Analysis of Genetic Variation Indicates DNA Shape Involvement in Purifying Selection

Xiaofei Wang,[†,1] Tianyin Zhou,[*,†,‡,1] Zeba Wunderlich,[2] Matthew T. Maurano,[3] Angela H. DePace,[4] Sergey V. Nuzhdin,[1] and Remo Rohs[*,1,5]

[1]Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA

[2]Department of Developmental and Cell Biology, University of California Irvine, Irvine, CA

[3]Institute for Systems Genetics, New York University Medical Center, New York, NY

[4]Department of Systems Biology, Harvard Medical School, Boston, MA

[5]Departments of Chemistry, Physics and Astronomy, and Computer Science, University of Southern California, Los Angeles, CA

[†]These authors contributed equally to this work.

[‡]Present address: Google Inc., Mountain View, CA

***Corresponding authors:** E-mails: tianyinz@alumni.usc.edu; rohs@usc.edu.

Associate editor: Banu Ozkan

Abstract

Noncoding DNA sequences, which play various roles in gene expression and regulation, are under evolutionary pressure. Gene regulation requires specific protein–DNA binding events, and our previous studies showed that both DNA sequence and shape readout are employed by transcription factors (TFs) to achieve DNA binding specificity. By investigating the shape-disrupting properties of single nucleotide polymorphisms (SNPs) in human regulatory regions, we established a link between disruptive local DNA shape changes and loss of specific TF binding. Furthermore, we described cases where disease-associated SNPs may alter TF binding through DNA shape changes. This link led us to hypothesize that local DNA shape within and around TF binding sites is under selection pressure. To verify this hypothesis, we analyzed SNP data derived from 216 natural strains of *Drosophila melanogaster*. Comparing SNPs located in functional and nonfunctional regions within experimentally validated cis-regulatory modules (CRMs) from *D. melanogaster* that are active in the blastoderm stage of development, we found that SNPs within functional regions tended to cause smaller DNA shape variations. Furthermore, SNPs with higher minor allele frequency were more likely to result in smaller DNA shape variations. The same analysis based on a large number of SNPs in putative CRMs of the *D. melanogaster* genome derived from DNase I accessibility data confirmed these observations. Taken together, our results indicate that common SNPs in functional regions tend to maintain DNA shape, whereas shape-disrupting SNPs are more likely to be eliminated through purifying selection.

Key words: **single nucleotide polymorphism, SNP, noncoding region, DNA structure, protein–DNA recognition, transcription factor.**

Article

Introduction

Biomedical disease research has focused primarily on identifying pathogenic nonsynonymous variants in protein-coding regions of the genome because of the absence of functional annotation of noncoding variants (Ward and Kellis 2012). However, noncoding genetic variants can also be pathogenic, as shown by numerous genome-wide association studies (Welter et al. 2014). Noncoding variants may affect disease pathogenesis through various gene regulatory mechanisms, such as transcription factor (TF) binding, RNA splicing, and mRNA degradation (Faustino and Cooper 2003; Abelson et al. 2005; Maurano et al. 2012).

TFs regulate gene expression by binding to specific TF binding sites (TFBSs), which are traditionally described by position weight matrices (PWMs) representing the independent binding-affinity contributions of nucleotides at each position of the TFBS (Stormo 2000, 2013). Nucleotide mutations within and near TFBSs, or in TFBS-rich regions, can potentially disrupt TF binding and consequentially up- or down-regulate gene expression (Mogno et al. 2013). It is, therefore, not surprising that TFBS regions are under strong selection pressure (Andolfatto 2005). The three-dimensional structure of DNA, or "DNA shape", is an important determinant of specific TF binding (Rohs et al. 2009, 2010). TFs employ base readout (direct contacts between amino acids and functional groups of the base pairs) and shape readout (recognition of three-dimensional DNA structure) to achieve DNA binding specificity (Slattery et al. 2014). Adding DNA shape features to models of TF–DNA binding increases the prediction accuracy of TF binding (Abe et al. 2015; Zhou et al. 2015; Yang et al. 2017).

Here, by analyzing the connection between DNA shape and TF binding using single nucleotide polymorphisms

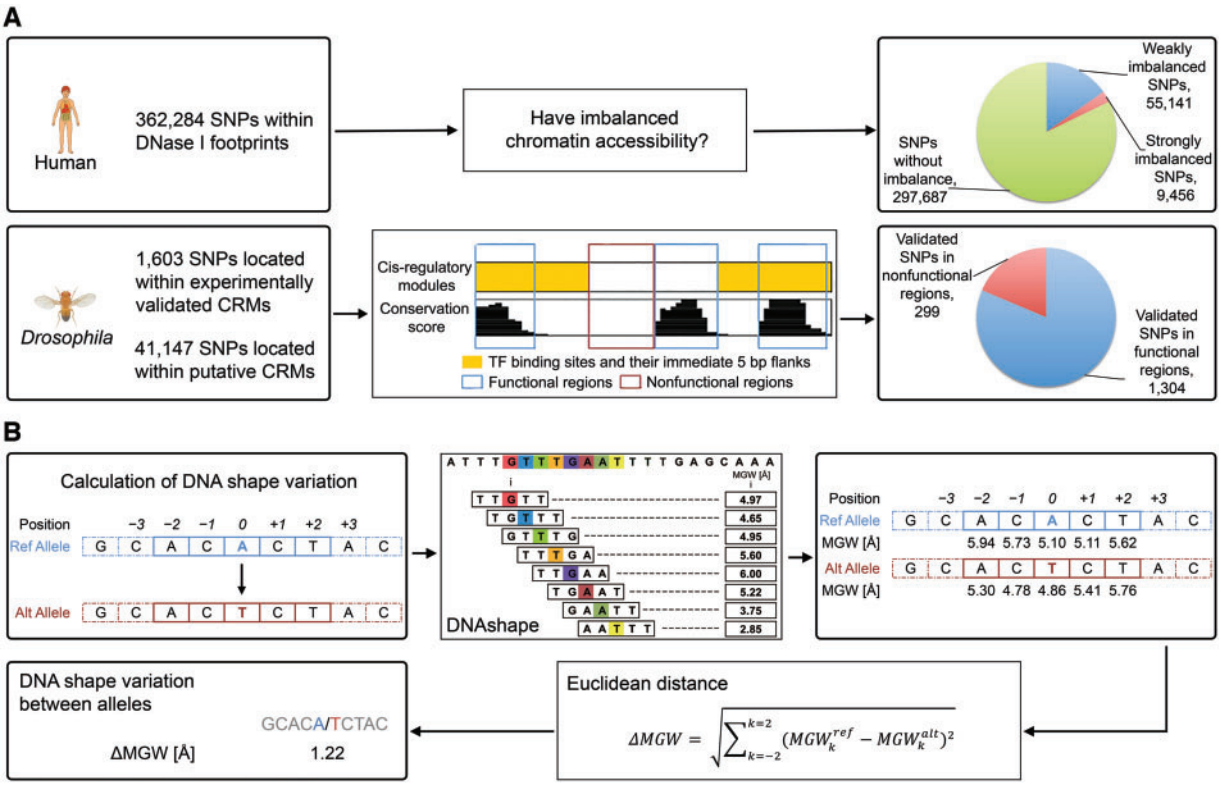

Fig. 1. Pipeline for evaluation of SNP effects on DNA shape. (*A*) Human SNPs derived from DNase-seq data were divided into three groups, 1) strongly imbalanced SNPs, 2) weakly imbalanced SNPs, and 3) SNPs without imbalance, according to their effect on DNA accessibility. *Drosophila* SNPs, called from 216 natural strains of *D. melanogaster* and located within blastoderm stage-active CRMs, were divided into two groups, a) SNPs in functional regions and b) SNPs in nonfunctional regions, based on the criteria shown in the center. The same analysis was repeated for a larger number of *Drosophila* SNPs in putative CRMs. (*B*) Example calculation of DNA shape variation for one SNP. One single-nucleotide variant at position 0 would result in DNA shape changes at the five nucleotide positions centered around the variant. First, vectors of MGW for each allele were predicted using DNAshape (Zhou et al. 2013). Euclidean distances between MGWs of the two alleles were calculated as the DNA shape variation of the SNP (see Materials and Methods).

