frontiers in Chemistry

ORIGINAL RESEARCH
published: 18 January 2019
doi: 10.3389/fchem.2018.00666

Trinucleotide Base Pair Stacking Free Energy for Understanding TF-DNA Recognition and the Functions of SNPs

Gen Li, Yuan Quan, Xiaocong Wang, Rong Liu, Lihua Bie, Jun Gao and Hong-Yu Zhang*

Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan, China

OPEN ACCESS

Edited by:
Yong Wang,
Ningbo University, China

Reviewed by:
Dingguo Xu,
Sichuan University, China
Qing-Chuan Zheng,
Jilin University, China

***Correspondence:**
Jun Gao
gaojun@mail.hzau.edu.cn

Specialty section:
This article was submitted to Theoretical and Computational Chemistry, a section of the journal Frontiers in Chemistry

Received: *03 December 2018*
Accepted: *21 December 2018*
Published: *18 January 2019*

Citation:
Li G, Quan Y, Wang X, Liu R, Bie L, Gao J and Zhang H-Y (2019) Trinucleotide Base Pair Stacking Free Energy for Understanding TF-DNA Recognition and the Functions of SNPs. Front. Chem. 6:666. doi: 10.3389/fchem.2018.00666

Single nucleotide polymorphisms (SNPs) affect base pair stacking, which is the primary factor for maintaining the stability of DNA. However, the mechanism of how SNPs lead to phenotype variations is still unclear. In this work, we connected SNPs and base pair stacking by a 3-mer base pair stacking free energy matrix. The SNPs with large base pair stacking free energy differences led to phenotype variations. A molecular dynamics (MD) simulation was then applied. Our results showed that base pair stacking played an important role in the transcription factor (TF)-DNA interaction. Changes in DNA structure mainly originate from TF-DNA interactions, and with the increased base pair stacking free energy, the structure of DNA approaches its free type, although its binding affinity was increased by the SNP. In addition, quantitative models using base pair stacking features revealed that base pair stacking can be used to predict TF binding specificity. As such, our work combined knowledge from bioinformatics and structural biology and provided a new understanding of the relationship between SNPs and phenotype variations. The 3-mer base pair stacking free energy matrix is useful in high-throughput screening of SNPs and predicting TF-DNA binding affinity.

Keywords: base stacking, free energy, single nucleotide polymorphisms, molecular dynamics simulation, binding specificity, transcription factor

INTRODUCTION

The stacking of adjacent base pairs and the pairing between complementary bases via hydrogen bonding are fundamentally related to the sequence and shape properties of DNA and critically influence the configurations, stabilities, and other properties of DNA(Yakovchuk et al., 2006; Hase and Zacharias, 2016; Kilchherr et al., 2016). The structure of a DNA sequence at the transcription factor (TF) binding site affects the interactions between these molecules (Rohs et al., 2009; Stormo and Zhao, 2010). In addition to current methodologies (Garvie and Wolberger, 2001; Slattery et al., 2014), base pair stacking free energy provides a novel way for understanding TF-DNA interactions. Determining base pair stacking free energy for a DNA sequence is critical for substantiating this underlying relationship. In the past few decades, several methods have been developed for determining base pair stacking free energy, including optical spectroscopic techniques (Warshaw and Tinoco, 1966), NMR spectroscopy (Chan and Nelson, 1969), and self-diffusion NMR (Stokkeland and Stilbs, 1985). More recently, base pair stacking free energies were measured in DNA fragments, such as beacon kinetics (Aalberts et al., 2003), thermal denaturation (Guckian et al., 1996), and mechanical unzipping (Huguet et al., 2010).

A single nucleotide polymorphism (SNP) is a common mutation phenomenon in the human genome (Clamp et al., 2007; Kimchi-sarfaty et al., 2007; Lu et al., 2015) and can significantly influence interactions between DNA and TFs, leading to related disease or phenotype variations (Deplancke et al., 2006; Bass et al., 2015). A single mutation in a SNP affects the base pair stacking free energies of two consecutive dinucleotides (2-mers), which compose a trinucleotide (3-mer) with the mutation site located at the central position. Studying 3-mers provides more comprehensive information than studying two 2-mers (Santalucia and Hicks, 2004; Taghavi et al., 2017). Another reason of selection 3-mer is because 3 base pairs are the minimal unit to describe the base-pairs stacking change of single SNP. However, measuring base pair stacking free energy for 3-mers is beyond the capability of current experimental approaches and remains a challenge to the scientific community (Sponer et al., 2013).

Base pair stacking is the primary factor for maintaining the stability of DNA structures (Yakovchuk et al., 2006). MD simulations (Hase and Zacharias, 2016) and quantum mechanics (QM) calculations (Parker et al., 2013) reveal that changes in the base pair stacking free energy affect many DNA parameters, including the twist, slide, groove, and bend of DNA. Recent studies showed that the structural changes caused by SNPs affect the binding affinity of protein-DNA complexes (Arkova et al., 2016). Based on the view that shape readout plays an important role in TF-DNA interactions (Slattery et al., 2011; Gordan et al., 2013; Yang et al., 2014), Zhou et al. developed a quantitative model that utilized sequences as well as DNA shape features and achieved a higher accuracy than traditional sequence models (Zhao et al., 2012; Mordelet et al., 2013; Zhou et al., 2015). Thus, we hypothesize that changes in the base pair stacking could cause DNA structural variations and transcriptional regulation disorders, which would eventually disrupt the TF-DNA interactions and lead to various diseases or phenotype variations. Therefore, measuring base pair stacking free energy is greatly beneficial for studying the underlying relationship between base-pair stacking and related disease or phenotype variations.

In this study, a 3-mer base pair stacking free energy matrix was constructed to calculate the base pair stacking free energy of 3-mers based on those of 2-mers. A 3-mer base pair stacking free energy difference matrix was then built to establish the correlation between the base pair stacking free energy and SNPs. Statistically significant variants from the GWAS database (GWASdb) were analyzed to identify the relationship between the base pair stacking free energies and the SNPs related to phenotype variations. The phenotype variations were enriched in the regions that possessed large differences in the base pair stacking free energies. Next, MD simulations revealed that changes in the base pair stacking free energies led to function variations via structural changes in the DNA, including twist, slide, and groove. Lastly, base pair stacking free energy was combined with experimental sequence data to generate a 1-mer+ΔG_s model for quantitatively predicting TF-DNA binding affinities, which exhibited a higher accuracy efficiency than 1-mer model. Our study revealed the significance of base pair stacking free energies for tri- or longer nucleotides and their relationships with the function of SNPs. We believe that the 3-mer base pair stacking free energy matrix may pave a new way for understanding and predicting TF-DNA interactions.

MATERIALS AND METHODS

Generation of the 3-mer Base Pair Stacking Free Energy Matrix

A base pair stacking interaction involves adjacent base pairs, which means we need two 2-mers to represent a complete stacking interaction. Although 2-mer base pair stacking has been studied (Warshaw and Tinoco, 1966; Chan and Nelson, 1969; Stokkeland and Stilbs, 1985; Guckian et al., 1996; Aalberts et al., 2003; Huguet et al., 2010), 3-mer base pair stacking is poorly studied, partly due to the limitations of experimental techniques and the computing power of quantum mechanics (Sponer et al., 2013). To calculate the 3-mer base pair stacking free energy, we combined the base pair stacking free energies for the two consecutive 2-mers in the 3-mer (Santalucia and Hicks, 2004; Taghavi et al., 2017):

$$\Delta G_{ABC} = \Delta G_{AB} + \Delta G_{BC} \tag{1}$$

where ΔG_{ABC} is the base pair stacking free energy of three consecutive nucleotides, ABC, and ΔG_{AB} and ΔG_{BC} are the stacking free energies of two adjacent nucleotides within the three consecutive nucleotides, AB and BC, respectively. Our 3-mer base pair stacking free energy matrix was constructed using the 2-mer base stacking energies from Protozanova et al.'s experimental data (Protozanova et al., 2004).

