BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 13 2011, pages 1758–1764
doi:10.1093/bioinformatics/btr290

Genome analysis

Advance Access publication May 6, 2011

Genome-wide DNA sequence polymorphisms facilitate nucleosome positioning in yeast

Zhiming Dai*, Xianhua Dai* and Qian Xiang
Department of Electronic, School of Information Science and Technology, Sun Yat-Sen University, Guangzhou 510006, China

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The intrinsic DNA sequence is an important determinant of nucleosome positioning. Some DNA sequence patterns can facilitate nucleosome formation, while others can inhibit nucleosome formation. Nucleosome positioning influences the overall rate of sequence evolution. However, its impacts on specific patterns of sequence evolution are still poorly understood.

Results: Here, we examined whether nucleosomal DNA and nucleosome-depleted DNA show distinct polymorphism patterns to maintain adequate nucleosome architecture on a genome scale in yeast. We found that sequence polymorphisms in nucleosomal DNA tend to facilitate nucleosome formation, whereas polymorphisms in nucleosome-depleted DNA tend to inhibit nucleosome formation, which is especially evident at nucleosome-disfavored sequences in nucleosome-free regions at both ends of genes. Sequence polymorphisms facilitating nucleosome positioning correspond to stable nucleosome positioning. These results reveal that sequence polymorphisms are under selective constraints to maintain nucleosome positioning.

Contact: zhimdai@gmail.com; issdxh@mail.sysu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 18, 2011; revised on April 11, 2011; accepted on May 2, 2011

1 INTRODUCTION

The nucleosome is the fundamental unit of eukaryotic chromatin, which consists of a histone octamer around which a stretch of 147 bp DNA is wrapped. Most of the DNA is wrapped in nucleosomes, with adjacent nucleosomes generally separated by ~20–50 bp of unwrapped linker DNA. The nucleosome limits the accessibility of nucleosomal DNA to regulatory DNA-binding proteins. Nucleosome positioning thus plays an essential role in the regulation of diverse cellular processes, including DNA replication, recombination, repair, transcription, chromosome segregation and cell division. Nucleosome positioning is determined by multiple factors, including the intrinsic DNA sequence, DNA methylation, histone variants, post-translational modifications, transcription factors and chromatin remodelers (Segal and Widom, 2009).

Sequence evolution can be under selection to facilitate specific cellular processes. There is evidence for selection on synonymous codons required for accurate translation (Akashi, 2003; Stoletzki and Eyre-Walker, 2007). Sequence patterns in exons are also under selection to facilitate correct splicing (Parmley and Hurst, 2007; Zhang *et al.*, 2008). Genome-wide analysis on nucleosome positioning in yeast has revealed that linker DNA evolves more slowly than nucleosomal DNA (Warnecke *et al.*, 2008; Washietl *et al.*, 2008). A possible explanation is that nucleosomes occlude the DNA from access to most DNA repair proteins, resulting in low repair efficiency in nucleosomal DNA (Ataian and Krebs, 2006; Thoma, 2005).

Experimental evidence indicates that DNA sequences differ greatly in their abilities to bend and twist (Widom, 2001). Consequently, some DNA sequence patterns facilitate nucleosome formation, whereas others distort nucleosome formation (Field *et al.*, 2008). The intrinsic DNA sequence is one dominant factor responsible for nucleosome positioning (Segal *et al.*, 2006). Linker DNA-based nucleosome depletion signals have been shown to be more critical for nucleosome positioning than nucleosome formation signals (Peckham *et al.*, 2007; Yuan and Liu, 2008). An alternative explanation for the slower sequence evolution rates of linker DNA is that the extent of sequence evolution in linker DNA is under selection to maintain the nucleosomal organization. Although the impact of nucleosome positioning on the overall rate of sequence evolution is clear, its impacts on specific patterns of sequence evolution remain to be elucidated.

In this study, we investigated into the relationship between nucleosome positioning and nucleotide polymorphism patterns within *Saccharomyces cerevisiae* populations. We analyzed genome-wide single nucleotide polymorphism (SNP) data in *S.cerevisiae* populations. We found distinct polymorphism biases between nucleosomal DNA and nucleosome-depleted DNA. Polymorphism patterns that facilitate nucleosome formation are preferred in nucleosomal DNA, while polymorphism patterns with nucleosome-excluding properties are favored in nucleosome-depleted DNA. We also found that nucleotide polymorphism patterns that facilitate nucleosome positioning correspond to stable nucleosome positioning.

2 METHODS

We used the genome-wide resequencing dataset of 38 aligned strains of *S.cerevisiae* that were generated by the *Saccharomyces* Genome Resequencing Project (Liti *et al.*, 2009). These strains were sampled from different ecological niches and from locations on different continents. These strains include lab, pathogenic, baking, wine, food spoilage, natural fermentation, sake, probiotic and plant isolates. These data were collected by a combination of ABI and Illumina GA (Solexa) sequencing, which detected a total of 2 125 945 high-quality individual SNPs, which were grouped into

*To whom correspondence should be addressed.

235 127 distinct sites in the reference strain S288c. There are 12 possible nucleotide polymorphism types relative to the reference strain (i.e. A → C, A → G, A → T, C → A, C → G, C → T, G → A, G → C, G → T, T → A, T → C, T → G). For each of SNP sites, there are three possible polymorphism types relative to the reference strain.

