mathematics

Article

General Designs Reveal a Purine-Pyrimidine Structural Code in Human DNA

Dana Cohen

Ronin Institute, Montclair, NJ 07043, USA; dana.cohen@ronininstitute.org

Abstract: The human genome carries a vast amount of information within its DNA sequences. The chemical bases A, T, C, and G are the basic units of information content, that are arranged into patterns and codes. Expansive areas of the genome contain codes that are not yet well understood. To decipher these, mathematical and computational tools are applied here to study genomic signatures or general designs of sequences. A novel binary components analysis is devised and utilized. This seeks to isolate the physical and chemical properties of DNA bases, which reveals sequence design and function. Here, information theory tools break down the information content within DNA bases, in order to study them in isolation for their genomic signatures and non-random properties. In this way, the RY (purine/pyrimidine), WS (weak/strong), and KM (keto/amino) general designs are observed in the sequences. The results show that RY, KM, and WS components have a similar and stable overall profile across all human chromosomes. It reveals that the RY property of a sequence is most distant from randomness in the human genome with respect to the genomic signatures. This is true across all human chromosomes. It is concluded that there exists a widespread potential RY code, and furthermore, that this is likely a structural code. Ascertaining this feature of general design, and potential RY structural code has far-reaching implications. This is because it aids in the understanding of cell biology, growth, and development, as well as downstream in the study of human disease and potential drug design.

Keywords: genomic signature; general designs; dinucleotide; non-coding; DNA sequence; genomic DNA; human genome; purine; pyrimidine

MSC: 92B05; 92D20

Citation: Cohen, D. General Designs Reveal a Purine-Pyrimidine Structural Code in Human DNA. *Mathematics* **2022**, *10*, 2723. https://doi.org/10.3390/math10152723

Academic Editor: Alexander Zeifman

Received: 2 July 2022
Accepted: 28 July 2022
Published: 1 August 2022

1. Introduction

1.1. The Genome as an Information Source Containing Patterns and Codes

A genome is both a data source and a program to construct an organism. It functions through codes contained within the DNA sequence [1]. DNA is comprised of four different chemical bases, and the well-known triplet code functions within the protein-coding sequences of the genome. However, in the human genome, coding DNA comprises less than 2% of the sequence. The vast majority, 98%, of the sequence does not code for protein, having other characteristics [2], patterns, and codes that are not yet well understood [3]. We know that this vast array of sequences contains much more information and functionality for gene regulation [4,5]. Furthermore, all genomic DNA including coding sequences possess other (than triplet code) inherent characteristics, or general designs. To decipher these general patterns and codes in the DNA, mathematical and computational methods can be applied [6].

In this study, we employ an alternative philosophy through which to view DNA, and devise a novel experiment in order to further understand the general designs of genomic sequences. The idea was to treat the DNA as a binary information system, with some likeness to a computer system. We call this novel approach a binary components analysis.

1.2. Genomic Signatures and General Designs

Genomic signatures are a powerful tool for assessing general designs of sequences. These signatures are pervasive and stable in genomes [7], and provide an excellent tool for characterizing sequences [8]. This was shown in early experiments where it was observed that the set of dinucleotide odds ratios or 'general design' is stable and is a property of the DNA within a given organism. [9]. Analysis has demonstrated that general designs of random samples of a given genome are substantially more similar to each other than to those of sequences from other organisms. It has also been found that dinucleotide relative abundance profiles are remarkably constant across human chromosomes [10]. The set of dinucleotide odds ratios makes up a unique signature for a given genome, and this is species-specific [11].

Genomic signatures are known from previous research to be extremely stable within a given species. Dinucleotide relative abundance profiles depict the presence of inherent patterns in sequences relative to their random expectation. These relative abundance profiles go far beyond an analysis of sequence composition. So, these describe general designs and baseline patterns and codes within the sequence. They relate to basic mechanisms, sequence assembly, and evolutionary constraints. Since these are baseline codes, they are extremely stable and pervasive.

Dinucleotides can provide basic information about structural and chemical tendencies, and functionality [12]. Odds ratios show if a dinucleotide is enhanced or suppressed beyond random expectation. It, therefore, reveals a deeper dimension of insight, as it describes sequence assembly tendencies, and the random/non-random characteristics of the sequence [13]. The absolute relative abundance is the overall profile or average across all possible dinucleotides. It can be adapted to show how distant from randomness a sequence is in a general sense.

1.3. The Power of the Binary Components Analysis

DNA possesses layers of information. In order to further characterize these, a novel experiment was designed, which broke down the information content of the DNA into binary components. The philosophy used in this research is to treat the DNA as a binary information system, thereby, making use of information theory tools [14,15].

This method seeks to isolate information held within the four bases (letters) of the DNA, to better understand the general designs of the genome, and decipher underlying codes. When doing so with DNA sequences (as opposed to digital information), this subdivision is based on the real chemical or physical properties of the bases/nucleic acids. This is similar to how a computer system separates data into a binary system.

The general design (dinucleotide signatures) analysis is then applied to form the separate binary components. These base properties are then studied in isolation for their genomic signatures, and also for relative random (or non-random) features in the DNA sequences. This reveals the relative importance of each of these isolated properties.

It is important to clarify that by binarizing, we do not suggest that the human genome (or any genome) is simple. Quite the contrary, the codes contained with it are highly complex and multi-layered in nature. Translation of genomic DNA into a binary sequence, and applying the binary components analysis is conducted only as a way of analyzing, extracting, and isolating information. It presents a potentially powerful method, as it breaks down information into component parts.

1.4. The Physical and Chemical Properties of DNA Bases

There are three well-defined categories of DNA base properties [16,17] that are isolated for the binary components analysis (see Figure 1). The first is a physical property: purines and pyrimidines (RY) [18,19]. It describes the chemical ring structure of the bases, with purines possessing a two-ring structure, and pyrimidines a one-ring (see Figure 1a). The purine/pyrimidine content of DNA is known to influence its secondary structure [20,21]. The second is keto and amino bases (KM), which define the chemical property of tautomerism [22] (see Figure 1b). The third property is weak and strong bases (WS), which defines hydrogen-bonding between the base pairs, with C and G pairing with three hydrogen bonds, and A and T with two hydrogen bonds [23] (see Figure 1c).

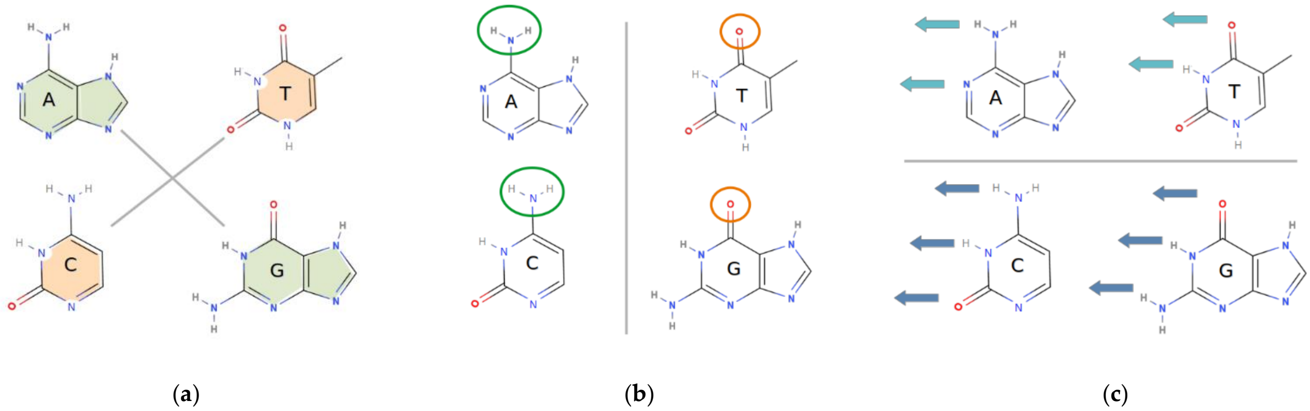


Figure 1. The chemistry behind, and reason for the conversion to binary components. The four DNA bases, A, T, C, and G, each contain distinctive, and yet overlapping chemical and physical properties. This diagram illustrates those properties. It shows the subdivision into three separate categories, which are subsequently used for the binary components analysis. The binary components analysis is about breaking down the information content of the DNA bases into binary parts or components. There is an isolation of targeted chemical and physical properties at the exclusion of other properties. Here, information content of the bases of DNA is isolated into parts so that they can be studied separately. (**a**) The DNA bases contain either a two-ring or one-ring molecular structure. The diagram shows the division into two-ring (green) or one-ring (orange) structures. The two-ring structures are called purines, and A and G are purines (R). The one-ring structures are pyrimidines, C and T and pyrimidines (Y). The presence of purines and pyrimidines in DNA, and the ratios of these in the DNA, will affect the secondary and tertiary local structure of molecules. (**b**) The next category is the chemical property of an amino or keto group. The bases each contain one of these. The diagram shows the amino group (-NH2) circled in green, and the keto group (CO) circled in orange at the relevant position. The bases A and C contain an amino (M) group, whereas G and T contain a keto (K) group. This chemical property affects the hydrogen-bonding capability of the bases, and specifically the hydrogen-bonding donor-acceptor patterns positioned at the major groove of the DNA. These hydrogen-bonding donor-acceptor sites are there for the binding of other particles, such as regulatory proteins to the DNA, during a variety of biological processes. (**c**) The last category is the weak or strong hydrogen-bonding capacity of the DNA bases. This is with reference to the number of hydrogen bonds between the base pairs of the DNA. A and G are weak (W) bases because these contain only two hydrogen bonds between the complementary base pairs, and C and G are strong (S) bases, as these contain three hydrogen bonds between the base pairs. The diagram shows the number of potential hydrogen bonds at the relevant position on each base as blue arrows. The dark blue arrows depict three potential sites, and light blue arrows two potential sites. This property can also influence minor groove hydrogen-bonding patterns. These minor groove sites are for potential binding of other particles to the DNA.

