nature biotechnology

ChIP-nexus enables improved detection of *in vivo* transcription factor binding footprints

Qiye He[1,3,4], Jeff Johnston[1,4] & Julia Zeitlinger[1,2]

Understanding how eukaryotic enhancers are bound and regulated by specific combinations of transcription factors is still a major challenge. To better map transcription factor binding genome-wide at nucleotide resolution *in vivo*, we have developed a robust ChIP-exo protocol called ChIP-nexus (chromatin immunoprecipitation experiments with nucleotide resolution through exonuclease, unique barcode and single ligation), which utilizes an efficient DNA self-circularization step during library preparation. Application of ChIP-nexus to four proteins—human TBP and *Drosophila* NFkB, Twist and Max—shows that it outperforms existing ChIP protocols in resolution and specificity, pinpoints relevant binding sites within enhancers containing multiple binding motifs, and allows for the analysis of *in vivo* binding specificities. Notably, we show that Max frequently interacts with DNA sequences next to its motif, and that this binding pattern correlates with local DNA-sequence features such as DNA shape. ChIP-nexus will be broadly applicable to the study of *in vivo* transcription factor binding specificity and its relationship to *cis*-regulatory changes in humans and model organisms.

The ability to precisely map transcription factor binding footprints *in vivo* at single-nucleotide resolution is essential for an understanding of the mechanisms of combinatorial control by transcription factors[1]. Occupancy by specific transcription factors can be mapped by ChIP coupled to deep sequencing (ChIP-seq), but the resolution of this technique is limited by the minimal DNA-fragment size required for unique alignment to the genome[2]. In a clever improvement to ChIP-seq called ChIP-exo, the immunoprecipitated chromatin fragments are treated with λ-exonuclease, which digests one strand of the double-stranded DNA in a 5′-to-3′ direction and stops when it encounters a cross-linked protein[3,4]. In this manner the exact bases bordering a DNA-bound protein (the 'stop bases') can be mapped at essentially nucleotide resolution, enabling new biological insights[3,5,6]. However, we found significant technical hurdles in applying ChIP-exo. The additional wash and digestion steps reduce the amount of DNA that can be recovered compared to conventional ChIP-seq experiments, which is critical for the quality of a ChIP library. For amplification during library preparation, DNA fragments must go through two inefficient ligation steps to acquire adaptors on both ends. Low amounts of starting DNA often lead to overamplification artifacts during PCR, which result in noisy data that are not reproducible[7,8]. Another hurdle is that the original ChIP-exo protocol is designed for the SOLiD platform, although Illumina versions have recently become available[9,10].

Here we describe a more robust and reproducible, Illumina-based ChIP-exo protocol. As λ-exonuclease digestion of ChIP DNA mostly yields single-stranded DNA and requires the retention of strand information, we combined the standard ChIP-exo protocol with the library-preparation protocol from the iCLIP method for mapping RNA-protein interactions[11] to improve the efficiency with which DNA fragments are incorporated into the library. In addition, we added a unique, randomized barcode to the adaptor that enables monitoring of overamplification[7,8]. This combined protocol, called ChIP-nexus, is more efficient because it requires only one successful ligation per DNA fragment. Although ChIP-nexus adaptors were designed to be ligated to both DNA ends as in conventional ChIP-seq and ChIP-exo protocols, a library product is still generated if the adaptor is ligated to only one end. This is because λ-exonuclease digests the 5′ end of each strand regardless of whether an adaptor is present, and thus a single ChIP-nexus adaptor on the 3′ end is sufficient. The fragment is then circularized, which brings Illumina library primers to the digested end. Because intramolecular circularization is far more efficient than intermolecular ligation, library generation is more efficient than in a classical library preparation protocol where two independent ligations are required to generate a library product. As a result, ChIP-nexus produces high-quality libraries without requiring more starting material than conventional ChIP-seq experiments. The protocol is outlined in **Figure 1a** and in the Online Methods. A detailed protocol is available as **Supplementary Protocol 1** or from http://research.stowers.org/zeitlingerlab.

We compared the results from the ChIP-nexus protocol to published results on human TBP obtained with the original ChIP-exo protocol adapted for the Illumina sequencing platform[9]. Our ChIP-nexus experiments were performed using the same number of K562 cells and the same TBP antibody as in the previous study, and the locations of the stop bases on each strand were plotted. As exemplified by the RPS12 locus[9], ChIP-nexus produced visibly better results (**Fig. 1b**).

[1]Stowers Institute for Medical Research, Kansas City, Missouri, USA. [2]Department of Pathology, Kansas University Medical Center, Kansas City, Kansas, USA. [3]Present address: Institute of Neurosciences, Chinese Academy of Sciences, Shanghai, P.R. China. [4]These authors contributed equally to this work. Correspondence should be addressed to J.Z. (jbz@stowers.org).

Received 11 March 2014; accepted 8 December 2014; published online 9 March 2015; doi:10.1038/nbt.3121

When the previously published ChIP-exo data were plotted in the same way, they showed signs of overamplification: the reads often occurred in extremely high numbers at a certain position, without reads detected at neighboring positions. In contrast, ChIP-nexus produced a signal across the entire promoter region in a pattern that can be observed with regular ChIP-exo data only after averaging across many genes. Thus, although the overall readout is comparable to that obtained with the original ChIP-exo protocol,