We used the genome-wide nucleosome positioning dataset of *S.cerevisiae* that were generated by Chip-Seq (Mavrich *et al.*, 2008). A total of 53 026 nucleosome locations were identified by at least three sequencing reads of >100 bp each. We also used other three independent datasets of nucleosome positioning (Lee *et al.*, 2007; Shivaswamy *et al.*, 2008; Whitehouse *et al.*, 2007) to test the robustness to choice of datasets, which contain 70 868, 49 043 and 63 026 nucleosomes, respectively (see Supplementary Table S1 for the summary of nucleosome positioning data). The reference strain is the strain on which nucleosome locations were measured. The main results in this study were tested on each of the four independent datasets of nucleosome positioning, and could be reproduced on each of the four datasets. The main conclusions in this study are robust to the choice of datasets.

For most genes, there is a wide nucleosome-free region (NFR) at their both 5′ ends and 3′ ends. These NFRs play important roles in nucleosome positioning in gene bodies (Mavrich *et al.*, 2008). We classified genome-wide nucleosome-depleted regions as 5′ end NFR, 3′ end NFR and linker DNA. The 5′ end NFR was defined as the first nucleosome-depleted region immediately upstream of transcription start site (TSS). The 3′ end NFR was defined as the first nucleosome-depleted region immediately downstream of transcription termination site. Linker DNA was defined as the nucleosome-depleted regions except 5′ end NFR and 3′ end NFR. We classified every SNP site into four categories according to its location, including nucleosomal DNA SNP, 5′ end NFR SNP, 3′ end NFR SNP and linker DNA SNP (see Supplementary Table S1 for the frequency and density of each category of SNPs).

Nucleosome fuzziness was measured by the SD of sequencing read locations that define each nucleosome position (Mavrich *et al.*, 2008). Nucleosome fuzziness is a measure of how delocalized or spread out a nucleosome position is. We calculated for each promoter or each coding region the average fuzziness of its nucleosomes. We compared genes where most nucleotide polymorphisms on promoters (or coding regions) facilitate nucleosome positioning with the other genes in terms of their average nucleosome fuzziness. Promoter region was defined as the region 455 bp upstream of TSS (the upstream region was truncated if it overlapped with neighboring genes). This threshold was chosen according to the observation that the median intergenic distance in *S.cerevisiae* is 455 bp (Kristiansson *et al.*, 2009).

We used linker DNA potential (LDP) data of every one of the 1024 possible sequences of length 5 (i.e. 5mers). Genome-wide occupancy of nucleosomes assembled on purified yeast genomic DNA was measured by Kaplan *et al.* (2009). This *in vitro* nucleosome map is determined only by the intrinsic sequence preferences of nucleosomes. For each 5mer, they calculated the reciprocal of the average *in vitro* nucleosome occupancy across all its instances in the map, where this reciprocal average occupancy is then scaled to a probability by dividing it by the sum of all such reciprocal occupancies across all 5mers. The resulting values range from 6.74×10^{-4} (for 5mer GCAGC) to 1.48×10^{-3} (for 5mer AAAAA). We termed these resulting values as LDP. The nucleosome-disfavored sequences have relatively high LDP and the nucleosome-favored sequences have relatively low LDP. The nucleosome occupancy predicted by the model using the LDP of 5mers alone shows per-base pair correlation of 0.876 with *in vivo* nucleosome occupancy data (Kaplan *et al.*, 2009). This result demonstrates that the LDP of 5mers is highly representative of the intrinsic sequence preferences of nucleosomes. High LDP values correspond to high sequence preferences of nucleosome depletion, whereas low LDP values correspond to high sequence preferences of nucleosome formation.

Gene coordinate data were downloaded from the *Saccharomyces* Genome Database (http://www.yeastgenome.org). Using this data, we distinguished coding regions from non-coding regions. The TSS data was taken from David *et al.* (2006). Transcription factor binding data was taken from Harbison *et al.* (2004), which includes the indentified exact binding sites at promoters for each transcription factor. A *P*-value cutoff of 0.001 was used to define the set of genes bound by a particular transcription factor. The dataset includes 9678 binding sites for 101 transcription factors. Using these data, we identified SNP sites that are in the transcription factor binding sites (TFBSs) for a separate analysis.

3 RESULTS

3.1 LDP of polymorphisms

We mainly used three types of datasets. First, we used the genome-wide nucleosome positioning dataset of *S.cerevisiae* in which a total of 53 026 nucleosome locations were identified (Mavrich *et al.*, 2008). Second, we used the SNP data from 38 aligned strains of *S.cerevisiae* (Liti *et al.*, 2009). A total of 2 125 945 high-quality individual SNPs relative to the reference strain S288c were identified, which were grouped into 235 127 distinct sites on the reference strain. The reference strain is the strain on which nucleosome locations were measured (Mavrich *et al.*, 2008). Assuming that nucleosome positioning should be evolutionarily conserved among strains, we could use the nucleosome positioning data on the reference strain to study the impacts of nucleosome positioning on sequence polymorphisms that are relative to the reference strain. This assumption is reasonable as only ∼10% of the nucleosome positioning shows divergence between *S.cerevisiae and S.paradoxus* (Tirosh *et al.*, 2010; Tsankov *et al.*, 2010), and nucleosome divergence should be lower among strains of *S.cerevisiae*. Third, we used LDP data of every one of the 1024 possible 5mers (Kaplan *et al.*, 2009). High LDP values correspond to high sequence preferences of nucleosome depletion, whereas low LDP values correspond to high sequence preferences of nucleosome formation (see Section 2 for details).

