Local DNA Topography Correlates with Functional Noncoding Regions of the Human Genome

Stephen C.J. Parker[1], **Loren Hansen**[1,2], **Hatice Ozel Abaan**[3], **Thomas D. Tullius**[1,4], and **Elliott H. Margulies**[3]

[1]Bioinformatics Program, Boston University, Boston, MA 02215, USA

[2]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, USA

[3]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

[4]Department of Chemistry, Boston University, Boston, MA 02215, USA

Abstract

The three-dimensional molecular structure of DNA, specifically the shape of the backbone and grooves of genomic DNA, can be dramatically affected by nucleotide changes, which can cause differences in protein binding affinity and phenotype. We developed an algorithm to measure constraint on the basis of similarity of DNA topography among multiple species using hydroxyl radical cleavage patterns to interrogate the solvent accessible surface area of DNA. This algorithm found that 12% of bases in the human genome are evolutionarily constrained—double the number detected by nucleotide sequence-based algorithms. Topography-informed constrained regions correlated with functional non-coding elements, including enhancers, better than regions identified solely on the basis of nucleotide sequence. These results support the idea that the molecular shape of DNA is under selection and can identify evolutionary history.

Genomic sequences that code for proteins are relatively well understood, but comprise only ~2% of the human genome (1). Many functions are encoded in the remaining ~98% non-coding portion of the genome, but little is known about how functional non-coding information is specified (2). It has been hypothesized that functional regions are likely to be evolutionarily constrained because of their importance to the organism (3–11). Nonetheless, evolutionary sequence-constraint algorithms fail to identify many non-coding functional elements (12–15). This may be because these methods only analyze the primary nucleotide sequence of a genome (i.e., the order of A's, C's, G's, and T's). However, DNA is a molecule with a three-dimensional structure that varies according to the nucleotide sequence (16–18).

We used a previously developed method based on the hydroxyl radical cleavage pattern of DNA (19) to quantitatively evaluate how the structure of DNA varies throughout a genome. This approach can be used to predict the shape of the DNA backbone and grooves of genomic DNA at single-nucleotide resolution (20). We call this pattern the structural profile of a DNA region. Structural profiles reveal that the relationship of DNA structure to the corresponding DNA sequence is not always simple. While similar sequences often adopt similar structures (Fig. 1A) divergent nucleotide sequences can have similar local structures (Fig. 1B) (20). Conversely (albeit less frequently), similar sequences can adopt very different structures (Fig. 1C). These observations indicate that DNA regions that differ on the basis of the order of

Correspondence to: Thomas D. Tullius; Elliott H. Margulies.

nucleotides may be similar in structure, which suggests they may perform similar biological functions.

We quantified the effect of single-base substitutions on DNA structure by computing the structural profiles of all possible 11-base pair (bp) sequences (4,194,304 in total), measuring the similarity between profiles for all pairs of 11-bp sequences that differ only by a single substitution at the central nucleotide (21). A histogram of structural differences for all 11-bp sequences reveals a range of effects from minor to drastic (Fig. 1D–G). In some cases, the one-base change in an 11-bp sequence has a minor effect, but the sequence may have a dramatically different structure if another base is substituted.

Because current sequence-constraint algorithms do not compute the effect of a base change on DNA structure, we developed a computer program, Chai, that incorporates structural information. This method works in a manner analogous to the DNA sequence-based binomial conservation (binCons) algorithm (8) but instead of computing the binomial probability of observed base substitutions between species, Chai calculates the difference between DNA structural profiles as a measure of similarity (21). We used the binCons and Chai algorithms to analyze 30 Mb of high-quality, comparative sequence data for 36 different species [the ENCODE pilot project regions (15,22)]. We defined a false discovery rate (FDR) on the basis of a neutral (or null) alignment (10,22) in an identical manner for assessing both sequence- and structure-informed conservation, so that each type of conserved region was identified with equal statistical confidence. We found that at any given FDR, the structure-informed Chai algorithm identified more evolutionarily constrained bases compared to binCons (Fig. 2A,Table S1). For example, at a 5% FDR, binCons identified 6.7% of bases as constrained, corresponding with previous findings of sequence-based mammalian constraint (8–11,15,22, 23). Chai identified nearly twice as many bases as constrained (12%) (Fig. 2B) while identifying 89% of the regions identified with the binCons method.

