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]

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

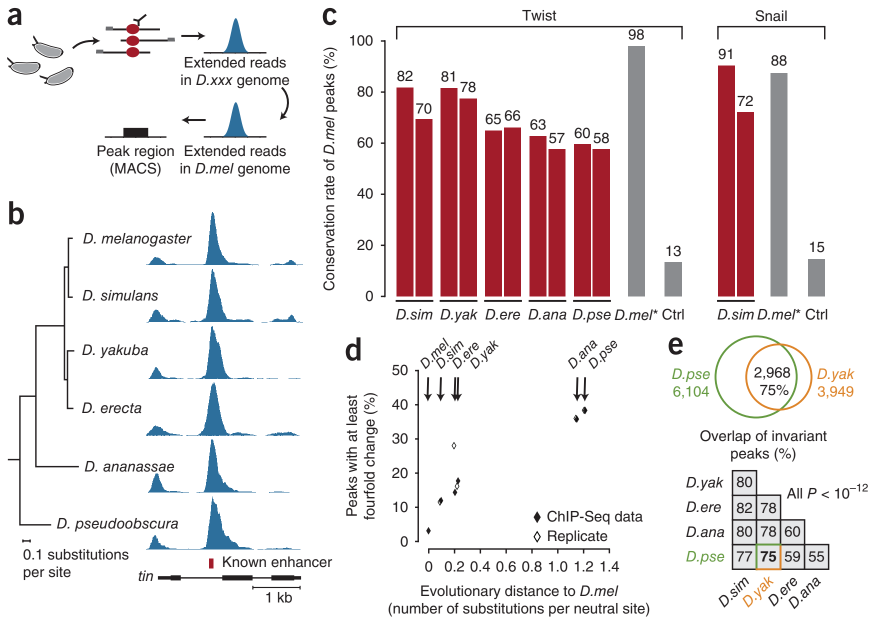

Figure 1 Evolutionary constraints on Twist binding across six *Drosophila* species. (**a**) Overview of the comparative ChIP-Seq pipeline. We directly translated the genomic coordinates of matched reads to *D. melanogaster* for peak calling and analysis (see **Supplementary Tables 3–6** for alternatives). (**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.

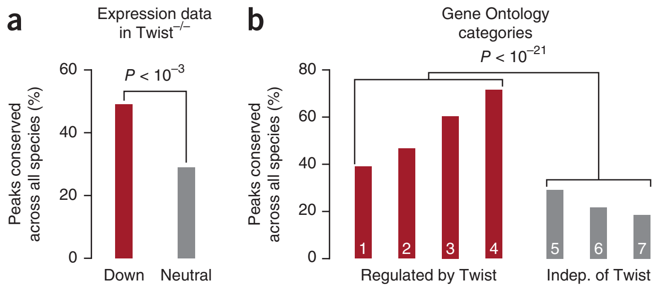

Figure 2 High conservation of functional Twist binding across six *Drosophila* species. (**a**) Preferential conservation of peaks near genes that are downregulated at least twofold in *twist* mutant embryos (red) compared to control genes that do not change (gray; data from a previous study[10]); the 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**.

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

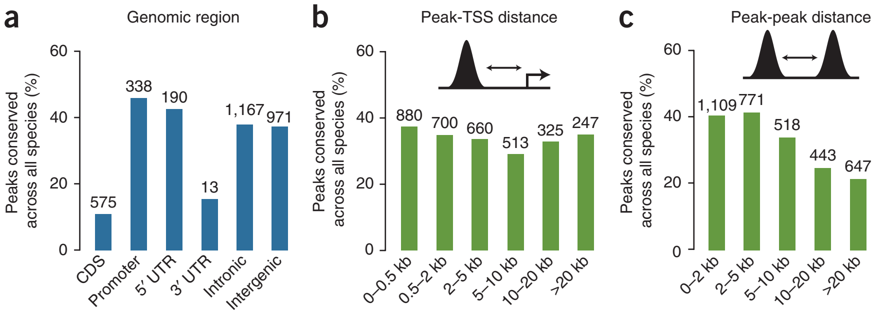

Figure 3 Preferential conservation of clustered binding peaks. (**a**) Conservation rates (percent of *D. melanogaster* peaks that are conserved across all six species) for peaks in different genomic regions. CDS, coding-sequence; UTR, untranslated region. The number of *D. melanogaster* peaks in each 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.