For one 5mer, there are three possible polymorphism types relative to the reference strain at each position (see the example 5mer AGGCT in Fig. 1). The 5mer can be converted into other 15 possible 5mers through SNP. These possible 5mers differ in their LDP values (l in Fig. 1 is for LDP values of the 15 possible 5mers). Considering all non-reference strains, we counted the number for each of three possible polymorphism types relative to the reference strain at each SNP site, respectively. For each of the 1024 5mers, we detected all its instances that cover SNP sites across the genome, and classified these instances into four categories according to their locations, including nucleosomal DNA instances, 5′ end NFR instances, 3′ end NFR instances and linker DNA instances (see Section 2). For each category of one 5mer, we summed up the above counted numbers for each of the three possible polymorphism types at each of the five positions from all its instances (C_n for nucleosomal DNA in Fig. 1, the index n represents nucleosomal DNA), and calculated the nucleotide polymorphism frequency (f_n for nucleosomal DNA in Fig. 1).

$$f(j,i) = C(j,i)/(C(1,i) + \cdots + C(4,i)) \quad (1)$$

Take the first column in f_n of the 5mer AGGCT in Figure 1 for example, considering all instances covering SNP sites, at the first position, 10% of the nucleotide polymorphisms in nucleosomal DNA are from base A to base C (from 5mer AGGCT to CGGCT), 77% of the nucleotide polymorphisms in nucleosomal DNA are

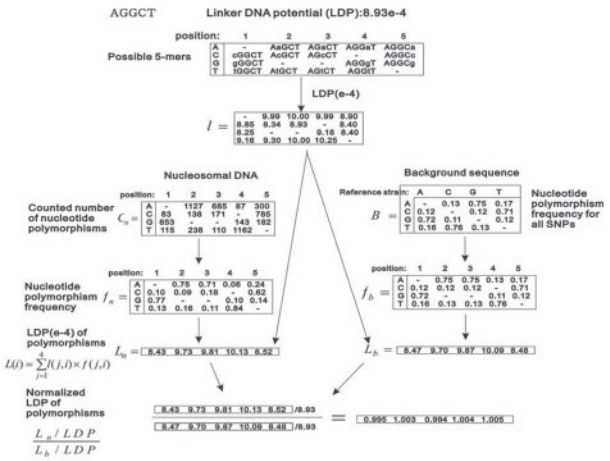

Fig. 1. The calculation of LDP of polymorphisms for one example 5mer AGGCT. This is an example for nucleosomal DNA. The same procedure is carried out for the other three categories: 5′ end NFR, 3′ end NFR and linker DNA. The 5mer AGGCT can be converted into other 15 possible 5mers through SNP. These 5mers differ in their LDP values. We counted the numbers for each polymorphism type at each of the five positions from all its instances, and calculated the nucleotide polymorphism frequency. We also calculated background nucleotide polymorphism frequency relative to the reference strain for all SNPs. The value in the background frequency matrix represents the frequency of polymorphism from the column-indicated base (reference strain) to the row-indicated base. LDP values of polymorphisms were calculated in nucleosomal DNA L_n and background L_b, respectively. L_n were normalized by LDP value of the original 5mer and L_b. These normalized LDP values are independent of the LDP value of the original 5mer before nucleotide polymorphism, and are only dependent of nucleotide polymorphism frequency values and the LDP values of the corresponding 5mers that are converted into. They represent the changing trends of nucleotide polymorphism towards high LDP values compared to those of background polymorphisms.

from base A to base G (from 5mer AGGCT to GGGCT) and 13% of the nucleotide polymorphisms in nucleosomal DNA are from base A to base T (from 5mer AGGCT to TGGCT). The nucleotide polymorphism frequency corresponds to each of the 15 possible 5mers that are converted into. We multiplied the nucleotide polymorphism frequency by the LDP value of the corresponding 5mer after polymorphism, and termed the summation of the three resulting values at each position as LDP of nucleotide polymorphism (L_n for nucleosomal DNA in Fig. 1).

$$L(i) = l(1,i) \times f(1,i) + \cdots + l(4,i) \times f(4,i) \quad (2)$$

As similar sequence patterns show similar preferences of nucleosomes (Kaplan *et al.*, 2009), the LDP values of nucleotide polymorphism should be partly dependent on the LDP value of the original 5mer before polymorphism. To control for this effect, we divided LDP values of nucleotide polymorphism by LDP value of the original 5mer (8.93×10^{-4} for the example 5mer AGGCT in Fig. 1). The resulting values represent the changing trends of LDP of nucleotide polymorphism. However, the resulting LDP values of nucleotide polymorphism might be confounded by genome-wide background nucleotide polymorphisms. First, if LDP values of nucleotide polymorphism in nucleosomal DNA for one 5mer are as high as those in nucleosome-depleted DNA, it indicates that the changing trend towards high LDP values of nucleotide polymorphism is just a general feature of genome-wide background nucleotide polymorphisms, not characteristic to nucleotide polymorphisms in nucleosome-depleted DNA. Second, for one 5mer with extremely high or low LDP value, the LDP values of 5mers converted into through SNP are more likely to be lower or higher compared with the original LDP value. To evaluate whether polymorphisms facilitate or inhibit nucleosome positioning, we should compare changing trends of LDP values of nucleotide polymorphisms to those of background polymorphisms, that is, to compare LDP of nucleotide polymorphism with those of genome-wide background nucleotide polymorphisms. Considering the two points above, we next normalized LDP of nucleotide polymorphism by genome-wide background levels (Fig. 1). We calculated nucleotide polymorphism frequency relative to the reference strain for all 2 125 945 SNPs (B in Fig. 1). Using these universal background polymorphism frequency values and the corresponding LDP values, we calculated LDP of background nucleotide polymorphism for each 5mer as the method above [Equation (2)]. We divided LDP values of nucleotide polymorphism (L_n, L_l, $L_{5\mathrm{NFR}}$, $L_{3\mathrm{NFR}}$) for each 5mer by those of background polymorphisms (L_b) (Fig. 1). These normalized LDP values are independent of the LDP value of the original 5mer before nucleotide polymorphism, and are only dependent of nucleotide polymorphism frequency values and the LDP value of the corresponding 5mers that are converted into. They represent the changing trends of LDP values of nucleotide polymorphisms compared with those of background polymorphisms. In the following analysis, we termed the normalized LDP values of nucleotide polymorphisms as LDP of polymorphisms.

