

Impact of chromatin structure on sequence variability in the human genome

Michael Y. Tolstorukov^{1,2}, Natalia Volfovsky³, Robert M. Stephens³, and Peter J. Park^{1,2,4,*}

¹Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA

²Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA

³Advanced Biomedical Computing Center, Information Systems Program, SAIC-Frederick, National Cancer Institute at Frederick, Frederick, MD, USA

⁴HST Informatics Program at Children's Hospital Boston, Boston, Massachusetts, USA.

*To whom correspondence should be addressed

Abstract

DNA sequence variations in individual genomes within the same species give rise to different phenotypes. One mechanism in this process is the alteration of chromatin structure due to sequence variation that impacts gene regulation downstream. In this study, we compose a high-confidence collection of human indels and SNPs based on the analysis of a large set of publicly available sequencing data and investigate whether the DNA loci associated with stable nucleosome positions are protected against sequence mutations. We address how the sequence variation is reflected in the occupancy profiles of nucleosomes of different types at regulatory sequences and genome-wide. We find that indels are depleted around nucleosome positions of all considered types; SNPs, on the other hand, are enriched around the positions of bulk nucleosomes but depleted around the positions preferentially occupied by epigenetically modified nucleosomes. Such a behavior indicates an increased level of conservation for the sequences associated with epigenetically modified nucleosomes and highlights complex organization of the human chromatin.

Introduction

Growing evidence indicates that structural organization of chromatin, including the presence of regular nucleosome-positioning patterns, are crucial for faithful gene regulation¹⁻³. There is an on-going debate about the role of DNA sequence in establishing such patterns *in vivo* in different organisms⁴⁻⁷. In this regard, analysis of sequence variation associated with stable nucleosome positions that are common across a cell population can provide important clues. In particular, mutations in genomic DNA can disrupt nucleosome positioning signal encoded in DNA as well as alter the binding sites of transcription factors in the linkers. If the presence of a nucleosome at a specific location is functionally important, a mutation in that region should be excluded from the genome due to natural selection. On the other hand, the presence of a nucleosome can affect the efficiency of DNA repair or change the rate at which mutations appear in that sequence by protecting it from damaging agents^{8,9}. Therefore, the positions of stable nucleosomes are likely to be correlated with sites of alterations in density of sequence variation along the genome.

Two types of genomic sequence variation are the most relevant in this context: single nucleotide polymorphisms (SNPs) and short insertions and deletions (indels). A recent analysis of sequence variability in the yeast genome has shown that the SNP density is higher by 10-15% in the DNA fragments associated with nucleosome cores than in linkers¹⁰. An analysis of the SNP distribution in the human genome has revealed the presence of a periodic signal close to the nucleosomal length in the promoter proximal regions as well as increased SNP density in the closed chromatin enriched with nucleosomes^{11,12}. Association of both SNPs and indels with chromatin structure was recently characterized for regions around gene starts in the Medaka genome¹³. It was shown that unlike the SNP density, the density of indels is decreased within stable nucleosome positions as compared to linker DNA.

Association of sequence variation with nucleosome organization in the human genome has not been studied comprehensively yet. Earlier studies focused exclusively on SNPs¹¹; a direct comparison of the genome variability and nucleosome occupancy profile was not possible for the human genome due to the lack of genome-scale nucleosome profiles. However, recent advances in high-throughput sequencing technology have made it possible to map nucleosomes and to accurately identify sequence variants on a genome-scale in humans¹⁴⁻¹⁷.

To this end, we collected sequencing data available from the NCBI Trace Archive and composed a high-confidence non-redundant data set of SNPs and indels from 1 to 100 bp in length (see Methods for detail). These comprise data from different sequencing centers obtained for multiple unrelated genomes and thus any biases due to genome sampling are significantly reduced. We also used a recently published set of genome variations based on the analysis of 8 individual genomes for the validation of our findings¹⁷ (results for the ‘8-genome’ set are presented in Supplementary Material). Nucleosome occupancy has been profiled in the human genome for several types of epigenetically modified nucleosomes and for ‘bulk’ nucleosomes not selected for any histone variant or modification^{14,15}. Based on these data, we have recently identified with high resolution the stable positions for bulk nucleosomes and for the nucleosomes containing the H2A.Z histone variant and the histone H3 tri-methylated at lysine 4 (Ref 7). The H2A.Z and H3K4me3 nucleosomes are associated with transcriptional activation and are enriched at gene starts, while the bulk nucleosomes are distributed throughout the genome, making the combination of these sets well-suited for our analysis.