In this study, we observe that the KM base classification also reflects another extremely crucial property. This is the hydrogen bond donor-acceptor pattern in the major groove of DNA. Additionally, it is noted that RY is a physical structural property while WS/KM is a chemical property. A conversion of DNA sequences from an ATCG sequence into each of RY, WS, and KM sequences is a breakdown of information content, which isolates one type of property (see Figure 2) [24].

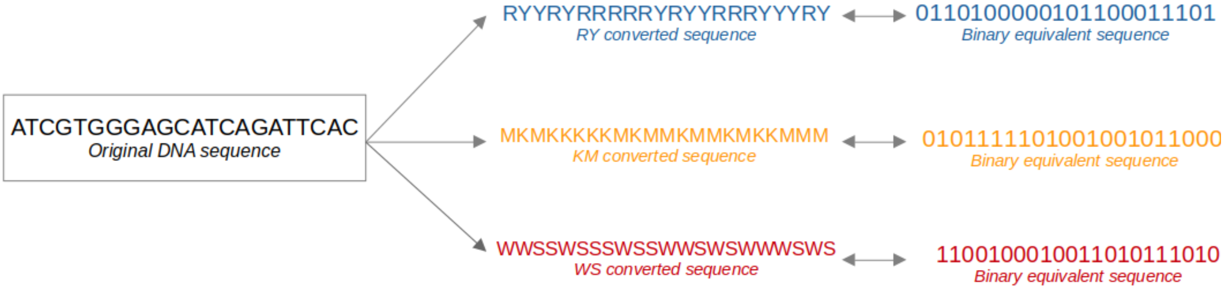


Figure 2. Conversion of original DNA sequence into binary sequences. The original DNA sequence is converted or translated into three different sequences. The diagram shows conversion to a purine/pyrimidine (RY) sequence, keto/amino (KM) sequence, and weak/strong (WS) sequence. It also depicts an equivalent binary sequence for each represented by zero or one, where the assignment of either a zero or one is arbitrary. Each of the RY, KM, and WS conversions, is a separate type of conversion to a binary 'type' of sequence. RY sequence: A and G are converted to R, and C and T are converted to Y; KM sequence: A and C are converted to M, and G and T are converted to K; WS sequence: A and T are converted to W, and C and G are converted to S.

In this study, we conceptually view the DNA sequence as an information stream, and an encoding or decoding system in which there are exactly two possible states. A binary system offers a simple and elegant way for computers to work, and in the context of DNA, the same may be true, even though the information contained within the bases is more complex. Since the DNA sequence carries codes within it, this may be viewed and treated in principle in a similar way to a computer system. The separation of base properties into binary values is powerful, as chemical and physical properties are isolated. The significance of these isolated properties can in turn be studied within the DNA sequence. It lends itself to the study of general design analysis and elucidation of random/non-random properties of sequences.

The RY, WS, and KM properties of nucleotides are long and well established. The conversion of DNA into these is a known entity. Early experiments sought to gain insight through these conversions. A more recent experiment [25] has binarized codons, in order to gain an understanding of how codons operate, and their transmission to amino acids. These, of course, are applied specifically to coding DNA.

The research carried out is novel in the application and implementation of the binary conversions. Here, the goal is to extract information from genomic DNA, to gain insight into how this DNA is encoded. We seek to make progress in deciphering new codes and understanding genomic DNA on a global scale. A complex and multi-layered code(s) which is not understood. The aim is to use this binary system as a tool to extract information. The utilization of binary sequences to understand general designs is novel. Here, the concept of general designs is developed in a new way, to assess the relative importance of RY, WS, and KM in the sequence, by assessing the relative non-randomness of these components.

1.5. Aim of Experiment and Importance of Findings

The aim of this work is to explore concepts and methods for the analysis of genomic DNA sequences, to deepen our understanding of general designs. To this end, we analyze human chromosomal DNA, adding insight and dimension to previous research. The human genome is used since it permits large-scale analysis, is well characterized, and is a genome of importance. Genomic signatures are studied, utilizing dinucleotide odds ratios, and relative abundance profiles showing the distance from the randomness of sequences. This is conducted for each of the RY, KM, and WS binary components in order to compare them.

It is concluded that the RY property of a sequence is the least random in all chromosomes, and therefore, there exists a pervasive, widespread RY code in genomic DNA. This is likely a structural-based code since the RY sequence is a strong determinant of local DNA structure. This has far-reaching implications for our understanding of how genomic DNA functions. Since the genome contains the information and program to construct the organism, in a wider sense this aids in our understanding of cell biology, growth, and the development of humans. It can also have implications for identifying diseases, and potential drug design.

2. Methods, Data and Concepts

2.1. Obtaining Genomic Sequences from Human Genome Database

The DNA sequences for this project were taken from the NCBI human genome database, build 38. The 24 human chromosomes sequences were obtained and treated individually. Each chromosome was 'divided' into 100 kb discrete segments of DNA. There were non-overlapping continuous portions. All sequence data was analyzed from the 5′ to 3′ end, in this direction for both strands, therefore, the total length or number of mononucleotides for both strands amounted to 200,000, for each 100 kb portion. The 100 kb portions form a large dataset (*n*) are synonymous with the 100 kb portions previously utilized for dinucleotide relative abundance studies. These large segments are used because they form long sequence lengths that are a large enough sequence length for the various analyses carried out in this research. The logic behind the 100 kb sequence windows is that these stretches of DNA are of suitable length to determine relative over- or under-representation of dinucleotides in sequences, with statistical analysis. This was gained by experience from previous research carried out with genomic signatures, and this present analysis builds on this.

2.2. Mononucleotide and Dinucleotide Frequencies

Mononucleotide and dinucleotide frequencies were calculated for each of the 100 kb double stranded portions. Mononucleotides are the classic A, T, C, and G chemical bases represented by these four letters. There are a total of sixteen different possible dinucleotides; ApA, ApT, ApC, ApG, TpA, TpT, TpC, TpG, CpA, CpT, CpC, CpG, GpA, GpT, GpC and GpG. Each of these dinucleotides was determined stepwise along the sequence from the 5′ to the 3′ along the transcribed strand, and also the anti-sense strand from the 5′ to 3′ end. For each base step, the next base was determined. The length of the segments is 100 kb, and so for the sense strand this is 100,000 bases, plus 100,000 bases for the anti-sense strand, so this is a total of 200,000 mononucleotides for one chromosomal segment. The total number of dinucleotides is 198,000. These calculations were performed for all 100 kb sequence portions individually within a chromosome, and for each of the 24 chromosomes separately.

2.3. Genomic Signatures: Odds Ratios and Relative Abundance Profiles

The dinucleotide representation is a value that can be used to assess dinucleotide contrasts while taking into consideration the mononucleotide composition of the sequence. This describes the frequency of each dinucleotide in the sequences, above or below the random expectation. Dinucleotide representation was calculated by using an odds ratio [11]. The odds ratio was calculated for each of the 100 kb fragments along the length of a given chromosome, and this was repeated for all of the twenty-four chromosomes using the odds ratio.

$$\text{Dinucleotide odds ratio:} \quad \rho_{XY} = f_{XY}/f_Xf_Y$$

f_X is the frequency of the nucleotide X within the sequence and f_{XY} is the frequency of the dinucleotide XpY within the sequence. The result obtained from a frequency or count of nucleotides (and dinucleotides) is then multiplied by *n* (where *n* = length of sequence) in order to standardize the odds ratio. A value of $\varrho_{XY} > 1$ indicates over-representation of the dinucleotides, whereas $\varrho_{XY} < 1$ indicates under-representation. In a random sequence (IE. a shuffled sequence) the ϱ_{XY} values for all the dinucleotides approach 1.0. [13,24]. The odds ratios of the sixteen dinucleotides form dinucleotide relative abundance profiles, whose different from 1, provide a measure of deviation from randomness. It has been determined from previous experiences [10] that for a random sequence the ϱ_{XY} values have the following relationship: the deviation from 1 is approximately $1/\sqrt{n}$. For n~200,00, $|\varrho_{XY} - 1| = 0.0022$.

The composition or frequency of dinucleotides in a given sequence provides some information about its basic make-up. This is because the dinucleotide is the simplest word or pattern, and this describes in simple form the language of DNA. The dinucleotide odds ratio, however, is a measure of suppression or enhancement of that dinucleotide in any given sequence, since it takes into consideration (and normalizes) the nucleotide composition of the sequence. It may also be utilized as a measure of non-randomness/randomness, since it describes the proportion of each dinucleotide, above or below the random expectation given constituent mononucleotides. This is important since it reveals tendencies of sequence assembly, structure, and functionality. It reveals general designs and takes together a DNA signature. The entire complement of (sixteen) dinucleotide odds ratios forms a genomic signature profile for any given sequence.

A measure of distance between two sequences either within or across organisms has been referred to as the average absolute dinucleotide relative abundance difference [9,11]. This has been used to cross compare different genomic signature of different sequences and is calculated as follows:

Average absolute dinucleotide relative abundance difference:

$$\delta\,(f, g) = 1/16 \sum_{XY} |\varrho_{XY}\,(f) - \varrho_{XY}\,(g)|$$

They may be used as a measure of distance between two sequences either within or across organisms, as it compares the genomic signatures, and has been referred to as the average absolute dinucleotide relative abundance difference. Here, f is one sequence type whereas g is another. For example, f may be a human sequence whereas g a mouse sequence. $\varrho(f)$ is the odds ratio value for a dinucleotide for one sequence type. $\varrho(g)$ is the odds ratio for the same dinucleotide within the second sequence type. The sum, here, extends over all sixteen possible dinucleotides.