(SNPs) called from human DNase-seq data, we found that SNP-induced allelic differences of TF binding can be partially explained by SNP-induced local DNA shape changes. This observation prompted us to hypothesize that DNA shape in TFBS regions is under purifying selection pressure. To test this hypothesis, we analyzed SNPs derived from 216 natural strains of *Drosophila melanogaster*. Comparing SNPs located in functional versus nonfunctional regions and SNPs with high versus low minor allele frequency (MAF), we statistically showed that DNA shape is under purifying selection.

Results

Effect of SNPs on DNA Shape Varies with Allele Type and Local DNA Context

To predict the effects of SNPs on local DNA shape, we used the DNAshape method (Zhou et al. 2013) to analyze human and *Drosophila* data sets (fig. 1*A*). We used the predicted minor groove width (MGW) as predominant DNA shape feature for evaluating the functional implications of DNA shape readout (fig. 1*B*) (Parker et al. 2009). Given that the predicted MGW at each nucleotide position is a continuous value, we defined the DNA shape change for each SNP, in the form of ΔMGW, as the Euclidean distance between vectors of MGW profiles of the two alleles (see Materials and Methods). We observed that the effect of SNPs on local DNA shape patterns varied with allele type and local DNA context. At one extreme of the spectrum were SNPs that had a weak effect on local shape patterns (fig. 2*A* and *B*). When a TF recognizes the DNA shape pattern of its binding site, this recognition is likely to be preserved for both the reference and alternative allele of those SNPs. At the other extreme were SNPs that completely disrupted the local DNA shape (fig. 2*C* and *D*). We suggest that loss of function is more likely to occur with those shape-disrupting SNPs. Between these two extremes were SNPs that caused modest variations in local DNA shape, which would likely still affect TF binding (fig. 2*E* and *F*). Subtle changes in MGW lead to changes in electrostatic potential, which will affect DNA shape readout (Chiu et al. 2017).

SNPs with Imbalanced DNA Accessibility Induced Larger DNA Shape Changes Compared with SNPs without Imbalance

To assess the effect of SNP-induced DNA shape changes on TF binding, we analyzed the correlation between these shape changes and DNA accessibility changes in an in vivo human data set. DNA accessibility is a hallmark of specific TF binding (Maurano et al. 2015); if alleles differ in their degree of DNase I

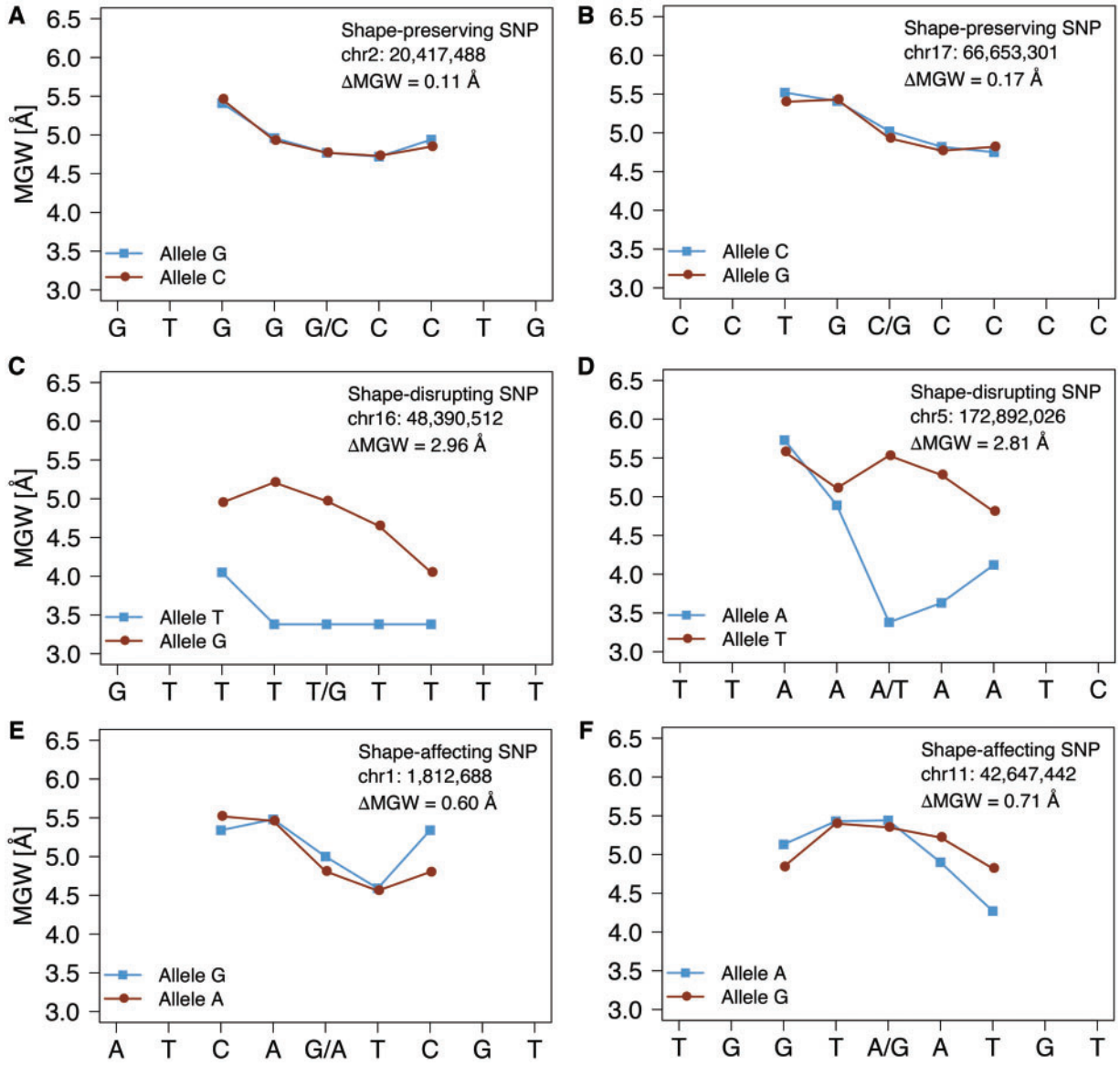

FIG. 2. Local effect of SNPs on MGW profiles. Effects of SNPs on their surrounding MGW patterns varied with allele type and local DNA context. MGW patterns of local DNA region for two alternate alleles of SNPs were plotted in blue and red, respectively. At one extreme of the spectrum were SNPs (*A* and *B*) that had very small effects on local MGW. At the other extreme were SNPs (*C* and *D*) that completely disrupted the local MGW geometry. Between these two extremes were SNPs that led to an intermediate extent of variation in local DNA shape, whereas potentially still affecting TF binding (*E* and *F*).

accessibility, then we can infer that they will differ in their degree of TF binding. By analyzing DNase-seq signals at heterozygous sites, Maurano et al. (2015) classified SNPs located in DNase I hypersensitive sites into three groups: 1) strongly imbalanced SNPs (0.1% false discovery rate [FDR] and >70% imbalance), 2) weakly imbalanced SNPs (5% FDR), and 3) SNPs without imbalance, as defined by (Maurano et al. 2015).

