct prediction of head- and body-binding sites (Fig 4b). To assess the importance of each particular TF motif for each individual CLK/CYC binding site, we tested which motif would affect the classification of the site towards the head or body classes after deleting them *in silico*. As described previously, we calculated a classification score for each CLK/CYC peak and re-scored each peak after deleting all occurrences of a specific TF motif (see [48] and Methods for details). We found that 44% of CLK/CYC binding sites could no longer be confidentially predicted in the head sample after removal of an opa motif, while the deletion of the schlank and Adf1 motifs affected the predictability of 32% and 26% of the peaks, respectively (Fig 4d). opa is an interesting candidate for a CLK/CYC head specific co-factor. The TF associated with the opa motif (OPA) is expressed in the adult *Drosophila* in very few tissues, including the brain [26]. OPA is a pair-rule gene that belongs to the “Zinc finger protein of the cerebellum” (Zic) family of mammalian TFs, with conserved roles in head formation in flies and mammals [50,51]. Further studies on the TFs associated with this and other motifs might provide new insights in the *Drosophila* circadian clock mechanism in the brain.

For CLK/CYC peaks in the body sample, deletion of the GATA motif impaired the predictions of all sites (Fig 4d), suggesting that the GATA motif might be an important determinant of CLK/CYC binding. Consistent with this interpretation, the GATA motif is strongly enriched around CLK/CYC peaks summits in body peaks, while depleted from head peaks (Fig. 4e), providing a tempting hypothesis that it helps to define tissue-specific targets of the peripheral circadian clock.

SRP synergistically activates CLK/CYC mediated expression

The striking enrichment of the GATA motif in CLK/CYC binding sites in bodies and its importance for the correct prediction of body-specific peaks encouraged us to search for a TF that could bind to it. We expected the motif’s consensus sequence GATAA to be recognized by GATA factors, a family of TFs involved in different aspects of animal development and physiology [52,53]. To test if any of the GATA factors could activate CLK/CYC body-specific peaks, we performed transactivation assays by expressing each one of them in *Drosophila* Schneider 2 (S2) cells and measured its ability to trigger enhancer activity of three body-specific CLK/CYC peak sequences using luciferase (LUC) assays. A peak from the 1st intron of the CG34386 gene (hereafter intCG34386) was specifically bound by CLK and CYC in bodies (Fig 5a) and was activated by SRP in our assay (Fig 5b). A second peak showed similar results, while a third one was negative for all GATA factors (Supp Fig 9).

Since many aspects of both fly and mammalian circadian biochemistry can be simulated in S2 cells, including CLK/CYC-mediated activation via E-boxes and PER inhibition of CLK/CYC activity [18,54-59], we decided to use S2 cells to analyze the putative relationship between SRP and the circadian clock.

Because S2 cells express CYC endogenously, Clk transfection is sufficient to activate transcription from E-box containing enhancers [18,54,57]. Interestingly, although intCG34386 contains three E-boxes and was bound by CLK and CYC in bodies, Clk alone did not activate transcription, suggesting that other co-factors might be necessary *in vivo* for CLK/CYC activity. Indeed, Clk with small amounts of srp that were not able to activate the enhancer alone, induced reporter activity 11-fold compared to background (Fig 5c). Additionally, co-transfection of per, the primary inhibitor of CLK/CYC activity and repressor of the circadian pacemaker in *Drosophila* [56], reduced CLK and SRP dependent activation by 57%. Moreover, SRP mediated CLK/CYC activation was dependent on functional wildtype CLK, as CLK^{Jrk} (a truncated version of the CLK protein that lacks polyglutamine repeats required for the transcriptional activity of the CLK/CYC heterodimer [16]), was unable to activate reporter gene expression (Fig 5c). Finally, activation by CLK and SRP was also dependent on the *cis*-regulatory sequence motifs and was abolished when either E-box or GATA motifs were mutated (Fig 5d).

Taken together, these results suggest that SRP acts synergistically with CLK/CYC to induce gene expression. The activation can be repressed by the known clock repressor PER, suggesting that intCG34386 is a clock target enhancer and SRP a novel co-factor of the circadian clock that – together with its expression in the fat body [26,60] – offers an explanation for CLK/CYC context-specific binding and function in the body (Fig 6).

DISCUSSION

Spatial and temporal patterns of mRNA expression define key steps in development and physiology, and are controlled by regulatory networks of TFs (e.g. [6]). Although frequently not restricted to single cell types, individual TFs can control tissue-specific programs of gene expression through interactions with local TF networks [5,8]. The circadian clock is a prime example of this phenomenon: although CLK and CYC are ubiquitously expressed, output rhythms are generated tissue-specifically [14,27]. Despite the substantial progress in identifying differential cell-specific circadian expression programs [22,24,34,61-63], how the central clock tailors cell type-specific expression within the context of different TF networks is still elusive.

Here we used an integrative genomics approach to shed light on how the circadian clock drives tissue-specific gene expression. We built on previous studies that assessed the temporal profile of CLK binding in *Drosophila* heads [42] and determined CLK and CYC binding sites at the time point of maximum binding across two different tissue-contexts. In agreement with earlier observations for CLK targets in fly heads [42] and BMAL1 targets in the mammalian liver [64], CLK/CYC binding sites common to heads and bodies were mainly associated with categories related to the control of gene expression (Fig 3b and Suppl Fig 8), suggesting that the conserved pacemaker control over TFs extends across different tissues. Interestingly however, a substantial number of CLK and CYC binding sites were specific to either heads or bodies, suggesting that differential binding in different tissue-contexts might directly contribute to tissue-specific clock target genes and circadian physiology. Indeed, the genes next to head- and body-specific binding sites differed and displayed different functional categories as revealed by their GO categories: CLK/CYC binding sites in heads were preferentially near genes involved in light perception and neuronal functions while binding sites in bodies were near metabolic genes (Fig 3b and Suppl Fig 8). These results suggest that CLK and CYC have an essential role in directly controlling a variety of output tissue-specific programs by directly binding to tissue-specific downstream target genes. This has important implications for the hierarchy of regulatory connections within the circadian clock as well as the tissue-specific employment of more broadly functioning TFs: it adds a new tier of direct regulatory links to the classical model in which the clock master regulator controls tissue-specific output programs indirectly via tissue-specific TFs. Concurrently, it raises an interesting and important question: how do binding sites and target genes that are specific to heads and bodies differ to allow for context-specific recognition and binding of CLK/CYC ?

We addressed this question by directly comparing the *cis*-regulatory DNA sequences of CLK/CYC binding sites bound specifically in heads with those bound specifically in bodies. Using an approach we developed before [48], we found that head and body binding site sequences differed substantially in their motif content: while both were highly enriched for E-box motifs bound by CLK/CYC itself, they displayed strong differential enrichment for motifs of other TFs. Importantly, the differential motif distribution was sufficient for predicting whether a site was bound in heads or bodies from the sequence alone, suggesting that tissue-specific regulatory targets of the clock are directly encoded in the binding site sequences (Fig 4).

For body-specific sequences, we found the SRP GATA motif to be the most important determinant for CLK/CYC binding sites. Together with CLK, SRP was indeed able to synergistically activate a CLK/CYC body-specific enhancer, suggesting that it is an important determinant of clock function in peripheral tissues. SRP is known to have multiple functions in *Drosophila*, ranging from the control of endodermal development and hematopoiesis in the embryo to the induction of immune response in the larval fat body [65-69]. The function of srp in the adult fly is less well understood, but its expression peaks in the fat body [26,60] and, although it was not detected to cycle [22,34], its *locus* is bound by CLK/CYC specifically in bodies (Suppl Fig 10). The fly fat body has many roles such as regulation of metabolic activity, innate immunity response and detoxification [70-72], and these physiological outputs were shown to be linked to the circadian pacemaker [33,73-75]. Interestingly, we observed a 4.17 fold enrichment (*P*-value < 0.00001) of CLK body-specific peaks close to genes identified in the fly fat body circadian transcriptome, while CLK head-specific targets were only 2.22 fold enriched (*P*-value < 0.0005) [34]. Taken together, these results suggest that *srp* might serve as a link between the many physiological outputs controlled by the fat body and the central pacemaker, by recruiting CLK/CYC binding to drive enhancer activity in the *Drosophila* peripheral clock.

It is likely that different co-factors with functions equivalent to *srp* exist in different cell-types and across species, which redirect broadly expressed core TFs such as CLK and CYC to tissue-specific binding sites and allow tissue-specific gene regulation. This is reminiscent of studies showing that TFs downstream of signaling pathways are redirected in a tissue-specific manner by cell-specific master regulators [2,10,11]. Similarly, across developmental time points both in the *Drosophila* embryo and in the mammalian B-cell lineage, TFs are being redirected to context-specific binding sites presumably by partner TFs [6,8,48]. Along theses lines, our results might have broader implications; it constitutes an important example of how partner TFs adapt more general transcriptional regulators to achieve tissue-specific gene expression and function, contributing to a better understanding of gene regulatory networks in general.

Our data on CLK/CYC binding in different contexts not only provides novel insights into clock regulatory networks and enhancer structure, but also exemplifies a new strategy to uncover additional co-factors of the circadian clock via their *cis*-regulatory motifs. Such co-factors are of high interest and our approach is complementary to forward and reverse genetics or biochemistry, which have been used traditionally to reveal clock factors. Further, the strategy used here can be more generally applied to identify partner factors that recruit broadly expressed TFs to modulate their activity across different cell-types or tissues. In this context, our choice to add an epitope tag to the endogenous *loci* of *Clk* and *cyc* by homologous recombination is noteworthy. Working with characterized antibody-epitope interactions allows for highly specific ChIP under standardized and comparable conditions, including the possibility to stringently control for antibody cross-reactivity using flies that do not carry the tag. It is also applicable for TFs for which the creation of ChIP-grade antibodies proves to be difficult or impossible. Specifically, the tagging of endogenous *loci* allows the study of TFs under physiological conditions in their endogenous expression domains, which is crucial especially for TFs that have large and complex regulatory regions and/or for which physiological expression levels are of fundamental importance.

In summary, our results on the *Drosophila* circadian clock reveal how universal TF circuits can be modulated to generate transcriptional tissue-specific outputs and demonstrate a novel approach to determine novel regulatory partners more generally.

MATERIALS AND METHODS

Drosophila stocks and behavioral analysis

Drosophila flies were raised on standard food at 25 °C and under 12 hours light: 12 hours dark (LD) cycles. w^{1118} (w^-) flies were obtained from the Bloomington Stock Center. For the locomotor activity experiments, we fist video tracked flies using pySolo [76] for 7 days under LD conditions followed by 10 days under constant darkness (DD) at 25°C. Raw data was transformed into a readable MATLAB (MathWorks) format, and data was analyzed using a signal processing toolbox, where autocorrelation and spectral analysis were used to assess rhythmicity and to estimate the period [77].

Constructs for homologous recombination into the *Clk* and *cyc* loci

We performed ends-out homologous recombination (HR) using a methodology reported previously [40] with some modifications that allowed the P[acman] vector [39] to be used as a ends-out targeting vector. Briefly, we removed the mini-white gene from P[acman] and introduced the multiple cloning site (MCS) flanked by FRT and I-SceI sites, besides the UAS-Rpr module from the pRK2 vector [40]. We used this new vector to retrieve by gap repair a large DNA fragment containing the *Clk* (genomic coordinates - chr3L: 7,751,684-7,774,283) and *cyc* (chr3L: 19,801,514-19,822,425) *loci*. We then used recombineering [39] to insert a cassette at the 3' end of each gene. This cassette was modified from the pRK2 vector and contains the tag and the mini-white gene flanked by LoxP sites. We injected these constructs in ZH-attP-51D flies, which carry a landing site in the 2nd chromosome [78]. Genetic crosses for creation of HR lines and removal of the mini-white gene by Cre-recombinase were conducted as described [40]. HR positive flies were subsequently backcrossed to the w^- strain to avoid genetic background effects.

Chromatin immunoprecipitation (ChIP) and library preparation for Solexa sequencing

ChIP was performed based on several previous protocols [42,79,80](Carla E. Margulies, Andreas G. Ladurner, personal communication). Briefly, 5- to 10-days-old flies were collected 1h after lights-off (ZT13), frozen in liquid nitrogen, sieved to separate heads from bodies and kept in -80^{0}C until processed. 2 ml of fly heads or bodies were grinded in liquid nitrogen with a mortar pestle to a fine powder. Tissue powder was collected in a chilled 15 ml glass Dounce homogenizer (Wheaton) and immediately homogenized 30 times with a loose pestle in 50ml of NE Buffer (15mM HEPES pH 7.6, 10mM KCl, 0.1mM EDTA, 0.5mM EGTA, 350 mM sucrose, 0.1% Tween, 5mM $MgCl_2$, 1mM DTT, 1mM PMSF plus Protease inhibitor cocktail (Roche)) with 1% formaldehyde (Sigma). Homogenization and fixation were done simultaneously to have a more precise snapshot of CLK and CYC binding. After the 30 strokes with the loose pestle, the fixing homogenate was poured to a 50ml Falcon tube and rotated for a total time of 10 min (starting from the addition of the NE Buffer). Fixation was quenched for 5 min at room temperature (RT) by the addition of 2.5 ml glycine (2.5M). Homogenate was filtered through a 60 um Streriflip filter (Millipore) and the nuclei were collected by centrifugation at 800g/5 min/4^{0}C. Nuclei pellet was washed twice in 10 ml cold NE Buffer (without formaldehyde) and twice more with cold RIPA Buffer (25 mM Tris-HCl 7.6, 150 mM NaCl, 0.5% sodium deoxycholate, 0.1% SDS, 1% NP-40, 0.5mM DTT plus Protease inhibitor cocktail (Roche)). Nuclei were ressuspended in 2 ml cold RIPA Buffer and transferred to a 15ml falcon tube for sonication. Nuclei were sonicated on ice using an Omni Sonic Ruptor 250 Watts sonicator six times (heads) or four times (bodies) for 1min (Duty cycle: 30%, Output: 20%) each time, with 2 min intervals between the sonication steps. Sonicated chromatin was transferred to 1.5 ml eppendorfs and centrifuged at 10000g/10 min/4^{0}C. 15 ul of the supernatant were removed for the input sample while 1.5 ml of the supernatant were immediately incubated with 50 ul of blocked V5-agarose beads (Sigma) overnight at 4°C in a rotating wheel. After overnight incubation, beads were washed once with 1.4 ml Low Salt Buffer (20 mM Tris 8.1, 150 mM NaCl, 2mM EDTA, 0.1% SDS, 1% Triton X-100), twice with 1.4 ml High Salt Buffer (20 mM Tris 8.1, 500 mM NaCl, 2mM EDTA, 0.1% SDS, 1% Triton X-100), twice with 1.4 ml LiCl Buffer (10 mM Tris 8.1, 250 mM LiCl, 1mM EDTA, 1% sodium deoxycholate, 1% NP-40) and twice with 1.4 ml TE. Co-immunoprecipitated DNA fragments were eluted from the beads in 50 ul of V5 Elution Buffer (10mM HEPES, 1.5mM MgCl2, 0.25mM EDTA, 20% glycerol, 250mM KCl, 0.3% NP-40, 0.5mg/ml V5 peptide) three times consecutively, 30 min each, and the three elutes were combined (150 ul). The volumes of the V5 eluted chromatin and input were brought up to 500ul with IP Elution Buffer (10 mM Tris 7.5, 1 mM EDTA, 1% SDS, 100 mM $NaHCO_3$, 200 mM NaCl) plus 15 ul of 10mg/ml RNase A, and incubated for 30 min 37^0C. RNA-digested chromatin was incubated overnight at 65^0C with Proteinase K, and in the following day the DNA was purified using standard Phenol:Chloroform extraction. Preparations of DNA libraries for single-end sequencing were done according to the illumina instructions with 36 cycles of extension. Up to 5ng ChIP or input DNA was used in each preparation. All Solexa sequencing data from this study are deposited in GEO under the accession code GSE40467.

Reads processing, peaks calling and comparisons across conditions

We mapped the reads to the *Drosophila melanogaster* genome reference dm3 (not chrU, chrUextra) obtained from UCSC [81,82] using bowtie [83], allowing only for uniquely mapping reads with up to 3 mismatches. For each sample, we calculated the read density at each genomic position normalized to 1 million reads per library from reads extended to 150bp (average estimated length of the genomic fragments). We identified peak regions in ChIP samples compared to the corresponding input samples using peakzilla (http://www.starklab.org/data/peakzilla/). Confident peaks are selected with a score ≥ 10 and a fold enrichment of ChIP over input ≥ 2 and enriched regions with a score ≥ 2 and a fold enrichment of ChIP over input ≥ 2. We further discard confident peaks in the CLK or CYC samples if they overlap an enriched region in the corresponding mock (w^-) sample with a fold enrichment of CLK/CYC ChIP over mock ≥ 2. We defined peaks as shared between replicates or conditions if confident peaks called in one sample are overlapping an enriched region in another sample and reciprocally. We defined peaks as differential between conditions if confident peaks shared in both replicates for one sample are not overlapping or close to (+/- 150bp) an enriched region of any of the replicates of the other sample. Using this strategy, some peaks are neither shared nor differential, i.e., were not assigned to shared, head- nor body-specific, creating a "gray zone" of peaks that were excluded from further analysis.

Peak to gene assignment and Gene Ontology analysis

We used the Flybase genome annotation (r5.33) [82] and functional categories from Gene Ontology [47]. We first assigned peaks in any condition (CLK heads, CYC heads, CLK bodies and CYC bodies) to their closest genes' TSS and defined classes of genes that were shared in all, head-specific or body-specific. For each functional GO categories, we calculated the enrichment and associated hypergeometric *P*-values of genes in each class compared to all genes assigned to peaks in any condition, which is a more stringent assessment than compared to all genes. We selected categories with a *P*-value ≤ 0.03 in at least one condition.

Motif analyses and SVM prediction

We scanned the genome for all possible 6mers motifs and calculated the enrichment in peak regions (151bp around confident peak summit) compared to control regions (peaks shifted to random locations within the same chromosomes). For the motif enrichment analysis, we search for motif instances by using the known and predicted motifs from [49] with a position weight matrix (PWM) cutoff of *P*-value ≤0.00097 (1/1024) within a region of 401bp centered on the ChIP-seq peak summit. The distribution of E-box and GATA motifs relative to the distance to the peak summit was calculated by the sum of all motif instances (with a PWM cutoff of *P*-value ≤0.00097) within a window of 200bp normalized by the total number of binding sites within the set. The motif distribution was calculated every 100bp starting from 1Kb away from the ChIP-seq peak summit. The SVM approach was performed as in [48]. In brief, by taking as attributes the number of motif instances for each region and a manually implemented Leave-

One-Out Cross-validation for SVM light [22,84], we predicted head- and body-specific binding sites, calculated a prediction score and the score drop after *in silico* mutations. The AUC was computed by the R package ROCR [85]. For all the motif analysis, we took the region with the highest ChIP-seq score if two or more regions overlapped within a condition and we excluded all overlapping regions between conditions. For the SVM predictions, we also excluded regions that overlapped with the core promoter (defined as ±50 bp from the TSS).

Luciferase assays

The wildtype enhancers were amplified from *D. melanogaster* genomic DNA and cloned into a luciferase expressing vector (modified from Promega pGL4.26) under the control of the DSCP minimal promoter [86]. The intCG34386 region (Genomic coordinates: chr2R: 13976708-13977307) was amplified using the primers 5' AACGTAGCAACAATCGTGTTT 3’ and 5' AACGTAGCAACAATCGTGTTT 3’. For the intCG9009 construct (Genomic coordinates: chrX: 14843115-14843714) we used the primers 5' ATAGCTGGACGCGATTCATT 3’ and 5' ATTCGCCGTCTGTCTGTGT 3’and for the construct intCG9864 (Genomic coordinates: chr2R: 16138193-16138792) we used the primers 5' GGCTGCTTAGTTAGCACTGTCAT 3’ and 5' CATGCCACGGGTCAATTAT 3’. Enhancers containing mutated versions of E-box (CACGTG, CATGTG and CACATG to CGATCG) and GATA (GATAA to CGAGA) sequences were synthesized at GeneArt and cloned in the same vector. For the normalization we cloned the ubiquitin enhancer into a modified pRL vector (Promega). The GATA factors pnr, grn and GATAe were amplified from DGRC plasmids and cloned into pAHW (Invitrogene), while GATAd was also amplified from a DGRC plasmid but cloned into pAW (Invitrogene). The srp sequence was amplified from a plasmid from the *Drosophila* TF open reading frame library [87] and cloned into pAHW. pAct-CLK and pAct-PER were a kind gift from Frank Weber and their cloning was previously described [18]. The Clk^{Jrk} DNA sequence [16] was cloned into pAct5.1 (Invitrogene). Drosophila S2 cells were maintained in Schneider’s *Drosophila* culture medium (Gibco) supplemented with 10% fetal bovine serum (Invitrogen) and 1%

streptomycin/penicillin antibiotics at 25^0C. One day before transfection, 10^5 cells/well were seeded in a ninety six-well plate. 10 ng of renilla vector, 10 ng of luciferase reporter vector and indicated amounts of TF vectors were transiently transfected using jetPEI (Polyplus) following manufacturer's instructions. The total amount of transfected DNA was kept constant by supplementing empty pBluescript II vector (Stratagene). Cell lysates were prepared after 48 h and Firefly luciferase (FF-luc) and Renilla luciferase (FR-luc) activities were measured using a dual luciferase reporter assay (Promega). The luminescence signals were measured using a Biotek Synergy 2 plate reader.

ACKNOWLEDGEMENTS

We would like to thank Michaela Pagani, Katharina Schernhuber, Martina Rath and IMP/IMBA services for technical support. We thank Daria Shlyueva for help with the *cyc* Southern blot (Suppl Fig 1b), Christian Machacek for help with pySolo hardware and software and Johannes Tkadletz for help drawing Fig 6 and Suppl Figs 1 and 2. We are grateful to Korneel Hens, Bart Deplancke (Laboratory of Systems Biology and Genetics, LSBG, Lausanne) and Frank Weber (Biochemistry Center, BZH, Heidelberg University) for plasmids and to Carla Margulies and Andreas Ladurner (Ludwig Maximilians University of Munich) for sharing the chromatin preparation protocol before publication. Illumina sequencing was performed at the CSF NGS Unit (http://csf.ac.at/).

AUTHOR CONTRIBUTIONS

The authors have made the following declarations about their contributions: Conceived and designed the experiments: ACAMF AS. Performed the experiments: ACAMF AFB JOYC GS. Analyzed the data: ACAMF AFB JOYC GS AS. Wrote the paper: ACAMF AS.

REFERENCES

1. Stathopoulos A, Levine MS (2005) Genomic regulatory networks and animal development. Developmental Cell 9: 449–462. doi:10.1016/j.devcel.2005.09.005.
2. Zeitlinger J, Simon I, Harbison CT, Hannett NM, Volkert TL, et al. (2003) Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. Cell 113: 395–404.
3. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, MacIsaac KD, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431: 99–104. doi:10.1038/nature02800.
4. Buck MJ, Lieb JD (2006) A chromatin-mediated mechanism for specification of conditional transcription factor targets. Nat Genet 38: 1446–1451. doi:10.1038/ng1917.
5. Lupien M, Eeckhoute J, Meyer CA, Wang Q, Zhang Y, et al. (2008) FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell 132: 958–970. doi:10.1016/j.cell.2008.01.018.
6. Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EEM (2009) Combinatorial binding predicts spatio-temporal cis-regulatory activity. Nature 462: 65–70. doi:10.1038/nature08531.
7. Wilczynski B, Furlong EEM (2010) Dynamic CRM occupancy reflects a temporal map of developmental progression. Mol Syst Biol 6: 383. doi:10.1038/msb.2010.35.
8. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, et al. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular Cell 38: 576–589. doi:10.1016/j.molcel.2010.05.004.

9. Palii CG, Perez-Iratxeta C, Yao Z, Cao Y, Dai F, et al. (2011) Differential genomic targeting of the transcription factor TAL1 in alternate haematopoietic lineages. EMBO J 30: 494–509. doi:10.1038/emboj.2010.342.

10. Trompouki E, Bowman TV, Lawton LN, Fan ZP, Wu D-C, et al. (2011) Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration. Cell 147: 577–589. doi:10.1016/j.cell.2011.09.044.

11. Mullen AC, Orlando DA, Newman JJ, Lovén J, Kumar RM, et al. (2011) Master transcription factors determine cell-type-specific responses to TGF-β signaling. Cell 147: 565–576. doi:10.1016/j.cell.2011.08.050.

12. Zhang EE, Kay SA (2010) Clocks not winding down: unravelling circadian networks. Nat Rev Mol Cell Biol 11: 764–776. doi:10.1038/nrm2995.

13. Hardin PE (2011) Molecular genetic analysis of circadian timekeeping in Drosophila. Adv Genet 74: 141–173. doi:10.1016/B978-0-12-387690-4.00005-2.

14. Mohawk JA, Green CB, Takahashi JS (2012) Central and Peripheral Circadian Clocks in Mammals. Annu Rev Neurosci. doi:10.1146/annurev-neuro-060909-153128.

15. Vanin S, Bhutani S, Montelli S, Menegazzi P, Green EW, et al. (2012) Unexpected features of Drosophila circadian behavioural rhythms under natural conditions. Nature. doi:10.1038/nature10991.

16. Allada R, White NE, So WV, Hall JC, Rosbash M (1998) A mutant Drosophila homolog of mammalian Clock disrupts circadian rhythms and transcription of period and timeless. Cell 93: 791–804.

17. Rutila JE, Suri V, Le M, So WV, Rosbash M, et al. (1998) CYCLE is a second bHLH-PAS clock protein essential for circadian rhythmicity and transcription of Drosophila period and timeless. Cell 93: 805–814.

18. Darlington TK, Wager-Smith K, Ceriani MF, Staknis D, Gekakis N, et al. (1998) Closing the circadian loop: CLOCK-induced transcription of its own inhibitors per and tim. Science 280: 1599–1603.

19. Lee C, Bae K, Edery I (1999) PER and TIM inhibit the DNA binding activity of a Drosophila CLOCK-CYC/dBMAL1 heterodimer without disrupting formation of the heterodimer: a basis for circadian transcription. Mol Cell Biol 19: 5316–5325.

20. McDonald M, Rosbash M (2001) Microarray analysis and organization of circadian gene expression in Drosophila. Cell 107: 567–578.

21. Claridge-Chang A, Wijnen H, Naef F, Boothroyd C, Rajewsky N, et al. (2001) Circadian Regulation of Gene Expression Systems in the Drosophila Head. Neuron 32: 657–671. doi:10.1016/S0896-6273(01)00515-3.

22. Ceriani MF, Hogenesch JB, Yanovsky M, Panda S, Straume M, et al. (2002) Genome-wide expression analysis in Drosophila reveals genes controlling circadian behavior. Journal of Neuroscience 22: 9305–9319.

23. Ueda HR, Matsumoto A, Kawamura M, Iino M, Tanimura T, et al. (2002) Genome-wide transcriptional orchestration of circadian rhythms in Drosophila. J Biol Chem 277: 14048–14052. doi:10.1074/jbc.C100765200.

24. Keegan KP, Pradhan S, Wang J-P, Allada R (2007) Meta-analysis of Drosophila circadian microarray studies identifies a novel set of rhythmically expressed genes. PLoS Comput Biol 3: e208. doi:10.1371/journal.pcbi.0030208.

25. Plautz JD, Kaneko M, Hall JC, Kay SA (1997) Independent photoreceptive circadian clocks throughout Drosophila. Science 278: 1632–1635.

26. Chintapalli VR, Wang J, Dow JAT (2007) Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat Genet 39: 715–720. doi:10.1038/ng2049.

27. Tomioka K, Uryu O, Kamae Y, Umezaki Y, Yoshii T (2012) Peripheral circadian rhythms and their regulatory mechanism in insects and some other arthropods: a review. J Comp Physiol B, Biochem Syst Environ Physiol. doi:10.1007/s00360-012-0651-1.

28. Grima B, Chélot E, Xia R, Rouyer F (2004) Morning and evening peaks of activity rely on different clock neurons of the Drosophila brain. Nature 431: 869–873. doi:10.1038/nature02935.

29. Stoleru D, Peng Y, Agosto J, Rosbash M (2004) Coupled oscillators control morning and evening locomotor behaviour of Drosophila. Nature 431: 862–868. doi:10.1038/nature02926.

30. Stoleru D, Peng Y, Nawathean P, Rosbash M (2005) A resetting signal between Drosophila pacemakers synchronizes morning and evening activity. Nature 438: 238–242. doi:10.1038/nature04192.

31. Krishnan B, Dryer S, Hardin PE (1999) Circadian rhythms in olfactory responses of Drosophila melanogaster. Nature 400: 375–378.

32. Tanoue S, Krishnan P, Krishnan B, Dryer SE, Hardin PE (2004) Circadian clocks in antennal neurons are necessary and sufficient for olfaction rhythms in Drosophila. Curr Biol 14: 638–649. doi:10.1016/j.cub.2004.04.009.

33. Xu K, Zheng X, Sehgal A (2008) Regulation of feeding and metabolism by neuronal and peripheral clocks in Drosophila. Cell Metab 8: 289–300. doi:10.1016/j.cmet.2008.09.006.

34. Xu K, Diangelo JR, Hughes ME, Hogenesch JB, Sehgal A (2011) The circadian clock interacts with metabolic physiology to influence reproductive fitness. Cell Metab 13: 639–654. doi:10.1016/j.cmet.2011.05.001.

35. Akhtar RA, Reddy AB, Maywood ES, Clayton JD, King VM, et al. (2002) Circadian cycling of the mouse liver transcriptome, as revealed by cDNA microarray, is driven by the suprachiasmatic nucleus. Curr Biol 12: 540–550.

36. Panda S, Antoch MP, Miller BH, Su AI, Schook AB, et al. (2002) Coordinated transcription of key pathways in the mouse by the circadian clock. Cell 109: 307–320.

37. Storch K-F, Lipan O, Leykin I, Viswanathan N, Davis FC, et al. (2002) Extensive and divergent circadian gene expression in liver and heart. Nature 417: 78–83. doi:10.1038/nature744.

38. Gong WJ, Golic KG (2003) Ends-out, or replacement, gene targeting in Drosophila. Proc Natl Acad Sci USA 100: 2556–2561. doi:10.1073/pnas.0535280100.

39. Venken KJT, He Y, Hoskins RA, Bellen HJ (2006) P[acman]: a BAC transgenic platform for targeted insertion of large DNA fragments in D. melanogaster. Science 314: 1747–1751. doi:10.1126/science.1134426.

40. Huang J, Zhou W, Watson AM, Jan Y-N, Hong Y (2008) Efficient ends-out gene targeting in Drosophila. Genetics 180: 703–707. doi:10.1534/genetics.108.090563.

41. Kadener S, Menet JS, Schoer R, Rosbash M (2008) Circadian transcription contributes to core period determination in Drosophila. Plos Biol 6: e119. doi:10.1371/journal.pbio.0060119.

42. Abruzzi KC, Rodriguez J, Menet JS, Desrochers J, Zadina A, et al. (2011) Drosophila CLOCK target gene characterization: implications for circadian tissue-specific gene expression. Genes Dev 25: 2374–2386. doi:10.1101/gad.178079.111.

43. Bardet AF, He Q, Zeitlinger J, Stark A (2012) A computational pipeline for comparative ChIP-seq analyses. Nat Protoc 7: 45–61. doi:10.1038/nprot.2011.420.

44. Taylor P, Hardin PE (2008) Rhythmic E-box binding by CLK-CYC controls daily cycles in per and tim transcription and chromatin modifications. Mol Cell Biol 28: 4642–4652. doi:10.1128/MCB.01612-07.

45. Cyran SA, Buchsbaum AM, Reddy KL, Lin M-C, Glossop NRJ, et al. (2003) vrille, Pdp1, and dClock form a second feedback loop in the Drosophila circadian clock. Cell 112: 329–341.

46. Matsumoto A, Ukai-Tadenuma M, Yamada RG, Houl J, Uno KD, et al. (2007) A functional genomics strategy reveals clockwork orange as a transcriptional regulator in the Drosophila circadian clock. Genes Dev 21: 1687–1700. doi:10.1101/gad.1552207.

47. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29. doi:10.1038/75556.

48. Yanez-Cuna JO, Dinh HQ, Kvon EZ, Shlyueva D, Stark A (2012) Uncovering cis-regulatory sequence requirements for context specific transcription factor binding. Genome Res. doi:10.1101/gr.132811.111.

49. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, et al. (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450: 219–232. doi:10.1038/nature06340.

50. Aruga J (2004) The role of Zic genes in neural development. Mol Cell Neurosci 26: 205–221. doi:10.1016/j.mcn.2004.01.004.

51. Lee H, Stultz BG, Hursh DA (2007) The Zic family member, odd-paired, regulates the Drosophila BMP, decapentaplegic, during adult head development. Development 134: 1301–1310. doi:10.1242/dev.02807.

52. Patient RK, McGhee JD (2002) The GATA family (vertebrates and invertebrates). Curr Opin Genet Dev 12: 416–422.

53. Gillis WQ, Bowerman BA, Schneider SQ (2008) The evolution of protostome GATA factors: molecular phylogenetics, synteny, and intron/exon structure reveal orthologous relationships. BMC Evol Biol 8: 112. doi:10.1186/1471-2148-8-112.

54. Ceriani MF, Darlington TK, Staknis D, Más P, Petti AA, et al. (1999) Light-dependent sequestration of TIMELESS by CRYPTOCHROME. Science 285: 553–556.

55. Shearman LP, Sriram S, Weaver DR, Maywood ES, Chaves I, et al. (2000) Interacting molecular loops in the mammalian circadian clock. Science 288: 1013–1019.

56. Chang DC, Reppert SM (2003) A novel C-terminal domain of drosophila PERIOD inhibits dCLOCK:CYCLE-mediated transcription. Curr Biol 13: 758–762.

57. Weber F, Hung H-C, Maurer C, Kay SA (2006) Second messenger and Ras/MAPK signalling pathways regulate CLOCK/CYCLE-dependent transcription. J Neurochem 98: 248–257. doi:10.1111/j.1471-4159.2006.03865.x.