It permits cross-comparison of different sequence types within and between organisms. Beyond this the dinucleotide relative abundance profiles can also be used to demonstrate a departure from randomness of genomic DNA sequences. The average absolute dinucleotide relative abundance may be adapted to measure the difference between a real genomic sequence and a randomized one. This permits a measure of distance from randomness, or random model.

Distance from randomness:

$$\lambda\,(f, g) = 1/16 \sum_{XY} |\varrho_{XY}\,(f) - 1|$$

This is with respect to all possible dinucleotides. Therefore, the absolute dinucleotide relative abundance (δ) profile is adapted here to measure the general deviation from the randomized model. The distance from randomness adaptation (λ) can be employed under certain conditions. This would be relevant where n (sequence length) is long enough. Where n is large, for instance if $n >= 100{,}000$, for any given dinucleotide, the odds ratio for a randomized DNA sequence approaches 1. For random sequences of equivalent mononucleotide composition, the dinucleotide odds ratio will have a value of one. In this instance, where f is a 'real' tested genomic DNA sequence, and g is an equivalent randomized (or shuffled) sequence, and n is large, ϱ_{XY} (g) approaches 1. Therefore, instead of subtracting two different sequences, the dinucleotide odds ratios of real actual sequences are subtracted from the theoretical odds ratio of a randomized equivalent sequence. For such a sequence sample, the distance from measure (λ) may be used. This is useful for ascertaining the relative ‘randomness’ or non-randomness profiles for sets of sequences, which may be used for cross comparison.

The sum here extends over all sixteen dinucleotides, and therefore, this calculation is the average of odds ratio values for all sixteen possible dinucleotides for a particular sequence. Additionally, this calculation does in fact compare two types of sequence. However, instead of comparing two real genomic DNA sequences, a real sequence is compared with its random equivalent, since the theoretical odds ratio value for the random sequence is 1.0. Therefore, this value may also be regarded as an average deviation of dinucleotides from the expectation of a random sequence of equivalent mononucleotide proportions. This calculation is referred to as the ‘distance from randomness’ (λ value) of the sequence in this research. The further this value is from zero, the further away the value is from the random (or shuffled) sequence. This total distance from the randomness value was calculated for each of the chromosomal sequence datasets.

2.4. Components Analysis: Odds Ratios, Relative Abundance and Distance from Randomness

For the binary components analysis, all the genomic sequences were ‘translated’ or converted into three different sequences. Here, for each conversion type specific properties of the bases are isolated (see Figure 2). The first dataset is the translation of the original ATCG sequence to a purine/pyrimidine (R/Y) sequence. This is the physical structural property of the bases of DNA. Here, the bases A and G were converted to R (purines), and C and T were converted to Y (pyrimidines). This yield two possible states, and hence a binary sequence, where each base is assigned as either R, or Y. This was performed for all 100 Kb sequence segments, and across all 24 chromosomes. The second dataset is the translation of the original sequence into a weak/strong (W/S) sequence, where A and T are converted to W (weak bases), and C and G are converted to S (strong bases). This deals with the chemical property of hydrogen bonding between the base pairs. The third dataset is the translation of the original sequence into keto/amino (K/M) sequence, where A and C are converted to M (amino bases), and G and T are converted to M (keto bases). This deals with the chemical property of donor-acceptor sites on the major groove of DNA. This translation for the components analysis, therefore, yielded three distinct sequence datasets in addition to the original ATCG dataset for the genomic DNA sequences. This results in three separate converted/translated sequences which were treated as separate entities for analyses. This was conducted so that the relative importance of these two subdivisions of nucleotide properties could be assessed individually.

Now for each of these three binary components datasets, RY, KM, and WS, there are four possible dinucleotides (instead of the sixteen dinucleotides). The dinucleotides for each of these datasets are as follows:

1. The purine/pyrimidine RY dataset: RpR, RpY, YpR, and YpY;
2. The weak/strong WS dataset: WpW, WpS, SpW, and SpS;
3. The keto/amino dataset: KpK, KpM, MpK, and MpM.

An equivalent data analysis was carried out for each of the three different binary components, as described above for the ATCG original sequence. The calculations we adapted for four dinucleotides, instead of sixteen dinucleotides for the original DNA sequence. This includes the odds ratio, dinucleotide relative abundance, and distance from randomness. The adaptions for these are as follows:

Average absolute dinucleotide relative abundance difference for the binary components would be:

$$\delta\ (f, g) = 1/4 \sum_{XY} |\varrho_{XY}\ (f) - \varrho_{XY}\ (g)|$$

Therefore, the distance from randomness is:

$$\lambda\,(f, g) = 1/4 \sum_{XY} |\varrho_{XY}\,(f) - 1|$$

Odds ratios, and distance from randomness values were calculated for each of the RY, KM, and WS datasets separately, utilizing these expressions. These were performed for the 100 kb sequence segments, as before for double stranded DNA, and as described about for each of the 24 chromosomes individually. Then statistical analysis was carried out to compares the three binary components. The components were analyzed for their dinucleotide enhancement and suppression. Additionally the distance from randomness profiles were compared for relative importance of the binary components in chromosomal DNA, in order to understand which of the chemical and physical properties are most significant.

2.5. Sequence Analysis and Statistics

A Python script was used to process the DNA sequences for each chromosome, and segmenting the sequences (for the calculations), and implementing this for forward and reverse strands. This was conducted for calculating mononucleotide and dinucleotide frequency, as well as odds ratios, relative abundance and distance from randomness values. For each chromosome the 100 kb sequence portions were taken as separated entities. The dinucleotide frequencies, odds ratios, and relative abundance calculation were taken separately for each 100 kb sequence fragment, and forward and reverse strands used from 5′ to 3′ end. Then for each chromosome statistics were calculated, including mean, median, standard deviation, variance, and quartiles, for all the fragments within each chromosome. This resulted in statistics and analysis for each chromosome separately.

Further statistical analysis was carried out to determine whether there is significant difference between the RY, KM and WS binary component sequence datasets. This significance testing was performed using IBM SPSS statistics software. To this end ANOVA tests were applied to the relative abundance values for all sequence segments (sample set) within any one given chromosome. The ANOVAs were carried out separately for each of the 24 human chromosomes.

Ho: *The null hypothesis of is that there is no difference between the RY, KM and WS datasets.*

H1: *The alternate hypothesis is that there is a significant difference.*

24 separate one-way within-subject ANOVA-s were performed, comparing the RY, KM and KM analysis types in each chromosome separately (see Table A1), evaluated at the 5% level. Since in all instances the sphericity assumption—evaluated in advance by the Maulchy test—was indicated to be violated, the reported results below are all based on the Greenhouse-Geisser correction. To further explore the differences between RY, KM and WS dataset, the ANOVA-s were followed up in the form of post hoc multiple pairwise comparisons corrected by the Bonferroni method.

2.6. Real DNA Sequences Verses a Random Model

A random model is generated for the dinucleotide relative abundance profiles. This is conducted in order to analyze the results, and see if the output is significantly different to the random expectation. To this end, each of the discrete 100 kb sequence segments were randomized, and this was performed for sequence data for all of the chromosomes. The EMBOSS ‘Shuffleseq’ tool was used to generate a randomized DNA sequence, while retaining the mononucleotide composition. This generated a set of random or shuffled sequences that were equivalent to each of the 100 kb sequence segments. This random dataset was then converted to binary component RY, WS, and KM sequences, exactly as with the original ‘real’ dataset, and the same algorithm applied to generate odds ratios and dinucleotide relative abundance profiles, utilizing the identical pipeline.

Statistical analysis was carried out to test the difference between real and random models for each of the binary components (RY, WS, and KM), in a like-for-like comparison. It was carried out for the dinucleotide relative abundance within each chromosome. A two-tailed paired t-test, at the 5% significance level was used.

Ho: The null hypothesis (H0) is that there is no difference between real and random sequences for each binary component (RY/WS/KM).

H1: The alternate hypothesis is that there is a difference between real and random sequences.

3. Results and Discussion

3.1. Distance from Randomness Comparison of Binary Components RY, KM, and WS

Dinucleotide distance from randomness (relative abundance (λ)) results show that each of the RY, WS, and KM binary components have very distinct profiles within any given human chromosome (see Figure 3). This same overall trend is seen across all the 24 chromosomes, so that each of the components is similar in its genomic signature and distance from randomness values across all the chromosomes. This is true also for the original DNA (ATCG) sequence, and so the binary components are also remarkably constant across human chromosomes. The RY component has more variation across the chromosomes than the other components. Overall, the genomic signature for each of the RY, WS, and KM components is stable and pervasive across the chromosomes, and behaves in a similar way with regard to the original ATCG sequence.

