Genome Wide Analysis of Dinucleotide Distribution along the Genomes of Species and its Biological Implication

Zhicheng Cai, Sirui Liu, Yue Xue, Hui Quan, Ling Zhang, Yi Qin Gao

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

Abstract

Genome compositions vary among species and nucleotides are unevenly distributed in the genomes in correlation with genomic functions. The multi-scale organizations of dinucleotides in the genome and their evolution are important genomic features informative on biological function and evolution, but remain to be fully analyzed. Here, we investigated the distributions of dinucleotides, especially that of CpG due to its biological importance, in a variety of species. Among all dinucleotides, we found that CpG is the most unevenly distributed and the distributions of all dinucleotides are correlated and organized in blocks in high species, suggesting their biological impact on regulation. By comparing the local density fluctuations and the hierarchical distribution of CpG at different scales of genomic lengths, we found that CpG distributions of different species have distinct characteristics. The clustering of species based on the CpG distribution is consistent with the phylogenetic tree. Interestingly, the heterogeneity of CpG density appears to correlate with species' body temperature control. We propose a phase separation hypothesis to explain the dependence of chromatin structure and body temperature range on the genome sequence.

Introduction

The genome composition and organization play an important role in various cellular processes. To adapt to the environment, the genomes of organisms undergo

drastic mutations, leading to developmental complexity in organisms.[42,43] Many studies were conducted to explore the relationship between genome organizations and diverse phenotypes of organisms.[44-47]

With the development of next-generation sequencing, the genomes of a large number of species become 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[1]. Accordingly, thermal stability of DNA double helix was reported to be influenced by G+C contents[2]. Besides, nucleotide composition also influences genes and their distribution. The GC-rich regions often have high gene densities in eukaryotes, and the GC-rich genes in grass are usually related to basic metabolic processes and biotic stress responses[3]. Human promoters are divided into two classes according to the CpG density[4]. 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[5-6] 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[7-8]. In addition, among all dinucleotides, CpG has a higher tendency to be unevenly distributed, and aggregates to form the CpG islands (CGIs). Like isochores, based on the distribution of CGIs, the human and mouse genomes can be divided into two types of large domains: CGI(gene) forest domains with high CGI(gene) density and CGI(gene) prairie domains with low CGI (gene) density[9]. 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[10-11], detrended fluctuation analysis[12] and wavelet transform[13-14].

Previous analyses mainly focused on segments of the DNA chain such as coding sequences, but the global distribution of dinucleotides on the whole chromosome has been rarely discussed. Since the chromatin structure is proposed to be hierarchical and fractal[15-16], it is attempting to speculate that the dinucleotides density averaged at different lengths along the genome also exhibits fractal properties if a sequence-structure relation exists. Compartmentalization of chromatin structure was shown to correlate with CpG density[9]. It is thus interesting to interrogate how long-range

correlations of dinucleotide distributions discussed above and the global organization of dinucleotides evolve among species and affect phenotypes.

It becomes increasingly accepted that the chromatin 3D structure plays an important role in gene regulation and cellular functions[17-18], and many factors contributing to the chromatin structure formation have been explored. A sequence-based model[9] was proposed to explain the chromatin structure formation from the domain segregation perspective, which provides a framework for the exploration of chromatin structure formation in various cellular processes. To verify and generalize this model, an analysis of the sequence-structure relationship in different species is needed.

In this study, we use several methods to analyze the multi-scale distributions of all sixteen dinucleotides in a number of species from different classes and their effect on the chromatin structure revealed by HiC measurement. We first focus on the CpG dinucleotide due to its known biological importance. 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. On the larger scale, we divided the genomes into CGI-rich and CGI-poor domains with distinctly different densities of high CpG density loci, expanding the definition of CGI forest and prairie domains in human and mouse genomes to all species studied here. Among different species, the CGI-rich and CGI-poor domains are different in their lengths, CpG densities, and CpG density coefficients of variance. The distribution of CpG is also characterized by the Multi-scale Entropy and the Pearson correlation in a scale-continuous way. We further analyzed the distributions of all dinucleotides, and discuss their possible organizations along the genomes. Interestingly, we found that the CpG distribution profile correlates nicely with the degree of a species' chromatin structural segregation and body temperature control. A phase separation theory is proposed to explain such a correlation and biological consequences are discussed.

Materials and Methods

Hilbert-Huang transform

A brief explanation of Hilbert-Huang transform (HHT)[19] is given here and the details can be found in the reference. 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 data set, 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 get the next mode h_2. Repeating this procedure, we can get modes of different frequencies: h_1, h_2, h_3...h_n, in the order of the decreasing frequency.

Variability

To quantify the extent of the dinucleotide density fluctuations at high frequencies, the variability of decomposed series is defined. As the decomposed series fluctuates around the value zero, we first calculated the absolute values of the decomposed series. Then we 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 variation 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 variability of different genomes, see SI). 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[9] 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[9]). 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.

Multi-scale entropy analysis

To quantify the heterogeneity of the dinucleotide distribution, we calculated the multi-scale entropy[20-21] of the dinucleotide density. For a data series $X=\{x_1,\ldots,x_i,\ldots,x_N\}$ of length N, its m-length vectors are

$$u_m(i)=\{x_i,x_{i+1},\ldots,x_{i+m-1}\},\ 0<i<N-m$$

in which $n^m_i(r)$ is defined as the number of m-length vectors $u_m(j)$ that satisfies $d(u_m(i)\ ,\ u_m(j))<r$. Here d is the distance between two vectors defined as the maximum absolute difference between their components:

$$d(u_m(i), u_m(j))=max\{|x_{i+k} - x_{j+k}|:\ 0\leq k \leq m-1\}.$$

Multi-scale entropy (SE) of the data series X is then defined as:

$$SE(m,r,N) = ln\frac{\sum_{i=1}^{N-m} n_i^m}{\sum_{i=1}^{N-m} n_i^{m+1}}$$

m and r are chosen in this study to be 2 and 1.5 following reference 20.

A higher value of SE indicates a more heterogeneous distribution of X. The data series X needs to be normalized so that its heterogeneity rather than its integral fluctuation is quantified. To quantify the property of the data series at different length scales, the data can be averaged at different window sizes to yield differently coarse-grained data series, and the multi-scale entropy of each coarse-grained data series can then be calculated.

Result

We first analyzed the data on CpG dinucleotide. Among the 16 dinucleotides, CpG has many special properties such as its often low density in the genome[22] and richness in many human promoters in the form of CpG islands.[4] In addition, high gene densities are often found in the CpG-rich regions.

Local fluctuations of CpG density

We compared the local 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 the CpG Islands. Next, we decomposed CpG density data series of different species along the sequence into fluctuations at different frequencies by Hilbert-Huang transform. We then defined and calculated the variability (see method) of the CpG density fluctuation with the highest frequency, a large value of which corresponds to a large amplitude difference in CpG density fluctuation along the sequence.

As shown in Figure 1, the CpG variability can effectively cluster species into different groups. Among the species investigated, birds possess the highest variability, and bacteria have the lowest. The variability of mammals, reptiles, fish, plants and invertebrates are intermediate, approximately in the order following the phylogenetic tree. From the ranking of variability, alligator is similar to those of mammals, and 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.[23] 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.