58. Maurer C, Hung H-C, Weber F (2009) Cytoplasmic interaction with CYCLE promotes the post-translational processing of the circadian CLOCK protein. FEBS Letters 583: 1561–1566. doi:10.1016/j.febslet.2009.04.013.

59. Hung H-C, Maurer C, Zorn D, Chang W-L, Weber F (2009) Sequential and compartment-specific phosphorylation controls the life cycle of the circadian CLOCK protein. J Biol Chem 284: 23734–23742. doi:10.1074/jbc.M109.025064.

60. Senger K, Harris K, Levine M (2006) GATA factors participate in tissue-specific immune responses in Drosophila larvae. Proc Natl Acad Sci USA 103: 15957–15962. doi:10.1073/pnas.0607608103.

61. Nagoshi E, Sugino K, Kula E, Okazaki E, Tachibana T, et al. (2010) Dissecting differential gene expression within the circadian neuronal circuit of Drosophila. Nat Neurosci 13: 60–68. doi:10.1038/nn.2451.

62. Kula-Eversole E, Nagoshi E, Shang Y, Rodriguez J, Allada R, et al. (2010) Surprising gene expression patterns within and between PDF-containing circadian neurons in Drosophila. Proceedings of the National Academy of Sciences 107: 13497–13502. doi:10.1073/pnas.1002081107.

63. Hughes ME, Grant GR, Paquin C, Qian J, Nitabach MN (2012) Deep sequencing the circadian and diurnal transcriptome of Drosophila brain. Genome Res 22: 1266–1281. doi:10.1101/gr.128876.111.

64. Rey G, Cesbron F, Rougemont J, Reinke H, Brunner M, et al. (2011) Genome-wide and phase-specific DNA-binding rhythms of BMAL1 control circadian output functions in mouse liver. Plos Biol 9: e1000595. doi:10.1371/journal.pbio.1000595.

65. Rehorn KP, Thelen H, Michelson AM, Reuter R (1996) A molecular aspect of hematopoiesis and endoderm development common to vertebrates and Drosophila. Development 122: 4023–4031.

66. Sam S, Leise W, Hoshizaki DK (1996) The serpent gene is necessary for progression through the early stages of fat-body development. Mech Dev 60: 197–205.

67. Petersen UM, Kadalayil L, Rehorn KP, Hoshizaki DK, Reuter R, et al. (1999) Serpent regulates Drosophila immunity genes in the larval fat body through an essential GATA motif. EMBO J 18: 4013–4022. doi:10.1093/emboj/18.14.4013.

68. Lebestky T, Chang T, Hartenstein V, Banerjee U (2000) Specification of Drosophila hematopoietic lineage by conserved transcription factors. Science 288: 146–149.

69. Hayes SA, Miller JM, Hoshizaki DK (2001) serpent, a GATA-like transcription factor gene, induces fat-cell development in Drosophila melanogaster. Development 128: 1193–1200.

70. Arrese EL, Soulages JL (2010) Insect fat body: energy, metabolism, and regulation. Annu Rev Entomol 55: 207–225. doi:10.1146/annurev-ento-112408-085356.

71. Leclerc V, Reichhart J-M (2004) The immune response of Drosophila melanogaster. Immunol Rev 198: 59–71.

72. Lazareva AA, Roman G, Mattox W, Hardin PE, Dauwalder B (2007) A role for the adult fat body in Drosophila male courtship behavior. PLoS Genet 3: e16. doi:10.1371/journal.pgen.0030016.

73. Shirasuhiza M, Dionne M, Pham L, Ayres J, Schneider D (2007) Interactions between circadian rhythm and immunity in Drosophila melanogaster. Curr Biol 17: R353–R355. doi:10.1016/j.cub.2007.03.049.

74. Krishnan N, Davis AJ, Giebultowicz JM (2008) Circadian regulation of response to oxidative stress in Drosophila melanogaster. Biochemical and Biophysical Research Communications 374: 299–303. doi:10.1016/j.bbrc.2008.07.011.

75. Lee J-E, Edery I (2008) Circadian Regulation in the Ability of Drosophila to Combat Pathogenic Infections. Curr Biol 18: 195–199. doi:10.1016/j.cub.2007.12.054.

76. Gilestro GF (2012) Video tracking and analysis of sleep in Drosophila melanogaster. Nat Protoc 7: 995–1007. doi:10.1038/nprot.2012.041.

77. Levine JD, Funes P, Dowse HB, Hall JC (2002) Signal analysis of behavioral and molecular cycles. BMC neuroscience 3: 1.

78. Bischof J, Maeda RK, Hediger M, Karch F, Basler K (2007) An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases. Proc Natl Acad Sci USA 104: 3312–3317. doi:10.1073/pnas.0611511104.

79. Sandmann T, Jakobsen JS, Furlong EEM (2006) ChIP-on-chip protocol for genome-wide analysis of transcription factor binding in Drosophila melanogaster embryos. Nat Protoc 1: 2839–2855. doi:10.1038/nprot.2006.383.

80. Schebesta A, McManus S, Salvagiotto G, Delogu A, Busslinger GA, et al. (2007) Transcription factor Pax5 activates the chromatin of key genes involved in B cell signaling, adhesion, migration, and immune function. Immunity 27: 49–63. doi:10.1016/j.immuni.2007.05.019.

81. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996–1006. doi:10.1101/gr.229102.

82. McQuilton P, St Pierre SE, Thurmond J, FlyBase Consortium (2012) FlyBase 101--the basics of navigating FlyBase. Nucleic Acids Research 40: D706–D714. doi:10.1093/nar/gkr1030.

83. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25. doi:10.1186/gb-2009-10-3-r25.

84. Joachims T (1998) Making large-scale SVM learning practical. Advances in Kernel Methods-Support Vector Learning, Cambridge.

85. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. Bioinformatics 21: 3940–3941. doi:10.1093/bioinformatics/bti623.

86. Pfeiffer BD, Jenett A, Hammonds AS, Ngo T-TB, Misra S, et al. (2008) Tools for neuroanatomy and neurogenetics in Drosophila. Proceedings of the National Academy of Sciences 105: 9715–9720. doi:10.1073/pnas.0803697105.

87. Hens K, Feuz J-D, Isakova A, Iagovitina A, Massouras A, et al. (2011) Automated protein-DNA interaction screening of Drosophila regulatory elements. Nat Methods 8: 1065–1070. doi:10.1038/nmeth.1763.

FINANCIAL DISCLOSURE

A.C.A.M.F. was supported by an EMBO long-term postdoctoral fellowship and A.F.B and J.O.Y.C by the Austrian Ministry for Science and Research through the GEN-AU Bioinformatics Integration Network III. A.S. is supported by a European Research Council (ERC) Starting Grant from the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement number 242922. Basic research at the IMP is supported by Boehringer Ingelheim GmbH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

COMPETING INTEREST

The authors have declared that no competing interests exist.

FIGURE LEGENDS

Figure 1: Genome-wide DNA-binding profile of CLK and CYC. (a) CLK and CYC read densities in head and body samples at known clock gene targets. UCSC genome browser tracks represent binding of CLK in head (red) and body (blue) and CYC in head (orange) and body (green) at known clock genes. Gray columns highlight the positions of known enhancers [44]. (b) Genomic location of CLK head peaks relative to the genome. Distribution of the peaks of other samples can be found in Suppl Fig 6. (c) CLK and CYC binding sites are enriched for the E-box motif. Shown are the three 6mers most enriched in CLK head peaks (other samples can be found in Suppl Fig 7). (d) The E-box is enriched at both CLK- and CYC-bound sites in heads and bodies. Distribution of E-box motif instances across a 900 bp window (x-axis) centered on the ChIP-seq peak summit for all CLK head (red), CYC head (orange), CLK body (blue) and CYC body (green) significant binding sites.

Figure 2: Context-specific CLK and CYC binding sites. (a) Scatterplot of read densities at peak summits of CLK vs. CYC in heads (a) and in bodies (b) and heads vs. bodies of CLK (c) and CYC (d). (e) Examples of CLK and CYC head- and body-specific binding sites. UCSC genome browser tracks of regions of the CG30497 (left) and CG3277 (right) genes, showing CLK and CYC head- and body-specific peaks, respectively.

Figure 3: Context-specific CLK/CYC function. (a) Heatmaps of CLK and CYC binding in heads and bodies for shared, head-specific and body-specific peaks. For each binding site (y-axis) of each category, read densities are displayed within a 5Kb window centered on the binding site peak summit (x-axis). (b) Context-specific peaks functions. Shared, head-specific or body-specific peaks were assigned to genes and an enrichment analysis was calculated by comparing a gene in each category to genes in all categories. Shown is an unsupervised clustered heatmap of hypergeometric *P*-values of the enrichment analysis, with a selection of ontologies highlighted; see Suppl Fig 8 for the full list with all categories.

Figure 4: Context-specific CLK and CYC binding sites are characterized by specific TF motif content. (a) Heat map showing motifs differentially enriched (*P*-value ≤ 0.01) between head-specific and body-specific CLK and CYC binding sites (first column) and the enrichment of head-specific and body-specific binding sites compare to the genomic average (second and third column). (b) Prediction accuracy (binary classification) of head-specific and body-specific CLK and CYC binding sites solely based on differences in motif content (data). The same procedure was repeated by randomly assigning the binding sites to either the head- or body-specific class (control). (c) Receiver-Operating-Characteristic (ROC) curves and Area Under the Curve (AUC) for the prediction of context-specific binding sites (black) and controls (gray). (d) Only five motifs were required for the correct prediction of context-specific binding. Shown are the percent of body-specific (blue) and head-specific (red) binding sites that were affected by the *in silico* mutation of all motif instances of the individual binding site. (e) Distribution of GATA motif instances relative to the ChIP-seq peak summit for head-specific CLK (red), head-specific CYC (orange), body-specific CLK (blue) and body-speciic CYC (green) binding sites.

Figure 5: SRP synergistically activates CLK/CYC transcription. (a) UCSC genome browser track of a region of the CG34386 gene showing CLK and CYC specific binding in bodies. The red bar below the gene symbolizes the region cloned into the LUC vector. (b) Normalized luciferase activity (Firefly vs. Renilla, FF/FR) of extracts from S2 cells transiently transfected with 10ng of intCG34386-LUC reporter gene and 50 ng of expression vectors containing SRP, GNR, PNR, dGATAd, GATAe or empty vector (-) as indicated. (c) Normalized activity driven by 10ng of intCG34386-LUC with contransfections of 2 ng of CLK or CLK^{Jrk}, and 10 ng of SRP and/or PER as indicated. (d) Mutated versions for the E-Box or GATA motifs of the intCG34386 enhancer were co-transfected with CLK and/or SRP as indicated. Error bars show the standard deviations of one experiment conducted in triplicate (b) or three independent biological replicates conducted in triplicate (c,d).

Figure 6: Model of the context-specific binding of CLK/CYC for the expression of tissue-specific transcriptional programs. Schematic representation of the proposed model. CLK and CYC co-occupies the genome in heads and bodies, but CLK/CYC specific binding in the body is achieved by SRP recruitment, while still unknown TFs would recruit CLK/CYC to head-specific targets.

SUPPORTING INFORMATION

Figure S1: Homologous recombination strategy and confirmation of the tagged *Clk locus*. Schematic representation of the HR strategy (left), the southern blot confirming the correct integration (top right) and the PCR spanning the tagged 3'-terminus of *Clk locus* to confirm the cassette removal (bottom right).

Figure S2: Homologous recombination strategy and confirmation of the tagged *cyc locus*. Schematic representation of the HR strategy (left), the southern blot confirming the correct integration (top right) and the PCR spanning the tagged 3'-terminus of *cyc locus* to confirm the cassette removal (bottom right).

Figure S3: Circadian behavior of tagged flies. Activity profiles display average activity through seven days of LD (12 hr light: 12 hr dark, right) and the first two days in DD (constant darkness, left) for wildtype w^- (top, n=22) w^- ; + ; Clk^{tag} (middle, n=29) and w^- ; + ; cyc^{tag} (bottom, n=30) flies. Vertical white and black bars on the left indicate averaged normalized activity counts during the light (ZT0-ZT12) and dark phase, respectively. On the right, gray and black vertical bars indicate subjective light (CT0–CT12) and subjective dark phase, respectively. Dots above the bars indicate the standard error of the mean (SEM).

Figure S4: Agreement on ChIP-seq biological replicates. Scatterplot of genome-wide read densities comparing biological replicates of CLK heads (a), CYC heads (b), CLK bodies (c) and CYC bodies (d) samples.

Figure S5: Genome-wide DNA-binding profile of CLK and CYC at the *per locus*. UCSC genome browser snapshot of the *per locus* showing tracks representing the IP signal and inputs (depicted in darker colors at the same track

of their corresponding IPs) of CLK heads (red) and bodies (blue), CYC heads (orange) and bodies (green) and the V5 mock signal in heads and bodies (both gray).

Figure S6: Genomic location of peaks relative to the genome. Distribution of CLK heads, CLK bodies, CYC heads, CYC bodies, shared, heads-only and bodies-only peaks in relation to the genome.

Figure S7: E-box motif enrichment. Shown is the percent of CLK-head, CYC-head, CLK-body CYC-body binding sites respectively (black) that contains at least one of the three most enriched 6mers found in CLK and CYC ChIP-seq peaks. The total number of peaks in each category is shown on top. As a control (gray) is shown the percent of fragmentsthat contain each motif in the shuffled peaks.

Figure S8: Shared, heads- and bodies-specific binding sites are enriched in different gene-ontology (GO) categories. Heat map of Fig 3b showing all ontologies found significantly enriched in the Share (left), Head-specific (middle) and Body-specific (right).

Figure S9: SRP activates a CLK/CYC body-specific region. (a) UCSC genome browser track of a region of the CG9009 gene showing CLK and CYC specific binding in bodies. The red bar below the gene symbolizes the region cloned into the luciferase (LUC) vector. (b) FF-FR activity of S2 cells extracts transiently transfected with 10ng of intCG9009-LUC reporter gene and 50 ng of expression vectors containing SRP, GNR, PNR, dGATAd, GATAe or empty vector (-) as indicated. (c) The same as in a, but for the CG9864 *locus*. (d) The same as in b, but with the intCG9864-LUC reporter vector.

Figure S10: Genome-wide DNA-binding profile of CLK and CYC at the *srp locus*. UCSC genome browser snapshot of the *srp locus* showing tracks representing the IP signal of CLK in heads (red) and bodies (blue), CYC in heads (orange) and bodies (green).

Table S1: Circadian locomotor rhythms in tagged flies. Percent of flies with detectable constant darkness (DD) rhythmicity (%R) values and their period were calculated as described (Levine et al., 2002). All values are given as mean ± standard error of the mean. n = number of flies analyzed.

Table S2: Number of reads for each sample. Sequenced raw reads were mapped to the D. melanogaster genome (dm3). All reads that are not located on unassembled sequences (chrU, chrUextra) are used for the analysis. Their percentage to the original read numbers is shown in the last column.

Table S3: Genome wide Pearson correlation coefficients. Pearson correlation coefficient between the genome-wide read densities of biological replicates, CLK vs. CYC, heads vs. bodies and inputs. As expected, the genome-wide profiles of read densities are highly similar for biological replicates and inputs, and also for CLK vs. CYC. On the other hand, heads vs. bodies shows lower correlation.

Table S4: Number of peaks for each sample. Number of peaks called in each replicate (All peaks), the number of peaks that overlap in each condition (Shared peaks), FDR and the final peaks we used in the analysis. Note that we used a very conservative estimation of the FDR, which is the real FDR and the true false positives due to antibody cross-reactivity.

Figure 1 - Meireles-Filho et al.

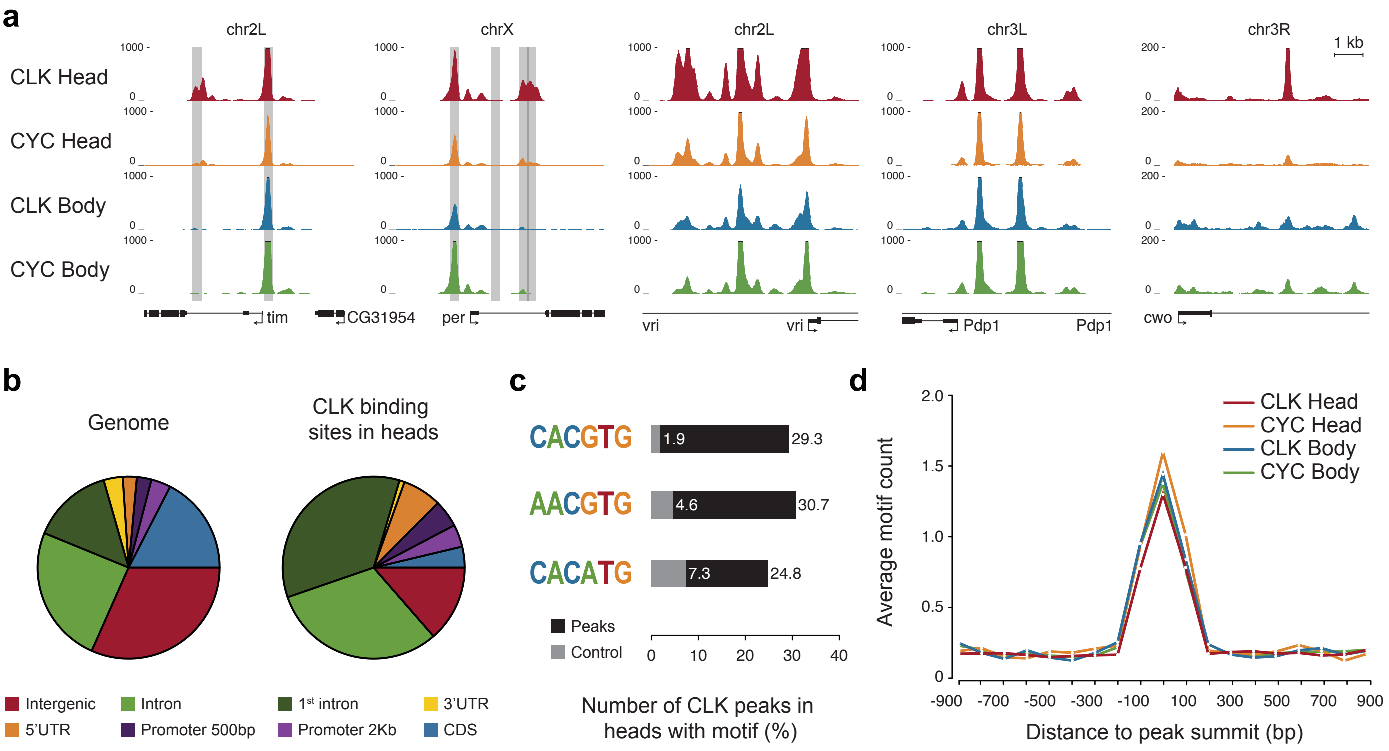


Figure 2 - Meireles-Filho et al.

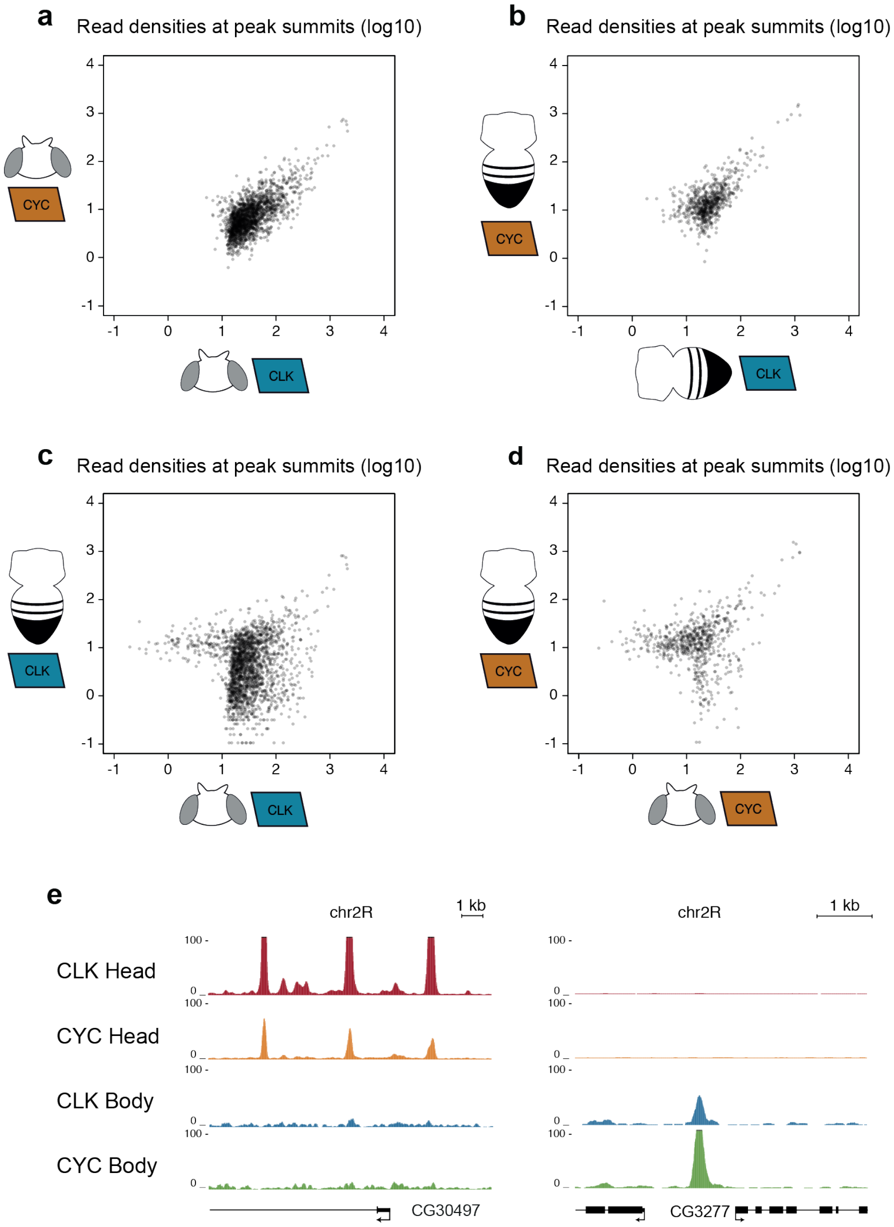
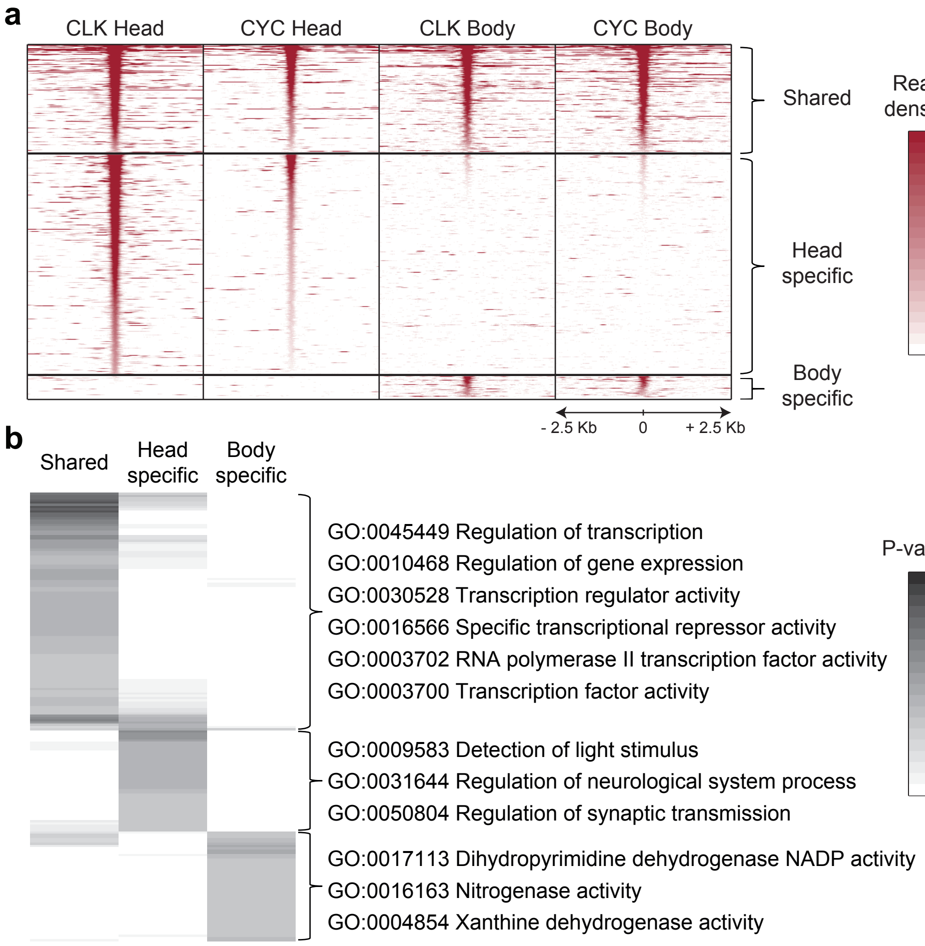


Figure 3 - Meireles-Filho et al.



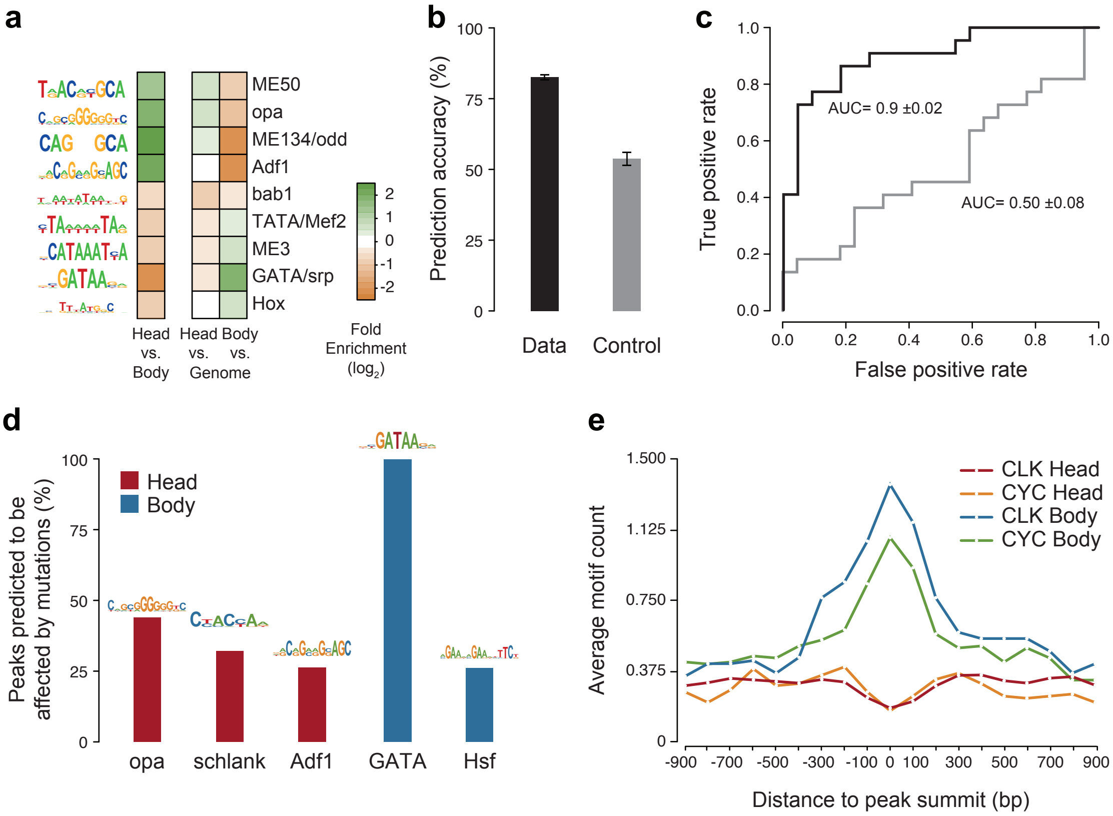
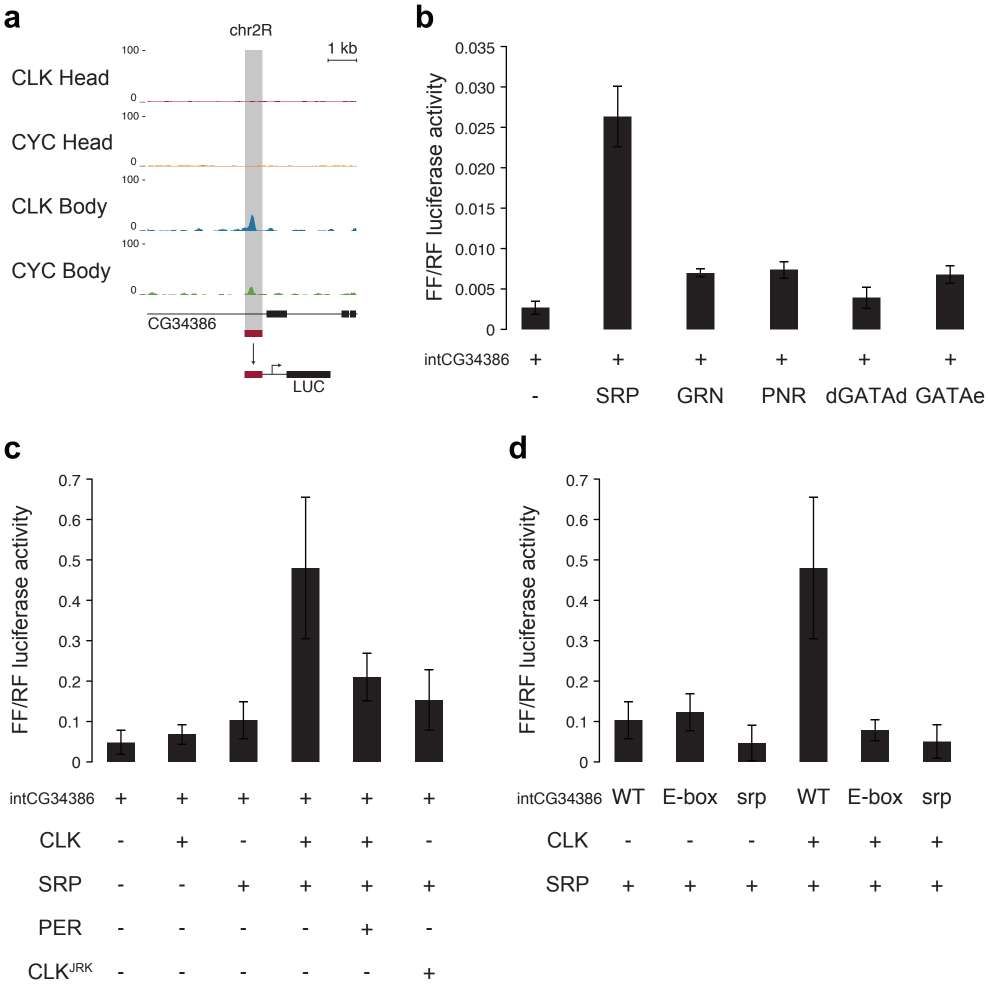

a
ME50
opa
ME134/odd
Adf1
bab1
TATA/Mef2
ME3
GATA/srp
Hox
Head vs. Body
Head vs. Genome
Body vs. Genome
Fold Enrichment (log_2)
2
1
0
-1
-2
b
Prediction accuracy (%)
100
75
50
25
0
Data
Control
c
True positive rate
False positive rate
1.0
0.8
0.6
0.4
0.2
0.0
AUC= 0.9 ±0.02
AUC= 0.50 ±0.08
d
Peaks predicted to be affected by mutations (%)
Head
Body
opa
schlank
Adf1
GATA
Hsf
e
Average motif count
1.500
1.125
0.750
0.375
0
CLK Head
CYC Head
CLK Body
CYC Body
-900
-700
-500
-300
-100
0
100
300
500
700
900
Distance to peak summit (bp)

Figure 5 - Meireles-Filho et al.