We next determined the extent to which evolutionarily constrained regions identified by Chai and binCons harbor functional elements. We examined deoxyribonuclease I (DNase I)-hypersensitive sites (15) and predicted transcriptional enhancers identified using chromatin modification patterns (24) from published studies (21) as examples of non-coding functional elements that frequently exhibit little sequence similarity. Compared to binCons regions, we found that Chai-detected regions overlap a higher proportion of DNase I-hypersensitive sites (78% for Chai versus 50% for binCons) and predicted enhancers (84% for Chai versus 59% for binCons) (Fig. 2C). To test whether the increased correlation with functional sequences was due to the additional territory identified by Chai relative to binCons, we tuned each algorithm for equal base coverage as opposed to equal statistical confidence and found that Chai-detected regions still overlap a significantly greater number of functional non-coding elements compared to binCons-detected regions ($p < 10^{-12}$; Fisher's exact test) (Fig. S1, Fig. 2A, Table S1).

Focusing our analysis on regions identified only by Chai and not by binCons (Chai-only regions) resulted in a statistically significant overrepresentation of noncoding functional sequences [$p < 0.01$, genome structure correction (GSC) statistic] [see (21) for a description of the GSC statistic] in the Chai-only regions and a statistically significant underrepresentation ($p < 10^{-9}$; GSC statistic) of coding regions (Fig. 2C). This suggests that some of the functional information in the non-coding portion of the genome is conferred by DNA structure as well as by the nucleotide sequence.

We thus examined if non-coding nucleotide substitutions inducing changes in DNA structure impact the biological function of a sequence. We focused on the DNA-binding properties of the Zif268 protein, a mammalian transcription factor that consists of three zinc fingers that

wrap around DNA in the major groove (25). Binding affinity data (26) were compared with structural profiles for the 15 sites identified to bind wild type Zif268 and the Zif268 REDV mutant (21). For both proteins the structural profiles of the high-affinity sequence motifs were similar, while low-affinity motifs had a different structural profile (Fig. S2A, Fig. S3A).

We ranked each Zif268 REDV motif by the extent of DNA structural difference relative to the best binding site, and found that binding affinity correlated with this ranking (r=0.75; p = 6.9 $\times 10^{-4}$; t-test; Fig. S2B). In other words, low-affinity binding sites differed dramatically in structure from high-affinity sites, while sites with intermediate affinity showed less structural difference. We observed a similarly high correlation for the wild type protein, which has different binding preferences (r = 0.74; p = 7.5 $\times 10^{-4}$; t-test; Fig. S3B). Analysis of DNA-binding site structural profiles and binding affinity for the archaeal transcriptional regulator Ss-LrpB revealed similar trends to those found for Zif268 (Fig. S4).

We next tested whether mutations that change the molecular topography of non-coding genomic regions have phenotypic consequences, using data from the PhenCode Project (27), which organizes information from several locus-specific mutation databases. We gathered 734 non-coding single-nucleotide variants in the human genome associated with a phenotype. For each variant, we calculated the difference in the structural profile between the mutant and non-mutant sequence over an 11-bp window centered on the variant nucleotide (similar to the analysis above). For comparison, a distribution of baseline variation in DNA topography was computed for 16,832 neutrally-evolving single-nucleotide polymorphisms (SNPs) (21). The phenotype-associated distribution was significantly correlated with larger changes in structure (p<3×10^{-4}; Wilcoxon rank sum test) relative to the baseline distribution, and contained an additional high structure-change peak (Fig. 3). These results indicate that phenotype-associated mutations tend to induce larger changes in the structural profile of non-coding DNA than does baseline neutral variation.

We also identified phenotype-associated SNPs causing the top 10% structural changes and the lowest 10%. The phenotype-associated mutations with the highest changes in DNA structure occurred significantly more often (p<10^{-2}, Fisher's exact test) in evolutionarily constrained regions of the genome (56% for high structure-change regions versus 29% for low structure-change regions) (21). This suggests that non-coding DNA may be under selective constraint, which may prevent changes in DNA structure. Because the severity of structural change might help identify functional SNPs, we constructed a database of changes in the structural profile for all known SNPs in the human genome (21).