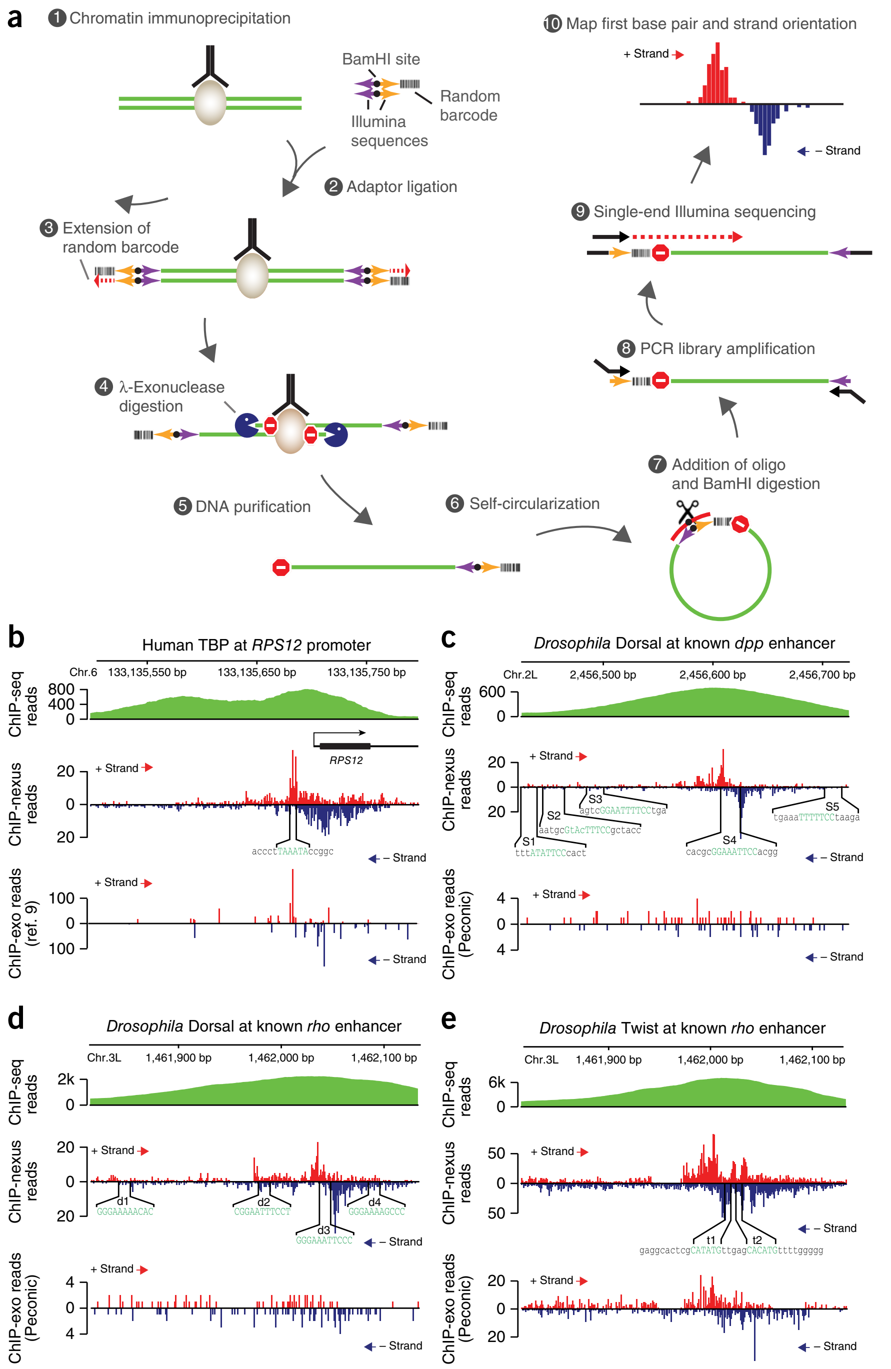

Figure 1 Superior performance of ChIP-nexus in discovering relevant binding footprints for transcription factors. (**a**) Outline of ChIP-nexus. (1) The transcription factor of interest is immunoprecipitated from chromatin fragments with antibodies in the same way as during conventional ChIP-seq experiments. (2) While the DNA is still bound to the antibodies, the DNA ends are repaired, dA-tailed and ligated to a special adaptor that contains a pair of sequences for library amplification (purple and orange arrows indicate the correct orientation required for the DNA to be functional), a BamHI site (black dots) for linearization and a nine-nucleotide barcode containing five random bases and four fixed bases to remove reads resulting from overamplification of library DNA. The barcode is part of a 5′ overhang that reduces adaptor-adaptor ligation. (3) After the adaptor-ligation step, the 5′ overhang is filled, copying the random barcode and generating blunt ends for λ-exonuclease digestion. (4) λ-Exonuclease (blue 'Pac-Man' symbols) digests until it encounters a physical barrier such as a cross-linked protein-DNA complex ('do not enter' signs represent 'stop bases'). (5) Single-stranded DNA is eluted and purified. (6) Self-circularization places the barcode next to the stop base. (7) An oligonucleotide (red arc) is paired with the region around the BamHI site for BamHI digestion (black scissor). (8) The digestion results in relinearized DNA fragments with suitable Illumina sequences on both ends that are ready for PCR library amplification. (9) Using single-end sequencing with the standard Illumina primer, each fragment is sequenced: first the barcode, then the genomic sequence starting with the stop base. (10) After alignment of the genomic sequences, reads with identical start positions and identical barcodes are removed. The final output is the position, number and strand orientation of the stop bases. The frequencies of stop bases on the positive strand are shown in red, and those on the negative strand are shown in blue. (**b**–**e**) Comparison of conventional ChIP-seq data (extended reads), ChIP-nexus data (raw stop base reads) and data generated using the original ChIP-exo protocol (raw stop base reads). (**b**) TBP profiles in human K562 cells at the *RPS12* promoter. Although ChIP-nexus and ChIP-exo generally agree on TBP binding footprints, ChIP-nexus provides better coverage and richer details than ChIP-exo, which shows signs of overamplification as large numbers of reads accumulate at a few discrete bases. (**c**) Dorsal profiles at the *Drosophila melanogaster dpp* enhancer. Five 'strong' dorsal binding sites (S1–S5) were previously mapped by *in vitro* DNase footprinting[12]. Note that ChIP-nexus identified S4 as the only site with significant Dorsal binding *in vivo*. ChIP-exo performed by Peconic did not detect any clear Dorsal footprint within the enhancer, in part because of the low read counts obtained. (**d**) Dorsal profiles at the *rho* NEE enhancer. Four Dorsal binding sites (d1–d4) were previously mapped by *in vitro* DNase footprinting[14]. Note that ChIP-nexus identified d3 as the strongest dorsal binding site *in vivo*, consistent with its close proximity to two Twist binding sites. Again, the original ChIP-exo protocol did not detect any clear Dorsal footprint within the enhancer. (**e**) Twist profiles at the same *rho* enhancer. Note that ChIP-nexus showed strong Twist footprints surrounding the two Twist binding sites (t1 and t2)[14]. In this case, ChIP-exo performed by Peconic identified a similar Twist footprint. This shows that the Peconic experiments, which were performed with the same chromatin extracts as the Dorsal experiments, worked in principle but were less robust than our ChIP-nexus experiments. Chr., chromosome.