These data allow us to address the question of how sequence variations are distributed relative to nucleosome positions on a genome-scale. For the first time, we consider epigenetically modified and bulk nucleosomes separately and find that the level of sequence variation depends on the nucleosome type. We also compare patterns of sequence variability and their association with chromatin structure in different regulatory genome regions such as transcription start and end sites (TSS and TES) and splicing sites.

Results

1. Distribution of indels and SNPs around stable nucleosome positions genome-wide

Distributions of the genome variation instances around stable nucleosome positions follow different patterns for indels and SNPs (Figure 1). Frequencies of indels are decreased inside core sequences compared to linker DNA for all types of nucleosomes (Figure 1A). The distribution of SNP frequency, however, is more complex: the SNP frequency is higher inside bulk nucleosomes, while it does not show significant variation inside H2A.Z and H3K4me3 nucleosome sequences (Figure 1B, Supplementary Table 1).

The coordination of indel and SNP occurrences with nucleosome positioning is further illustrated in Figures 1C and 1D, where genome-wide autocorrelations of indel and SNP distributions are shown. Autocorrelation is a measure of probability to find two and more instances of a variation separated by the specific distance in the genome. Therefore, if periodic patterns exist in the distribution of the variations, it should be reflected in the autocorrelation function. Unlike the monotonic autocorrelation plot for SNPs, the plot for indels features two pronounced local maxima at 170 bp and 318 bp, which agrees well with nucleosomal repeat length in the human chromatin¹⁸.

We performed a number of further analyses to confirm our results. We checked that the frequency profiles around stable nucleosome positions share the same features when indels were split into insertions and deletions and SNPs were split into transitions and transversions and analyzed independently (Supplementary Figure 1A-D). Since nucleosomes are known to favor GC-rich sequences^{19,20} we stratified nucleosome positions by GC-content and verified that the distribution of indels and SNPs is similar for GC-rich and GC-poor sequences (Supplementary Figure 1E-G).

A 10-bp periodic pattern in the dinucleotide distribution has been found in many organisms²¹⁻²³. Since a 5-bp shift would disrupt the sequence patterns determining rotational phasing of nucleosomes while the 10-bp shift would preserve them^{21,24,25}, one may expect the 5-bp indels to be excluded from the core nucleosome sequences more than the 10-bp indels. However, we do not observe such dependence in the ratio of indel occurrences in nucleosome cores and linkers (Supplementary Figure 2A,C). We do not completely rule out the stronger exclusion of the 5-bp indels than that of the 10-bp ones because the absolute number of occurrences of indels longer than 5 bp is relatively small in our dataset. A more likely explanation, however, is that number of the sequences that exhibit the 10-bp periodic pattern is relatively low in the human genome⁷. Another observation from the analysis of the occurrences of indels of different length inside and outside nucleosome cores is that the indels of 1-bp in length are excluded from the nucleosome sequences as well as longer indels, unlike what was reported for the Medaka genome (see Supplementary Material for more details).

2. Nucleosome positioning and genome variation density at splicing sites

Intron-exon and exon-intron boundaries are among the mostly conserved genomic regions. Nucleosome positioning in these regions was recently studied^{26,27}. Our analysis reveals that the nucleosome density profiles differ at intron-exon and exon-intron junctions, while the patterns of SNP and indel frequencies are similar (Figure 2). We observe a pronounced stable nucleosome position at the exon-intron junction; at the intron-exon boundary, we see a trough in the nucleosome density, flanked by two positioned nucleosomes. This difference in the profiles is consistent with the distribution of nucleosome-disfavoring sequences, which has a stronger and wider peak at the intron-exon junctions than at the exon-intron junctions²⁶.

Distributions of SNPs and indels reach minima at both intron-exon and exon-intron junctions and feature wide troughs on the exon side of the splice site. The appearance of the trough inside exons is consistent with their coding function and the presence of the conserved regulatory elements such as exonic splicing enhancers and silencers in these regions²⁸. A noteworthy feature in the indel distribution is the presence of a narrow peak about 10 bp from the splicing site on the intron side of the intron-exon boundary. We verified that this peak is not a data processing artifact (Supplementary Figure 3). Although the exact nature of this peak is currently not clear, we note that it co-localizes with the trough in the nucleosome density.

Near splicing sites, indels are on average excluded from stable nucleosome positions, as expected from the genome-wide pattern (Figure 1). In contrast, the distribution of SNPs near splicing sites deviates from its genome-wide pattern and does not show any increase at the nucleosome position, even though more than half of the nucleosomes at the splicing sites are bulk in our data set. These observations suggest that sequence variation around splicing sites is driven by the required conservation of splicing signals rather than by the nucleosomal patterns. In other words, strong selective pressure at specific genomic locations can overcome the features in the genome variation profile imposed by nucleosome positioning.