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

In our previous work, forest and prairie domains were defined based on the unevenness of CGI distribution along the DNA sequence[9]. These two sequential domains effectively reflect the linear segregation in the genome of not only CGI densities, but also genetic, epigenetic, and structural properties. However, since for 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 forest-prairie domain definition to the CGI-rich and CGI-poor domains to compare

the difference in high-order sequential unevenness of CpG dinucleotide(see methods). We make use of properties of the prairies 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, and generalized CGI-poor domains have low CpG density and are deficient in CpG Islands. The newly defined domains are in good accordance with previous definition (the CGI-rich domains overlap ratio with forest domains is 94% for human and 87% for mouse). For 38 representative species, we evaluated their sequence unevenness 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 CVs (coefficient of variance) of the CpG density.

Among all the representative species, 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 lack in noncoding elements, which is consistent with the fact that it contains very few CGI-poor domains. For eukaryotes, 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 megabase 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). Though fishes and amphibians have low amounts and short lengths of CGI-poor domains, 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 the two species mentioned above and close to birds and mammals. 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. In contrast, the average CpG density levels and density fluctuations in low CpG density domains are similar between mammals and birds. 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, and plants specifically cluster together among eukaryotes. Moreover, cold-blooded species are also distinctly discriminated against warm-blooded ones. 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 clustered together with coelacanth, which is an important link in vertebrate evolution from fishes to tetrapod.

Multi-scale analysis of CpG distribution

The alternative appearance of CGI-rich and CGI-poor domains in genomes clearly shows the heterogeneity of CpG distribution on mega-base length scales. Next, we calculate the multi-scale entropy SE[20] (see method) of CpG density averaged at different lengths to describe the CpG distribution in a scale-continuous way. A higher value of this entropy corresponds to a more heterogeneous density distribution along the DNA sequence.

The trends of multi-scale entropy at different average lengths differ substantially among species (Figure 3 and SI). SEs of plants and invertebrates decrease in a linear way on the semi-logarithmic graph, suggesting their distribution of CpG becomes more uniform with the increasing average length and no specific mosaic CpG distribution exists in their genome. The SEs of fish, amphibians and reptiles also decrease with the increase of length, but much slower than those of plants and invertebrates. In contrast, SEs of mammals decrease slightly at short genome lengths and then increase at lengths over 100 kb, suggesting a heterogeneous distribution of CpG at a large length scale. Like mammals, SEs of birds also increase with genome length, but to a lesser extent than mammals, leading to largely flat SE curves. SEs of some birds slightly decrease at lengths over 1Mb. This result suggests that CpG distributions of mammals are more heterogeneous than those of birds at the mega-base length scale, consistent with mammals having longer CGI-rich and CGI-poor domains.

At average lengths larger than ~100kb, mammals and birds exhibit strongly heterogeneous CpG density distributions, while those of plants and invertebrates exhibit a low mosaicity, similar to the sequence variations composed of white noise. It is noted that the increase of SE with the average length is not due to higher noncoding DNA proportion in mammal and bird genome, as approximately 90% of the Takifugu genome is noncoding DNA but its SE decreases with the average

length. The result is also in agreement with the isochore theory that mammals and birds have more pronounced isochore structures[6].

Pearson correlation

Long-range Pearson correlations of the CpG density of different genomes are calculated to further investigate the scaling properties of the CpG distribution.

The CpG density correlations of higher species decay with the increase of genomic distance in the form of power-law (Figure 4A), whereas in lower species this power law decay only persists to a short distance if it does exist. Such a power law decay of correlation coefficient reflects the scale-free property of the CpG distribution in the higher but not lower species[24-25].

Although birds and mammals both exhibit power law decays, they differ substantially in that correlation coefficients of the former decrease more rapidly than those of the latter as the genomic distance increases.

CpG density distribution in different species

As seen in Figure 5, the distribution of average CpG density and CpG variability of the different species shows roughly a "L" shape. In general, along evolution, the CpG density decreases. For example, its value for bacteria and plants is much higher than that for fishes, reptiles, birds, and mammals. The CpG density difference among different bacteria, plants, and that between plants 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 slightly less than that of fishes and reptiles. The average CpG density being nearly constant at the later stage of evolution is consistent with a detailed balance condition[40] 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[39]), 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 variabilities for bacteria, plants, fishes and reptiles are all generally very similar and small. The CpG density variabilities of mammals and birds are much higher than those of lower species (Figure 5). It seems that genomes evolve in different ways among lower and higher species. In early evolution, the CpG density decreases with little change of the mosaicity and in the

later stage of evolution, the CpG density remains largely constant but the mosaicity increases.

Distribution of other dinucleotides

Next, we extend our analysis to all dinucleotides. As mentioned earlier, among 16 types of dinucleotides, CpG tends to possess the most uneven density distribution. Taking chromosome 1 of human as an example, CpG density has the largest 4th moment at various length scales (Figure 6A), indicating its highest probability of assuming extreme values among all the 16 dinucleotides. We also calculated the variability of all dinucleotide densities which reflects local fluctuations (Figure 6B). Again, CpG has the largest value.

One can also see that dinucleotides composed by C and G (that is CC, GG, CG, GC) or by A and T (that is AA, TT, AT, TA) have a 4th moment that is larger than the other eight dinucleotides such as AG which is composed of two types of nucleotides, one being A or T and the other being C or G. Besides CpG, distributions of other dinucleotides also become more heterogeneous as one moves from lower to higher species. For example, the long-range Pearson correlations of AC and CA of mammals and birds also exhibit the power-law decay. Meanwhile, the long-range Pearson correlations of CA of the birds, mammals and cold-blooded animals can be easily distinguished from each other (Figures 4B). Interestingly and consistent with the trend discussed above, among all the mammals, the density correlation function of platypus decays in a uniquely different way, possibly because of its transitional position on the phylogenetic tree.

The analysis above thus shows that nucleotides C and G tend to co-localize along the linear genome, so do nucleotides A and T. Nucleotides C and G density fluctuations are positively correlated[26] and are of an opposite trend to A and T density fluctuations. This relation can be seen from the calculated density fluctuations at large genome lengths of all dinucleotides. It is observed in Figure 6C that the distributions of all dinucleotides fluctuate in blocks and their density curves change synergistically. These results suggest that the nucleotide distribution is organized in large blocks. It is known that genes distribute along the genome in blocks and genes with similar functions tend to be located near each other[48,49]. Since genetic information contained in genome is expected to depend on the nucleotide arrangement, our analysis suggests that genetic information has a tendency to be stored block by block on the genome, which is readily used, repressed and changed in a concerted way. From the information storage perspective, the amount of information contained in a block is reflected by the variation and diversity of

nucleotide composition in the block. A larger density variation of CpG corresponds to richer genetic information and the appearance of mosaicity of the genome during evolution indicates the change towards a more complex gene regulation system.

Finally, we also calculated 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 dinucleotides. For example, as shown in Figure 7, the trinucleotides containing CpG such as CCG, CGC and ACG also show high coefficient of variance (CV), similar to the dinucleotide CpG. Meanwhile the trinucleotides such as ACA have high variability like the dinucleotides AC and CA. Similar phenomena are also observed for tetranucleotides (see SI). These results indicate that the heterogeneous nucleotide distribution is driven by that of dinucleotides. Interestingly, it is known that codon is dominantly determined by its first dinucleotide[52], supporting the importance of the dinucleotide composition in the DNA sequence.