Finally, we demonstrated that structural changes affect biological function in non-coding evolutionarily constrained regions identified by the Chai algorithm. We chose 12 predicted enhancer-containing regions (24): 5 regions overlap elements only detected by the Chai algorithm; 7 regions overlap a combination of Chai- and binCons-detected elements (Table S2). We cloned the 300 bp surrounding each of these genomic regions in a luciferase reporter construct and transfected them into 293T cells. Eight of the twelve constructs displayed luciferase activity that was significantly greater (p ≤ 0.05; Wilcoxon rank sum test) than that of random control sequences (Fig. 4). Three of the 5 constructs only overlapping Chai-detected elements were positive (Table S2).

Given the plethora of regulatory functions that a genome encodes and the three-dimensional genomic architecture required to orchestrate these events (28), it may not be surprising that there is widespread conservation of local DNA topography. Perhaps only a subset of local structural configurations can accommodate the functional requirements of a genomic locus [for example see (29)]. Once the molecular topography of a locus is permissive to a regulatory function, this structure may be maintained within the genome. Our high-resolution topography-

based constraint-detection method reveals that structure-informed constraint is widespread in the human genome and that these regions overlap known non-coding functional sites. Because different DNA sequences can have similar local structures (20), these regions might escape detection with sequence-based conservation–identification methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References and Notes