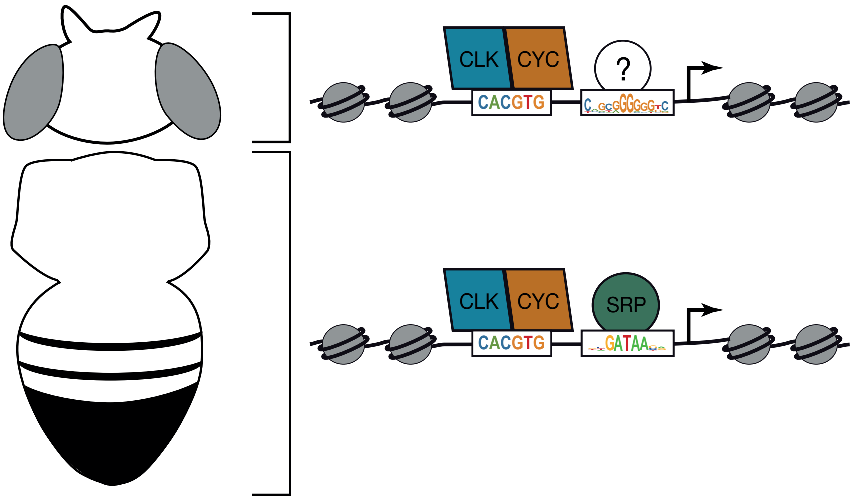
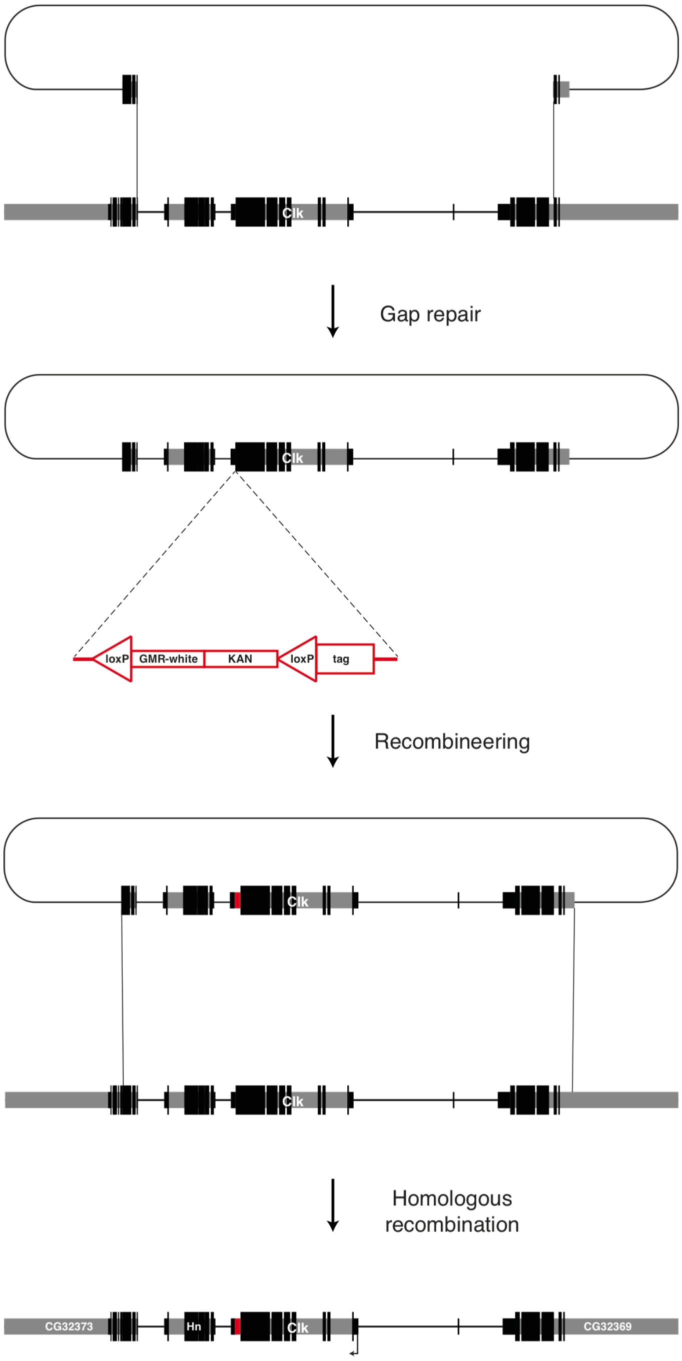
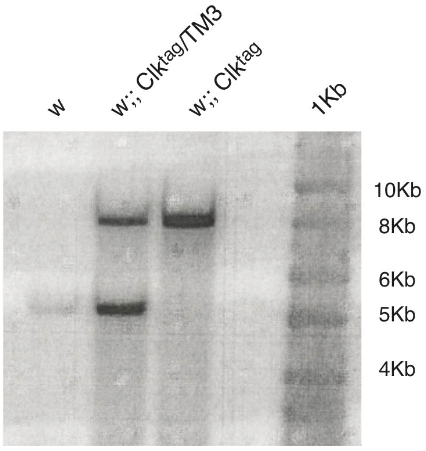

CLK
CYC
?
CACGTG
C GCGGGGGGTC
CLK
CYC
SRP
CACGTG
GATAA

Figure S1 – Meireles-Filho et al





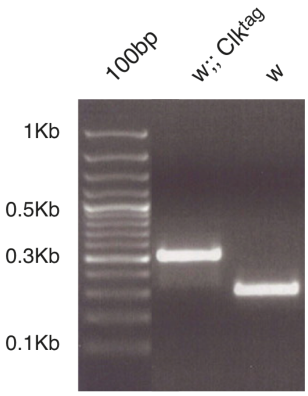


Figure S2 – Meireles-Filho et al

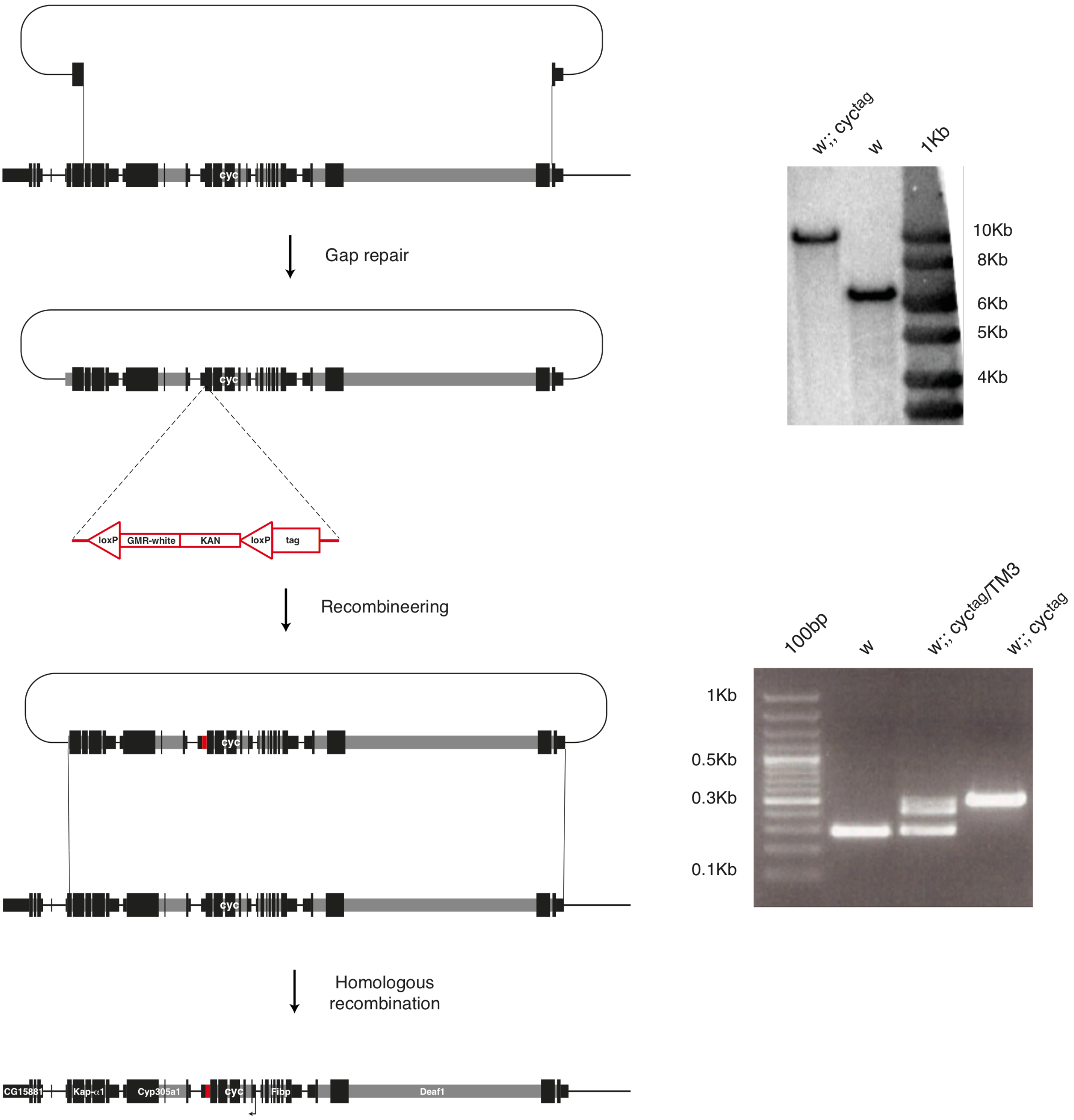
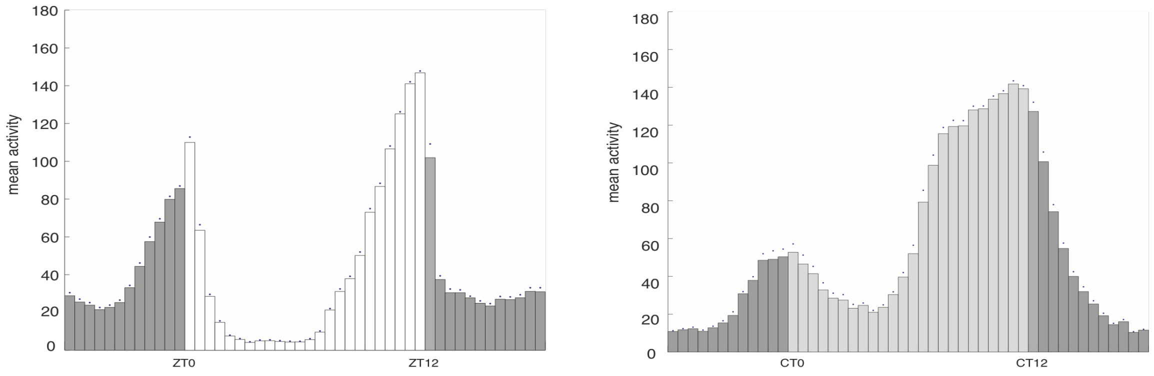
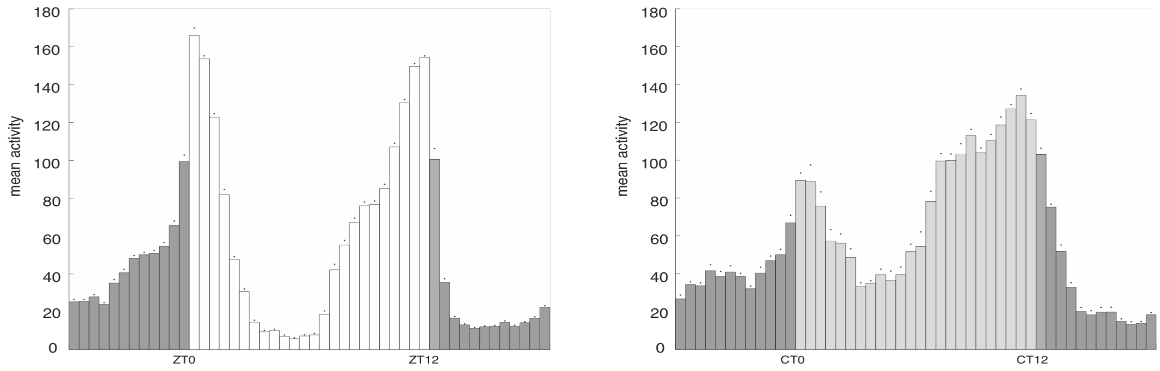


Figure S3 – Meireles-Filho et al

w



w; +; Clk^{tag}



w; +; cyc^{tag}

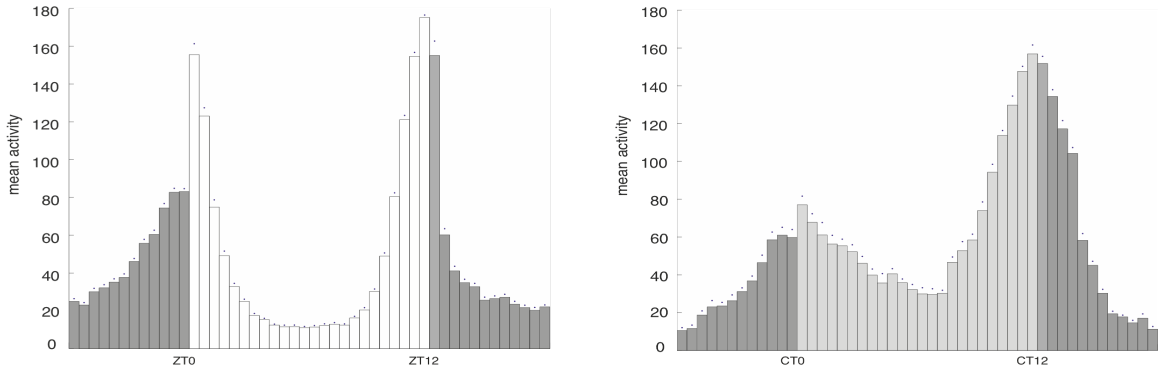
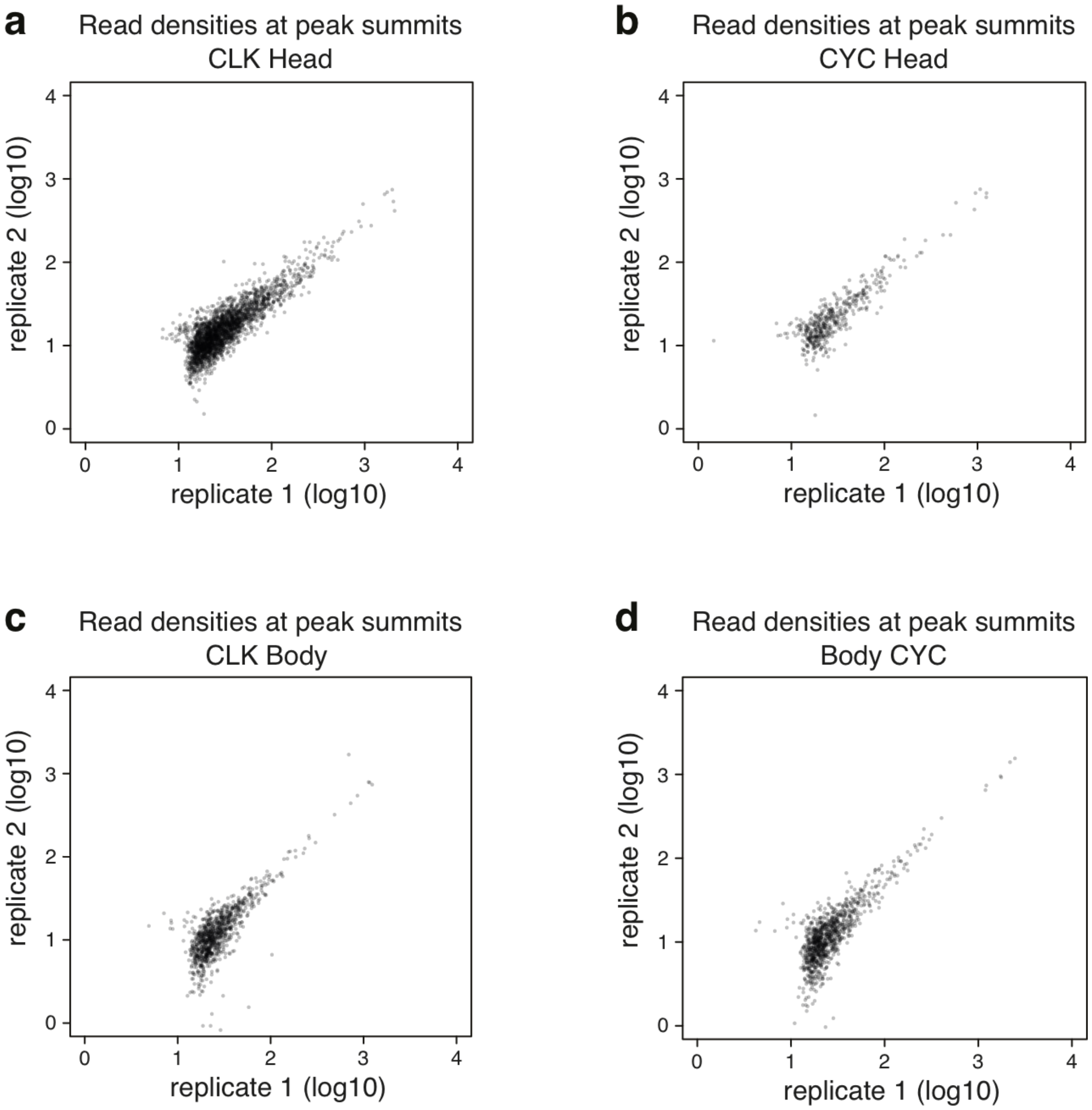


Figure S4 – Meireles-Filho et al



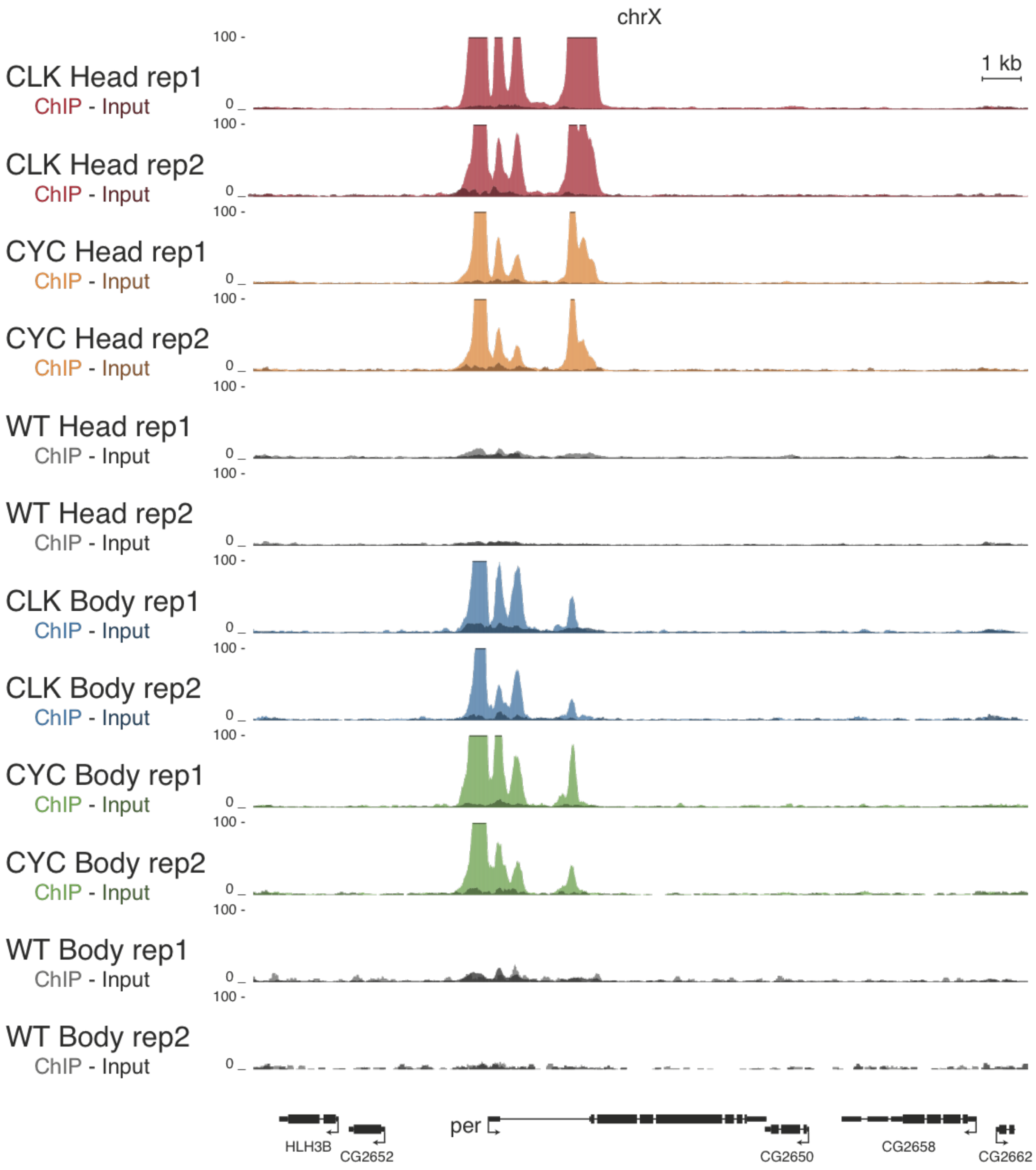

chrX
1 kb
CLK Head rep1
ChIP - Input
CLK Head rep2
ChIP - Input
CYC Head rep1
ChIP - Input
CYC Head rep2
ChIP - Input
WT Head rep1
ChIP - Input
WT Head rep2
ChIP - Input
CLK Body rep1
ChIP - Input
CLK Body rep2
ChIP - Input
CYC Body rep1
ChIP - Input
CYC Body rep2
ChIP - Input
WT Body rep1
ChIP - Input
WT Body rep2
ChIP - Input
100
0
HLH3B
CG2652
per
CG2650
CG2658
CG2662

Figure S6 – Meireles-Filho et al

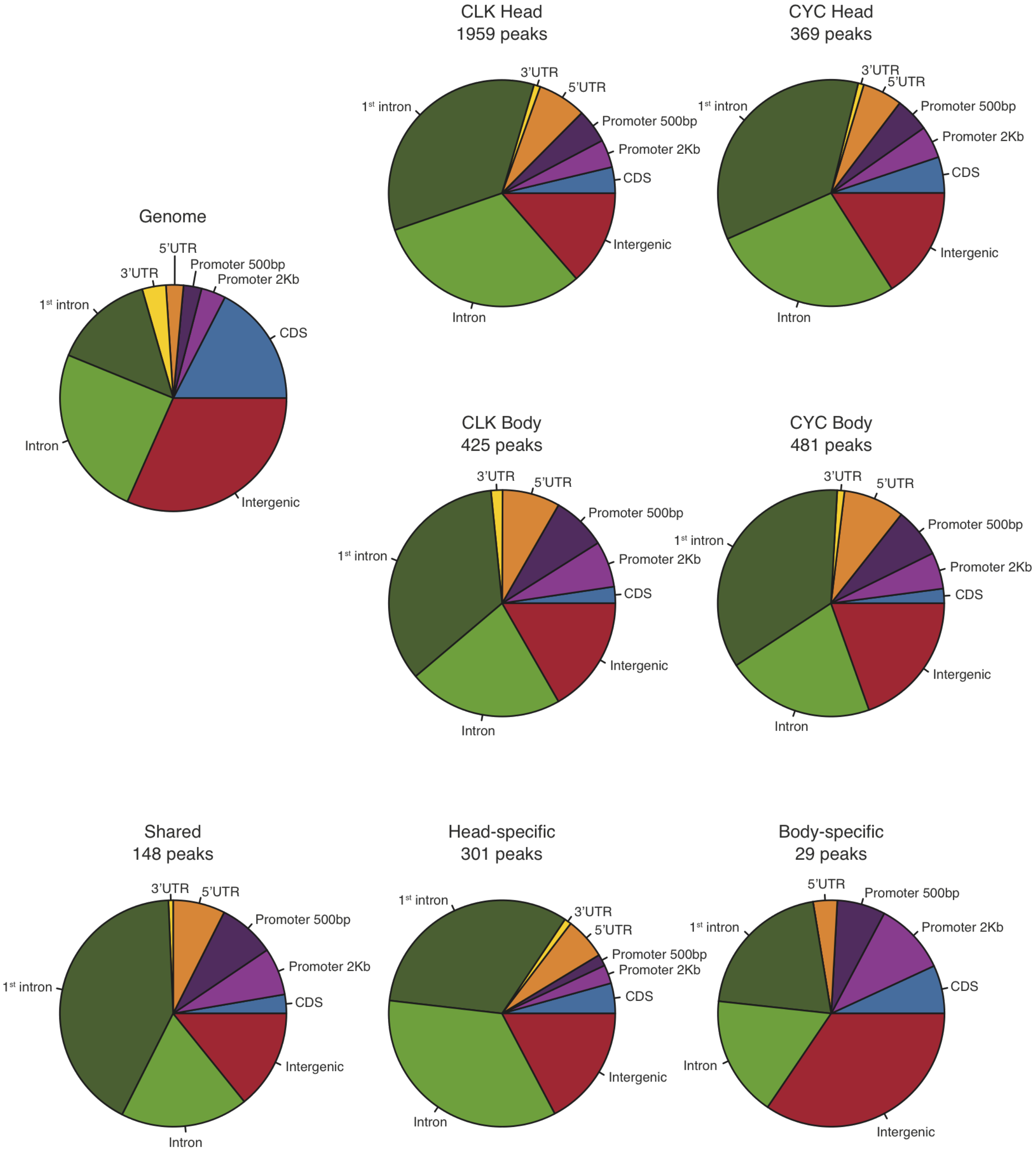


Figure S7 – Meireles-Filho et al

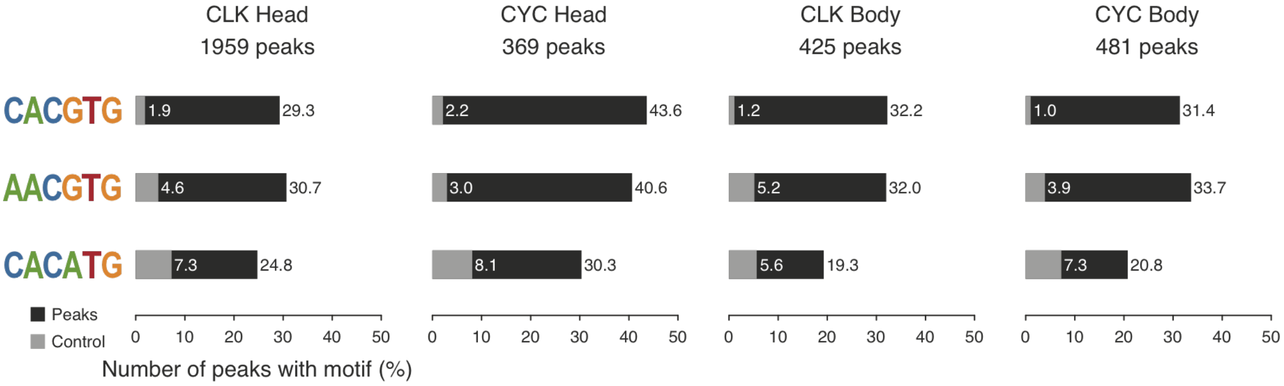
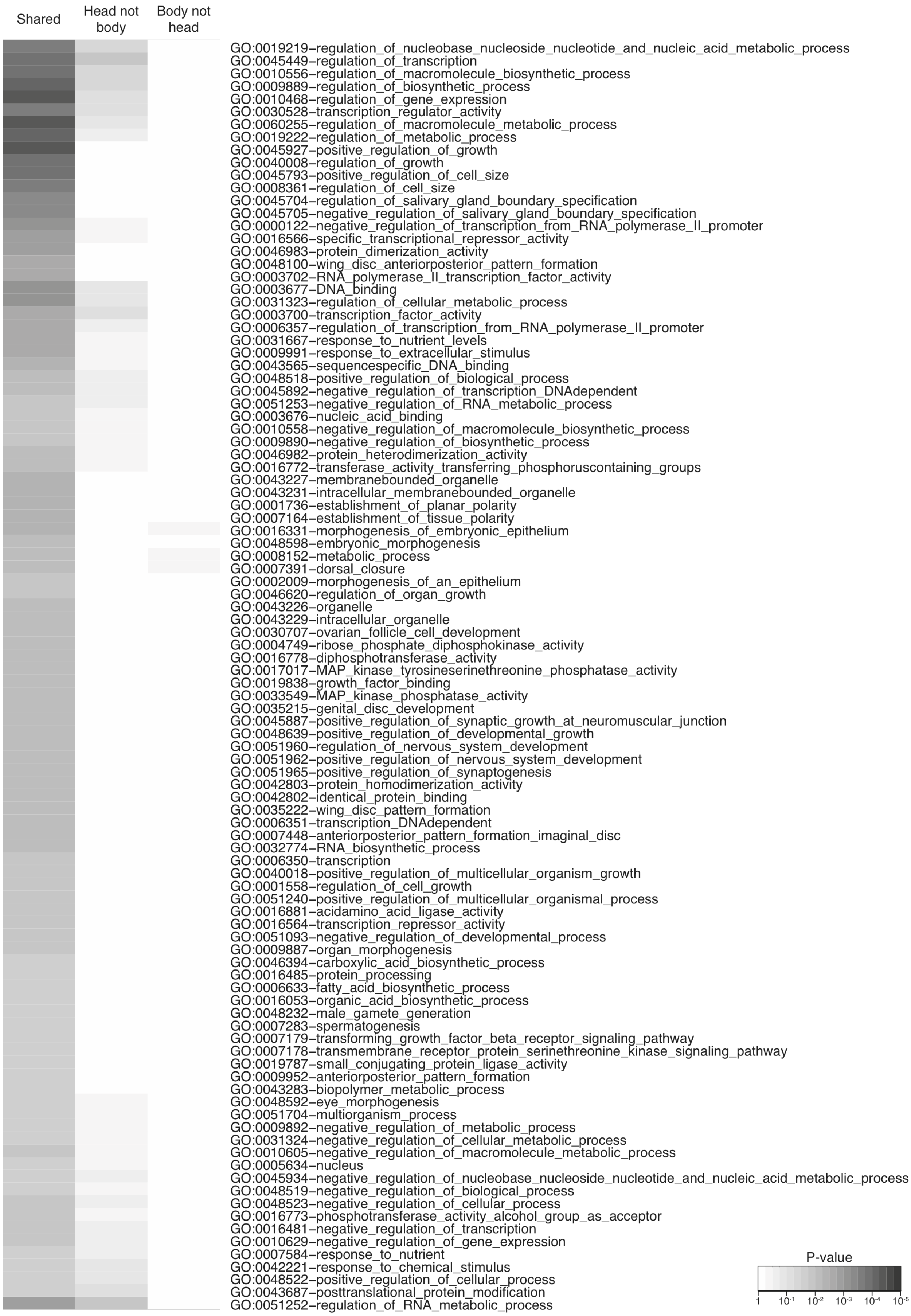


Figure S8 – Meireles-Filho et al



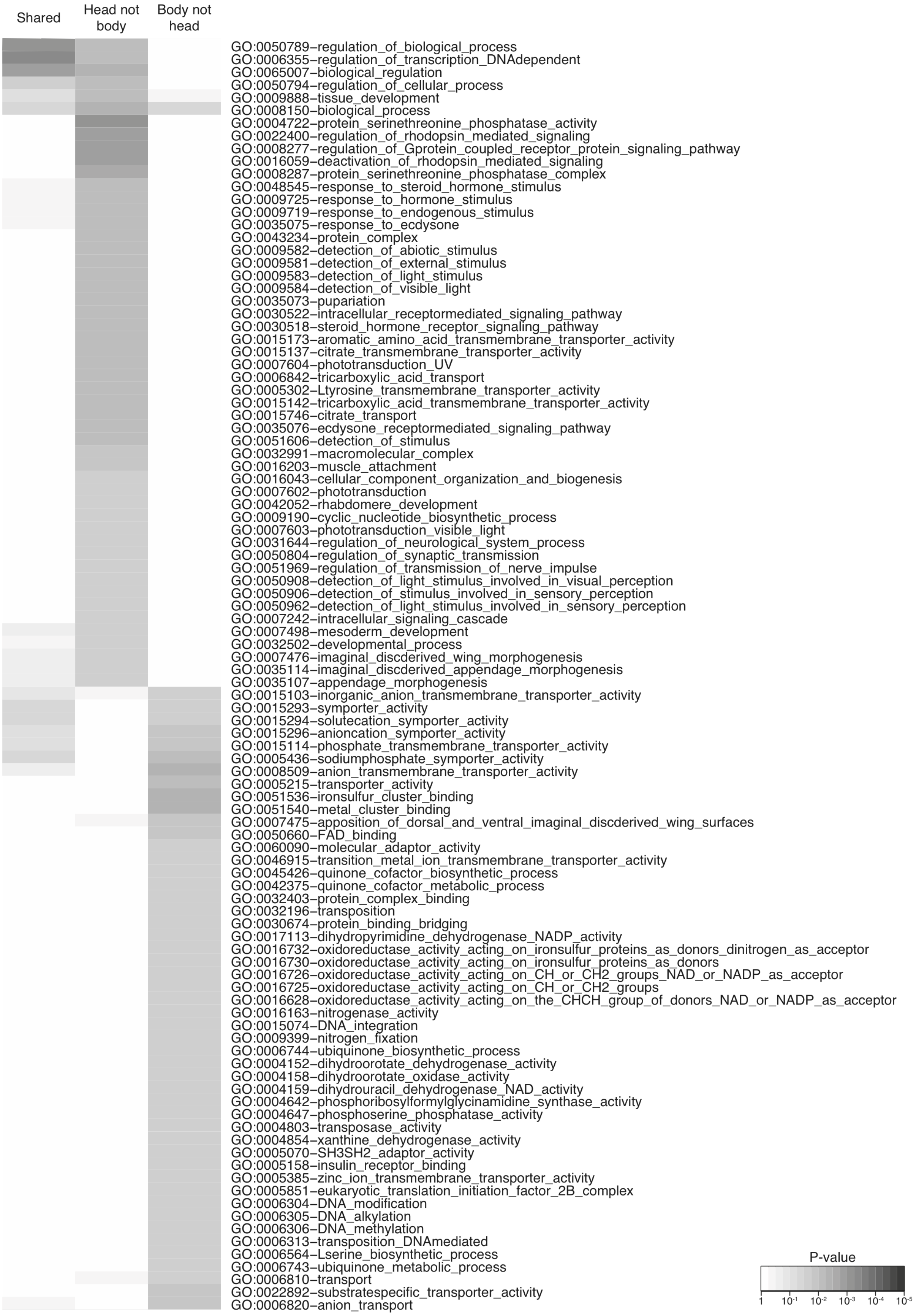
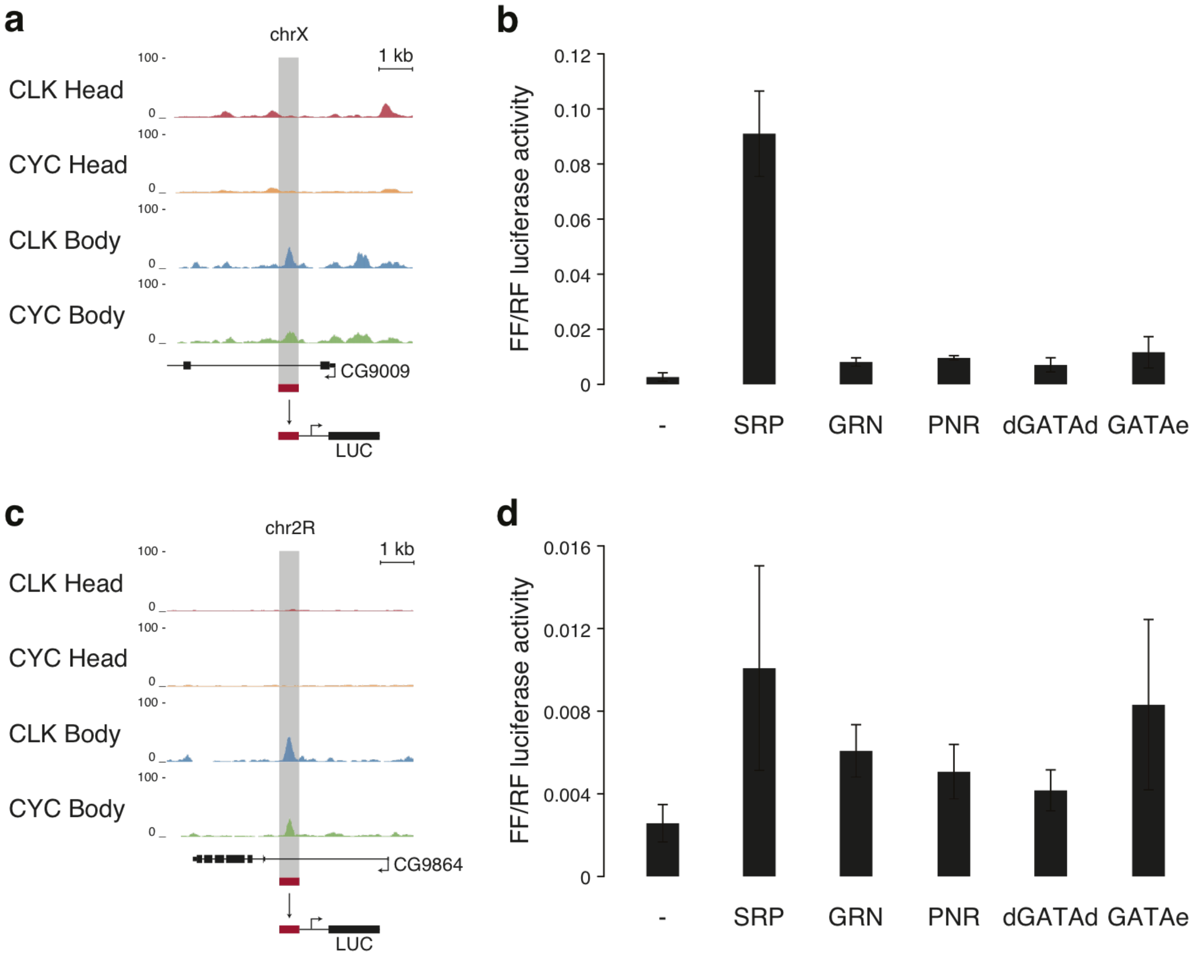