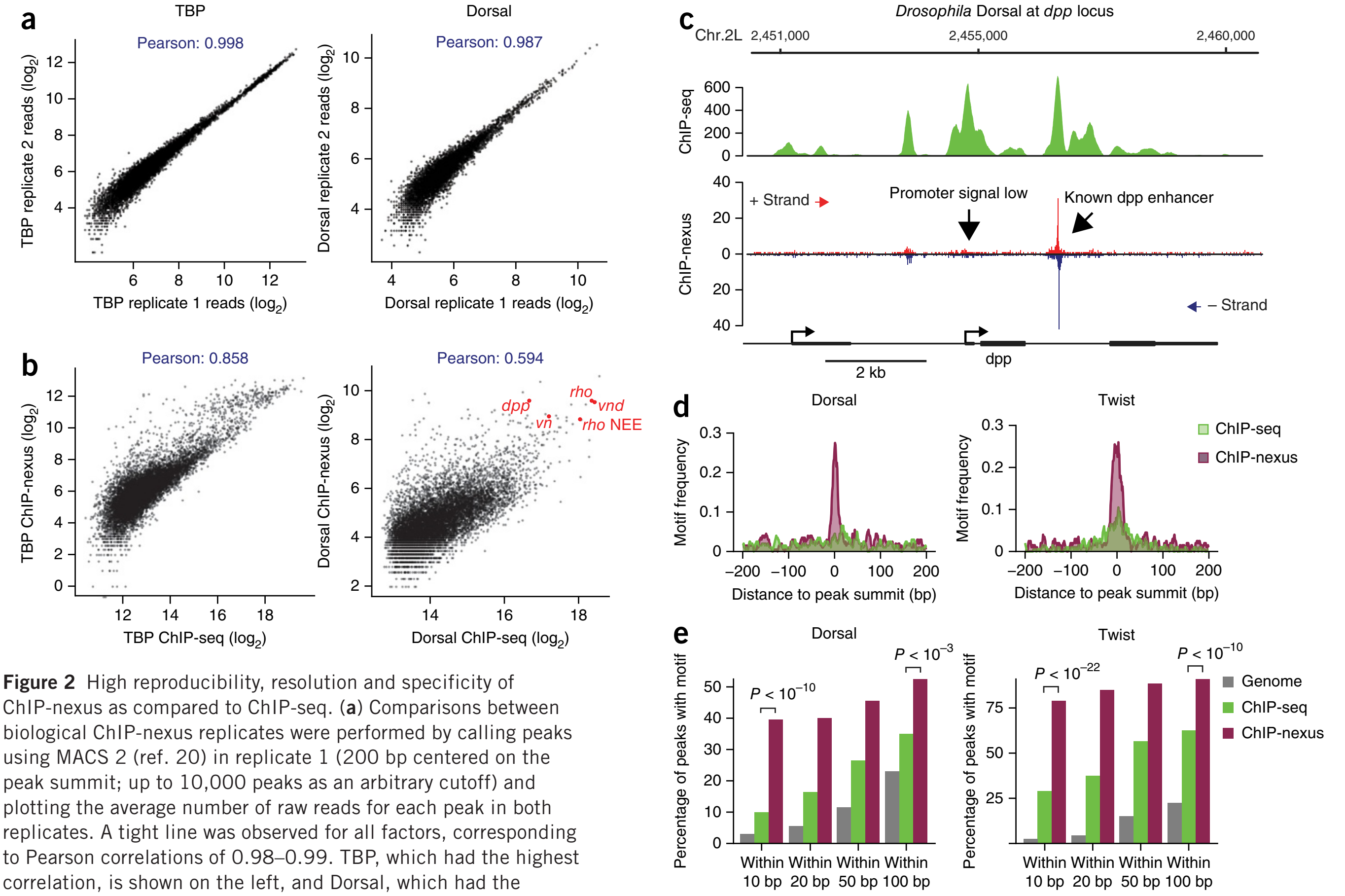

Figure 2 High reproducibility, resolution and specificity of ChIP-nexus as compared to ChIP-seq. (**a**) Comparisons between biological ChIP-nexus replicates were performed by calling peaks using MACS 2 (ref. 20) in replicate 1 (200 bp centered on the peak summit; up to 10,000 peaks as an arbitrary cutoff) and plotting the average number of raw reads for each peak in both replicates. A tight line was observed for all factors, corresponding to Pearson correlations of 0.98–0.99. TBP, which had the highest correlation, is shown on the left, and Dorsal, which had the lowest correlation, is shown on the right. (**b**) Comparison between ChIP-seq and ChIP-nexus. Peaks were called in the ChIP-seq data as in **a**, and reads in these peaks from ChIP-seq and ChIP-nexus data are shown as scatter plots. As can be seen, for both TBP and Twist, there was overall good correlation between the bulk data (Pearson correlations between 0.5 and 0.9). However, the ChIP-nexus data showed an increased signal for a fraction of peaks. (**c**) Examination of individual examples shows that the ChIP-nexus signal was indeed highly specific. For example, the known *dpp* enhancer as shown in **Figure 2** had a strong ChIP-nexus footprint, whereas the signal at the *dpp* promoter, which was equally high in the ChIP-seq data, had much lower and more broadly distributed ChIP-nexus reads without any typical footprint (arrows). (**d**) Frequency distribution of consensus motifs in peaks identified by ChIP-seq (green) and ChIP-nexus (purple). Shown are the examples of Dorsal (left), for which ChIP-nexus showed a dramatic increase in motifs directly at the summit of the peaks, and Twist (right), for which ChIP-nexus showed a more moderate improvement in motif frequency over ChIP-seq. (**e**) Quantification of the motif frequency in random genomic regions, in ChIP-seq peaks and in ChIP-nexus peaks in increasing windows from the peaks' summits for Dorsal and Twist. ChIP-nexus performed much better within a small interval from the peak summit (within 10 bp on either side) (χ^2 test: Dorsal, $P < 10^{-11}$; Twist, $P < 10^{-14}$), underscoring the increased specificity of ChIP-nexus. But even at wider intervals (within 100 bp on either side of the summit), ChIP-nexus peaks contained more motifs (χ^2 test: Dorsal, $P < 2 \times 10^{-3}$; Twist, $P < 10^{-5}$), which suggests that ChIP-nexus has higher specificity than ChIP-seq.

ChIP-nexus produces higher quality data that can be analyzed at the single-gene level.

Next, we studied transcription factors in the early *Drosophila* embryo, where many well-characterized enhancers allowed us to assess the performance of ChIP-nexus compared to other techniques. One of the best-studied transcriptional regulatory networks is dorso-ventral patterning, which is controlled by the activity gradient of Dorsal, the homolog of the vertebrate transcription factor NFκB. One well-characterized enhancer is located in an intron of *decapentaplegic* (*dpp*) and is ventrally repressed by Dorsal[12]. *In vitro* footprinting has shown that Dorsal binds to multiple binding sites in the enhancer, but simultaneous mutation of two specific sites (S3 and S4) almost completely abolishes ventral repression[12]. Our ChIP-nexus data showed a clear footprint of Dorsal at the previously mapped S4 binding site, but not at S3 or other previously mapped Dorsal sites (**Fig. 1c**), which suggested that S4 is the most critical site for *dpp* repression. We also noted that the boundaries of the ChIP-nexus footprint were similar to those of DNase footprints *in vitro*[12], extending beyond the NFkB consensus motif by a similar number of nucleotides.

To further test whether ChIP-nexus footprints are preferentially found at critical binding sites, we also analyzed Dorsal interactions at the extensively characterized *rhomboid* (*rho*) enhancer, which drives expression in the neuroectoderm (NEE)[13–15]. *In vitro* footprinting has revealed four Dorsal sites in the *rho* NEE enhancer (d1–d4), and simultaneous mutation of d2, d3 and d4 almost completely abolishes the enhancer activity[14]. ChIP-nexus showed a strong Dorsal footprint directly over the d3 binding site, whereas weaker footprints were found at the other Dorsal binding sites (**Fig. 1d**). Indeed, d3 is probably the most important Dorsal binding site, because of its proximity to two E-box motifs[13,16]. Both E-boxes can be bound by the basic helix-loop-helix (bHLH) transcription factor Twist *in vitro*[14] and are important for enhancer activity *in vivo*[14,17]. We therefore tested whether ChIP-nexus with Twist could identify these two binding sites. Indeed, prominent ChIP-nexus footprints of

Twist were found exactly over the two known binding sites next to the d3 Dorsal site (**Fig. 1e**).

In order to compare these results to results from ChIP-exo, we had Peconic LLC perform ChIP-exo experiments with Dorsal and Twist using the original ChIP-exo protocol. Although both experiments were performed in biological replicates from the same chromatin extracts, Twist showed a footprint at the *rho* NEE enhancer, whereas Dorsal did not show footprints at known target sites