3. Nucleosome positioning is coordinated with indel and SNP distributions at transcription start and end sites.

Comparison of the distributions of SNPs, indels, and stable nucleosome positions around transcription starts reveals two levels of coordination between genome variability and nucleosome positioning (Figure 3A). First, the overall increase in the nucleosome density around TSS is correlated with the decrease in density of both SNPs and indels in this region. It should be noted that the increase in the nucleosome density corresponds to stable positions only and may not represent the overall density of nucleosomes. Also, higher accessibility of open chromatin at TSS for nuclease digestion used to produce mono-nucleosome fragments for sequencing can contribute to the appearance of such an increase.

Second, genome variation and nucleosome profiles are negatively correlated at the level of individual nucleosome positions, especially at the nucleosome-free region and at the +1 nucleosome position downstream of TSS. The Pearson correlation between genome variations and nucleosome occupancy in the 1.5 kb region around TSS clearly indicates that both SNPs and indels are depleted at stable nucleosome positions at gene starts (Supplementary Table 2). Here, the profiles were de-trended before the calculation of correlation coefficients (see Methods for detail), and therefore our results are not influenced by the ‘overall’ coordination described above.

The exact location of the +1 position for bulk nucleosomes has been shown to depend on the transcription status of the gene¹⁵. The genes that are highly transcribed in a broad range of tissues often have their TSS encompassed by CpG islands^{29,30}, implying that the transcription status of those genes is reflected in the underlying DNA sequences. In this context, it is interesting to compare the profiles of the sequence variation and nucleosome positioning around TSS for the CpG and non-CpG genes. We focus this analysis on bulk nucleosomes because most of the epigenetic nucleosomes considered in the current study are associated with the transcriptionally active genes and the nucleosome occupancy profiles are nearly identical around TSS of CpG and non-CpG genes for these nucleosomes⁷.

Since the number of stable nucleosome positions determined from the experimental data for bulk nucleosomes is not sufficient to obtain a reliable average profile around TSS for each gene group, we treat all sequenced tags as independent nucleosome fragments (Figure 3B). This approach allows increased statistical power to detect small changes in average profiles although

it may reduce accuracy at the level of individual nucleosomes. This comparison shows that the +1 position of bulk nucleosomes is shifted downstream in CpG genes as compared to non-CpG genes, as expected for the genes with increased expression level¹⁵. For both CpG and non-CpG genes, we find that the minimum in the indel distribution aligns well with the +1 nucleosomes in their respective group (Figure 3B). This shift of the minimum in the indel profile indicates that the nucleosome positioning at TSS of CpG genes has evolved together with DNA sequence, presumably to accommodate high levels of transcription in a broad range of tissues³⁰. The distribution of SNPs does not exhibit the same level of coordination with nucleosome occupancy for CpG and non-CpG genes (Supplementary Figure 4), in accordance with the lower correlation between SNP and nucleosome density observed earlier (Supplementary Table 2).

Around TES, indel density is negatively correlated with stable nucleosome positions, while SNP density is positively correlated (Figure 3C, Supplementary Table 2). Since most nucleosomes at TES are bulk, the positive correlation between SNPs and nucleosome positions agrees with our finding that the SNP occurrence is higher on average inside the core sequences of the bulk nucleosomes (Figure 1B).

4. Different distributions of SNPs for bulk and epigenetic nucleosomes

There are two possible explanations for the differences in SNP occurrence profiles around bulk and epigenetic nucleosomes observed in our analysis. One possibility is that the sequences associated with epigenetic nucleosomes we consider here are themselves conserved to a higher extent than the positions of ‘less important’ bulk nucleosomes. Another possibility for the lower frequency of SNPs detected for epigenetic nucleosome positions, however, is simply the higher conservation of the TSS regions, where most of such nucleosomes are located. To clarify this issue, we calculated the distributions of genome variations around nucleosome positions of each type in the regions that are proximal to and distant from TSS (Figure 4).

We observe a clear decrease in SNP density for the epigenetic nucleosomes but not bulk nucleosomes in the TSS-proximal region (Figure 4A). Although the number of bulk nucleosomes in this region is small compared to that of epigenetic nucleosomes, there is no clear dip at TSS proximal regions, consistent with the first explanation above. The statistical significance of the

difference in the SNP density inside nucleosome core and linker sequences supports this conclusion, showing that only epigenetic nucleosomes at TSS are associated with the significant changes in the SNP density ($P < 0.01$, Supplementary Table 1).

