Hierarchical dinucleotide distribution in genome along evolution and its effect on chromatin packing

Zhicheng Cai[123], Yueying He[123], Sirui Liu[123], Yue Xue[123], Hui Quan[123], Ling Zhang[12], Yi Qin Gao[123]*

1.Beijing National Laboratory for Molecular Sciences, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China

2.Biomedical Pioneering Innovation Center (BIOPIC), Peking University, Beijing 100871, China

3.Beijing Advanced Innovation Center for Genomics (ICG), Peking University, Beijing 100871, China

*Corresponding author. E-mail: gaoyq@pku.edu.cn

Running title:

Hierarchical CpG distribution in evolution

Key words:

CpG distribution, CGI-rich/poor domain, phylogenetic tree, chromatin structure, phase separation

Abstract

Dinucleotide densities and their distribution patterns vary significantly among species. Previous studies revealed that CpG is susceptible to methylation, enriched at topologically associating domains (TADs) boundaries and its distribution along the genome correlates with chromatin compartmentalization. However, the multi-scale organizations of CpG in the linear genome, their role in chromatin organization, and how they change along the evolution are only partially understood. By comparing the CpG distribution at different genomic length scales, we quantify the difference between the CpG distributions of different species and evaluate how the hierarchical uneven CpG distribution appears in evolution. The clustering of species based on the CpG distribution is consistent with the phylogenetic tree. Interestingly, we found the CpG distribution and chromatin structure to be correlated in many different length scales, especially for mammals and avians, consistent with the mosaic CpG distribution in the genomes of these species.

Introduction

The genome composition and organization play important roles in various cellular processes. To adapt to the environment, the genomes of organisms undergo drastic mutations, leading to developmental complexity in organisms(Suzuki and Nijhout 2006; Liu et al. 2014). Many studies were conducted to explore the relationship between genome organizations and diverse phenotypes of organisms (Kim et al. 2004; Pigliucci 2010; Kulminski et al. 2013; Rinker et al. 2019).

Owing to the development of next-generation sequencing, the genomes of a large number of species are now available. They are different in sizes, karyotypes and base compositions. Particularly, it has been long known that the proportion of nucleotides varies significantly among species. For example, there is a substantial variation in average G+C contents among different species. In prokaryotes, G+C contents were reported to be positively correlated to the optimal growing temperature(Musto et al. 2004). Accordingly, thermal stability of DNA double helix was reported to be influenced by G+C contents(Yakovchuk et al. 2006), although thermophilic archaea with extreme low G+C contents also exist. Besides, the nucleotide composition also influences gene functions and regulation. The GC-rich genes in grass are usually related to basic metabolic processes and biotic stress responses(Tatarinova et al. 2010). In yeast, the AT-rich sequences are ubiquitous in promoter regions and incorporate poorly into nucleosomes, and are thus important for transcription initiation. Besides, human promoters are divided into two classes according to the CpG density(Saxonov et al. 2006). Genes with high CpG density promoters are generally expressed in more tissues than those with low CpG density promoters.

Moreover, the nucleotides and dinucleotides are not uniformly distributed along the genome. The proportions of C and G vary along the chromosomes over a large genomic length scale. The chromosomes of warm-blooded vertebrates are divided into

isochores(Bernardi et al. 1985; Bernardi 1993) which are different DNA segments with homogeneous G+C content and are separated by the sharp content transition. Isochores correlate with genome features such as gene density and replication timing(Costantini et al. 2006; Costantini and Musto 2017). Among dinucleotides, CpG distribution is the most thoroughly studied, for its biological significance. CpG is deficient in the genomes of vertebrates, probably because DNA methylation occurs predominantly at CpG. Besides, CpG is enriched at many human promoters, and high gene densities are often found in the CpG-rich regions in the human genome. Accordingly, CpG has a higher tendency to be unevenly and hierarchically distributed, i.e. CpG aggregates to form the CpG islands(CGIs), the distribution of which is also heterogeneous and correlated with gene density on the genome. Based on the distribution of CGIs, the human and mouse genomes can be divided into two types of Mb-sized domains: CGI(gene) forest domains with high CGI(gene) density and CGI(gene) prairie domains with low CGI(gene) density(Liu et al. 2018). Consistent with such multi-scale uneven CpG distributions, long-range correlations have been found to exist in the distributions of nucleotides and dinucleotides by methods such as power spectrum(Voss 1992; Buldyrev 2006), detrended fluctuation analysis(Peng et al. 1992) and wavelet transform(Audit et al. 2001; Arneodo et al. 2011). However, studies are yet to be expanded beyond limited number of model species and to cover the entire genomes but not limited DNA regions such as exons and introns. To find out how multi-scale CpG distributions change along evolution and its connection to gene regulation in different species, a comprehensive study of CpG distributions in the genomes covering species from different taxa is strongly desired.

It becomes increasingly accepted that the chromatin 3D structure plays an important role in gene regulation and cellular functions(Lieberman-Aiden et al. 2009; Bickmore and van Steensel 2013). In the meantime, many factors contributing to the chromatin structure formation have been explored. For example, it was suggested that compartments are fine-

scale structures of chromatin which are correlated with transcription states(Rowley et al. 2017). Topologically associating domains (TADs) are highly conserved structures among cell types in mammals, the boundaries of which often correspond to CTCF loop anchors(Dixon et al. 2012; Rowley and Corces 2016). Recent studies have revealed the influence of linear DNA sequence on 3D chromatin organizations. TADs boundaries often correlate with the presence of CGI, and the CGI distribution along the genome is correlated with compartmentalization in human and mouse cells(Liu et al. 2018). A unified and quantitative analysis on how the linear CpG distribution affects the 3D chromatin organizations at various length-scales is expected to shed more light on the mechanism of chromatin structure formation as revealed by Hi-C and imaging studies.

A sequence-based model(Liu et al. 2018) was proposed to explain the chromatin structure formation from the domain segregation perspective, which provides a framework for the exploration of chromatin structure formation, especially compartmentalization, in various cellular processes. To verify and generalize this model, different species besides human needs to be examined. It is also interesting to interrogate how this sequence-structure relation affects the different phenotypes of species.

In this study, we use quantitative methods to analyze the multi-scale CpG distributions on the genomes of a number of species from different taxa. We showed that using the disparity of CpG local density fluctuation one can effectively cluster species into different groups, consistent with their positions on the phylogenetic tree. The distribution of CpG is also characterized by the Multi-scale Entropy and the Pearson correlation in a scale-continuous manner. More importantly, we quantified the relation between CpG distribution, 3D chromatin organizations and gene expression activity for a number of exemplary species. Our results also show that the CpG distribution profile correlates well with the degree of a species' chromatin structural segregation and body temperature control.

Results

CpG has many special properties among the 16 dinucleotides, such as its often low density in the genome(Cooper and Gerber-Huber 1985) and richness in many human promoters in the form of CpG islands(CGIs) (Saxonov et al. 2006). We therefore first focused on the CpG dinucleotide due to these known relation to biological functions. Next, we extended our analysis to all dinucleotides.

Distribution of CpG along the genomes of different species

We compared the CpG density heterogeneity of different species based on the amplitude variation of its density fluctuation along the DNA sequence. Here the CpG density was averaged using a 1000bp window which is close to the average length of CGI. Next, we decomposed CpG density data series of different species along the DNA sequence into fluctuations at different frequencies by Hilbert-Huang transform (see method and SI). We then defined and calculated the variability (see method) of the CpG density fluctuation at the highest frequency, a large value of which corresponds to a large amplitude difference in CpG density fluctuation along the DNA sequence.

As shown in Figure 1 and Figure S4, the CpG variability can be used to effectively cluster species into different groups. Among the species investigated, birds possess the highest variability, and bacteria have the lowest. The variabilities of mammals, reptiles, fishes, plants and invertebrates are intermediate. Such an order approximately follows the phylogenetic tree. From the ranking of variability, alligator is close to mammals, and is thus closer to those of birds than to other reptiles (especially, lizards). Interestingly, according to the phylogenetic tree, birds did evolve from reptiles and alligators are more closely related to birds than to other reptiles(Crawford et al. 2012). The ranking of reptile variabilities also fits

to their positions in the phylogenetic tree. In addition, platypuses have the lowest variability among mammals, again consistent with its position in the phylogenetic tree.