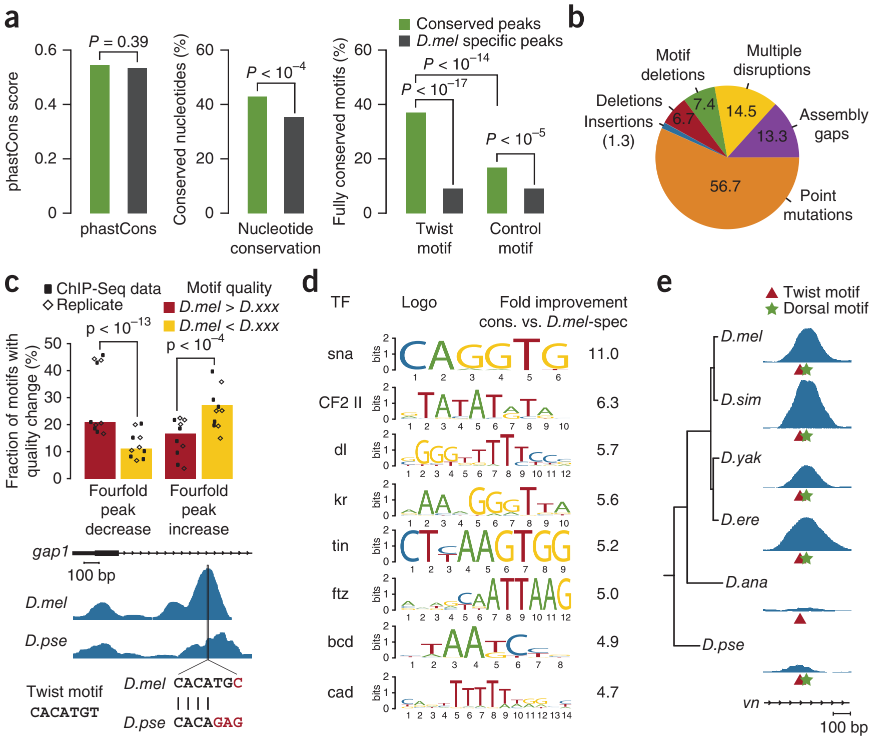

Figure 4 Twist binding depends on the sequence motifs of Twist and its partner transcription factors. (**a**) Twist binding peaks shared across all species (conserved) or *D. melanogaster*–specific (*D.mel*-spec) peaks have similar overall phastCons scores (left; Wilcoxon $P = 0.39$) and nucleotide conservation (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

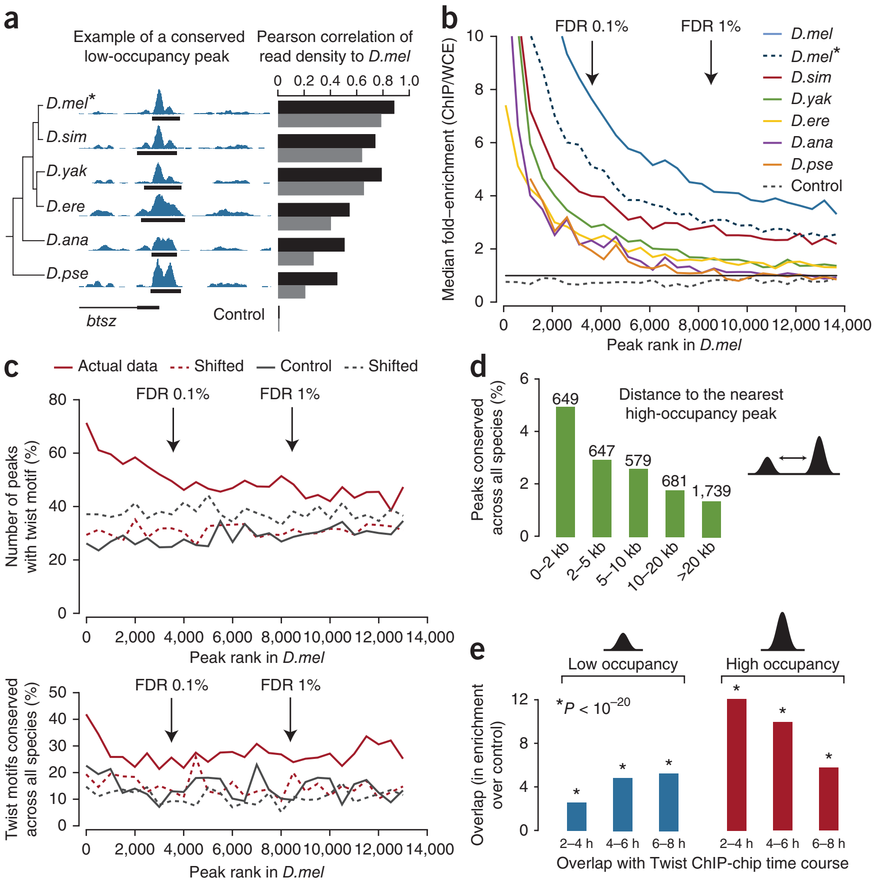

Figure 5 Conservation of low-occupancy peaks. (**a**) The similarity of Twist binding (blue) extends beyond the peak (black bar) at the *btsz* (*bitesize*) locus (left). Read densities are similar across species (black) even when excluding peak regions (gray). *D. mel**, biological replicate; control, 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 $P < 10^{-20}$).

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).

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).

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 immunostainings. 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