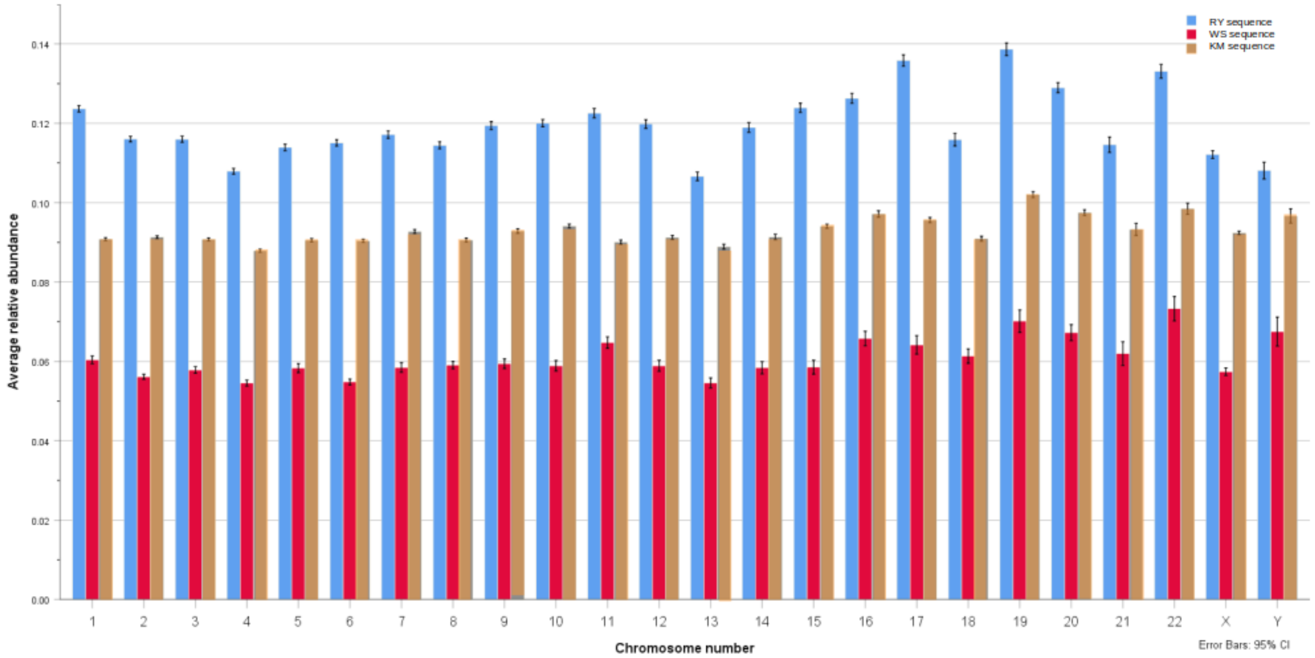


Figure 3. Binary components' distance from randomness. The dinucleotide relative abundance—distance from randomness values show that each of the RY, WS, and KM binary components have very distinct profiles within any given human chromosome. A value of zero is synonymous with a randomized sequence or model. Within each of the human chromosomes, the RY binary sequence component is the least random, followed by the KM sequence, and then the WS sequence, which is the closest to randomness. This suggests that RY is the most prominent and important feature in terms of general designs of DNA. Since purines/pyrimidines are structural properties and determinants of DNA secondary structure, this suggests a potential RY structural code that is prevalent in human DNA in general, and across all chromosomes. KM and WS components

determine the chemical properties of the bases, including hydrogen-bonding capability (and patterns), and this property is closer to randomness, and therefore, of lesser prominence.

A distance from the randomness (λ) value of zero is synonymous with a randomized sequence or model. The higher the value, the more distant from randomness the sequence is considered to be. Within each of the human chromosomes, the RY binary component is the least random, followed by the KM sequence, and then the WS sequence, which is closest to randomness. For instance, for chromosome 22, the distance from the randomness (λ) value is: RY, 0.133; KM, 0.098; and WS, 0.073. This general trend of the RY binary component being the least random, followed by the KM, and the WS, is true for all the chromosomes. Given that for each chromosome, synonymous sequences are analyzed, while the sequences are converted/translated into RY, KM, and WS, we can see the relative random or non-random characteristics of each of these components.

The binary components were found to be significantly different from each other within each chromosome. In order to investigate the significance of these results, one-way ANOVAs were performed, comparing the RY, KM, and WS's relative abundance values within each of the 24 chromosomes (see Appendix A, Table A1). In all instances the sphericity assumption—evaluated in advance by the Maulchy test—was indicated to be violated, the reported results are all based on the Greenhouse–Geisser correction. The results revealed that for all 24 chromosomes a highly significant effect, and the effect size being large in all cases, therefore, the RY, KM, and WS datasets are significantly different from each other in human genomic DNA.

To further explore the differences between analysis types, ANOVAs were followed up in the form of post hoc multiple pairwise comparisons corrected by the Bonferroni method. Once again, similar results were found in all instances, with all comparisons being found to be statistically significant at the 1% level. For each of the chromosomes, the same patterns emerged, where RY analysis types showed the highest relative abundance (and least random) values on average, followed by KM having approximately 20–25% lower values, and finally, the WS type having the lowest and closest to random values.

A comparison of real versus random sequences was carried out for the dinucleotide relative abundance profile for each of the RY, WS, and KM sequences. The results showed that for each of these, the real sequence dataset is significantly different to the random one at the 5% level of significance (see S4: file Statistics_random_model.xlsx for results tables). This effect is true across all of the chromosomes. Even though the WS component is generally closest to the random model, it is still significantly different from it. The results were that p-values are consistently close to zero, and the dinucleotides are neither enhanced nor suppressed in the random model. In contrast to this, with the real DNA sequences, the odds ratios revealed consistent enhancement or suppression as outlined. Furthermore, the behavior of the randomized sequence model is as expected and in line with previous observations where the odds ratios are close to zero.

This result suggests that RY is the most prominent and important feature in terms of general designs of DNA. Since RY sequence is associated with DNA structure [20,21], it implies that structure is the least random and most important feature of chromosomal and genomic DNA. This physical property of the structure is less random, and therefore, more significant than the chemical properties (WS/KM). Purines and pyrimidines are determinants of a DNA's local secondary and tertiary structure, and so this suggests a potential RY structural code that is prevalent in human DNA in general, and across all chromosomes.

In contrast to this, the KM and WS components determine the chemical properties of the bases, including hydrogen-bonding capability (and patterns). These include the hydrogen-bonding donor-acceptor sites on the major groove of DNA and hydrogen-bonding between the two strands.

It is known that the RY make-up of DNA determines the relative flexibility/rigidity of the molecule, and the ability to bend. At any given segment of the DNA, this will change depending on the sequence. The analysis here of the non-randomness of identical sequences with respect to each of these three components seeks to determine their relative general importance in these sequences.

A potential KM or WS chemical code is generally of lesser importance in comparison to RY. The results indicate that hydrogen-bonding patterns are the least important feature in the genomic DNA across all the chromosomes. Furthermore, the WS signature is the most random (relatively speaking) of the three. This means that the KM signature, which relates to tautomerism patterns and hydrogen-bonding patterns on the major groove, is more important in the chromosomal DNA. Given existing knowledge of the biological function of the vast proportion of genomic DNA sequences, this makes sense. The vast amount of DNA sequences are non-coding, and contain particle (transcription factor) binding sites and regulatory elements [26,27], and the sequences reflect this. This result indicates that the local structure of the DNA is more important than its chemical properties for this functionality.

3.2. Genomic Signatures and Odds Ratios for the Binary Components RY, KM, and WS

The dinucleotide is an extremely powerful tool for understanding DNA. It is the most basic description of a sequence beyond composition, as it is the simplest motif or pattern, considering each base step and its nearest neighbor. Dinucleotides can provide basic information about structural and chemical tendencies, and functionality. The dinucleotide odds ratio, however, reveals a deeper dimension of insight, as it describes the sequence assembly tendencies, and the random or non-random characteristics of the sequence. Dinucleotides and the odds ratio profile reflect deeply sequenced patterns and codes. Dinucleotide odds ratio differences from a value of one (or zero, where results are scaled to zero) reflect contrasts between the actual dinucleotide frequencies, and those expected from random sequences of equivalent mononucleotide frequencies [13,24]. The more distant the value is from one, the greater the departure from the random expectation. Values above one mean a given dinucleotide is over-represented, while a value less than one means it is under-represented.

The results of this study and odds ratio profiles for the binary components reveal much about the human chromosomal DNA. The odds ratios for the RY component show that all four dinucleotides display an equally non-random profile (see Figure 4a). However, RpY/YpR dinucleotides are enhanced above the random expectation given the nucleotide content, and RpR/YpY are suppressed, or below the random expectation. This profile is consistent for all the chromosomes. The RY property of the DNA influences the secondary structure and tertiary structure of localized regions of the DNA [12]. The relative rigidity and flexibility of the molecule are dependent on RY [28]. Specifically, the RpY/YpR dinucleotides are flexible, while RpR/YpY are rigid, and so, in general, flexibility is enhanced while rigidity is suppressed in human DNA. The RY genomic signature likely reflects a necessary balance of rigidity and flexibility of DNA structure, and an optimal general structural design that is required for biological function.