We calculated and plotted the distribution of ΔMGW values of SNPs for all three groups (fig. 3). Although the ΔMGW distributions for all three groups were similar in shape, the distributions for imbalanced SNPs were significantly shifted towards larger ΔMGW values (without imbalance vs. strongly imbalanced: Mann–Whitney $P = 1.44 \times 10^{-12}$, shuffling test $P = 0.023$ using "bogus" MGW predictions; without imbalance vs. weakly imbalanced: Mann–Whitney $P = 1.40 \times 10^{-3}$, shuffling test $P = 0.201$; strongly imbalanced vs. weakly imbalanced: Mann–Whitney $P = 4.10 \times 10^{-8}$, shuffling test $P = 0.011$; see Materials and Methods for shuffling test; supplementary fig. S1A, Supplementary Material online). As the group of SNPs became increasingly imbalanced, the distribution shifted towards larger ΔMGW values. This observation revealed an association between imbalanced SNPs and increased ΔMGW, and indicated that drastic DNA shape changes could lead to loss of specific TF binding.

Disease-Associated SNPs Potentially Alter TF Binding through Shape Changes

We have illustrated that alternate alleles of SNPs could result in different DNA shape patterns, thereby potentially

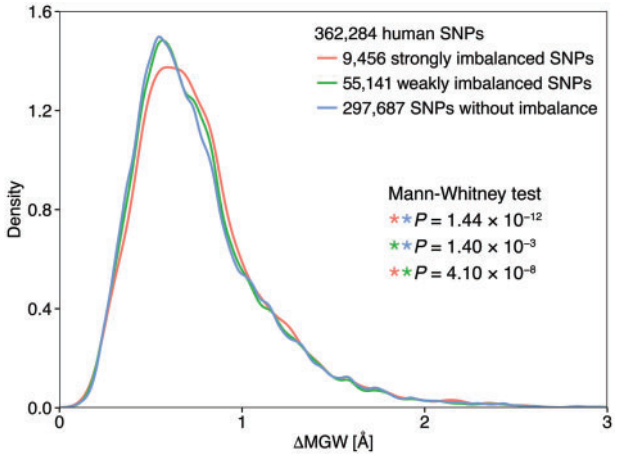

Fig. 3. Distribution of MGW changes for strongly imbalanced SNPs, weakly imbalanced SNPs, and SNPs without imbalance in human. Distributions of ΔMGW values for imbalanced SNPs (red and green plots) were shifted rightward compared with SNPs without imbalance (blue plot). The more imbalanced the SNPs were, the larger the ΔMGW or change in DNA shape was. Asterisks are color-coded to indicate the SNP distributions being compared. Sample sizes for all groups are listed in the legend.

influencing TF binding through effects on shape readout and possibly leading to disease due to altered expression of the target genes (Maurano et al. 2012; Mogno et al. 2013). Thus, the pathogenesis of some diseases may be attributed to the loss of preferred DNA shape at TFBSs. However, loss of the preferred DNA shape at TFBSs is not the only possible pathogenic mechanism for disease-associated SNPs. Therefore, we did not attempt to correlate disease with DNA shape changes generally, but instead analyzed the effects on DNA shape by SNPs that were shown in the literature to affect TF binding. Here we highlight a few examples in which shape-disrupting SNPs could potentially cause disease.

For example, the T allele of SNP rs339331 increases the DNA binding affinity of HOXB13, leading to overexpression of Regulatory Factor X 6 (RFX6), which promotes prostate cancer cell growth and invasion (Huang et al. 2014). Examining the effect of the allele on local DNA shape (fig. 4A), we found that the risk allele T induced a narrower minor groove in the core-binding site TTTTAT. The ΔMGW of approximately 1.3 Å is at the right tail of the distribution in figure 3, indicating a large ΔMGW for this SNP. This finding is consistent with our previous studies showing that MGW plays a role in achieving homeodomain binding specificity (Slattery et al. 2011; Abe et al. 2015; Dror et al. 2015).

Another SNP, rs6893009, located in a PU.1 binding site, is a strong binding quantitative trait locus as measured by ChIP-seq experiments (Tehranchi et al. 2016). These ChIP-seq measurements found that the SNP was in perfect linkage disequilibrium with a Crohn's disease-associated SNP, rs4958847 (Parkes et al. 2007). Interestingly, we previously showed that DNA shape features help to guide binding of PU.1 (Barozzi et al. 2014). Investigation of how rs6803009 affects local DNA shape showed that the SNP caused a large change in MGW of approximately 1.3 Å (fig. 4B), consistent with our previous finding that MGW was a predominant structural determinant of PU.1 binding.

Our analysis also revealed medium-to-strong MGW changes caused by SNPs located in the c-MYB, GATA3, ER-α, and TCF7L2 binding sites (fig. 4C–F), all of which are disease-associated SNPs with evidence of disrupted TF binding (Jin et al. 2010; Miyoshi et al. 2010; Alipanahi et al. 2015; Mathelier et al. 2015).

DNA Shape in Functional Regulatory Regions of the *Drosophila* Genome Is More Conserved

The above results suggested a possible link between DNA shape changes and loss of TF–DNA binding, leading us to hypothesize that DNA shape in TFBS regions is under purifying selection. To test this hypothesis, we analyzed SNP data derived from 216 natural strains of *D. melanogaster* to uncover the signal of DNA shape selection. We used these data because of the large numbers of sequenced individuals, annotated cis-regulatory modules (CRMs), and available PWMs for TFs in the *D. melanogaster* genome.

First, we identified SNPs within experimentally verified CRMs in the *Drosophila* genome that regulate gene expression during the blastoderm stage, as annotated in the REDfly database (see Materials and Methods) (Gallo et al. 2011). We focused on these CRMs because we were able to identify a high-confidence set of functioning TFs at the blastoderm stage, whereas such a set of TFs was unavailable for other developmental stages. CRMs are composed of TFBSs and intervening sequences. We assumed that within CRMs, TFBSs and their flanking regions as well as nucleotide positions with high conservation scores are functionally important. This assumption enabled us to divide CRMs into two distinct DNA fragment sets, that is, functional and nonfunctional regions (see Materials and Methods).