Building the Phenotype Variation Related 3-mer SNPs Dataset

We downloaded SNPs related to human phenotypes (traits) from the GWASdb (http://jjwanglab.org/gwasdb, before August 2015). The GWASdb is the most widely used GWAS result database (Li et al., 2012), and it combines the National Human Genome Research Institute (NHGRI) GWAS Catalog, the tables and supplementary materials of manuscripts archived in the NHGRI GWAS Catalog, and the database of Genotypes and Phenotypes (dbGaP). To obtain the SNPs notably related to phenotype variations, only statistically significant ($P \leq 1 \times 10^{-8}$) variants were included. A total of 25,029 SNPs, which included 883 human traits, were used for the analysis. To study the changes in the 3-mer base pair stacking free energy caused by the SNPs, the SNP adjacent nucleotides were obtained from the BioMart of Ensembl database (version 88) (Yates et al., 2016). To analyze the SNPs that were located at TF binding sites, we used RegulomeDB (http://www.regulomedb.org) to filter out the SNPs in the motif (Boyle et al., 2012), and there were 10123 SNPs in total. The raw data can be found in the **Supplementary Data Sheet**.

Building the High-Resolution TF-DNA Complex Crystal Structure and JASPAR Dataset

The crystal structures of the TF-DNA interaction complexes used in this work were obtained from the Protein Data Bank and were published before 11 January 2017. There were 81 crystal structures that contained both TF and DNA with ≤ 2.0 Å resolution and no chemical modifications, mismatches or drugs. All of the TF-DNA interaction crystal structure complexes are listed in **Table S1**. The TFBS dataset contains nucleotide sequences that are within a distance of 3.5 Å of the TF in the high-resolution TF-DNA complex crystal structure dataset. The JASPAR database dataset consists of 593 non-redundant core nucleotide sequences in the JASPAR database that were download from http://jaspar.binf.ku.dk.

According to the work by Bass et al. (2015), we first selected the SNPs related to human phenotype variations with the highest supporting evidence score. Second, the corresponding TF-DNA complexes from our high-resolution TF-DNA complex crystal structure dataset were screened. Only 2 SNPs (MUT_17 and MUT_190) were reserved. MUT_17 was located in a specific binding site of Hepatocyte nuclear factor 4 alpha (HNF4α), a transcription factor containing zinc finger motifs, with force field parameters that still need improvements (Santos-martins et al., 2014). MUT_190 was located in a specific binding site of MEIS1. The structure of MEIS1 is similar to that of HNF4α, which contains a DNA binding domain but not a zinc finger (Jolma et al., 2015). There is an A→ G SNP in the binding site (ACT*A*TCGA → ACT*G*TCGA) that is located in the MCP-1 promoter sequence (−2511 to −2528) at −2518, and this SNP increases MEIS1 binding affinity and leads to hepatitis C virus (HCV)-related liver disease (Bass et al., 2015). The MEIS1 complex (PDB ID: 4XRM) was adopted as a template to construct the structures of mutated-complexes.

Molecular Dynamics Simulation Protocol

The MD simulations of the constructed systems were performed by using the NAMD software package (Kal et al., 1999) with AMBER ff14SB (Hornak et al., 2006) and parmbsc1 force fields (Ivani et al., 2016). Both the wild-type and mutated complexes and the free DNA were embedded in a box-shaped (72 × 68 × 84 Å^3) bath of water molecules, and there was a layer of TIP3P water 12 Å in each direction from the atom with the largest coordinate in that direction. The system was neutralized with sodium cations. Na$^+$ and Cl$^-$ ion pairs were then added to reach a physiological salt concentration of 0.15 M. The solvated complex was equilibrated by carrying out a series of 1,000 steps of energy minimization with 10 kcal/mol/Å^2 restraints on the backbone, after 1000 steps of minimization without restraints, 310 ps of heating restricted 2 kcal/mol/Å^2 on the backbone from 0 to 310 K and 1 ns of density equilibration with NVT followed by 200 ns of constant pressure equilibration at 310 K. The system was equilibrated using an NPT ensemble at 310 K and pressure at 1 atm (1 atm = 101.3 kPa). All the simulations were run with SHAKE on hydrogen atoms, a 2 fs time step and a Langevin thermostat for temperature control and pressure control. Periodic boundary conditions and the Particle-Mesh-Ewald (PME)(Essmann et al., 1998) algorithm were adopted to compute the long range electrostatic forces, and the cutoff was set as 12 Å. Trajectory frames were collected at every 5 ps for a total of 50 ns. Curves+ software (Swaminathan et al., 1990; Blanchet et al., 2011) was employed to calculate the base pair parameters to define the geometry of the DNA. The values of the MD geometries presented here ignore the terminal base pairs of the oligomers since these often suffer from local deformations (Etheve et al., 2016). The standard value for the DNA structures used the data from Olson et al. (2001).

Model of TF Binding Affinity Prediction

For a DNA sequence of length K, the 1-mer feature was used to represent each nucleotide position, and the target sequence was seen as a binary vector with a length of 4K. For example, one nucleotide position was encoded as 0 0 0 1, which indicated A, T, G, and C, respectively, and a value of 1 represented the occurrence. Regarding the base pair stacking free energy feature, a sliding-window approach to the DNA sequence calculated the base pair stacking with every 3-mer. We used the genomic context PBM (gcPBM) data from Zhou et al. to train and test the 1-mer model, the 1-mer+shape model, and the 1-mer+ΔG_s model. The gcPBM was derived from the Gene Expression Omnibus (GEO) under the accession number GSE59845 (Zhou et al., 2015), which contained 36-bp genomic sequences. The gcPBM data for each TF were converted into a matrix after preprocessing and feature encoding. The first column of this matrix contained the natural logarithm of the fluorescence signal intensities of the PBM probes, and the remaining columns contained the encoded features. The E-SVR algorithm in the LIBSVM toolkit (Chang and Lin, 2011; Claesen et al., 2014) was used to train the linear regression model to predict the natural logarithm of the gcPBM signal intensities based on the encoded sequence and base pair stacking free energy. The total length of the DNA base pair stacking vectors was 34 due to the unavailability of the values at two positions at the end. To obtain unbiased performance estimates of the regression models in each dataset, a nested 10-fold cross-validation procedure was implemented. The details of the gcPBM raw data processing methods were described by Zhou et al. (2015). The time-consuming of the model were tested on CPU E5-2683v3.

RESULTS AND DISCUSSION

Building the 3-mer Base Pair Stacking Free Energy Matrix

The base pair stacking free energies for 3-mers were calculated by the sum of the base pair stacking free energies for two consecutive 2-mers (see Equation 1 in the Materials and Methods section). This strategy was used also by SantaLucia et al. (Santalucia and Hicks, 2004) and Taghavi et al. (2017) and showed reliable accuracies. Currently, base pair stacking free energies for 2-mers have been extensively studied, both theoretically and experimentally. Friedman and Honig calculated 2-mer base pair stacking free energies theoretically and reported values ranging from −7.79 to −4.36 kcal/mol (Friedman and Honig, 1995), whereas, Protozanova et al. measured them experimentally in a nicked DNA duplex, with values ranging from −2.17 to −0.19 kcal/mol (Protozanova et al., 2004), and later, Kilchherr et al.'s experimental results values ranged from −3.42 to −0.78 kcal/mol (Kilchherr et al., 2016). Since the theoretical 2-mer stacking free energies have a tendency for overestimations (Hase and Zacharias, 2016), we used the 2-mers stacking free energies from

Protozanova et al.'s study to calculate the 3-mer base pair stacking free energies for their wide acceptances (Hase and Zacharias, 2016).

All 64 combinations of the 3-mer base pair stacking free energies were computed, and a 3-mer base pair stacking free energy matrix was created (**Figure 1A**). Generally, the stacking interaction between the GC base pair is stronger than that between the TA base pair (Geggier and Vologodskii, 2010). In 2-mers, GC and TA have the most negative and positive values for base pair stacking free energies, respectively. Similarly, in 3-mers, GGC and GCC base pairs have the most negative base pair stacking free energies, while TAG and CTA have the most positive base pair stacking free energies.

A 3-mer base pair stacking free energy difference matrix was also developed to study the changes in the base pair stacking free energies for SNPs ($\Delta\Delta G_S$). In this matrix, the changes in the base pair stacking free energies for 64 trinucleotides and 4 possible mutations at each central nucleotide were calculated from the 3-mer base pair stacking free energy matrix. As shown in **Figure 1B**, the base pair stacking free energies for the 3-mers decreased after mutations to G and C occurred and increased after mutations to A and T. The maximum changes were CTA→ CGA (−1.09 kcal/mol) and TCG→ TAG (+1.09 kcal/mol). Furthermore, the value of the $\Delta\Delta G_s$ was always lower for a SNP with a mutation to G or C compared to the same SNP with a mutation to A or T.