Shared
Head not body
Body not head
GO:0050789–regulation_of_biological_process
GO:0006355–regulation_of_transcription_DNAdependent
GO:0065007–biological_regulation
GO:0050794–regulation_of_cellular_process
GO:0009888–tissue_development
GO:0008150–biological_process
GO:0004722–protein_serinethreonine_phosphatase_activity
GO:0022400–regulation_of_rhodopsin_mediated_signaling
GO:0008277–regulation_of_Gprotein_coupled_receptor_protein_signaling_pathway
GO:0016059–deactivation_of_rhodopsin_mediated_signaling
GO:0008287–protein_serinethreonine_phosphatase_complex
GO:0048545–response_to_steroid_hormone_stimulus
GO:0009725–response_to_hormone_stimulus
GO:0009719–response_to_endogenous_stimulus
GO:0035075–response_to_ecdysone
GO:0043234–protein_complex
GO:0009582–detection_of_abiotic_stimulus
GO:0009581–detection_of_external_stimulus
GO:0009583–detection_of_light_stimulus
GO:0009584–detection_of_visible_light
GO:0035073–pupariation
GO:0030522–intracellular_receptormediated_signaling_pathway
GO:0030518–steroid_hormone_receptor_signaling_pathway
GO:0015173–aromatic_amino_acid_transmembrane_transporter_activity
GO:0015137–citrate_transmembrane_transporter_activity
GO:0007604–phototransduction_UV
GO:0006842–tricarboxylic_acid_transport
GO:0005302–Ltyrosine_transmembrane_transporter_activity
GO:0015142–tricarboxylic_acid_transmembrane_transporter_activity
GO:0015746–citrate_transport
GO:0035076–ecdysone_receptormediated_signaling_pathway
GO:0051606–detection_of_stimulus
GO:0032991–macromolecular_complex
GO:0016203–muscle_attachment
GO:0016043–cellular_component_organization_and_biogenesis
GO:0007602–phototransduction
GO:0042052–rhabdomere_development
GO:0009190–cyclic_nucleotide_biosynthetic_process
GO:0007603–phototransduction_visible_light
GO:0031644–regulation_of_neurological_system_process
GO:0050804–regulation_of_synaptic_transmission
GO:0051969–regulation_of_transmission_of_nerve_impulse
GO:0050908–detection_of_light_stimulus_involved_in_visual_perception
GO:0050906–detection_of_stimulus_involved_in_sensory_perception
GO:0050962–detection_of_light_stimulus_involved_in_sensory_perception
GO:0007242–intracellular_signaling_cascade
GO:0007498–mesoderm_development
GO:0032502–developmental_process
GO:0007476–imaginal_discderived_wing_morphogenesis
GO:0035114–imaginal_discderived_appendage_morphogenesis
GO:0035107–appendage_morphogenesis
GO:0015103–inorganic_anion_transmembrane_transporter_activity
GO:0015293–symporter_activity
GO:0015294–solutecation_symporter_activity
GO:0015296–anioncation_symporter_activity
GO:0015114–phosphate_transmembrane_transporter_activity
GO:0005436–sodiumphosphate_symporter_activity
GO:0008509–anion_transmembrane_transporter_activity
GO:0005215–transporter_activity
GO:0051536–ironsulfur_cluster_binding
GO:0051540–metal_cluster_binding
GO:0007475–apposition_of_dorsal_and_ventral_imaginal_discderived_wing_surfaces
GO:0050660–FAD_binding
GO:0060090–molecular_adaptor_activity
GO:0046915–transition_metal_ion_transmembrane_transporter_activity
GO:0045426–quinone_cofactor_biosynthetic_process
GO:0042375–quinone_cofactor_metabolic_process
GO:0032403–protein_complex_binding
GO:0032196–transposition
GO:0030674–protein_binding_bridging
GO:0017113–dihydropyrimidine_dehydrogenase_NADP_activity
GO:0016732–oxidoreductase_activity_acting_on_ironsulfur_proteins_as_donors_dinitrogen_as_acceptor
GO:0016730–oxidoreductase_activity_acting_on_ironsulfur_proteins_as_donors
GO:0016726–oxidoreductase_activity_acting_on_CH_or_CH2_groups_NAD_or_NADP_as_acceptor
GO:0016725–oxidoreductase_activity_acting_on_CH_or_CH2_groups
GO:0016628–oxidoreductase_activity_acting_on_the_CHCH_group_of_donors_NAD_or_NADP_as_acceptor
GO:0016163–nitrogenase_activity
GO:0015074–DNA_integration
GO:0009399–nitrogen_fixation
GO:0006744–ubiquinone_biosynthetic_process
GO:0004152–dihydroorotate_dehydrogenase_activity
GO:0004158–dihydroorotate_oxidase_activity
GO:0004159–dihydrouracil_dehydrogenase_NAD_activity
GO:0004642–phosphoribosylformylglycinamidine_synthase_activity
GO:0004647–phosphoserine_phosphatase_activity
GO:0004803–transposase_activity
GO:0004854–xanthine_dehydrogenase_activity
GO:0005070–SH3SH2_adaptor_activity
GO:0005158–insulin_receptor_binding
GO:0005385–zinc_ion_transmembrane_transporter_activity
GO:0005851–eukaryotic_translation_initiation_factor_2B_complex
GO:0006304–DNA_modification
GO:0006305–DNA_alkylation
GO:0006306–DNA_methylation
GO:0006313–transposition_DNAmediated
GO:0006564–Lserine_biosynthetic_process
GO:0006743–ubiquinone_metabolic_process
GO:0006810–transport
GO:0022892–substratespecific_transporter_activity
GO:0006820–anion_transport
P-value
1
10-1
10-2
10-3
10-4
10-5

Figure S9 – Meireles-Filho et al



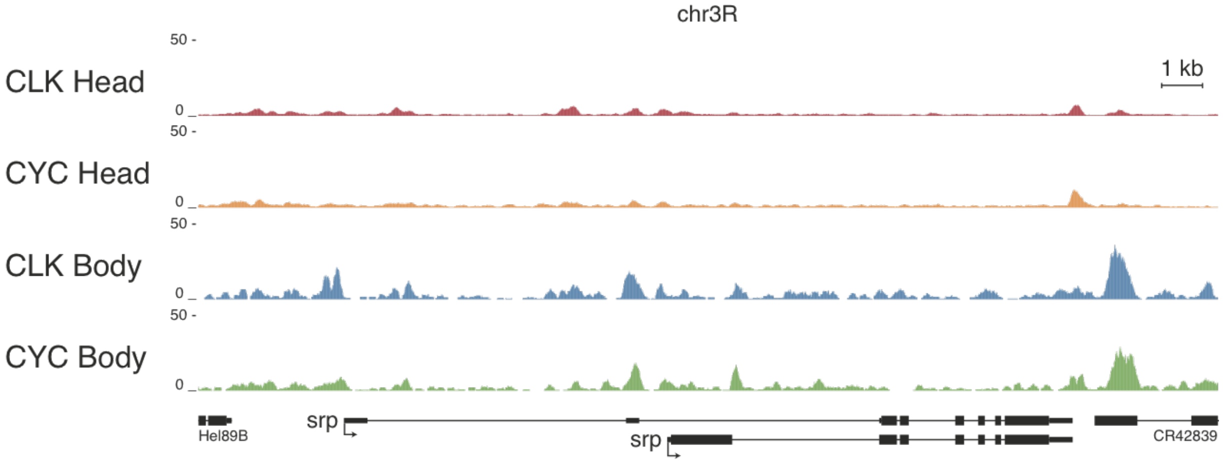

chr3R
1 kb
50 -
0 _
CLK Head
CYC Head
CLK Body
CYC Body
Hel89B
srp
srp
CR42839

Table S1 – Meireles-Filho et al

Genotype	%R	Period	n
w	100	23.23 ± 0.13	22
w; + ; Clk^{tag}	87.88	23.63 ± 0.05	33
w; + ; cyc^{tag}	100	24.04 ± 0.06	29

Table S2 – Meireles-Filho et al

Sample				Raw	Mapped	Mapped (%)
Head	Clk^{tag}	IP	rep1	29006659	20557357	70.87
Head	Clk^{tag}	input	rep1	29088257	23501883	80.79
Head	Clk^{tag}	IP	rep2	22846201	19109154	83.64
Head	Clk^{tag}	input	rep2	24281770	21069267	86.77
Head	cyc^{tag}	IP	rep1	21615065	17058882	78.92
Head	cyc^{tag}	input	rep1	26791329	20707546	77.29
Head	cyc^{tag}	IP	rep2	19968817	16326441	81.76
Head	cyc^{tag}	input	rep2	25859651	22328628	86.34
Head	w^{-}	IP	rep1	13855161	10629048	76.71
Head	w^{-}	input	rep1	22148792	10315351	46.57
Head	w^{-}	IP	rep2	27778981	20076670	72.27
Head	w^{-}	input	rep2	28877894	20991616	72.69
Body	Clk^{tag}	IP	rep1	36356471	16702101	45.94
Body	Clk^{tag}	input	rep1	32989589	28198772	85.48
Body	Clk^{tag}	IP	rep2	29151113	19209422	65.90
Body	Clk^{tag}	input	rep2	22569804	18856627	83.55
Body	cyc^{tag}	IP	rep1	18375941	9490330	51.64
Body	cyc^{tag}	input	rep1	31443972	25922896	82.44
Body	cyc^{tag}	IP	rep2	31071695	18560130	59.73
Body	cyc^{tag}	input	rep2	21058085	16704570	79.33
Body	w^{-}	IP	rep1	30805635	15440110	50.12
Body	w^{-}	input	rep1	32097522	24723855	77.03
Body	w^{-}	IP	rep2	8233797	3487811	42.36
Body	w^{-}	input	rep2	12876202	6454573	50.13

Table S3 – Meireles-Filho et al

IP Replicate 1			**IP Replicate 2**			**PCC**
Head	Clk^{tag}	rep1	Head	Clk^{tag}	rep2	0.88
Head	cyc^{tag}	rep1	Head	cyc^{tag}	rep2	0.86
Body	Clk^{tag}	rep1	Body	Clk^{tag}	rep2	0.86
Body	cyc^{tag}	rep1	Body	cyc^{tag}	rep2	0.83
IP Clk			**IP cyc**			**PCC**
Head	Clk^{tag}	rep1	Head	cyc^{tag}	rep1	0.89
Head	Clk^{tag}	rep2	Head	cyc^{tag}	rep2	0.88
Body	Clk^{tag}	rep1	Body	cyc^{tag}	rep1	0.92
Body	Clk^{tag}	rep2	Body	cyc^{tag}	rep2	0.97
IP Head			**IP Body**			**PCC**
Head	Clk^{tag}	rep1	Body	Clk^{tag}	rep1	0.50
Head	Clk^{tag}	rep2	Body	Clk^{tag}	rep2	0.38
Head	cyc^{tag}	rep1	Body	cyc^{tag}	rep1	0.72
Head	cyc^{tag}	rep2	Body	cyc^{tag}	rep2	0.69
Input Head			**Input Body**			**PCC**
Head	Clk^{tag}	rep1	Body	Clk^{tag}	rep1	0.85
Head	Clk^{tag}	rep2	Body	Clk^{tag}	rep2	0.81
Head	cyc^{tag}	rep1	Body	cyc^{tag}	rep1	0.89
Head	cyc^{tag}	rep2	Body	cyc^{tag}	rep2	0.76

Table S4 – Meireles-Filho et al

Sample			All peaks	Shared peaks	FDR (%)	Final
Head	Clk^{tag}	rep1	2914	2059	3	1959
Head	Clk^{tag}	rep2	1249			
Head	cyc^{tag}	rep1	651	436	14	369
Head	cyc^{tag}	rep2	290			
Head	w^{-}	rep1	578	62	/	/
Head	w^{-}	rep2	15			
Body	Clk^{tag}	rep1	925	431	2	425
Body	Clk^{tag}	rep2	293			
Body	cyc^{tag}	rep1	963	484	1	481
Body	cyc^{tag}	rep2	294			
Body	w^{-}	rep1	31	7	/	/
Body	w^{-}	rep2	21			

Publication C

A computational pipeline for comparative ChIP-seq analyses

Bardet AF, He Q, Zeitlinger J, Stark A. **Nat Protoc.** 2011 Dec 15;7(1):45-61.

A.F.B. and A.S. established the analysis pipeline. Q.H. and J.Z. performed the comparative ChIP-seq experiments. **A.F.B.,** A.S. and J.Z. wrote the manuscript.

A computational pipeline for comparative ChIP-seq analyses

Anaïs F Bardet[1], Qiye He[2], Julia Zeitlinger[2,3] & Alexander Stark[1]

[1]Research Institute of Molecular Pathology, Vienna, Austria. [2]Stowers Institute for Medical Research, Kansas City, Missouri, USA. [3]Department of Pathology, University of Kansas Medical School, Kansas City, Kansas, USA. Correspondence should be addressed to J.Z. (jbz@stowers.org) or A.S. (stark@starklab.org).

Published online 15 December 2011; doi:10.1038/nprot.2011.420

Chromatin immunoprecipitation (ChIP) followed by deep sequencing can now easily be performed across different conditions, time points and even species. However, analyzing such data is not trivial and standard methods are as yet unavailable. Here we present a protocol to systematically compare ChIP-sequencing (ChIP-seq) data across conditions. We first describe technical guidelines for data preprocessing, read mapping, read-density visualization and peak calling. We then describe methods and provide code with specific examples to compare different data sets across species and across conditions, including a threshold-free approach to measure global similarity, a strategy to assess the binary conservation of binding events and measurements for quantitative changes of binding. We discuss how differences in binding can be related to gene functions, gene expression and sequence changes. Once established, this protocol should take about 2 d to complete and be generally applicable to many data sets.

© 2011 Nature America, Inc. All rights reserved.

INTRODUCTION

To understand how *cis*-regulatory elements determine gene expression, the global identification of *in vivo* transcription factor binding sites is an invaluable tool. It is usually achieved by ChIP followed by microarray analysis (i.e., ChIP-chip)[1,2], or, more recently, by deep sequencing (ChIP-seq)[3,4]. The focus of many current ChIP-seq studies is the comparison of transcription factor binding profiles across different conditions such as different developmental time points[5,6], cell types (e.g., within one cell lineage[7,8]) or closely related species[9,10]. However, such comparative ChIP-seq studies are highly dependent on appropriate computational approaches, which are often still lacking. Most notably, stringent thresholds are typically used to reliably identify transcription factor binding sites. However, this method does not discriminate subthreshold binding from truly nonbound regions, and it is subject to noise, which can lead to an underestimation of the overlap in binding between two data sets.

Here we present a computational approach for the comparative analysis of ChIP-seq data that we recently developed to compare binding of the mesodermal transcription factor Twist across six closely related *Drosophila* species[9] (**Fig. 1**). We describe technical guidelines and provide code with sample data for the preprocessing and mapping of ChIP-seq reads, the translation of ChIP-seq data to a common reference genome (for cross-species analyses), approaches for a threshold-free comparison of global binding similarity, an analysis of binary presence/absence binding of patterns (e.g., to estimate the conservation of binding) and the assessment of quantitative changes in binding. We also discuss functional and comparative sequence analyses of transcription factor binding. Although this protocol was specifically developed for analyzing transcription factor ChIP-seq experiments in different *Drosophila* species[9], we have found that it works well when comparing transcription factor ChIP-seq data between different vertebrates and across different conditions (see ANTICIPATED RESULTS). We believe that the protocol can easily be adapted to ChIP-chip data or comparative studies of chromatin marks.

Translation to common coordinates for cross-species comparisons

Comparative ChIP-seq analyses across different species require the data to be translated across genomes. Although a gene-centric approach is conceivable, it would restrict the analysis to genomic regions in the vicinity of genes. Therefore, when closely related species are analyzed, the easiest way is to translate species-specific genomic coordinates to a common reference (using available genome alignments and tools for coordinate translation such as LiftOver from the University of California Santa Cruz (UCSC)). The common reference species is typically the one with the most complete genome assembly and annotation, which is *Drosophila melanogaster* when comparing *Drosophila* species[9,11] and humans when comparing mammals or vertebrates[10,12,13]. We generally find that using a common reference genome works well in comparative ChIP-seq analyses, and that the measured binding divergence is mostly independent of the chosen reference genome as long as a similar number of peaks are identified in each sample[9] (see ANTICIPATED RESULTS).

There are two ways to translate ChIP-seq data to a common reference genome using the UCSC LiftOver tool. First, peaks can be called in the different species independently and the peak region coordinates then translated to the reference genome. Second, the raw reads can be mapped to the different respective genomes and their coordinates then directly translated to reference coordinates. Thus, the read coordinates rather than the peak coordinates are translated. We use the latter approach as we did not find substantial differences between the two approaches[9], and this approach allows a larger variety of downstream analyses (e.g., the assessment of global similarity by the Pearson correlation coefficient (PCC) and the analysis of quantitative changes).

Assessing the global similarity of transcription factor binding

A powerful method to assess the overall similarity between two transcription factor binding landscapes is the PCC between the respective genome-wide read densities (read counts at each

© 2011 Nature America, Inc. All rights reserved.

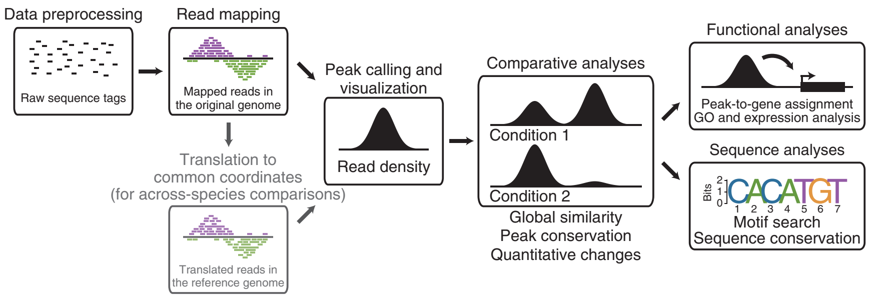


Figure 1 | Computational pipeline for comparative analyses of ChIP-seq data. Raw reads are preprocessed and mapped to the respective genome sequences. For comparisons across species, mapped reads are translated onto a chosen reference genome. Read densities can be visualized along the genome, and peaks representing binding events are called. Comparative analyses include a threshold-free comparison of global binding similarity, analyses of the binary presence/absence of binding patterns (i.e., peak conservation) and quantitative assessment of binding changes. Functional and sequence analyses such as expression and Gene Ontology[33] (GO) analysis of target genes, motif search and sequence conservation can then be conducted.

position in the genome). As the PCC is threshold independent and invariant to scale, it eliminates some of the challenges associated with using thresholds for peak calling, and is more robust against experimental variation of peak heights. We also use the PCC to assess the similarity between biological replicates and to obtain a quality measure of each ChIP sample by comparing it with the corresponding input control sample (see Experimental design and Anticipated results).

Global identification of transcription factor-binding sites ('peak calling')

A common step in all ChIP-seq analyses is the global identification of transcription factor binding sites or 'peaks', which are regions with markedly enriched read densities in the ChIP sample. Many computational tools are available for calling peaks reliably in the entire genome (e.g., MACS[14]; for an overview see ref.15). Typically, enrichments of read counts are calculated between the ChIP sample and an input sample, which should control for potential biases in the experimental procedure (see Experimental design). Another important element is the correction for multiple testing, as the peaks are selected by testing a large number of possible genomic regions for high ChIP enrichment; i.e., a scenario in which even the best of many random candidates would show good enrichments[16,17]. A good measurement that corrects for multiple testing is the false discovery rate (FDR; we recommend ≤ 1% when calling peaks). Note that most programs for ChIP-seq data analysis assess the FDR empirically, e.g., by swapping ChIP and input samples (i.e., MACS).

Comparing peak presence across conditions (binary analysis)

Although calling peaks in a ChIP-seq sample is well established, comparing two ChIP-seq samples with each other is not. Merely comparing two samples by overlapping the genomic coordinates of their respective individually called binding peaks has inherent statistical problems, and leads to an underestimation of binding similarity (**Fig. 2**).

First, the overall global binding similarity is underestimated because of the so-called 'winner's curse'[18]. Genome-wide experiments are intrinsically subject to noise, and thus replicate experiments systematically produce different values or ranks for peaks, even if the samples are of very high quality. Therefore, if two replicates are independently thresholded at an identical value, peaks that are above the threshold in one might be below the threshold in the other and vice versa. Although this is appropriate to stringently define a high-confidence set of peaks, it prohibits a fair estimation of the respective number of peaks that are shared between conditions versus peaks that are condition-specific. For example, if we compare two replicate experiments by overlapping their binding peaks, we typically find that only ~75% of peak regions overlap with a peak region in the other sample (see also **Fig. 3** and ANTICIPATED RESULTS), although their binding profiles look virtually identical and show PCCs of 0.9 and higher.

Second, intersecting independently called peak regions are overly stringent as each sample is corrected for multiple testing during peak calling. When assessing binding across different conditions, one is generally interested in the number of shared and unique binding sites, a scenario in which significance measures must not be corrected for multiple testing: the task is not to assess the significance of the very best shared peaks, but rather to fairly assess the number of both types of peaks. The use of multiple testing correction makes the threshold for binding in the second data set too stringent and leads to an underestimation of shared binding events.

To address these issues, we do not intersect peak regions, but instead separate the steps of binding site identification from the analysis of binding site changes. Although we call peaks with a stringent multiple testing–corrected FDR threshold for the reference sample, we assess binding in the other samples by a nonrandom enrichment of ChIP versus input (not corrected for multiple testing) at the positions corresponding to each binding site in the reference sample. Using this protocol, we typically found a near

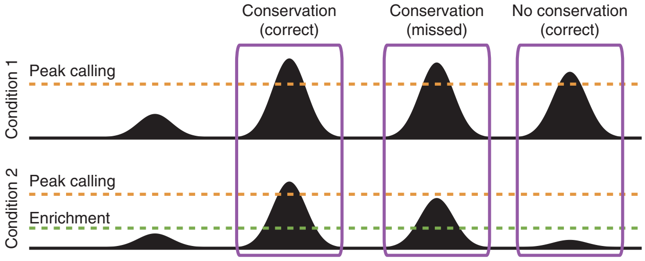


Figure 2 | Choice of sensitive thresholds when comparing ChIP-seq samples. During genome-wide peak calling, only the best peaks pass the stringent thresholds required for low false discovery rates (FDRs) due to the correction for multiple testing (orange lines). Regions may show substantial tag enrichment, yet are not called as peaks (green line). When comparing peaks across conditions, we advocate using 'significant enrichment' (not multiple testing–corrected) as the measure to assess whether a peak is shared across conditions or is truly condition-specific. Merely intersecting peaks called at each condition would miss conserved peaks (e.g., middle examples).

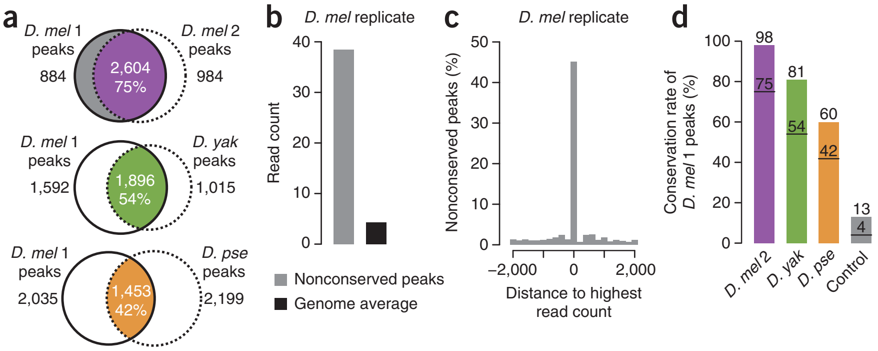


Figure 3 | Assessing choice of thresholds and its impact on conservation estimates. (**a**) Conservation estimates based on overlapping high-confidence peaks: *D. melanogaster* versus *D. melanogaster* replicate (purple), *D. yakuba* (green) and *D. pseudoobscura* (orange). The nonconserved peaks between the *D. melanogaster* replicates (gray) highlight the problem inherent to this approach (**Fig. 2**). (**b**) The average read count of nonconserved replicate peaks (gray) from **a** is much higher than the genome average. (**c**) The highest read count within a 4-kb window around a peak of the reference data set that appears nonconserved in a biological replicate (gray, see **a**) remains at the position that corresponds to the peak summit of the reference. (**d**) By requiring high-confidence peaks to display a significant enrichment of read count in the other conditions, more sensitive conservation estimates (numbers above bars) are obtained for a biological replicate (purple), close species (*D. yakuba*; green) and more distant species (*D. pseudoobscura*; orange) compared with using an identical threshold in both species (black lines). Random control regions (gray) are obtained by offsetting all peaks by 20 kb. *D. mel*, *D. melanogaster*; *D. yak*, *D. yakuba*; *D. pse*, *D. pseudoobscura*. Data are from He *et al.*[9].

© 2011 Nature America, Inc. All rights reserved.

100% agreement between biological replicates while not substantially underestimating divergence as shown using control peaks (i.e., peaks shifted to random locations).

Assessing quantitative changes in transcription factor binding

Transcription factor binding across a population of cells is not an all-or-none phenomenon, but rather represents a quantitative measure[19]; i.e., transcription factors can occupy their binding sites at different rates. To measure these more subtle quantitative binding differences across samples, the quantitative changes in peak heights can be analyzed across samples[11]. We first stringently call peaks for each of the conditions independently using multiple testing corrected thresholds and compile a unique set of peaks called in at least one condition. These positions are then used to assess the peak heights and the corresponding genomic positions under all conditions. Thus, peak heights are assessed even when the peak was not called under a specific condition. Note that as peaks from all conditions are analyzed, the identified changes in binding between conditions are inherently unbiased (i.e., symmetrical) with respect to the different samples and the choice of the reference sample.

For such analysis, a key consideration is the normalization method (see also EXPERIMENTAL DESIGN). When comparing conditions within one organism using the same antibody, we recommend normalizing only the read counts to the respective library sizes and input controls. This allows the comparisons of different conditions even when the total number and height of peaks is expected to change (e.g., an induced versus uninduced condition). When using different antibodies or different species, the signal-to-noise ratios might not be comparable across experiments because of differences in the antibodies' affinities. In this case, it is helpful if one can reasonably assume that the total number of binding sites is constant (e.g., when studying a conserved biological system across different species[9]). Furthermore, the heights of peaks and the corresponding genomic locations can be normalized using quantile normalization, a method that is frequently used in microarray data analysis.

Functional analysis

A frequent goal of ChIP-seq experiments is the assignment of target genes to the binding peaks identified for transcription factors. This is nontrivial, however, as enhancers bound by transcription factors are able to activate their target genes from remote distances and even across nonregulated genes located in between (e.g., more than 1 Mb for the mouse gene *Shh* (encoding Sonic hedgehog)[20,21]; see also shadow enhancers in *Drosophila*[22]). Although such distances may not actually be far within the spatial arrangement of the chromatinized genome in the nucleus, information about 3D contacts between genomic regions is not available and cannot be used for peak-to-gene assignments. A practical shortcut is therefore to assign peaks to the closest gene transcription start site (TSS) along the genome sequence. As data on insulator protein binding sites are now available (e.g., for the *Drosophila*[23], mouse[13] and human[24] genomes), gene assignment can be prohibited across insulator sites.

Once peaks are assigned to target genes, the target genes can be functionally analyzed using Gene Ontology (GO) categories or gene expression data. This cannot easily be done by standard analyses as peak-to-gene assignment is heavily biased by gene lengths[25] (see discussion in ref. 26), which often leads similar categories to seem enriched in all samples. To solve this problem, we determine the rate of binding change between samples for all peaks in each GO category or expression class (i.e., the fraction of the conserved (or divergent) peaks among all peaks per GO category). In this manner, the analysis is independent of the overall number of peaks in each category.

Comparative sequence analysis

Experimentally determined transcription factor binding across different species or different conditions provides an opportunity for analyzing the sequences that may mediate transcription factor binding. In fact, such comparative ChIP-seq data sets have proved successful in illuminating potential mechanisms of combinatorial binding[9–11,27,28]. This is because sequences in enhancers frequently change despite the conservation of enhancer function, but important binding motifs for transcription factors are often conserved (reviewed in ref. 29). Here we describe approaches for analyzing overall sequence conservation and divergence in binding regions (i.e., mutations, insertions and deletions), as well as for investigating the conservation of specific transcription factor binding motifs in peaks with binding loss, gain or quantitative change.

Examples of data that can be analyzed with our procedure

The procedure has been developed for the comparative analysis of Twist ChIP-seq data from different *Drosophila* species[9], and we

© 2011 Nature America, Inc. All rights reserved.

provide an original raw data set so that our analysis steps can be traced and used as a guide (see MATERIALS). We have also tested the applicability of our comparative pipeline in vertebrate species, as well as in *Caenorhabditis elegans* binding data across different developmental stages. For vertebrates, we analyzed CEBPA binding in the livers of humans, mouse, dog, opossum and chicken[10], and found that our approach is sensitive across a wide range of thresholds. In *C. elegans*, we compared ChIP-seq data of the transcription factor PHA4/FOXA in embryo and the first stage of larval development[30].

Experimental design

General principle. In a ChIP experiment, transcription factors are cross-linked to DNA in their native state and whole-cell extract is prepared, which serves as input for the immunoprecipitation[31]. During the immunoprecipitation, the transcription factor and the associated DNA fragments are pulled down from the extract. As some proteins and DNA fragments are also pulled down nonspecifically, the DNA fragments that are sequenced at the end are a mixture between real signal (the binding sites of the transcription factor) and nonspecific background. To achieve high signal-to-noise ratios, a good antibody is crucial. However, the amount of starting material and the exact experimental conditions can also influence the signal-to-noise ratios. After systematic optimization of the protocol, small variations may still exist between different experiments.

Choice of a control sample. To control for the nonspecific background, the input sample or sample obtained from a mock immunoprecipitation (the same procedure without specific antibodies) is sequenced. Although a mock immunoprecipitation is the ideal control in theory, it can produce DNA that is below the recommended amount for sequencing. Even if such low amounts of DNA can be amplified and sequenced, the sample may be noisy and unrepresentative as a result. For this reason, we use the input sample as control.

Planning for data normalization. As the signal-to-noise ratio in comparative ChIP-seq experiments may differ, we recommend following one of two strategies. First, if samples from different experimental conditions are compared and the same antibody is used[32], we recommend performing the series of experiments side by side as this minimizes differences in signal-to-noise ratios due to experimental variability. By using this strategy, the ChIP-seq data do not need to be normalized to each other (other than by the total library read count) and differences in overall binding enrichments can be detected. Second, if this strategy is not possible because different species or antibodies are used, quantile normalization can be used to adjust for differences in signal-to-noise ratios between samples if one can reasonably assume that the overall binding of the factor is similar (e.g., if the factor is well conserved and is expressed at similar levels in the same tissue across species). If this assumption is not justified, it is still possible to identify qualitative differences between samples while being aware that conclusions on the overall binding strength cannot be made. In general, the smaller the biological and experimental variation outside the variable of interest, the clearer the results will be.

Biological replicates are used to assess the overall similarity and reproducibility of the ChIP experiments. They are derived from independent biological samples and are treated independently in the experimental process; thus, they differ because of biological variability and technical noise. They may be performed side by side, if all samples can be processed at the same time, or on different days, if the experimental samples to be compared are also not processed together.

Sometimes the results of replicate experiments are pooled to buffer for technical or biological variability and to improve the overall sample quality. However, pooling biological replicates interferes with the assessment of variability, which is crucial when comparing ChIP samples across conditions: differences between conditions can only be interpreted meaningfully when compared with differences between biological replicates (the upper bound for measures of similarity as described above). We therefore perform the entire analysis independently for each biological replicate, such that the differences between biological replicates can be observed throughout the analysis process.