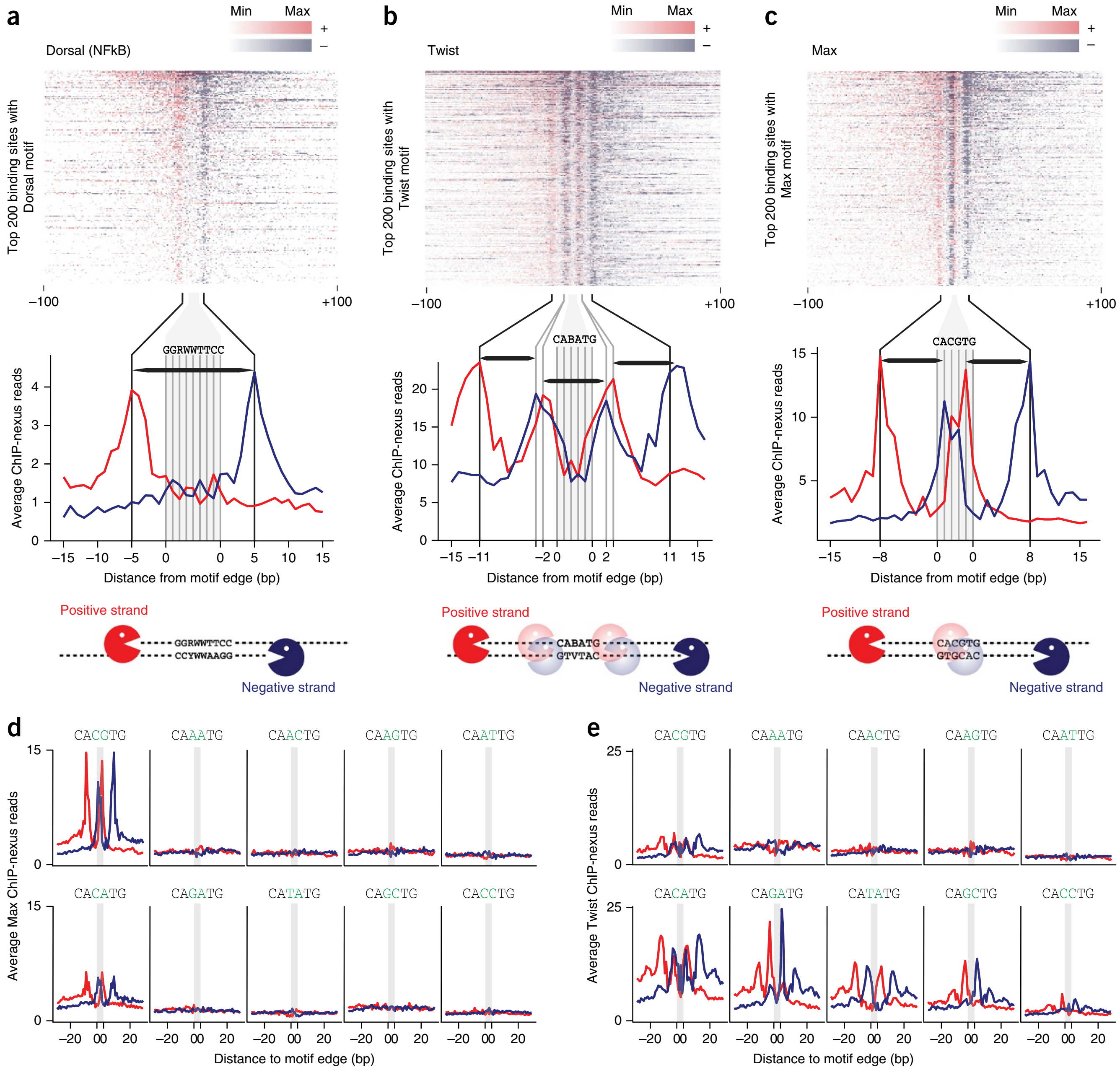

Figure 3 Analysis of Dorsal, Twist and Max *in vivo* footprints. (**a**–**c**) For each factor, the top 200 motifs with the highest ChIP-nexus read counts were selected (shown as a heatmap). The footprints showed a consistent boundary on the positive strand (red) and the negative strand (blue) around each motif. The zoomed-in average profile (bottom) revealed that the footprints were wider than the motifs. A schematic representation of the digestion pattern is shown at the bottom of each panel, with Pac-Man symbols representing λ-exonuclease. (**a**) The ChIP-nexus footprint for Dorsal (NFkB) on its canonical motif (GGRWWTTCC with up to one mismatch) extended on average 5 bp away from the motif edge. Thus, the average dorsal footprint was 18 bp long (horizontal black bar in line graph). (**b**) The Twist ChIP-nexus footprint on the E-box motif CABATG (no mismatch) had two outside boundaries, one at 11 bp and one 2 bp from the motif edge, suggesting interactions with flanking DNA sequences. Each portion of the footprint was about 8–9 bp long (horizontal black bars in line graph). (**c**) The Max ChIP-nexus footprint on its canonical E-box motif (CACGTG, no mismatch) had an outside boundary 8 bp from the motif edge, as well as a boundary inside the motif (at the A-T base), suggesting two partial footprints (horizontal black bars in line graph). (**d**,**e**) Average Max and Twist ChIP-nexus footprints at the top 200 sites for all possible E-box variants (CANNTG). Each variant profile includes its reverse complement. (**d**) Max bound specifically to the canonical CACGTG motif and, to a lesser extent, to the CACATG motif. Note that the Max footprint shape looks identical in the two motifs. (**e**) In contrast, the Twist binding specificity and footprint shape were more complex. Notably, the outer boundary at –11 bp was stronger at the CATATG and CACATG motifs, whereas the inner boundary at –2 bp was stronger at the CAGATG motif.

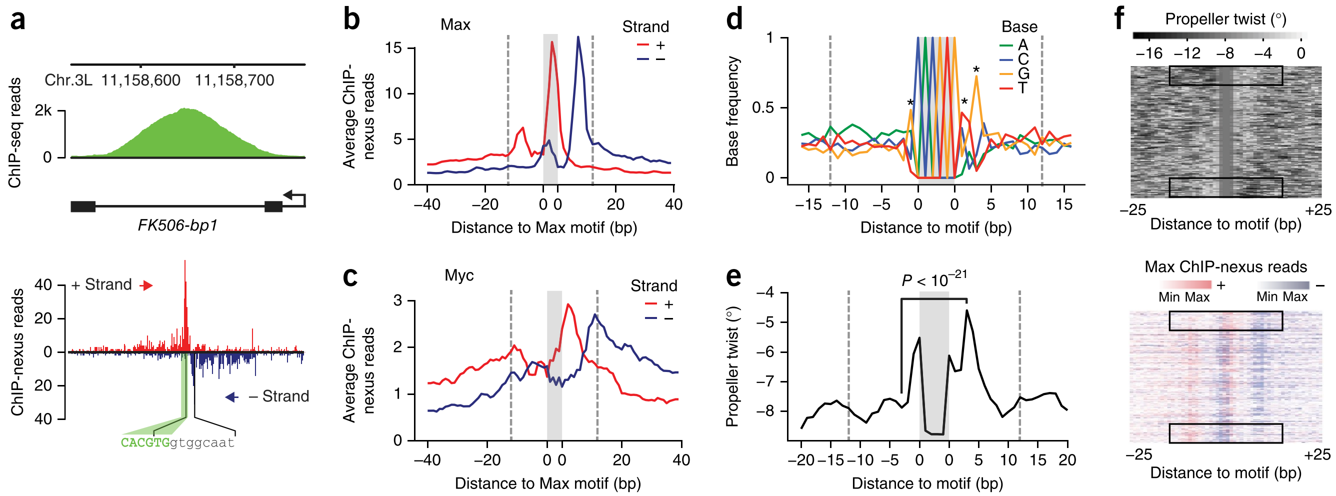

Figure 4 Favored interaction side of Max at E-box motifs correlates with DNA features in the flanking sequences. (**a**) Single-gene examples of the ChIP-nexus footprints show that the Max profile indeed consisted of two separate footprints, one of which was frequently dominant. For example, in the *FK506-bp1* intron, the Max footprint (black brackets) was found to the right of the E-box motif (green). (**b**) Average Max ChIP-nexus profile at the top 200 CACGTG motifs after each footprint had been oriented such that the higher signal was to the right. The area of the motif is shaded in gray, and the extended area of the footprint is demarcated with dashed lines (at 12 bp from the motif to include most reads from the footprint). (**c**) Average Myc ChIP-nexus profile at the same motifs shown in **b** shows that Myc's footprint was generally localized to the same side of the motif as that of Max. (**d**) Average base composition of the oriented E-box motifs from **b**. Significant differences in nucleotides within the area of the footprint are marked with asterisks (χ^2 test; $P < 10^{-24}$ for the G to the right, and $P < 10^{-12}$ for all others). The consensus sequence for orientation to the right was RCACGTGYTG. (**e**) The oriented sequences also showed a marked difference in predicted DNA shape, notably the propeller twist score for a base pair (measured in degrees of rotation). At the third position from the motif, the difference was the highest (paired *t*-test, $P < 10^{-21}$). Note that on the favored interaction side, the predicted propeller twist was more neutral (seen as a peak because of the negative scale). (**f**) Differences in DNA propeller twist in regions flanking the E-box motif correlated with the Max ChIP-nexus footprint level. The top 200 motifs were ordered by the difference in the mean DNA propeller twist measurements in the 6 bp flanking the E-box on either side (top panel). The Max ChIP-nexus heatmap with the same order of motifs (lower panel) shows that the favored interaction side was most pronounced when there was an asymmetry in the DNA propeller twist around the motif (black boxes).

(**Fig. 1c**–**e**). The reduced quality of the Dorsal experiment was due in part to the lower read number obtained (**Fig. 1c**,**d**). However, even the Twist ChIP-exo experiment, which had read counts comparable to those in our ChIP-nexus data, showed a less precise footprint (**Fig. 1e** and **Supplementary Fig.1**), which supported our conclusion that ChIP-nexus produces better results at the single-gene level.

The strong concordance between ChIP-nexus binding and previously characterized sites suggests that ChIP-nexus is an effective approach that can pinpoint critical binding sites within an enhancer. Our analyses of Dorsal also suggested that its *in vivo* binding sites might differ from those bound *in vitro*, consistent with studies on other transcription factors[18,19].

To test the robustness of the ChIP-nexus protocol, we analyzed the correlation between replicates at bound regions. Although peak-finding algorithms are not designed for ChIP-nexus data[2,3,9], we found that MACS[20] (version 2) and Peakzilla[2] identified thousands of binding peaks in all cases. When a maximum of 10,000 peaks was used, the ChIP-nexus reads were highly correlated between replicates (**Fig. 2a**; Pearson correlations: TBP, 0.998; Dorsal, 0.986; and Twist, 0.993), which shows that our ChIP-nexus data are highly reproducible.