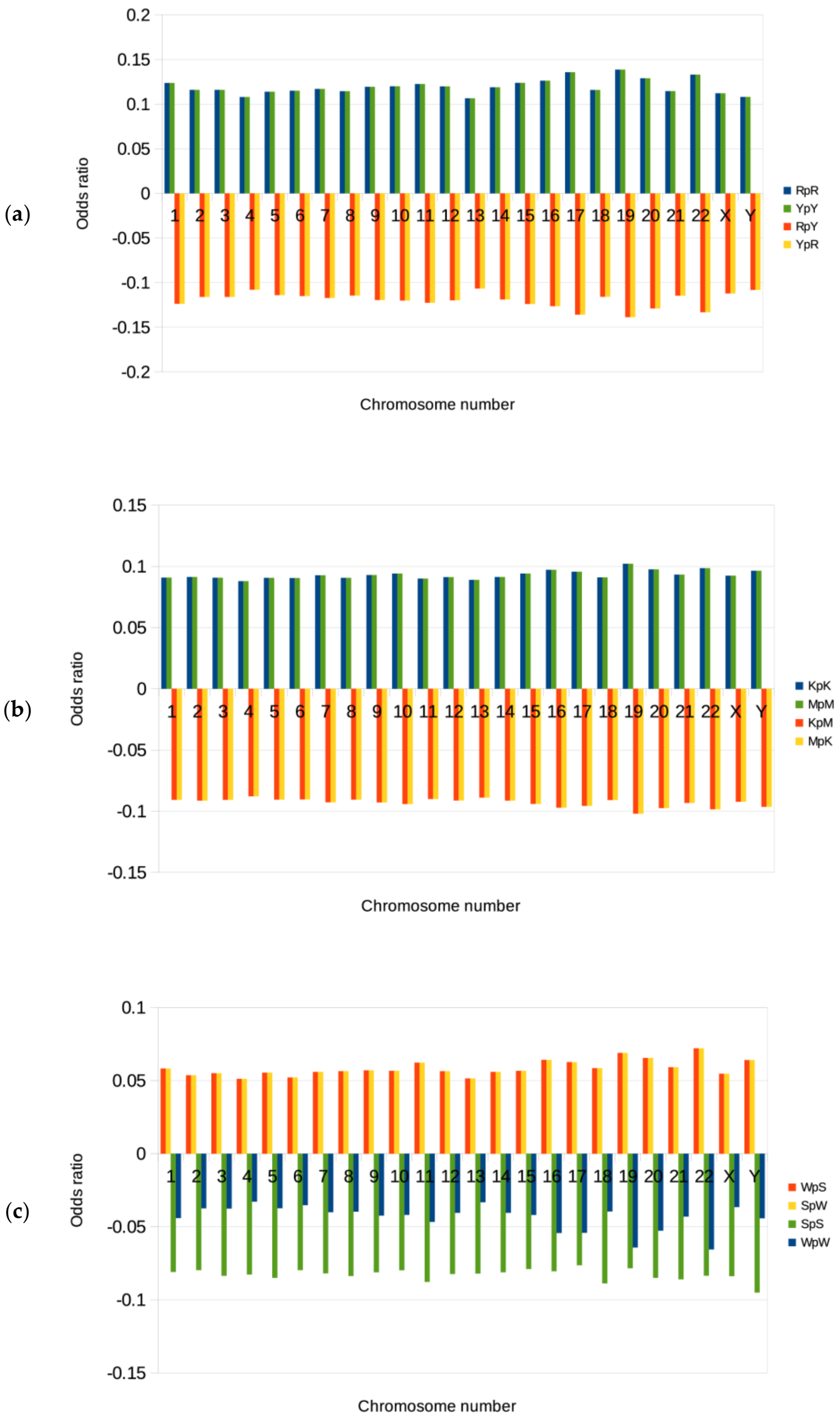


Figure 4. Dinucleotide odds ratios for the binary components. These charts show odds ratios for the four different dinucleotides in each of the RY, KM, and WS sequence datasets, across all the human chromosomes. The odds ratios shown are the mean values, scaled to zero for each of the chromosomes. A value of zero is synonymous with a randomized sequence or model. A value above zero for a given dinucleotide means that it is over-represented or enhanced in the dataset, and if it is less than zero, it is under-represented or suppressed. (**a**) RY binary component: In this sequence dataset, there are four possible dinucleotides, which are RpY, RpR, YpR, and YpY. The dinucleotides RpY/YpR are enhanced above the random expectation given the nucleotide content,

and RpR/YpY are suppressed, or below the random expectation. This profile is consistent for all the 24 chromosomes. RpY/YpR dinucleotides are flexible, while RpR/YpY are rigid, and so, in general, flexibility is enhanced while rigidity is suppressed in human chromosomal DNA. (**b**) KM binary component: KpM/MpK are suppressed, while KpK/MpM are enhanced across all the chromosomes. This pattern directly affects hydrogen-bonding donor-acceptor sites on the major groove of the DNA. Here, either sequential donors or sequential acceptors are favored. It is likely that this pattern is generally more favorable for the binding of particles to the DNA. (**c**) WS binary component: WpW and SpS are suppressed, while WpS and SpW are enhanced across all the chromosomes. Genomic DNA in general favors and enhances dinucleotides with a heterogeneous hydrogen-bonding donor-acceptor pattern in the minor groove of the DNA, and also a heterogeneous number of hydrogen-bonds between the base pairs. In contrast, homogeneous patterns are suppressed. It may be that this heterogeneous pattern of hydrogen bonding between base pairs is favored as it increases the stability of local secondary helical structures.

The RY binary component distinguishes and is directly connected with the physical structure of the DNA. It is a fundamental biological principle that structure is connected with function. More specifically though, structure affects the recognition of particles, which is a crucial property [29,30]. DNA functions via binding to other biological particles, such as a wide array of transcription factors, it is necessary, in particular for gene regulation, as well as other functions of the DNA [31,32]. The fact that the RY component is least random with respect to the dinucleotides' relative abundance profile when compared to the other components places structure and biological recognition as the least random and most significant properties of the DNA. It is this property that is most prominent for the general functionality of genomic DNA. This result points strongly to RY patterns forming a general underlying code, which is a novel idea and observation.

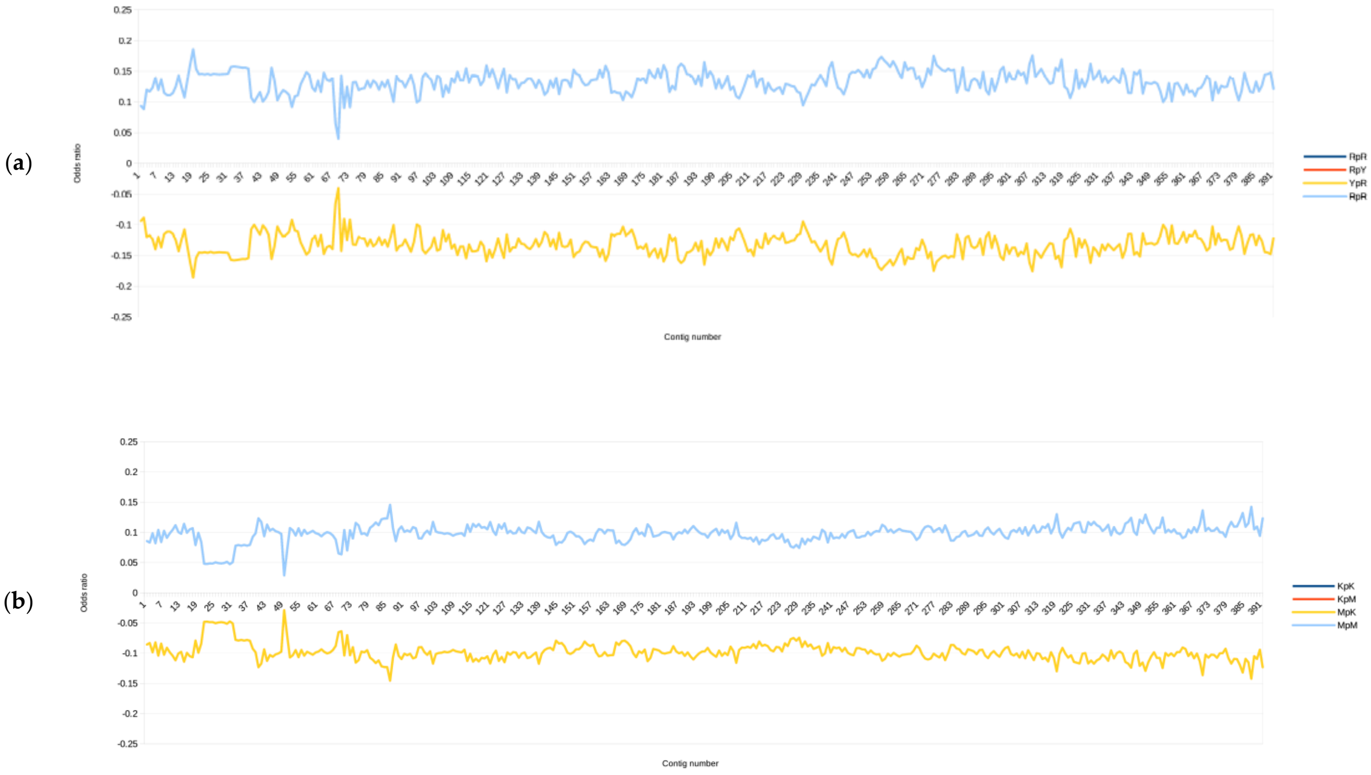
The KM binary component distinguishes the chemical properties of hydrogen-bonding donor-acceptor patterns [33] in the major groove of DNA. K (G/T) bases possess a hydrogen-bonding acceptor site, whereas M (A/C), has a hydrogen-bonding donor site [34,35]. The odds ratio results show that KpM and MpK are suppressed, while KpK and MpM are enhanced across all the chromosomes (see Figure 4b). Therefore, for KM sequences, homogeneous dinucleotides are enhanced, and heterogeneous ones are suppressed. This pattern directly affects hydrogen-bonding donor-acceptor sites, where either sequential donors or sequential acceptors are favored in the major groove. Since the major groove of the DNA is the primary location of direct chemical bonding interaction of the DNA with other biological particles (such as regulatory proteins), this pattern is likely to assist in such bonding. It may be related to the recognition and specificity of binding to biological particles [36]. This result points to some general features of chromosomal DNA with respect to the KM binary component, and the bias for these dinucleotide patterns points to a potential bias in the hydrogen-bonding code. Protein-DNA binding is an essential functionality of the DNA.

The WS binary component distinguishes the chemical properties of hydrogen bonding between the base pairs of DNA [22]. This is a property that also affects the overall stability of the double helix and its ability to separate (the two helical strands) for necessary biological functions. The WS component additionally differentiates between the presence or absence of a hydrogen bond donor in the minor groove of the DNA. The odds ratio results show that WpW and SpS are suppressed, while WpS and SpW are enhanced across all the chromosomes (see Figure 4c). This is also in line with numerous observations that CpG (SpS) is generally suppressed [37,38] in many genomes. This means that genomic DNA favors and enhances dinucleotides with a heterogeneous hydrogen-bonding donor-acceptor pattern in the minor groove of the DNA, and also a heterogeneous number of weak or strong hydrogen bonds between the base pairs. This observation in the context of hydrogen-bonding donor-acceptor sites, though, is novel. In contrast, homogeneous patterns are suppressed.