1. International Human Genome Sequencing Consortium. Nature 2004;431:931. [PubMed: 15496913]
2. Wilson MD, et al. Science 2008;322:434. [PubMed: 18787134]
3. Elnitski L, et al. Genome Res 2003;13:64. [PubMed: 12529307]
4. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Nature 2003;423:241. [PubMed: 12748633]
5. Loots GG, et al. Science 2000;288:136. [PubMed: 10753117]
6. Pennacchio LA, Rubin EM. Nat. Rev. Genet 2001;2:100. [PubMed: 11253049]
7. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE. Nat. Genet 2000;26:225. [PubMed: 11017083]
8. Margulies EH, Blanchette M, Haussler D, Green ED. Genome Res 2003;13:2507. [PubMed: 14656959]
9. Siepel A, et al. Genome Res 2005;15:1034. [PubMed: 16024819]
10. Cooper GM, et al. Genome Res 2005;15:901. [PubMed: 15965027]
11. Asthana S, Roytberg M, Stamatoyannopoulos J, Sunyaev S. PLoS Comp. Biol 2007;3:e254.
12. Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. Science 2006;312:276. [PubMed: 16556802]
13. McGaughey DM, et al. Genome Res 2008;18:252. [PubMed: 18071029]
14. Petrykowska HM, Vockley CM, Elnitski L. Genome Res 2008;18:1238. [PubMed: 18436892]
15. ENCODE Project Consortium. Nature 2007;447:799. [PubMed: 17571346]
16. Ohyama, T., editor. DNA Conformation and Transcription. Georgetown TX: Landes Bioscience/Eurekah.com; 2005.
17. Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB. Proc. Natl. Acad. Sci. USA 1998;95:11163. [PubMed: 9736707]
18. Dickerson RE. Methods Enzymol 1992;211:67. [PubMed: 1406328]
19. Price MA, Tullius TD. Methods Enzymol 1992;212:194. [PubMed: 1325598]
20. Greenbaum JA, Pang B, Tullius TD. Genome Res 2007;17:947. [PubMed: 17568010]
21. Materials and methods are available as supporting material on *Science* Online.
22. Margulies EH, et al. Genome Res 2007;17:760. [PubMed: 17567995]
23. Mouse Genome Sequencing Consortium. Nature 2002;420:520. [PubMed: 12466850]
24. Heintzman ND, et al. Nature Genet 2007;39:311. [PubMed: 17277777]
25. Pavletich NP, Pabo CO. Science 1991;252:809. [PubMed: 2028256]
26. Bulyk ML, Johnson PLF, Church GM. Nucleic Acids Res 2002;30:1255. [PubMed: 11861919]
27. Giardine B, et al. Human Mutation 2007;28:554. [PubMed: 17326095]
28. Misteli T. Bioessays 2005;27:477. [PubMed: 15832379]
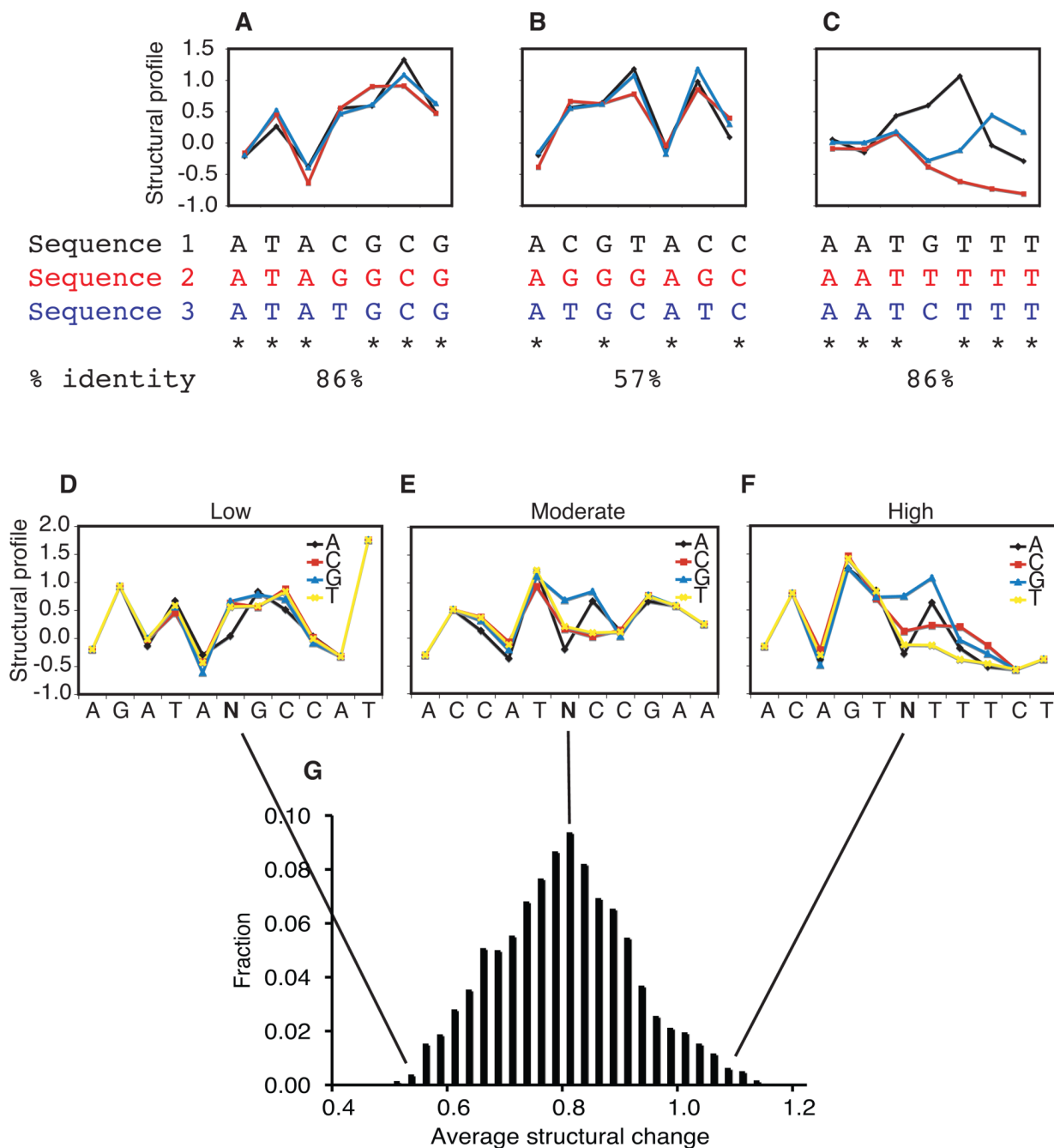29. Joshi R, et al. Cell 2007;131:530. [PubMed: 17981120]