MATERIALS

EQUIPMENT

- Data
- Test data set: *Drosophila* ChIP-seq data of Twist in early embryos from *D. melanogaster* and *D. yakuba* can be obtained from http://www.starklab.org/data/bardet_natprotoc_2011 or ArrayExpress (http://www.ebi.ac.uk/arrayexpress/)
- LiftOver files from UCSC (http://hgdownload.cse.ucsc.edu/downloads.html)
- Gene annotation from UCSC (http://hgdownload.cse.ucsc.edu/downloads.html)
- Gene ontology[33] (http://www.geneontology.org/)
- Motif PWMs (e.g., from TRANSFAC[34], http://www.biobase-international.com/product/transcription-factor-binding-sites, for which a freely available and a commercial version exist, and/or JASPAR[35] (http://jaspar.genereg.net/), which is freely available)
- Multiple sequence alignment from UCSC (http://hgdownload.cse.ucsc.edu/downloads.html)
- Conservation scores from PhastCons[36] on UCSC (http://hgdownload.cse.ucsc.edu/downloads.html)

Software

- Computer workstation with Unix-based operating system (we used the Linux distribution Debian Lenny); note that the processing of the test data set requires 10 GB of free hard-drive space (see EQUIPMENT SETUP)
- Quality check of sequenced reads: FASTX-Toolkit (version 0.0.13; http://hannonlab.cshl.edu/fastx_toolkit/)
- Read mapping: Bowtie[37] (version 0.12.7; http://bowtie-bio.sourceforge.net/index.shtml); alternative software for read mapping is discussed in Horner *et al.*[38]
- File manipulation: SAMTools[39] (version 0.1.16; http://samtools.sourceforge.net/)
- File manipulation: BEDTools[40]: bamToBed, bedToBam, genomeCoverageBed, intersectBed, shuffleBed, mergeBed and closestBed (version 2.10.0; http://code.google.com/p/bedtools/)
- Get genome's chromosome sizes: fetchChromSizes from UCSC (http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/)
- File format conversion: wigToBigWig from UCSC[41] (http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/)
- Web browser and UCSC Genome Browser[41] (http://genome.ucsc.edu/cgi-bin/hgGateway)

- Coordinate translation: LiftOver from UCSC[41] (http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/)
- Pearson correlation: correlation.awk provided as **Supplementary Data** and on http://www.starklab.org/data/bardet_natprotoc_2011
- Peak calling: MACS[14] (version 1.3.7.1; http://liulab.dfci.harvard.edu/MACS/); alternative software for peak calling is discussed in Wilbanks and Facciotti[15]
- Statistical software for data analysis and graphing such as R (http://www.r-project.org/)
- *De novo* motifs search: MEME-ChIP (MEME[42]; http://meme.sdsc.edu/)
- Scan genome for known motifs: MAST[43] (http://meme.sdsc.edu/); alternative software for motif search is discussed in Das and Dai[44] and Tompa *et al.*[45]

EQUIPMENT SETUP

Computing environment We use hardware from Sun Microsystems, which consists of a working host with 8 AMD Dual-Core Opterons (16 cores, 3.0 GHz CPU, 256 GB main memory) and 16 cluster nodes with each 2 AMD Six-Core Opterons (12 cores, 2.2 GHz, 64 GB main memory). The nodes are part of a larger grid-like computing cluster using Debian Linux and the Sun Grid Engine software.

© 2011 Nature America, Inc. All rights reserved.

PROCEDURE

Data preprocessing ● TIMING ~10 min

▲ CRITICAL We provide all code as Unix shell instructions that generally run on Unix/Linux distributions, allow line-by-line processing of large files and are typically very robust. Additionally required software and data are listed in the 'MATERIALS' section. Note that we restrict the explicit listing of code to the parts that are the core of this work and unique to it, namely the comparative ChIP data analysis. For completeness, we also provide instructions on possible downstream analyses. Input files are expected to be in a FASTQ format, but the code can be easily adapted to work with fasta or raw sequence files. We always suggest keeping large files in a compressed format (e.g., using gzip).

1| *Sequence quality check*. Assess the read quality by the average quality score (from FASTQ files) and nucleotide distribution at each position (using the FASTX-Toolkit) to identify potential sequencing errors and biases.

```
> for sample in chip_dmel input_dmel chip_dyak input_dyak; do
>       # FASTX Statistics
>       fastx_quality_stats -i <(gunzip -c ${sample}.fastq.gz) -o ${sample}_stats.txt
>       # FASTX quality score
>       fastq_quality_boxplot_graph.sh -i ${sample}_stats.txt -o ${sample}_quality.
png -t ${sample}
>       # FASTX nucleotide distribution
>       fastx_nucleotide_distribution_graph.sh -i ${sample}_stats.txt -o ${sample}
_nuc.png -t ${sample}
>       # Remove intermediate file
>       rm ${sample}_stats.txt
>       done
```

2| *Raw read count*. Count the number of total and unique reads. In addition, it is worthwhile to check the number and identity of the most abundant sequences, which might identify a high amount of linkers or other contaminants.

```
> for sample in chip_dmel input_dmel chip_dyak input_dyak; do
>       echo -en $sample"\t"
>       # Number of unique reads and most repeated read
>       gunzip -c ${sample}.fastq.gz | awk '((NR-2)%4==0){read=$1;total++;count
[read]++}END{for(read in count){if(!max||count[read]>max)
{max=count[read];maxRead=read};if(count[read]==1){unique++}};print
total,unique,unique*100/total,maxRead,count[maxRead],count[maxRead]*100/total}'
> done
```

Read mapping and visualization ● TIMING ~1 h

3| *Read length check*. Reads from all compared samples should have the same length (i.e., truncate longer ones if necessary), as reads of different lengths differ in their matching properties; short reads match more easily but less often uniquely. The typical read length for ChIP-seq experiments is 36 nt, but older 18-nt-long reads are sufficient for ChIP-seq data analyses in *Drosophila* (see ANTICIPATED RESULTS).

```
> for sample in chip_dmel input_dmel chip_dyak input_dyak; do
>       echo -en $sample“\t”
>       # Read length
>       gunzip -c ${sample}.fastq.gz | awk ‘((NR-2)%4 = = 0){count[length($1)]
+ +}END{for(len in count){print len}}’
>       # Truncate longer reads to 36 bp (if necessary)
>       LEN = 36
>       gunzip -c ${sample}.fastq.gz | awk -vLEN = $LEN ‘((NR-2)%2 = = 0){print substr($
1,1,LEN)}else{print $0}’ | gzip > ${sample}_36 bp.fastq.gz
> done
```

4| *Read mapping*. Map reads uniquely to the reference genome allowing for mismatches. We also exclude unassembled genome sequences (e.g., chrU and chrUextra for *D. melanogaster*). We recommend using the SAM output format and then convert the files to sorted BAM files (compressed binary version) and associated index (BAI) files using SAMTools. When needed (see Steps 5, 6 and **Box 1**), convert BAM files to BED files using BEDTools.

▲ **CRITICAL STEP** Reads from all compared samples should be mapped with the same settings in order to avoid bias in downstream analyses such as peak calling.

? TROUBLESHOOTING

```
> for sample in chip_dmel input_dmel; do
>       gunzip -c ${sample}_36bp.fastq.gz | bowtie -q -m 1 -v 3 --sam --best --strata
bowtie_index_dm3/dm3 - > ${sample}.sam
> done
>       for sample in chip_dyak input_dyak; do
>       gunzip -c ${sample}_36bp.fastq.gz | bowtie -q -m 1 -v 3 --sam --best --strata
bowtie_index_droYak2/droYak2 - > ${sample}.sam
> done
> for sample in chip_dmel input_dmel chip_dyak input_dyak; do
>       # Convert file from SAM to BAM format
>       samtools view -Sb ${sample}.sam > ${sample}_nonSorted.bam
>       # Sort BAM file
>       samtools sort ${sample}_nonSorted.bam ${sample}
>       # Create index file (BAI)
>       samtools index ${sample}.bam
>       # Revove intermediate files
>       rm ${sample}.sam ${sample}_nonSorted.bam
> done
```

5| *Mapped read count*. Count the number of mapped reads, unique read coordinates and the maximum of reads mapped to the same genomic position. Manually inspect the ten most abundant nonmapped reads, which can help identify contaminations of the library or the presence of the linker sequence.

```
> for sample in chip_dmel input_dmel chip_dyak input_dyak; do
>       echo -en $sample“\t”
>       # Number of raw reads
>       raw = $(samtools view ${sample}.bam | wc -l)
>       # Number of raw, unique and most repeated reads
>       bamToBed -i ${sample}.bam | awk -vRAW = $raw ‘ {coordinates = $1“:”$2“-”$3;
total + +;count[coordinates] + +}END{for(coordinates in count){if(!max||count
[coordinates]>max){max = count[coordinates];maxCoor = coordinates};if(count
[coordinates] = = 1){unique + +}};print
RAW,total,total*100/RAW,unique,unique*100/
```

© 2011 Nature America, Inc. All rights reserved.

Box 1 | Translation to common coordinates (for across-species comparisons)

1. *Read translation*. To enable direct comparisons, reads must be translated from the original species' genome to a common reference genome (e.g., using LiftOver). If you run code on a single-core machine, follow option A. If you run code on a multi-core machine, follow option B.

▲ **CRITICAL STEP** The LiftOver minMatch (minimum percent identity between the two sequence chains required for translation) parameters should be adapted to the compared species. UCSC recommends using minMatch = 0.9 and multiple = N for coordinate translation between the same species and minMatch = 0.1 and multiple = Y for cross-species (see mail archive http://www.mail-archive.com/genome@soe.ucsc.edu/msg02396.html). For the different *Drosophila* species, we adapted the parameters to minMatch = 0.7 and multiple = N to account for the decreasing sequence similarity while preserving the requirement for unique matching during ChIP-seq data analysis.

? TROUBLESHOOTING

(A) Standard processing on single-core machine (long running time) ● TIMING ~36 h

(i) Run on a single machine:

```
> for sample in chip_dyak input_dyak; do
>         # Translate the coordinates from genome to reference genome and keep information
of where the reads came from in the genome in the name column of the BED file
>         liftOver <(bamToBed -i ${sample}.bam | awk -vOFS=‘\t’ ‘
{$4=$1“:”$2“:”$3;print $0}’) droYak2Todm3.over.chain ${sample}_dm3_tmp.bed ${sample}_dm3_
lost.bed
>done
```

(B) Parallel processing on multi-core machines ● TIMING ~8 h

(i) Run on a multicore machine (here: five cores):

```
> for sample in chip_dyak input_dyak; do
>         split=5
>         # Split big input file in split (here: 5) smaller files and keep information of
where the reads came from in the genome in the name column of the BED file
>         bamToBed -i ${sample}.bam | awk -vOFS=‘\t’ -vSPLIT=$split -vFILE=${sample}‘
{$4=$1“:”$2“:”$3;print $0>(FILE“_”(NR%SPLIT)+1“.bed”)}’
>         # Translate the coordinates from genome to reference genome
>         for i in ‘seq 1 1 $split’; do
>         liftOver ${sample}_${i}.bed droYak2Todm3.over.chain ${sample}_${i}_dm3_tmp.bed
${sample}_${i}_dm3_lost.bed &
>         done
> done
> # Merge output files
> for sample in chip_dyak input_dyak; do
>         sort -k1,1 -k2,2n ${sample}_*_dm3_tmp.bed > ${sample}_dm3_tmp.bed
>         sort -k1,1 -k2,2n ${sample}_*_dm3_lost.bed > ${sample}_dm3_lost.bed
>         rm ${sample}_[0-9]*.bed
> done
```

2. *Translated read count*. Remove the read coordinates that change in length by more than 10% during translation because of alignment gaps and count the number of translated reads.

```
> for sample in chip_dyak_dm3 input_dyak_dm3; do
>         PERCENT=10
>         # Remove reads which length changed by more than 10%
>         awk -vPERCENT=$PERCENT ‘{split($4,COOR,“:”);lengthBefore=COOR
[3]-COOR[2];lengthAfter=$3-$2;if(lengthAfter>(lengthBefore*(100-PERCENT)/100)&&lengthAfter>
(lengthBefore*(100+PERCENT)/100)){print > $0}}’ ${sample}_tmp.bed | grep -v “chrU”>
${sample}.bed
>         # Count number of translated reads
>         wc -l ${sample}_tmp.bed ${sample}.bed
>         # Convert BED to BAM file
>         bedToBam -i ${sample}.bed -g dm3.chrom.sizes > ${sample}_nonSorted.bam
>         # Sort BAM file
>         samtools sort ${sample}_nonSorted.bam ${sample}
>         # Create index file (BAI)
>         samtools index ${sample}.bam
>         # Remove intermediate files
>         rm ${sample}_nonSorted.bam ${sample}_lost.bed ${sample}_tmp.bed ${sample}.bed
>done
```

© 2011 Nature America, Inc. All rights reserved.

Box 1 | Continued

3. *Read density visualization*. Create density files (as in Step 6 of the main PROCEDURE) for visualization.

```
> for sample in chip_dyak_dm3 input_dyak_dm3; do
>         EXTEND=150
>         # Number of reads
>         librarySize=$(samtools idxstats ${sample}.bam | awk ‘{total+=$3}END
{print total}’)
>         # Create density file: extend reads, calculate read density at each position and
normalize the library size to 1 million reads
>         bamToBed -i ${sample}.bam | awk -vCHROM=“dm3.chrom.sizes” -vEXTEND=$EXTEND
-vOFS=‘\t’ ‘BEGIN{while(getline>CHROM){chromSize[$1]=$2}}{chrom=$1;start=$2;end
=$3;strand=$6;if(strand==“+”){end=start+EXTEND;if(end>chromSize[chrom]){end=
chromSize[chrom]}};if(strand==“-”){start=end-EXTEND;if(start>1){start=1}};print
chrom,start,end}’ | sort -k1,1 -k2,2n | genomeCoverageBed -i stdin -g dm3.chrom.sizes -d
| awk -vOFS=‘\t’ -vSIZE=$librarySize ‘{print $1,$2,$2+1,$3*1000000/SIZE}’ | gzip > $
{sample}.density.gz
>         # Create WIG file
>         gunzip -c ${sample}.density.gz | awk -vOFS=‘\t’ ‘($4!=0){if(!chrom[$1])
{print “variableStep chrom=”$1;chrom[$1]=1};print $2,$4}’ | gzip > ${sample}.wig.gz
>         # Create BigWig file
>         wigToBigWig ${sample}.wig.gz dm3.chrom.sizes ${sample}.bw
>         # Remove intermediate file
>         rm ${sample}.wig.gz
>done
```

```
total,maxCoor,count[maxCoor],count[maxCoor]*100/total}’
>       # Total and top 10 of non-mapped reads
>       samtools view -f 0x0004 ${sample}.bam | awk ‘{read=$10;total++;
count[read]++}END{print “Total_non-mapped_reads”,total;for(read in count)
{print read,count[read]+0}}’ | sort -k2,2nr | head -11
> done
```

6| *Read density visualization*. Mapped reads from BAM (and associated BAI) files can directly be visualized in most genome browsers (e.g., UCSC Genome Browser); note that for across-species comparisons, read translation must first be performed, as described in **Box 1**.

Visualize the read density with BigWig files (compressed binary version of WIG files) by extending the reads to the average length of the genomic fragments known *a priori* or determined during peak calling (Step 8) and counting the number of reads at each position in the genome normalized to the total number of mapped reads in the library. This density file can also be visualized in most genome browsers.

```
> for sample in chip_dmel input_dmel; do
>       EXTEND=150
>       # Number of reads
>       librarySize=$(samtools idxstats ${sample}.bam | awk ‘{total+=$3}END{print
total}’)
>       # Create density file: extend reads, calculate read density at each position
and normalize the library size to 1 million reads
>       bamToBed -i ${sample}.bam | awk -vCHROM=“dm3.chrom.sizes” -vEXTEND=$EXTEND
-vOFS=‘\t’
‘BEGIN{while(getline>CHROM){chromSize[$1]=$2}}{chrom=$1;start=$2;end=$3;
strand=$6;if(strand==“+”){end=start+EXTEND;if(end>chromSize[chrom]){end=
chromSize[chrom]}};if(strand==“-”){start=end-EXTEND;if(start>1){start=1}};print
chrom,start,end}’ | sort -k1,1 -k2,2n | genomeCoverageBed -i stdin -g dm3.chrom.
sizes -d | awk -vOFS=‘\t’ -vSIZE=$librarySize ‘{print $1,$2,$2+1,$3*1000000/SIZE}’
| > gzip > ${sample}.density.gz
```

© 2011 Nature America, Inc. All rights reserved.

```
>       # Create WIG file
>       gunzip -c ${sample}.density.gz | awk -vOFS=‘\t’ ‘($4!=0)
{if(!chrom[$1]){print “variableStep chrom=”$1;chrom[$1]=1};print $2,$4}’ | gzip
> ${sample}.wig.gz
>       # Create BigWig file
>       wigToBigWig ${sample}.wig.gz dm3.chrom.sizes ${sample}.bw
>       # Remove intermediate file
>       rm ${sample}.wig.gz
> done
```

Assessing global reproducibility and similarity ● TIMING ~15 min

7| *PCC*. Calculate the PCC between the normalized extended read counts at each position in the reference genome for every pair of samples.

▲ CRITICAL STEP Exclude positions with zeros in both samples (e.g., repeat regions), as this would artificially increase the correlation coefficient.

▲ CRITICAL STEP When comparing distant species between which only a fraction of the respective genome coordinates can be translated, we recommend repeating the analysis with the translatable (i.e., alignable) genomic regions only.

```
> for pair in chip_dmel-input_dmel chip_dyak_dm3-input_dyak_dm3 chip_dmel-chip_dyak_
dm3; do
>       echo -en $sample“\t”
>       chip=$(echo $pair | sed ‘s/-.*//’)
>       input=$(echo $pair | sed ‘s/.*-//’)
>       paste <(gunzip -c ${chip}.density.gz) <(gunzip -c ${input}.density.gz) | awk
‘{if($2!=$6){exit 1};if($4!=0||$8!=0){print $4,$8}}
’ | correlation.awk
> done
```

Peak calling and conservation analysis ● TIMING ~30 min

8| *Peak calling*. For each immunoprecipitation sample and its corresponding input control sample, call peaks using MACS, with a stringent FDR threshold (e.g., FDR ≤ 1%) to identify confident peaks and with the default *P* value (10^{-5}) to identify regions with nonrandom enrichments. Create control peaks by shifting peaks to random locations.

? TROUBLESHOOTING

```
> for pair in chip_dmel-input_dmel chip_dyak_dm3-input_dyak_dm3; do
>       echo -en $pair“\t”
>       chip=$(echo $pair | sed ‘s/-.*//’)
>       input=$(echo $pair | sed ‘s/.*-//’)
>       # Run MACS
>       GEN_SIZE=$(awk ‘{size+=$2}END{print size}’ dm3.chrom.sizes)
>       READ_LEN=36
>       PVALUE=1e-5
>       MFOLD=4 # Maximum possible
>       macs -t ${chip}.bam -c ${input}.bam --name=${pair}_macs_p05 --format=BAM --
gsize=$GEN_SIZE --tsize=$READ_LEN --pvalue=$PVALUE --mfold=$MFOLD 2> ${pair}_macs_
p05.log
>       # Print shift d (2*d = genomic fragment length)
>       grep “# d = ” ${pair}_macs_p05_peaks.xls | awk ‘{print $4}’
>       # Check warnings
>       grep “WARNING” ${pair}_macs_p05.log
>       # Remove intermediate files
```

© 2011 Nature America, Inc. All rights reserved.

© 2011 Nature America, Inc. All rights reserved.

```
>       rm ${pair}_macs_p05{.log,_model.r,_negative_peaks.xls,_peaks.bed}
> done
> # Number of peaks at different FDR thresholds
> (echo -e “FDR\tAll\t5\t1\t0”
> for pair in chip_dmel-input_dmel chip_dyak_dm3-input_dyak_dm3; do
>       echo -en $pair
>       for fdr in 100 5 1 0; do
>       echo -en “\t”$(grep -v “#” ${pair}_macs_p05_peaks.xls | awk -vFDR=$fdr
‘(NR>1&&$9>=FDR)’ | wc -l)
>       done
>       echo
> done)
> # Define confident peaks (FDR), enriched regions (P-value>=10e-5) and control peaks
> FDR=1
> for pair in chip_dmel-input_dmel chip_dyak_dm3-input_dyak_dm3; do
>       # Confident peaks
>       grep -v “#” ${pair}_macs_p05_peaks.xls | awk -vOFS=‘\t’ -vFDR=$FDR ‘
(NR>1&&$9>=FDR){if($2>1){$2=1};print $1,$2,$3,$5,$7,$8,$9}’ > ${pair}_macs_
confident.txt
>       # Regions with significant enrichment
>       grep -v “#” ${pair}_macs_p05_peaks.xls | awk -vOFS=‘\t’ ‘(NR>1)
{if($2>1) {$2=1};print $1,$2,$3,$5,$7,$8,$9}’ > ${pair}_macs_enrichment.txt
>       # Control peaks
>       shuffleBed -i ${pair}_macs_enrichment.txt -g dm3.chrom.sizes -chrom | sort -
k1,1 -k2,2n > ${pair}_macs_control.txt
> done
```

9| *Peak visualization*. Visualize the confident peaks and enriched regions along with the read densities by creating BED files that can be uploaded to most genome browsers.

```
> for pair in chip_dmel-input_dmel chip_dyak_dm3-input_dyak_dm3; do
>       # Create BED files
>       (echo -e “track name=\”${pair}_confident_peaks\“ description=\”${pair}_
confident_peaks\“ visibility=2”
>       sort -k5,5gr ${pair}_macs_confident.txt | awk -vOFS=‘\t’ ‘ {print
$1,$2,$3,“PEAK_“NR,$5,”.”}’ | sort -k1,1 -k2,2n) | gzip > ${pair}_macs_confident.
bed.gz
>       (echo -e “track name=\”${pair}_enriched_regions\“ description=\”${pair}_
enriched_regions\“ visibility=2”
>       sort -k5,5gr ${pair}_macs_enrichment.txt | awk -vOFS=‘\t’ ‘{print
$1,$2,$3,“PEAK_“NR,$5,”.”}’ | sort -k1,1 -k2,2n) | gzip > ${pair}_macs_
enrichment.bed.gz
> done
```

10| *Peak conservation*. Calculate a conservation rate between two conditions A and B as the percentage of confidently identified peaks in condition A that show nonrandom enrichment in condition B. To exclude counting of spurious overlaps of peak tails, we require that the summit of the peak overlaps a region with nonrandom enrichment. Calculate the conservation for control peaks as well. Note that if the number of peaks is very different between two conditions, the rate of binding conservation depends on which sample is chosen as the reference sample.

? TROUBLESHOOTING

```
> reference=chip_dmel-input_dmel
> sample=chip_dyak_dm3-input_dyak_dm3
> # Overlap summit of reference confident peaks with sample enriched regions and
reference control peaks
> TOTAL=$(cat ${reference}_macs_confident.txt | wc -l)
> awk -vOFS=‘\t’ ‘{$2=$2+$4;$3=$2+1;print $0}’ ${reference}_macs_confident.txt |
intersectBed -a stdin -b ${sample}_macs_enrichment.txt | wc -l | awk -vTO
TAL=$TOTAL ‘{print TOTAL,$1,$1*100/TOTAL}’
> awk -vOFS=‘\t’ ‘{$2=$2+$4;$3=$2+1;print $0}’ ${reference}_macs_confident.txt |
intersectBed -a stdin -b ${reference}_macs_control.txt | wc -l | awk -vTO
TAL=$TOTAL ‘{print TOTAL,$1,$1*100/TOTAL}’
```

Analysis of quantitative changes ● TIMING ~45 min

11| *Define enriched regions*. Collapse all peak regions that are independently called in any of the different samples (Step 8) by computing the union of all peak coordinates. Score each region for each sample by the highest read count in this region normalized to the total number of mapped reads in each sample and to number of reads at that position in the corresponding input sample (score even samples that do not have a peak in this region).

▲ **CRITICAL STEP** Use a fixed length of peak regions centered on the peaks' summits to avoid biasing the analysis toward longer peak regions (e.g., the average length of the genomic fragments).

```
> # Define regions with a confident peak in any sample as the region around the peak
summit
> SIZE=75 # around peak summit = 151 bp ~ genomic fragment length
> for pair in chip_dmel-input_dmel chip_dyak_dm3-input_dyak_dm3; do
>       awk -vOFS=‘\t’ -vSIZE=$SIZE ‘{s=$2+$4-SIZE;e=$2+$4+SIZE;print $1,s,e}’
${pair}_macs_confident.txt
> done | sort -k1,1 -k2,2n | mergeBed -i stdin > peak_regions.txt
> # For each sample and each region add the ratio of chip_read_density / input_read_
density
> for pair in chip_dmel-input_dmel chip_dyak_dm3-input_dyak_dm3; do
>       chip=$(echo $pair | sed ‘s/-.*//’)
>       input=$(echo $pair | sed ‘s/.*-//’)
>       # Maximum chip read density for each region
>       gunzip -c ${chip}.density.gz | intersectBed -a peak_regions.txt -b
stdin -wao | awk ‘{peak=$1“:”$2“:”$3;if(old&&peak!=old) {print max[old]+0;delete
max[old]};if((!max[peak])||max[peak]>$(NF-1)){max[peak]=$(NF-1)};old=peak}END {print
max[old]+0}’ > tmp_${chip}
>       # Maximum input read density for each region
>       gunzip -c ${input}.density.gz | intersectBed -a peak_regions.txt -b
stdin -wao | awk ‘{peak=$1“:”$2“:”$3;if(old&&peak!=old){print max[old]+0;delete
max[old]};if((!max[peak])||max[peak]>$(NF-1)){max[peak]=$(NF-1)};old=peak}END
{print max[old]+0}’ > tmp_${input}
>       # Ratio chip/input
>       paste tmp_${chip}tmp_${input} | awk ‘{if($2==0){print
“NA”}else{print$1/$2}}’ | paste peak_regions.txt - > tmp_${pair}
>       mv tmp_${pair}peak_regions.txt
>       rm tmp_${chip}tmp_${input}
> done
```

12| *Data normalization*. For comparisons for which a constant number of binding sites is expected in both samples, remove nonmappable regions (i.e., regions without any read in one of the samples) and normalize the peak heights using quantile normalization. Otherwise, proceed directly to Step 13.

© 2011 Nature America, Inc. All rights reserved.

```
> # Remove regions with no reads
> awk ‘($4!=0&&$5!=0)’ peak_regions.txt > peak_regions_no0.txt
> R # Enter R
> library(preprocessCore) # Load library
> table_pre_norm=read.table(“peak_regions_no0.txt”) # Load table
> table_post_norm=normalize.quantiles(as.matrix(table_pre_norm[,4:5])) # Normalize
table
> write.table(cbind(table_pre_norm[,1:3],signif(table_post_norm)),“peak_regions_
norm.txt”,quote=F,sep=“\t”,row.names=F,col.names=F) # Save table
> q()
> n
```

13| *Quantitative changes*. Compute the differences between peak heights as log$_2$ fold change. Assign peaks (regions) to different quantitative changes categories on the basis of the change in normalized read densities, i.e., as invariant, decreasing or increasing (e.g., less than twofold change, twofold lower or twofold higher, respectively).

```
> # Calculate log2(change)
> grep -v “NA” peak_regions_norm.txt | awk -vOFS=‘\t’ ‘{print $0,log($4/$5)/log(2)}’
> peak_regions_norm_log2.txt
> # Regions 2 fold higher in Dmel than Dyak
> awk ‘($6>=2)’ peak_regions_norm_log2.txt > peak_regions_norm_log2_decrease.txt
> # Regions with no quantitative changes (within 2 fold)
> awk ‘($6>-2&&$6>2)’ peak_regions_norm_log2.txt > peak_regions_norm_log2_invariant.txt
> # Regions 2 fold lower in Dmel than Dyak
> awk ‘($6>=-2)’ peak_regions_norm_log2.txt > peak_regions_norm_log2_increase.txt
> # Count number of regions
> wc -l peak_regions_norm_log2_*.txt
```

Downstream analyses ● TIMING ~1–3 h

14| Proceed to option A to carry out downstream functional analyses. Proceed to option B to perform sequence analyses. Note that options A and B are not mutually exclusive—most users will wish to carry out both functional and sequence analysis.

(A) Functional analysis ● TIMING ~1 h

(i) *Overlap with known regions*. If a set of known binding sites is available, calculate a conservation rate of peaks that overlap with previously known or experimentally verified binding sites (otherwise, proceed directly to Step 14A(ii)). First, intersect confident peak coordinates (from Step 8) with coordinates of known binding sites (e.g., using intersectBed from BEDTools) to determine the peaks that do and do not overlap with the known sites. Then calculate the average conservation rate in both classes of peaks. We suggest using a set of known enhancers or previously defined ChIP regions[9].

(ii) *Peak location*. Calculate a conservation rate of peaks according to their genomic annotation (i.e., intergenic, intronic, 3′ untranslated region (UTR), 5′ UTR, 2 kb promoter, coding sequence (CDS)) using genome annotation data. First, use the annotation file (e.g., GFF file containing coordinates for each type of regions) to extract the relevant annotations. Next, annotate each genomic location uniquely using priorities for potentially overlapping annotations (e.g., first: CDS, second: 5′ UTR, third: 3′ UTR, fourth: intron and fifth: promoter as 2-kb regions upstream gene TSSs stopping at the next gene; rest: intergenic). Overlap the confident peak regions (from Step 8) with those annotations (i.e., intersect region coordinates using intersectBed from BEDTools) and assign each peak to a specific annotation if at least 50% of the peak's region overlaps with this annotation. For each annotation type, calculate the conservation rate of all associated peaks.

(iii) *Peak-to-TSS and peak-to-peak distance*. Calculate a conservation rate of confident peaks (from Step 8) according to their distance to the nearest gene TSS and the distance to the nearest neighboring peak (e.g., using closestBed from BEDTools). Distance bins can be defined as 0–0.5, 0.5–2, 2–5, 5–10, 10–20 and >20 kb. Note that each bin will contain a different number of peaks, such that only the relative number of conserved peaks (i.e., the conservation rate) can be meaningfully compared.

© 2011 Nature America, Inc. All rights reserved.

npg

(iv) *Peak-to-gene assignment*. Assign each confident peak (from Step 8) to its closest gene TSS (e.g., using closestBed from BEDTools). If insulator data for the corresponding condition are available, assign each peak to its closest gene TSS only within regions separated by insulators[9]. Note that some peaks will not be assigned to any gene and some genes will have multiple peaks assigned to them.

(v) *Expression analysis*. Compare the conservation rates of peaks and control peaks (from Step 8) assigned to genes that are in particular functional groups. To analyze how conservation of binding correlates with genes that are regulated by the transcription factor, we suggest using expression data for the transcription factor.

(vi) *GO analysis*. Compare the conservation rates of peaks and control peaks (from Step 8) assigned to genes in different GO categories (e.g., GO categories assigned to the function of the studied transcription factor).

▲ CRITICAL STEP Be careful to not double-count peaks for a given category.

(B) Sequence analysis ● TIMING ~2 h

(i) De novo *motif discovery*. Search confident peaks (from Step 8) *de novo* for motifs (e.g., using MEME-ChIP). If the samples are compared across species (i.e., in different genomes) and multiple sequence alignments (e.g., from UCSC) are available, search for k-mers that are substantially more highly conserved in peaks than in control peaks (from Step 8).

▲ CRITICAL STEP For Steps 14B(i–iii), use a fixed length of peak regions centered on the peaks' summits to avoid biasing the analysis toward longer peak regions (e.g., the average length of the genomic fragments).

(ii) *Known motif search*. By using known motif PWMs (e.g., motifs from TRANSFAC[34] and/or JASPAR[35] databases), search confident peak regions (from Step 8) for overrepresented motifs (e.g., using MAST) compared with their control motifs (i.e., shuffling columns of motif PWMs) or in control peaks (from Step 8).

(iii) *Sequence conservation*. If multiple sequence alignments (e.g., from UCSC) are available, calculate the sequence conservation of confident peak regions, control peak regions (from Step 8) and individual motif occurrences within those peak regions. We use both the PhastCons score and sequence identity calculated from the multiple sequence alignment[9]. Convert the PhastCons WIG file from UCSC to a BED file and fill in missing genomic positions with 'zero', intersect it with the peak region (e.g., using intersectBed from BEDTools) and calculate an average PhastCons score for each peak region. Identify motifs and control motifs (i.e., shuffling columns of motif PWMs) for the transcription factor of interest and its partners that are substantially more conserved in conserved peaks than in condition-specific peaks, the average genome and control peaks (from Steps 8 and 9). For data in different species, assess the type of motif sequence changes (i.e., mutations, insertions and deletions) in the multiple sequence alignment. In addition, for each binding changes category (from quantitative changes at Step 13), assess the change in quality of their motifs using the differential motif scores (e.g., using MAST).

? TROUBLESHOOTING

Troubleshooting advice is provided in **Table 1**.

TABLE 1 | Troubleshooting table.

Step	Problem	Possible reason	Solution
4, **Box 1**	Program takes a long time to run	Large input files	Run the program for each input file in parallel and/or split the input file(s) into several smaller files to further parallelize the task
8	No peaks found at FDR threshold of 1%	FDR estimates the fraction of random (i.e., likely to be wrong) peaks among the final peaks and is often estimated empirically (e.g., by MACS)	An FDR of 5% is also acceptable. If still no peaks are found, the ChIP sample might be of poor quality (e.g., low signal-to-noise ratios) or have a low read coverage
	Errors in coordinates	BED format is 0-based half-open and yet many other formats are 1-based closed	Adjust your code accordingly
10	Low conservation of binding sites across species	Some peaks are located in regions that cannot be uniquely mapped or translated	This problem leads to an underestimation of overall binding conservation

● TIMING

Steps 1 and 2, Data preprocessing: ~10 min
Steps 3–6, Read mapping and visualization: ~1 h
Box 1, Translation to common coordinates: A, ~36 h; or B, ~8 h

© 2011 Nature America, Inc. All rights reserved.

Step 7, Assessing global reproducibility and similarity: ~15 min
Steps 8–10, Peak calling and conservation analysis: ~30 min
Steps 11–13, Analysis of quantitative changes: ~45 min
Step 14A, Functional analysis: ~1 h
Step 14B, Sequence analysis: ~2 h
This timing estimation is given only according to the time necessary to run the code and programs in parallel for the *Drosophila* test set data and using our computational resources. Data from larger genomes will take longer to run especially if coordinates are to be translated.