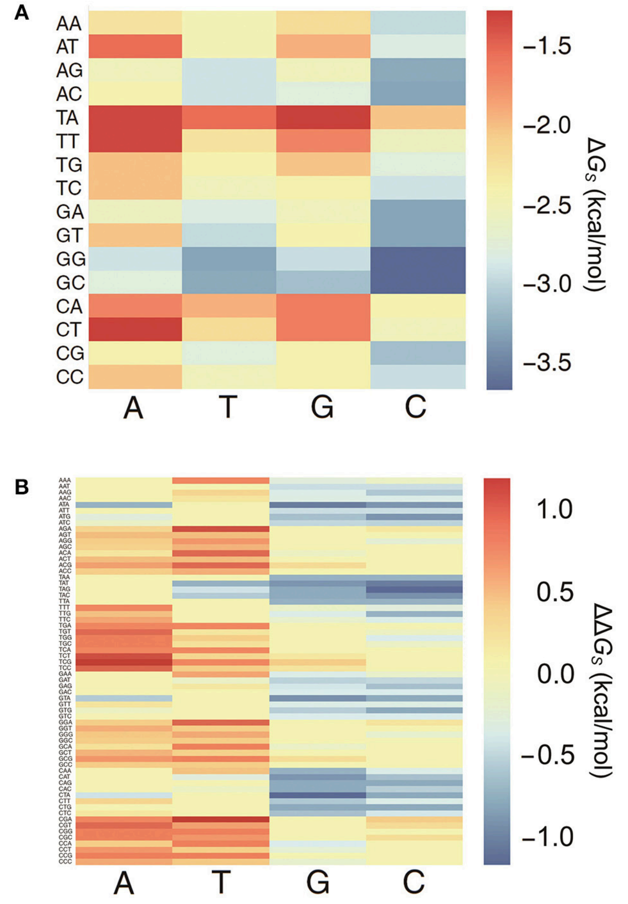

FIGURE 1 | The heat map of 3-mer base pair stacking matrixes. **(A)** Heat map of the 3-mer base pair stacking free energy matrix. The X-axis is the last nucleotide in the 3-mer, and the Y-axis is the first two nucleotides. **(B)** Heat map of the 3-mer base pair stacking free energy difference matrix. The X-axis is the mutation at the middle position in the 3-mer, and the Y-axis is the 64 combinations of 3-mers. The values for the individual entries in the 3-mer base pair stacking free energy matrix and the 3-mer base pair stacking free energy difference matrix are listed in **Tables S2**, **S3**.

Investigating the Relationship Between the 3-mer Base Pair Stacking Free Energy Difference Matrix and the Phenotype Variation

Recent advances in genome-wide association studies (GWAS) in genetics have enabled us to identify thousands of genetic variants that are associated with phenotype variations. It is well known that SNPs are closely linked with various phenotypes or traits (Kimchi-sarfaty et al., 2007; Helyar et al., 2011; Gutierrez-arzaluz et al., 2017), such as obesity and age-related macular degeneration(Sangiovanni et al., 2017; Dong et al., 2018). Herein, several assumptions were made to explore the relationships between SNPs and phenotype variations.

The first assumption was that the enhanced of the base pair stacking free energy was related to the phenotype variation. To validate this assumption, a phenotype variation-related 3-mer SNPs dataset was generated. There were 25,029 SNPs in the dataset, which involved 883 human traits. The base pair stacking free energy difference of each 3-mer SNP was obtained via the 3-mer base pair stacking free energy difference matrix. Interestingly, 47.83% of the variants (**Table S4**) showed enhanced base pair stacking interactions after mutation, whereas this ratio increased to 51.87% (hypergeometric test, $P < 10^{-20}$) when only including variants with mutations in the TF binding sites. This increase may imply that the TFs are more sensitive to enhanced of base pair stacking. To confirm this hypothesis, we constructed two datasets (**Table 1**). The first was a high-resolution crystal structure dataset of TF binding, which had 81 binding sequences from the TF-DNA crystal structure complexes. The second one was the DNA binding motifs from the JASPAR database, which had 593 sequences of the TF binding motif. The AT contents in these two datasets were 54.90 and 53.94%, respectively. Therefore, we concluded that the relatively high ratio of the TF binding site was from not only the enhanced of the base pair stacking but also the high AT content of the TF binding site since the AT pairs have more room for enhancing the base pair stacking interactions during mutations.

The second assumption was that the scale of the base stacking difference was related to the phenotype variation. First, we focused on the SNPs in TF binding sites, since it may affect TF-DNA interactions. The changes in the base pair stacking free energy ($|\Delta\Delta G_s|$) of the phenotype variation-related 3-mer SNPs dataset (for TF binding site only) were obtained from the 3-mer base pair stacking free energy difference matrix and were manually screened. As shown in **Figure 2**, we categorized the variants of the $|\Delta\Delta G_s|$ values into three bins, namely, 0.0–0.3, 0.3–0.6, and larger than 0.6 kcal/mol. The ratios of the $|\Delta\Delta G_s|$

TABLE 1 | The ratio of AT and GC in the two different datasets.

Dataset	Ratio (%)	
	AT	GC
TFBS	54.90(538)	45.10(442)
JASPAR database	53.94(772)	46.06(654)

The TFBS dataset is the nucleotide sequences that were within a distance of 3.5 Å of the TF in our high-resolution TF-DNA complex crystal structure dataset. The JASPAR database dataset consists of the nonredundant core nucleotide sequences in the JASPAR database.

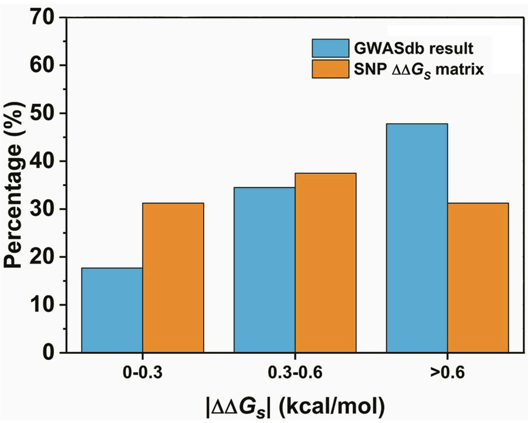

FIGURE 2 | The distribution of the base pair stacking free energy differences. The SNPs of the TF binding sites are labeled in blue, and the distribution of the 3-mer base pair stacking free energy difference matrix is labeled in orange. Student's test of the two series yielded a value of $P < 10^{-9}$.

values were 17.7, 34.5, and 47.8%, respectively. Interestingly, the total number of the SNPs with a $|\Delta\Delta G_s|$ value lager than 0.6 kcal/mol was almost half of all the phenotype variation-related SNPs (the overall SNP results that had the same trend as the SNP located at the TF binding sites are found in **Figure S1**). At the same time, we calculated the $|\Delta\Delta G_s|$ value distribution of the 3-mer base pair stacking free energy difference matrix. It was nearly evenly distributed, and the ratio of >0.6 kcal/mol was ∼30%. Student's test on the two series yielded a value of $P < 10^{-9}$, indicating that the SNPs located in TF binding sites with a larger $|\Delta\Delta G_s|$ had a higher ratio. To confirm this finding, the SNPs were recategorized based on their mutation types (**Figure 3A**). There was a marked preference A→ G, C→ T, G→ A, and T→ C mutations, which accounted for 70.0% of all SNPs. Interestingly, the average $|\Delta\Delta G_s|$ values for these four types of mutations were also the top 4, as shown in **Figure 3B**. Furthermore, as shown in both **Figures 3C,D**, the distribution of the $|\Delta\Delta G_s|$ for SNPs in the 4 preferred mutation types displays a similar trend to that found in all the chosen variants, in which a larger $|\Delta\Delta G_s|$ occupies a higher ratio.