3.2 Distinct LDP values of polymorphisms between nucleosomal DNA and nucleosome-depleted DNA

We examined whether the polymorphisms on non-reference strains are under constraints to maintain the nucleosome organization on the reference strain. If this is the case, the polymorphism types that facilitate nucleosome positioning should be favored, that is, polymorphism types that facilitate nucleosome formation should be favored in nucleosomal DNA, and polymorphism types that facilitate nucleosome depletion should be favored in nucleosome-depleted DNA. The instances of one 5mer should show different LDP values of polymorphisms between nucleosomal DNA and nucleosome-depleted DNA. Considering all 1024 5mers (5 LDP values of polymorphisms per 5mer, a total of 5120 values), LDP of polymorphisms in nucleosome-depleted DNA are significantly higher than those of nucleosomal DNA (Mann–Whitney U-test, $P < 10^{-9}$ for nucleosomal DNA versus linker DNA, $P < 10^{-30}$ for nucleosomal DNA versus 5′ end NFR DNA, $P < 10^{-44}$ for nucleosomal DNA versus 3′ end NFR DNA, Fig. 2A). To exclude any potentially unforeseen experimental bias in the nucleosome data we used, we performed the above analysis on another three independent datasets of nucleosome positioning (Lee *et al.*, 2007; Shivaswamy *et al.*, 2008; Whitehouse *et al.*, 2007), respectively. We found that the distinction of LDP values of polymorphisms between nucleosome-depleted DNA and nucleosomal DNA is robust to the choice of datasets (Supplementary Fig. S1). To test any potentially artifact caused by the choice of reference strain, we performed the analysis above using reference nucleotide instead of reference strain. For every SNP site, we identified the most conserved nucleotide among all strains as the reference nucleotide, and calculated

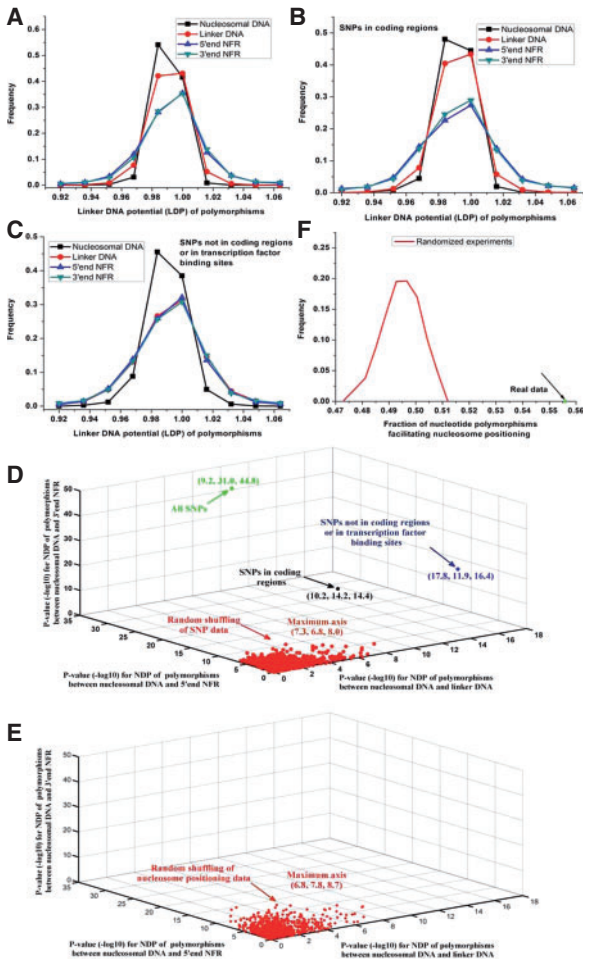

Fig. 2. Distinct sequence polymorphism patterns between nucleosomal DNA and nucleosome-depleted DNA. (**A**) Distributions of LDP values of polymorphisms ($n=5120$) of all 1024 5mers are presented for nucleosomal DNA, linker DNA, 5′end NFR and 3′ end NFR. (**B**) Same as (A), but for SNP sites that are only in coding regions. (**C**) Same as (A), but exclusion of SNP sites that are either in coding regions or in TFBSs. (**D**) Scatter comparison of *P*-values (Mann–Whitney *U*-test, −log10 transformed) that evaluate the difference in LDP values of polymorphisms between nucleosomal DNA and nucleosome-depleted DNA. *X*-axis is the *P*-values between nucleosomal DNA and linker DNA, *y*-axis is the *P*-values between nucleosomal DNA and 5′end NFR, and *z*-axis is the *P*-values between nucleosomal DNA and 3′ end NFR. The green dot is the *P*-values for (A), the black dot is the *P*-values for (B) and the blue dot is the *P*-values for (C). The red dots are the *P*-values for the 1000 randomized experiments shuffling the SNP data. (**E**) Same as (D), but for *P*-values for the 1000 randomized experiments shuffling the nucleosome positioning data. (**F**) Distributions of fraction of sequence polymorphisms that facilitates nucleosome positioning. The green dot is for the realistic SNP data, while the red line depicts the distributions for 1000 randomized experiments shuffling the SNP data.

the frequency of each of the three possible polymorphism types relative to the reference nucleotide, respectively. For SNP sites with more than one nucleotide being the most conserved, we chose one at random. We reasoned that the conserved nucleotides should correspond to the conserved nucleosome positioning, and we could test whether polymorphisms relative to the conserved nucleotides facilitate the conserved nucleosome positioning. We found that LDP of polymorphisms in nucleosome-depleted DNA are still significantly higher than those of nucleosomal DNA (Mann–Whitney *U*-test, $P<10^{-10}$ for nucleosomal DNA versus linker DNA, $P<10^{-29}$ for nucleosomal DNA versus 5′ end NFR DNA, $P<10^{-35}$ for nucleosomal DNA versus 3′ end NFR DNA, Supplementary Fig. S2).