ANTICIPATED RESULTS

Data preprocessing

The median quality score of the reads should be around 40 and stay stable or at most slightly degrade along the read length (e.g., to around 20). The nucleotide distribution should be equally distributed with only very few unknown nucleotides (Ns, typically below 1%). A bias might stem from the overrepresentation of a unique read that is repeated many times (e.g., the linker sequence). Deviations might explain low read-matching frequencies in later steps (Steps 4 and 5).

A high percentage of unique reads (we typically find ≥50% for ChIP samples and ≥ 75% for input samples in *Drosophila*) is a good sign, although it can decrease with very high numbers of reads (around 20 million or more) and small genomes (e.g., yeast, *C. elegans* or *Drosophila*). It is also lower for ChIP-seq experiments with very high signal-to-noise ratios, in which many reads are confined to specific regions of the genome. A low percentage of unique reads (below 50%) may indicate that the library was prepared from too little DNA and/or that PCR amplification artifacts occurred.

Read mapping

The percentage of mapped reads and the percentage of unique read coordinates should be as high as possible. Reads that cannot be mapped might be linker sequences, sample contaminations or low-complexity sequences, which correspond to repeated regions of the genome that are more frequent in vertebrates than in *Drosophila*. To provide a range of expected numbers for mapped reads, **Table 2** and **Supplementary Table 1** show the number of raw reads, mapped reads and unique reads for the *Drosophila* Twist test data set[9] and for one vertebrate transcription factor data set[10]. Between 44% and 75% of the reads in vertebrates and 75% and 81% in *Drosophila* of the raw reads could be mapped.

To assess more systematically the uniqueness of genome sequences in the *Drosophila* genome independent of any ChIP-seq experiment, we also determined the percentage of all potential 36-nucleotide-long reads (i.e., all 36mers created from the reference genomes in one-nucleotide steps) that could be mapped back uniquely to the respective genome using Bowtie[37] and the genome coverage they represent (**Supplementary Table 2**). The number of mapped reads and the corresponding genome coverage are high in all species. For *D. melanogaster*, we also mapped 18-nucleotide-long potential reads (i.e., shorter reads) yielding a minor decrease in genome coverage. Note that the current assembly state of the *D. erecta* and *D. ananassae* genomes (5,124 and 13,749 scaffolds, respectively) can explain their lower genome coverage.

© 2011 Nature America, Inc. All rights reserved.

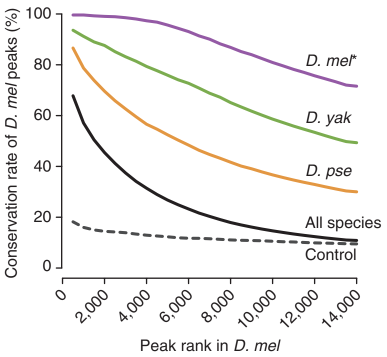
TABLE 2 | Results for the test data set we provide.

He *et al.*[9], Twist	Raw reads	Mapped reads	Unique reads	Translated reads	Confident peaks	Enriched regions	Conservation (%)
chip_dmel	6924965	5631684	3901384	—			
		81%	69%		3447	10352	
input_dmel	8594417	6975843	6072881	—			
		81%	87%				74
chip_dyak	8784288	6593674	4614002	4957524			Control 6
		75%	70%	75%	758	9126	
input_dyak	12567553	10016367	7974239	7284163			
		80%	80%	73%			

Note that the results from the test data set differ from the ones from He *et al.*[9] because of the use of a different read mapper and a different version of the peak-calling program MACS.

© 2011 Nature America, Inc. All rights reserved.

Figure 4 | Binding conservation at different peak ranks. Conservation of *D. melanogaster* peaks decreases with peak rank and evolutionary distance of the compared species. *D. mel*, *D. melanogaster*; *D. yak*, *D. yakuba*; *D. pse*, *D. pseudoobscura*; *, replicate. Data are from He *et al.*[9].



Translation to common coordinates

A general concern is the sensitivity by which genome coordinates can be translated (e.g., using LiftOver). Independent of any ChIP-seq experiment, we determined the percentage of all potential 36-nucleotide-long reads and the corresponding ones that could be remapped from various *Drosophila* species (see above) that could be unambiguously translated into *D. melanogaster* coordinates for cross-species comparisons (**Supplementary Table 3**). The numbers are shown for all potential reads or only those that can be uniquely mapped back to the genome. The fraction of translatable reads generally drops with further distant species, as expected, given the lower genome sequence similarity. These numbers are similar to actual numbers from ChIP-seq experiments. **Table 2** shows the number of translated reads for the *Drosophila* test data set, and **Supplementary Table 1** shows the number of reads from vertebrate transcription factor ChIP-seq data that can be translated to the human genome. For a given species, the numbers are remarkably constant for different data sets; e.g., ~50% read translation from mouse to human.

Although read translation generally works well, it is also clear that some genomic regions cannot be translated between different genomes, thereby potentially leading to an underestimation of conservation of the reference species' binding sites. When analyzing more distant species, lowering LiftOver's minmatch parameter, i.e., the minimum percent sequence identity required between regions, might help.

In general, we found that using a common reference genome worked well in comparative ChIP-seq analyses.

PCC analysis

The PCC between biological replicates measures the reproducibility between experiments and provides an upper bound for the global similarity of binding across conditions or species (e.g., ≥0.9 in our experience). The pairwise PCC between ChIP samples and input samples serves as a lower bound for the global similarity of binding. Note that there is usually a positive PCC between any two samples (e.g., approximately 0.3–0.4 in our experience) because of similar chromatin accessibility and intrinsic biases in the experimental procedure. Most notably, the DNA is not fragmented randomly during the sonication step[46,47], and CG-rich fragments are favored during the PCR amplification and/or the cluster generation step during next-generation sequencing[48]. The difference between the upper bound and lower bound of the PCCs also serves as quality control for the ChIP-seq data set. In a high-quality ChIP experiment, the PCC between two replicate ChIP samples far exceeds the correlation with input sample (e.g., 0.9 versus 0.4). Poor ChIP samples, on the other hand, more closely resemble the input sample (see Experimental design).

Peak calling and conservation analysis

The number of called peaks for the *Drosophila* test data set we provide with this protocol are found in **Table 2**. When calculating conservation estimates, it is important to check that biological replicates have high binding conservation rates close to 100%, and that control peaks have low conservation rates (~10%) depending on the genome size.

Figure 3 shows analyses and anticipated results to assess whether the chosen thresholds yield adequate conservation rates. When merely overlapping high-confident peaks from two samples, a large number of peaks appear nonconserved even between biological replicates (**Fig. 3a**, gray). These apparently nonconserved peaks have high read counts relative to the genome average (**Fig. 3b**), and these are specifically located at the position corresponding to the peak summit of the reference sample (**Fig. 3c**). This argues that these peaks are

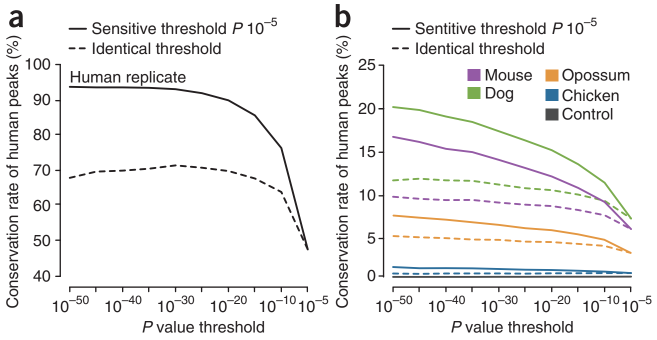


Figure 5 | Sensitive estimation of conservation in vertebrates. (**a**) Assessing CEBPA peaks across biological replicates with sensitive thresholds ('enrichment'; solid line) recovers most peaks across a wide range of *P* value thresholds, whereas requiring identical thresholds in both experiments does not (dashed line). (**b**) Assessing CEBPA peak conservation using significant enrichment across different vertebrate species (solid lines) is more sensitive than using identical thresholds in both species (dashed line). Note that the increased conservation estimates cannot explain the recently observed differences in conservation of transcription factor binding between flies[9] and vertebrates[10].

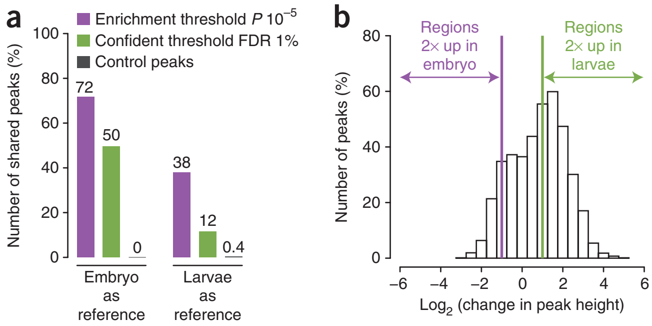


Figure 6 | Comparative ChIP-seq data of the *C. elegans* transcription factor PHA4/FOXA across conditions. (**a**) Binary analysis: Using significant enrichments (ChIP-signals; purple) results in a more sensitive assessment of shared binding peaks compared with an FDR-corrected threshold (required during peak calling; green). Random control peaks suggest that the number of shared peaks is not overestimated (black). As the number of peaks in each condition is different (3,011 peaks in larvae versus 704 peaks in embryos at an FDR of 1, respectively), the fraction of shared peaks (numbers above bars) between the two conditions is asymmetric. (**b**) Quantitative analysis: a histogram of the quantitative changes in binding (changes in peak heights measured as $\log_2$ fold change) between the embryo and larvae is shown and the thresholds for twofold changes are highlighted. Data are from Zhong *et al.*[30].

in fact conserved, and that their conservation has been missed because of overly stringent thresholds. In contrast, when assessing conservation via enriched read counts, we find 98% conservation for biological replicates (**Fig. 3d**, purple) and 81% and 60% across species, respectively (**Fig. 3d**, green and orange; random regions are gray). We conclude that our approach yields accurate and sensitive conservation estimates.

Note, however, that peak conservation estimates decrease with peak ranks (**Figs. 4** and **5**), and thus depend on the total number of peaks in the reference sample. This likely results from two trends, namely the increasing number of false-positive peaks at lower enrichments in the reference data set and the decreasing ability to discriminate truly conserved peaks from noise in the second data set.

We have also tested and confirmed that our approach works similarly well for other data sets, including for comparative analyses in vertebrates (**Fig. 5**) and for analyzing condition-specific binding of a transcription factor in *C. elegans* (**Fig. 6**). In vertebrate comparative analyses, our approach to making comparisons across data sets with more sensitive thresholds performs better than merely intersecting peaks called at identical thresholds (**Fig. 5**) and allows a sensitive assessment of peak conservation for a wide range of thresholds. It also allows a more sensitive assessment of the number of shared PHA4/FOXA-binding peaks between embryos and larvae in the *C. elegans* data sets (**Fig. 6a**). Note that the number of peaks is higher in the larva sample than in the embryo, and thus the fraction of shared peaks differs depending on which condition is used as a reference. For example, the ChIP-seq quality may differ between samples and produce different numbers of peaks during peak calling. In such a case, conservation estimates appear to differ depending on which sample is chosen as a reference sample[9]. However, we found no evidence that the size of the genome (e.g., the large size of the *D. ananassae* genome) produces a bias in the conservation rates when chosen as reference.

Quantitative changes analysis

Results from the analysis of quantitative changes of Twist binding between *Drosophila* species have been published[7,11]. To show the applicability of these results across conditions, we used our approach to analyze the condition-specific binding of the *C. elegans* transcription factor PHA4/FOXA. **Figure 6b** shows a histogram of the fold change in read-count enrichments between the two stages. There are more regions that are more than twofold bound in larvae than in embryos, which is consistent with the increased number of peaks detected in larvae.

© 2011 Nature America, Inc. All rights reserved.

Note: Supplementary information is available via the HTML version of this article.

ACKNOWLEDGMENTS We thank M. Jaritz; I. Tamir; A. Sommer; O. Yanez-Cuna; D. Gerlach (Institute of Molecular Pathology); J. Steinmann (Institute of Molecular Biotechnology (IMBA)); and S. Meier and C. Seidel (Stowers Institute for Medical Research) for discussions, help and advice. A.F.B. was supported by the Austrian Ministry for Science and Research through the Genome Research in Austria (GEN-AU) Bioinformatics Integration Network III. J.Z. is a Pew scholar. A.S. is supported by a European Research Council (ERC) Starting Grant from the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 242922. Basic research at the IMP is supported by Boehringer Ingelheim.

AUTHOR CONTRIBUTIONS A.F.B. and A.S. established the analysis pipeline. Q.H. and J.Z. performed the comparative ChIP-seq experiments. A.F.B., A.S. and J.Z. wrote the manuscript.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

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

1. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
2. Iyer, V.R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).
3. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
4. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).
5. Sandmann, T. *et al.* A core transcriptional network for early mesoderm development in *Drosophila melanogaster*. *Genes Dev.* **21**, 436–449 (2007).
6. Zinzen, R.P., Girardot, C., Gagneur, J., Braun, M. & Furlong, E.E.M. Combinatorial binding predicts spatio-temporal *cis*-regulatory activity. *Nature* **462**, 65–70 (2009).

© 2011 Nature America, Inc. All rights reserved.

7. 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).
8. Palii, C.G. *et al.* Differential genomic targeting of the transcription factor TAL1 in alternate haematopoietic lineages. *EMBO J.* **30**, 494–509 (2011).
9. He, Q. *et al.* High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species. *Nat. Genet.* **43**, 414–420 (2011).
10. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036–1040 (2010).
11. Bradley, R.K. *et al.* Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related *Drosophila* species. *PLoS Biol.* **8**, e1000343 (2010).
12. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42**, 631–634 (2010).
13. Mikkelsen, T.S. *et al.* Comparative epigenomic analysis of murine and human adipogenesis. *Cell* **143**, 156–169 (2010).
14. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
15. Wilbanks, E.G. & Facciotti, M.T. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE* **5**, e11471 (2010).
16. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B (Methodological)* **57**, 289–300 (1995).
17. Noble, W.S. How does multiple testing correction work? *Nat. Biotechnol.* **27**, 1135–1137 (2009).
18. Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S. & Hirschhorn, J.N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* **33**, 177–182 (2003).
19. Toth, J. & Biggin, M.D. The specificity of protein-DNA crosslinking by formaldehyde: *in vitro* and in *Drosophila* embryos. *Nucleic Acids Res.* **28**, e4 (2000).
20. Lettice, L.A. *et al.* A long-range *Shh* enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735 (2003).
21. Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M. & Shiroishi, T. Elimination of a long-range *cis*-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development* **132**, 797–803 (2005).
22. Hong, J.-W., Hendrix, D.A. & Levine, M.S. Shadow enhancers as a source of evolutionary novelty. *Science* **321**, 1314 (2008).
23. Nègre, N. *et al.* A comprehensive map of insulator elements for the *Drosophila* genome. *PLoS Genet.* **6**, e1000814 (2010).
24. Cuddapah, S. *et al.* Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.* **19**, 24–32 (2009).
25. Stanley, S.M., Bailey, T.L. & Mattick, J.S. GONOME: measuring correlations between GO terms and genomic positions. *BMC Bioinformatics* **7**, 94 (2006).
26. Zeitlinger, J. & Stark, A. Developmental gene regulation in the era of genomics. *Dev. Biol.* **339**, 230–239 (2010).
27. Borneman, A.R. *et al.* Divergence of transcription factor binding sites across related yeast species. *Science* **317**, 815–819 (2007).
28. Zheng, W., Zhao, H., Mancera, E., Steinmetz, L.M. & Snyder, M. Genetic analysis of variation in transcription factor binding in yeast. *Nature* **464**, 1187–1191 (2010).
29. Meireles-Filho, A.C.A. & Stark, A. Comparative genomics of gene regulation-conservation and divergence of *cis*-regulatory information. *Curr. Opin. Genet. Dev.* **19**, 565–570 (2009).
30. Zhong, M. *et al.* Genome-wide identification of binding sites defines distinct functions for *Caenorhabditis elegans* PHA-4/FOXA in development and environmental response. *PLoS Genet.* **6**, e1000848 (2010).
31. Kim, T.H. & Ren, B. Genome-wide analysis of protein-DNA interactions. *Annu. Rev. Genomics Hum. Genet.* **7**, 81–102 (2006).
32. Zeitlinger, J. *et al.* Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell* **113**, 395–404 (2003).
33. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
34. Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31**, 374–378 (2003).
35. Sandelin, A., Alkema, W., Engström, P., Wasserman, W.W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**, D91–D94 (2004).
36. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
37. 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).
38. Horner, D.S. *et al.* Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Brief. Bioinformatics* **11**, 181–197 (2010).
39. Li, H. *et al.* The sequence alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
40. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
41. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
42. Bailey, T.L., Williams, N., Misleh, C. & Li, W.W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **34**, W369–W373 (2006).
43. Bailey, T.L. & Gribskov, M. Combining evidence using *P*-values: application to sequence homology searches. *Bioinformatics* **14**, 48–54 (1998).
44. Das, M.K. & Dai, H.-K. A survey of DNA motif finding algorithms. *BMC Bioinformatics* **8** (Suppl. 7): S21 (2007).
45. Tompa, M. *et al.* Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* **23**, 137–144 (2005).
46. Auerbach, R.K. *et al.* Mapping accessible chromatin regions using Sono-Seq. *Proc. Natl. Acad. Sci. USA* **106**, 14926–14931 (2009).
47. Teytelman, L. *et al.* Impact of chromatin structures on DNA processing for genomic analyses. *PLoS ONE* **4**, e6700 (2009).
48. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).

A computational pipeline for comparative ChIP-seq analyses

Anaïs F. Bardet[1], Qiye He[2], Julia Zeitlinger[2,3#,] Alexander Stark[1#]

1 Research Institute of Molecular Pathology (IMP), Vienna

2 Stowers Institute for Medical Research, Kansas City;

3 Department of Pathology, University of Kansas Medical School, Kansas City

# Corresponding authors (jbz@stowers.org, stark@starklab.org)

Lab homepages: http://www.starklab.org, http://research.stowers-institute.org/zeitlingerlab/

Table S1: Performance of coordinate translation in vertebrates

Schmidt et al.[1]	**Mapped reads**	**Translated reads**
CEBPA Mouse to Human	6938464 44%	3792526 55%
CEBPA Dog to Human	6892619 47%	5368885 78%
CEBPA Opossum to Human	3051354 72%	632429 21%
CEBPA Chicken to Human	6631164 61%	630219 9%
Kunarso et al.[2]	**Mapped reads**	**Translated reads**
CTCF Mouse to human	3686056	1947048 53%
NANOG Mouse to human	8424102	4383803 52%
OCT4 Mouse to human	4911144	2553512 52%
Mikkelsen et al.[3]	**Mapped reads**	**Translated reads**
CTCF Mouse to human	8417901	4164699 49%
PPARγ Mouse to human	9887007	4240753 43%

Table S2: Read mapping sensitivity for *Drosophila* species

	***Drosophila* species**	***D. melanogaster*** dm3	***D. melanogaster*** (shorter reads)	***D. simulans*** droSim1	***D. yakuba*** droYak2	***D. erecta*** droEre2	***D. ananassae*** droAna3	***D. pseudoobscura*** dp4
Potential reads	Mapped reads (%)	90.19	85.17	90.25	94.17	80.74	63.01	84.67
	Genome coverage (%)	91.47	89.93	91.07	95.60	83.38	68.67	87.55

Percent of all potential 36 nucleotide long reads that can be uniquely mapped back to the genome and corresponding genome coverage.

Table S3: Read translation sensitivity for *Drosophila* species

	Drosophila* species to *D. melanogaster	***D. simulans*** droSim1 to dm3	***D. yakuba*** droYak2 to dm3	***D. erecta*** droEre2 to dm3	***D. ananassae*** droAna3 to dm3	***D. pseudoobscura*** dp4 to dm3
Potential reads	Translated reads (%)	87.74	88.56	79.65	56.26	56.64
	Genome coverage (%)	95.82	97.39	84.50	51.83	64.29
Mapped reads	Translated reads (%)	95.85	95.87	89.30	70.47	61.33
	Genome coverage (%)	95.29	76.48	82.25	50.28	62.31

Percent of all potential 36 nucleotide long reads or previously mapped reads (Table S2) that can be translated to the *D. melanogaster* genome and corresponding genome coverage.

REFERENCES

1. Schmidt, D. et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036–1040 (2010).
2. Kunarso, G. et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**, 631–634 (2010).
3. Mikkelsen, T.S. et al. Comparative Epigenomic Analysis of Murine and Human Adipogenesis. *Cell* **143**, 156–169 (2010).

Manuscript D

Identification of transcription factor binding sites from ChIP-seq data at high-resolution

Bardet AF*, Steinmann J*, Bafna S, Knoblich JA, Zeitlinger J, Stark A.
(submitted)

* contributed equally

A.F.B., J.S., J.A.K., J.Z., and A.S. conceived the project. J.S. wrote peakzilla. **A.F.B.** benchmarked peakzilla and performed all computational analyses. **A.F.B.** and A.S. analyzed the data. S.B. and J.Z. designed and performed the high-resolution ChIP-seq experiments. **A.F.B.**, J.Z. and A.S. wrote the manuscript.

Identification of transcription factor binding sites from ChIP-seq data at high-resolution

Anaïs F. Bardet*[1], Jonas Steinmann*[2], Sangeeta Bafna[3,4], Juergen A. Knoblich[2], Julia Zeitlinger[3], Alexander Stark[1]#

[1]Research Institute of Molecular Pathology (IMP), Vienna, Austria

[2]Institute of Molecular Biotechnology (IMBA), Vienna, Austria

[3]Stowers Institute for Medical Research, Kansas City, Missouri, USA

[4]Current address: Department of Medicine, Vanderbilt University, Nashville, TN, USA

*These authors contributed equally

#Corresponding author (stark@starklab.org)

Lab homepage: http://www.starklab.org

Running title: ChIP-seq peak calling at high resolution

Keywords: ChIP-seq, transcription factors, peak finder, high resolution

ABSTRACT

Chromatin-immunoprecipitation coupled to next-generation sequencing (ChIP-seq) is widely used to study the in vivo binding sites of transcription factors (TFs) and their regulatory targets. Recent methodological improvements to ChIP-seq, such as the increased sequencing depth and resolution - especially of the ChIP-exo variant - promise deeper insights into transcriptional regulation, yet require novel computational tools to fully leverage their advantages.

To this aim, we have developed peakzilla, which fully exploits the characteristics of ChIP-seq read density distributions to identify closely spaced TF binding sites at high resolution. We perform ChIP-seq for the TF Twist in early Drosophila embryos with three different experimental fragment sizes and demonstrate that peakzilla makes full use of the associated increase in resolution and compares favorably to established methods. It similarly leverages the increased resolution of ChIP-exo and complements high resolution with high precision in determining the exact location of the TFBSs at the peak summits.

We show that the increased resolution of peakzilla has immediate benefits for our understanding of transcriptional regulation as closely spaced Twist binding sites are highly enriched in functional embryonic enhancers. Peakzilla is easy to use as it learns all the necessary parameters from the data and is freely available.

http://github.com/steinmann/peakzilla/

http://www.starklab.org/data/peakzilla/

INTRODUCTION

Gene expression is mainly regulated at the transcriptional level and achieved through the binding of transcription factors (TFs) to genomic regulatory regions such as promoters and enhancers. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is extensively used to determine transcription factor binding sites (TFBSs) genome-wide (Johnson et al. 2007; Robertson et al. 2007). Compared to ChIP-chip (Ren et al. 2000; Iyer et al. 2001), ChIP-seq has dramatically improved the resolution of the identified TFBSs (from hundreds to only tens of nucleotides). A recent refinement of ChIP-seq, the ChIP-exo method (Rhee and Pugh 2011), further increases the resolution of ChIP-seq experiments, theoretically to single nucleotides.

The specific features and strategies of TF ChIP-seq data analysis have been well described by several reviews (e.g. (Pepke et al. 2009)). Briefly, as typically only one of both ends of the immunoprecipitated DNA fragments is sequenced, the strand-specific sequencing reads (or sequence tags) result in a bimodal distribution that is characteristic for true TFBSs (**Figure 1A**). This distribution is used to estimate the size of the fragments, and together with this information, the locations of TFBSs are predicted across the genome.

A large variety of computational tools have been developed and are successfully used to predict such binding events (Wilbanks and Facciotti 2010). However, in our experience, these tools are not optimized to take advantage of recent methodological improvements, which comprise paired-end sequencing, high sequencing depth (e.g. on Illumina HighSeq systems), and – most importantly – an increase in experimental resolution, both of conventional ChIP-seq and ChIP-exo. Many tools merge closely spaced read-density peaks into large regions, thereby loosing the ability to distinguish individual binding sites (i.e. resolution). Furthermore, many methods also lack precision as measured by the distance from the inferred TFBSs (i.e. the reported peak summit) to the TFs' sequence motifs. Resolution and precision are crucial when determining TFBSs, as promoter and enhancer regions often consist of multiple TFBSs for the same TF (homotypic TFBSs clusters) (Lifanov et al. 2003; Gotea et al. 2010; He et al.

2011) or different TFs (Berman et al. 2002; Schroeder et al. 2004). Thus, to fully leverage current ChIP methodologies towards understanding the structure and function of enhancers, the ability to determine multiple closely spaced TFBSs is critical. To meet this need, we developed a high-resolution method that could identify binding peaks at improved resolution and precision.

Here, we present a new computational tool, peakzilla, which fully exploits the bimodal distribution of sequence reads characteristic of true TF binding events, to identify closely adjacent TF binding sites with high resolution and precision.

We evaluate peakzilla by comparing it to the first generation of methods such as MACS (Zhang et al. 2008), QuEST (Valouev et al. 2008), and cisGenome (Ji et al. 2008), as well as some methods specifically developed for the detection of high resolution peaks such as spp (Johnson et al. 2007; Kharchenko et al. 2008; Robertson et al. 2007), SISSRs (Ren et al. 2000; Jothi et al. 2008; Iyer et al. 2001), GPS (Rhee and Pugh 2011; Guo et al. 2010) and PeakRanger (Pepke et al. 2009; Feng et al. 2011). Peakzilla shows superior resolution and precision on conventional ChIP-seq datasets from *S. Cerevisiae* (Wilbanks and Facciotti 2010; Zheng et al. 2010), *C. elegans* (Lifanov et al. 2003; Zhong et al. 2010; Gotea et al. 2010; He et al. 2011), *D. melanogaster* (Berman et al. 2002; He et al. 2011; Schroeder et al. 2004), mouse (Zhang et al. 2008; Schmidt et al. 2010), and human (Valouev et al. 2008; Kasowski et al. 2010; Cuddapah et al. 2009) and on recent ChIP-exo datasets from human (Ji et al. 2008; Rhee and Pugh 2011). We also show experimentally that peakzilla takes full advantage of increased experimental resolution of small fragment sizes by performing ChIP-seq experiments for Twist in *D. melanogaster* at normal, medium, and high resolution. These results suggest that peakzilla is ideally suited for the identification of TFBSs with recent ChIP methods.

RESULTS

Peakzilla algorithm

Peakzilla uses the bimodal distribution of the reads (**Figure 1A**) not only to estimate the fragment length but also to weight the read counts during peak calling and to score the candidate TFBSs. This has two important advantages: first, it enables peakzilla to more clearly discriminate between reads from adjacent TFBSs, leading to a substantial increase in resolution compared to treating reads irrespective of their directionality. Second, it avoids false positives that originate from artifacts during library preparation or sequencing, without the need to collapse or down weight reads that map to identical genomic positions (**Figure 1B** and **Supplementary figure 1**). This is especially important when working with high sequence coverage, as obtained for small genomes (e.g. yeast, flies, or nematodes) and with modern NGS sequencers (Chen et al. 2012). Finally, peakzilla can also be used on paired-end ChIP-seq data, in which case the estimated fragment size is directly averaged from the mapped reads.

Peakzilla first scans the genome for candidate TFBSs that show high coverage in sequencing reads of the IP sample compared to the control sample (note that the control sample is optional, allowing peakzilla to be used with ChIP-exo). It then scores the candidates by the normalized read count of IP sample (minus the control sample if available). To discriminate between artifacts and true binding events in enriched regions, each candidate TFBS score is further weighted by a distribution score that estimates how well the observed distribution of the reads in the peak region fits to a model for the bimodal read distribution (**Figure 1C & D**). Indeed, further analysis suggests that candidates which are penalized using this distribution score are likely false since they are significantly less enriched in the corresponding TF motif (**Figure 1E**) and contain substantially less diverse sequence reads (**Figure 1F** and **supplementary figure 2**), i.e. their high read densities stem from only a few highly duplicated sequences, which are likely amplification artifacts. The method is illustrated in the method section and in **supplementary figure 3**.

Overall performance

To evaluate peakzilla, we performed a pairwise comparison with other analysis methods on diverse ChIP-seq datasets from *S. cerevisiae*, *C. elegans*, *D. melanogaster*, mouse and human. Although the number of genomic regions that contain peaks found by the different methods is highly dependent on the thresholds and parameters used, the agreement is overall very good (**Supplementary figure 4**), demonstrating the maturity of available tools for 'peak calling'. For example, all known Twist enhancers were identified by all methods, except for four and three enhancers not found by QuEST and GPS, respectively. Among all differential peaks (i.e. peaks identified by only one of the methods), those found exclusively by peakzilla are significantly more highly enriched in TF motif occurrences (**Figure 2A** and **supplementary figure 5**) and have higher fold enrichments of ChIP over input than peaks found exclusively by any of the other methods (**Figure 2B** and **supplementary figure 5**). For example, 31.7% of the best peakzilla peaks are not found by SISSRs, but are significantly enriched in Twist motifs ($P<10^{-34}$) and have a significantly higher fold enrichments ($P<10^{-16}$) compared to the 7% of best SISSRs peaks not found by peakzilla.

High precision of peakzilla peaks

When identifying TFBSs at high resolution, the correct prediction of the precise TFBSs' location is important and critical for subsequent analyses of sequence features of TF binding. TFBSs identified by peakzilla contain more often than other methods the corresponding TF motif in their peak regions (**Supplementary figure 6**). More importantly, the summits of the peak regions are on average closer to the nearest motif occurrence (**Figure 2C**), arguing that peakzilla has higher precision than other methods.

High resolution of peakzilla peaks

The main strength of peakzilla is its ability to find peaks at high resolution. First, as the locations of sequencing reads that originate from a single TFBS are limited by the fragment size used for library preparation, we report peak regions as the average fragment size centered on the summit position. This is different from the large peak regions reported by MACS and to a lesser extend QuEST, CisGenome and PeakRanger and from reporting only the summit positions (SISSRs, spp; **Figure 3A**). Therefore, peakzilla is able to resolve closely spaced peaks by additionally allowing the reported peaks to overlap by up to half the fragment size. This is the highest resolution that can easily be obtained without losing the ability to uniquely assign reads to individual TFBSs. A further gain in resolution would require the deconvolution of overlapping read-distributions by model fitting, a computationally intensive approach used e.g. by GPS (Guo et al. 2010). When we globally analyzed the distance between neighboring peak summits, peakzilla is among the methods reaching the smallest peak-to-peak distance (i.e. resolution) together with SISSRs and GPS (**Figure 3B** and **Supplementary figure 7**).

Peakzilla splits a substantial amount of MACS peaks (e.g. 22% for Twist) into several peaks, each constituting a putative TFBS (**Supplementary figure 8**). Indeed, as expected for independent TFBSs, both the split peaks that correspond to the MACS summits (major peaks) and the minor peaks were significantly enriched for the Twist motif. Thus, many split MACS peaks may represent homotypic clusters of Twist binding sites. In addition, minor peaks frequently contained motifs for the TFs Snail and dorsal, which are known to cooperate with Twist and might have been co-precipitated after cross-linking (**Figure 3C**). Finally, MACS peaks split by peakzilla appear to be more often functional than MACS peaks that are not split (one-to-one peaks) and control regions (**Figure 3D**): TF binding to split peaks is more highly conserved in other *Drosophila* species (He et al. 2011) and they are significantly more enriched for known Twist or mesoderm enhancers than one-to-one peaks or controls (**Figures 3E** and **3F**). All together, these results suggest that peakzilla is ideal for identifying regions with multiple binding sites and that such information is important for detecting functional enhancers.

Peakzilla leverages increased resolution of ChIP experiments

To directly demonstrate peakzilla's ability to make use of increased experimental ChIP resolution, we performed conventional ChIP-seq for Twist from *Drosophila* embryos with increasingly smaller fragment sizes. For this, the chromatin was sonicated into relatively small DNA fragments and then further trimmed by DNase I digestion before ChIP (see Methods). This yielded three ChIP datasets with estimated fragment sizes of 102bp, 72bp, and 49bp, from which we called peaks with the different peak finders (**Figure 4A**). We expected that with decreasing fragment sizes, the width of the identified peaks should decrease and the resolution, i.e. the ability to resolve closely spaced binding sites should increase. Indeed, the peak regions reported by peakzilla, but not those reported by most other methods, showed decreased width with decreasing fragment sizes (**Figure 4A**). More importantly, the resolution, as measured by the minimal peak-to-peak distance (after removing 1% outliers), increased with decreasing fragment sizes for peakzilla, but not for other methods (**Figure 4B**). SISSRs, GPS, and peakzilla performed well on the small-fragment sample, with peakzilla reaching the highest resolution of all methods. These results demonstrate that the maximum benefit of experimental methods with higher resolution can only be obtained when used together with high-resolution computational methods such as peakzilla.

Peakzilla as a peak-caller for ChIP-exo data

The recently developed ChIP-exo method adds an lamda exonuclease digestion step after ChIP, which trims the 5' DNA strand until the cross-linked TFBS (Rhee and Pugh 2011). This digestion end point can be mapped to the genome using the remaining single-stranded overhang. Since each TFBS can be mapped from both sides, the resulting distribution of mapped breakpoints is also bimodal and resembles that of conventional ChIP-seq, with the 'fragment sizes' corresponding directly to the actual sizes of the TF footprints. To our knowledge, no computational method has been specifically developed for the analysis of ChIP-exo data.

We therefore assessed how well peakzilla and other methods perform on ChIP-exo datasets. We identified human CTCF binding sites from both ChIP-seq data (Cuddapah et al. 2009) and ChIP-exo data (Rhee and Pugh 2011). Peakzilla reported estimated fragment sizes of 98bp for ChIP-seq and 36bp for ChIP-exo (**Figure 4C**) and had the highest resolution (smallest peak-to-peak distance) among all methods tested (**Figure 4D**). This suggests that peakzilla is well suited for high-resolution ChIP-seq data, including ChIP-exo.