In contrast, SNPs with a low $|\Delta\Delta G_s|$ were speculated to be disfavored. A total of 590 variants in the two lowest ratio mutation types displayed a significantly different distribution from those in the preferred mutation types. Variants with a $|\Delta\Delta G_s|$ smaller than 0.3 kcal/mol accounted for over 70% of the total in these two mutation types (**Figure 4**). Therefore, SNPs with a smaller $|\Delta\Delta G_s|$ appeared to be disfavored in our GWASdb results. Since all of our chosen variants were statistically significant, this implied that the SNPs in the TF-binding sites with a larger $|\Delta\Delta G_s|$ were more likely to lead to a phenotype variation. In summary, our study with statistically significant variants from the GWASdb showed that phenotype variation prefers SNPs with a large $|\Delta\Delta G_s|$ and made a solid case, where a 3-mer base pair stacking free energy matrix and a 3-mer base pair stacking free energy difference matrix helped to probe the relationship between base pair stacking free energy and diseases.

Simulation of the Relationship Between Base Pair Stacking and TF-DNA Interactions

As discussed earlier, SNPs result in changes in base pair stacking free energies. Recent studies showed that the base pair stacking free energy is closely related to DNA structure (Luscombe et al., 2001; Baker and Grant, 2007; Gu et al., 2015), which means it is likely that base pair stacking free energies affect TF-DNA interactions via DNA structural changes. However, the impact of SNPs on the structure of DNA is poorly studied, especially the relationship between the change in the base pair stacking free energy caused by the SNP and TF-DNA interaction. To explore this potential relationship, we searched the entire PDB, but did not find a TF-DNA cocrystal complex that had crystal structures for both wild-type and mutated complexes. To solve this problem, we combined our existing high-resolution TF-DNA complex crystal structure dataset with the study by Bass et al. (2015). The SNPs with the highest supporting evidence scores listed in the study by Bass et al. were selected. These SNPs are all located in the regulatory region and have been studied for their influence on TF-DNA binding affinities (Bass et al., 2015). Only two TF-DNA complexes with these SNPs in the regulatory region have available cocrystal structures, namely, HNF4α and Meis homeobox 1 (MEIS1). However, HNF4α was excluded due to inaccurate force field parameters for the zinc finger motifs (Santos-martins et al., 2014).

MEIS1 plays an essential role in the development and function of vertebrate organs (Shen et al., 1999). It is a homodimer of the TALE type homeobox transcription factor that regulates gene expression by binding specific DNA sequences (Jolma et al., 2015) (**Figure 5**). There is a SNP at the binding site (ACT**A**TCGA → ACT**G**TCGA) that causes an increase in the binding affinity and results in hepatitis C virus (HCV)-related liver disease (Bass et al., 2015). More interestingly, this mutation (A→ G) was one of the 4 preferred mutation types in our chosen variants from the GWASdb (**Figure 3A**). The cocrystal structure for the MEIS1 complex (PDB ID: 4XRM) was adopted as a template to construct the structures of the mutated complexes.

To verify our molecular models for the MEIS1 complexes, both the Helmholtz (ΔH_b) and binding free energies (ΔG_b) for the wild-type and mutated complexes were calculated. As shown in **Table S5**, the ΔH_b for the mutated complex was −6.27 kcal/mol. After considering the entropic penalty (Chang et al.,

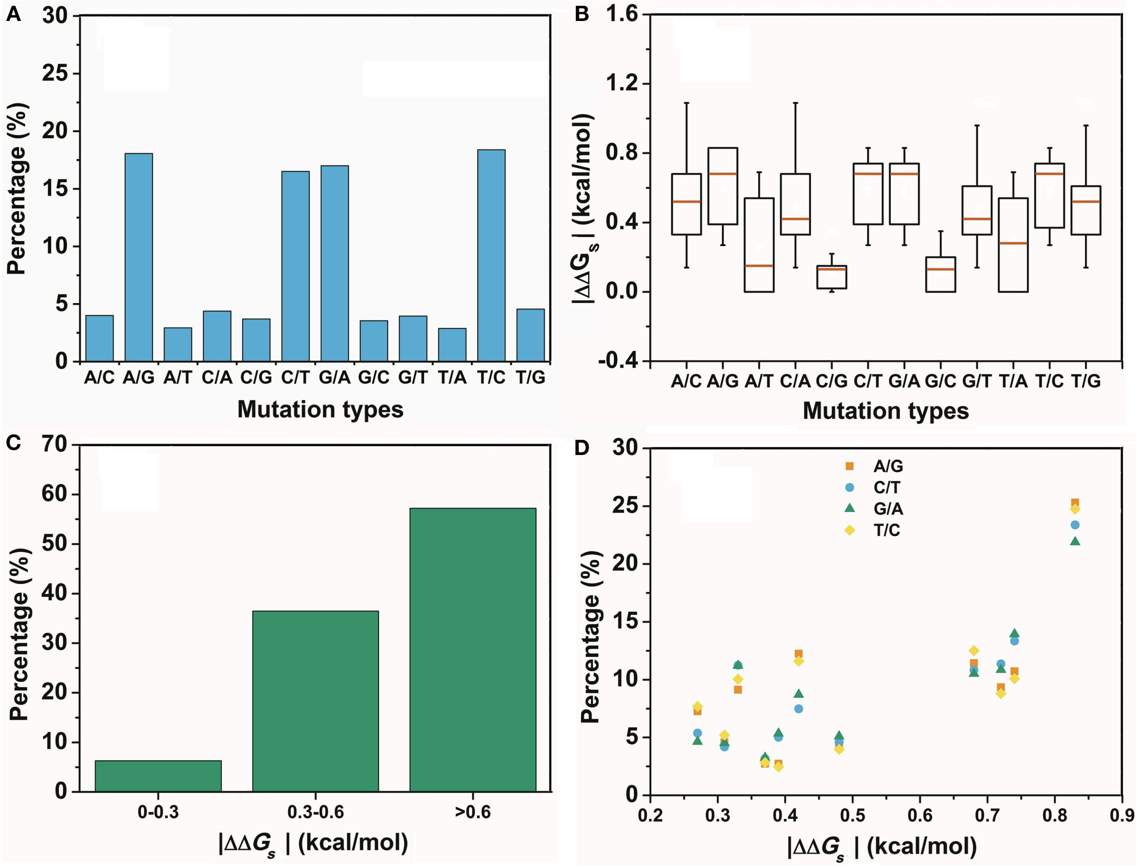

FIGURE 3 | The distribution of the different mutation types. **(A)** The percentage of the different mutation types in the GWASdb result. **(B)** The average $|\Delta\Delta G_S|$ for the variants chosen from the GWASdb in the different mutation types. **(C)** The percentages of the variants in the top 4 mutation types of **(A)** with different $|\Delta\Delta G_S|$ ranges. **(D)** The distributions of the $|\Delta\Delta G_S|$ for the top 4 mutations types in **(A)**.

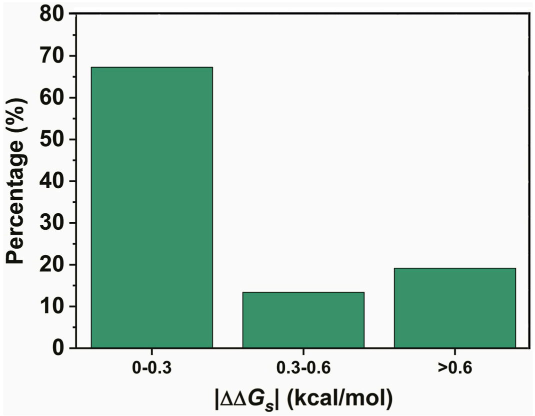

FIGURE 4 | The percentages of the variants in the 2 mutation types with the lowest ratio mutation types with different $|\Delta\Delta G_S|$ ranges.

2007), the ΔG_b for the mutated complex was still −1.42 kcal/mol. This was consistent with the experimental measurements, in which the binding affinity for the MEIS1 complex increases after the A→ G mutation in the TF binding site (Bass et al., 2015). Our MM/GBSA results correlated well with the Montclare's DNA mutation binding affinity experiments, which report values ranging from −5.0 to −1.3 kcal/mol (Montclare et al., 2001).