We further tested the distinction of LDP values of polymorphisms between nucleosome-depleted DNA and nucleosomal DNA. Nucleotide polymorphisms in coding regions are under selection to maintain specific functions. Although nucleotide polymorphisms are generally assumed to occur more or less independently of selection in non-coding regions, TFBSs tend to be evolutionarily conserved to guarantee transcription factor binding. We distinguished coding regions from non-coding regions using gene coordinate data, and used the genome-wide dataset of 9678 binding sites for 101 transcription factors (Harbison *et al.*, 2004). We repeated the analysis as above for three types of SNP sites respectively: those in coding regions, those in TFBSs and the others. Note that as TFBSs lie upstream of genes, we considered SNPs in TFBSs in terms of only three categories, including nucleosomal DNA SNPs, 5′ end NFR SNPs and linker DNA SNPs, but not 3′ end NFR SNPs. In addition, though NFRs lie at both 5′ end and 3′ end of genes, they still overlap with coding regions. In other words, some SNP sites in coding region lie in NFRs. We could thus compare LDP values of polymorphisms of NFR DNA with those in nucleosomal DNA for SNP sites in coding regions. For SNPs in coding regions, LDP values of polymorphisms in nucleosome-depleted DNA are still significantly higher than those of nucleosomal DNA (Mann–Whitney *U*-test, $P<10^{-10}$ for nucleosomal DNA versus linker DNA, $P<10^{-14}$ for nucleosomal DNA versus 5′ end NFR DNA, $P<10^{-14}$ for nucleosomal DNA versus 3′ end NFR DNA, Fig. 2B). Similar results could be reproduced when excluding SNP sites that are in coding regions or in TFBSs (Mann–Whitney *U*-test, $P<10^{-17}$ for nucleosomal DNA versus linker DNA, $P<10^{-11}$ for nucleosomal DNA versus 5′ end NFR DNA, $P<10^{-16}$ for nucleosomal DNA versus 3′ end NFR DNA, Fig. 2C). However, for SNPs in TFBSs, LDP values of polymorphisms in nucleosome-depleted DNA are comparable with those of nucleosomal DNA (Mann–Whitney *U*-test, $P=0.07$ for nucleosomal DNA versus linker DNA, $P=0.26$ for nucleosomal DNA versus 5′ end NFR DNA, Supplementary Fig. S3), indicating less constraints of nucleosome positioning on nucleotide polymorphisms in TFBSs.

We repeated the above analysis by randomly shuffling SNP data and nucleosome positioning data, respectively. If nucleotide polymorphisms are not under constraints of nucleosome positioning, the random perturbation on both datasets should not weaken the distinction of LDP values of polymorphisms between nucleosome-depleted DNA and nucleosomal DNA, and the statistical significance should be as high as those in realistic SNP data. First, we shuffled the SNP data. Fixing the number of SNPs on each SNP site, for every SNP, we substituted the nucleotide in the non-reference strain with the possible three nucleotides at random according to the nucleotide in the reference strain and the background polymorphism frequency calculated above (*B* in Fig. 1). In this way, we generated a new randomized SNP map. We repeated this randomized experiment 1000 times, generating a total of 1000 randomized SNP maps. We found that the statistical differences of LDP values of polymorphisms between nucleosome-depleted DNA and nucleosomal DNA for all these randomized SNP maps are weaker than those of realistic SNP data (Fig. 2D). Second, we shuffled the nucleosome positioning data. Fixing the number

of SNP sites for each nucleosomal category (i.e. nucleosomal DNA, 5′ end NFR, 3′ end NFR and linker DNA), we randomly shuffled the category type for each SNP site. Repeating this randomized experiment 1000 times, we generated a total of 1000 randomized maps. We found that the statistical differences of LDP values of polymorphisms between nucleosome-depleted DNA and nucleosomal DNA for all these randomized SNP maps are weaker than those of realistic SNP data (Fig. 2E).

Finally, we quantitatively evaluated the constraints of nucleosome positioning on nucleotide polymorphism. For each polymorphism type at each position of one 5mer, we examined whether it facilitates nucleosome positioning or not by comparing its corresponding LDP value with background level. If $l(j,i)$ is smaller than $L_b(i)$ in nucleosomal DNA or $l(j,i)$ is higher than $L_b(i)$ in nucleosome-depleted DNA (Fig. 1), the corresponding polymorphism type was considered to facilitate nucleosome positioning. Otherwise, the corresponding polymorphism type was considered to inhibit nucleosome positioning. We counted the number of sequence polymorphisms that facilitate or inhibit nucleosome positioning, which were then subtracted by the corresponding number expected by the background polymorphism frequency. In this procedure, to control for the background polymorphism, we only considered those satisfy $f(j,i) > f_b(j,i)$, that is, the counted number must be higher than the expected number. We found that nucleotide polymorphisms facilitating nucleosome positioning are more prevalent than those inhibiting nucleosome positioning, occupying 55.6% of the total. The fraction value is 55.8 and 55.3%, when we analyze nucleosomal DNA alone and nucleosome-depleted DNA alone. These results suggest that both nucleosomal DNA and nucleosome-depleted DNA are under constraints to facilitate nucleosome positioning. To evaluate the statistical significance of the prevalence of nucleotide polymorphisms facilitating nucleosome positioning, we repeated the randomized experiment shuffling the SNP data as above 1000 times. We found that the fraction of sequence polymorphisms facilitating nucleosome positioning is lower in randomized experiments, ranging from 47.3% to 51.2% (Fig. 2F). Comparing the number of sequence polymorphisms that facilitate or inhibit nucleosome positioning in realistic data with those in randomized experiments, we found that the prevalence of nucleotide polymorphisms facilitating nucleosome positioning in realistic data is more statistically significant than those in randomized experiments ($P \approx 0$ for each of the 1000 randomized experiments, chi-square test).