As a different measure, we also analyzed the Multi-scale Entropy and Pearson correlation for CpG density of the genomes of different species to compare their CpG distribution in a scale-continuous manner. Consistently, mammals and birds have larger Multi-scale Entropy than bacteria, plants, invertebrates and fishes at large genome length scales (Figure S2). Besides, Pearson correlation of the CpG density of mammals and birds decay with the increase of genomic distance roughly in a form of power law (Figure S3), whereas in bacteria, plants, invertebrates and fishes this power law decay only persists to a short distance if it does exist. These results all indicate a more heterogeneous CpG distribution in mammals and birds than in bacteria, plants, invertebrates and fishes.

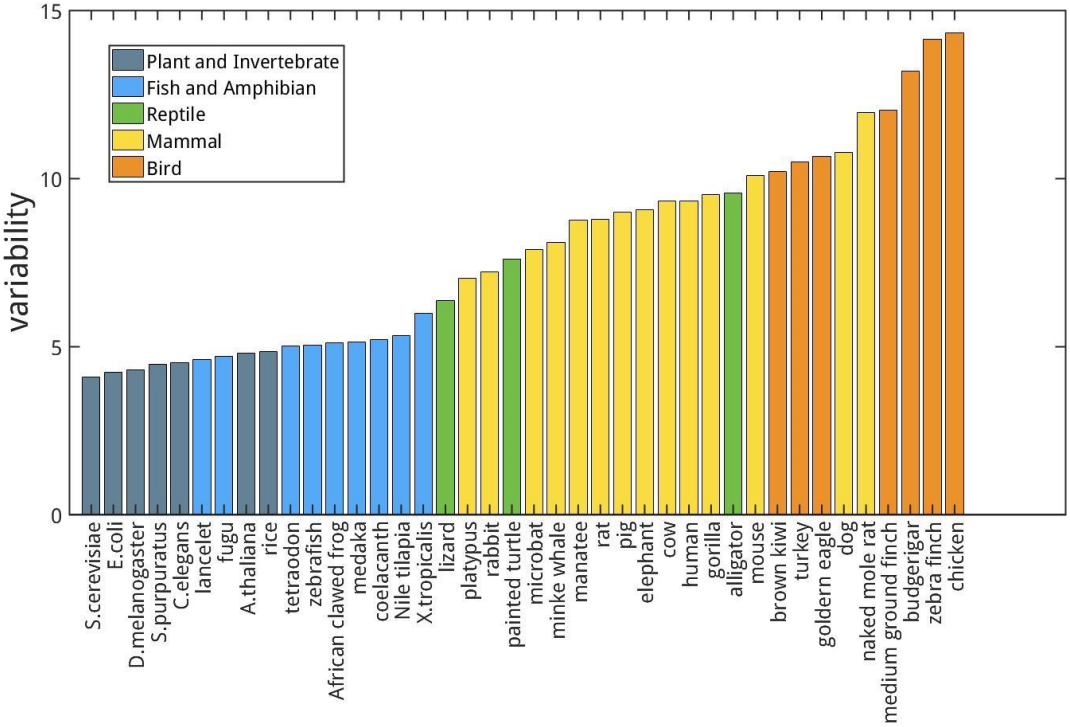


Figure 1. Variabilities of the CpG density of the longest chromosome for different species.

Division of CGI-rich and CGI-poor domains in different species

In our previous study, CGI forest and prairie domains were defined in mouse and human based on the unevenness of CGI distribution along the DNA sequence(Liu et al. 2018). These two sequential domains effectively reflect the linear segregation in the genome of not only CGI densities, but also genetic, epigenetic, and structural properties. In this study, we generalize the definition of CGI forest and prairie domains to all species termed CGI-rich and CGI-poor domains.

Since for genomes of most species traditionally defined CGIs cannot be identified because of high CpG density or a largely even distribution of CpG, a method is needed to generalize the CGI forest-prairie domain definition (see method). We make use of properties of the prairie domains of human and mouse to define the “CGI-poor clusters”, and the generalized CGI-rich and CGI-poor domains are defined so that CGI-rich domains are the longest possible domains that possess little “CGI-poor clusters”. The generalized CGI-rich domains are regions in which high CpG density loci cluster, while the generalized CGI-poor domains have low CpG density and are deficient in CGIs. The newly defined domains are in good accordance with previous definition (the CGI-rich domains overlap ratio with the CGI forest domains is 94% for human and 87% for mouse). For representative species, we evaluated their sequence properties, including the proportion of CGI-poor domain, CGI-poor domain average length, average CpG densities of the two types of domains, and the corresponding Coefficients of Variation (CV) of the CpG density.

Among the species analyzed, E. coli has very few CGI-poor domains with very low average CGI-poor domain lengths. The CpG densities of E. coli are significantly higher than multicellular eukaryotes. The E. coli genome is gene-rich and lacks noncoding elements, which is consistent with the fact that it contains very few CGI-poor domains. For eukaryotes, fishes and amphibians also have low amounts and short lengths of CGI-poor domains, but

their CpG densities in both CGI-poor and CGI-rich domains are lower than the corresponding domains in prokaryotes and invertebrates. Reptiles possess more CGI-poor domains and the CpG density levels of their CGI-poor domains are lower than fishes and amphibians, close to birds and mammals. In fact, birds and mammals are significantly different from other species, with a higher proportion of core CGI-poor domains (0.229±0.06 compared to 0.003±0.007) and a longer average CGI-poor domain length (1.76±0.63 Mb compared to 0.23±0.21 Mb), indicating an uneven CpG density distribution at the Mb level. The density fluctuation of the low CpG density regions is small, indicating that these regions have largely uniformly distributed CpG dinucleotide (with an average standard deviation of 27.6±4.3 Mb^{-1} for mammals and birds, 55.0±86.4 Mb^{-1} for other species). Birds and mammals also differ from each other in sequential properties. Mammals have longer average CGI-poor domain lengths (2.03± 0.57 Mb) than birds (1.25± 0.36 Mb). The average CpG densities for CGI-rich domains of mammals are slightly higher than those of birds, and their CGI-rich domain CpG densities vary significantly less.

More generally, hierarchical clustering yielded a dendrogram for different species (Figure 2), which is in reasonable accordance with the phylogenetic tree. It can be seen from Figure 2 that E. coli can be distinguished from eukaryotes. Among eukaryotes, plants specifically cluster together. Moreover, fishes and amphibians are also distinctly discriminated against birds and mammals. Birds and mammals roughly separate from each other except for turkey and brown kiwi of which the CpG density fluctuations at the 1kb average length in CGI-rich domains are also similar to those of mammals. Interestingly, lizard and painted turtle are grouped closely with coelacanth, which is an important link in vertebrate evolution from fishes to tetrapod.

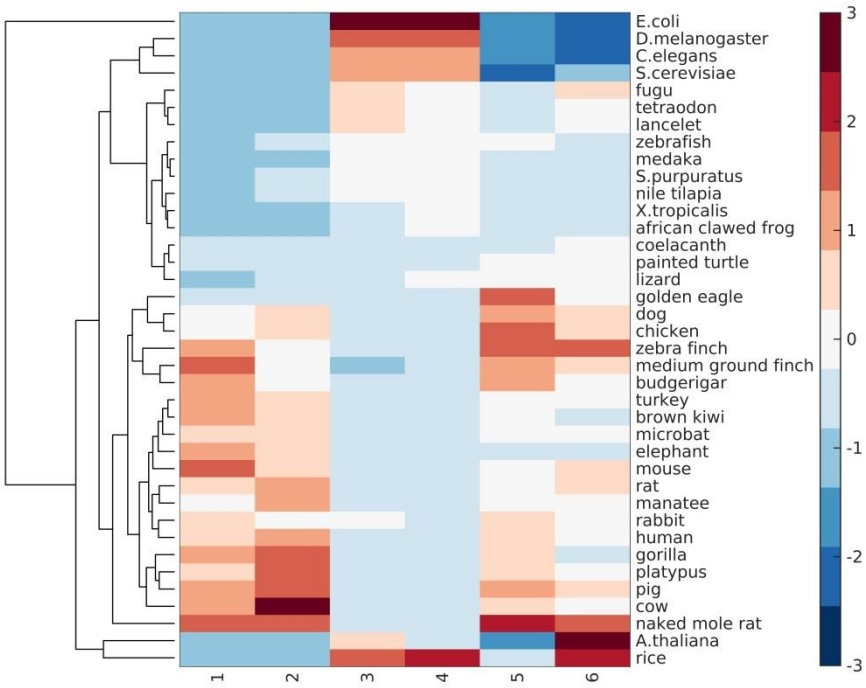


Figure 2. Dendrogram for 38 species from different classes. Six columns represent the proportions of CGI-poor domain, CGI-poor domain average length, average CpG densities of CGI-rich and poor domains, and the corresponding Coefficients of Variation (CV) of CpG density respectively. The color bar shows the column-scaled values.