Likewise, far from TSS, the epigenetic nucleosome positions are not associated with an increase in SNP rate, unlike the bulk nucleosomes that show a clear increase (Figure 4B). The relatively flat SNP density profiles for epigenetic nucleosomes in this region could be a result of a shift in the positions of such nucleosomes in the CD4+ T cells profiled here as compared to those in the germ-line cells where mutations accumulate. The fact that the positions of the bulk nucleosomes in the same TSS-distant regions are clearly reflected in the SNP density profile argues against this assumption. However, in the absence of the nucleosome positioning data for the germ-line cells, we can confidently state the difference in the SNP distribution around bulk and epigenetic nucleosomes only for the TSS proximal regions.

A possible bias in our analysis can also come from the fact that different fractions of the epigenetic and bulk nucleosomes are located in the coding regions of the genome, which are under strong selective pressure. Therefore, we directly compared the SNP densities for the epigenetic and bulk nucleosome positions occurring inside the exons of the annotated genes (Figure 4C, Supplementary Table 1). The results show that the density of SNPs around the epigenetic nucleosomes is decreased significantly as compared to the linkers, while it is increased for bulk nucleosomes ($P < 0.05$). The trend remains the same, albeit less pronounced, when non-coding regions of the genes are considered (Supplementary Figure 5). Also, in line with the results of other analyses presented in this study, the density of indels is decreased regardless of the nucleosome location relative to TSS or coding regions (Supplementary Figure 5).

The sites of increased variability in genomic sequences can also be identified by comparison of the genomes of closely related species. A recent study of the variation between three primate genomes demonstrated relevance of such an approach to the analysis of chromatin properties, revealing the correlation between the substitution rate and nucleosome occupancy³¹. Therefore, to validate our results further, we composed a set of indels and SNPs based on mapping of the

DNA sequences from four primate genomes to the human genome assembly (see Method for details). A detailed comparison of the variations in sequences from several primate genomes and in sequences from individual human genomes will be carried out elsewhere. However, using the data set based on the alignment of primate sequences, we confirmed the main results presented in this paper, showing in particular that the increase in SNP density is associated with bulk nucleosomes only (Supplementary Figures 6-8).

Discussion

Availability of the stable nucleosome positions, i.e. genomic positions preferentially occupied by the histones within a cell population, allows an investigation of the interplay between chromatin structure and genome sequence variability. Our results indicate that while indels are depleted on average in all types of nucleosomes at TSS, TES, and genome-wide, SNPs exhibits a more intricate behavior. The density of SNPs is increased in the core sequences associated with bulk but not epigenetic nucleosomes (Figure 1). Consistent with this, SNPs are negatively correlated with nucleosome occupancy at TSS and positively correlated with nucleosome occupancy at TES (Figure 3 and Supplementary Table 1), where the majority of the nucleosomes in our set are epigenetic and bulk, respectively.

The positive correlation between SNP density and nucleosome occupancy was reported previously for the Medaka and yeast genomes^{10,13}. A similar effect was recently reported for the SNPs identified as variations in the sequences of the three primate genomes including the human genome³¹. We note that the nucleosome positions used in earlier studies correspond to the bulk positions in our notation; thus, the results between the previous studies and ours are consistent. However, in the current paper, we show that this rule does not hold in a number of important cases in the human genome. For example, SNP density is negatively correlated with nucleosome occupancy in the genomic regions that are under strong selective pressure, such as exon-intron boundaries (Figure 2). At the same time, our analysis shows that the overall conservation of the regulatory regions alone cannot explain the changes in the mutation density associated with the presence of nucleosomes. Indeed, the epigenetic nucleosomes are associated with a decrease in the SNP density in the same regions where bulk nucleosomes are associated with the increase in the SNP density (Figure 4). These findings have far-reaching biological implications suggesting

that, at least for some classes of the epigenetically modified nucleosomes in the human genome, the rules of sequence-directed positioning are different and likely to be more pronounced than for bulk nucleosomes as discussed below.

We observed a weaker coordination between SNPs and indels at TSS than that reported for the Medaka genome¹³. One reason for this may be that human nucleosome occupancy data are available only for the CD4+ T-cells, while mutations that can affect genotype occur in germ-line cells. Although the clear dependence of the sequence variation frequencies on the distance from the stable nucleosome positions (Figures 1, 4) confirms the validity of our analysis, the cell-type difference may reduce the correlation between the nucleosome and genome variation profiles. Another reason may be the more complex regulation of gene expression in the human genome as compared to the Medaka genome. For example, we show that the minima in the indel profiles are shifted in CpG and non-CpG genes and correspond to the nucleosome positions +1 in each group of genes (Figure 3B). This should also contribute to the diffused minimum in the indel profile for all genes (Figure 3A).