We next analyzed the relationship between ChIP-nexus and ChIP-seq signals. The Pearson correlation of the reads was lower than that between replicates but was still very high (**Fig. 2b**; TBP, 0.858; Dorsal, 0.594). Scatter plots confirmed that the bulk signal was similar in ChIP-nexus and ChIP-seq but that many bound regions had higher signals in the ChIP-nexus data (**Fig. 2b**). Regions with higher ChIP-nexus:ChIP-seq ratios included many known Dorsal enhancers (e.g., *rho* NEE, *dpp*, *zen*, *vnd* and *vn*), whereas regions with lower ChIP-nexus:ChIP-seq signal ratios often lacked a specific footprint, which indicated that they might have been enriched through unspecific binding to open chromatin. For instance, the *dpp* promoter showed high Dorsal ChIP-seq enrichments comparable to those of the known *dpp* enhancer, but it had no specific footprint in the ChIP-nexus data (**Fig. 2c**).

To test more systematically whether ChIP-nexus indeed has increased specificity and resolution compared to ChIP-seq, we analyzed the presence and location of consensus binding motifs within peaks (**Fig. 2c**,**d**). Among the top 200 Dorsal and Twist ChIP-nexus binding peaks, the corresponding consensus motif was found directly at the center of the ChIP-nexus binding peaks much more frequently than at the center of the ChIP-seq binding peaks (**Fig. 2d**), underscoring the increased resolution. Indeed, within 10 bp of the peak summit, there was a significant improvement in motif enrichment in the ChIP-nexus data compared to the ChIP-seq data (χ^2 test: Dorsal, $P < 10^{-10}$; Twist, $P < 10^{-22}$; **Fig. 2e**). Yet even at 100 bp from the summit, ChIP-nexus still had significantly greater motif enrichment than ChIP-seq (χ^2 test: Dorsal, $P < 10^{-3}$; Twist, $P < 10^{-10}$; **Fig. 2e**), supporting the notion that ChIP-nexus has not only improved resolution but also improved specificity.

We next examined the binding profile of the footprint of Dorsal when bound to a Dorsal consensus binding motif (GGRWWTTCC). Using the 200 motifs with the highest ChIP-nexus counts, we generated the average Dorsal footprint (**Fig. 3a**). It was very similar to the footprints on known Dorsal targets, with the boundaries located five nucleotides upstream of the motif. This distance is consistent with the crystal structure of NFkB, which also suggests that the footprint was wider than the binding sequence[16]. Whether λ-exonuclease stops exactly at the protein-DNA boundary or a few nucleotides before remains unclear.

Next we analyzed the ChIP-nexus footprint of Twist over the known binding motifs (CABATG; thus CATATG, CACATG or CAGATG). We found that Twist had two boundaries, one located 11 nucleotides upstream and the other 2 nucleotides upstream of the motif (**Fig. 3b**), indicating interactions between Twist and the DNA flanking sequences outside the binding motif. To obtain further insights into the binding of bHLH transcription factors in general, we then analyzed Max, which binds to the palindromic E-box CACGTG either as a homodimer or as a heterodimer with other bHLH proteins such as Myc[21,22]. The average ChIP-nexus footprint of Max had a second set of boundaries located 8 bp upstream of the motif (**Fig. 3c**), again indicating interactions with flanking DNA sequences. The crystal structures of Max-Max, Max-Myc and Max-Mad included only 6 bp flanking either side of the E-box motif and did not use full-length Max or Myc[23,24]. However, *in vitro* footprinting assays of Max and Myc show protection of four to six bases beyond the motif[25,26], consistent with our results.

We next tested whether the binding footprints of Max and Twist vary across E-box variants of the pattern CANNTG (**Fig. 3d**,**e**). For each possible middle sequence, we selected the 200 motifs with the highest ChIP-nexus read counts. As expected, Max binding was strongest at the canonical CACGTG motif. A weaker but similar pattern was detected at the CACATG motif (**Fig. 3d**), consistent with its binding specificity as measured by a bacterial one-hybrid system[27]. Consistent with previous data[17,27], Twist binding occurred at multiple E-boxes (**Fig. 3e**). But the shapes of these footprints varied in that the outer boundary (at 11 bp from the motif) was dominant at the CATATG motif and, to a lesser extent, the CACATG motif, the two motifs with the highest evolutionary conservation across *Drosophila* species[17]. In contrast, the inner boundary (at 2 bp from the motif) was more prominent at the CAGATG motif. Although the basis for these differences in footprints is unknown, the results might indicate an unappreciated specificity in the way transcription factors are detected *in vivo*.

The average footprint of Max suggested interactions with flanking DNA sequences on both sides of the motif. Inspection of the footprints at individual genes, however, suggested that Max often favored interaction at one side of the motif (**Fig. 4a**). A favored interaction side was also found for Twist, especially at the CATATG motif (**Supplementary Fig. 1**), but here we focus on the analysis of Max.

To analyze the basis for the Max binding asymmetry, we determined the dominant side for each of the top 200 Max binding footprints (on the basis of the difference in read counts observed between the right and left side of each motif). Because the CACGTG motif is palindromic and thus not strand specific, we oriented the binding footprints such that the dominant side was to the right of the motif. The average footprint observed after the motifs were oriented is shown in **Figure 4b**. We then searched for differences between the left side and the right side.

To test whether binding to a half-site might reflect the binding of Max as a heterodimer with its partner Myc, we performed ChIP-nexus with Myc. We assumed that if the Myc-Max heterodimer determined the orientation, the trend of the Myc footprint would be opposite that of the Max footprint at the oriented binding sites (i.e., the higher signal would be found to the left of the motif). Although there were differences between the binding footprints, the Myc profile, like the Max profile, was oriented to the right (**Fig. 4c**), which suggested that the favored interaction side was not determined by heterodimer orientation.

Next, we searched for differences in the DNA sequences surrounding the Max motif that could explain the favored interaction side (**Fig. 4d**,**e**). We found that the base composition showed significant biases next to the E-box (indicated by asterisks in **Fig. 4d**), which created a directional motif of the consensus RCACGTGYTG. The nucleotide biases outside the motif could either mediate direct contacts with the Max-Myc dimer or indirectly affect the protein interactions through the overall DNA shape[28]. Indeed, the specificity of bHLH factors has been shown to correlate with parameters of DNA shape in flanking sequences[19,29]. We therefore examined predicted DNA shape parameters[30] for all 200 sequences and found that the 'propeller twist', a measurement of the relative rotation between two paired bases, was on average significantly stronger at the less favored interaction side (**Fig. 4e**; paired *t*-test, $P < 10^{-21}$). To visualize the correlation between propeller twist and favored interaction side, we sorted our 200 Max footprints on the basis of the difference in propeller twist between the two sides and then plotted the Max footprint in the same order (**Fig. 4f**). This showed that a strong asymmetry with regard to the propeller twist was highly correlated with the favored interaction side.

In summary, ChIP-nexus achieved increased resolution compared to conventional ChIP-seq and enhanced robustness compared to ChIP-exo, providing a detailed view of the *in vivo* binding landscape of transcription factors. ChIP-nexus uses a similar amount of cells as ChIP-seq, but it pinpoints binding sites within individual enhancers more precisely and provides new information on how different motif variants are bound *in vivo*. Although high-resolution *in vivo* binding data can also be obtained by digital genomic footprinting[31], that method requires substantially more sequencing depth and does not reveal the identity of the bound transcription factors.

The increased resolution suggests that the Max binding footprint was influenced by DNA sequences flanking the motif and that this interaction was stronger at one side of the motif. The favored interaction side correlated with differences in specific nucleotides, as well as in parameters of DNA shape, and might explain why the reads from conventional ChIP-seq experiments often do not peak directly over the binding motif (e.g., Twist at the CATATG motif[17]). Although we cannot exclude the possibility that the favored side is the preferred side for cross-linking by formaldehyde, it is unlikely that this is the only explanation. It is becoming more and more evident that local DNA features around a motif contribute to the specificity of protein-DNA interactions, whether measured *in vitro* without formaldehyde cross-linking[19] or *in vivo* using reporter assays[32]. Thus, it is possible that Max indeed has a favored interaction side *in vivo*, but whether this preference has a functional consequence is not known.