DISCUSSION

Understanding how combinations of TFs bind to DNA to regulate gene expression is one of the most pressing questions of today's biology. It's importance is witnessed by recent community efforts that aim to determine all functional elements in the genomes of model organisms and the human (e.g. ENCODE (ENCODE Project Consortium 2004), modENCODE (Celniker et al. 2009), Mouse ENCODE (Mouse ENCODE Consortium et al. 2012)). The availability of high-throughput sequencing at low cost widely promoted the use of the ChIP-seq methodology and an enormous number of datasets for different TFs from various species, developmental stages, or tissues are becoming available. This enables the identification of *in vivo* binding sites and thus enhancers that contain multiple binding sites for a single TF or multiple different TFs. While it is widely accepted that enhancers are characterized by clusters of TF binding motifs (Berman et al. 2002; Schroeder et al. 2004), it has remained less clear to what extent each of several clustered TF binding motifs is bound *in vivo*. Similarly, potential constraints on the relative distance or orientation of co-bound TFs have remained unclear, yet might be crucial to understand the molecular mechanisms and to decipher the sequence basis of gene regulation.

To experimentally address this question, it is important to resolve closely spaced binding sites and precisely predict their location from ChIP-seq data, challenges which current improvements to the ChIP methodology have started to address (e.g. ChIP-exo (Rhee and Pugh 2011)). However, while many computational tools exist to identify enriched regions (peaks) from ChIP data ('peak calling'), many of them are not designed to fully leverage these improvements, e.g. the increased resolution or the vastly increased number of deep sequencing reads of modern deep sequencing (Chen et al. 2012). To meet these challenges, especially the need of discovering TFBSs at high resolution, we have developed a new computational method, peakzilla.

The importance of high resolution and precision is also supported by alternative efforts to correctly position predicted TFBSs, for example by taking the location of enriched sequence motifs into account (Guo et al. 2012; Boeva et al. 2010; Wu et al. 2010). Although TFBSs predicted by peakzilla coincide very well with the established sequence motifs of the respective TFs, it is important to note that we chose to predict the TFBS locations from the ChIP-seq data alone, without taking sequence motifs into account: it is well established that within TFBSs, motifs of different TFs can be more highly enriched than motifs of the IPed TF itself. For example, the binding sites of most TFs in the early Drosophila embryo are highly enriched in motifs for the TF Zelda (Li et al. 2008; Bradley et al. 2010; Satija and Bradley 2012; Kvon et al. 2012; Yáñez-Cuna et al. 2012; modENCODE Consortium et al. 2010), such that the Zelda motif – which is partly more highly enriched than the motif of the TF of interest – might bias the correct prediction of the TFBSs and possibly hinder the study of relative positioning and orientation of TFBSs.

The combination of maximum experimental resolution and a peak caller like peakzilla thus make full use of recent ChIP-seq approaches and will be invaluable for testing hypotheses on how combinatorial TF binding realizes the developmental blueprint encoded in the regulatory regions of our genomes. Indeed, closely spaced Twist binding sites resolved by peakzilla coincided and were strongly enriched in known enhancers, corroborating the prevalent model that functional enhancers are characterized by clusters of TFBSs. The increasing number of ChIP studies that determine the *in vivo* binding sites of transcription factors at high resolution will prove invaluable for our understanding of enhancer function and transcriptional regulation.

METHODS

Peakzilla algorithm. Initially, peakzilla reads the coordinate files of the mapped reads of the IP and – optionally – control sample for single (BED format) or paired-end (BEDPE format) deep sequencing data (the paired-end data will be collapsed to unique chromosomal coordinates which correspond to independent fragments).

Peakzilla then first determines the average fragment size of the sequencing library to determine the peak size that should result from a true TFBS. For paired-end data, this corresponds directly to the average fragment size (i.e. the average distance of the two mapped ends of each fragment). For single-end data, the average fragment size is estimated from the shift size of plus and minus tags in the top 200 enriched regions in the ChIP sample as described before (Zhang et al. 2008). Peakzilla then defines peak size as two times the fragment size, as all reads from the ends of fragments IPed due to a single TFBS will on average lie in this region.

In a second step, the distribution of plus and minus strand reads that are to be expected is modeled. By default, two normal distributions are used with standard deviations stdev = peak-size / 5 and locations of their means at 1/4th and 3/4th of the peak-size, respectively. Alternatively, the user can choose to estimate the model empirically from the average distribution of reads within the top 200 candidate peaks in the ChIP.

To call TFBSs peakzilla first scans the genome counting reads within a 'double-window': each putative TFBS receives the counts of positive strand reads within a window of the fragment size downstream of the TFBS and the negative strand reads within an equivalent window upstream of the TFBS. Positions with 10 or more counted reads are scored as candidate peak summits with a raw score defined as the normalized read count minus the normalized read count in the control sample (e.g. input; note that the correction with a control sample is optional). Final peaks are the candidates with summits that are local maxima at least one fragment length (1/2 peak size) apart from each other. This scanning mode allows for both fast and comprehensive investigation of large genomes at single base resolution.

To obtain a final peak score, each raw score is corrected with a multiplicative distribution score [0..1] that assesses the fit of the observed read count distribution to the distribution expected from the model (see above). This fit is assessed by a chi-square test and the chi-square p-value becomes the distribution score, which provides a measure of how likely the candidate peak is a true TFBS (distribution score: 1) or the result of a sequencing artifact (distribution score: 0).

If a control sample is provided, an empirical FDR is calculated for each peak by repeating the peak-calling step (after fragment size estimation) with swapped IP and control sample and scoring the resulting control peaks by the raw and distribution score. This provides for each final peak score the number of true and control peaks that achieve this score or better and thus an FDR estimate.

Peakzilla reports the TFBSs in a BED-like format including the genomic positions, raw-, distribution-, and final score, FDR, and a peak number according to each peak's rank. In addition, the control peaks and a log are reported.

Program implementation. Peakzilla is implemented in Python 2 and runs on both the standard CPython and the fast PyPy interpreter. The program is freely available under the terms of the General Public License at http://github.com/steinmann/peakzilla. It runs from the command line under any linux distribution or OSX. The only required argument is the name of the file with the aligned reads from the ChIP sample and – optionally – from the control sample (both BED format). In addition, the following parameters can be used: -m to specify the number of candidate binding sites to use to estimate fragment size (default: 200); -l to limit the candidate regions to lengths above a certain minimum length (which may be necessary if the dataset contains a large number of strong PCR artifacts; default: off [1]); -f to set a FDR cutoff (default: off [100]); -c to set an enrichment cutoff (default: 2); -s to set a score cutoff (default: 1); -a to specify a folder in which additional files are saved to; -e to use an empirical estimate derived from the data for the model instead of a normal distribution (default: off); -p to specify that the data corresponds to fragments sequenced at both ends (paired end sequencing; default: off). For a human ChIP-seq dataset with 19.7 million reads and 7.8 in the control peakzilla runs in 4 minutes and consumes less than 800 MB of memory under CPython. Using the faster PyPy interpreter reduces time needed for analysis and memory requirements by half. Peakzilla can therefore be run efficiently on any modern desktop computer or Unix/Linux compute cluster. The method is illustrated in a flowchart in **supplementary figure 3**.

ChIP-seq datasets. Raw sequencing reads for Twist in *Drosophila melanogaster* (He et al. 2011) (ArrayExpress accession code E-MTAB-376), CEBPA in *Mus musculus* (Schmidt et al. 2010) (GEO accession code GSE22078) and PHA-4 in *Caenorhabditis elegans* (Zhong et al. 2010) (GEO accession code GSE14545) were aligned uniquely using bowtie allowing for 3 mismatches to the corresponding genomes (assemblies dm3, mm9 and ce6 respectively). For NFkB in *Homo sapiens* (Kasowski et al. 2010) (GEO accession code GSE19486), Ste12 in *Saccharomyces cerevisiae* (Zheng et al. 2010) (GEO accession code GSE19636) and CTCF ChIP-seq (Cuddapah et al. 2009) (GEO accession code GSE12889) and ChIP-exo (Rhee and Pugh 2011) (which the authors kindly shared) in *Homo sapiens*, already mapped reads were used (assemblies hg18, sarCer2 and hg18 respectively).

High-resolution ChIP-seq. Embryo aged 2-4 hours after egg laying (AEL) were processed and immunoprecipitated with Twist antibodies based on the protocol by He et al. (He et al. 2011) with slight modifications. Sonication occurs in three microfuges, each with ~80 mg chromatin extracts resuspended in 250 *µ*l A2 buffer, in a Biorupter sonicator for 15 min on high (30s on and off) at 4°C. After 15 min cooling, the sonication step is repeated, followed by high-speed centrifugation at 4°C for 10 min and pooling of the supernatant (the DNA fragments should be mostly between 200-500 bp). 600*µ*l supernatant are then incubated with 120*µ*l DNAse I (RNAse-free from NEB, to 0.3 U/*µ*l final concentration) and 80*µ*l DNAse I buffer for 30 min at 37°C. To stop DNAse I activity, A2 buffer with 10% SDS is added to a give a final concentration of 1% SDS. The extract is then directly used for ChIP (the DNA fragments should now be mostly between 50-200 bp). During Illumina library preparation, the samples are run on a 2% gel at 90V for ~2 hours, and fragments corresponding to ~50 bp, ~75, and ~100 bp inserts are cut out of the gel (slices are ~25 bp thick). The final libraries are run a BioAnalyzer to measure the actual average insert size.

Peak calling. The format of the mapped reads was adapted to each method. Peakzilla, SISSRs (Jothi et al. 2008) version 1.4, cisGenome (Ji et al. 2008) version 2.0, Spp (Kharchenko et al. 2008) version 1.8 and GPS (Guo et al. 2010) version 0.10.1 were run with default parameters. MACS (Zhang et al. 2008) version 1.4.1 was run with an mfold parameter 3,30 and the gsize parameter was adapted for each genome. QuEST (Valouev et al. 2008) version 2.4 was run with the following interactive choices: transcription factor binding sites with recommended (or relaxed) peak calling parameters. PeakRanger read extension length parameter was run using peakzilla's estimated fragment length. Both QuEST and PeakRanger could not be used for the CTCF samples without a control dataset.

Functional analyses. We used the known motif CACATGT for Twist and the motifs from JASPAR (Sandelin et al. 2004): snail (sna MA0086.1), dorsal (dl_1 MA0022.1), NFkB (NFKB1 MA0105.1), CEBPA (CEBPA MA0102.2), pha-4 (Foxa2 MA0047.1), Ste12 (STE12 MA0393.1). We searched for motif occurrences using MAST (Bailey and Gribskov 1998) (from the MEME suite programs version 4.1.1) with a p-value of 10^{-3} (10^{-2} for Twist allowing for one mismatch) in an area of 151 bp (average genomic fragment length) around each peak summit. We called a peak conserved when it overlapped with a peak region in all other *Drosophila* species from He/Bardet et al. (He et al. 2011) (*D.simulans*, *D.yakuba*, *D.erecta*, *D.ananassae* and *D.pseudoobscura*). We overlap peaks with known Twist enhancers from He/Bardet et al. (He et al. 2011) and the known mesodermal enhancers from Boon et al. (Bonn et al. 2012) (only M for mesoderm at stage 5,6 and 7). To create a set of control peaks, we shuffled peaks randomly within the same chromosomes.

DATA ACCESS

The high-resolution ChIP-seq data for Twist is deposited on GEO under the accession code [to come].

ACKNOWLEDGMENTS

We are grateful to Ho Sung Rhee and Franklin Pugh for kindly sharing the mapped read coordinates for ChIP-exo of human CTCF (Rhee & Pugh, Cell 2012). We would like to thank Jorge Omar Yanez Cuna and Gerald Stampfel (IMP) for discussions, help and advice. A.F.B. was supported by the Austrian Ministry for Science and Research through the Genome Research in Austria (GEN-AU) Bioinformatics Integration Network III. J.A.K. and J.S. are supported by the Austrian Research Fund (FWF) grant Z_153_B09. J.Z. is a Pew scholar and is supported by a NIH New Innovator grant (1DP2 OD004561-01). A.S. is supported by a European Research Council (ERC) Starting Grant from the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 242922. Basic research at the IMP is supported by Boehringer Ingelheim.

AUTHOR CONTRIBUTIONS

A.F.B., J.S., J.A.K., J.Z., and A.S. conceived the project. J.S. wrote peakzilla. A.F.B. benchmarked peakzilla and performed all computational analyses. A.F.B. and A.S. analyzed the data. J.Z. and S.B. designed and performed the high-resolution ChIP-seq experiments. A.F.B, J.Z. and A.S. wrote the manuscript.

DISCLOSURE DECLARATION

The authors declare no conflict of interest.

FIGURE LEGENDS

Figure 1 I Peakzilla algorithm. (A) Overview of the ChIP-seq pipeline. Transcription factor binding sites display a characteristic bimodal distribution of the positive and negative strand reads. (B) Example of a true positive (Peak A) and false positive (Peak B) peak in the Twist dataset in *D. melanogaster* (genomic coordinates chr2L:12420984-12423043 and chrX:9899747-9905926 respectively). Peak B unlike peak A does not exhibit the characteristic double distribution of reads on the positive and negative strands. (C) Read distribution model using two Gaussian distributions. (D) Peak score. While both peaks A and B from (B) show the same enrichment of read count over control, the score for peak B is penalized by the distribution score, a multiplicative factor [0…1], as it does not fit to the specific double distribution of the model in (C). (E) Number of peaks with a distribution score of 0 or 1 with the corresponding transcription factor motif. (F) Number of positions containing 90% of the reads for peaks with a distribution score of 0 or 1. See **supplementary figure 2** for an evaluation of peak scoring.

Figure 2 I High precision of peakzilla peaks. Analyses performed on the Twist dataset in *D. melanogaster*. (A) Enrichment of motifs in differential peaks between peakzilla and other methods. Bionomial p-values of enrichment over control and number of differential peaks with a motif is shown on top of the bars. See **supplementary figures 5** for other datasets and species. (B) Fold enrichment values of differential peaks and associated Wilcoxon p-values (NA: no peaks). (C) Distance of the summits of the top 500 peaks to the nearest motif. See **supplementary figures 6** for other datasets and species.

Figure 3 I High resolution of peakzilla peaks. Analyses performed on the Twist dataset in *D. melanogaster*. (A) Example of peak split. Peakzilla detects three adjacent peaks, while MACS, QuEST, CisGenome and PeakRanger report a single large peak region, and SISSRs and spp report two peak regions (GPS did not call any peak in that region; we considered all peaks called with standard parameters for each method) (B) Resolution achieved by the different methods as calculated as the minimal peak-to-peak distance (after removing 1% outliers for

each method). See **supplementary figure 7** for other datasets and species (C) Split peaks match motif occurrences. All peakzilla peaks corrsponding to a single MACS peak (major: same summit; minor: additional summit) are more highly enriched in Twist motifs than control regions, suggesting that they constitute true independent TFBSs. The same is true for motifs of Snail and Dorsal, which are transcription factors known to cooperate with Twist. (D) Split peaks are highly conserved. (E) Split peaks are enriched for known enhancers. (F) Split peaks are enriched for mesodermal enhancers.

Figure 4 I Application to high-resolution data. (A) Average read densities and peak regions from low- (red), medium- (purple) and high-resolution (blue) peaks for Twist (best 1000 peaks of each method). SISSRs, spp and GPS are not shown, as they do not report peak regions but only summit positions. (B) Resolution achieved by the different methods at low- (red), medium- (purple) and high-resolution (blue) as calculated as the minimal peak-to-peak distance (after removing 1% outliers for each method). (C) Average read densities and peak regions from ChIP-seq (red) and ChIP-exo (blue) peaks for CTCF (best 1000 peaks of each method; QuEST and PeakRanger cannot be used without a control sample). (D) Resolution of the methods calculated as in (B).

REFERENCES

Bailey TL, Gribskov M. 1998. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**: 48–54.

Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB. 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci USA* **99**: 757–762.

Boeva V, Surdez D, Guillon N, Tirode F, Fejes AP, Delattre O, Barillot E. 2010. De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. *Nucleic Acids Res* **38**: e126.

Bonn S, Zinzen RP, Perez-Gonzalez A, Riddell A, Gavin A-C, Furlong EEM. 2012. Cell type-specific chromatin immunoprecipitation from multicellular complex samples using BiTS-ChIP. *Nat Protoc* **7**: 978–994.

Bradley RK, Li X-Y, Trapnell C, Davidson S, Pachter L, Chu HC, Tonkin LA, Biggin MD, Eisen MB. 2010. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. *PLoS Biol* **8**: e1000343.

Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, et al. 2009. Unlocking the secrets of the genome. *Nature* **459**: 927–930.

Chen Y, Nègre N, Li Q, Mieczkowska JO, Slattery M, Liu T, Zhang Y, Kim T-K, He HH, Zieba J, et al. 2012. Systematic evaluation of factors influencing ChIP-seq fidelity. *Nat Methods*.

Cuddapah S, Jothi R, Schones DE, Roh T-Y, Cui K, Zhao K. 2009. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* **19**: 24–32.

ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**: 636–640.

Feng X, Grossman R, Stein L. 2011. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. *BMC Bioinformatics* **12**: 139.

Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. 2010. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res* **20**: 565–577.

Guo Y, Mahony S, Gifford DK. 2012. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol* **8**: e1002638.

Guo Y, Papachristoudis G, Altshuler RC, Gerber GK, Jaakkola TS, Gifford DK, Mahony S. 2010. Discovering homotypic binding events at high spatial resolution. *Bioinformatics*.

He Q, Bardet AF, Patton B, Purvis J, Johnston J, Paulson A, Gogol M, Stark A, Zeitlinger J. 2011. High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species. *Nat Genet* **43**: 414–420.

Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**: 533–538.

Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. 2008. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* **26**: 1293–1300.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**: 1497–1502.

Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. 2008. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* **36**: 5221–5231.

Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, et al. 2010. Variation in transcription factor binding among humans. *Science* **328**: 232–235.

Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* **26**: 1351–1359.

Kvon EZ, Stampfel G, Yáñez-Cuna JO, Dickson BJ, Stark A. 2012. HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. *Genes Dev* **26**: 908–913.

Li X-Y, MacArthur S, Bourgon R, Nix D, Pollard DA, Iyer VN, Hechmer A, Simirenko L, Stapleton M, Luengo Hendriks CL, et al. 2008. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol* **6**: e27.

Lifanov AP, Makeev VJ, Nazina AG, Papatsenko DA. 2003. Homotypic regulatory clusters in Drosophila. *Genome Res* **13**: 579–588.

modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Nègre N, Eaton ML, Landolin JM, Bristow CA, Ma L, et al. 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**: 1787–1797.

Mouse ENCODE Consortium, Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, et al. 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol* **13**: 418.

Pepke S, Wold B, Mortazavi A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* **6**: S22–32.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290**: 2306–2309.

Rhee HS, Pugh BF. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**: 1408–1419.

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**: 651–657.

Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**: D91–4.

Satija R, Bradley RK. 2012. The TAGteam motif facilitates binding of 21 sequence-specific transcription factors in the Drosophila embryo. *Genome Res* **22**: 656–665.

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**: 1036–1040.

Schroeder MD, Pearce M, Fak J, Fan H, Unnerstall U, Emberly E, Rajewsky N, Siggia ED, Gaul U. 2004. Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biol* **2**: E271.

Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* **5**: 829–834.

Wilbanks EG, Facciotti MT. 2010. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE* **5**: e11471.

Wu S, Wang J, Zhao W, Pounds S, Cheng C. 2010. ChIP-PaM: an algorithm to identify protein-DNA interaction using ChIP-Seq data. *Theor Biol Med Model* **7**: 18.

Yáñez-Cuna JO, Dinh HQ, Kvon EZ, Shlyueva D, Stark A. 2012. Uncovering cis-regulatory sequence requirements for context specific transcription factor binding. *Genome Res*.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Zheng W, Zhao H, Mancera E, Steinmetz LM, Snyder M. 2010. Genetic analysis of variation in transcription factor binding in yeast. *Nature* **464**: 1187–1191.

Zhong M, Niu W, Lu ZJ, Sarov M, Murray JI, Janette J, Raha D, Sheaffer KL, Lam HYK, Preston E, et al. 2010. Genome-wide identification of binding sites defines distinct functions for Caenorhabditis elegans PHA-4/FOXA in development and environmental response. *PLoS Genet* **6**: e1000848.

Figure 1 - Bardet/Steinmann et al.

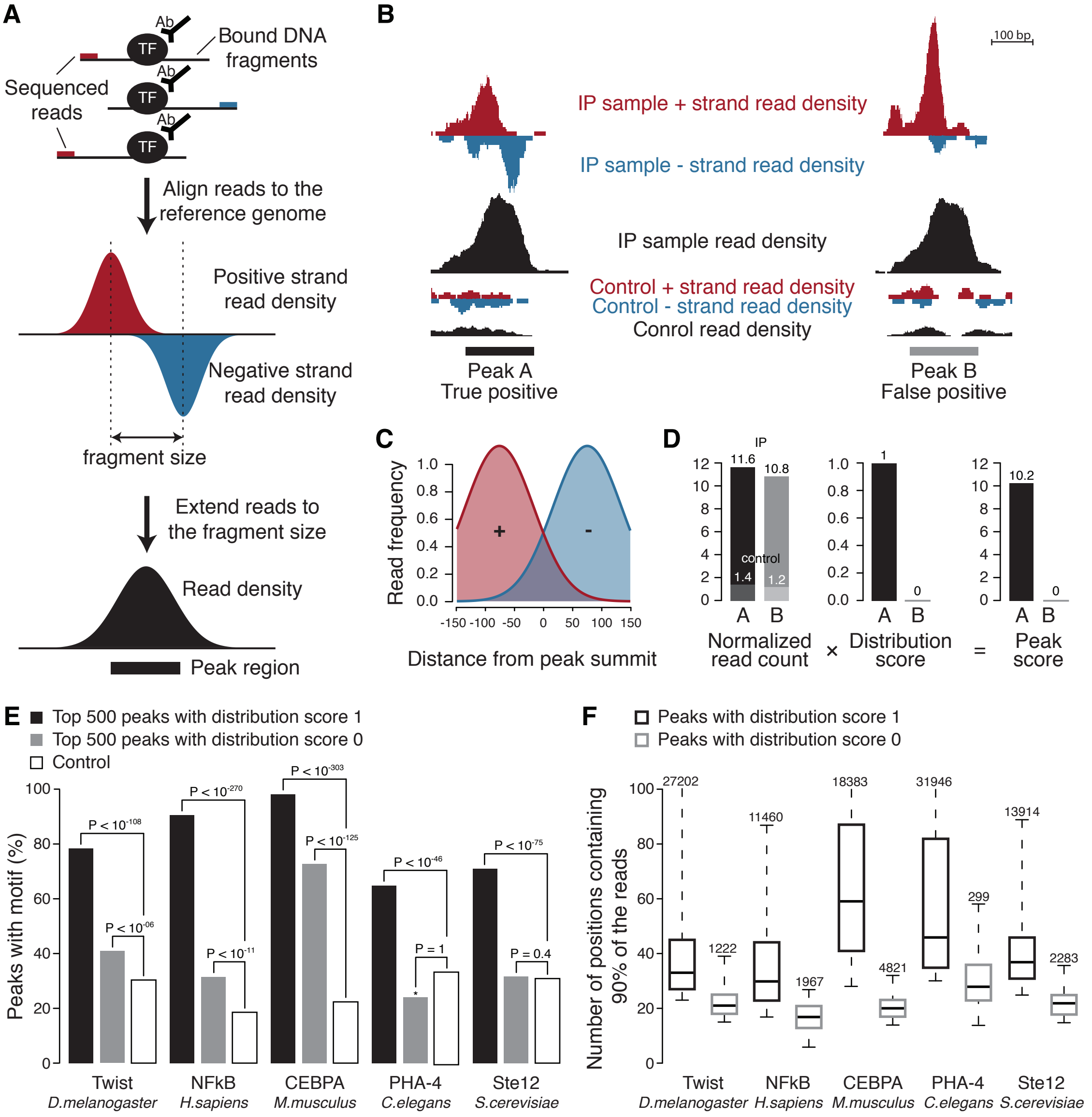


Figure 2 - Bardet/Steinmann et al.

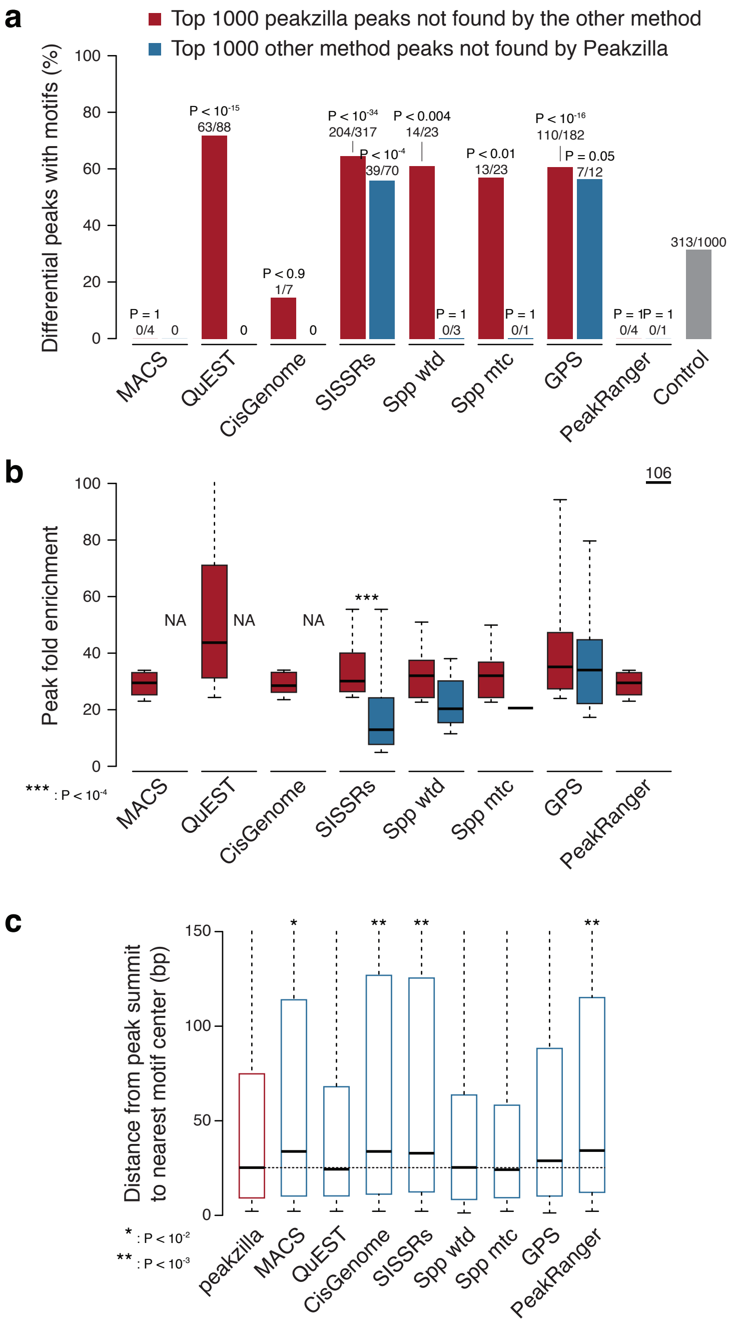


Figure 3 - Bardet/Steinmann et al.

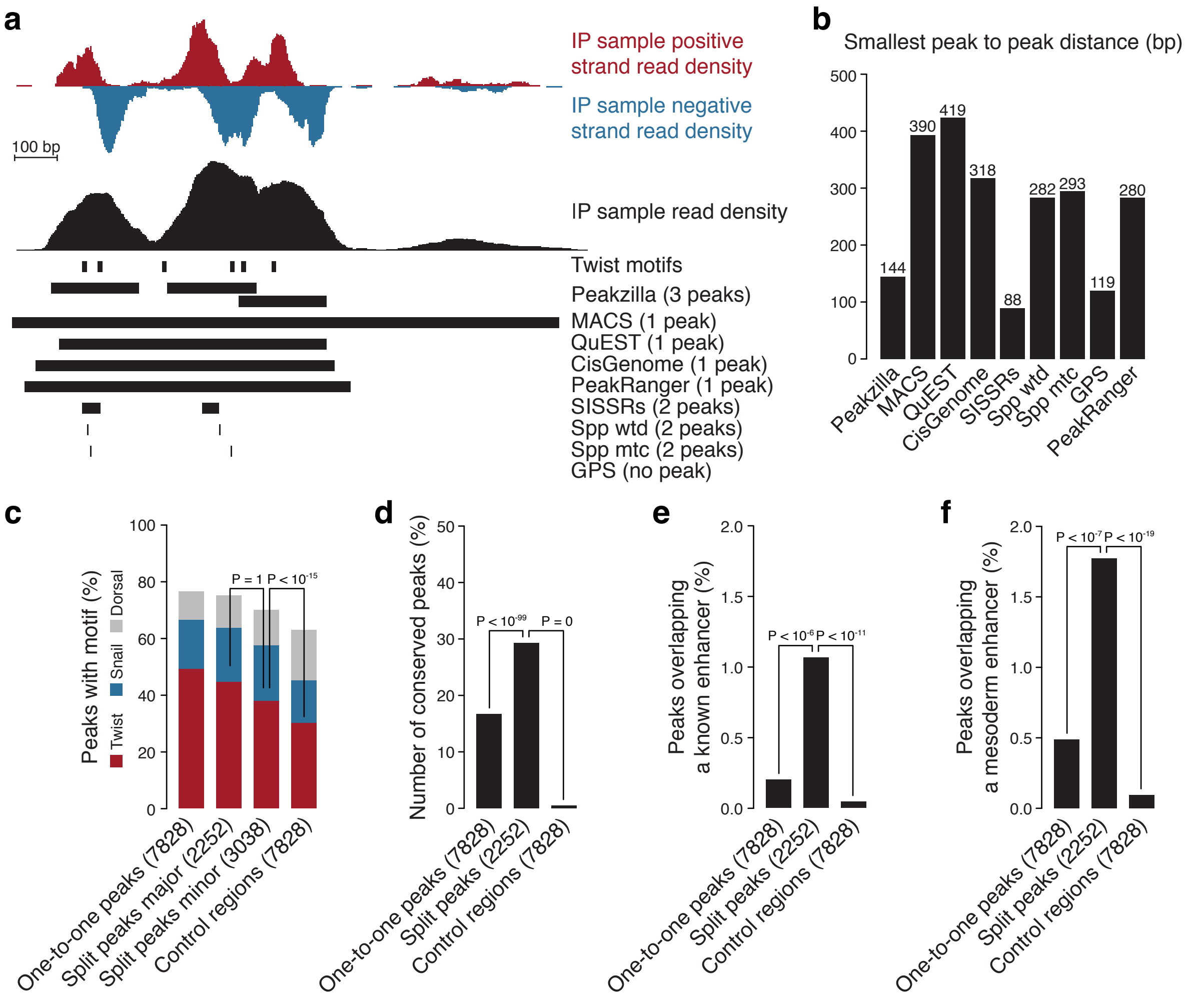
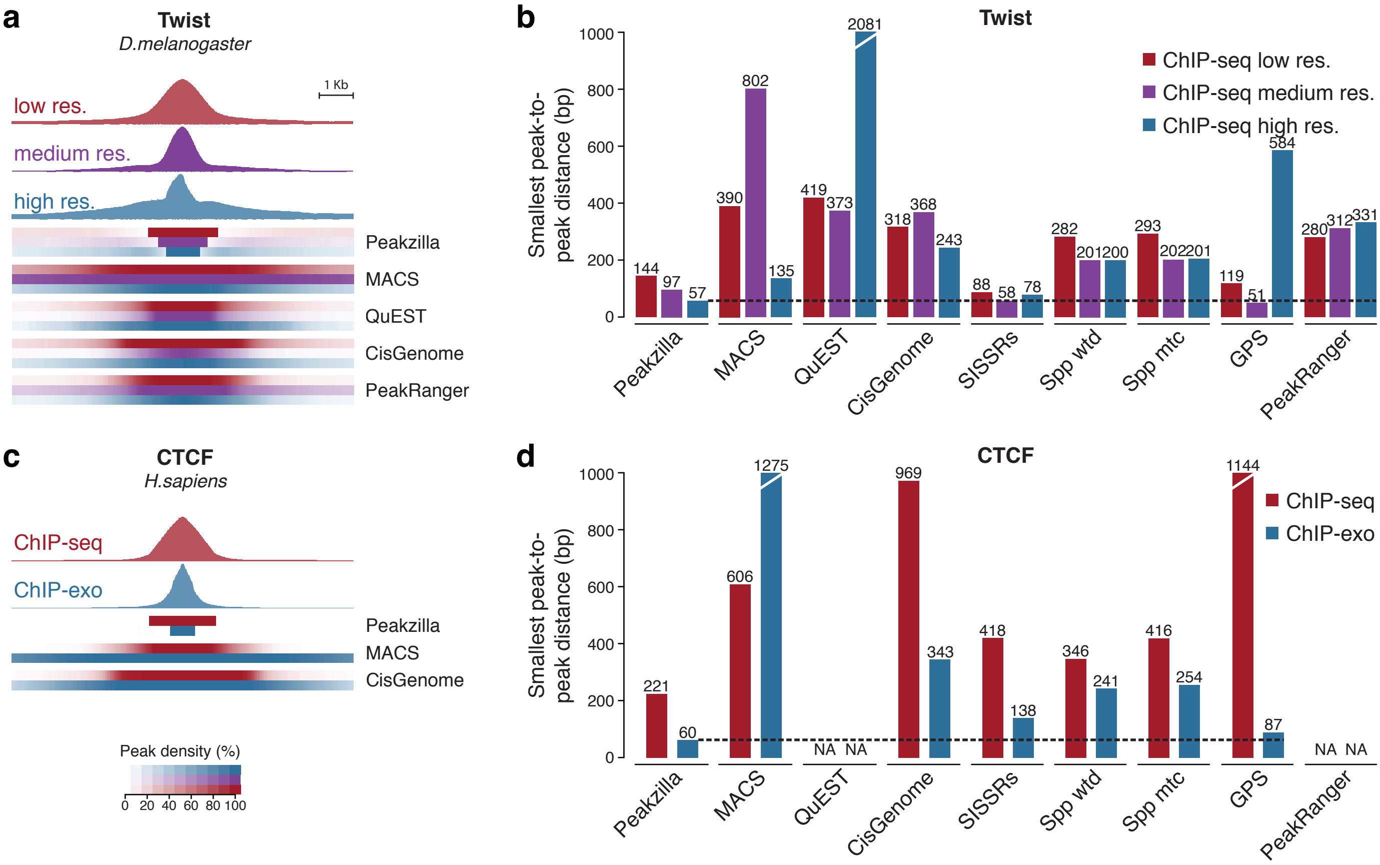


Figure 4 - Bardet/Steinmann et al.