It is interesting to consider why nucleosomal sequences in bulk are strongly depleted of one type of mutations, indels, while they are either only moderately depleted or even enriched in another type of mutations, SNPs. In general, two mechanisms are potentially responsible for the difference in the density of genome variations inside and outside nucleosomes³². One is the alteration of the mutation rate in nucleosomal DNA, e.g. due to physical interaction the nucleosomal DNA with histones^{8,10,13,33}. Another is that the DNA sequences that contain nucleosome positioning signals and/or binding sites of transcription factors are evolutionarily conserved to a higher extent than the adjacent DNA fragments³⁴. These mechanisms are not mutually exclusive and can both contribute as discussed below.

Our observation of roughly the same frequency of indels inside nucleosomes of different types (Figure 1) suggests alteration of mutation rate rather than action of purifying selection for indels. Indeed, our results provide little support for the hypothesis that the selection pressure excludes indels from nucleosomes. We did not detect a dependence of the nucleosome-to-linker ratio of the indel occurrences on indel length (Supplementary Figure 2), which would be suggestive of

this mechanism. Overall, our results indicate that the stable nucleosome positions are reflected in the indel frequency profile regardless of the local base composition or details of regulatory pathways in which a specific DNA locus is involved. This is illustrated by a shift of the nucleosome position +1 at starts of the CpG genes relative to the corresponding position at starts of the non-CpG genes in the indel frequency profile (Figure 3B). The sequence composition of the TSS proximal regions of CpG and non-CpG genes is quite different and CpG genes are actively transcribed in a broader range of cell types than the non-CpG genes³⁰, yet the nucleosome position +1 is reflected in the indel frequency profile in each of these groups.

On the other hand, the density of SNPs appears to be affected by natural selection. A single nucleotide mutation can disrupt a transcription factor binding site to interfere with regulatory pathways. It is less likely that such a mutation would significantly alter the positioning properties of a 147-bp sequence associated with a nucleosome. Furthermore, even if a mutation changes the position of a bulk nucleosome by several base pairs, this may not have any biological effect. As a result, mutations would be tolerated in the core sequence of bulk nucleosomes but would be excluded from the linkers where many transcription factors bind^{3,35,36}. In contrast, correct placement of epigenetically modified nucleosomes is important for gene regulation, and the positions preferentially occupied by these nucleosomes are likely to be conserved to the same or greater extent compared to the linker sequences. It should be emphasized that our results do not imply a complete absence of selective pressure on the bulk nucleosome sequences but rather that the pressure is stronger in linkers than in the nucleosomes of this type.

Neither do we suggest that the SNP occurrence rate is not changed in nucleosome core sequences. It is likely that the increased substitution rate is at least partly responsible for the higher density of SNPs in bulk nucleosomes as compared to the linkers. However, the substitution rate should be significantly lower in the epigenetic nucleosomes than that in bulk nucleosomes and even lower than that in linkers, so that our observations could be explained by the differences in the mutation rate. Although we cannot exclude such a model, the mechanism that would be responsible for the differences in substitution rate in the bulk and epigenetic nucleosomes does not seem feasible.

The interpretation of our results as a stronger conservation of epigenetic nucleosome positions, rather than the difference in mutation rates in the bulk and epigenetic nucleosomes is further supported by two lines of evidence. The fraction of SNPs rarely occurring in population, in particular those associated with only one genome in our data set, is higher for the epigenetic than for bulk nucleosomes (Supplementary Figure 9). This indicates a stronger selection against SNPs from the epigenetic nucleosomes. As discussed above, we also observe a clear drop in SNP density at the nucleosome positions coinciding with exon-intron boundaries (Figure 2), which is likely to result from the strong selection pressure acting on the splicing sites. Since the greater part of the nucleosomes proximal to exon-intron junctions are bulk, the anti-correlation of SNP frequency with nucleosome occupancy argues against the idea that the presence of nucleosomes of this type necessarily increases the SNP accumulation rate.

Taken together, our results suggest that a combination of purifying selection acting on biologically important sequences and the alteration of the mutation rate in nucleosomal DNA determine the pattern of sequence variation in the human genome (Figure 5). Further studies are required, however, to unambiguously prove or disprove the involvement of the above mechanisms in the evolution of nucleosome positioning sequences in the human genome. In particular, characterization of molecular mechanisms that can underlie chromatin-directed mutational bias will undoubtedly advance our understanding of the principles of genome evolution.