The high resolution and robustness of the protocol open the possibility for a more extended analysis of the *in vivo* binding site specificity of transcription factors. For example, ChIP-nexus is ideally suited for the identification of single-nucleotide polymorphisms that alter transcription factor binding, either across species or between individuals in a population. Furthermore, because it precisely identifies which binding motif is bound *in vivo*, it will help in identifying the influence of nucleosomes, other transcription factors or DNA methylation on the *in vivo* binding of transcription factors. Therefore, ChIP-nexus could become a useful tool for untangling the mechanisms of combinatorial regulation.

METHODS

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

Accession codes. GEO: GSE55306.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank M. Blanchette and A. Stark for discussions and R. Krumlauf and R. Mohan for critical comments on the manuscript. This work was funded by the US National Institutes of Health (New Innovator Award 1DP2 OD004561 to J.Z.) and the Stowers Institute for Medical Research.

AUTHOR CONTRIBUTIONS

Q.H. and J.Z. conceived and designed the ChIP-nexus protocol. Q.H. performed all experiments. J.J. developed all computational analysis tools. Q.H., J.J. and J.Z. analyzed and interpreted the data and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

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

1. Spitz, F. & Furlong, E.E. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
2. Bardet, A.F. *et al.* Identification of transcription factor binding sites from ChIP-seq data at high resolution. *Bioinformatics* **29**, 2705–2713 (2013).
3. Rhee, H.S. & Pugh, B.F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408–1419 (2011).
4. Rhee, H.S. & Pugh, B.F. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr. Protoc. Mol. Biol.* Chapter 21, Unit 21.24 (2012).
5. Rhee, H.S. & Pugh, B.F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483**, 295–301 (2012).
6. Yen, K., Vinayachandran, V., Batta, K., Koerber, R.T. & Pugh, B.F. Genome-wide nucleosome specificity and directionality of chromatin remodelers. *Cell* **149**, 1461–1473 (2012).
7. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2012).
8. Casbon, J.A., Osborne, R.J., Brenner, S. & Lichtenstein, C.P. A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res.* **39**, e81 (2011).
9. Venters, B.J. & Pugh, B.F. Genomic organization of human transcription initiation complexes. *Nature* **502**, 53–58 (2013).
10. Serandour, A.A., Brown, G.D., Cohen, J.D. & Carroll, J.S. Development of an Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties. *Genome Biol.* **14**, R147 (2013).
11. König, J. *et al.* iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* **17**, 909–915 (2010).
12. Huang, J.D., Schwyter, D.H., Shirokawa, J.M. & Courey, A.J. The interplay between multiple enhancer and silencer elements defines the pattern of decapentaplegic expression. *Genes Dev.* **7**, 694–704 (1993).
13. Fakhouri, W.D. *et al.* Deciphering a transcriptional regulatory code: modeling short-range repression in the *Drosophila* embryo. *Mol. Syst. Biol.* **6**, 341 (2010).
14. Ip, Y.T., Park, R.E., Kosman, D., Bier, E. & Levine, M. The dorsal gradient morphogen regulates stripes of rhomboid expression in the presumptive neuroectoderm of the *Drosophila* embryo. *Genes Dev.* **6**, 1728–1739 (1992).
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. Szymanski, P. & Levine, M. Multiple modes of dorsal-bHLH transcriptional synergy in the *Drosophila* embryo. *EMBO J.* **14**, 2229–2238 (1995).
17. Ozdemir, A. *et al.* High resolution mapping of Twist to DNA in *Drosophila* embryos: efficient functional analysis and evolutionary conservation. *Genome Res.* **21**, 566–577 (2011).
18. Liu, X., Lee, C.K., Granek, J.A., Clarke, N.D. & Lieb, J.D. Whole-genome comparison of Leu3 binding *in vitro* and *in vivo* reveals the importance of nucleosome occupancy in target site selection. *Genome Res.* **16**, 1517–1528 (2006).
19. Gordân, R. *et al.* Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.* **3**, 1093–1104 (2013).
20. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X.S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).
21. Blackwood, E.M. & Eisenman, R.N. Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc. *Science* **251**, 1211–1217 (1991).
22. Prendergast, G.C., Lawe, D. & Ziff, E.B. Association of Myn, the murine homolog of max, with c-Myc stimulates methylation-sensitive DNA binding and ras cotransformation. *Cell* **65**, 395–407 (1991).
23. Ferré-D'Amaré, A.R., Prendergast, G.C., Ziff, E.B. & Burley, S.K. Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. *Nature* **363**, 38–45 (1993).
24. Nair, S.K. & Burley, S.K. X-ray structures of Myc-Max and Mad-Max recognizing DNA. Molecular bases of regulation by proto-oncogenic transcription factors. *Cell* **112**, 193–205 (2003).
25. Walhout, A.J., Gubbels, J.M., Bernards, R., van der Vliet, P.C. & Timmers, H.T. c-Myc/Max heterodimers bind cooperatively to the E-box sequences located in the first intron of the rat ornithine decarboxylase (ODC) gene. *Nucleic Acids Res.* **25**, 1493–1501 (1997).
26. Wechsler, D.S., Papoulas, O., Dang, C.V. & Kingston, R.E. Differential binding of c-Myc and Max to nucleosomal DNA. *Mol. Cell. Biol.* **14**, 4097–4107 (1994).
27. Zhu, L.J. *et al.* FlyFactorSurvey: a database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res.* **39**, D111–D117 (2011).
28. Rohs, R. *et al.* The role of DNA shape in protein-DNA recognition. *Nature* **461**, 1248–1253 (2009).
29. Yang, L. *et al.* TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res.* **42**, D148–D155 (2014).
30. Zhou, T. *et al.* DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res.* **41**, W56–W62 (2013).
31. Hesselberth, J.R. *et al.* Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nat. Methods* **6**, 283–289 (2009).
32. White, M.A., Myers, C.A., Corbo, J.C. & Cohen, B.A. Massively parallel *in vivo* enhancer assay reveals that highly local features determine the *cis*-regulatory function of ChIP-seq peaks. *Proc. Natl. Acad. Sci. USA* **110**, 11952–11957 (2013).

ONLINE METHODS

Preparation of K562 cells. K562 cells from ATCC were grown at 37 °C in 5% CO_2 with humidity in Iscove's DMEM with 10% FBS. Ten million cells were harvested for each ChIP-seq or ChIP-nexus experiment. Cells were cross-linked in 1% formaldehyde (in 50 mM HEPES-KOH, pH 7.5, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA) and rotated for 10 min at room temperature. We quenched cross-linking by adding glycine to 0.125 M to cells and rotating them for 5 min at room temperature. Cells were spun down, washed with PBS, resuspended in A1 buffer (15 mM HEPES, pH 7.5, 15 mM NaCl, 60 mM KCl, 4 mM $MgCl_2$, 0.5% Triton X-100, 0.5 mM dithiothreitol), transferred to a Wheaton Dounce homogenizer and broken down by 20 strokes with each pestle. Homogenates were spun down at 3,000*g* and washed three times with A1 buffer and once with A2 buffer (15 mM HEPES, pH 7.5, 140 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 1% Triton X-100, 0.1% sodium deoxycholate, 1% SDS, 0.5% *N*-lauroylsarcosine sodium). Nuclei were resuspended in 0.7 ml A2 buffer. The chromatin was fragmented with a Bioruptor by two rounds of 15 min of sonication at high power. Chromatin was cleared by centrifugation, and the supernatant was used for ChIP.

Preparation of *Drosophila* embryos. *D. melanogaster* embryos from Oregon-R flies raised and kept at 25 °C and 60% humidity were collected on apple plates. The apple plates were placed into fly cages for 2 h and then incubated for another 2 h outside the cage such that the embryos were aged 2–4 h after egg laying. Embryo collections and whole-cell extract preparations were performed as previously described[33,34]. About 0.1 g of fixed embryos was used per ChIP-seq or ChIP-nexus experiment.