Discussion

In summary, the sequence properties of lower (mainly bacteria, plants, invertebrates and fishes) and higher species (birds and mammals) are distinctly different in terms of the dinucleotide distribution. The CpG density of the former shows low variabilities and a decreasing trend (as a function of genomic distance) in multi-scale entropy. Its Pearson correlation coefficient does not exhibit a long-range power law decay. In contrast, the CpG density of higher species has relatively high CpG variabilities, a partially increasing trend of multi-scale entropy with the genomic length, and a long-range power law decay, all of which indicate a more heterogeneous CpG distribution. From the perspective of CGI-rich and CGI-poor domains, higher species have longer and a larger proportion of CGI-poor domains than lower ones. Consistently, their overall average CpG density is also lower than that of lower species. Among higher species, birds and mammals can also be distinguished from each other according to their sequence mosaicity, as birds have a higher CpG variability, a shorter CGI-poor domain length, and a more rapid decay in Pearson correlation coefficients of the CpG density than mammals.

Therefore, CpG distributions of different genomes have several distinctly different characteristics. Interestingly, species with shorter and smaller proportions of CGI-poor domains are more dominantly cold-blooded, living with a wide body temperature range. It seems that CGI-poor domains rarely exist in the genome of these species, but become ubiquitous in the higher species. For birds and mammals, their long CGI-poor domain lengths, relatively high proportions of CGI-poor

domains and significant CpG density differences between 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 show a correlation with body temperature control of different species (Figure 1). Warm-blooded species (birds and mammals) have higher DNA sequence variability values 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[27-28]. Consistently, the CpG density variability of the alligator 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 the variability of fishes from the tropical and polar region (see SI). The tropical fishes were found to have a higher sequence variability than the polar fishes. Besides, the brown bear has a higher sequence variability than the polar bear. 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[29-30]. 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 which show that a high CpG density is correlated with the low variability (Figure 5). It is likely that the methylation level correlates with CpG density and its fluctuation.

Our analysis also 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 5) can cluster species corresponding to their positions in the phylogenetic tree. As for 3D chromatin structure, our early study 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 TAD formation as well as DNA methylation in human and mouse[9].

Based on the observations mentioned above we propose a hypothesis to explain the relation between the body temperature range and nucleotide heterogeneity of different species[40]: Especially for mammals and birds, the chromatin is composed by two types of regions with distinct nucleotide compositions and thus sequence

properties: CGI-rich and poor domains, resembling a random block copolymer. The inactive CGI-poor domains tend to segregate from CGI-rich domains[9, 31]. This domain segregation process is an entropy-driven process accelerated by the higher temperature[32]. The more heterogeneous chromatin of species such as mammals and birds has a stronger tendency to form a domain-segregated structure, and is thus more sensitive to temperature changes. Evidence does exist that a higher temperature corresponds to more condensed chromatin structures. For example, it was observed that the rice genome has fewer Hi-C contact frequencies between compartment A and B at higher temperatures[33]. In a different study, it was observed that the chromatin structure of the cell became condensed after the heat shock stimulation[35]. Besides, the calculated[34] separation ratio of the CGI-rich and CGI-poor domains of mouse embryo's chromatins during the embryo development agrees well with the temperature change of mouse embryo from the experiment, i.e., a higher separation ratio corresponds to a higher temperature.

To further examine the relation between DNA sequence heterogeneity and structural segregation, we calculated the chromatin contact frequency decay for different species. We found that in general the sequence difference does correspond to the chromatin structure difference for the various species investigated (Figure 8). The contact frequency decay exponent is conserved between mouse and human, and among many cell types of human. At the short length scale, the HiC contact frequency for mouse and human samples decays more slowly than species such as yeast, A.thaliana and Drosphila, indicating a more segregated structure of mouse and human at these length scales. In contrast, at length scales larger than megabases, the HiC contact frequency for the lower species decays more slowly, indicating the higher frequency of long range contacts in lower species than mammals.

As mentioned above, a close relation does exist between the mosaic CGI and gene distributions in human and mouse genomes[9]. The segregation of the CGI/gene rich-poor domains largely determines the chromatin compartmentalization, with a large number of alternating compartments A and B (Figure 9). In contrast, it is well known that chromosomes of most plants are simply partitioned into three compartments: two compartments A near the telomere and one compartment B near the centromere[33]. We calculated the gene density of rice and A.thaliana along the genome and found their gene density is correlated with compartment partition: gene density is higher in compartment A than B (Figure 9). However, no correlation exists between CpG density and compartment partition since the CpG distribution is nearly uniform in plants (Figure 3). Similar gene density-compartment correlation was also observed in archaea. These results suggest the possible 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 higher species like mammals but not in plants. The separation of the chromatin structure into different compartments following the gene density distribution is highly conserved across many different species, and indicates the close relation between the uneven 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, consistent with a function-driven gene and DNA sequence redistribution in evolution. Since genes of similar functions tend to form spatial contacts, it would be interesting to examine the cross-talk between 3D chromatin structure and DNA linear sequence in different species and in evolution.

Finally, it is noted that besides the body temperature, there are other putative correlations between the biological function and sequence mosaicity of a species. Recently, it was proposed that a stable phase separation of genome is correlated to differentiation, senescence and diseases such as cancer[9]. A high mosaicity of a genome appears to correspond to stable differentiation. Consistent with this theory, with a low genome mosaicity (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[36-37]. Moreover, fishes and reptiles can grow up and develop throughout their lives. For example, lizards can regenerate tails[41]. Among mammals, whale, which has a low variability (Figure 1), is the largest mammal on earth and famous for its longevity and ability of suppressing cancer[38]. Similarly, elephants are also known for their longevity, resistance to aging and cancer, and indeterminate body size. In contrast, birds and mammals with a high sequence mosaicity are difficult to reprogramming[50], and more prone to cancer than invertebrates and plants[51].

Conclusion

In this study, we explored the features of genome sequence in different species. In evolution, the genome gradually loses CpG dinucleotide and gains in the unevenness of their distributions along the genome. The distribution of longer DNA sequences appears to be largely determined by that of the dinucleotides. The importance of dinucleotide distribution in determining DNA sequence properties is consistent with dinucleotide's dominant role in codons[52]. 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, ratio and compositions of the CGI-rich and CGI-poor domains, we showed that genomes of warm-blooded species are more heterogeneous than those of cold-blooded ones. We propose a phase separation model in which CGI-rich and CGI-poor domains gradually segregate from each other in cellular processes such as differentiation and senescence. Due to the genome mosaicity, the chromatin 3D structures of warm-blooded species are more (stably) segregated than the ones of cold-blooded species, which makes them more sensitive to the temperature. We speculate this difference in genome sequence segregation to have an effect on differentiation, senescence, and maybe susceptibility to cancer among species. The many correlations found in this study call for extensive experiment verifications.

Acknowledgement

This work was supported by National Natural Science Foundation of China [21573006, 21233002 and 91427304 to Yi Qin Gao] and the National Key Research and Development Program of China [2017YFA0204702 to Yi Qin Gao].

Author Contributions

Yi Qin Gao designed the research. Zhicheng Cai, Sirui Liu, 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.