Fig. 1.
DNA topography vs. nucleotide sequence. (**A–C**) Hydroxyl radical cleavage patterns (referred to here as the structural profile) and corresponding color-matched sequence alignments (below each panel). Asterisks indicate identical columns in the sequence alignment. (**D–G**) Single-base substitutions have a range of effects on local DNA structure. For all possible 11-bp nucleotide sequences, we computationally changed the middle position to each possible base and measured its effect on the structural profile. These changes were quantitatively measured by calculating the mean Euclidean distance (referred to here as the average structural change) (21), where low values indicate similar structure profiles and large values indicate different structure profiles. (**G**) The distribution of average structure changes observed for all 11-bp

sequences. Arrows indicate what we classify as low, moderate, or high average structure changes, with representative 11-bp sequences for each shown in (**D**), (**E**), and (**F**), respectively. For (**D–F**), the y-axis indicates the structural profile at each nucleotide position in the 11-bp sequence (x-axis). The structural profile for each sequence containing a different base in the middle position (noted by an 'N') is plotted in a different color. The structural profile patterns are increasingly different in changing from low to high structural change [(**E–G**)].
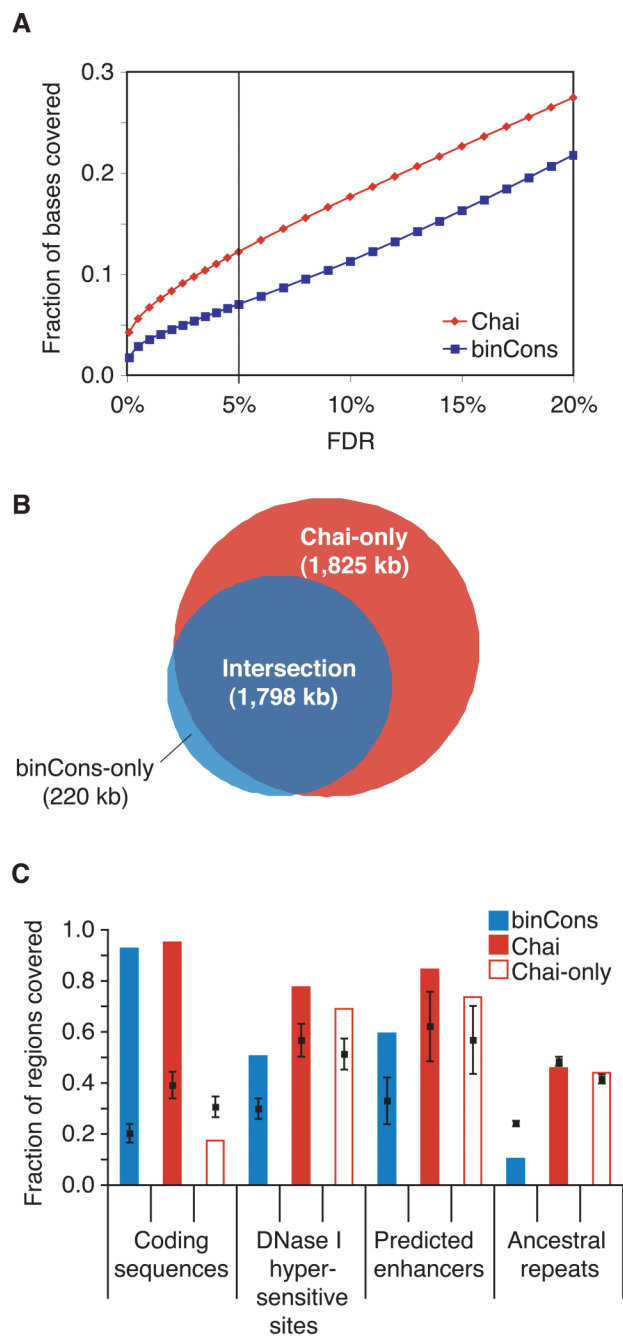
A

B

C

Fig. 2.
Structure-informed evolutionarily constrained regions. (**A**) Plot showing the fraction of bases identified as evolutionarily constrained at various FDRs (21) by the DNA sequence-based algorithm binCons (blue line) and the structure-informed algorithm Chai (red line). (**B**) Venn diagram showing nucleotides identified by each algorithm [Chai (red) and binCons (blue)] at 5% FDR (gray vertical line in **A**). Chai detects nearly the same set of bases as binCons identifies (intersection), plus a nearly equal number of additional bases (Chai-only). (**C**) Bar graph showing the fraction of different types of genomic regions that overlap binCons- (blue), Chai- (red), and Chai-only-detected (white) constrained elements. Black points are the mean of a null

distribution constructed with the GSC method (21); error bars represent 95% confidence intervals.