Methods

We identified SNPs and indels by comparing trace sequences with the sequence of the reference human genome (NCBI version 36.2). The trace data from the human libraries produced in 8 different sequencing centers (Agencourt Biosciences (ABC), Baylor College of Medicine (BCM), Celera (CRA), Cold Spring Harbor Laboratory –Watson Genome (CSHL), J. Craig Venter Institute (JCVI), Santa Cruz Genome Center (SC), Whitehead Center for Biomedical Research (WIBR), Washington University Sequencing Center (WUGSC) – referred to as sources) were downloaded from the Trace Archive (NCBI). The traces were mapped to the genomic reference DNA using GMAP³⁷ software and the high score alignments were detected by the previously described procedure³⁸. The GMAP alignments were parsed using the following parameters (i) distance of the reported variation from the end of the alignment – more than 20 bp; (ii) perfect alignment of 5 bp of flanking regions on both sides of the variations. All SNPs and indels of lengths from 1 bp to 100 bp were taken for analysis. The repeats content of the indels/SNPs loci was analyzed by comparing variations positions with the RepeatMasker annotation of Human genome. The indels that have lengths of more than 5 bp and contain mono- and dinucleotide repeats were filtered out from the final set. All variations were reported on the positive strand, so each chromosomal position represents a separate event of specific length, type (SNP, insertion or deletion) and allele. The events corresponding to SNPs and indels were clustered separately by the 5'-end for each source and for all sources together. The final data set includes 907,324 indel and 4,068,654 SNP events that were supported by at least 3 traces covering the variation from at least 2 sources (Supplementary Table 3). The histogram showing the frequency of SNP/indel events relative to the reference sequence is shown in Supplementary Figures 10A,B. The distribution of the indel lengths is shown in Supplementary Figure 10C.

We also used a recently published set¹⁷ of indels and SNPs based on analysis of 8 human genomes for comparison and validation of the results (the results of the analysis obtained for this dataset, shown in Supplementary Figures 11-13, support our findings described above). The genomic positions of the sequence variation events in the ‘8-genome’ set originally are presented in the coordinates that correspond to the human genome build hg17. The coordinates were converted to the hg18 coordinate frame with UCSC utility liftOver. Both data sets were generated by the analyses of the sequence alignments of the trace sequences to the reference

Human genome. We found that the overlaps between the genome loci from both data sets constitute ~50-60% (Supplementary Figure 14). The difference between the sets is explained by the differences in the original traces data used for variation analysis and the definition of the indels/SNPs calling parameters. The first data set was produced based on the libraries from 8 centers including the ABC, which was the only source of the second data set (list of all libraries is given in Supplementary Material). The distinction in the indel/SNP calling procedure includes different alignment tools (GMAP and ssahaSNP) applied in the analysis and different classification of the alignments (see Supplementary Material for detail).

For further validation of our results we identified sets of primate SNPs and indels following a similar procedure as that used for the human variations. To this end, the traces from Trace Archive were downloaded for chimpanzee, rhesus, orangutan, and gorilla genomes. The downloaded sequences were mapped to the reference human genome (hg18) and the resulting alignments were scanned for SNPs and indels. The obtained sets of sequence variations were filtered for simple repeats and the instances of variations supported by at least three traces and having frequencies less than 50% were retained for the analysis. This approach allowed us to compose the high-confidence sets of 5,586,505 primate SNPs and 1,059,367 primate indels which were similar in size to the corresponding sets identified from the analysis of human traces only.

Stable positions for nucleosomes bearing H3K4me3 mark (28,976 positions), H2A.Z variant (17,667 positions), and bulk nucleosomes not selected for a specific epigenetic mark or histone variant (27,486 positions) were taken from a recent analysis of ChIP-Seq and MNase-Seq data⁷. In addition we composed an aggregative set of nucleosome positions that comprises all three individual sets. In the cases of two or more positions located closer than 150 bp to each other only the position that is associated with the largest number of sequenced tags was retained. The final set included 63,554 nucleosome positions. We considered several subsets of the nucleosomes. The nucleosome positions proximal and distant to TSS were identified as those located less than 1 kb and more than 2 kb from the closest transcription start site respectively. The GC-rich and GC-poor nucleosomes were identified as those having GC-content higher than 55% and lower than 45% respectively. The sizes of each set are given in Supplementary Table 4.

The autocorrelations in the sequence variation positioning were computed for different lag distances for each chromosome separately and then averaged genome-wide accounting for chromosome sizes, similar to our previous analysis of nucleosome positions⁷. The frequency profiles of genome variations around stable nucleosome positions represent the indel and SNP occurrences normalized to the number of nucleosome positions in the corresponding set and smoothed in the 75-bp running window. The frequency profiles of the nucleosome positions and sequence variations around TSS, TES, and splicing sites were normalized to the number of genes or exons in the corresponding sets. The genes were oriented in the direction of transcription prior to averaging. Smoothing in the 100-bp running window was used for TSS, TES profiles and for nucleosome frequency in the splicing site profiles. The smaller running window of 11-bp was used in case of the genome variation profiles around splicing sites to allow for a better resolution in this case. The profiles were scaled to the interval from zero to one for easier comparison. Additional loess smoothing in 11-bp window, which does not affect positions of the major minima and maxima on the plots, was applied to reduce the jaggedness in the TSS, TES, and splicing site profiles. For the calculation of Pearson correlations between nucleosome and sequence variation frequencies and creating heatmaps, the profiles were de-trended by subtracting the same profile smoothed in the 750-bp running window.