Figures

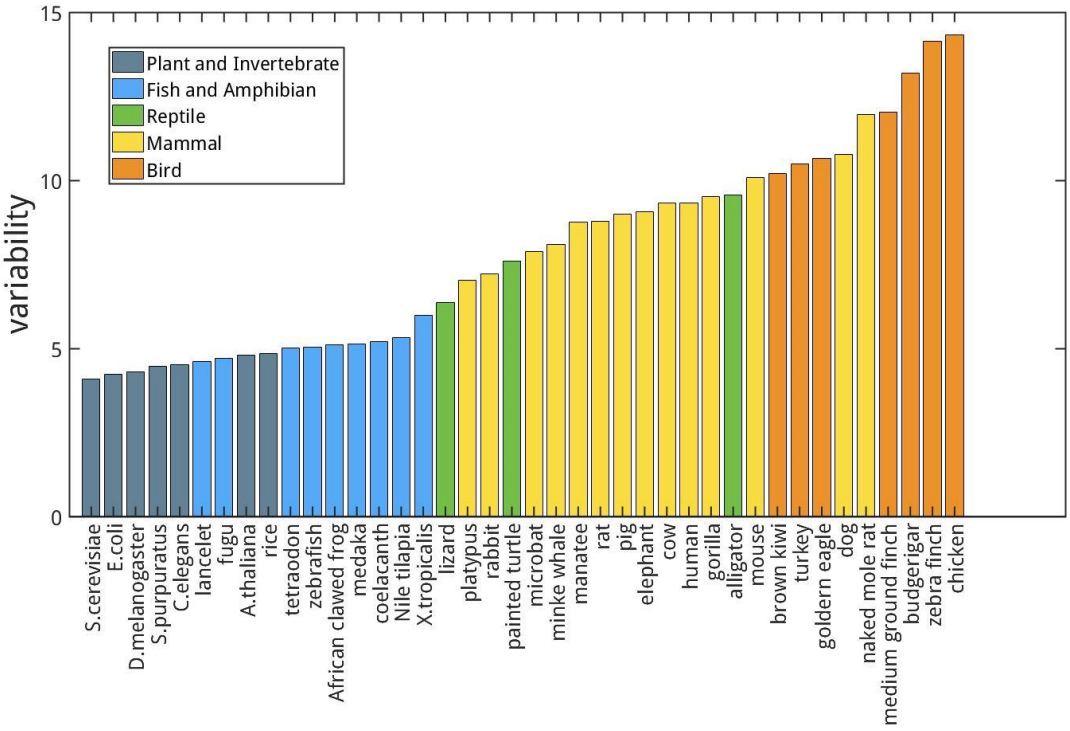

Figure 1. Variability of the CpG density for different species.

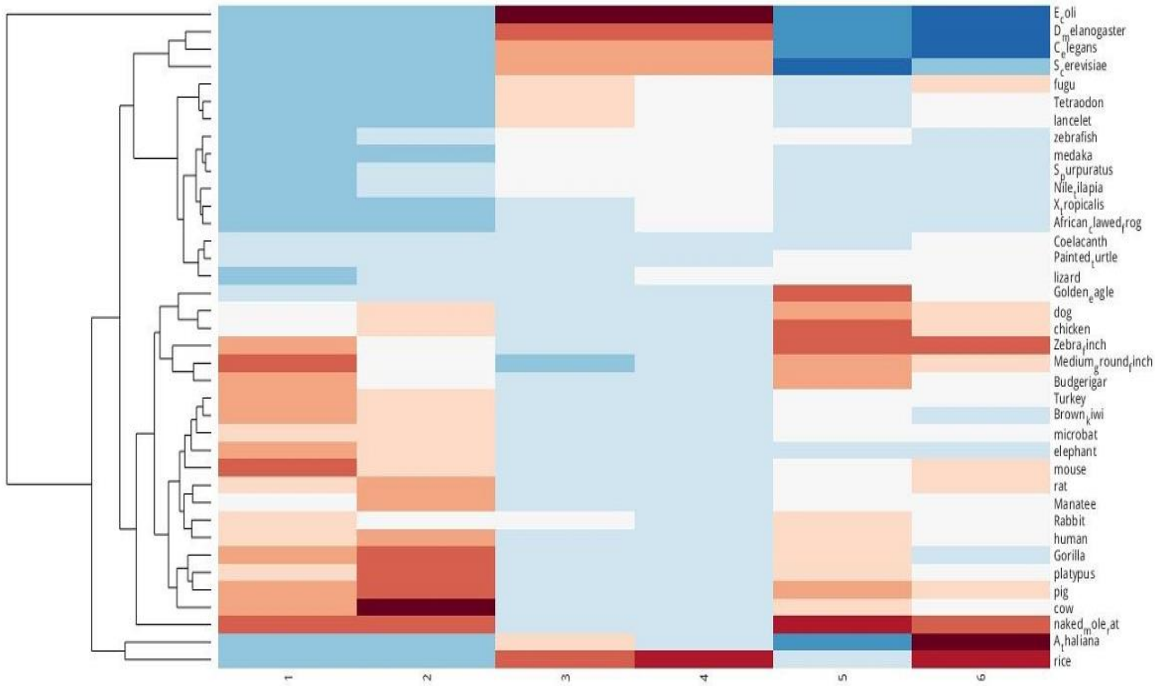

Figure 2. Dendrogram for 38 species from different classes. Six columns represent the proportions of CGI-poor domain, CGI-poor domain average length, average CpG densities of CGI-rich and poor domains, and the corresponding CVs of CpG density respectively.

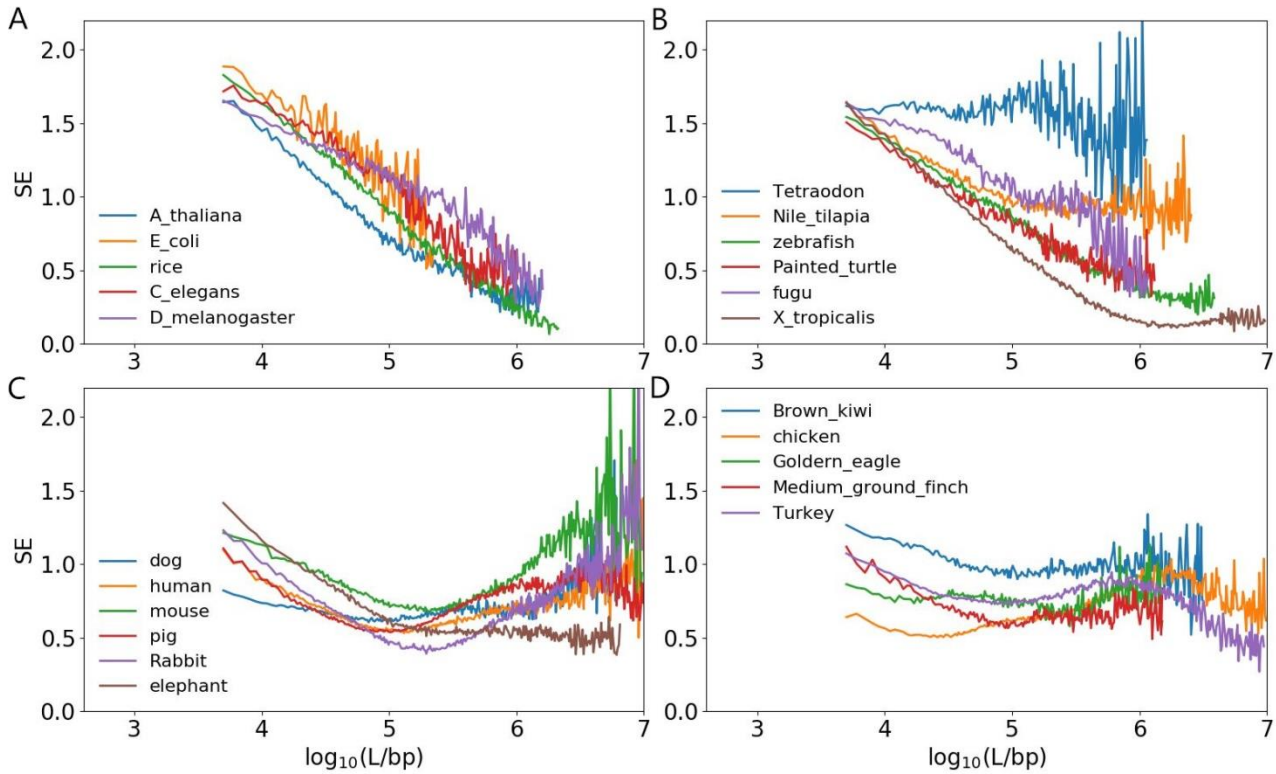

Figure 3. Multi-scale entropy (SE) of species belonging to (A) plants and invertebrates, (B) fishes, amphibians and reptiles, (C) mammals, (D) birds.