A

B
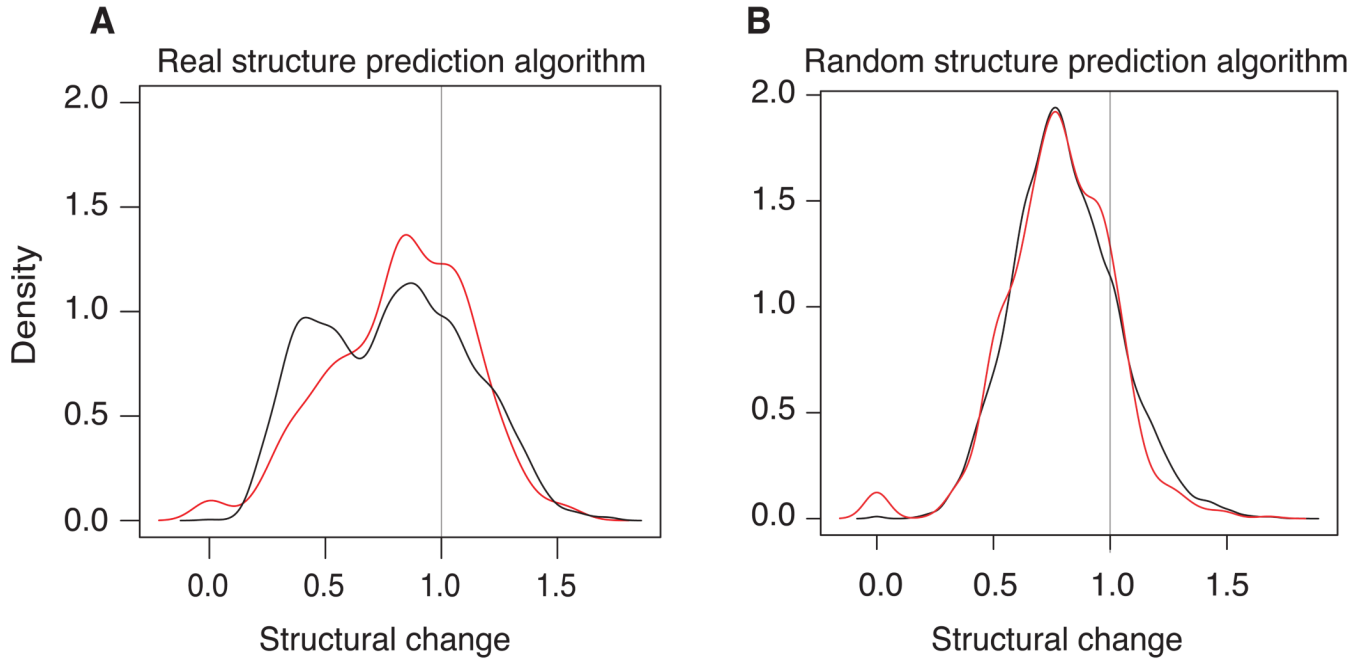
Fig. 3.
The distribution of DNA structural changes for phenotype-associated variants (red line) and neutral variants (black line) (21) with the real structure prediction algorithm (**A**) and a randomized version (**B**). (**A**) The phenotype-associated distribution was shifted significantly to the right ($p < 3 \times 10^{-4}$; Wilcoxon rank sum test) and contains an additional high structure change peak relative to the neutral distribution. (**B**) No significant shift ($p > 0.05$; Wilcoxon rank sum test) was observed with a randomized version of the structure prediction algorithm.