Next, we conducted comparative studies on SNPs in those two regions. We plotted the distributions of ΔMGW values for SNPs in functional and nonfunctional regions within *Drosophila* CRMs that are active in the blastoderm developmental stage (fig. 5A). Compared with the ΔMGW distribution for functional regions (red), the distribution for nonfunctional regions (blue) was significantly shifted towards larger ΔMGW values (Mann–Whitney $P = 2.62 \times 10^{-4}$; shuffling test $P = 0.001$; supplementary fig. S1B, Supplementary Material online). When we compared the distributions of shuffled ΔMGW values using "bogus" MGW predictions, the shift was not significant (see Materials and Methods; fig. 5B shows one shuffling). This result indicated that SNPs in functional regions were less likely to induce drastic DNA shape changes, implying that SNPs that greatly change MGW in these regions were removed by purifying selection. Results of multi-allelic analysis confirmed this notion that local DNA shape was more conserved in functional than in nonfunctional regions (see Materials and Methods; supplementary fig. S2, Supplementary Material online).

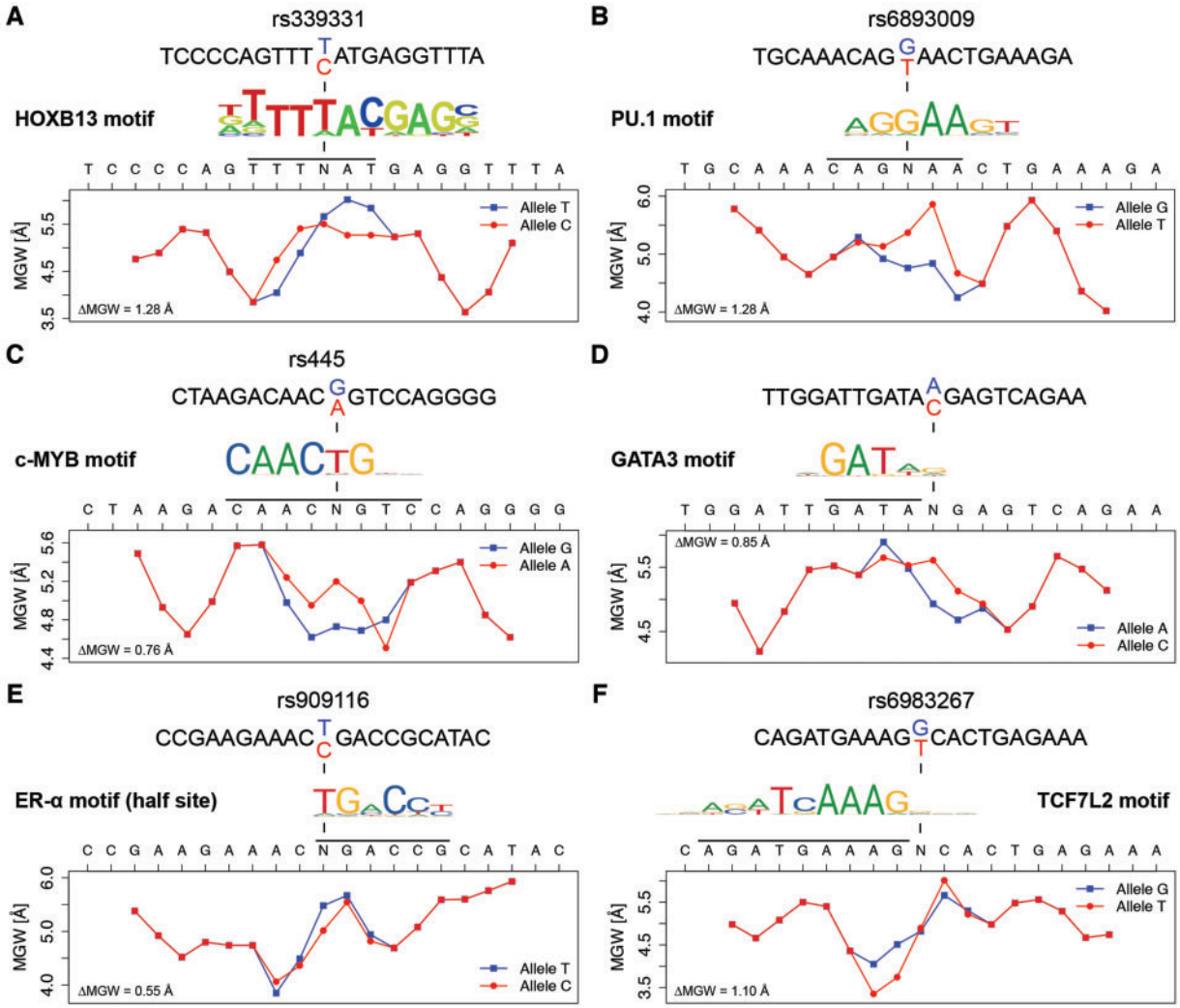

FIG. 4. DNA shape variation caused by disease-associated SNPs. (*A*) DNA shape variation caused by SNP rs339331 in the HOXB13 binding site. HOXB13 prefers binding to a narrower MGW induced by risk allele T. (*B*) DNA shape variation caused by SNP rs6893009 in the PU.1 binding site. The SNP caused large variance in MGW, which was previously reported to be a predominant structural determinant of PU.1 binding. DNA shape variation caused by (*C*) SNP rs445 in the c-MYB binding site, (*D*) a SNP in the GATA3 binding site, (*E*) SNP rs909116 in the ER-α binding site, and (*F*) SNP rs6983267 in the TCF7L2 binding site.

SNPs with Higher Minor Allele Frequency Tend to Relate to Smaller DNA Shape Change

We further classified SNPs in functional regions within CRMs into high- and low-frequency groups based on their MAFs. To create similarly sized groups, we used a frequency cutoff of 0.04. Plotting the ΔMGW distributions for SNPs in these two groups, we observed that SNPs with higher MAFs tended to have smaller variations of MGW (fig. 5*C*; Mann–Whitney $P = 3.97 \times 10^{-2}$; shuffling test $P = 0.045$; supplementary fig. S1*B*, Supplementary Material online). Although this observation implies that SNPs that greatly change DNA shape are less likely to have large MAFs, we caution that the biological effect is likely small for most SNPs. To ensure that the observed negative correlation between MAF and ΔMGW values was meaningful, we used SNPs in nonfunctional regions as a negative control group. Using this negative control, we compared distributions of ΔMGW values for the high- and low-frequency groups as we did for SNPs in functional regions. In this case, no negative correlation between MAF and ΔMGW could be found (fig. 5*D*; Mann–Whitney $P = 0.822$; shuffling test $P = 0.819$; supplementary fig. S1*B*, Supplementary Material online). These results further support our hypothesis that purifying selection acts on DNA shape near TFBSs.