Taken together, we observed distinct nucleotide polymorphism patterns between nucleosomal DNA and nucleosome-depleted DNA. Polymorphisms that facilitate nucleosome formation are favored in nucleosomal DNA, whereas polymorphisms that inhibit nucleosome formation are favored in nucleosome-depleted DNA.

3.3 Sequences with high LDP values in NFR regions have nucleosome-disfavored polymorphisms

We asked whether some specific 5mers are under strong constraints to conserve the nucleosome organization. As stated above, the LDP values of polymorphisms have been normalized by the initial LDP values and the LDP values of background polymorphisms. They represent the changing trends of nucleotide polymorphisms toward high LDP values compared with those of background polymorphisms. A high initial LDP value does not necessarily mean that the LDP value of polymorphisms will be high. As nucleosome-disfavored sequences in NFR regions have been shown to be more

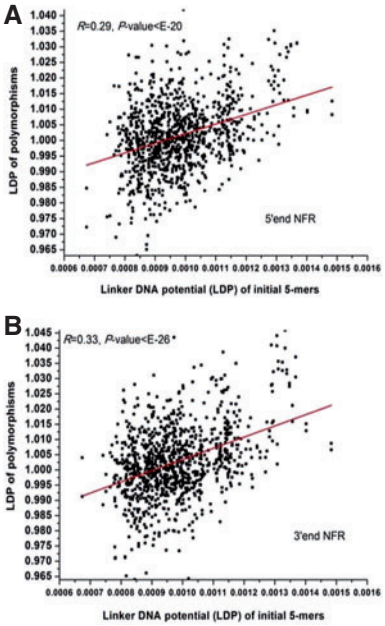

Fig. 3. Sequences with high LDP values in NFR regions have nucleosome-disfavored polymorphisms. (**A**) Shown is a scatter plot comparison between original LDP values and average LDP values of polymorphisms in 5′ end NFR. The Spearman correlation and the statistical significance of the scatter plot are indicated. Note that the LDP values of polymorphisms represent the changing trends of nucleotide polymorphism toward high LDP values compared with those of background polymorphisms. (**B**) Same as (A), but in 3′end NFR.

critical for nucleosome positioning (Peckham *et al.*, 2007; Yuan and Liu, 2008), we asked whether sequences with high LDP values in NFR regions are under strong constraints compared with those with low LDP values, having nucleosome-disfavored polymorphisms for the maintenance of nucleosome organization. As LDP values of polymorphisms of each 5mer is a vector of length 5, we used its average value (Supplementary Table S2). We found that LDP values of original 5mers show a significantly positive correlation with LDP values of polymorphisms in both NFR regions (Fig. 3). Similar positive correlation in both NFR regions can be observed in the three additional datasets of nucleosome positioning (Supplementary Figs S4–S6). However, no significant correlation between LDP values of original 5mers and LDP values of polymorphisms was observed in linker DNA (Supplementary Fig. S7A).

We next asked whether nucleosome-favored sequences in nucleosomal DNA are under strong constraints compared with nucleosome-disfavored sequences. We found no significant correlation between LDP values of original 5mers and LDP values of polymorphisms in nucleosomal DNA (Supplementary Fig. S7B). These results suggest that nucleosome-favored sequences and nucleosome-disfavored sequences in nucleosomal DNA are under comparable constraints to conserve the nucleosome organization.

3.4 Polymorphisms facilitating nucleosome positioning correspond to stable nucleosome positioning

We tested whether sequence polymorphisms that are under constraints of nucleosome positioning in turn stabilize nucleosome positioning. First, we estimated the extent of nucleosome positioning constraints on sequence polymorphism for each promoter or each

coding region. For each SNP site, we used a method similar as above to calculate its LDP values of polymorphisms (Supplementary Fig. S8). We used the average of LDP values of polymorphisms normalized by those of genome-wide background polymorphisms to represent the LDP changing trends of nucleotide polymorphism. If the average resulting value in nucleosomal DNA is smaller than 1 or that in nucleosome-depleted DNA is higher than 1, this SNP site was considered to facilitate nucleosome positioning. For simplicity, we calculated for each promoter or each coding region the fraction of its nucleotide polymorphisms that facilitate nucleosome positioning relative to background polymorphisms. We identified the top 10% promoters or coding regions with the highest fraction values as promoters or coding regions with strong nucleosome positioning constraints on nucleotide polymorphisms. Next, we used nucleosome fuzziness to represent the degree of stableness of nucleosome positioning. Nucleosome positioning is not static, and unstable nucleosome positioning should make the nucleosome spread out over a broad region along DNA. Nucleosome fuzziness is a measure of how delocalized or spread out a nucleosome position is. Low nucleosome fuzziness corresponds to stable nucleosome positioning. Nucleosome fuzziness data are available for each nucleosome on the reference strain (Mavrich *et al.*, 2008). Note that nucleosome fuzziness is not a measure of nucleosome variation among strains. We calculated for each promoter or each coding region, the average fuzziness of its nucleosomes. We found that strong nucleosome positioning constraints on nucleotide polymorphism correspond to stable nucleosome positioning (i.e. low nucleosome fuzziness) both in promoters and in coding regions (the first columns in Fig. 4A and B). This correspondence also exists when we used the top 20% promoters. Similar results can be reproduced in the three independent datasets of nucleosome positioning (the second to the fourth columns in Fig. 4A and B).