The localized heterogeneity of hydrogen bonding between the two strands of DNA may increase the stability of a double helix, while at the same time permitting the separation of the two strands. This is because too many strong (three hydrogen bond) connections between the base pairs make it more difficult for the two strands to separate, while too many weak (two hydrogen bond) connections make it too easy, destabilizing the double helix. We conclude that this is the likely reason why DNA favors and drives heterogeneity for this feature.

Regarding the minor groove surface hydrogen bond donor-acceptor sites, the DNA enhances a variety of potential donor-acceptor patterns [39]. DNA functions via contact with a vast host of protein particles, and these interactions rely on hydrogen bonding. While the majority of interactions occur at the major groove, the minor groove is also utilized for this. Heterogeneity is enhanced here too, and the outcome is a greater variety of localized donor-acceptor sites.

Next, we consider whether the binary components of odds ratio observations are stable and consistent throughout the length of the chromosomes. Genomic signatures for the binary components are stable across genomic DNA. Within a given chromosome, the RY, KM, or WS odds ratio values are constant for sequence segments along the chromosomal DNA (see Figure 5). As an example, using chromosome 22, there is some limited fluctuation, however, the general trends persist. This is true for all chromosomes and is in line with previous studies that show this stability with the original (ATCG) DNA sequence. Those dinucleotides that are over-represented are this way throughout the chromosome. Those that are under-represented, possess this characteristic across the entire chromosome. This further demonstrates the stability of the binary components of the genomic signature within human chromosomes.



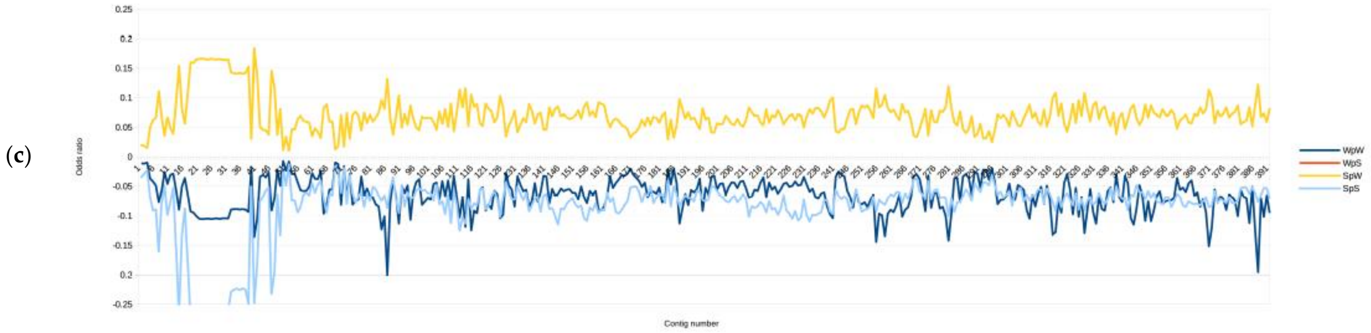


Figure 5. Dinucleotide odds ratios for each of the binary components, for 100 KB sequence segments in chromosome 22. These three graphs show traces of each of the binary components of dinucleotides across their entire length. We see that the dinucleotide's odds ratio patterns are pervasive and stable across the chromosome. This is true for all three RY, KM, and WS components. The enhancements and suppression of the dinucleotides are seen to be constant. Those dinucleotides that are over-represented are this way throughout the chromosome. Those that are under-represented possess this characteristic across the entire chromosome. This further demonstrates the stability of the genomic signature within a human chromosome. These graphs for chromosome 22 are given as an example, however, the same stability and constant profile are true across all the human chromosomes. (**a**) RY binary component: The dinucleotides RpY/YpR are enhanced, and RpR/YpY are suppressed throughout the chromosome, and this profile is remarkably constant and stable. (**b**) KM binary component: The dinucleotides KpM/MpK are suppressed, while KpK/MpM are enhanced throughout chromosome 22. (**c**) WS binary component: Towards the telomeres and particularly at one end, there is more variation in the signature, however, it still holds. Telomeres display more repeats and sequence structures that are different, and the signatures highlight this. The WS component fluctuates the most in this regard. The WS component shows more variations, particularly towards the telomeres.

For the RY binary component (see Figure 5a), the dinucleotides RpY/YpR are enhanced, and RpR/YpY are suppressed throughout the length of chromosome 22, and this profile is remarkably constant and stable. For the KM binary component (see Figure 5b), the dinucleotides KpM/MpK are suppressed, while KpK/MpM are enhanced throughout.

Regarding the WS binary component (see Figure 5c), however, towards the telomeres, particularly at one end, there is more variation in the signature, however, it still holds. For all three binary components, the sequence approaching the telomeres shows more signature fluctuations. This is likely due to the nature of the telomeres, which may have more repeat sequences, and sequence structures that are different [40]. The WS component fluctuates the most in this regard. The WS component shows more variation in this regard.

3.3. Information Content: Patterns and Codes in the DNA

DNA contains information and a program for building the organism, in this case, the human. This information content is both compartmentalized and layered. We know that the vast array (more than 98%) of human DNA sequences are non-coding and do not function via the triplet code. The non-coding DNA functions in gene regulation. These sequences are dense with information content, and therefore inherently have patterns and codes. This study seeks to gain more understanding of the general designs of human DNA and the layers of code it contains. It seeks to build general rules as to how DNA sequences carry information. The results provide a new dimension to general design properties via the binary component analysis.

Dinucleotide odds ratios reflect the chemistry of dinucleotide stacking energies or what might be viewed as sequence assembly tendencies. Additionally, these reflect the structural properties of the DNA and the base-step conformational preferences, in any given genome or genomic sequence type. The stability of the genomic signature also suggests that there may be physical or chemical factors that impose limits on the general compositional variation of the genome.

Since there exists an inherent connection between sequence, structure, and function, analysis of the DNA sequence and its signatures permits the understanding of structure and function. The dinucleotide biases provide a genome signature that is characteristic of the bulk properties of an organism's DNA. Furthermore, patterns and signatures help build rules for how these sequences are constructed or assembled. It is thought that these properties of the genomic signature may be due to the existence of genome-wide factors. Examples include the replication and repair machinery, mutational tendencies, and structural tendencies of genomic DNA. However, the genomic signatures also reveal more about the information contained within the DNA.

There is a connection between DNA functionality and the non-randomness of sequence. The RY binary component is the least random, and it suggests an RY code. This is likely a structural code. In the genomic DNA, a chemical hydrogen-bonding code is also present, but of lesser importance. Structure affects recognition rather than direct chemical bonding. This recognition is crucial for information transfer. Since this is pervasive across all the human chromosomes, we propose that RY is more important than presently believed as an information content system for DNA, and is by far the most crucial aspect of its information content. If we consider that DNA is an information-containing and programming system, we conclude that recognition is key, and propose that the structure itself can be a code.

3.4. The Use of Binary Components Analysis: Theory of Breakdown of Information Content

The philosophy used in this research is to treat the DNA as a binary information system, thereby making use of information theory tools. In a computer system, a binary code is used to represent any data including text, images, or computer processor instructions. When this is used, it assigns a sequence of binary digits or bits to each piece of data. It permits the computer to both store the data, and also process it. This concept is applied here to DNA as it helps reveal sequence design and function.

Information theory [41] represents communication systems analysis. This theory permits analysis of the various components of such systems. It generally encompasses an information source, with an encoder, a channel that contains random noise, then a decoder, and a receiver. Within genomes, Shannon information is analyzed in two to ten letters of DNA [42]. In this case, 'word' matches are far greater than equivalent randomized sequences. Genomic signatures, such as those applied in this research, are considered to possess Shannon information [43] as they comply with such a model.

With respect to genomic DNA, the receiver is a vast array of biological particles that bind to it. Among these are transcription factors, transcription, replication, and repair machinery. The information source is the DNA. The aim of this research is to decipher basic codes, and so the information source is the focus of interest. In this way, information theory has been applied via genomic signature analysis for the global characterization of genomic chromosomal DNA. This research marks a step forward in understanding non-coding DNA.

The DNA bases and their information content can be sub-divided via a binary system, similar to the way a computer stores binary information, as described by information theory. This principle is applied to the chemical and physical properties of DNA, and is applied with the intention of studying these properties separately and in isolation. The information is contained within it, literally within the bases or nucleotides. Therefore, separating this into the simplest form, and utilizing binary components is a powerful methodology.

The genome is a vast source of information that is multi-layered with many patterns and codes. By breaking down complex and layered information into the most basic binary parts, we made progress in understanding DNA. Combined with this, the genomic signatures and distances from randomness have revealed new insights into the nature of the DNA. The approach here has been to understand general designs, patterns, and signatures. The results have deepened our insight into general designs, and also developed concepts and novel analyses for deciphering the information and programs contained in the DNA.

3.5. Assumptions, limitations and Future Research

Since sequence and function are intrinsically connected, it is possible to draw conclusions regarding the relative importance of these chemical and physical DNA properties. There are, however, some assumptions and limitations. The dinucleotide relative abundance profile is a measure of the average enhancement or suppression of all possible dinucleotides in a given sequence. It is, therefore, also a measure of the overall randomness (or non-randomness). For the RY, KM, and WS component analyses, dinucleotide relative abundance profiles were calculated for each of these, for the exact same sequences, in order to evaluate the relative non-randomness with respect to each component. Here, the assumption (and interpretation) is that relative non-randomness is synonymous with the greater importance of that component. A further assumption is the connection of components with chemical or physical properties. This includes that RY is synonymous with structure. It makes sense in terms of existing evidence regarding DNA structure and its connection with purine and pyrimidine properties of the base sequence of DNA. A further assumption is that the KM and WS characteristics are synonymous with hydrogen-bonding patterns. Again, this makes sense in terms of chemical properties, however, the coherence is not necessarily exact, since KM and WS may have properties in addition to hydrogen-bonding capability.