Large Data Set of SNPs in Putative CRMs Confirmed the Observations for SNPs in Validated CRMs

We repeated the same analysis on a larger data set of SNPs in putative CRMs of the *Drosophila* genome defined based on DNase I accessibility (see Materials and Methods). In comparison to the ΔMGW distribution for functional regions, the distribution for nonfunctional regions was significantly shifted towards larger ΔMGW values (Mann–Whitney $P = 5.31 \times 10^{-29}$; fig. 6*A*). The shift between distributions of shuffled ΔMGW values that were calculated from "bogus" MGW predictions was not significant (fig. 6*B*). This analysis

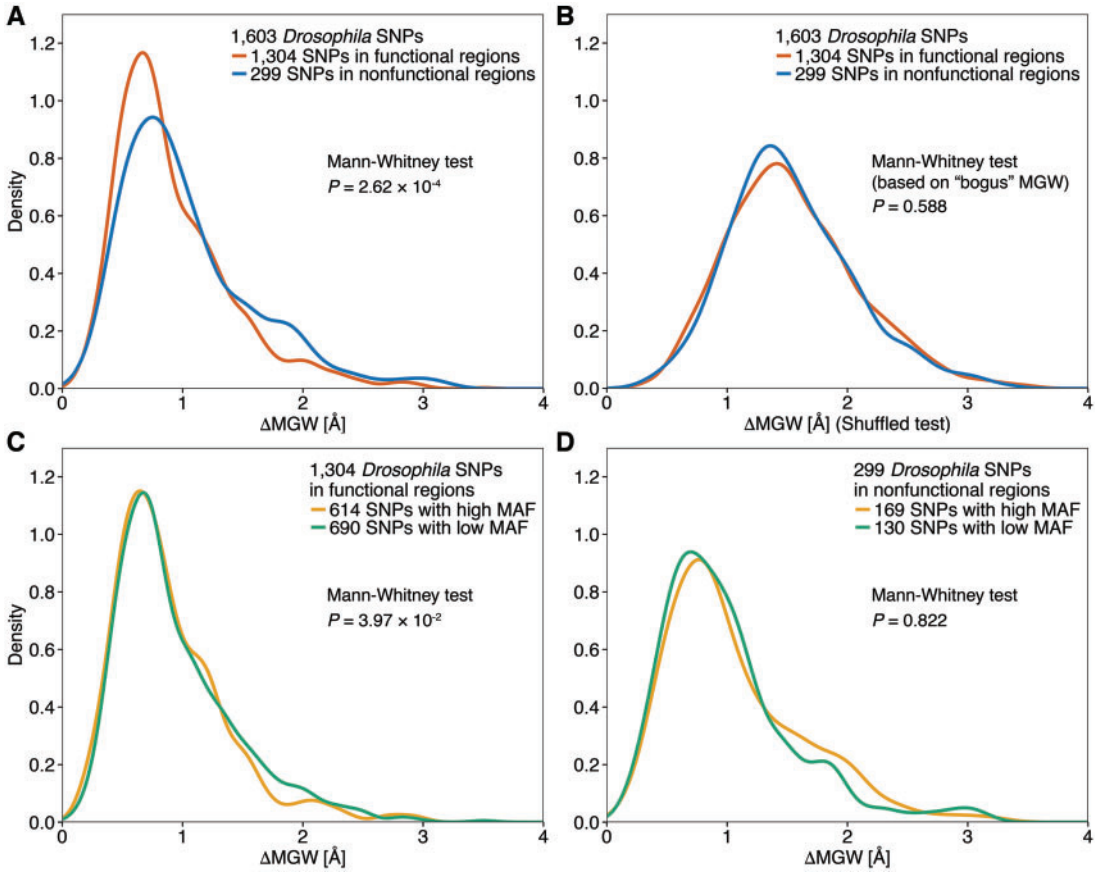

FIG. 5. Distributions of MGW changes for *Drosophila* SNPs in experimentally validated CRMs at different locations and with different MAFs. (*A*) Distribution of ΔMGW values for SNPs in functional and nonfunctional regions (see Materials and Methods for definition) using the DNAshape-derived MGW. Compared with the distribution for functional regions (red plot), the distribution for nonfunctional regions (blue plot) was significantly shifted rightward, indicating that SNPs induced greater changes in ΔMGW in nonfunctional than in functional regions. (*B*) Distribution of ΔMGW values for SNPs in functional and nonfunctional regions, using one of the shuffled MGW predictions. Using arbitrarily shuffled MGW, no signal of purifying selection emerged. (*C*) Distribution of ΔMGW values for SNPs with high and low MAF in functional regions. Distribution of ΔMGW values for low MAF was significantly shifted towards the right. (*D*) Distribution of ΔMGW values for SNPs with high and low MAF in nonfunctional regions. Distributions of these two groups exhibited no significant difference. Sample sizes for all groups are listed in the legends.

also confirmed the negative correlation between MAF and ΔMGW (Mann–Whitney $P = 6.27 \times 10^{-15}$; fig. 6*C*). Using SNPs in nonfunctional regions as negative control, no negative correlation between MAF and ΔMGW could be found (fig. 6*D*). Thus, this complementary analysis using a large data set of SNPs in putative CRMs reaffirmed our observations for SNPs in experimentally validated CRMs.

Discussion

Comparative studies have shown that at least 3–8% of all nucleotides in the human genome are under purifying selection, and many of these nucleotides are located in noncoding regions of the genome (ENCODE Project Consortium 2012). The conservation level of a nucleotide position reflects its functional importance. However, nucleotide sequence may not be the sole target of selective pressure (Parker et al. 2009). In this study, our findings suggest that purifying selection also acts on DNA shape. Due to the methodology currently available to probe structural features of the genome (Zhou et al. 2013; Li et al. 2017), only local DNA shape could be analyzed in this study. We, therefore, did not address global genomic topological information, such as data obtained from chromosome conformation capture-based experiments (Dekker 2016; Dekker and Mirny 2016), which is also likely to be under evolutionary selection (Kovina et al. 2017).

In DNase-seq data, allelic imbalance of a SNP reflects a change in protein binding at the locus. We showed that allelicly imbalanced SNPs in human DNase I hypersensitive sites tended to induce slightly larger changes in DNA shape than SNPs without allelic imbalance in DNA accessibility in vivo. This shift in DNA shape change suggested a role for DNA shape in determining TF binding specificity and sheds light on the mechanism of DNA shape-targeted purifying selection. Although statistically significant, the magnitude of the shift of shape variation was relatively small. One possible explanation for this result is that, although many TFs employ DNA shape readout in addition to base readout, not all TFs use shape readout to the same extent (Yang et al. 2017). By aggregating data from regions where many different TFs bind, signals from TFs with different levels of dependence on DNA