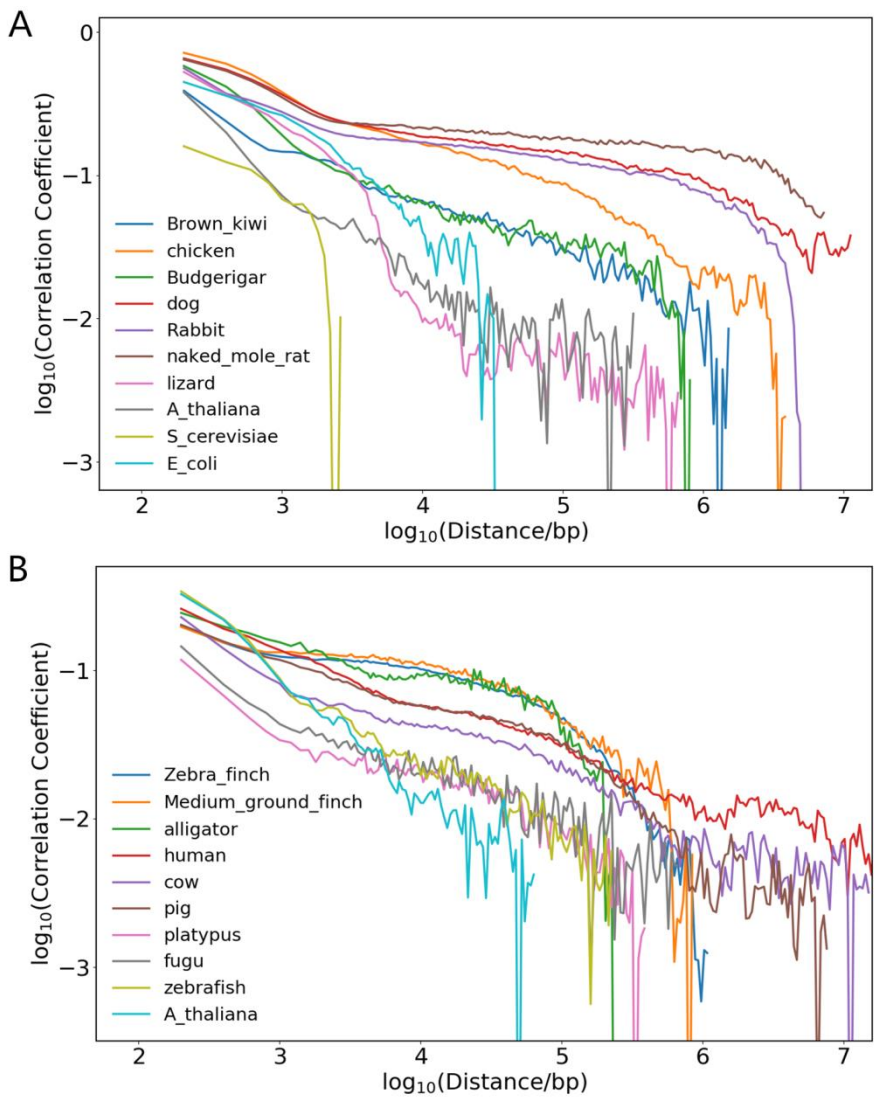

Figure 4. Long-range Pearson correlation of dinucleotide density distribution of different species

(A) Long-range Pearson correlation of CpG density distribution.

(B) Long-range Pearson correlation of CpA density distribution.

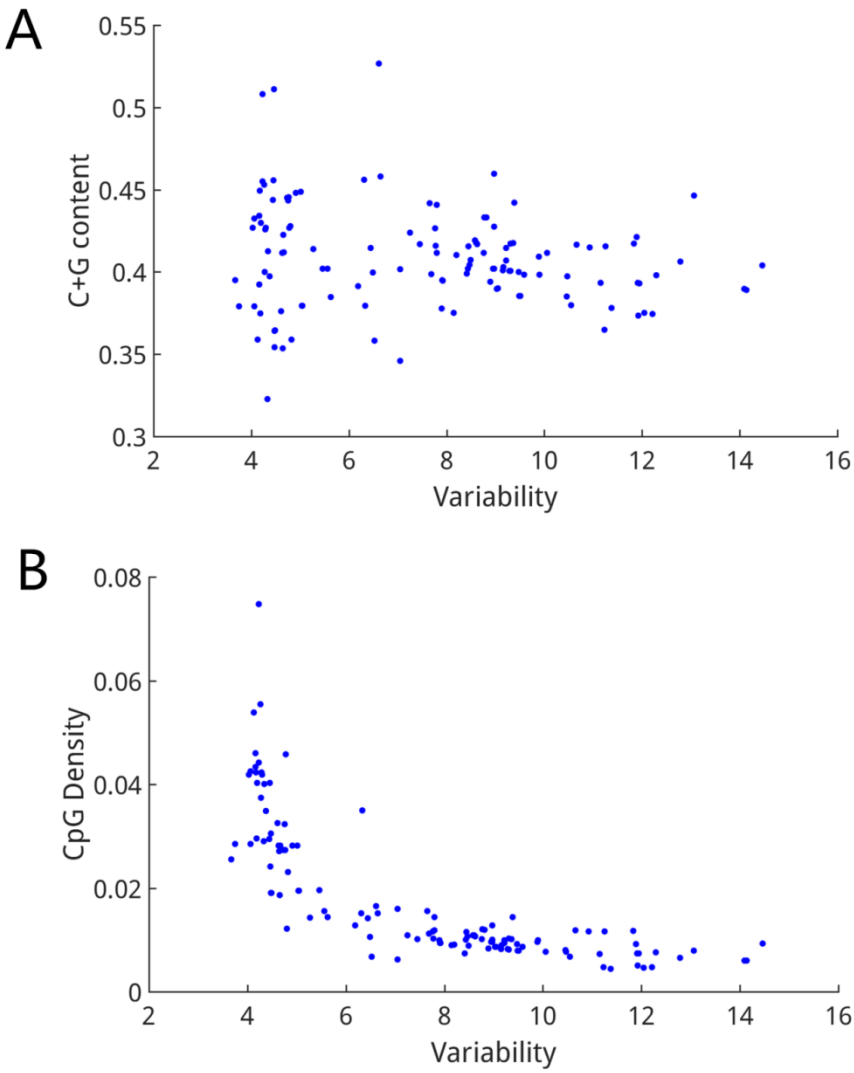

Figure 5.

(A) The scatter plot for C+G content and CpG variability of various species.

(B) The scatter plot for CpG density and CpG variability of various species.