Preparation of *Drosophila* S2 cells. S2 cells from Invitrogen were grown at 25 °C in HyClone SFX-Insect Cell Culture Media with 1× penicillin and streptomycin (Sigma-Aldrich). About 20 million cells were harvested for each ChIP-seq or ChIP-nexus experiment. S2 sells were cross-linked with 1% formaldehyde for 10 min at room temperature. Formaldehyde was quenched with 0.125 M glycine for 5 min. Cells were washed with PBS, resuspended in Orlando and Paro's Buffer A (0.25% Triton X-100, 10 mM EDTA, 0.5 mM EGTA, 10 mM Tris-HCl, pH 8.0) and rotated for 10 min at room temperature. Nuclei were spun down and resuspended in RIPA buffer (10 mM Tris-HCl, pH 8.0, 140 mM NaCl, 0.1% SDS, 0.1% sodium deoxycholate, 0.5% sarkosyl, 1% Triton X-100). The chromatin was fragmented with a Bioruptor by two rounds of 15 min of sonication at high power. Chromatin was cleared by centrifugation, and the supernatant was used for ChIP.

Chromatin immunoprecipitation. Chromatin immunoprecipitations were performed in biological duplicate as previously described[35] with rabbit polyclonal antibodies to TBP (sc-204X, 3 μg/ChIP), Dorsal (20 μg/ChIP), Twist (10 μg/ChIP), Max (sc-28209, 8 μg/ChIP) and Myc (sc-28207, 8 μg/ChIP). Custom rabbit polyclonal antibodies to Dorsal protein (amino acids 39–346) and Twist protein (C-terminal amino acids 340–490) were produced by GenScript. The ChIP-seq patterns obtained with these antibodies matched those obtained previously[34]. Enrichments for each transcription factor of interest were confirmed at known target sites by real-time PCR (StepOnePlus, Applied Biosystems) before library preparation. Primers are available upon request.

ChIP-nexus oligonucleotides. Nex_adaptor_UBamHI: /5Phos/GATCGGAAGAGCACACGTCTGGATCCACGACGCTCTTCC.

Nex_adaptor_BN5BamHI: /5Phos/TCAGNNNNNAGATCGGAAGAGCGTCGTGGATCCAGACGTGTGCTCTTCCGATCT.

To anneal the two Nex_adaptor oligonucleotides, we mixed 50 μmol of each in 1× TE buffer with 50 mM NaCl and placed them in a thermocycler at 95 °C for 5 min, after which the temperature was ramped down to 25 °C at a rate of ~3.5 °C/min and then held at 25 °C for 30 min.

Nex_cut_BamHI (cut oligo): GAAGAGCGTCGTGGATCCAGACGTG.

Nex_primer_U (universal PCR primer with 3′ phosphoro-thioate bond): AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T.

Nex_primer_B01 (barcoded PCR primer with 3′ phosphoro-thioate bond; other barcodes may be used): CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T.

ChIP-nexus digestion steps. Digestion with λ-exonuclease was carried out using a modified version of the published ChIP-exo protocol[3,4], and the chromatin was immunoprecipitated on Dynabeads.

The chromatin was first washed five times with the following buffers: wash buffer A (10 mM Tris-EDTA, 0.1% Triton X-100), wash buffer B (150 mM NaCl, 20 mM Tris-HCl, pH 8.0, 5 mM EDTA, 5.2% sucrose, 1.0% Triton X-100, 0.2% SDS), wash buffer C (250 mM NaCl, 5 mM Tris-HCl, pH 8.0, 25 mM HEPES, 0.5% Triton X-100, 0.05% sodium deoxycholate, 0.5 mM EDTA), wash buffer D (250 mM LiCl, 0.5% IGEPAL CA-630, 10 mM Tris-HCl, pH 8.0, 0.5% sodium deoxycholate, 10 mM EDTA) and Tris buffer (10 mM Tris, pH 7.5, 10 mM Tris, pH 8.0, or 10 mM Tris, pH 9.5, depending on the next enzymatic step).

After the last wash, residual buffer was drained before the next enzymatic reaction was initiated. These washing steps were repeated between all subsequent steps.

To repair the DNA ends, each sample was incubated at 12 °C for 30 min with 0.05 U/μl DNA polymerase I, large fragment (New England BioLabs, M0210), 0.15 U/μl T4 DNA polymerase (New England BioLabs, M0203), 0.5 U/μl T4 polynucleotide kinase (New England BioLabs, M0201) and 0.4 mmol/μl deoxynucleotide triphosphates (dNTPs) in 30–40 μl 1× NEB T4 ligase buffer (New England BioLabs, B0202). Incubation was followed by washing steps as above.

For dA tailing, each sample was incubated at 37 °C for 30 min with 0.3 U/μl Klenow fragment ($3'\rightarrow5'$ exo$^-$) (New England BioLabs, M0212) and 0.2 mmol/μl ATP in 50 μl 1× NEBuffer 2. Incubation was followed by washing steps as above.

The adaptors were then ligated by incubation at 25 °C for 60 min in 200 U/μl Quick T4 DNA ligase (New England BioLabs, M2200) and 60 nmol/μl Nex_adaptor in 50 μl 1× Quick Ligation Reaction Buffer (New England BioLabs, B6058S). Incubation was followed by washing steps as above.

To fill the ends of the adaptors, each sample was incubated at 37 °C for 30 min with 0.1 U/μl Klenow fragment ($3'\rightarrow5'$ exo$^-$) (New England BioLabs, M0212) and 0.1 mmol/μl dNTPs in 50 μl 1× NEBuffer 2. Incubation was followed by washing steps as above.

The ends were then trimmed by incubation at 12 °C for 5 min in 0.09 U/μl T4 DNA polymerase (New England BioLabs, M0203) and 0.1 mmol/μl dNTPs in 50 μl 1× NEBuffer 2, with incubation followed by washing steps as above.

For λ-exonuclease digestion, each sample was incubated at 37 °C for 60 min with constant agitation in 0.2 U/μl λ-exonuclease (New England BioLabs, M0262), 5% dimethyl sulfoxide (DMSO) and 0.1% Triton X-100 in 100 μl 1× Lambda Exonuclease Reaction Buffer (New England BioLabs, B0262S). Incubation was followed by washing steps as above.

Finally, RecJf exonuclease digestion occurred at 37 °C for 60 min with constant agitation in 0.75 U/μl RecJf exonuclease (New England BioLabs, M0264), 5% DMSO and 0.1% Triton X-100 in 100 μl 1× NEBuffer 2. After RecJf digestion, the Dynabeads were washed three times with RIPA buffer (50 mM HEPES, pH 7.5, 1 mM EDTA, 0.7% sodium deoxycholate, 1% IGEPAL CA-630, 0.5 M LiCl). DNA elution, reverse cross-linking, DNA purification and precipitation were performed as previously described[34,35].

ChIP-nexus library preparation. The library-preparation protocol was based on the iCLIP protocol[11]. After the DNA was purified and precipitated, each sample was dissolved in 11.25 μl H_2O, 1.5 μl 10× CircLigase buffer, 0.75 μl 1 mM ATP, 0.75 μl 50 mM $MnCl_2$, 0.75 μl CircLigase (Epicentre) and incubated at 60 °C for 60 min for self-circularization. To anneal the oligonucleotide complementary to the BamHI restriction site (cut oligo Nex_cut_BamHI), we added 26 μl H_2O, 5 μl FastDigest buffer (Fermentas) and 1 μl 10 μM cut oligo to each sample. The mixture was incubated on a thermocycler as follows: 95 °C for 5 min, then ramped down to 25 °C at a rate of ~3.5 °C/min and held at 25 °C for 30 min. For BamHI digestion, 3 μl Fastdigest BamHI (Fermentas) was added, and the sample was incubated at 37 °C for 30 min. The samples were then precipitated by the addition of 150 μl TE buffer, 30 μg glycogen, 20 μl 3 M/l sodium acetate (pH 5.5) and 500 μl 100% ethanol and incubated at −80 °C for 2.5 h. After centrifugation at 4 °C for 30 min at 16,100*g*, the samples were washed with 500 μl 80% ethanol, dried overnight at room temperature and resuspended in 36 μl H_2O.