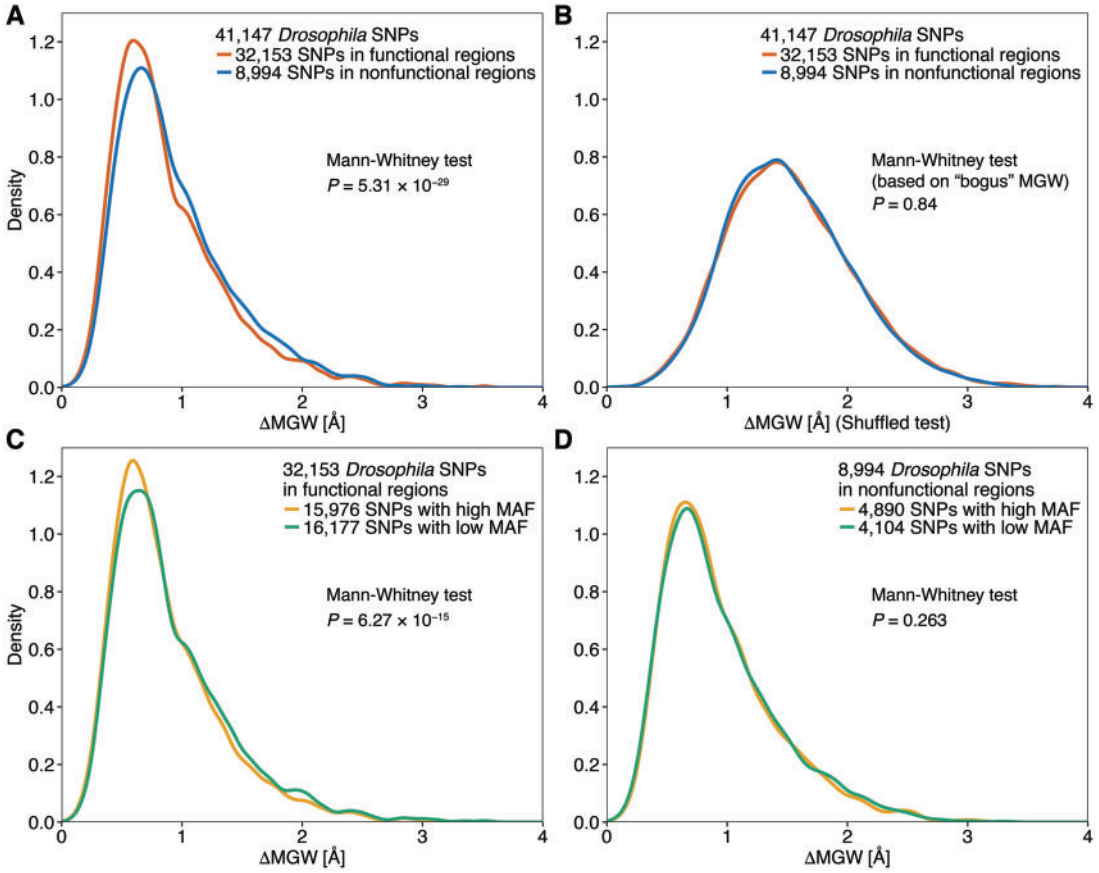

Fig. 6. Distributions of MGW changes for *Drosophila* SNPs in putative CRMs at different locations and with different MAFs. (*A*) Distribution of ΔMGW values for SNPs in functional and nonfunctional regions (see Materials and Methods for definition) using the DNAshape-derived MGW. Compared with the distribution for functional regions (red plot), the distribution for nonfunctional regions (blue plot) was significantly shifted rightward, indicating that SNPs induced greater changes in ΔMGW in nonfunctional than in functional regions. (*B*) Distribution of ΔMGW values for SNPs in functional and nonfunctional regions, using one of the shuffled MGW predictions. Using arbitrarily shuffled MGW, no signal of purifying selection emerged. (*C*) Distribution of ΔMGW values for SNPs with high and low MAF in functional regions. Distribution of ΔMGW values for low MAF was significantly shifted towards the right. (*D*) Distribution of ΔMGW values for SNPs with high and low MAF in nonfunctional regions. Distributions of these two groups exhibited no significant difference. Sample sizes for all groups are listed in the legends.

shape for binding specificity are aggregated together, which likely dilutes the overall signal.

The effect of DNA shape on TF binding is also supported by the work of other researchers. In a study of the effect of SNPs on TF binding, allele-specific binding events strongly correlated with TFBS alterations (Shi et al. 2016). Within each TF binding motif, certain positions were more sensitive to allele-specific binding events. For example, the binding of the CCAAT/enhancer-binding protein (CEBPB), which binds to an 11-base pair (bp) DNA motif as a dimer, was particularly sensitive to position 6 at the center of the motif (Shi et al. 2016). Interestingly, in separate work that used machine learning to study the position-dependent DNA shape importance in TF binding (Yang et al. 2017), we observed that one side of the CEBPB half-site exhibited stronger shape importance, coinciding with the central position identified by (Shi et al. 2016). Taken together, these findings suggest a mechanism by which DNA shape affects TF binding as a prerequisite for natural selection. Alterations in TF binding due to shape-disrupting SNPs could also cause disease, as illustrated above with examples of disease-associated SNPs.

Furthermore, we found statistical evidence that purifying selection acts on DNA shape features, based on SNP data derived from 216 natural strains of *D. melanogaster*. In functional TF binding regions, DNA tended to maintain its shape. We hypothesized that, if DNA shape is functional, then the more common a SNP is, the more likely it conserves the local DNA shape, as shape-disrupting SNPs would have been removed through purifying selection due to their deleteriousness. Our statistical analysis confirmed this hypothesis. Specifically, we found that in functional TF binding regions, SNPs with larger MAFs were more likely to result in smaller shape variations, whereas such a correlation was not present in nonfunctional regions. Although the difference in shape changes between high and low MAFs was small, it was statistically significant for SNPs in experimentally validated CRMs, with the significance level being much higher for a larger data set of SNPs in putative CRMs.

In summary, we propose that selection acts on TF binding partially through maintenance of DNA shape. This understanding adds a new perspective to the practical study of genome evolution. The accuracy of regulatory SNP prediction

has recently increased dramatically due to more efficient machine-learning methods (Alipanahi et al. 2015; Lee et al. 2015). As adding DNA shape information has been demonstrated to improve the modeling of TF–DNA binding specificities (Zhou et al. 2015; Mathelier et al. 2016; Yang et al. 2017), we envision that adding information of DNA shape variation can further improve the prediction accuracy in the identification of regulatory or pathogenic noncoding variants.

Materials and Methods

Single Nucleotide Polymorphism Data

Human

A total of 362,284 common variants called from 493 high-resolution DNase-seq profiles were considered in this study (Maurano et al. 2015). Among these variants, 64,597 SNPs were identified as allelicly imbalanced in TF occupancy and DNA accessibility in vivo, based on the allelicly imbalanced read counts in the DNase-seq data. Among these imbalanced SNPs, 55,141 were identified as weakly imbalanced (5% FDR), and 9,456 were strongly imbalanced (0.1% FDR and >70% imbalance; fig. 1A).

Drosophila

Genomic data for 216 natural strains of *D. melanogaster* from two resequencing projects (Mackay et al. 2012; Campo et al. 2013) were downloaded from NCBI Sequence Read Archive (accession numbers PRJNA36679 and PRJNA74721). For this vast amount of data, SNP calling as described in (Campo et al. 2013) yielded 2,605,315 SNPs that occurred at least twice among the 216 individuals. The MAF of each SNP was calculated from these data. Here, we only investigated SNPs located within blastoderm stage-active CRMs, representing regions in which TFs bind during the blastoderm stage. We obtained the set of CRMs annotated as "blastoderm embryo" in the REDfly database (Gallo et al. 2011). These CRMs have been experimentally validated to regulate gene expression during *Drosophila* embryonic development (Gallo et al. 2011). We filtered this list to eliminate regions <100 bp long, as described in (Su et al. 2010), which resulted in 342 regions. By requiring an overlap with these CRMs, we narrowed the number of SNPs down to 1,603 (fig. 1A).