Average CpG density and CpG distribution in different species

As seen in Figure 3, the scatterplot of average CpG density and CpG variability of the different species shows roughly an 'L' shape. In general, along evolution, the CpG density decreases. For example, its value for invertebrates and plants is in general much higher than that for fishes, reptiles, birds, and mammals. The CpG density difference among different invertebrates, plants, and that between invertebrates and vertebrates are all very large. Such a difference is small among different species of vertebrates: the average CpG density of mammals and birds is only slightly lower than that of fishes and reptiles. The average CpG

density being nearly constant during the evolution of mammals and birds is consistent with a detailed balance condition in CpG mutation, i.e.

$$u \times f = v \times (1 - f)$$

In the above equation, u is the mutation rate of CpG, f is the density, and v is the reproduction rate of CpG. Besides mutation, insertion of repeating sequences including CpG-rich ones into their genomes (partly from the transposition(Hwu et al. 1986)), could also be important in avoiding further CpG depletion in genomes and could explain the nearly constant average CpG density among mammals and birds.

Different from CpG density, the CpG variabilities of invertebrates and plants are all generally very similar and small. In contrast, the variabilities of mammals and birds are much higher than those of other species (Figure 3). Therefore, the genomes appear to evolve in different ways among different species. During the evolution of invertebrates and plants, the CpG density decreases with little change of the genome mosaicity and during the evolution of mammals and birds, the CpG density remains largely constant but the genome mosaicity increases. As a comparison, C+G content does not exhibit an clear trend of changes among species from different classes like CpG density (Figure 3).

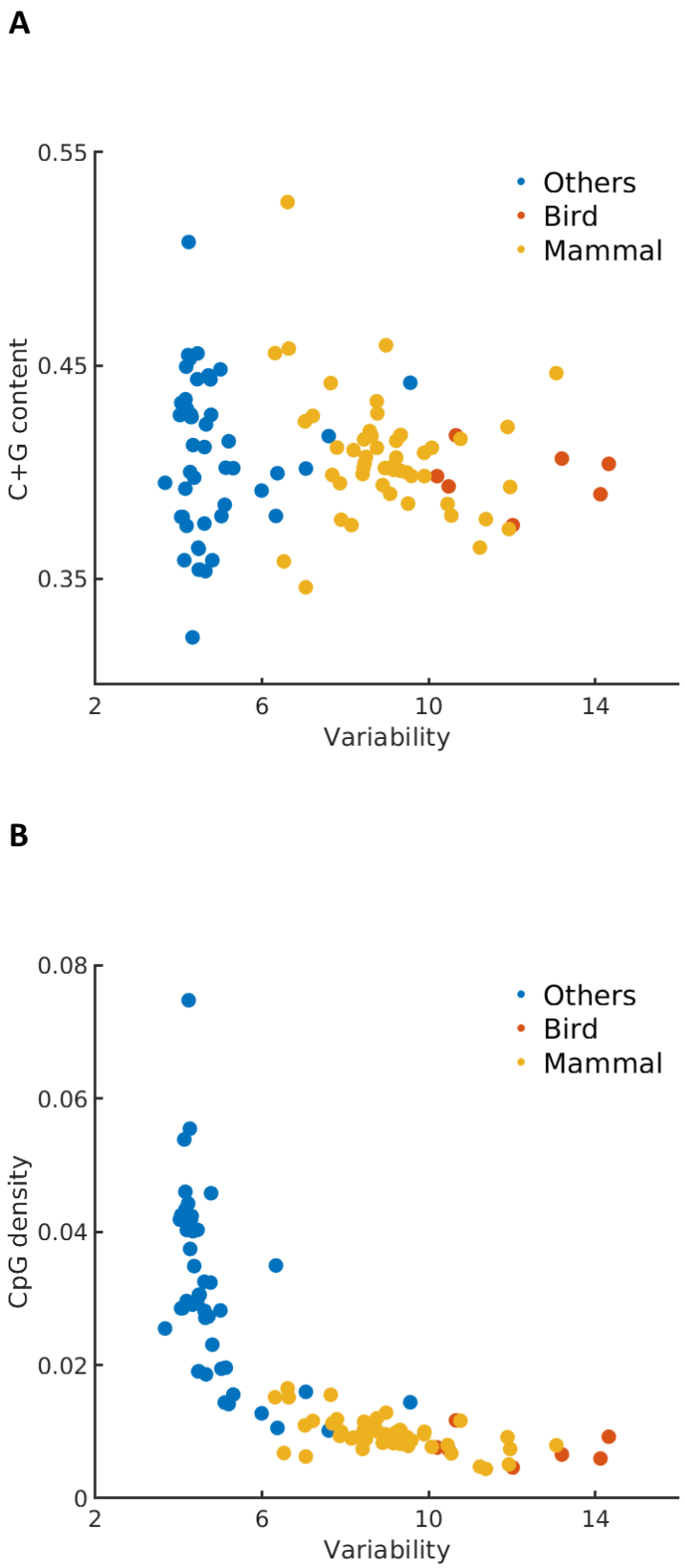


Figure 3. (A) The scatterplot for C+G content and CpG variability of various species. (B) The scatterplot for CpG density and CpG variability of various species.

CpG distribution and 3D chromatin organizations

Given that 3D structure is typically strongly influenced by 1D sequence (such as that seen in protein folding) and the DNA sequence appears to be of dramatically different properties for different species. In this section, we investigate how the CpG distribution correlates with the 3D chromatin organizations in different species. We collected the Hi-C datasets of archaea, yeast, A.thaliana, zebrafish, chicken, mouse and human. To compare the 2-dimensional Hi-C map with the CpG distribution along the linear genome, we factorized the Hi-C matrix into two 1-dimensional vectors by Nonnegative Matrix Factorization (NMF, see methods). We termed these two vectors "structure vectors" (i.e. W1 and W2), which reflect intensities of two anticorrelated Hi-C signals along the sequence. We tested the robustness of structure vectors and found structure vectors factorized by different Nonnegative Matrix Factorization methods to be similar to each other after scaling (Figure S6).

Next we performed continuous wavelet transformation for the structure vectors as well as for the CpG density along the sequence at multiple wavelet frequencies, to compare directly their fluctuations at varied length scales. Interestingly, the wavelet transform coefficients of structure vectors and CpG density of the same species largely resemble each other at many length scales (Figure 4 and S7). To quantify their similarity, we further calculated the Pearson correlation of the wavelet transform coefficients of structure vectors and CpG density. One structure vector is overall positively correlated, whereas the other is anticorrelated with CpG density (Figure 5A, B and S8), indicating that the DNA can be grouped into high and low CpG density groups in space and contacts tend to be formed within each of the two groups. In birds and mammals such as chicken, mouse and human, the correlation increases with the length scale, and the peak appears at the ~10Mb scale. In contrast, this correlation in other species such as A.thaliana and zebrafish does not increase with the length scale monotonously. Consistently, we also compared the correlation between

CpG-rich/poor domains and compartments A/B in different species and found that it is higher in human, mouse and chicken than in A.thaliana and zebrafish(Table S9). Such a result suggests that the correlation between CpG distribution and 3D chromatin organization increases along evolution, especially in large length scales, indicating the strengthened influence of globally CpG heterogeneous distribution on chromatin structure in birds and mammals than other species like fishes and plants, consistent with the increasing multi-scale heterogeneity of CpG distribution (Figures 1 and S2).

Then, we compared CpG distribution with the lengths of TADs and compartments in human(Figure 5C). At length scale of compartments, CpG density is highly correlated with the structure vectors in IMR90 cell line, whereas their correlation is weaker at the length scale corresponding to TADs. Consistently, the overlap between A/B compartments and CpG-rich/poor domains is more significant than the one between TADs and CpG-rich/poor domains (Figure S10). These observations suggest CpG density may play a more important role in compartment than in TADs formation.