The average parameters of the DNA structures of the wild-type and mutated complexes were then compared (Blanchet et al., 2011). We analyzed the average structure difference values (D-values) of the DNA parameter between the wild-type complex, the mutated complex, the mutated free DNA and the wild-type free DNA. First, the differences (D-values) in the twist and slide for the wild-type and mutated complexes and the mutated free DNA at the mutation site were small, whereas those for other sites were relatively larger (**Figures 6A,B**). This phenomenon showed that the stronger base pair stacking free energy (**Table 2**) made the DNA structure closer to that of the free DNA, since the inter parameter was directly related to the base pair stacking free energy, and it also showed that the base pair stacking free energy was a long-range allosteric effect (Gu et al., 2015). Second, the amplitude of the variation for left and right at the mutation site was inconsistent. As shown, the changes in the left side were obviously larger than those in the right side. For the complexes, the D-values of the twist and slide had a larger amplitude of

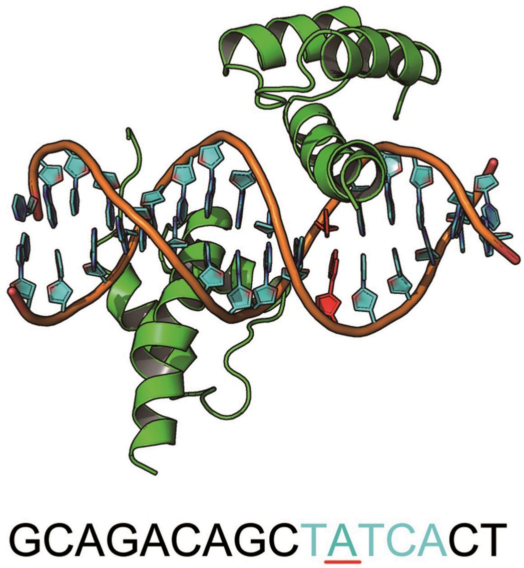

FIGURE 5 | The structure of MEIS1 and its DNA. The blue DNA sequence is the motif, and the red nucleotide is the mutation site.

variation than the free DNA. This suggested that the asymmetry of the left and right was due to the TF interaction (**Figure 5** shows that the left side of the mutation site was the binding location of the other monomer). Lastly, for the slide, there was a small difference between the wild-type and mutated complexes. However, the difference in the wild-type and mutated for the twist were more notable than those for the slide, which meant the effect of the SNP on the twist was greater than that on the slide for the TF-DNA interaction (Czapla et al., 2006; Cooper et al., 2008; Carvalho et al., 2014; Machado et al., 2015; Ngo et al., 2016). In addition, the average values of the probability distribution curves for both twist and slide for the mutated complex (**Figures 6C,D**) were closer to the standard values than those for the wild-type complex. In the meantime, the base pair stacking free energy calculated from the 3-mer base pair stacking free energy matrix for TGT in the mutated complex was more negative than that for TAT in the wild-type (**Table 2**). The more significant structural changes in the mutated MEIS1 complex than in the wild-type complex demonstrated a solid example that changes in the base pair stacking free energies resulted in DNA structural variation and altered the binding affinity for TF-DNA complexes. It also indicated that the interplay between protein and DNA plays an important role in the regulation of the variation of base stacking (Koshland, 1958; Ramakers et al., 2017).

Quantitative Modeling of TF Binding Specificities Using Base Pair Stacking

Protein-DNA binding is an essential biological process that is involved in DNA replication, restriction, and modification, transcriptional regulation, etc. (Halford and Marko, 2004). Increasing efforts have been made to understand how proteins recognize specific binding sites in the genome. Rohs et al. divided protein-DNA interactions into two major categories, including base readout and shape readout (Rohs et al., 2010). Base readout refers to proteins that recognize DNA by forming specific hydrogen bonds and hydrophobic contacts with bases in the major or minor grooves (Seeman et al., 1976), and shape readout refers to proteins that recognize DNA by sequence-dependent DNA structures and deformability (Travers, 1989; Shakked et al., 1994; Koudelka et al., 2006). Although high-throughput experimental methods, such as protein-binding microarrays (Berger et al., 2006), measure the binding affinities for tens of thousands of DNA sequences *in vitro* at the same time, it still takes extensive efforts to carry out these experiments. To achieve a higher efficiency, several theoretical models for predicting the TF-DNA binding were developed based on massive experimental data (Foat et al., 2006; Zhao and Stormo, 2011; Weirauch et al., 2013; Orenstein and Shamir, 2014). A 1-mer model, one of the earlier models, predicts TF-DNA binding affinity solely based on sequence information. This model displays a high efficiency, yet a low accuracy ($R^2 < 0.8$) when compared to more sophisticated models (Mordelet et al., 2013). Recently, the 1-mer+shape model (Zhou et al., 2015) was developed, which combines the sequence with the DNA structural information, such as the minor groove widths, propeller twists, rolls, and helix twists, as input features to achieve a higher accuracy ($R^2 > 0.9$) at the cost of a significantly increased computing time. The 1-mer+shape model requires not only predictions of the structural information, which are usually obtained from resource-hogging all-atom Monte Carlo simulations (Zhou et al., 2013), but also a significant amount of time to be trained and tested.

In this work, we confirmed, with an example, that base pair stacking free energy has an unambiguous relationship with TF-DNA interactions. In this respect, base pair stacking free energy might be used as a feature to predict TF-DNA binding affinity. Here, we extended the input matrix in the 1-mer model to include the base pair stacking free energies for all 3-mers in the DNA sequence to build the 1-mer+ΔG_s model. A 10-fold cross-validation was performed on the gcPBM data for three human basic helix–loop–helix TFs, including Mad1 (Mxd1)–Max (Mad), Max–Max (Max), and c-Myc–Max (Myc)(Mordelet et al., 2013), to compare our 1-mer+ΔG_s model with both the 1-mer and 1-mer+shape models. The person correlation coefficient (PCC) obtained by the 1-mer+ΔG_s model improved markedly compared with that of the 1-mer model. The PCC values were > 0.9 for all three TFs (Mad: PCC = 0.92, Max: PCC = 0.92, Myc: PCC = 0.91). Although the 1-mer+shape model demonstrated a higher Pearson correlation coefficient, it required significantly more features and running time (**Figure 7A**). By contrast, both the number of features per nucleotide position and the running time for the 1-mer+ΔG_s model were close to those required for the traditional 1-mer model (**Figure 7B**). Therefore, the 1-mer+ΔG_s model proved to be both accurate and efficient for predicting TF-DNA binding affinities, which might be exceedingly beneficial for preliminary disease screening.

Furthermore, as seen in the 3-mer base pair stacking free energy matrix, base pair stacking is sequence-dependent and is related to the DNA structures in the TF binding site. Thus, our 1-mer+ΔG_s model is deeply associated with the DNA's sequence and structural information, and the base pair stacking free