Alternatively, with a less stringent criterion, we were able to identify a larger number of putative CRMs based on DNase I accessible regions found during *Drosophila* embryonic development (Thomas et al. 2011). We selected regions that were accessible during stage 5 and "developmentally dynamic", that is differentially accessible at reported developmental time points (Thomas et al. 2011). We further filtered this list to eliminate regions <100 bp long (Su et al. 2010), resulting in 4,699 regions. Requiring an overlap with these putative CRMs yielded 41,147 SNPs (fig. 1A). Although this definition of putative CRMs likely included false positives, it allowed testing of our hypothesis based on a large number of SNPs (fig. 6).

Definition of Functional and Nonfunctional Regions in *Drosophila*

Functional genomic regions are more likely than nonfunctional regions to be under selection (ENCODE Project Consortium 2012). PhastCons (Siepel et al. 2005) is a widely used approach to identify evolutionarily conserved elements based on multiple sequence alignment and a phylogenetic tree. This approach produces continuous values for conservation scores for each nucleotide position of the genome. The higher the score is, the more conserved the respective nucleotide position is.

We defined the functional and nonfunctional regions within CRMs based on the following criteria (fig. 1A):

(1) A nucleotide position with phastCons conservation score >0.1 was considered to be in a functional region.
(2) A nucleotide position that had a conservation score ≤0.1 and was not located within any of the identified TFBSs or their immediate 5-bp flanking regions was considered to be in a nonfunctional region.

We excluded TFBSs with low conservation scores from nonfunctional regions to rule out the possibility of underestimation of conservation levels calculated by the sequence-based method. PhastCons conservation scores for alignments of genomes of 14 insects from *D. melanogaster* (dm3 assembly) at each nucleotide position were downloaded from the UCSC Genome Browser (Tyner et al. 2017).

TFBSs in CRMs were located through motif scans. To search for motif matches, we used PWMs of the 34 principal TFs (supplementary table S1, Supplementary Material online) that are active during the blastoderm stage of *Drosophila* development, as defined by FlyFactorSurvey (Zhu et al. 2011) and the PATSER program (Stormo et al. 1982). We used a GC content of 0.406, corresponding to the intergenic GC content of *D. melanogaster* (Berman et al. 2002), and a *P*-value cutoff of 0.001.

Genome-Wide Prediction of DNA Shape

The MGW values used in this study were predicted with DNAshape, a high-throughput method for predicting DNA structural features (Chiu et al. 2016; Zhou et al. 2013). DNAshape uses a precomputed pentamer query table that stores the DNA shape information for all 512 unique pentamers (fig. 1B). We focused in this study on MGW due to its well-established role in DNA shape readout and because MGW, unless other dinucleotide-based features such as helix twist, propeller twist, or roll, is defined over a region of several nucleotides.

Calculation of Euclidean Distance of Minor Groove Width between Two Alleles

Pentamer-based MGW prediction with the DNAshape approach was shown to agree well with experimental structural data (Zhou et al. 2013). Based on this modeling assumption, one single-nucleotide variant would result in DNA shape changes of the five consecutive nucleotide positions centered around it. Thus, prediction of MGW at these five positions relies on the 9-mer sequence context (fig. 1B). For any SNP, we

denoted the 9-mer sequence context as $S_{-4}S_{-3}S_{-2}S_{-1}S_0S_1S_2S_3S_4$, where S_0 is the SNP locus. We defined the DNA shape for each allele i of this SNP as $\vec{MGW}^i$, where

$$\vec{MGW}^i = \left(MGW^i_{-2},\ MGW^i_{-1},\ MGW^i_0,\ MGW^i_1,\ MGW^i_2\right)$$

We regarded the most frequently occurring allele as the reference allele and all other alleles as alternative alleles.

To quantify the variation of local MGW between two alternate alleles of a SNP, we calculated the Euclidean distance ΔMGW between the reference allele (*ref*) and alternative allele (*alt*), as follows:

$$\Delta MGW = \sqrt{\sum_{k=-2}^{k=2}\left(MGW^{ref}_k - MGW^{alt}_k\right)^2}$$

where k indicates the relative position to the SNP (fig. 1*B*). Quantification of the variation in local MGW can be expanded to multi-allelic SNPs by averaging ΔMGW values between the reference allele and each alternative allele. For example, consider a SNP having $n > 2$ alleles (one reference allele and $n-1$ alternative alleles). For each alternative allele alt_m, a ΔMGW_m value can be calculated using the above formula. The MGW change of the multi-allelic SNP can be defined as follows:

$$\Delta\vec{MGW} = \frac{1}{n-1}\sum_{m=1}^{n-1}\Delta MGW_m$$

Statistical Analysis

The Mann–Whitney *U* test, also called the Wilcoxon rank sum test, is a nonparametric statistical test that is used to analyze differences between the medians of two data sets. We applied the Mann–Whitney *U* test to evaluate the difference in ΔMGW distributions between SNPs with and without imbalance, between SNPs in functional and nonfunctional regions, and between SNPs with high and low MAFs. The *P*-value of each test was calculated.

Shuffling of the Pentamer Query Table and Statistical Verification

The DNAshape method predicts MGW based on a pentamer query table that is derived from data mining of all-atom Monte Carlo simulations (Zhou et al. 2013). To rule out the possibility that the statistical significance of detected purifying selection signals is an artifact of associating each pentamer with a floating number (i.e., MGW), we generated 1,000 shuffled pentamer tables. For each shuffled table, we derived the "bogus" ΔMGW values for each SNP (shuffled ΔMGWs) and computed a new *P*-value for every statistical test that we conducted.

We calculated the shuffling test *P*-value as the ratio between the number of new *P*-values that were lower than the originally reported *P*-value and the number of shuffles (1,000). In other words, the shuffling test *P*-value was the probability that the distribution difference can be observed by randomly associating a pentamer with a floating number. Statistical verification using the shuffled pentamer tables indicated that the observed purifying selection indeed acted on DNA shape rather than DNA sequence.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors thank all members of the Rohs laboratory for valuable input. This work was supported by the National Institutes of Health (U01GM103804 to A.H.D., S.V.N., and R.R.; R01GM106056 to R.R.; K99/R00HD073191 to Z.W.), two Andrew Viterbi Fellowships (to X.W. and T.Z.), and an Alfred P. Sloan Research Fellowship (to R.R.).

Author Contributions

X.W. and T.Z. designed and conducted the data analysis. T.Z. and R.R. conceived the study with contributions from A.H.D. and S.V.N. Z.W. and A.H.D. determined cis-regulatory regions and TFBSs. M.M. contributed human SNP data. S.V.N. contributed *Drosophila* genes and SNP mapping. X.W., T.Z., Z.W., and R.R. wrote the manuscript with help from all authors. T.Z. and R.R. directed this study.

References

Abe N, Dror I, Yang L, Slattery M, Zhou T, Bussemaker HJ, Rohs R, Mann RS. 2015. Deconvolving the recognition of DNA shape from sequence. *Cell* 161(2): 307–318.