For PCR amplification, 10 μl 5× Phusion buffer, 1.5 μl 10 mM dNTP, 1 μl each of 10 μM universal and barcoded PCR primers (Nex_primer_U and

doi:10.1038/nbt.3121

Nex_primer_B01), and 0.5 µl Phusion Polymerase (New England BioLabs, M0530) were added to each sample in a total volume of 50 µl. The DNA was amplified with the following program: 98 °C for 30 s; 18× (98 °C for 10 s, 65 °C for 30 s, 72 °C for 30 s); 72 °C for 5 min. To remove contaminating adaptor dimers, the PCR products were run on a 2% agarose gel. The adaptor dimers usually formed a thin, bright band migrating at the front edge of the library DNA, which formed a smear. The library DNA was carefully sliced out, purified with a MinElute kit (Qiagen, 28006) and eluted into 12 µl elution buffer. After Bioanalyzer analysis, libraries were sequenced on an Illumina HiSeq platform with the single-end sequencing primer over 50 cycles of extension according to the manufacturer's instructions.

Data processing for ChIP-nexus samples. Sequencing reads passing the default Illumina quality filter (CASAVA v1.8.2) were further filtered for the presence of the fixed barcode CTGA starting at read position 6. The random and fixed barcode sequences were then removed (read positions 1–9), with the 5-bp random barcode sequence for each read retained separately. Adaptor sequences from the right end were then trimmed using the cutadapt tool[36]. All reads at least 22 bp in length after adaptor trimming were then aligned to the appropriate reference genome (dm3 for *D. melanogaster* and hg19 for *Homo sapiens*) using bowtie v1.0.0 (ref. 37). Only uniquely aligning reads with a maximum of two mismatches were kept. To remove duplicates, we removed reads with identical alignment coordinates (chromosome, start position and strand) and identical random barcodes using R[38] and Bioconductor[39]. All reads were then split by strand orientation, and a genome-wide count of the start positions (λ-exonuclease's stop position) was calculated for each strand.

Data processing for ChIP-exo samples. The published ChIP-exo TBP samples from human K562 cells[9] were downloaded from the Sequence Read Archive (run SRR770743, accession number SRX248184; run SRR770744, accession number SRX248185) and aligned to the UCSC hg19 reference genome using the same parameters as for ChIP-nexus samples. Peconic provided aligned BAM files for both Dorsal and Twist ChIP-exo replicates. Aligned reads for all ChIP-exo experiments were separated by strand and reduced to the first sequenced base (λ-exonuclease's stop position), and genome-wide counts for read start positions were calculated.

Data processing for ChIP-seq samples. ChIP-seq reads were aligned to the appropriate reference genome (dm3 or hg19) using the same parameters as for the ChIP-nexus samples. After alignment, reads were extended in the 5′-to-3′ direction to each sample's estimated library insert size as determined by a Bioanalyzer. These extensions were 136 bp for Dorsal, 124 bp for Twist, 83 bp for Max and 74 bp for TBP. After extension, genome-wide coverage values were calculated.

Reference genome modification for *Drosophila* Oregon-R embryos. Multiple SNPs in our Oregon-R strain resulted in gaps in read coverage at a number of regions of interest (including the *rho* enhancer used as an example). To correct this, we combined the Dorsal and Twist ChIP-seq samples and realigned them to the reference genome while allowing up to three mismatches. Samtools[40] was then used to identify variants genome-wide using the following parameters:

```
samtools mpileup -uD -f dm3.fasta embryo_
combined_chipseq.bam|bcftools view -vcg
```

The identified single-allele variants were then used to create a modified reference genome matching the sequence of our Oregon-R strain. ChIP-seq and ChIP-nexus samples for Dorsal and Twist were aligned to this modified reference genome. As Peconic did not provide the unaligned reads for the Dorsal ChIP-exo data, we were able to perform this read recovery procedure only on our ChIP-seq and ChIP-nexus data.

Peak calling. MACS v2.0.10 (ref. 20) was run on the ChIP-nexus replicate 1 samples and the ChIP-seq samples for TBP, Dorsal, Twist and Max using the following parameters:

```
macs2 callpeak -g dm --keep-dup=all --call-summits
```

Resulting peak summits were sorted by score, and a maximum of 10,000 were retained per sample.

Comparison scatter plots. For each scatter plot, the peaks detected in the sample on the *x*-axis were resized to 201 bp centered at the summit. Each peak was scored using the genome-wide coverage values for the two samples. For ChIP-seq, these coverage values were calculated using the entire extended fragment size. For ChIP-nexus and ChIP-exo, coverage values were calculated using only the first base pair of each aligned fragment. Pearson correlations were calculated using the raw values before log transformation.

ChIP-nexus and ChIP-seq motif presence. For Dorsal, Twist and Max, we used the top 200 peaks according to MACS score. Motif frequency plots were generated by scoring each position in the genome as either 1 or 0 on the basis of the presence of a consensus motif for each factor. These consensus motifs were GGRWWTTCC with up to one mismatch for Dorsal, CABATG with no mismatches for Twist and CACGTG with no mismatches for Max. The average motif presence around the top 200 peak summits was then calculated and plotted for both ChIP-seq and ChIP-nexus (replicate 1) samples.

For each peak, the distance from the peak summit to the nearest consensus motif was calculated. For distance thresholds of 10, 20, 50 and 100 bp, a two-sided χ^2 test was used to test for a significant difference in the proportion of peaks near a consensus motif between ChIP-nexus and ChIP-seq.

Motif average profiles and heatmaps. For each factor, we scored all non-overlapping instances of its motif with up to one mismatch for ChIP-nexus signal (replicate 1) by summing the total reads from both strands in a fixed region centered on the motif (29 bp for Dorsal, 15 bp for Max and 51 bp for Twist). The heatmaps of the top 200 motifs were oriented such that the motif was on the positive strand and were sorted by total reads in a 50-bp window centered on the motif. Positive and negative strand reads (relative to the strand of the motif) were normalized from zero reads (minimum) to the read value at the 98th percentile or higher (maximum) for display.

We constructed the E-box specificity plots shown in **Figure 3** by separately averaging the positive-strand and negative-strand ChIP-nexus signals among the top-scoring 200 non-overlapping instances of each unique E-box motif CANNTG. We scored each motif by summing the ChIP-nexus reads in a 50-bp window centered on the motif.

To analyze the favored interaction side of Max (**Fig. 4**), we scored the same top 200 Max motifs described above for ChIP-nexus signal on the left and right sides on the basis of the observed average pattern. We calculated the left-side signal by summing the positive-strand reads in a region 9 bp wide centered 8 bp upstream of the motif and the negative-strand reads in a region 9 bp wide centered on the motif +1 position. We calculated the right-side signal by summing the positive-strand reads in a region 9 bp wide centered on the motif +4 position and the negative-strand reads in a region 9 bp wide centered 8 bp downstream of the motif. Each motif was then oriented so that the side with the higher signal was to the right of the motif.

Analysis of DNA shape. Genome-wide DNA shape parameters were collected for the positive strand of the *D. melanogaster* UCSC dm3 reference genome. First, all 1,024 DNA pentamers were uploaded to the DNA Shape Web service[30] to obtain predictions for minor groove width and propeller twist. For both DNA shape parameters, a single value was provided for the center base of each pentamer. We applied these values genome-wide by aligning the pentamers to the positive strand of the reference genome.

To order the top 200 Max-bound E-box motifs by the difference in DNA propeller twist (**Fig. 4f**), we calculated the mean propeller twist for the 6 bp immediately to the left and right of the motif. The motifs were then ordered by the difference between right and left mean propeller twist.

Availability of data, analysis code and experimental protocol. All analysis codes used for data processing and figure generation are available via GitHub at https://github.com/zeitlingerlab. In addition, we have prepared a Linux virtual machine containing all software tools, analysis codes, raw data and processed data used in this study. Instructions for accessing the virtual machine via Amazon Web Services, as well as a detailed ChIP-nexus protocol, can be found at our website (http://research.stowers.org/zeitlingerlab). Individual sample accession numbers and alignment statistics are available in **Supplementary Table 1**.

33. Sandmann, T. *et al.* A core transcriptional network for early mesoderm development in *Drosophila melanogaster*. *Genes Dev.* **21**, 436–449 (2007).
34. 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).
35. 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).
36. Martin, M. Cutadapt removes adaptor sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
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. R Development Core Team R: a language and environment for statistical computing. http://www.R-project.org/ (2013).
39. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
40. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

doi:10.1038/nbt.3121