In this research, we have sought to investigate the existence of baseline codes in genomic DNA. The RY code that was discovered in this analysis is widespread in genomic DNA. Any such baseline encoding is going to be stable and present in all relevant sequences, where the code is present. For instance, while protein-coding sequences are different from each other, they code for a huge variety of proteins. However, the triplet-code itself is pervasive in all of these, so too with the RY code, it is present in genomic DNA in a general sense, while the sequence is varied. This will of course be true in different individuals of a given organism, in this case humans. However, there are sequence variations in genomic DNA between individuals. These include both CNVs and SNPs. These are of great interest in studying disease, for instance [44]. Such variations also assist us in furthering our understanding of information content and transfer in the genome. CNVs, in particular, are marked by differences in DNA structure. While the genomic DNA of all individuals in the organism would possess the RY code, analysis of this variation is interesting, and holds potential for future research.

This research forms a strong basis and foundation for further research in decoding genomic DNA. These results pertain to a large-scale global analysis of chromosomal DNA. For future work, it is of interest to analyze protein-coding sequences of the RY code, and other binary components. The aim would be to see how these compare with the baseline genome sequence, and also to other genomic regions. For this, an appropriate method would be to devise and analyze short sequence fragments. Such analysis could also have future applications for the identification of different genomic sequences.

With regard to coding DNA and the triplet code, this method also can be applied to deepen our understanding. Beyond this, is future research that binarizes DNA sequences, and examines if the DNA->RNA->Protein model generates a pattern, or connecting code. This would be profoundly interesting, and has an impact on the fundamentals of biology.

Yet another area for further application of this result is the study and identification of hypothetical formation of triple-stranded DNA [45]. This type of formation clearly has structural characteristics that define it, and so an RY structural code is likely a determinant of this tendency. The RY code and base-step dinucleotides are relevant to the secondary and tertiary structure of DNA. The formation of such alternative helical structures may be due to a bias, and particular application of the RY code. This is a strong area for further investigation.

4. Conclusions

The use of binary components has been used to break down information content within the bases of DNA, similar to the use of computer binary code for the transmission of complex information. This may be layered or even multi-dimensional. This principle is to isolate information stored within DNA bases into binary components, decipher patterns and codes, and gain insight into their functionality.

This is a different way of thinking since we are so accustomed to the one-dimensional relationships of the triplet code, which does not apply to the vast majority of genomic sequences, yet there is so much depth to the information stored there. General designs are powerful, and coupling them with a binary system allows us to isolate the information content and decipher the language of the genome.

This research makes use of the inherent relationship between sequence, structure, and function; therefore, biological meaning is successfully derived from sequence data. This aids in the deciphering of codes in non-coding DNA. It also demonstrates the value of large-scale sequence analysis for general designs of the genome. The RY, KM, and WS binary components analyses reveal the relative importance of physical and chemical properties of the bases of the DNA. These are three known categories of the bases of DNA. However, here we bring to light some alternative properties within these known categories, and utilize them in a novel way.

In genomic DNA across all chromosomes, it is concluded that an RY- structural code is present. This is likely based on structural recognition of particles, which likely reflects the general functionality of genomic DNA. This research marks a step forward towards understanding the language of DNA.

This analysis is extremely important, since the underlying rules and codes for the vast majority of genomic (non-coding) DNA are not yet understood. These sequences carry the program for building an organism, and are known to be key to the regulation of genes. The mechanisms of DNA function in gene regulation and protein-DNA binding are not well understood. The elucidation of general designs, patterns, and codes is necessary in order to understand gene regulation, and general DNA functionality.

The outcome is far-reaching, and includes a better understanding of general cell biology, growth, and development. The wider impact includes insight into genetic causes of human disease, and a leap forward in drug design. Furthermore, while codes and general designs have some variation between species, general rules and principles are likely similar across species, since these are fundamental biological principles.

Future investigation will be the analysis of the nature of the RY structural code. DNA sequences may be examined further for patterns, including more complex sequence patterns, signatures, and their related structural features. The structure of these RY patterns may also be analyzed, in order to gain insight into the structural code. On a wider scope, different genomic locations within the human genome may be studied for DNA signatures, and an investigation conducted to see if there are any differences, for instance, between coding and non-coding DNA. The signature analysis will also be further expanded to other organisms.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/math10152723/s1, S1: statistics_full_output.xlsx; S2: results_data_tables.xlsx; S3: results_data_tables_random_model.xlsx; S4: Results for real versus random model data can be found in the following attached file: Statistics_random_model.xlsx.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article or Supplementary Materials.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

ANOVA results table for distance from randomness of binary components can be found in the following attached file.

Table A1. Results of the one-way within subject ANOVA and the follow-up pairwise comparisons.

Chromosome No.	ANOVA	RY vs. WS	RY vs. KM	WS vs. KM
1.	F(1.87, 4303.25) = 9896.044, $p < 0.001$, $\eta_p^2 = 0.811$	0.063	0.033	−0.03
2.	F(1.94, 4669.58) = 13948.031, $p < 0.001$, $\eta_p^2 = 0.853$	0.06	0.025	−0.035
3.	F(1.85, 3662.67) = 10741.541, $p < 0.001$, $\eta_p^2 = 0.844$	0.058	0.025	−0.033
4.	F(1.87, 3550.4) = 9521.713, $p < 0.001$, $\eta_p^2 = 0.834$	0.053	0.02	−0.033
5.	F(1.63, 2954.66) = 6436.407, $p < 0.001$, $\eta_p^2 = 0.78$	0.056	0.023	−0.032
6.	F(1.95, 3319.54) = 10573.787, $p < 0.001$, $\eta_p^2 = 0.862$	0.06	0.025	−0.036
7.	F(1.72, 2733.99) = 6212.169, $p < 0.001$, $\eta_p^2 = 0.796$	0.059	0.024	−0.034
8.	F(1.84, 2662.38) = 6442.536, $p < 0.001$, $\eta_p^2 = 0.817$	0.055	0.024	−0.031
9.	F(1.77, 2155.44) = 5441.261, $p < 0.001$, $\eta_p^2 = 0.817$	0.06	0.027	−0.033
10.	F(1.66, 2216.09) = 5461.652, $p < 0.001$, $\eta_p^2 = 0.804$	0.061	0.026	−0.035
11.	F(1.8, 2423.01) = 4060.935, $p < 0.001$, $\eta_p^2 = 0.751$	0.058	0.032	−0.025
12.	F(1.77, 2357.78) = 4766.399, $p < 0.001$, $\eta_p^2 = 0.782$	0.061	0.029	−0.032
13	F(1.72, 1681.74) = 3222.249, $p < 0.001$, $\eta_p^2 = 0.767$	0.052	0.018	−0.034
14.	F(1.7, 1534.44) = 3065.551, $p < 0.001$, $\eta_p^2 = 0.772$	0.061	0.028	−0.033
15.	F(1.63, 1377.82) = 3458.144, $p < 0.001$, $\eta_p^2 = 0.804$	0.065	0.03	−0.036
16.	F(1.63, 1329.35) = 3076.435, $p < 0.001$, $\eta_p^2 = 0.79$	0.061	0.029	−0.031
17.	F(1.51, 1250.69) = 2419.446, $p < 0.001$, $\eta_p^2 = 0.745$	0.072	0.04	−0.031
18.	F(1.65, 1316.46) = 2266.129, $p < 0.001$, $\eta_p^2 = 0.739$	0.055	0.025	−0.03
19.	F(1.64, 956.8) = 1505.212, $p < 0.001$, $\eta_p^2 = 0.721$	0.069	0.037	−0.032
20.	F(1.59, 1013.05) = 2682.895, $p < 0.001$, $\eta_p^2 = 0.808$	0.062	0.031	−0.03
21.	F(1.52, 605.95) = 692.452, $p < 0.001$, $\eta_p^2 = 0.634$	0.053	0.021	−0.031
22.	F(1.49, 581.68) = 806.02, $p < 0.001$, $\eta_p^2 = 0.674$	0.06	0.035	−0.025
X.	F(1.93, 2979.63) = 5842.198, $p < 0.001$, $\eta_p^2 = 0.791$	0.055	0.02	−0.035
Y.	F(1.79, 471.43) = 291.074, $p < 0.001$, $\eta_p^2 = 0.525$	0.041	0.011	−0.029

Note: η_p^2 partial eta squared; All ANOVA results are based on the Greenhouse-Geisser correction. All and post hoc pairwise comparisons are corrected using the Bonferroni method and are statistically significant at $p < 0.001$.

Table A2. Descriptive statistics of distance from randomness (relative abundance profile) analysis for each chromosome.