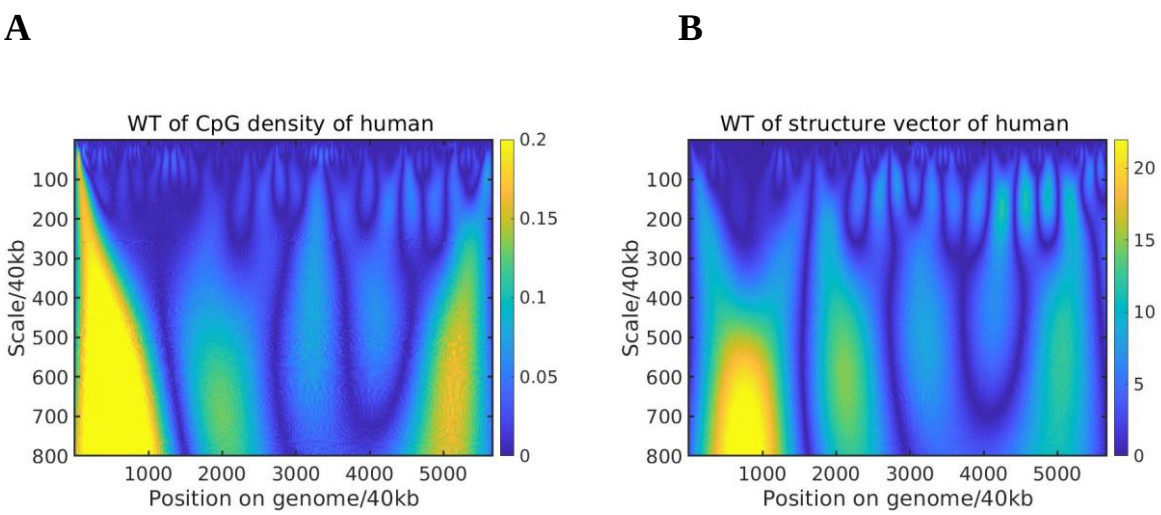


Figure 4. Wavelet Transform of (A) CpG density and (B) structure vector along the chr1 of human liver cell at different length scales.

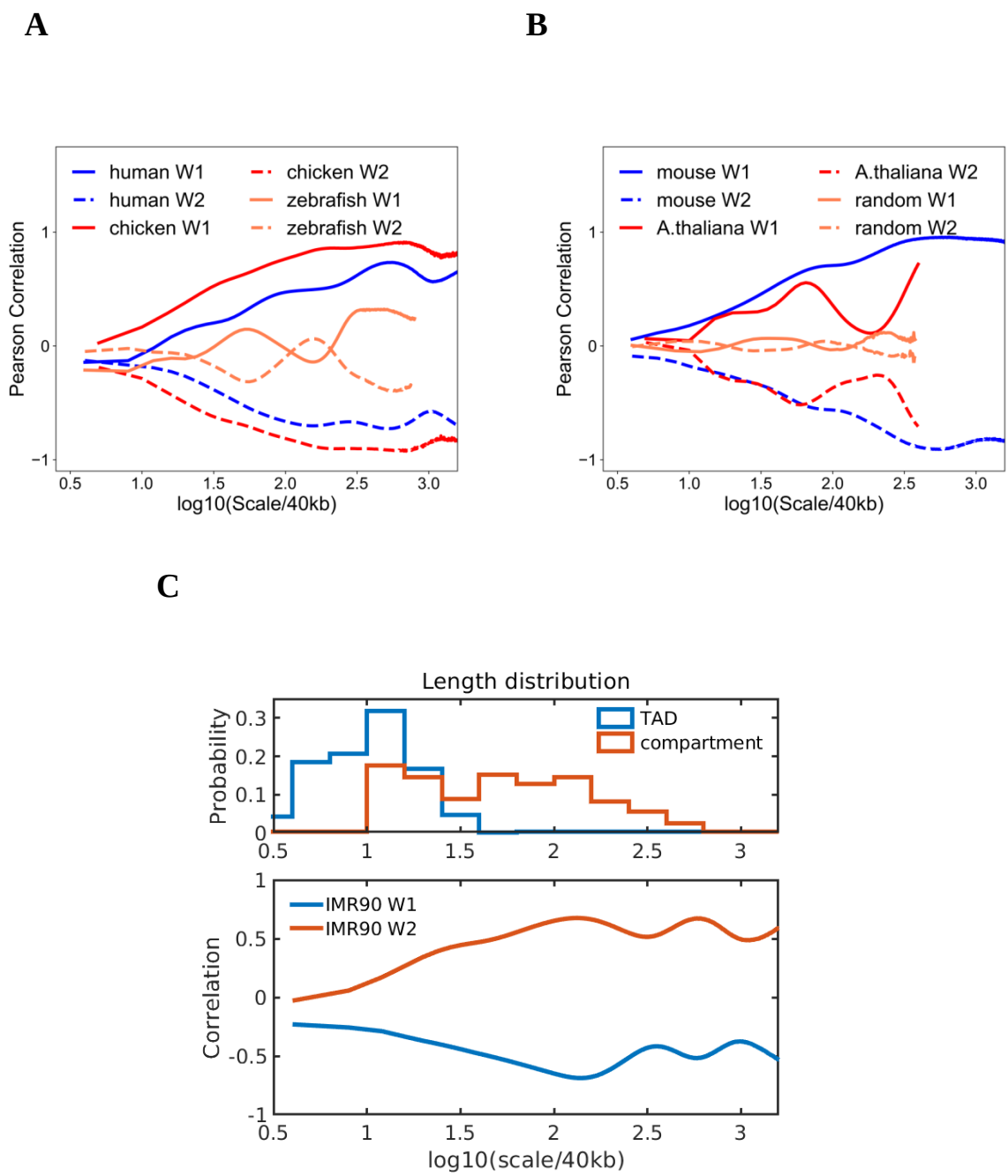


Figure 5. (A)(B) Pearson correlation of Wavelet Transform coefficient of structure vectors (i.e. W1 and W2) and CpG density of different species compared with random. X axis is the length scale of Wavelets. (C) Top: the length distribution of compartments and TADs of human respectively, Bottom: Pearson correlation of Wavelet Transform coefficient of structure vectors and CpG density of chr1 of IMR90 cell line.

Since CpG distribution and 3D chromatin organization is highly correlated in species with uneven CpG distribution, one may speculate the chromatin structure to be more segregated in mammals and birds. To examine this hypothesis, we calculated the chromatin contact frequency decay for different species. One can see from Figure S11 that contact frequency decay exponents are very similar between mouse and human, and among different cell types of human. At the short length scale, the Hi-C contact frequency for mouse and human samples decays more slowly than species such as A.thaliana, yeast and drosophila, indicating a more segregated chromatin structure of mouse and human at these length scales, consistent with the formation of interactions within the linear DNA domains such as compartments(Liu et al. 2018). In contrast, at the length scale larger than ~Mb, the chromatin contact frequency for A.thaliana, yeast and drosophila decays more slowly, indicating that long-range contacts are more likely to be observed in these species than in human and mouse. Besides, we also calculated the interaction segregation ratio of compartments A/B for different species (Table S10). The interaction segregation ratio of human and mouse is higher than other species.

CpG distribution and gene expression level

It was previously reported that the gene expression level is higher in CpG-rich than in CpG-poor regions(Liu et al. 2018). It is interesting to examine at which scale(s) such a correlation between CpG density and the gene expression hold. We found that a reasonably strong correlation exists at the scales ranging from hundreds of thousands to millions of base pairs between gene expression and CpG distribution in species such as human and zebrafish which have an uneven CpG distribution(Figure 6 and S12). However, in species with largely even CpG distribution like plants, the gene expression level varies along the sequence and is only weakly correlated with the CpG distribution(Figure 6). Furthermore, we grouped genes

by CpG density and compared the expression levels of genes of different CpG densities. The dependence of gene expression level on CpG density is seen to be different for different species (Figure S13). For example, the expression levels of zebrafish and rice genes show a more pronounced peak at median CpG densities than that of human. These results suggest that the relation between CpG density and expression level becomes stronger in evolution, with their distribution along the genome becomes more heterogeneous.