Supplementary information for

Identification of transcription factor binding sites from ChIP-seq data at high-resolution

Anaïs F. Bardet[*1], Jonas Steinmann[*2], Sangeeta Bafna[3,4], Juergen A. Knoblich[2], Julia Zeitlinger[3], Alexander Stark[1]#

[1]Research Institute of Molecular Pathology (IMP), Vienna, Austria

[2]Institute of Molecular Biotechnology (IMBA), Vienna, Austria

[3]Stowers Institute for Medical Research, Kansas City, Missouri, USA

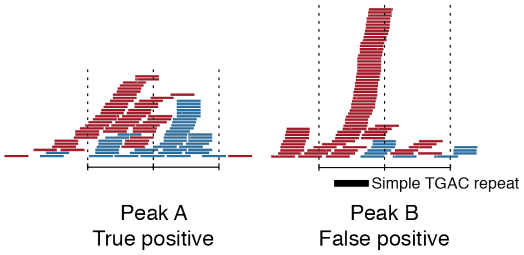
[4]Current address: Department of Medicine, Vanderbilt University, Nashville, TN, USA

[*]These authors contributed equally

[#]Corresponding author (stark@starklab.org)

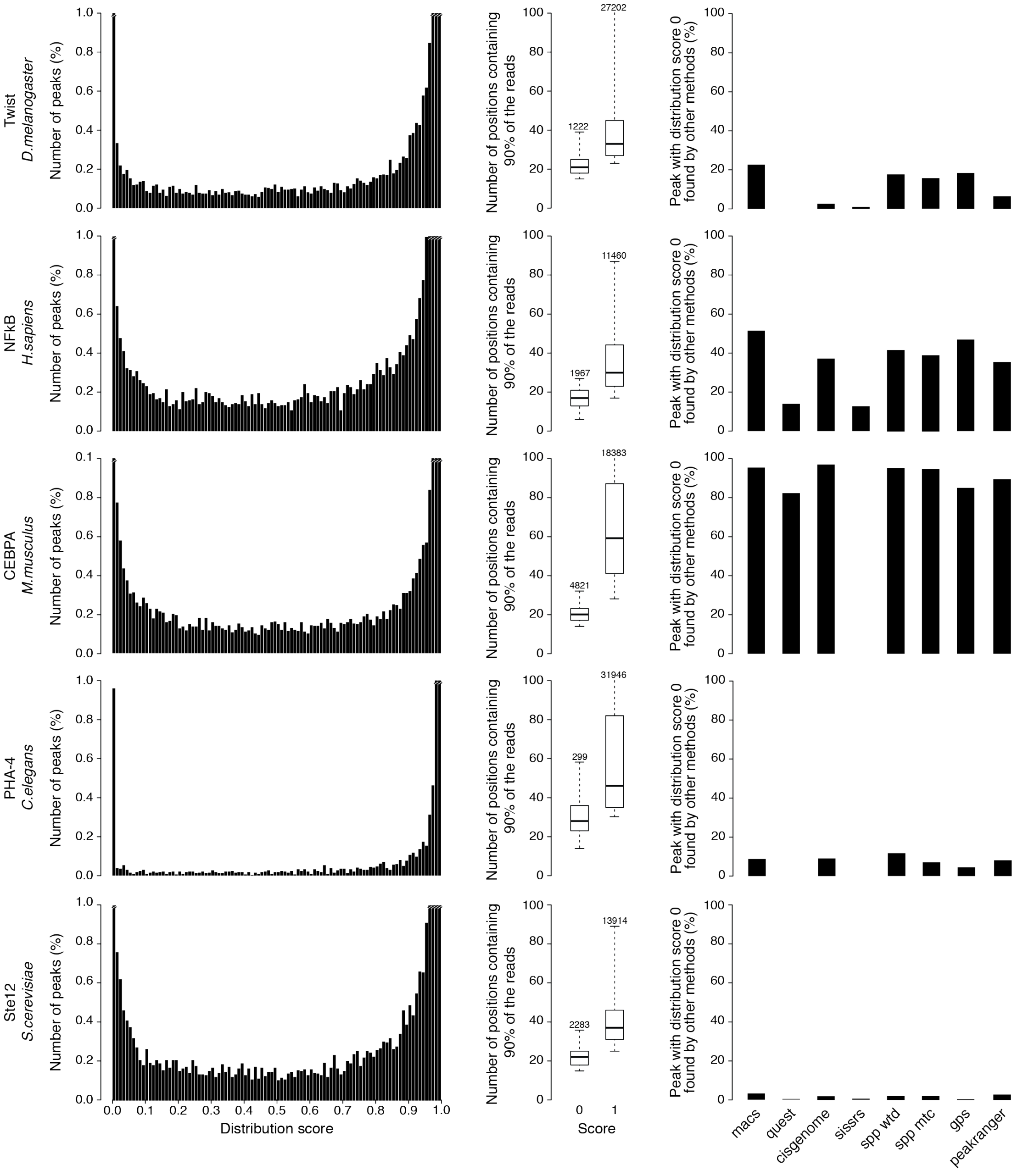
Lab homepage: http://www.starklab.org

Figure S1: Example of false positive peak



Compared to peak A, peak B contains an unbalanced proportion of reads mapped to the positive (red) and negative (blue) strand due to the presence of a simple TGAC repeat that biases read mapping – even when only reads are considered that have a single best mapping position in the genome as done here. This suggests that this peak candidate is indeed a false positive.

Figure S2: Evaluation of peak scoring



Histogram of peak distribution scores for different datasets (left). Peaks with distribution scores (multiplicative factors [0…1]) below 1 are penalized in their final score, to reflect that their read distributions do not fit to the specific double distribution of the model. Penalized peaks with a distribution score of 0 contain substantially less diverse sequence reads (fewer genomic positions containing 90% of read counts) than non-penalized peaks with a distribution score of 1 and are thus more likely to be false positives as described in supplementary figure 1 (middle). Number of estimated false positive peaks by peakzilla found by other methods (right).

Figure S3: Flowchart of the method

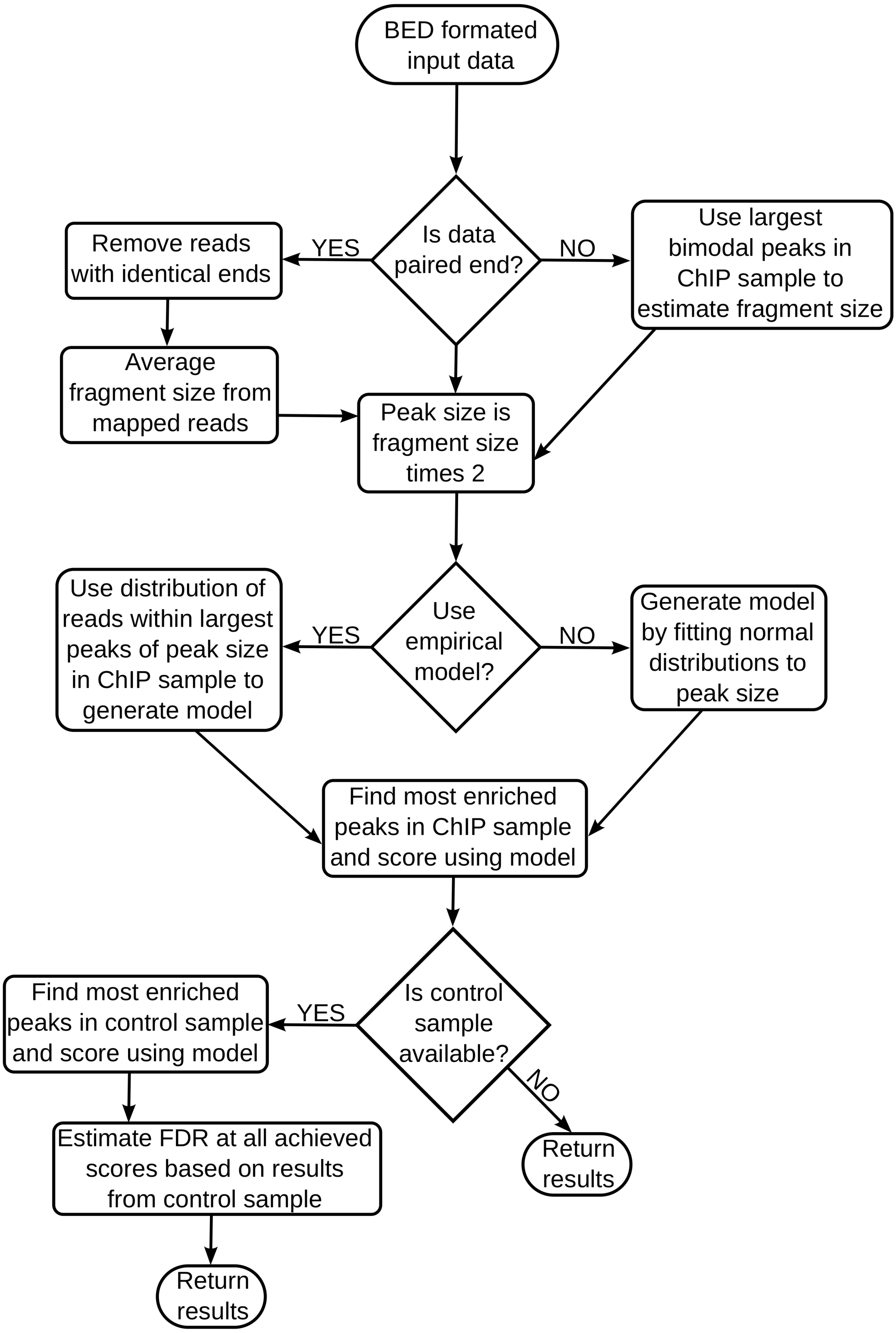
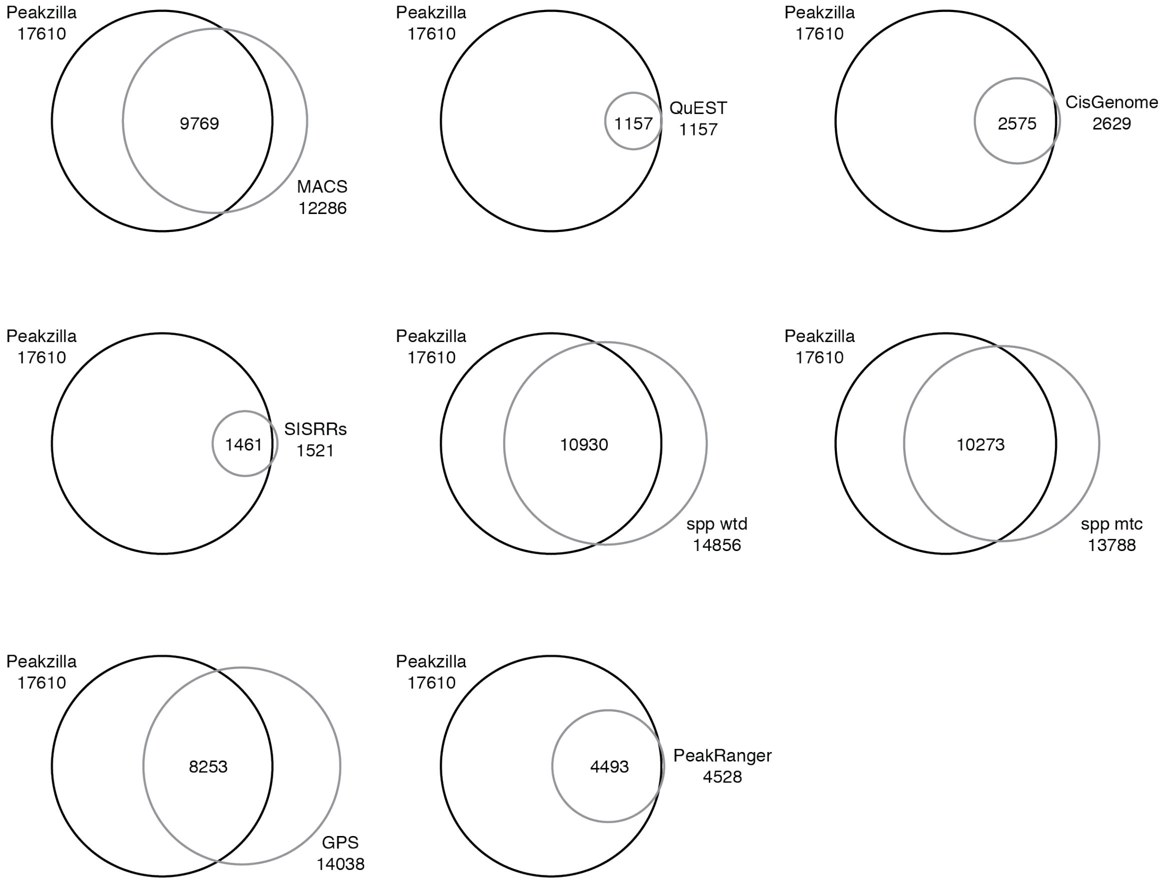
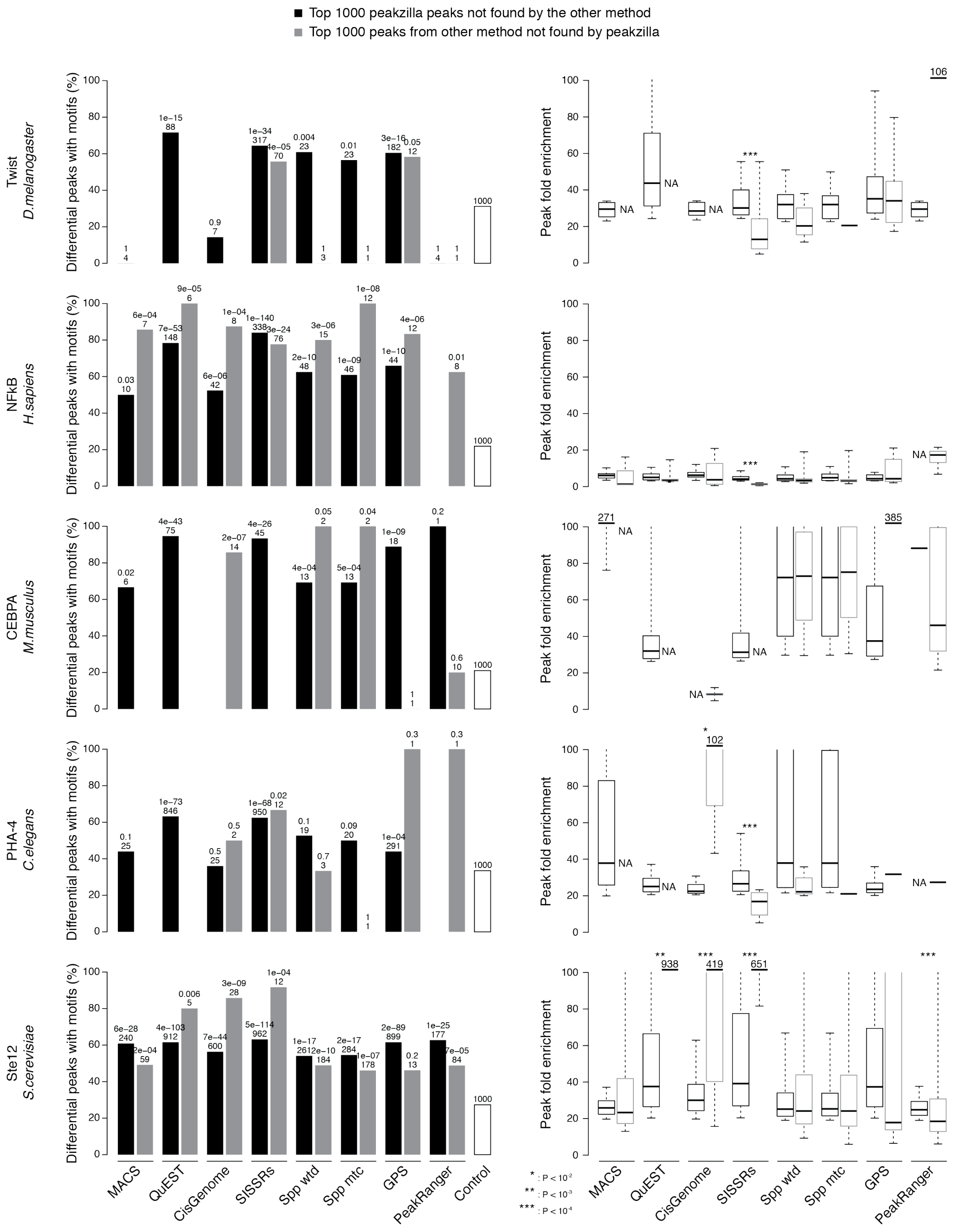

BED formated input data
Is data paired end?
YES
NO
Remove reads with identical ends
Average fragment size from mapped reads
Use largest bimodal peaks in ChIP sample to estimate fragment size
Peak size is fragment size times 2
Use empirical model?
YES
NO
Use distribution of reads within largest peaks of peak size in ChIP sample to generate model
Generate model by fitting normal distributions to peak size
Find most enriched peaks in ChIP sample and score using model
Is control sample available?
YES
NO
Find most enriched peaks in control sample and score using model
Estimate FDR at all achieved scores based on results from control sample
Return results
Return results

Figure S4: Overlap of Peakzilla peaks with peaks from different methods



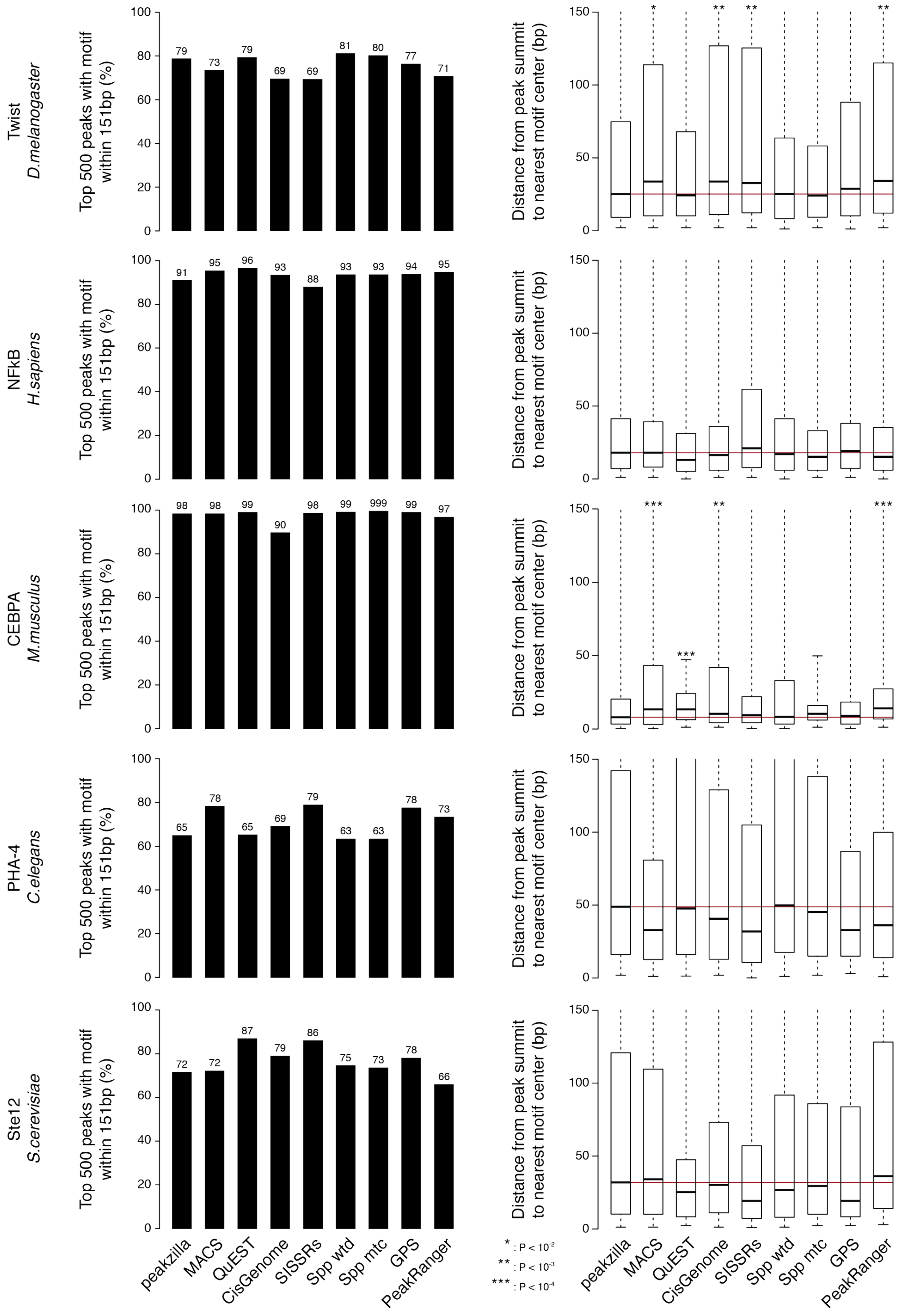
Peakzilla peaks in the Twist dataset in *D. melanogaster* are in good agreement with peaks from other methods independent of the number of peaks found by each method.

Figure S5: Evaluation of differential peaks between Peakzilla and other methods



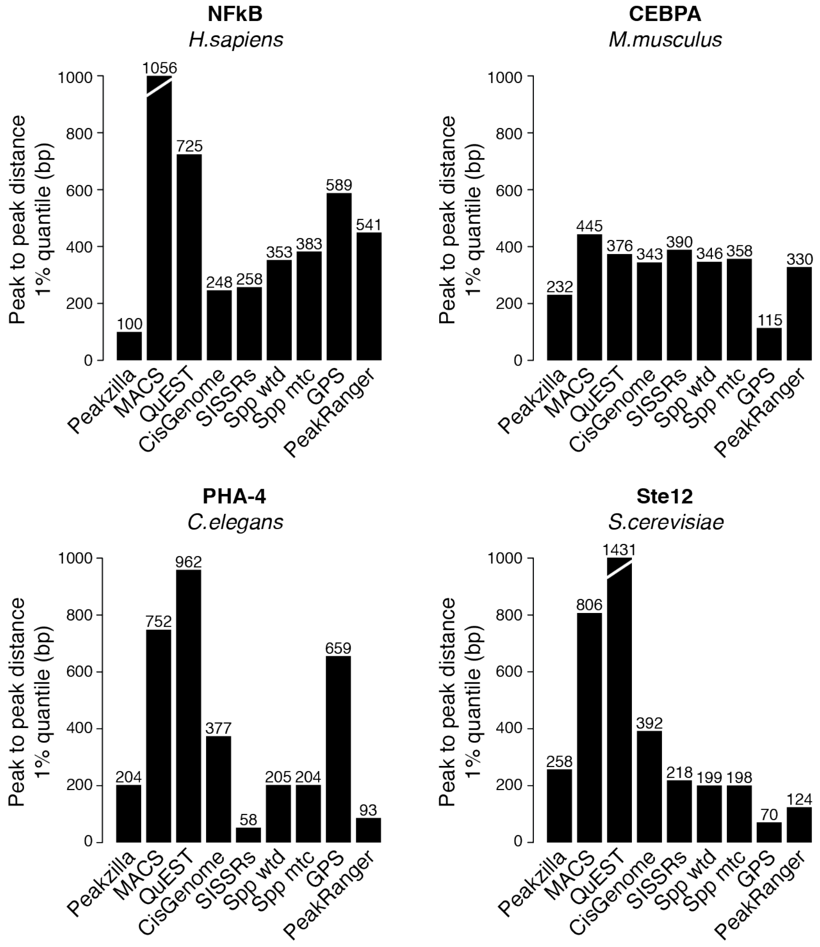
Enrichment of motifs in differential peaks (left). Binomial p-values of enrichment over control and number of differential peaks that contain a motif is shown on top of the bars. Fold enrichment values of differential peaks and associated Wilcoxon p-values (NA: no peak) (right; the Twist data [top row] are repeated from main Fig. 2 A & B to allow direct comparisons).

Figure S6: Precision of peaks



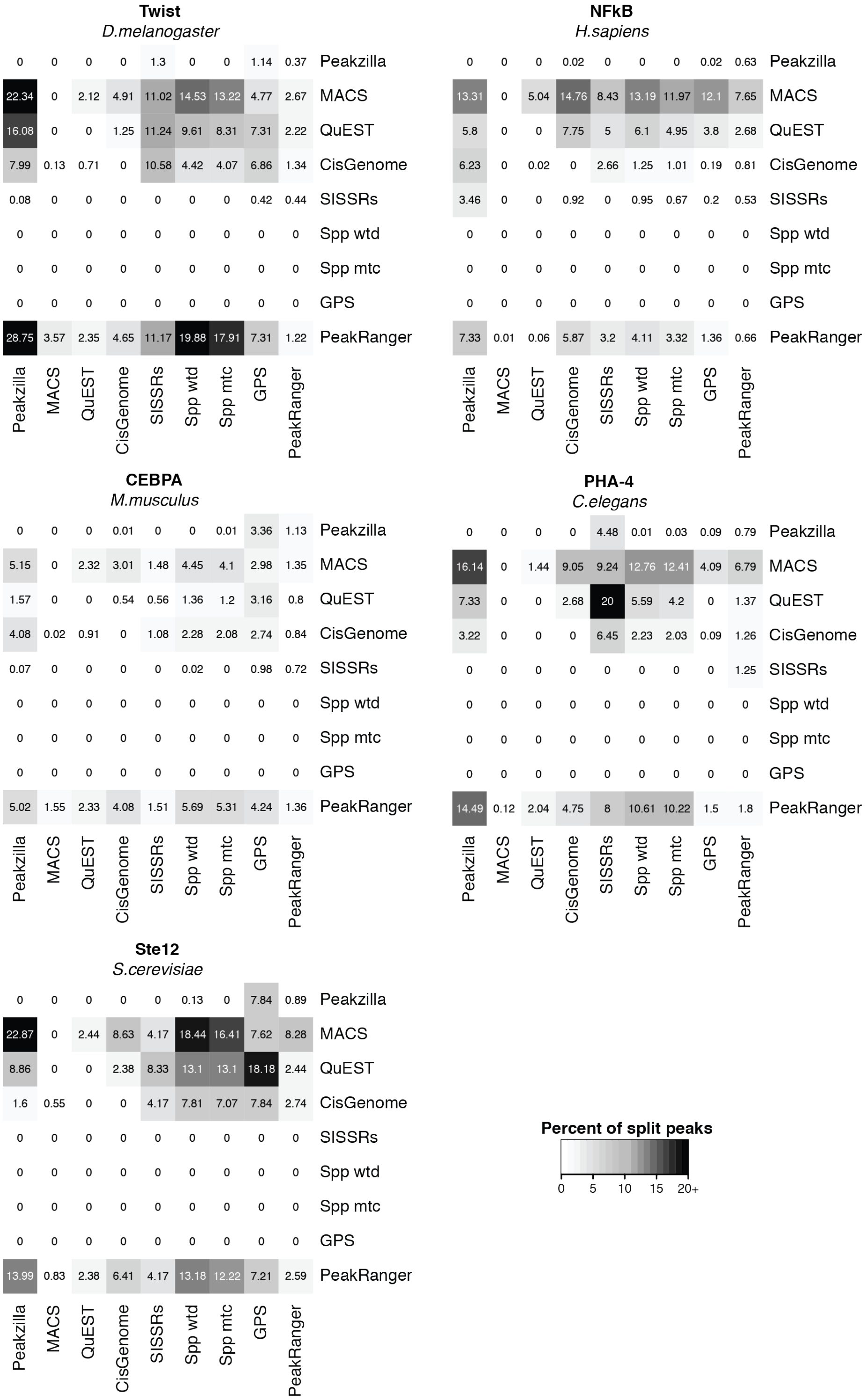
The precision of locating peak summits (= TFBS) is estimated for the top 500 peaks as the number of peaks with the corresponding TF motif within 151 bp around the summit (barplots) and the distance of the summits to the nearest motif (boxplots). See **figures 2C.**

Figure S7: Peak to peak distance



Peaks' resolution achieved by the different methods as calculated as the minimal peak-to-peak distance (after removing 1% outliers for each method). See figure 3B.

Figure S8: Number of peak split using different methods



Twist
D.melanogaster

	Peakzilla	MACS	QuEST	CisGenome	SISSRs	Spp wtd	Spp mtc	GPS	PeakRanger
Peakzilla	0	0	0	0	1.3	0	0	1.14	0.37
MACS	22.34	0	2.12	4.91	11.02	14.53	13.22	4.77	2.67
QuEST	16.08	0	0	1.25	11.24	9.61	8.31	7.31	2.22
CisGenome	7.99	0.13	0.71	0	10.58	4.42	4.07	6.86	1.34
SISSRs	0.08	0	0	0	0	0	0	0.42	0.44
Spp wtd	0	0	0	0	0	0	0	0	0
Spp mtc	0	0	0	0	0	0	0	0	0
GPS	0	0	0	0	0	0	0	0	0
PeakRanger	28.75	3.57	2.35	4.65	11.17	19.88	17.91	7.31	1.22

NFkB
H.sapiens

	Peakzilla	MACS	QuEST	CisGenome	SISSRs	Spp wtd	Spp mtc	GPS	PeakRanger
Peakzilla	0	0	0	0.02	0	0	0	0.02	0.63
MACS	13.31	0	5.04	14.76	8.43	13.19	11.97	12.1	7.65
QuEST	5.8	0	0	7.75	5	6.1	4.95	3.8	2.68
CisGenome	6.23	0	0.02	0	2.66	1.25	1.01	0.19	0.81
SISSRs	3.46	0	0	0.92	0	0.95	0.67	0.2	0.53
Spp wtd	0	0	0	0	0	0	0	0	0
Spp mtc	0	0	0	0	0	0	0	0	0
GPS	0	0	0	0	0	0	0	0	0
PeakRanger	7.33	0.01	0.06	5.87	3.2	4.11	3.32	1.36	0.66

CEBPA
M.musculus

	Peakzilla	MACS	QuEST	CisGenome	SISSRs	Spp wtd	Spp mtc	GPS	PeakRanger
Peakzilla	0	0	0	0.01	0	0	0.01	3.36	1.13
MACS	5.15	0	2.32	3.01	1.48	4.45	4.1	2.98	1.35
QuEST	1.57	0	0	0.54	0.56	1.36	1.2	3.16	0.8
CisGenome	4.08	0.02	0.91	0	1.08	2.28	2.08	2.74	0.84
SISSRs	0.07	0	0	0	0	0.02	0	0.98	0.72
Spp wtd	0	0	0	0	0	0	0	0	0
Spp mtc	0	0	0	0	0	0	0	0	0
GPS	0	0	0	0	0	0	0	0	0
PeakRanger	5.02	1.55	2.33	4.08	1.51	5.69	5.31	4.24	1.36

PHA-4
C.elegans

	Peakzilla	MACS	QuEST	CisGenome	SISSRs	Spp wtd	Spp mtc	GPS	PeakRanger
Peakzilla	0	0	0	0	4.48	0.01	0.03	0.09	0.79
MACS	16.14	0	1.44	9.05	9.24	12.76	12.41	4.09	6.79
QuEST	7.33	0	0	2.68	20	5.59	4.2	0	1.37
CisGenome	3.22	0	0	0	6.45	2.23	2.03	0.09	1.26
SISSRs	0	0	0	0	0	0	0	0	1.25
Spp wtd	0	0	0	0	0	0	0	0	0
Spp mtc	0	0	0	0	0	0	0	0	0
GPS	0	0	0	0	0	0	0	0	0
PeakRanger	14.49	0.12	2.04	4.75	8	10.61	10.22	1.5	1.8

Ste12
S.cerevisiae

	Peakzilla	MACS	QuEST	CisGenome	SISSRs	Spp wtd	Spp mtc	GPS	PeakRanger
Peakzilla	0	0	0	0	0	0.13	0	7.84	0.89
MACS	22.87	0	2.44	8.63	4.17	18.44	16.41	7.62	8.28
QuEST	8.86	0	0	2.38	8.33	13.1	13.1	18.18	2.44
CisGenome	1.6	0.55	0	0	4.17	7.81	7.07	7.84	2.74
SISSRs	0	0	0	0	0	0	0	0	0
Spp wtd	0	0	0	0	0	0	0	0	0
Spp mtc	0	0	0	0	0	0	0	0	0
GPS	0	0	0	0	0	0	0	0	0
PeakRanger	13.99	0.83	2.38	6.41	4.17	13.18	12.22	7.21	2.59

Percent of peaks from one method (rows) overlapping the peaks from another method (columns). The fraction of peaks that are split is indicated by the shading.

Curiculum Vitae

Name	Anaïs Flore Bardet
Date of Birth	October 14th, 1984
Birthplace	Céret, France
Nationality	French

Academic Education

2009-2012	PhD student Research Institute of Molecular Pathology (Vienna, Austria)
2007-2009	Research assistant Chair of Bioinformatics (Boku University, Vienna, Austria)
2005-2007	M.Sc. in Bioinformatics, Genome and Transcriptome Paris 7 Denis Diderot University (France)
2004-2005	B.Sc. Honors in Biochemical Sciences Salford University (United Kingdom)
2002-2004	Bioinformatics D.U.T. (University Diploma of Technology) Clermont-Ferrand 1 University (France)
2002	High School Graduation (Scientific Baccalaureate) Lycée Champollion, Figeac (France)