Acknowledgements

We thank S. Sunyaev, I. Adzhubey and G. Kryukov for the helpful discussions. This project has been funded in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

References

1. Jiang, C. & Pugh, B.F. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet* **10**, 161-72 (2009).
2. Schones, D.E. & Zhao, K. Genome-wide approaches to studying chromatin modifications. *Nat Rev Genet* **9**, 179-91 (2008).
3. Segal, E. & Widom, J. From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat Rev Genet* **10**, 443-56 (2009).
4. Kaplan, N. et al. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362-6 (2009).
5. Zhang, Y. et al. Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nat Struct Mol Biol* **16**, 847-52 (2009).
6. Valouev, A. et al. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res* **18**, 1051-63 (2008).
7. Tolstorukov, M.Y., Kharchenko, P.V., Goldman, J.A., Kingston, R.E. & Park, P.J. Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes. *Genome Res* **19**, 967-77 (2009).
8. Boulikas, T. Evolutionary consequences of nonrandom damage and repair of chromatin domains. *J Mol Evol* **35**, 156-80 (1992).
9. Suter, B. & Thoma, F. DNA-repair by photolyase reveals dynamic properties of nucleosome positioning in vivo. *J Mol Biol* **319**, 395-406 (2002).
10. Washietl, S., Machne, R. & Goldman, N. Evolutionary footprints of nucleosome positions in yeast. *Trends Genet* **24**, 583-7 (2008).
11. Higasa, K. & Hayashi, K. Periodicity of SNP distribution around transcription start sites. *BMC Genomics* **7**, 66 (2006).
12. Prendergast, J.G. et al. Chromatin structure and evolution in the human genome. *BMC Evol Biol* **7**, 72 (2007).
13. Sasaki, S. et al. Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. *Science* **323**, 401-4 (2009).
14. Barski, A. et al. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-37 (2007).
15. Schones, D.E. et al. Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887-98 (2008).
16. Jin, C. et al. H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nat Genet* **41**, 941-5 (2009).
17. Kidd, J.M. et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64 (2008).
18. Lohr, D., Corden, J., Tatchell, K., Kovacic, R.T. & Van Holde, K.E. Comparative subunit structure of HeLa, yeast, and chicken erythrocyte chromatin. *Proc Natl Acad Sci U S A* **74**, 79-83 (1977).
19. Peckham, H.E. et al. Nucleosome positioning signals in genomic DNA. *Genome Res* **17**, 1170-7 (2007).
20. Kharchenko, P.V., Woo, C.J., Tolstorukov, M.Y., Kingston, R.E. & Park, P.J. Nucleosome positioning in human HOX gene clusters. *Genome Res* **18**, 1554-61 (2008).
21. Segal, E. et al. A genomic code for nucleosome positioning. *Nature* **442**, 772-8 (2006).

22. Johnson, S.M., Tan, F.J., McCullough, H.L., Riordan, D.P. & Fire, A.Z. Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. *Genome Res* **16**, 1505-16 (2006).
23. Mavrich, T.N. et al. Nucleosome organization in the *Drosophila* genome. *Nature* **453**, 358-62 (2008).
24. Trifonov, E.N. & Sussman, J.L. The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc Natl Acad Sci U S A* **77**, 3816-20 (1980).
25. Satchwell, S.C., Drew, H.R. & Travers, A.A. Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol* **191**, 659-75 (1986).
26. Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol* **16**, 990-5 (2009).
27. Tilgner, H. et al. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* **16**, 996-1001 (2009).
28. Wang, Z. & Burge, C.B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *Rna* **14**, 802-13 (2008).
29. Bird, A.P. CpG-rich islands and the function of DNA methylation. *Nature* **321**, 209-13 (1986).
30. Zhu, J., He, F., Hu, S. & Yu, J. On the nature of human housekeeping genes. *Trends Genet* **24**, 481-4 (2008).
31. Ying, H., Epps, J., Williams, R. & Huttley, G. Evidence that localized variation in primate sequence divergence arises from an influence of nucleosome placement on DNA repair. *Mol Biol Evol* **27**, 637-49.
32. Semple, C.A. & Taylor, M.S. Molecular biology. The structure of change. *Science* **323**, 347-8 (2009).
33. Kogan, S. & Trifonov, E.N. Gene splice sites correlate with nucleosome positions. *Gene* **352**, 57-62 (2005).
34. Warnecke, T., Batada, N.N. & Hurst, L.D. The impact of the nucleosome code on protein-coding sequence evolution in yeast. *PLoS Genet* **4**, e1000250 (2008).
35. Albert, I. et al. Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* **446**, 572-6 (2007).
36. Lee, W. et al. A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet* **39**, 1235-44 (2007).
37. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-75 (2005).
38. Akagi, K., Li, J., Stephens, R.M., Volfovsky, N. & Symer, D.E. Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition. *Genome Res* **18**, 869-80 (2008).
39. Kuhn, R.M. et al. The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res* **37**, D755-61 (2009).
40. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-5 (2007).