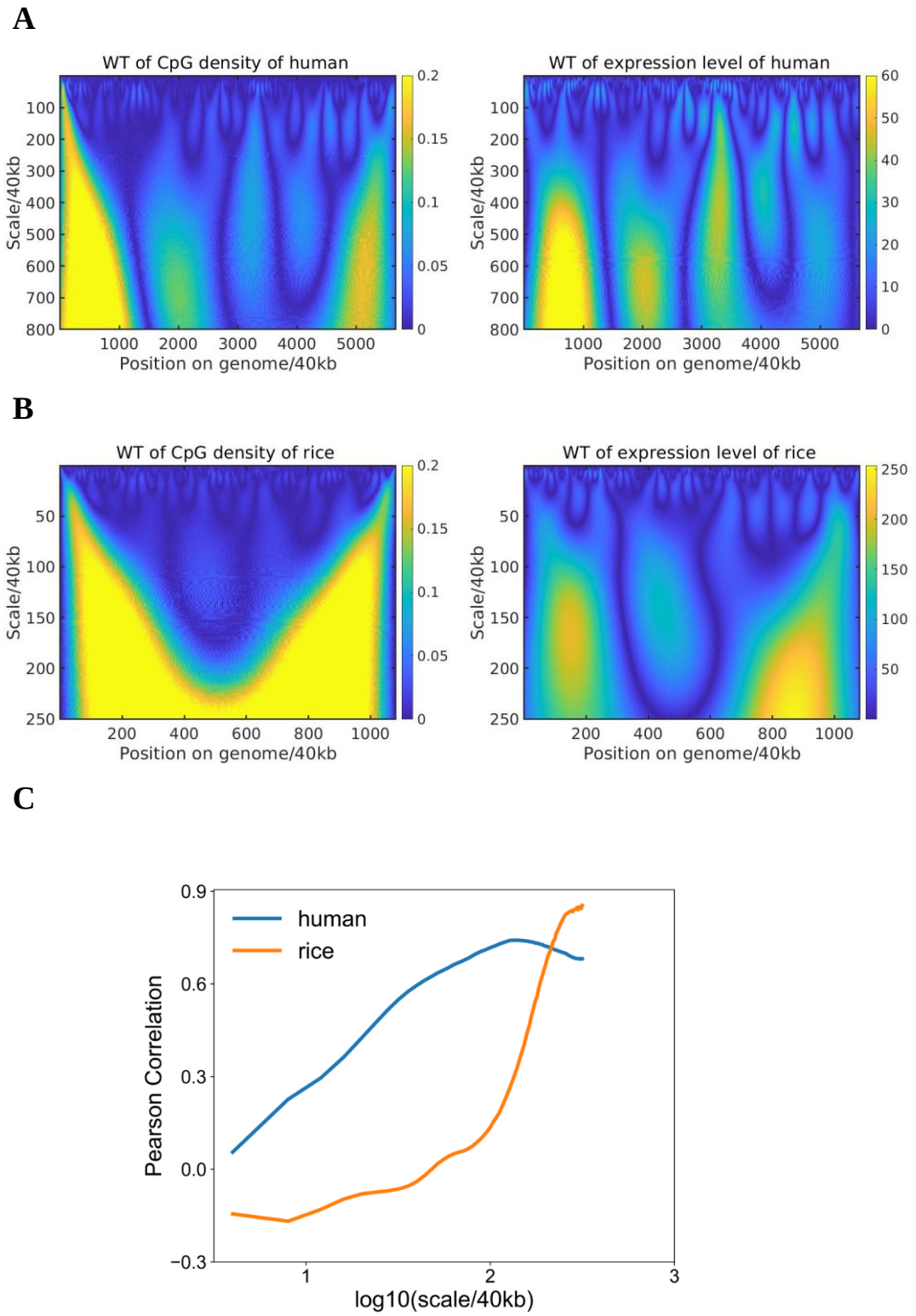


Figure 6. (A) Wavelet Transform of CpG density and expression level averaged at 40kb along chr1 of human liver cell. (B) Wavelet Transform of CpG density and expression level averaged at 40kb along chr1 of rice. (C) Pearson correlation of Wavelet transform coefficient of CpG density and expression level at multi-length scales of human and rice.

Distribution properties of other dinucleotides

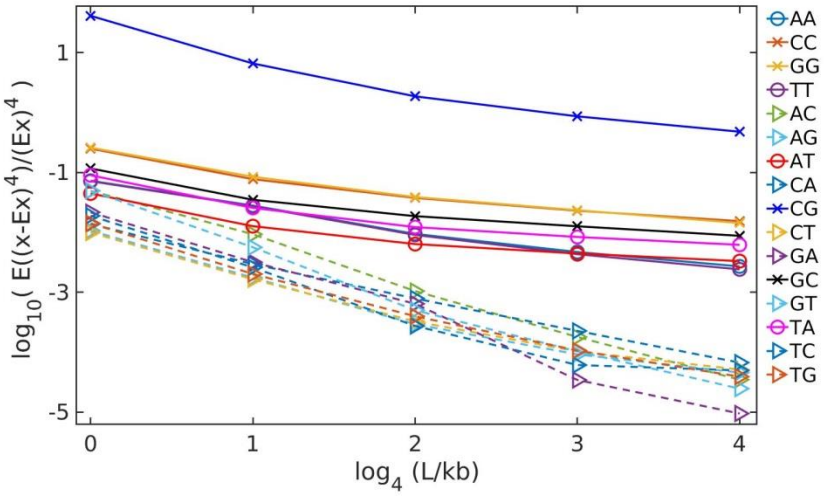
Finally, we extended our analysis to all 16 dinucleotides. Overall, we found that CpG tends to possess the highest density distribution fluctuation among the 16 different dinucleotides. For example, CpG density has the largest 4th moment at various length scales (Figure 7A), indicating its highest probability of assuming extreme values among all the 16 dinucleotides. We also calculated the variability of all dinucleotide densities. The result given in Figure 7B again shows that CpG has the largest local fluctuations.

We also found that dinucleotides composed by C and G (that is CpC, GpG, CpG, GpC) or by A and T (that is ApA, TpT, ApT, TpA) have a 4th moment that is larger than the other eight dinucleotides such as ApG which is composed of two types of nucleotides, one being A or T and the other being C or G (Figure 7). Besides CpG, distributions of other dinucleotides also become more heterogeneous changing from bacteria, invertebrates, plants, fishes to mammals and birds. For example, the Pearson correlations of CpA densities of mammals and birds but not bacteria, invertebrates, plants and fishes exhibit a long-range power law decay (Figure S3). Interestingly and consistent with the trend discussed above, among all the mammals, the density correlation of platypus decays more quickly than other mammals. The exponent of the decay function for platypus is similar to that of fishes, possibly because of its transitional position on the phylogenetic tree.

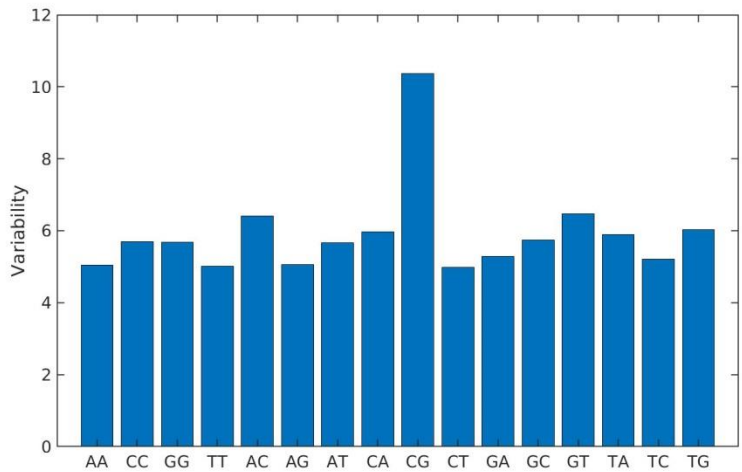
We further analyzed the density fluctuation of longer DNA sequences, for example, trinucleotides and tetra-nucleotides. The density fluctuation of these longer sequences appears to be largely determined by that of dinucleotides. For example, as shown in Figure 7C, the trinucleotides containing CpG such as CCG, CGC and CGA also show high Coefficients of Variation. Meanwhile the trinucleotides such as ACA have a high variability similar to those of AC and CA (Figure 7B and 7C). Similar phenomena are also observed for tetra-nucleotides

(Figure S14). Interestingly, it is known that codon is dominantly determined by its first dinucleotide(Hartman 1975), consistent with the observed importance of the dinucleotide composition in the DNA sequence.