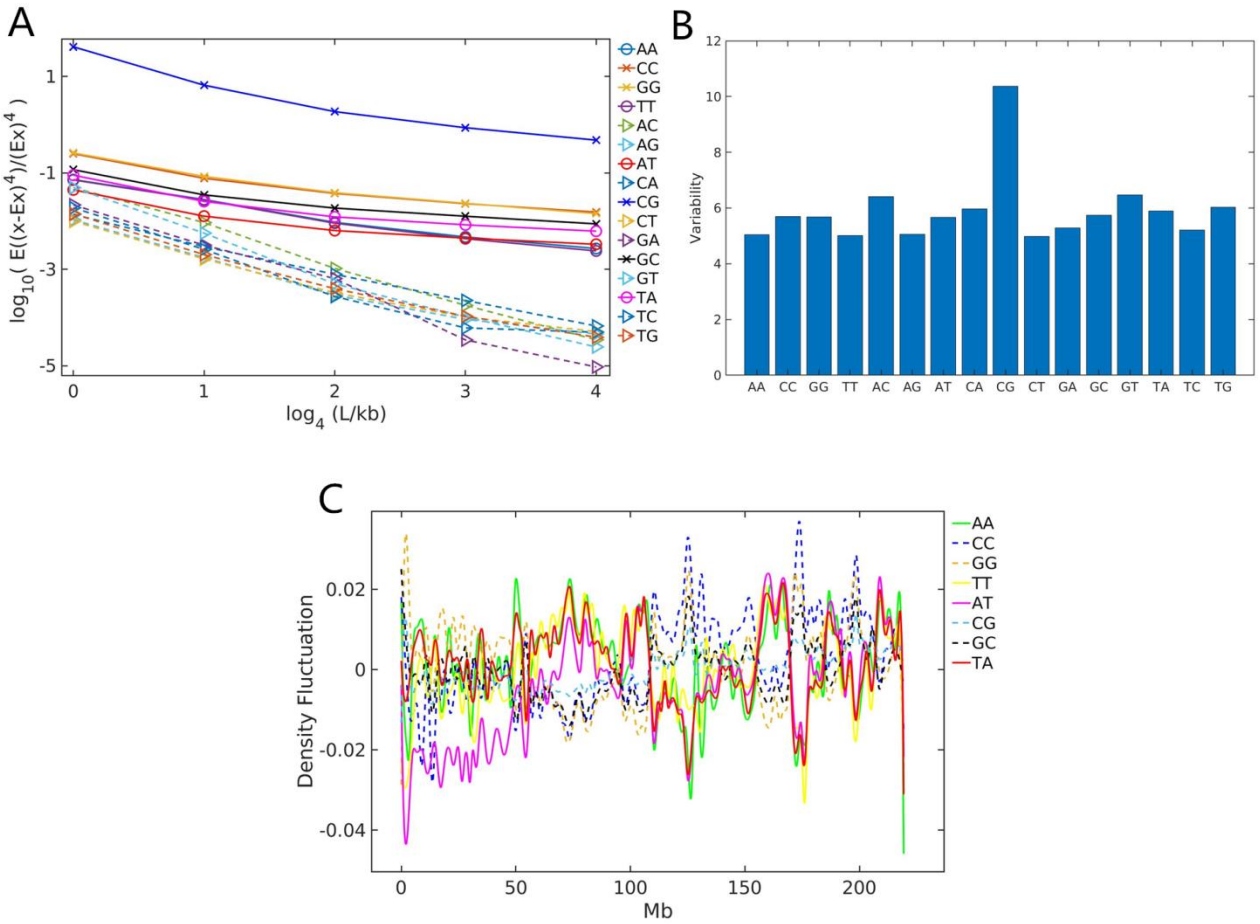

Figure 6. Distribution characteristics of 16 dinucleotides in human chr1

(A) The 4th moment of 16 dinucleotides density averaged at different length scales.

(B) Variability of 16 dinucleotides density averaged at 1kb.

(C) Density fluctuations at large length scale of 8 dinucleotides.

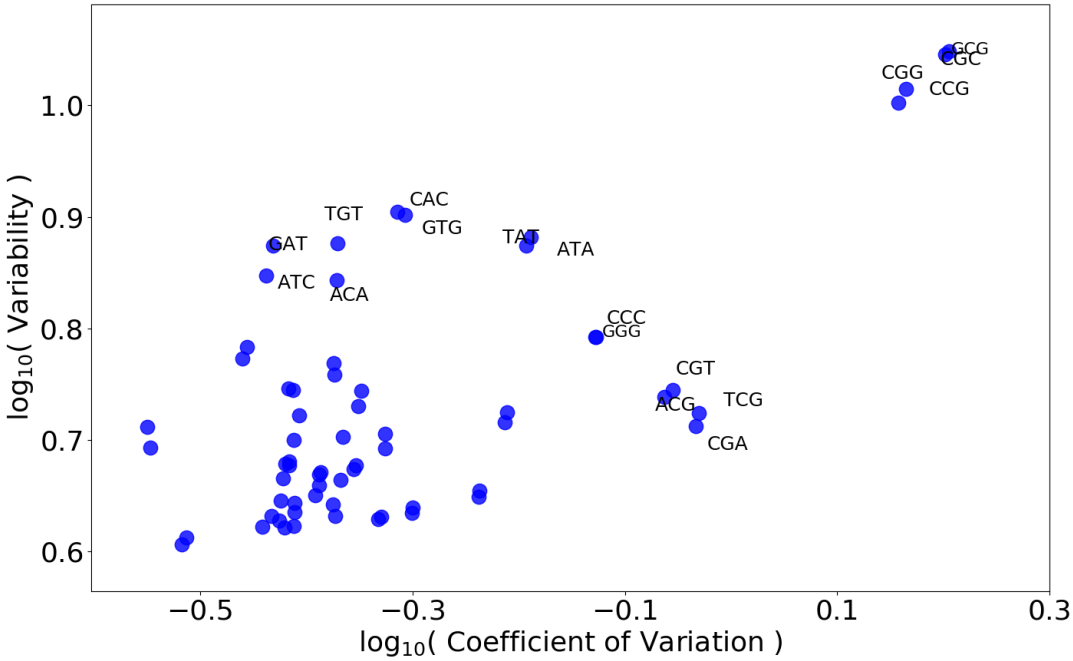

Figure 7. The scatter plot for variability and CV of 64 trinucleotides.

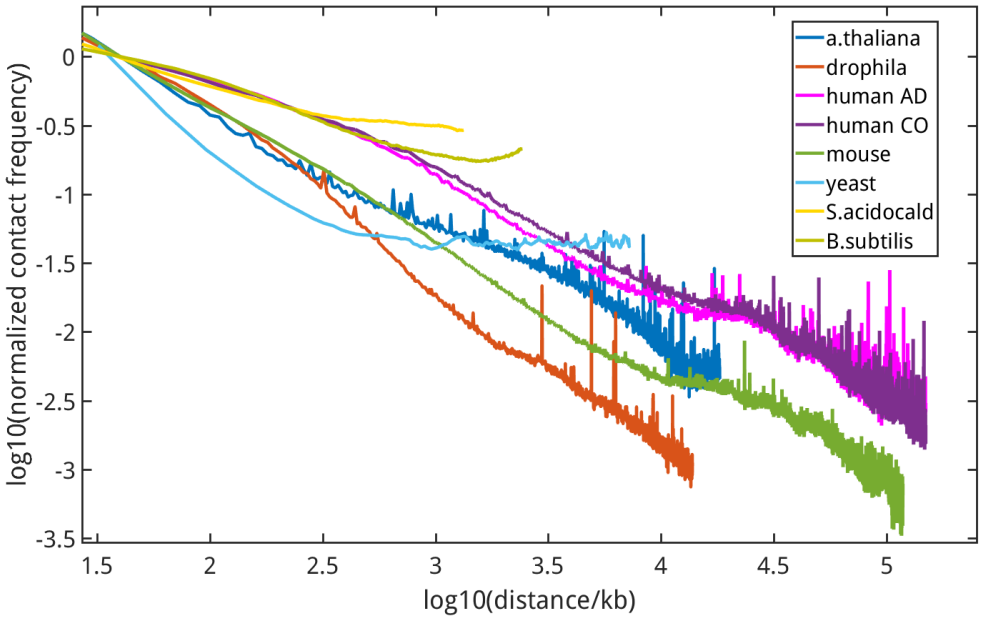

Figure 8. Contact frequencies of chromatin separated by different genome distances for various species.