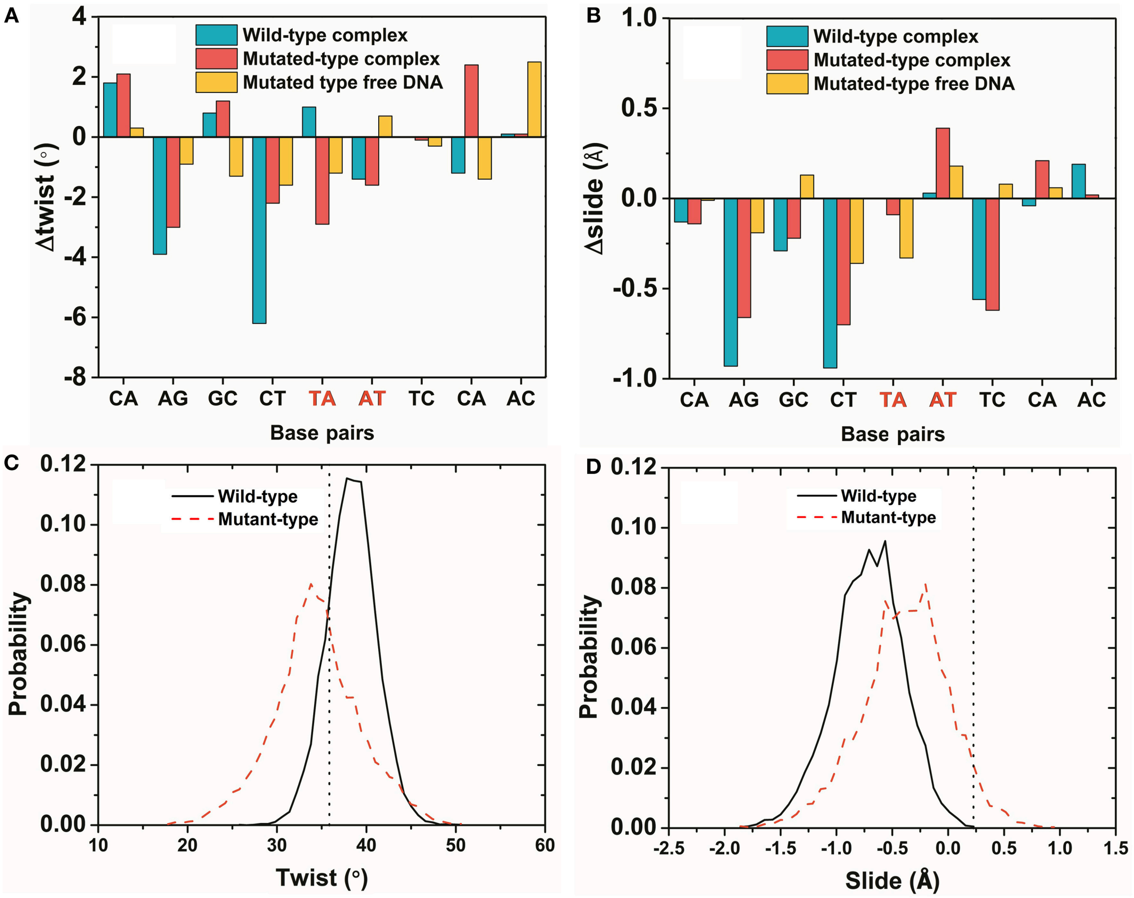

FIGURE 6 | The variation of DNA structure parameters. **(A,B)** The difference (D-values) in the twist and slide for the wild-type and mutated complexes and the mutated free DNA were obtained from the average structures of last 50 ns of the simulations. The mutated site is marked in red. **(C,D)** The probability distribution of a twist and slide at the mutation site for the wild-type and mutated complexes. The vertical dotted line is the standard value of the twist and slide. Most of the DNA structures had the same phenomenon as the twist and slide (**Figures S2**, **S3**).

TABLE 2 | Twist, slide, and base pair stacking free energy for the wild-type and mutated MEIS1 complexes.

Type	Twist[a] (°)	Slide[a] (Å)	ΔG_s[b] (kcal/mol)
Wild-type	38.3	−0.73	−1.53 (TAT)
Mutated-type	34.4	−0.53	−2.36 (TGT)
Standard Value[C]	36.0	0.23	/

[a] Twist and slide of the DNA structures obtained from the average structures. [b] ΔG_s was the calculated by the 3-mer base pair stacking free energy matrix for the 3-mer at the mutation site. [c] The standard value for the DNA structures used the data from Olson et al., 2001

energy, as a new prospect for understanding TF-DNA binding, has intrinsic connections to both base and shape readouts.

CONCLUSIONS

Base pair stacking free energy is an essential property of DNA, and it is intrinsically associated with DNA sequences and shapes. Since the sequence and shape information have already been successfully employed to understand TF-DNA interactions, base and shape readouts, base pair stacking free energy provides a new prospect in this area. In the present study, we presented an unambiguous relationship between base pair stacking free energy and TF-DNA interactions. Both a 3-mer base pair stacking free energy matrix and a 3-mer base pair stacking free energy difference matrix were constructed for establishing the matrices between the SNPs and their base pair stacking free energies. Our analyses of the variants from the GWASdb showed that mutations in SNPs with a larger $|\Delta\Delta G_s|$ had a higher probability of leading to a phenotype variation. MD simulations for the MEIS1 complexes demonstrated that the mutation in the TF-DNA binding site caused DNA structural changes and resulted in higher binding affinities. This mutation in the regulatory region was one of the four mutations with the largest $|\Delta\Delta G_s|$, which suggested that changes in the base pair stacking free energy might lead to phenotype variations via DNA structural changes. Lastly, we generated the 1-mer+ΔG_s model to apply base pair stacking free energy for predicting TF-DNA binding affinities, and it exhibited a higher accuracy than the traditional 1-mer model and a high efficiency compared to the 1-mer+shape model.

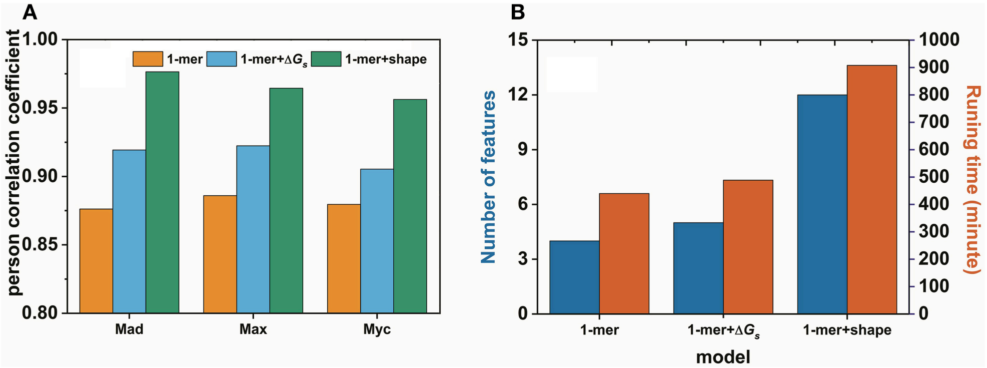

FIGURE 7 | Performance **(A)** and efficiency **(B)** comparisons for the 1-mer, 1-mer+ΔG_S, and 1-mer+shape models in predicting the binding affinities of Mad1 (Mxd1)–Max (Mad), Max–Max (Max), and c-Myc–Max (Myc) bound with DNA.

Our molecular dynamics simulation also indicated that the interplay between protein and DNA is important to the regulation of base stacking. This was in consistent with our finding that the SNPs with a larger base pair stacking free energy change led to phenotype variations. This is because a larger base pair stacking free energy change might affect the protein-DNA interactions more easily. We believe that the 3-mer base pair stacking free energy matrix and the 3-mer base pair stacking free energy difference matrix are useful for high-throughput SNP screening and for predicting TF-DNA binding affinities. Furthermore, as demonstrated in this work, proteins also play an important role in the TF-DNA interaction, and how they effect of TFs is still an open question. Improving the precision of 3-mer base pair stacking free energy is now being carried out using the QM/MM method (Warshel and Levitt, 1976).

AUTHOR CONTRIBUTIONS

GL and JG write the manuscript. XW helped and partially write the manuscript. RL and LB partially did simulation of TF binding specificities. YQ and H-YZ contributed on analysis the functions of SNPs.

FUNDING

This work was supported by the National Key R&D Program of China [Grant No. 2017YFB0203405], National Natural Science Foundation of China [No.21873034], Fundamental Research Funds for the Central Universities Project 2662018JC027) and Huazhong Agricultural University Scientific and Technological Self-innovation Foundation [Program No.201 5RC008].