We next investigated into the relationship between nucleosome positioning constraints on sequence polymorphisms and gene expression. Nucleosome positioning in promoter regions controls the binding of transcription-related proteins to DNA, and thus it is very critical for gene expression. Variation in gene expression is linked to changes of nucleosome organization in promoter regions (Choi and Kim, 2009). We asked whether nucleosome positioning constraints on nucleotide polymorphisms in promoter regions are also linked to variation in gene expression. We used seven measures for expression variability as in a previous study (Choi and Kim, 2009), including stochastic noise (stn), responsiveness (res), stress response (str), trans variability (trv), mutational variance (muv), interstrain variation (isv) and expression divergence (div). We found that strong nucleosome positioning constraints on nucleotide polymorphisms in promoter regions correspond to low variability of gene expression (Fig. 4C). However, when controlling for nucleosome fuzziness, this correspondence disappeared ($P > 0.05$, Mann–Whitney U-test; Supplementary Fig. S9), suggesting that this correspondence may be caused by the correspondence between stable nucleosome positioning and low variability of gene expression.

4 DISCUSSION

We have shown distinct nucleotide polymorphism patterns between nucleosomal DNA and nucleosome-depleted DNA: polymorphism patterns that facilitate nucleosome formation are favored in nucleosomal regions, whereas those inhibiting nucleosome formation are biased in nucleosome-depleted regions. This observation reveals the constraints of nucleosome positioning on sequence polymorphisms. We estimated that nucleosome positioning explains ~4% to ~8% more sequence polymorphisms than is expected by chance (Fig. 2F). Moreover, the proportion of sequence polymorphisms that can be explained by nucleosome positioning is comparable between nucleosomal DNA and nucleosome-depleted DNA, indicating that the constraints of nucleosome positioning on sequence polymorphisms are comparable between these two types of DNA.

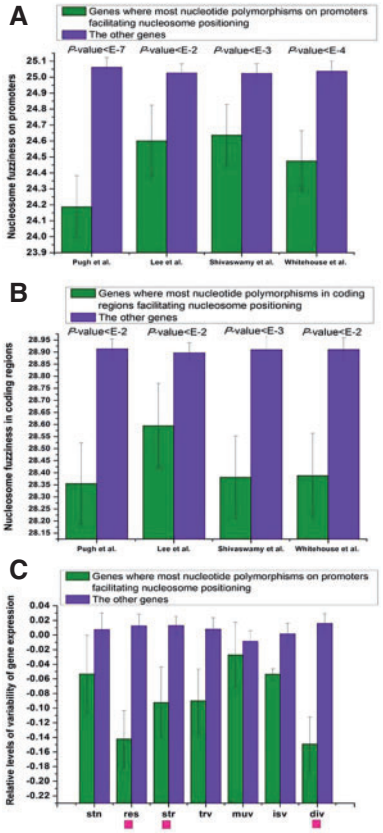

Fig. 4. Strong constraints of nucleosome positioning on nucleotide polymorphisms correspond to stable nucleosome positioning. (**A**) Average values that correspond to nucleosome fuzziness on promoters are shown for genes where most nucleotide polymorphisms on promoters facilitate nucleosome positioning and the other genes. *P*-values were from Mann–Whitney *U*-test. The results are shown for the four independent datasets of nucleosome positioning, including those of Pugh *et al.*, Lee *et al.*, Shivaswamy *et al.* and Whitehouse *et al.* Nucleosome fuzziness was used to represent nucleosome delocalization. Low nucleosome delocalization corresponds to high nucleosome stability. (**B**) Same as (A), but for coding regions. (**C**) Average values that correspond to seven measures for expression variability are shown for genes where most nucleotide polymorphisms on promoters facilitate nucleosome positioning and the other genes. The seven measures include stochastic noise (stn), responsiveness (res), stress response (str), trans variability (trv), mutational variance (muv), interstrain variation (isv) and expression divergence (div). Measures with a pink square on the bottom are statistically significant between the two gene classes ($P < 0.01$, Mann–Whitney *U*-test). Values were normalized using the function *zscore* in Matlab, such that their means are zero and SDs are one. Error bars were calculated by bootstrapping.

Nucleosome-disfavored sequences have been shown to play a more important role in nucleosome positioning compared with nucleosome-favored sequences (Peckham *et al.*, 2007; Yuan and Liu, 2008). Specifically, nucleosome-disfavored sequences in NFRs set a barrier for nucleosome positioning in gene bodies (Mavrich *et al.*, 2008). Indeed, we have shown that nucleosome-disfavored sequences in NFR regions are under strong constraints to have nucleosome-disfavored polymorphisms, facilitating nucleosome depletion. On the other hand, in nucleosomal DNA, nucleosome-favored and nucleosome-disfavored sequences are under similar constraints to have nucleosome-favored polymorphisms. These results demonstrate the different constraints of nucleosome positioning on nucleotide polymorphisms.