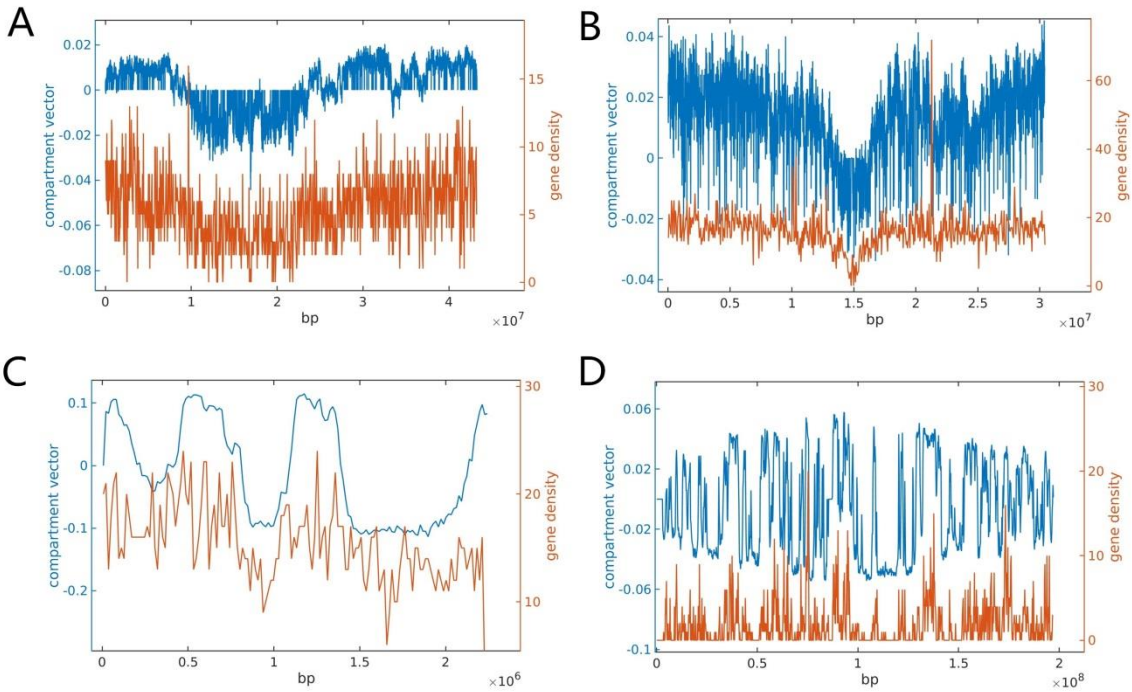

Figure 9. Comparison between chromosomal compartment vectors and gene density distribution of various species. (A) rice, (B) A.thaliana, (C) S.acidocaldarius, (D) mouse.

Reference

1. Musto, H.; Naya, H.; Zavala, A.; Romero, H.; Alvarez-Valín, F.; Bernardi, G., Correlations between Genomic Gc Levels and Optimal Growth Temperatures in Prokaryotes. *FEBS Letters* **2004**, *573*, 73-77.
2. Yakovchuk, P.; Protozanova, E.; Frank-Kamenetskii, M. D., Base-Stacking and Base-Pairing Contributions into Thermal Stability of the DNA Double Helix. *Nucleic Acids Research* **2006**, *34*, 1082-1082.
3. Tatarinova, T. V.; Alexandrov, N. N.; Bouck, J. B.; Feldmann, K. A., Gc3 Biology in Corn, Rice, Sorghum and Other Grasses. *BMC Genomics* **2010**, *11*, 308.
4. Saxonov, S.; Berg, P.; Brutlag, D. L., 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* **2006**, *103*, 1412-1417.
5. Bernardi, G., The Isochore Organization of the Human Genome and Its Evolutionary History--a Review. *Gene* **1993**, *135*, 57-66.
6. Bernardi, G.; Olofsson, B.; Filipski, J.; Zerial, M.; Salinas, J.; Cuny, G.; Meunierrotival, M.; Rodier, F., The Mosaic Genome of Warm-Blooded Vertebrates. *Science* **1985**, *228*, 953-958.
7. Costantini, M.; Clay, O.; Auletta, F.; Bernardi, G., An Isochore Map of Human Chromosomes. *Genome research* **2006**, *16*, 536-541.
8. Costantini, M.; Musto, H., The Isochores as a Fundamental Level of Genome Structure and Organization: A General Overview. *Journal of Molecular Evolution* **2017**, *84*, 93-103.
9. Liu, S.; Zhang, L.; Quan, H.; Tian, H.; Meng, L.; Yang, L.; Feng, H.; Gao, Y. Q., From 1d Sequence to 3d Chromatin Dynamics and Cellular Functions: A Phase Separation Perspective. *Nucleic Acids Research* **2018**, gky633-gky633.
10. Buldyrev, S. V., Power Law Correlations in DNA Sequences. In *Power Laws, Scale-Free Networks and Genome Biology*, Koonin, E. V.; Wolf, Y. I.; Karev, G. P., Eds. Springer US: Boston, MA, 2006; pp 123-164.
11. Voss, R. F., Evolution of Long-Range Fractal Correlations and 1/F Noise in DNA Base Sequences. *Physical Review Letters* **1992**, *68*, 3805-3808.
12. Peng, C. K.; Buldyrev, S. V.; Goldberger, A. L.; Havlin, S.; Sciortino, F.; Simons, M.; Stanley, H. E., Long-Range Correlations in Nucleotide Sequences. *Nature* **1992**, *356*, 168-170.
13. Arneodo, A.; Vaillant, C.; Audit, B.; Argoul, F.; d'Aubenton-Carafa, Y.; Thermes, C., Multi-Scale Coding of Genomic Information: From DNA Sequence to Genome Structure and Function. *Physics Reports* **2011**, *498*, 45-188.
14. Audit, B.; Thermes, C.; Vaillant, C.; d'Aubenton-Carafa, Y.; Muzy, J. F.; Arneodo, A., Long-Range Correlations in Genomic DNA: A Signature of the Nucleosomal Structure. *Physical Review Letters* **2001**, *86*, 2471-2474.
15. Grosberg, A.; Rabin, Y.; Havlin, S.; Neer, A., Crumpled Globule Model of the Three-Dimensional Structure of DNA. *Europhysics Letters (EPL)* **1993**, *23*, 373-378.
16. Gibcus, J.; Dekker, J., *The Hierarchy of the 3d Genome*, 2013; Vol. 49, p 773-82.
17. Lieberman-Aiden, E., et al., Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **2009**, *326*, 289-293.
18. Bickmore, Wendy A.; van Steensel, B., Genome Architecture: Domain Organization of Interphase Chromosomes. *Cell* **2013**, *152*, 1270-1284.
19. Huang, N. E.; Shen, Z.; Long, S. R.; Wu, M. C.; Shih, H. H.; Zheng, Q.; Yen, N. C.; Chi, C. T.; Liu, H. H., The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis. *Proceedings Mathematical Physical & Engineering Sciences* **1998**, *454*, 903-995.