A



B



C

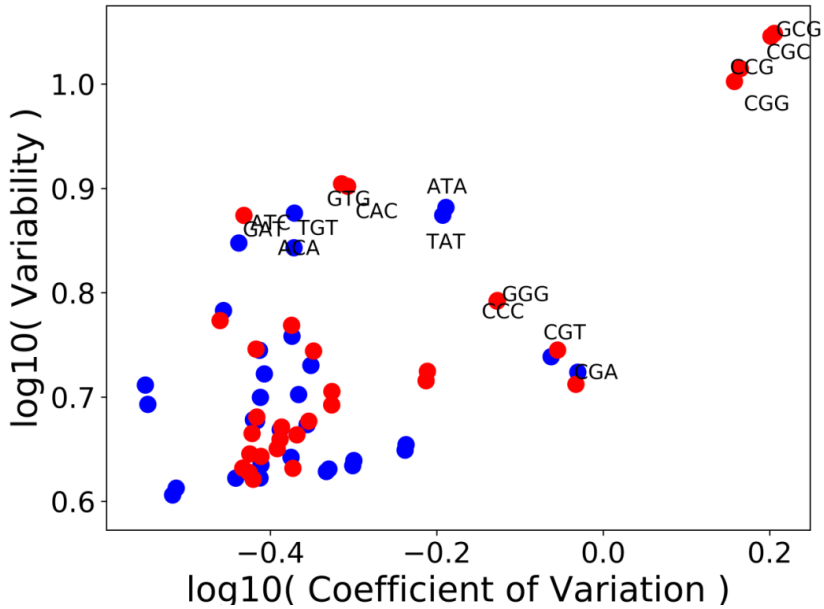


Figure 7. Distribution characteristics of dinucleotides and trinucleotides of human chr1. (A) 4th moment of 16 dinucleotides density averaged at different length scales. (B) Variability of 16 dinucleotides density averaged at 1kb. (C) The scatterplot for Variability and Coefficient of Variation of 64 trinucleotides. Blue: trinucleotides beginning with A or T; Red: trinucleotides beginning with C or G. Only trinucleotides with large variability or Coefficient of Variation are labeled.

Discussion

On one hand, chromatin structure formation and gene transcription regulation form complex networks and are influenced by many genetic and epigenetic factors. On the other hand, simple sequence information such as nucleotide and dinucleotide density distribution appears to have strong predictive power on these various biological functions. For example,

CpG dinucleotide has been shown to influence gene regulation and chromatin organization at different levels, ranging from single CpG through the methylation to CGI in promoter regions and TADs boundaries. CGI distribution along the genome was also shown to correlate with compartmentalization globally in human cells. These results show that the multi-level heterogenous CpG (among other dinucleotides) density distribution does code important information on chromatin organization and gene expression.

The sequence properties of bacteria, plants, invertebrates, fishes and reptiles are distinctly different from birds and mammals in terms of the dinucleotide distribution. The CpG density of the former shows low variabilities and a decreasing Multi-scale Entropy as a function of genomic distance. Consistently, its Pearson correlation coefficient does not exhibit a long-range power law decay. In contrast, the CpG density of birds and mammals has relatively high CpG variabilities, a Multi-scale Entropy generally increasing with the genomic length, and a wide-range power law decay, all of which indicate a more heterogeneous CpG distribution. From the perspective of CGI-rich and CGI-poor domains, birds and mammals have a larger proportion of and longer CGI-poor domains than other species. Consistently, birds' and mammals' overall average CpG density is also lower than that of other species.

Interestingly, species with shorter and smaller proportions of CGI-poor domains are more dominantly cold-blooded, living with a wide body temperature range. CGI-poor domains rarely exist in genomes of these species, but become ubiquitous in birds and mammals. Birds' and mammals' long CGI-poor domain lengths, relatively high proportions of CGI-poor domains and significant CpG density differences between these two types of domains suggest a high sequence heterogeneity. These two species are normally warm-blooded living with a narrow body temperature range.

Consistently, the values of CpG density variability also show a correlation with body temperature control of different species (Figure 1). Warm-blooded species (birds and

mammals) have higher DNA sequence variabilities than cold-blooded ones (fishes, amphibians and reptiles). Among warm-blooded species, DNA sequences of birds tend to have a higher variability than those of mammals. Furthermore, among the cold-blooded species, alligators are known as 'half warm-blooded animals' due to their maintenance of relative high body temperatures through basking(Seebacher et al. 2003; Tattersall et al. 2012). Consistently, the CpG density variability of the alligators resembles that of mammals and is very different from either turtles or lizards. To further investigate the possible relation between the body (or environmental, in the case of cold-blooded species) temperature and the CpG density variability, we also calculated and compared the variability of fishes from the tropical and polar regions (Table S4). The tropical fishes were found to have a higher sequence variability than the polar fishes. Interestingly, the brown bear appears to have a higher sequence variability than the polar bear (Table S3; Figure S15). Such a coincidence may suggest a possible relation between the living environment and the CpG distribution of the genomes of different species.

As a note, the DNA methylation level was also reported to be negatively correlated with the environment temperature for fishes and reptiles(Varriale and Bernardi 2006b; Varriale and Bernardi 2006a). It is well known that the majority of methylation occurs on cytosine and higher CpG density often results in the higher methylation level. Such an observation is consistent with our results for a high CpG density is correlated with the low variability (Figure 3). In fact, earlier analysis shows that the methylation level correlates with CpG density and its fluctuation(Liu et al. 2018).

As mentioned above, a close relation exists between the mosaic CGI and gene distributions in human and mouse genomes(Liu et al. 2018). The segregation of the CGI/gene -rich and CGI/gene-poor domains largely determines the chromatin compartmentalization, with a large number of alternating compartments A and B (Figure S16). In contrast, it is well

known that chromosomes of many plants are simply partitioned into three compartments: two compartments A near the telomere and one compartment B near the centromere(Liu et al. 2017). We calculated the gene density of rice and A.thaliana along the genome and found that their gene density correlates well with compartment partition and the gene density is higher in compartment A than in B (Figure S16). However, little correlation is seen between CpG density and compartment partition for these species as the CpG distribution is nearly uniform in plants (Figure 1 and S2). Similar gene density-compartment correlation is also observed for archaea (Figure S16). These results suggest an important role of the gene distribution along the linear genome in chromatin structure segregation and compartment formation, and suggest that the uneven CpG distribution further promoting the spatial partition of the chromosome in species such as mammals and birds but not plants. The separation of the chromatin structure into different compartments following the gene density distribution is highly conserved across different species, and indicates a connection between the CGI (and thus CpG, as well as other dinucleotides) distributions and the gene clustering along the linear genome. In fact, genes of similar functions also tend to segregate along the linear genome(Lewis 1992; Lewis 2004), consistent with a function-driven gene and DNA sequence redistribution in evolution. Since genes of similar functions tend to form spatial contacts(Belcastro et al. 2011; Ibn-Salem et al. 2017), it would be interesting to examine the cross-talk between 3D chromatin structure and DNA linear sequence in different species and in evolution.

Recently, it was proposed that a stable phase separation of genome is correlated to differentiation, senescence and diseases such as cancer(Liu et al. 2018). A high mosaicity of a genome appears to correspond to stable differentiation. Consistent with this theory, with a low genome mosaicity and a largely even CpG distribution (see discussion above and Figure 1), plants are prone to reprogramming and dedifferentiate. For example, plants can generate calluses in response to stresses, many of which are totipotent(C. Steward et al. 1958; Ikeuchi

et al. 2013). Moreover, fishes and reptiles can remain growing and developing throughout their lives. For example, lizards can regenerate tails(Baranowitz et al. 1979). Among mammals, whales, which have a low CpG variability (Figure 1), are the largest mammals on the earth and famous for the longevity and ability of suppressing cancer(Caulin and Maley 2011). Similarly, elephants are also known for their longevity, resistance to aging and cancer, and indeterminate body size. In contrast, cells of birds and mammals with a high sequence mosaicity are difficult to reprogram(Surani 2012), and more prone to cancer than invertebrates and plants(Albuquerque et al. 2018).