ACKNOWLEDGMENTS

The authors thank Raluca Gordân from Duke University and Remo Rohs and Tianyin Zhou from the University of Southern California for providing the preprocessed gcPBM data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2018.00666/full#supplementary-material

REFERENCES

Aalberts, D. P., Parman, J. M., and Goddard, N. L. (2003). Single-strand stacking free energy from DNA beacon kinetics. *Biophys. J.* 84, 3212–3217. doi: 10.1016/S0006-3495(03)70045-9

Arkova, O., Kuznetsov, N., Fedorova, O., and Savinkova, L. (2016). A real-time study of the interaction of TBP with a TATA box-containing duplex identical to an ancestral or minor allele of human gene LEP or TPI. *J. Biomol. Struct. Dyn.* 35, 3070–3081. doi: 10.1080/07391102.2016.1241190

Baker, C. M., and Grant, G. H. (2007). Role of aromatic amino acids in protein-nucleic acid recognition. *Biopolymers* 85, 456–470. doi: 10.1002/bip.20682

Bass, J. I. F., Sahni, N., Shrestha, S., Garcia-gonzalez, A., Mori, A., Bhat, N., et al. (2015). Human gene-centered transcription factor networks for enhancers and disease variants. *Cell* 161, 661–673. doi: 10.1016/j.cell.2015.03.003

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Rd, P. W. E., and Bulyk, M. L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* 24, 1429–1435. doi: 10.1038/nbt1246

Blanchet, C., Pasi, M., Zakrzewska, K., and Lavery, R. (2011). CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucl. Acids Res.* 39, W68–W73. doi: 10.1093/nar/gkr316

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., et al. (2012). Annotation of functional variation in personal genomes using REGULOMEDB. *Genome Res.* 22, 1790–1797. doi: 10.1101/gr.137323.112

Carvalho, A. T. P., Gouveia, L., Kanna, C. R., Warmlander, S. K. T. S., Platts, J. A., and Kamerlin, S. C. L. (2014). Understanding the structural and dynamic consequences of DNA epigenetic modifications: computational insights into cytosine methylation and hydroxymethylation. *Epigenetics* 9, 1604–1612. doi: 10.4161/15592294.2014.988043

Chan, S. I., and Nelson, J. H. (1969). Proton magnetic resonance studies of ribose dinucleoside monophoshates in aqueous solution. I. The nature of the base-stacking interaction in adenylyl 3'−5')adenosine. *J. Am. Chem. Soc.* 91, 168–183. doi: 10.1021/ja01029a033

Chang, C. C., and Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Sys. Technol.* 2, 1–27. doi: 10.1145/1961189.1961199

Chang, C. E. A., Chen, W., and Gilson, M. K., (2007). Ligand configurational entropy and protein binding. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1534–1539. doi: 10.1073/pnas.0610494104

Claesen, M., D. E., Smet, F., and Suykens, J. A. K., and, D. E., Moor, B. (2014). EnsembleSVM: a library for ensemble learning using support vector machines. *J. Mach. Learn. Res.* 15, 141–145.

Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M. F., et al. (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U.S.A* 104, 19428–19433. doi: 10.1073/pnas.0709013104

Cooper, V. R., Thonhauser, T., Puzder, A., Schroder, E., Lundqvist, B. I., and Langreth, D. C. (2008). Stacking interactions and the twist of DNA. *J. Am. Chem. Soc.* 130, 1304–1308. doi: 10.1021/ja0761941

Czapla, L., Swigon, D., and Olson, W. K. (2006). Sequence-dependent effects in the cyclization of short DNA. *J. Chem. Theor. Comput.* 2, 685–695. doi: 10.1021/ct060025+

Deplancke, B., Mukhopadhyay, A., AO, W. Y., Elewa, A. M., Grove, C. A., Martinez, N. J., et al. (2006). A gene-centered C. elegans protein-DNA interaction network. *Cell* 125, 1193–1205. doi: 10.1016/j.cell.2006.04.038

Dong, S. S., Zhang, Y. J., Chen, Y. X., Yao, S., Hao, R. H., Rong, Y., et al. (2018). Comprehensive review and annotation of susceptibility SNPs associated with obesity-related traits. *Obes. Rev.* 19, 917–930. doi: 10.1111/obr.12677

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1998). A smooth particle mesh ewald method. *J. Chem. Phys.* 103, 8577–8593. doi: 10.1063/1.470117

Etheve, L., Martin, J., and Lavery, R. (2016). Protein-DNA interfaces: a molecular dynamics analysis of time-dependent recognition processes for three transcription factors. *Nucl. Acids. Res.* 44, 9990–10002. doi: 10.1093/nar/gkw841

Foat, B. C., Morozov, A. V., and Bussemaker, H. J. (2006). Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* 22, e141–e149. doi: 10.1093/bioinformatics/btl223

Friedman, R. A., and Honig, B. (1995). A free energy analysis of nucleic acid base stacking in aqueous solution. *Biophys. J.* 69, 1528–1535. doi: 10.1016/S0006-3495(95)80023-8

Garvie, C. W., and Wolberger, C. (2001). Recognition of specific DNA sequences. *Mol. Cell* 8, 937–946. doi: 10.1016/S1097-2765(01)00392-6

Geggier, S., and Vologodskii, A. (2010). Sequence dependence of DNA bending rigidity. *Proc. Natl. Acad. Sci. U.S.A.* 107, 15421–15426. doi: 10.1073/pnas.1004809107

Gordan, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., et al. (2013). Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.* 3, 1093–1104. doi: 10.1016/j.celrep.2013.03.014

Gu, C., Zhang, J., Yang, Y. I., Chen, X., Ge, H., Sun, Y., et al (2015). DNA Structural correlation in short and long ranges. *J. Phys. Chem. B* 119, 13980–13990. doi: 10.1021/acs.jpcb.5b06217

Guckian, K. M., Schweitzer, B. A., Ren, R. X., Sheils, C. J., Paris, P. L., Tahmassebi, D. C., et al. (1996). Experimental measurement of aromatic stacking affinities in the context of duplex DNA. *J. Am. Chem. Soc.* 118, 8182–8183. doi: 10.1021/ja961733f

Gutierrez-arzaluz, L., Ramirez-palma, D., Buitron-cabrera, F., Rocha-rinza, T., Cortes-guzman, F., and Peon, J. (2017). Evolution of electron density towards the conical intersection of a nucleic acid purine. *Chem. Phys. Lett.* 683, 425–430. doi: 10.1016/j.cplett.2017.03.021

Halford, S. E., and Marko, J. F. (2004). How do site-specific DNA-binding proteins find their targets? *Nucl. Acids Res.* 32, 3040–3052. doi: 10.1093/nar/gkh624

Hase, F., and Zacharias, M. (2016). Free energy analysis and mechanism of base pair stacking in nicked DNA. *Nucl. Acids Res.* 44, 7100–7108. doi: 10.1093/nar/gkw607

Helyar, S. J., Hemmer-hansen, J., Bekkevold, D., Taylor, M. I., Ogden, R., Limborg, M. T., et al. (2011). Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Mol. Ecol. Resour.* 11, 123–136. doi: 10.1111/j.1755-0998.2010.02943.x

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006). Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins* 65, 712–725. doi: 10.1002/prot.21123

Huguet, J. M., Bizarro, C. V., Forns, N., Smith, S. B., Bustamante, C., and Ritort, F. (2010). Single-molecule derivation of salt dependent base-pair free energies in DNA. *Proc. Natl. Acad. Sci. U.S.A.* 107, 15431–15436. doi: 10.1073/pnas.1001454107

Ivani, I., Dans, P. D., Noy, A., P rez, A., Faustino, I., Hospital, A., et al. (2016). Parmbsc1: a refined force field for DNA simulations. *Nat. Methods* 13, 55–58. doi: 10.1038/nmeth.3658

Jolma, A., Yin, Y. M., Nitta, K. R., Dave, K., Popov, A., Taipale, M., et al. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* 527, 384–388. doi: 10.1038/nature15518

Kal, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., et al. (1999). NAMD2: greater scalability for parallel molecular dynamics. *J. Comput. Phys.* 151, 283–312. doi: 10.1006/jcph.1999.6201

Kilchherr, F., Wachauf, C., Pelz, B., Rief, M., Zacharias, M., and Dietz, H. (2016). Single-molecule dissection of stacking forces in DNA. *Science* 353:aaf5508. doi: 10.1126/science.aaf5508

Kimchi-sarfaty, C., Oh, J. M., Kim, I. W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V., et al. (2007). A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* 315, 525–528. doi: 10.1126/science.1135308

Koshland, D. E. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. U.S.A.* 44, 98–104. doi: 10.1073/pnas.44.2.98

Koudelka, G. B., Mauro, S. A., and Ciubotaru, M. (2006). Indirect readout of DNA sequence by proteins: the roles of DNA sequence-dependent intrinsic and extrinsic forces. *Prog. Nucl. Acid Res. Mol. Biol.* 81, 143–177. doi: 10.1016/S0079-6603(06)81004-4

Li, M. J., Wang, P., Liu, X., Lim, E. L., Wang, Z., Yeager, M., et al. (2012). GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucl. Acids Res.* 40, D1047–D1054. doi: 10.1093/nar/gkr1182