20. Costa, M.; Goldberger, A. L.; Peng, C. K., Multiscale Entropy Analysis of Biological Signals. *Physical Review E Statistical Nonlinear & Soft Matter Physics* **2005**, *71*, 021906.
21. Jiang, Y.; Peng, C. K.; Xu, Y., Hierarchical Entropy Analysis for Biological Signals. *Journal of Computational & Applied Mathematics* **2011**, *236*, 728-742.
22. Cooper, D. N.; Gerber-Huber, S., DNA Methylation and Cpg Suppression. *Cell Differentiation* **1985**, *17*, 199-205.
23. Crawford, N. G.; Faircloth, B. C.; Mccormack, J. E.; Brumfield, R. T.; Winker, K.; Glenn, T. C., More Than 1000 Ultraconserved Elements Provide Evidence That Turtles Are the Sister Group of Archosaurs. *Biology Letters* **2012**, *8*, 783.
24. Zhang, L.; Xie, W. J.; Liu, S.; Meng, L.; Gu, C.; Gao, Y. Q., DNA Methylation Landscape Reflects the Spatial Organization of Chromatin in Different Cells. *Biophysical Journal* **2017**, *113*, 1395-1404.
25. Clauset, A.; Shalizi, C. R.; Newman, M. E. J., Power-Law Distributions in Empirical Data. *Siam Review* **2009**, *51*, 661-703.
26. Baisnée, P.-F.; Hampson, S.; Baldi, P., *Why Are Complementary DNA Strands Symmetric?*, 2002; Vol. 18, p 1021-33.
27. Seebacher, F.; M Elsey, R.; L Trosclair, P., *Body Temperature Null Distributions in Reptiles with Nonzero Heat Capacity: Seasonal Thermoregulation in the American Alligator (Alligator Mississippiensis)*, 2003; Vol. 76, p 348-59.
28. Tattersall, G.; Sinclair, B.; Withers, P.; Fields, P.; Seebacher, F.; Cooper, C.; Maloney, S., *Coping with Thermal Challenges: Physiological Adaptations to Environmental Temperatures*, 2012; Vol. 2, p 2151-2202.
29. Varriale, A.; Bernardi, G., DNA Methylation in Reptiles. *Gene* **2006**, *385*, 122-127.
30. Varriale, A.; Bernardi, G., DNA Methylation and Body Temperature in Fishes. *Gene* **2006**, *385*, 111-121.
31. Boettiger, A. N.; Bintu, B.; Moffitt, J. R.; Wang, S.; Beliveau, B. J.; Fudenberg, G.; Imakaev, M.; Mirny, L. A.; Wu, C.-t.; Zhuang, X., Super-Resolution Imaging Reveals Distinct Chromatin Folding for Different Epigenetic States. *Nature* **2016**, *529*, 418.
32. Chandler, D., Interfaces and the Driving Force of Hydrophobic Assembly. *Nature* **2005**, *437*, 640-647.
33. Liu, C.; Cheng, Y.-J.; Wang, J.-W.; Weigel, D., *Prominent Topologically Associated Domains Differentiate Global Chromatin Packing in Rice from Arabidopsis*, 2017; Vol. 3.
34. Quan, H.; Liu, S.; Zhang, Y.; Xie, W.; Gao, Y. Q., Phase Separation During Mouse Early Embryonic Development and Underlying Genetic and Epigenetic Correlations. *bioRxiv* **2019**, 521401.
35. Huang, K.; Li, Y.; Shim, A. R.; Nap, R. J.; Agrawal, V.; Virk, R. K. A.; Eshein, A.; Almassalha, L. M.; Backman, V.; Szleifer, I., Physical and Data Structure of 3d Genome. *bioRxiv* **2019**, 596262.
36. Ikeuchi, M.; Sugimoto, K.; Iwase, A., *Plant Callus: Mechanisms of Induction and Repression*, 2013; Vol. 25.
37. C. Steward, F.; O. Mapes, M.; Mears, K., *Growth and Organized Development of Cultured Cells. Ii. Organization in Cultures Grown from Freely Suspended Cells*, 1958; Vol. 45.
38. Caulin, A. F.; Maley, C. C., Peto's Paradox: Evolution's Prescription for Cancer Prevention. *Trends in Ecology & Evolution* **2011**, *26*, 175-182.
39. Hwu, H. R.; Roberts, J. W.; Davidson, E. H.; Britten, R. J., Insertion and/or Deletion of Many Repeated DNA Sequences in Human and Higher Ape Evolution. *Proceedings of the National Academy of Sciences* **1986**, *83*, 3875-3879.
40. Amy R. Strom, Alexander V. Emelyanov, Mustafa Mir, Dmitry V. Fyodorov, Xavier Darzacq, Gary H. Karpen, Phase separation drives heterochromatin domain formation. *Nature* **2017**, 547, 241–245.
41. S. A. Baranowitz; P. F. A. Maderson; T. G. Connelly, Lizard and new tail regeneration: A Quantitative study. *Journal of Experimental Zoology* **1979**, 210.

42. Shiping Liu, Eline D.Lorenzen, Matteo Fumagalli, Bo Li, Kelley Harris, Zijun Xiong, LongZhou, Thorfinn Sand Korneliussen, Mehmet Somel, Courtney Babbitt,..,Population genomics reveal recent speciation and rapid evolutionary adaption in polar bears. *Cell* **2014**, 157, 4, 785-794.
43. Yuichiro Suzuki, H. Frederik Nijhout, Evolution of a Polyphenism by Genetic Accommodation *Science* **2006**, 311, 650-652.
44. Alexander M. Kulminski, Irina Culminskaya, and Anatoli I. Yashin,Inter-chromosomal level of genome organization and longevity-related phenotypes in humans. *Age (Dordr)* **2013**, 35(2), 501–518.
45. Hyunmi Kim, Andrey A. Panteleyev, Colin A.B. Jahoda, Yoshiyaki Ishii, Angela, M. Christiano, Genomic organization and analysis of the hair less gene in four hypotrichotic rat strains. *Mammalian Genome* **2004**, 15(12), 975–981.
46. Massimo Pigliucci, Genotype-phenotype mapping and the end of the 'gene as blueprint' metaphor. *Philosophical Transactions: Biological Sciences* **2010**, 365,1540.
47. David C. Rinker, Natalya K. Specian, Shu Zhao, John G. Gibbons, Polar bear evolution is marked by rapid changes in gene copy number in response to dietary shift. *PNAS* **2019**, 116 (27), 13446-13451.
48. E. B. Lewis, Regulation of the genes of the bithorax complex in Drosophila. *Cold Spring Harb Symp Quant Biol* **1985**, 50, 155-64.
49. E. B. Lewis, Clusters of master control genes regulate the development of higher organisms. *JAMA* **1992**, 267(11), 1524-31.
50. M. Azim Surani, Cellular reprogramming in pursuit of immortality. *Cell Stem Cell* **2012**, 11(6), 748-50.
51. Thales A.F. Albuquerque; Luisa Drummond do Val; Aoife Doherty; João Pedro de Magalhães, From humans to hydra: patterns of cancer across the tree of life. *Biol. Rev.* **2018**, 93, 1715–1734.
52. Hyman Hartman, Speculations on the evolution of the genetic code. *Origins of life* **1975**, 6, 423–427.