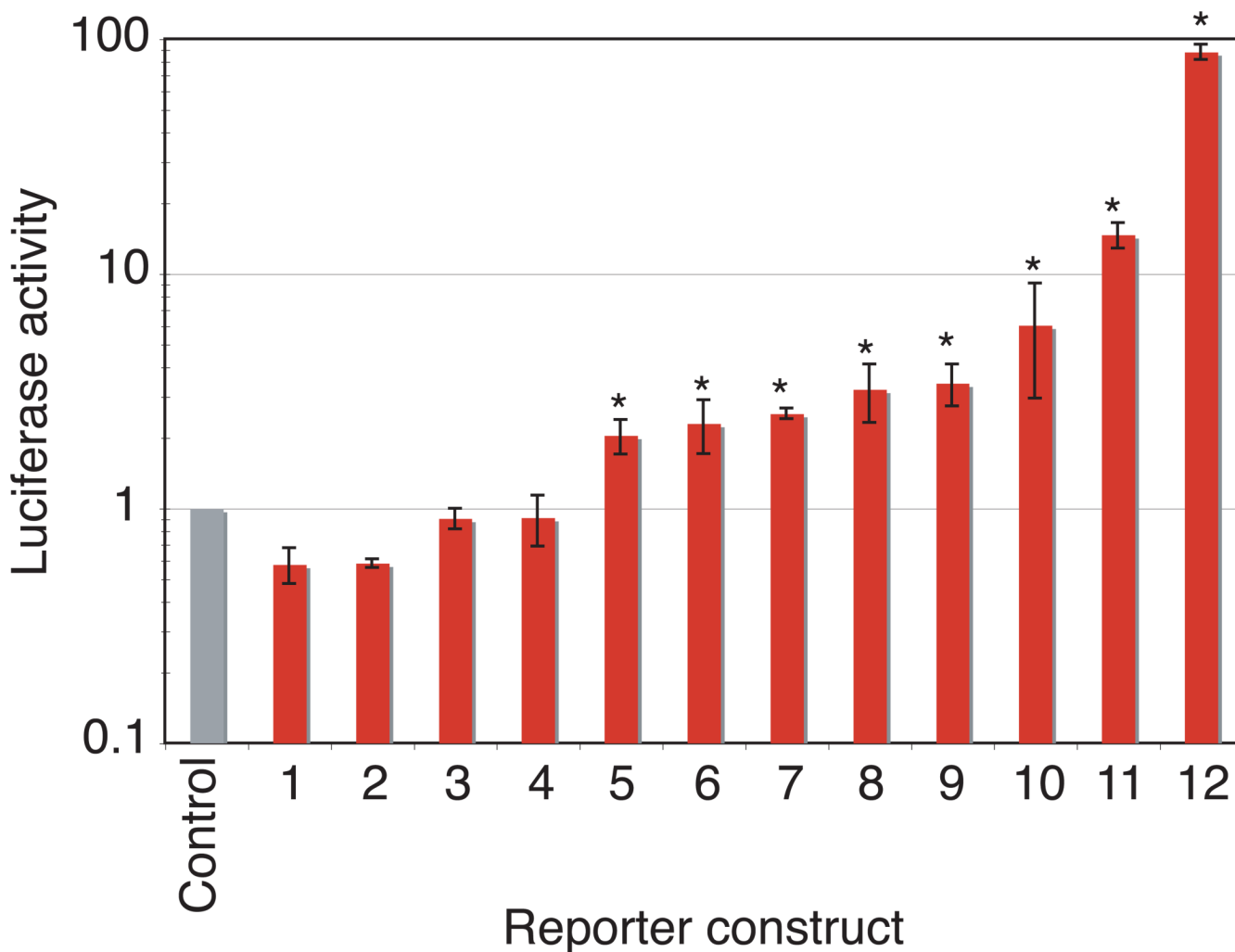
Fig. 4.
Luciferase-based reporter activity of 12 regions containing Chai-detected elements (21). These regions overlapped predicted enhancer regions. Plotted for each element is the luciferase activity relative to the median activity from 100 random control constructs (y-axis; see Fig. S5). Error bars represent one standard deviation from the mean of four experimental replicates, and asterisks denote $p \leq 0.05$ (Wilcoxon rank sum test).