Figure legend

Figure 1. Genome-wide distributions of indel and SNP events. (A,B) Distributions of indel (A) and SNP (B) frequencies around stable nucleosome positions. Results are shown for a combined set of nucleosome positions (grey) and for individual nucleosome sets: bulk (cyan), H2A.Z (blue), and H3K4me3 (red). The frequency profiles were normalized and smoothed as described in Methods. Black dashed line at position zero corresponds to the center of nucleosome position and red dashed lines at positions ± 73 bp give reference of nucleosomal size. (C,D) Auto-correlation profiles for indel (C) and SNP (D) occurrences. Thin grey lines correspond to the initial profile calculated with one base-pair lag increments and thick red line represents loess smoothing of the initial data. Two local maxima in the indel profile corresponding to mono- and di-nucleosomal sizes are indicated with numbers.

Figure 2. Distribution of indels (red), SNPs (green), and stable nucleosome positions from combined set (black) around intron-exon (A) and exon-intron (B) boundaries. Zero position in each plot corresponds to the position of boundary. Exonic coordinates were taken from the USCS track RefGene that reports known protein-coding genes from the NCBI mRNA sequences collection (RefSeq)^{39,40}. First and last exons were excluded from the analysis. Only genes for which no alternative start site was reported we considered in this analysis (14,946 genes). The combined nucleosome set ('all nucleosomes') was used to produce this plot. The frequency profiles were calculated as described in Methods. Heatmaps shown at the bottom panels represent de-trended profiles where large-scale variations were removed.

Figure 3. Distribution of indels (red), SNPs (green), and stable nucleosome positions (black) around TSS and TES of human genes. Profiles were calculated as described in Methods. Heatmaps shown at the bottom panels represent de-trended profiles where large-scale variations were removed. (A) Profiles around TSS (position zero). The combined nucleosome set ('all nucleosomes') was used to produce this plot. Genes were oriented in the direction of transcription in such a way that the up-stream region is shown on the left and the downstream region is shown on the right of TSS. (B) Profiles shown separately for the frequencies of indels (dark red and orange lines) and bulk nucleosomes (black and cyan lines) for the subsets of genes

associated and not associated with CpG islands at TSS. Black and cyan ovals represent nucleosomes at position +1 in CpG and non-CpG genes and are shown for a nucleosome size reference. Coordinates of CpG islands were taken from USCS genome browser annotation³⁹. (C) Profiles computed around TES (position zero) for all genes. The combined nucleosome set was used.

Figure 4. Distribution of SNP frequencies around stable nucleosome positions in the regions that are proximal (A) and distant (B) to the TSS of human genes and around nucleosome positions located within coding regions (C). TSS proximal and distant nucleosome positions were identified as those located less than 1 kb and more than 2 kb from the closest TSS respectively. Coding regions are defined according to the annotation of USCS genome browser³⁹. Normalized profiles are shown for the positions from the combined nucleosome set (grey) and for the individual nucleosome sets: bulk (cyan), H2A.Z (blue), and H3K4me3 (red). Vertical dashed lines at zero and ± 73 bp give reference of the nucleosome position and size.

Figure 5. Interplay of chromatin-mediated mutation bias and selection can shape sequence variation profile (*cf.* to schematic illustration in Ref. 32). (A) Bulk and epigenetically modified nucleosomes are represented with blue and red ovals. Green and orange lines represent mutation rate of SNPs and indels respectively, and black line represents selection pressure acting on the DNA sequence. (B) The significant difference in the indel rate inside and outside nucleosomes mainly determines the indel density profile observed in the genome (orange), while SNP density profile (green) is mainly affected by selection. Our results do not exclude the possibility that natural selection can affect the distribution of indels and that alteration of the mutation rate affects the distribution of SNPs. Rather, they indicate that these mechanisms are not the major factors shaping the resulting profiles.