Chromosome	**RY**		**WS**		**KM**		**N**
	Mean	**Std. Deviation**	**Mean**	**Std. Deviation**	**Mean**	**Std. Deviation**	
1.	0.124	0.020	0.060	0.024	0.091	0.009	2303
2.	0.116	0.017	0.056	0.016	0.091	0.009	2405

3.	0.116	0.018	0.058	0.019	0.091	0.009	1980
4.	0.108	0.016	0.055	0.017	0.088	0.009	1897
5.	0.114	0.018	0.058	0.025	0.091	0.009	1812
6.	0.115	0.017	0.055	0.016	0.090	0.008	1700
7.	0.117	0.019	0.058	0.024	0.093	0.010	1589
8.	0.114	0.017	0.059	0.018	0.091	0.009	1447
9.	0.119	0.018	0.059	0.022	0.093	0.010	1217
10.	0.120	0.017	0.059	0.025	0.094	0.009	1332
11.	0.123	0.022	0.065	0.027	0.090	0.010	1345
12.	0.120	0.020	0.059	0.026	0.091	0.010	1331
13.	0.107	0.017	0.055	0.020	0.089	0.010	979
14.	0.119	0.019	0.058	0.024	0.091	0.010	905
15.	0.124	0.017	0.059	0.026	0.094	0.008	846
16.	0.126	0.018	0.066	0.026	0.097	0.012	818
17.	0.136	0.021	0.064	0.034	0.096	0.010	829
18.	0.116	0.022	0.061	0.026	0.091	0.009	800
19.	0.139	0.020	0.070	0.035	0.102	0.009	584
20.	0.129	0.016	0.067	0.026	0.098	0.009	639
21.	0.115	0.020	0.062	0.030	0.093	0.015	400
22.	0.133	0.018	0.073	0.031	0.098	0.014	391
X.	0.112	0.020	0.057	0.019	0.092	0.009	1548
Y.	0.108	0.018	0.068	0.030	0.097	0.015	264

References

1. Locey, K.J.; White, E.P. Simple structural differences between coding and non-coding DNA. *PLoS ONE* **2011**, *6*, e14651.
2. Harrow, J.; Frankish, A.; Gonzalez, J.M.; Tapanari, E.; Diekhans, M.; Kokocinski, F.; Aken, B.L.; Barrell, D.; Zadissa, A.; Searle, S.; et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **2012**, *22*, 1760–1774.
3. Slattery, M.; Zhou, T.; Yang, L.; Dantas Machado, A.C.; Gordân, R.; Rohs, R. Absence of a simple code: How transcription factors read the genome. *Trends Biochem. Sci.* **2014**, *39*, 381–399.
4. Lee, C.M.; Barber, G.P.; Casper, J.; Clawson, H.; Diekhans, M.; Gonzalez, J.N.; Hinrichs, A.S.; Lee, B.T.; Nassar, L.R.; Powell, C.C.; et al. UCSC genome browser enters 20th year. *Nucleic Acids Res.* **2020**, *48*, D756–D761.
5. Fishilevich, S.; Nudel, R.; Rappaport, N.; Hadar, R.; Plaschkes, I.; Iny Stein, T.; Rosen, N.; Kohn, A.; Twik, M.; Safran, M.; et al. GeneHancer: Genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017**, *2017*, bax028.
6. Sternberg, R.V. DNA codes and information: Formal structures and relational causes. *Acta Biotheor.* **2008**, *56*, 205–232.
7. Jernigan, R.W.; Baran, R.H. Pervasive properties of the genomic signature. *BMC Genom.* **2002**, *3*, 23.
8. Karlin, S.; Ladunga, I. Comparisons of eukaryotic genomic sequences. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 12832–12836.
9. Karlin, S.; Burge, C. Dinucleotide relative abundance extremes: A genomic signature. *Trends Genet.* **1995**, *11*, 283–290.
10. Karlin, S.; Campbell, A.M.; Mrázek, J. Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.* **1998**, *32*, 185–225.
11. Karlin, S.; Mrázek, J. Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 10227–10232.
12. Ghannam, J.Y.; Wang, J.; Jan, A. Biochemistry, DNA Structure. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2022.
13. Burge, C.; Campbell, A.M.; Karlin, S. Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 1358–1362.
14. Travers, A.A.; Muskhelishvili, G.; Thompson, J.M. DNA information: From digital code to analogue structure. *Philos. Trans. A Math. Phys. Eng. Sci.* **2012**, *370*, 2960–2986.
15. Hood, L.; Galas, D. The digital code of DNA. *Nature* **2003**, *421*, 444–448.
16. Del Prado, A.; González-Rodríguez, D.; Wu, Y.L. Functional systems derived from nucleobase self-assembly. *Chem. Open* **2020**, *9*, 409–430.
17. Yagil, G. The over-representation of binary DNA tracts in seven sequenced chromosomes. *BMC Genom.* **2004**, *5*, 19.
18. Amano, N.; Ohfuku, Y.; Suzuki, M. Genomes and DNA conformation. *Biol. Chem.* **1997**, *378*, 1397–1404.
19. Bucher, P.; Yagil, G. Occurrence of oligopurine.oligopyrimidine tracts in eukaryotic and prokaryotic genes. *DNA Seq.* **1991**, *1*, 157–172.
20. Hunter, C.A. Sequence-dependent DNA structure. The role of base stacking interactions. *J. Mol. Biol.* **1993**, *230*, 1025–1054.
21. Chris RCalladine Horace RDrew Ben FLuisi Andrew, A. *Travers, Understanding DNA, the Molecule and How it Works*, 3rd ed.; Academic Press: Cambridge, MA, USA 2004.

22. Slocombe, L.; Al-Khalili, J.S.; Sacchi, M. Quantum and classical effects in DNA point mutations: Watson-Crick tautomerism in AT and GC base pairs. *Phys. Chem. Chem. Phys.* **2021**, *23*, 4141–4150.
23. Mo, Y. Probing the nature of hydrogen bonds in DNA base pairs. *J. Mol. Model.* **2006**, *12*, 665–672.
24. Shioiri, C.; Takahata, N. Skew of mononucleotide frequencies, relative abundance of dinucleotides, and DNA strand asymmetry. *J. Mol. Evol.* **2001**, *53*, 364–376.
25. Nemzer, L.R. A binary representation of the genetic code. *Biosystems* **2017**, *155*, 10–19.
26. Yu, C.P.; Kuo, C.H.; Nelson, C.W.; Chen, C.A.; Soh, Z.T.; Lin, J.J.; Hsiao, R.X.; Chang, C.Y.; Li, W.H. Discovering unknown human and mouse transcription factor binding sites and their characteristics from ChIP-seq data. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2026754118.
27. Xiong, L.; Kang, R.; Ding, R.; Kang, W.; Zhang, Y.; Liu, W.; Huang, Q.; Meng, J.; Guo, Z. Genome-wide Identification and Characterization of Enhancers Across 10 Human Tissues. *Int. J. Biol. Sci.* **2018**, *14*, 1321–1332.
28. Napoli, A.A.; Lawson, C.L.; Ebright, R.H.; Berman, H.M. Indirect readout of DNA sequence at the primary-kink site in the CAP-DNA complex: Recognition of pyrimidine-purine and purine-purine steps. *J. Mol. Biol.* **2006**, *357*, 173–183.
29. Pabo, C.O.; Nekludova, L. Geometric analysis and comparison of protein-DNA interfaces: Why is there no simple code for recognition? *J. Mol. Biol.* **2000**, *301*, 597–624.
30. Zhou, T.; Shen, N.; Yang, L.; Abe, N.; Horton, J.; Mann, R.S.; Bussemaker, H.J.; Gordân, R.; Rohs, R. Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 4654–4659.
31. Retureau, R.; Foloppe, N.; Elbahnsi, A.; Oguey, C.; Hartmann, B. A dynamic view of DNA structure within the nucleosome: Biological implications. *J. Struct. Biol.* **2020**, *211*, 107511.
32. Richmond, T.J.; Davey, C.A. The structure of DNA in the nucleosome core. *Nature* **2003**, *423*, 145–150.
33. Coulocheri, S.A.; Pigis, D.G.; Papavassiliou, K.A.; Papavassiliou, A.G. Hydrogen bonds in protein-DNA complexes: Where geometry meets plasticity. *Biochimie* **2007**, *89*, 1291–1303.
34. Gago, F. Stacking interactions and intercalative DNA binding. *Methods* **1998**, *14*, 277–292.
35. Seeman, N.C.; Rosenberg, J.M.; Rich, A. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. USA* **1976**, *73*, 804–808.
36. Cheng, A.C.; Chen, W.W.; Fuhrmann, C.N.; Frankel, A.D. Recognition of nucleic acid bases and base-pairs by hydrogen bonding to amino acid side-chains. *J. Mol. Biol.* **2003**, *327*, 781–796.
37. Youk, J.; An, Y.; Park, S.; Lee, J.K.; Ju, Y.S. The genome-wide landscape of C:G > T:A polymorphism at the CpG contexts in the human population. *BMC Genom.* **2020**, *21*, 270.
38. Cooper, D.N.; Gerber-Huber, S. DNA methylation and CpG suppression. *Cell Differ.* **1985**, *17*, 199–205.
39. Malik, F.K.; Guo, J.T. Insights into protein-DNA interactions from hydrogen bond energy-based comparative protein-ligand analyses. *Proteins* **2022**, *90*, 1303–1314.
40. Gershman, A.; Sauria, M.E.G.; Guitart, X.; Vollger, M.R.; Hook, P.W.; Hoyt, S.J.; Jain, M.; Shumate, A.; Razaghi, R.; Koren, S.; et al. Epigenetic patterns in a complete human genome. *Science* **2022**, *376*, eabj5089.
41. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
42. Chang, C.H.; Hsieh, L.C.; Chen, T.Y.; Chen, H.D.; Luo, L.; Lee, H.C. Shannon information in complete genomes. *J. Bioinform. Comput. Biol.* **2005**, *3*, 587–608.
43. Vinga, S. Information theory applications for biological sequence analysis. *Brief Bioinform.* **2014**, *15*, 376–389.
44. Zarrei, M.; MacDonald, J.R.; Merico, D.; Scherer, S.W. A copy number variation map of the human genome. *Nat. Rev. Genet.* **2015**, *16*, 172–183.
45. Matveishina, E.; Antonov, I.; Medvedeva, Y.A. Practical guidance in genome-wide RNA:DNA triple helix prediction. *Int. J. Mol. Sci.* **2020**, *21*, 830.