Abelson JF, Kwan KY, O'Roak BJ, Baek DY, Stillman AA, Morgan TM, Mathews CA, Pauls DL, Rasin MR, Gunel M, et al. 2005. Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science* 310(5746): 317–320.

Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol.* 33(8): 831–838.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. *Nature* 437(7062): 1149–1152.

Barozzi I, Simonatto M, Bonifacio S, Yang L, Rohs R, Ghisletti S, Natoli G. 2014. Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. *Mol Cell* 54(5): 844–857.

Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB. 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci U S A.* 99(2): 757–762.

Campo D, Lehmann K, Fjeldsted C, Souaiaia T, Kao J, Nuzhdin SV. 2013. Whole-genome sequencing of two North American *Drosophila melanogaster* populations reveals genetic differentiation and positive selection. *Mol Ecol.* 22(20): 5084–5097.

Chiu TP, Comoglio F, Zhou T, Yang L, Paro R, Rohs R. 2016. DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics* 32(8): 1211–1213.

Chiu TP, Rao S, Mann RS, Honig B, Rohs R. 2017. Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein-DNA binding. *Nucleic Acids Res.* 45(21): 12565–12576.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414): 57–74.

Dekker J. 2016. Mapping the 3D genome: aiming for consilience. *Nat Rev Mol Cell Biol.* 17(12): 741–742.

Dekker J, Mirny L. 2016. The 3D genome as moderator of chromosomal communication. *Cell* 164(6): 1110–1121.
Dror I, Golan T, Levy C, Rohs R, Mandel-Gutfreund Y. 2015. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res.* 25(9): 1268–1280.
Faustino NA, Cooper TA. 2003. Pre-mRNA splicing and human disease. *Genes Dev.* 17(4): 419–437.
Gallo SM, Gerrard DT, Miner D, Simich M, Des Soye B, Bergman CM, Halfon MS. 2011. REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. *Nucleic Acids Res.* 39(Database): D118–D123.
Huang Q, Whitington T, Gao P, Lindberg JF, Yang Y, Sun J, Vaisanen MR, Szulkin R, Annala M, Yan J, et al. 2014. A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding. *Nat Genet.* 46(2): 126–135.
Jin SH, Zhao HW, Yi Y, Nakata Y, Kalota A, Gewirtz AM. 2010. c-Myb binds MLL through menin in human leukemia cells and is an important driver of MLL-associated leukemogenesis. *J Clin Invest.* 120(2): 593–606.
Kovina AP, Petrova NV, Gushchanskaya ES, Dolgushin KV, Gerasimov ES, Galitsyna AA, Penin AA, Flyamer IM, Ioudinkova ES, Gavrilov AA, et al. 2017. Evolution of the genome 3D organization: comparison of fused and segregated globin gene clusters. *Mol Biol Evol.* 34(6): 1492–1504.
Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, Beer MA. 2015. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet.* 47(8): 955–961.
Li J, Sagendorf JM, Chiu TP, Pasi M, Perez A, Rohs R. 2017. Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. *Nucleic Acids Res.* 45(22): 12877–12887.
Mackay TF, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The *Drosophila melanogaster* genetic reference panel. *Nature* 482(7384): 173–178.
Mathelier A, Lefebvre C, Zhang AW, Arenillas DJ, Ding J, Wasserman WW, Shah SP. 2015. Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. *Genome Biol.* 16:84.
Mathelier A, Xin B, Chiu TP, Yang L, Rohs R, Wasserman WW. 2016. DNA shape features improve transcription factor binding site predictions in vivo. *Cell Syst.* 3(3): 278–286.
Maurano MT, Haugen E, Sandstrom R, Vierstra J, Shafer A, Kaul R, Stamatoyannopoulos JA. 2015. Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat Genet.* 47(12): 1393–1401.
Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337(6099): 1190–1195.
Miyoshi Y, Murase K, Saito M, Imamura M, Oh K. 2010. Mechanisms of estrogen receptor-alpha upregulation in breast cancers. *Med Mol Morphol.* 43(4): 193–196.
Mogno I, Kwasnieski JC, Cohen BA. 2013. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res.* 23(11): 1908–1915.
Parker SC, Hansen L, Abaan HO, Tullius TD, Margulies EH. 2009. Local DNA topography correlates with functional noncoding regions of the human genome. *Science* 324(5925): 389–392.
Parkes M, Barrett JC, Prescott NJ, Tremelling M, Anderson CA, Fisher SA, Roberts RG, Nimmo ER, Cummings FR, Soars D, et al. 2007. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat Genet.* 39(7): 830–832.
Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. 2010. Origins of specificity in protein-DNA recognition. *Annu Rev Biochem.* 79:233–269.
Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B. 2009. The role of DNA shape in protein–DNA recognition. *Nature* 461(7268): 1248–1253.
Shi W, Fornes O, Mathelier A, Wasserman WW. 2016. Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic Acids Res.* 44(21): 10106–10116.
Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15(8): 1034–1050.
Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, Zhou T, Rohs R, Honig B, Bussemaker HJ, et al. 2011. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* 147(6): 1270–1282.
Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordan R, Rohs R. 2014. Absence of a simple code: how transcription factors read the genome. *Trends Biochem Sci.* 39(9): 381–399.
Stormo GD. 2000. DNA binding sites: representation and discovery. *Bioinformatics* 16(1): 16–23.
Stormo GD. 2013. Modeling the specificity of protein-DNA interactions. *Quant Biol.* 1(2): 115–130.
Stormo GD, Schneider TD, Gold L, Ehrenfeucht A. 1982. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res.* 10(9): 2997–3011.
Su J, Teichmann SA, Down TA. 2010. Assessing computational methods of cis-regulatory module prediction. *PLoS Comput Biol.* 6(12): e1001020.
Tehranchi AK, Myrthil M, Martin T, Hie BL, Golan D, Fraser HB. 2016. Pooled ChIP-Seq links variation in transcription factor binding to complex disease risk. *Cell* 165(3): 730–741.
Thomas S, Li XY, Sabo PJ, Sandstrom R, Thurman RE, Canfield TK, Giste E, Fisher W, Hammonds A, Celniker SE, et al. 2011. Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. *Genome Biol.* 12(5): R43.
Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, Fischer CM, Gibson D, Gonzalez JN, Guruvadoo L, et al. 2017. The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.* 45(D1): D626–D634.
Ward LD, Kellis M. 2012. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol.* 30(11): 1095–1106.
Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42(D1): D1001–D1006.
Yang L, Orenstein Y, Jolma A, Yin Y, Taipale J, Shamir R, Rohs R. 2017. Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol Syst Biol.* 13(2): 910.
Zhou T, Shen N, Yang L, Abe N, Horton J, Mann RS, Bussemaker HJ, Gordan R, Rohs R. 2015. Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc Natl Acad Sci U S A.* 112(15): 4654–4659.
Zhou T, Yang L, Lu Y, Dror I, Dantas Machado AC, Ghane T, Di Felice R, Rohs R. 2013. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res.* 41(W1): W56–W62.
Zhu LJ, Christensen RG, Kazemian M, Hull CJ, Enuameh MS, Basciotta MD, Brasefield JA, Zhu C, Asriyan Y, Lapointe DS, et al. 2011. FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res.* 39(Suppl 1): D111–D117.