The intrinsic genomic sequence is an important determinant of nucleosome positioning. Nucleosome architecture in promoter regions is essential to control gene expression. Gene expression divergence in yeast has been shown to be coupled to evolution of DNA-encoded nucleosome organization (Field *et al.*, 2009). We have shown that strong constraints of nucleosome positioning on sequence polymorphisms correspond to stable nucleosome positioning and low variability of gene expression (Fig. 4). We speculated that stable nucleosome positioning imposes strong constraints on sequence polymorphisms, keeping its stable positioning. Genes having stable nucleosome positioning in promoter regions should have relatively constant chromatin structure. As variation of gene expression should require the alternation of chromatin structure in promoter regions, genes having stable nucleosome positioning in promoter regions have low variability of gene expression.

Although nucleosome positioning is not static, a considerable fraction of nucleosomes are well-positioned (Lee *et al.*, 2007). These well-positioned nucleosomes could better limit the accessibility of DNA regulatory elements compared with delocalized nucleosomes, preventing inappropriate biological processes from regulatory protein stochastic binding. We found that strong constraints of nucleosome positioning on sequence polymorphisms correspond to stable nucleosome positioning (i.e. well-positioned nucleosomes) and low variability of gene expression (Fig. 4). The polymorphism patterns that facilitate nucleosome positioning could help well-positioned nucleosomes protect DNA regulatory elements, leading to low variability of gene expression. During DNA-dependent biological processes (e.g. transcription), regulatory proteins (e.g. chromatin remodelers) are recruited to reposition nucleosomes to expose DNA regulatory elements by overriding the underlying nucleosome positioning DNA signals.

Sequence polymorphism is under constraints of nucleosome positioning in intergenic regions. As sequence polymorphisms in these regions are previously assumed to be independent of constraints, this finding reveals a new constraint at the DNA level in these regions. Our results give evidence for the important roles of DNA in nucleosome positioning, and also facilitate the understanding of a new selective constraint on sequence polymorphism on a genome-scale level.

ACKNOWLEDGEMENTS

We thank Yangyang Deng, Jiang Wang and Caisheng He for helpful discussions on the manuscript. We thank the three anonymous reviewers for helpful comments and suggestions on the manuscript.

Funding: National Natural Science Foundation of China (NSFC), Grant (60772132); Key project of Natural Science Foundation of Guangdong Province, Grant (8251027501000011); cultivation fund of major projects of Sun Yat-Sen University, Grant (10lgzd06); Yat-sen Innovative Talents Cultivation Program for Excellent Tutors.

Conflict of Interest: none declared.

REFERENCES

Akashi,H. (2003) Translational selection and yeast proteome evolution. *Genetics*, **164**, 1291–1303.

Ataian,Y. and Krebs,J.E. (2006) Five repair pathways in one context: chromatin modification during DNA repair. *Biochem. Cell Biol.*, **84**, 490–504.

Choi,J.K. and Kim,Y.J. (2009) Intrinsic variability of gene expression encoded in nucleosome positioning sequences. *Nat. Genet.*, **41**, 498–503.

David,L. *et al.* (2006) A high-resolution map of transcription in the yeast genome. *Proc. Natl Acad. Sci. USA*, **103**, 5320–5325.

Field,Y. *et al.* (2008) Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput. Biol.*, **4**, e1000216.

Field,Y. *et al.* (2009) Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. *Nat. Genet.*, **41**, 438–445.

Harbison,C.T. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Kaplan,N. *et al.* (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362–366.

Kristiansson,E. *et al.* (2009) Evolutionary forces act on promoter length: identification of enriched cis-regulatory elements. *Mol. Biol. Evol.*, **26**, 1299–1307.

Lee,W. *et al.* (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.*, **39**, 1235–1244.

Liti,G. *et al.* (2009) Population genomics of domestic and wild yeasts. *Nature*, **458**, 337–341.

Mavrich,T.N. *et al.* (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.*, **18**, 1073–1083.

Parmley,J.L. and Hurst,L.D. (2007) Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Mol. Biol. Evol.*, **24**, 1600–1603.

Peckham,H.E. *et al.* (2007) Nucleosome positioning signals in genomic DNA. *Genome Res.*, **17**, 1170–1177.

Segal,E. and Widom,J. (2009) What controls nucleosome positions? *Trends Genet.*, **25**, 335–343.

Segal,E. *et al.* (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.

Shivaswamy,S. *et al.* (2008) Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.*, **6**, e65.

Stoletzki,N. and Eyre-Walker,A. (2007) Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Mol. Biol. Evol.*, **24**, 374–381.

Thoma,F. (2005) Repair of UV lesions in nucleosomes–intrinsic properties and remodeling. *DNA Repair*, **4**, 855–869.

Tirosh,I. *et al.* (2010) Divergence of nucleosome positioning between two closely related yeast species: genetic basis and functional consequences. *Mol. Syst. Biol.*, **6**, 365.

Tsankov,A.M. *et al.* (2010) The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biol.*, **8**, e1000414.

Warnecke,T. *et al.* (2008) The impact of the nucleosome code on protein-coding sequence evolution in yeast. *PLoS Genet.*, **4**, e1000250.

Washietl,S. *et al.* (2008) Evolutionary footprints of nucleosome positions in yeast. *Trends Genet.*, **24**, 583–587.

Whitehouse,I. *et al.* (2007) Chromatin remodelling at promoters suppresses antisense transcription. *Nature*, **450**, 1031–1035.

Widom,J. (2001) Role of DNA sequence in nucleosome stability and dynamics. *Q. Rev. Biophys.*, **34**, 269–324.

Yuan,G.C. and Liu,J.S. (2008) Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput. Biol.*, **4**, e13.

Zhang,C. *et al.* (2008) RNA landscape of evolution for optimal exon and intron discrimination. *Proc. Natl Acad. Sci. USA*, **105**, 5797–5802.