Lu, Y. F., Mauger, D. M., Goldstein, D. B., Urban, T. J., Weeks, K. M., and Bradrick, S. S. (2015). IFNL3 mRNA structure is remodeled by a functional non-coding polymorphism associated with hepatitis C virus clearance. *Sci. Rep.* 5:16037. doi: 10.1038/srep16037

Luscombe, N. M., Laskowski, R. A., and Thornton, J. M. (2001). Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic Acids Res.* 29, 2860–2874. doi: 10.1093/nar/29.13.2860

Machado, A. C. D., Zhou, T. Y., Rao, S., Goel, P., Rastogi, C., Lazarovici, A., et al. (2015). Evolving insights on how cytosine methylation affects protein-DNA binding. *Brief. Funct. Genomics* 14, 61–73. doi: 10.1093/bfgp/elu040

Montclare, J. K., Sloan, L. S., and Schepartz, A. (2001). Electrostatic control of half-site spacing preferences by the cyclic AMP response element-binding protein CREB. *Nucl. Acids Res.* 29, 3311–3319. doi: 10.1093/nar/29.16.3311

Mordelet, F., Horton, J., Hartemink, A. J., Engelhardt, B. E., and Gordan, R. (2013). Stability selection for regression-based models of transcription factor-DNA binding specificity. *Bioinformatics* 29, i117–i125. doi: 10.1093/bioinformatics/btt221

Ngo, T. T. M., Yoo, J., Dai, Q., Zhang, Q., He, C., and Aksimentiev, A., et al (2016). Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nat. Commun.* 7:10813. doi: 10.1038/ncomms10813

Olson, W. K., Bansal, M., Burley, S. K., Dickerson, R. E., Gerstein, M., Harvey, S. C., et al. (2001). A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.* 313, 229–237. doi: 10.1006/jmbi.200 1.4987

Orenstein, Y., and Shamir, R. (2014). A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucl. Acids Res.* 42:e63. doi: 10.1093/nar/gku117

Parker, T. M., Hohenstein, E. G., Parrish, R. M., Hud, N. V., and Sherrill, C. D. (2013). Quantum-mechanical analysis of the energetic contributions to pi stacking in nucleic acids versus rise, twist, and slide. *J. Am. Chem. Soc.* 135, 1306–1316. doi: 10.1021/ja3063309

Protozanova, E., Yakovchuk, P., and Frank-kamenetskii, M. D. (2004). Stacked-unstacked equilibrium at the nick site of DNA. *J. Mol. Biol.* 342, 775–785. doi: 10.1016/j.jmb.2004.07.075

Ramakers, L. A., Hithell, G., May, J. J., Greetham, G. M., Donaldson, P. M., Towrie, M., et al. (2017). 2D-IR Spectroscopy Shows that optimised DNA minor groove binding of hoechst33258 follows an induced fit model. *J. Phys. Chem. B* 121, 1295–1303. doi: 10.1021/acs.jpcb.7b00345

Rohs, R., Jin, X. S., West, S. M., Joshi, R., Honig, B., and Mann, R. S. (2010). Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.* 79, 233–269. doi: 10.1146/annurev-biochem-060408-091030

Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. *Nature* 461, 1248–1253. doi: 10.1038/nature08473

Sangiovanni, J. P., Sangiovanni, P. M., Sapieha, P., and De guire, V. (2017). miRNAs, single nucleotide polymorphisms (SNPs) and age-related macular degeneration (AMD). *Clin. Chem. Lab. Med.* 55, 763–775. doi: 10.1515/cclm-2016-0898

Santalucia, J. J. R., and Hicks, D. (2004). The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.* 33, 415–440. doi: 10.1146/annurev.biophys.32.110601.141800

Santos-martins, D., Forli, S., Ramos, M. J., and Olson, A. J. (2014). AutoDock4(Zn): an improved AutoDock force field for small-molecule docking to zinc metalloproteins. *J. Chem. Inf. Model.* 54, 2371–2379. doi: 10.1021/ci500209e

Seeman, N. C., Rosenberg, J. M., and Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. U.S.A.* 73, 804–808. doi: 10.1073/pnas.73.3.804

Shakked, Z., Guzikevich-guerstein, G., Frolow, F., Rabinovich, D., Joachimiak, A., and Sigler, P. B. (1994). Determinants of repressor/operator recognition from the structure of the trp operator binding site. *Nature* 368, 469–473. doi: 10.1038/368469a0

Shen, W. F., Rozenfeld, S., Kwong, A., Kom, V. E. S., L. G., Lawrence, H. J., et al. (1999). HOXA9 forms triple complexes with PBX2 and MEIS1 in myeloid cells. *Mol. Cell. Biol.* 19, 3051–3061. doi: 10.1128/MCB.19.4.3051

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-alcala, P., Dror, I., et al. (2011). Cofactor binding evokes latent differences in DNA binding specificity between hox proteins. *Cell* 147, 1270–1282. doi: 10.1016/j.cell.2011.10.053

Slattery, M., Zhou, T. Y., Yang, L., Machado, A. C. D., Gordan, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.* 39, 381–399. doi: 10.1016/j.tibs.2014.07.002

Sponer, J., Sponer, J. E., Mladek, A., Jurecka, P., Banas, P., and Otyepka, M. (2013). Nature and magnitude of aromatic base stacking in DNA and RNA: quantum chemistry, molecular mechanics, and experiment. *Biopolymers* 99, 978–988. doi: 10.1002/bip.22322

Stokkeland, I., and Stilbs, P. (1985). A multicomponent self-diffusion NMR study of aggregation of nucleotides, nucleosides, nucleic acid bases and some derivatives in aqueous solution with divalent metal ions added. *Biophys. Chem.* 22, 65–75. doi: 10.1016/0301-4622(85)80026-0

Stormo, G. D., and Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.* 11, 751–760. doi: 10.1038/nrg2845

Swaminathan, S., Ravishanker, G., Beveridge, D. L., Lavery, R., Etchebest, C., Sklenar, H., et al. (1990). Conformational and helicoidal analysis of the molecular dynamics of proteins:"Curves," dials and windows for a 50 psec dynamic trajectory of BPTI. *Proteins* 8, 179–193. doi: 10.1002/prot.340080208

Taghavi, A., Van der schoot, P., and Berryman, J. T. (2017). DNA partitions into triplets under tension in the presence of organic cations, with sequence evolutionary age predicting the stability of the triplet phase. *Q. Rev. Biophys.* 50:e15. doi: 10.1017/S0033583517000130

Travers, A. A. (1989). DNA conformation and protein binding. *Annu. Rev. Biochem.* 58, 427–452. doi: 10.1146/annurev.bi.58.070189.002235

Warshaw, M. M., and Tinoco, I. JR. (1966). Optical properties of sixteen dinucleoside phosphates. *J. Mol. Biol.* 20, 29–38. doi: 10.1016/0022-2836(66)90115-X

Warshel, A., and Levitt, M. (1976). Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* 103, 227–249. doi: 10.1016/0022-2836(76)90311-9

Weirauch, M. T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T. R., et al. (2013). Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.* 31, 126–134. doi: 10.1038/nbt.2486

Yakovchuk, P., Protozanova, E., and Frankkamenetskii, M. D. (2006). Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res.* 34:564. doi: 10.1093/nar/gkj454

Yang, L., Zhou, T., Dror, I., Mathelier, A., Wasserman, W. W., Gordan, R., et al. (2014). TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucl. Acids Res.* 42, D148–D155. doi: 10.1093/nar/gkt1087

Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-silva, D., et al. 2016., Ensembl (2016). *Nucl. Acids Res* 44, D710–D716. doi: 10.1093/nar/gkv1157

Zhao, Y., Ruan, S. X., Pandey, M., and Stormo, G. D. (2012). Improved Models for Transcription Factor Binding Site Identification Using Nonindependent Interactions. *Genetics* 191, 781–790. doi: 10.1534/genetics.112.138685

Zhao, Y., and Stormo, G. D. (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat. Biotechnol.* 29, 480–483. doi: 10.1038/nbt.1893

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R. S., et al. (2015). Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. U.S.A.* 112, 4654–4659. doi: 10.1073/pnas.1422023112

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas machado, A. C., Ghane, T., et al. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucl. Acids Res.* 41, W56–W62. doi: 10.1093/nar/gkt437

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.