Finally, our analysis indicates that CpG density, consistent with its highly uneven distribution, is a better function and structure indicator than C+G content. CpG density but not C+G content (Figure 1 and 3) can cluster species corresponding to their positions in the phylogenetic tree. It was also found that the CGI-rich and CGI-poor domains correlate more strongly than isochores to the segregation of the genomic features such as compartment and TADs formation as well as DNA methylation in human and mouse(Liu et al. 2018), indicating CpG density correlates better than the C+G content to the chromatin structure formation. Since C to T mutation is believed to be associated with CpG methylation, which has been shown to be actively involved in gene regulation, these results all suggest that the CpG density is likely more directly connected to biological functions and evolution than C+G content.

Conclusion

In this study, we explored the nucleotide distribution features in the genome sequences of different species. In evolution, the genome gradually loses CpG dinucleotide and gains in the unevenness of their distributions along the genome. Among the dinucleotides, the density distribution of CpG shows the most prominent multi-scale heterogeneity. Based on this

distribution, we divided the genomes into the CGI-rich and CGI-poor domains with distinct compositions and properties. By analyzing the average lengths, ratios and compositions of the CGI-rich and CGI-poor domains, we showed that genomes of birds and mammals are more heterogeneous than those of bacteria, plants, invertebrates, fishes and reptiles. Furthermore, we found that the CpG distribution is closely correlated with 3D chromatin organization at different length scales, especially in birds and mammals, suggesting the increased role of DNA sequence in determining the chromatin 3D structure in evolution. In warm-blooded animals, it appears that chromatin organization is strongly coded in the DNA sequence, especially in the uneven dinucleotide distributions. Due to the genome mosaicity, the chromatin 3D structures of warm-blooded species are likely to be more (stably) segregated than the ones of cold-blooded species. The increased role of DNA sequence in determining the chromatin structure along evolution is consistent with a co-evolution of the 1D and 3D genomes. We speculate this difference in genome sequence segregation to have an effect on differentiation, senescence, and maybe susceptibility to cancer among species. The various correlations found in this study call for careful and extensive experiment verifications.

Materials and Methods

Data Source

In this study, we analyzed the genomes of 38 representative species including bacteria, plants, invertebrates, fishes, reptiles, mammals and birds (Table S1). Genomes of species were retrieved from UCSC genome browser and NCBI. Source and information of Hi-C datasets are listed in SI.

Hilbert-Huang transform

A brief explanation of Hilbert-Huang transform (HHT)(Huang et al. 1998) is given here and the details can be found in the reference and SI. Hilbert-Huang transform can decompose a data series into oscillatory modes of different frequencies. The difference between HHT and Fourier transform is that HHT can be applied to nonlinear and nonstationary series without the requirement of a priori basis.

For a given data series X(t) as input, the mean value series m_1 of the upper and lower envelope of X(t) is calculated using the cubic spline lines. The difference h_1 between the input X(t) and m_1 is called the first protomode

$$h_1 = X(t) - m_1$$

h_1 is used as input in the next iteration to yield a new set of h_1 until the following conditions are satisfied: (1) In the entire dataset, the number of extrema and the number of zero crossings must either be equal or differ at most by one. (2) At any data point, the mean value of the envelope consisting of the local maxima and the one consisting of the local minima is zero.

After the first mode h_1 has converged, the difference between the input X(t) and h_1, i.e. X(t)- h_1 can be used as the input for the next iteration to obtain the next mode h_2. Repeating this procedure, we can retrieve modes of different frequencies: h_1, h_2, h_3...h_n, in the order of the decreasing frequency.

Dinucleotide variability

We define the variability of decomposed dinucleotide density series to quantify the extent of the dinucleotide density fluctuations at high frequencies. Given that the decomposed series fluctuates around the value zero (Figure S1), we calculated the absolute values of the decomposed series and then divided the new (absolute value) series into the high amplitude

group in which the values are larger than a+3σ, and the low amplitude group with values smaller than a+σ. Here a and σ are average and standard deviation of the new series. CpG variability is defined as

$$V = \frac{a_{high}}{a_{low}}$$

where a_{high} and a_{low} are high and low amplitude group averages, respectively. (The threshold value a+3σ is chosen since 3σ is often used as a threshold in detecting outliers in statistics theory. Choice of different thresholds does not have significant influence on the relative order of CpG variability of different genomes, see Table S2). A higher variability for a genome indicates that the amplitude of local CpG density fluctuation varies more drastically along the genome.

Generalized definition of CGI forest and prairie domains

As many species do not have identified CGIs, we expanded the forest-prairie domains definition(Liu et al. 2018) in human and mouse to the CGI-rich and CGI-poor domains in all species using human and mouse genomes as references. Considering that forest and prairie domains are reflections of the high level of CpG density fluctuation along the sequence, we selected sequential units with significantly high CpG density as CGI-rich domain 'loci' at 200 bp, 10 kb and 500 kb length scales, respectively. We next define regions either with significantly low CpG density (the value of which is smaller than both 75% regions of the species and 5 percentile of human and mouse forest domains) or with both low CpG density variation (the value of which is smaller than 75% regions of the species and 5 percentile of human and mouse forest domains) and low CpG density as the CGI-poor 'clusters' at 10 kb scale. The CGI-rich and CGI-poor domains are then defined following previous procedures used for forest and prairie domains except that the critical distance is selected as the maximum distance, using which at least 95% CGI-poor 'clusters' would be classified into

CGI-poor domains. For species with CGI-poor ‘clusters’ taking up less than 1% of total length or with CGI-rich ‘loci’ taking up more than 60% of total length, we used instead a canonical critical distance (which is the critical neighboring CpG density peak distance defined following our previous work(Liu et al. 2018)). The generalized forest and prairie domains definitions are reflections of the alternation between regions enriched in high CpG density peaks and regions with low CpG densities and small fluctuations.

Non-negative Matrix Factorization

Non-negative matrix factorization (NMF)(Lee and Seung 1999) is a decomposition method for a matrix (i.e. multivariate data) that has been used widely in signal processing, image recognition(Wang et al. 2016; Du and Swamy 2019) and computational biology(Devarajan 2008). The aim of non-negative matrix factorization is to reproduce the observed data by combining a limited number of basis components.

A non-negative factorization of matrix X is an approximation of X by the product of two matrices W and H, which are constrained to have non-negative entries (i.e. $W \geqslant 0$, $H \geqslant 0$). This decomposition is thus an approximation, not an equality. The solution W and H matrices minimize the quadratic error between X and W*H. The number of rows of W and columns of H should be same as the ones of X respectively, and the number of columns of W and rows of H can be chosen as needed.

When it is applied to Hi-C map, each column of W are defined as structure vector for it reflects intensity of Hi-C signal along the DNA sequence. Unlike the compartment vector calculated by PCA, the values of structure vector are all positive. Using structure vectors to recapitulate the Hi-C map is more informative on the contacts between compartment A and B than compartment vectors, because there is no offset of positive and negative elements of matrices. In this study, we applied NMF to Observed/Expected Hi-C matrix. The number of

columns of W (structure vectors) is set as two, as these two structure vectors were found to be correlated with compartments A and B, respectively.

Hi-C map analysis

Raw Hi-C data was processed by the ICE procedure. We grouped chromatin loci according to genomic distance, then calculated the decay of average Hi-C contact frequency with the genomic distance. To calculate the structure vectors, the Observed/Expected Hi-C map was first calculated to ensure the average contact frequency at different genomic distance is normalized, before performing Non-negative matrix factorization.

Data Access

Genomes of species are publicly available in NCBI and UCSC genome browser. Code for CGI-rich/poor domain division, CpG variability calculation, Multi-scale entropy analysis and Nonnegative Matrix Factorization is available on request.

Acknowledgments

This work was supported by National Natural Science Foundation of China [92053202, 22050003, 21927901, 21821004] and the National Key Research and Development Program of China [2017YFA0204702].

Author Contributions

Yi Qin Gao designed the research. Zhicheng Cai, Sirui Liu, Yueying He, Yue Xue, Hui Quan, Ling Zhang performed the research. Zhicheng Cai, Sirui Liu and Yi Qin Gao wrote the manuscript.

Competing Interests

The authors declare that they have no competing interests.

Reference

Albuquerque TAF, Drummond do Val L, Doherty A, de Magalhães JP. 2018. From humans to hydra: patterns of cancer across the tree of life. *Biol Rev Camb Philos Soc* **93**: 1715-1734.

Arneodo A, Vaillant C, Audit B, Argoul F, d'Aubenton-Carafa Y, Thermes C. 2011. Multi-scale coding of genomic information: From DNA sequence to genome structure and function. *Physics Reports* **498**: 45-188.

Audit B, Thermes C, Vaillant C, d'Aubenton-Carafa Y, Muzy JF, Arneodo A. 2001. Long-Range Correlations in Genomic DNA: A Signature of the Nucleosomal Structure. *Physical Review Letters* **86**: 2471-2474.

Baranowitz SA, Maderson PFA, Connelly TG. 1979. Lizard and newt tail regeneration: A quantitative study. *Journal of Experimental Zoology* **210**: 17-37.

Belcastro V, Siciliano V, Gregoretti F, Mithbaokar P, Dharmalingam G, Berlingieri S, Iorio F, Oliva G, Polishchuck R, Brunetti-Pierri N et al. 2011. Transcriptional gene network inference from a massive dataset elucidates transcriptome organization and gene function. *Nucleic Acids Res* **39**: 8677-8688.

Bernardi G. 1993. The isochore organization of the human genome and its evolutionary history--a review. *Gene* **135**: 57-66.

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunierrotival M, Rodier F. 1985. The mosaic genome of warm-blooded vertebrates. *Science* **228**: 953-958.

Bickmore Wendy A, van Steensel B. 2013. Genome Architecture: Domain Organization of Interphase Chromosomes. *Cell* **152**: 1270-1284.

Buldyrev SV. 2006. Power Law Correlations in DNA Sequences. In *Power Laws, Scale-Free Networks and Genome Biology*, doi:10.1007/0-387-33916-7_9 (ed. EV Koonin, et al.), pp. 123-164. Springer US, Boston, MA.

C. Steward F, O. Mapes M, Mears K. 1958. *Growth and Organized Development of Cultured Cells. II. Organization in Cultures Grown from Freely Suspended Cells*.

Caulin AF, Maley CC. 2011. Peto's Paradox: evolution's prescription for cancer prevention. *Trends in Ecology & Evolution* **26**: 175-182.

Cooper DN, Gerber-Huber S. 1985. DNA methylation and CpG suppression. *Cell Differentiation* **17**: 199-205.

Costantini M, Clay O, Auletta F, Bernardi G. 2006. An isochore map of human chromosomes. *Genome research* **16**: 536-541.

Costantini M, Musto H. 2017. The Isochores as a Fundamental Level of Genome Structure and Organization: A General Overview. *Journal of Molecular Evolution* **84**: 93-103.

Crawford NG, Faircloth BC, Mccormack JE, Brumfield RT, Winker K, Glenn TC. 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters* **8**: 783.

Devarajan K. 2008. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS computational biology* **4**: e1000029.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376-380.

Du K-L, Swamy MNS. 2019. Nonnegative Matrix Factorization. In *Neural Networks and Statistical Learning*, doi:10.1007/978-1-4471-7452-3_14 (ed. K-L Du, MNS Swamy), pp. 427-445. Springer London, London.

Hartman H. 1975. Speculations on the evolution of the genetic code. *Origins of life* **6**: 423-427.

Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen NC, Chi CT, Liu HH. 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings Mathematical Physical & Engineering Sciences* **454**: 903-995.

Hwu HR, Roberts JW, Davidson EH, Britten RJ. 1986. Insertion and/or deletion of many repeated DNA sequences in human and higher ape evolution. *Proceedings of the National Academy of Sciences* **83**: 3875-3879.

Ibn-Salem J, Muro EM, Andrade-Navarro MA. 2017. Co-regulation of paralog genes in the three-dimensional chromatin architecture. *Nucleic acids research* **45**: 81-91.

Ikeuchi M, Sugimoto K, Iwase A. 2013. *Plant Callus: Mechanisms of Induction and Repression*.

Kim H, Panteleyev AA, Jahoda CAB, Ishii Y, Christiano AM. 2004. Genomic organization and analysis of the hairless gene in four hypotrichotic rat strains. *Mammalian Genome* **15**: 975-981.

Kulminski AM, Culminskaya I, Yashin AI. 2013. Inter-chromosomal level of genome organization and longevity-related phenotypes in humans. *AGE* **35**: 501-518.

Lee DD, Seung HS. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**: 788-791.

Lewis EB. 1992. Clusters of Master Control Genes Regulate the Development of Higher Organisms. *JAMA* **267**: 1524-1531.

Lewis EB. 2004. Regulation of the Genes of the Bithorax Complex in Drosophila. In *Genes, Development and Cancer: The Life and Work of Edward B Lewis*, doi:10.1007/978-1-4419-8981-9_16 (ed. HD Lipshitz), pp. 255-272. Springer US, Boston, MA.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO et al. 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**: 289-293.

Liu C, Cheng Y-J, Wang J-W, Weigel D. 2017. *Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis*.

Liu S, Lorenzen Eline D, Fumagalli M, Li B, Harris K, Xiong Z, Zhou L, Korneliussen Thorfinn S, Somel M, Babbitt C et al. 2014. Population Genomics Reveal Recent Speciation and Rapid Evolutionary Adaptation in Polar Bears. *Cell* **157**: 785-794.

Liu S, Zhang L, Quan H, Tian H, Meng L, Yang L, Feng H, Gao YQ. 2018. From 1D sequence to 3D chromatin dynamics and cellular functions: a phase separation perspective. *Nucleic Acids Research* doi:10.1093/nar/gky633: gky633-gky633.

Musto H, Naya H, Zavala A, Romero H, Alvarez-Valín F, Bernardi G. 2004. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS Letters* **573**: 73-77.

Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F, Simons M, Stanley HE. 1992. Long-range correlations in nucleotide sequences. *Nature* **356**: 168-170.

Pigliucci M. 2010. Genotype–phenotype mapping and the end of the 'genes as blueprint' metaphor. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**: 557-566.

Rinker DC, Specian NK, Zhao S, Gibbons JG. 2019. Polar bear evolution is marked by rapid changes in gene copy number in response to dietary shift. *Proceedings of the National Academy of Sciences* **116**: 13446.

Rowley MJ, Corces VG. 2016. The three-dimensional genome: principles and roles of long-distance interactions. *Current opinion in cell biology* **40**: 8-14.

Rowley MJ, Nichols MH, Lyu X, Ando-Kuri M, Rivera ISM, Hermetz K, Wang P, Ruan Y, Corces VG. 2017. Evolutionarily Conserved Principles Predict 3D Chromatin Organization. *Mol Cell* **67**: 837-852.e837.

Saxonov S, Berg P, Brutlag DL. 2006. Saxonov S, Berg P, Brutlag DL.. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci USA 103: 1412-1417. *Proceedings of the National Academy of Sciences* **103**: 1412-1417.

Seebacher F, M Elsey R, L Trosclair P. 2003. *Body Temperature Null Distributions in Reptiles with Nonzero Heat Capacity: Seasonal Thermoregulation in the American Alligator (Alligator mississippiensis).*

Surani MA. 2012. Cellular Reprogramming in Pursuit of Immortality. *Cell Stem Cell* **11**: 748-750.

Suzuki Y, Nijhout HF. 2006. Evolution of a Polyphenism by Genetic Accommodation. *Science* **311**: 650.

Tatarinova TV, Alexandrov NN, Bouck JB, Feldmann KA. 2010. GC3 biology in corn, rice, sorghum and other grasses. *BMC Genomics* **11**: 308.

Tattersall G, Sinclair B, Withers P, Fields P, Seebacher F, Cooper C, Maloney S. 2012. *Coping with Thermal Challenges: Physiological Adaptations to Environmental Temperatures.*

Varriale A, Bernardi G. 2006a. DNA methylation and body temperature in fishes. *Gene* **385**: 111-121.

Varriale A, Bernardi G. 2006b. DNA methylation in reptiles. *Gene* **385**: 122-127.

Voss RF. 1992. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Physical Review Letters* **68**: 3805-3808.

Wang T, Ye T, Gurrin C. 2016. Transfer Nonnegative Matrix Factorization for Image Representation. In *MultiMedia Modeling,* (ed. Q Tian, et al.), pp. 3-14. Springer International Publishing, Cham.

Yakovchuk P, Protozanova E, Frank-Kamenetskii MD. 2006. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research* **34**: 1082-1082.