## Characteristics of the DNAkmerQM database

To provide a general exploratory insight into our database, we analysed the loading values[1] from principal component analysis (PCA)[2], and describe the characteristics of our datasets. PCA has been utilised to reduce the dimensionality of the datasets to express it through a lesser number of new variables. In such data compression, PCA can output loading values, which are the weights that the original variables have for generating new variables. Here, variables that have a similar tendency are likely to have similar loading values. As a result, loading plots, which are obtained by plotting vectors from the origin to loading values of the first principle component (PC1) and the second principle component (PC2), were utilised before for analysing the tendency of variables in biology[2–4]. These plots indicate that variables are positively correlated when their vectors are close to each other and the angle between them is small. As a result, clusters of variables are formed and we may find hidden patterns in the database. By utilizing this, we analyzed relationships between features obtained from our calculations to provide general insight into the content and interrelation of DNAkmerQM features.

*Overall loading plots.* **Fig. S1** shows loading values from PCA of energy (E), differences of energy (dE), Mulliken charges (MullikC), Mulliken population (MullikD), and geometric features obtained by Curves+ (Curves). Each plot is fitted with an ellipsoid for clarity. **Fig. S1a** shows a loading plot, in which features are grouped by conformations. We found that loadings of B-DNA and A-DNA are along the PC2 direction while Z-DNA is along the PC1 direction. Therefore, B-DNA and A-DNA show a similar tendency while Z-DNA shows a different tendency from them. **Fig. S1b** shows a loading plot grouped by feature types. We found that energy terms (E) remain around the origin of coordinates, while differences in energy terms (dE) lie along the PC2 direction, indicating that they play different roles in overall PCA. On the other hand, loadings for Mulliken charges, Mulliken population, and geometric features overlap with each other, indicating that they have a similar tendency as a whole. **Figs. S1c,d** show a loading plot grouped by the nucleotide locations to which

features belong. The features at the $4^{th}$ nucleotide lie along the PC2 direction and they gradually begin to spread along the PC1 direction as they close to the $1^{st}$ nucleotide (**Fig. S1c**). This is also applied to the loading plot for the $4^{th}$ to $7^{th}$ nucleotides (**Fig. S1d**). These indicate that the similarity of features reduces as they close to the edges of DNA.

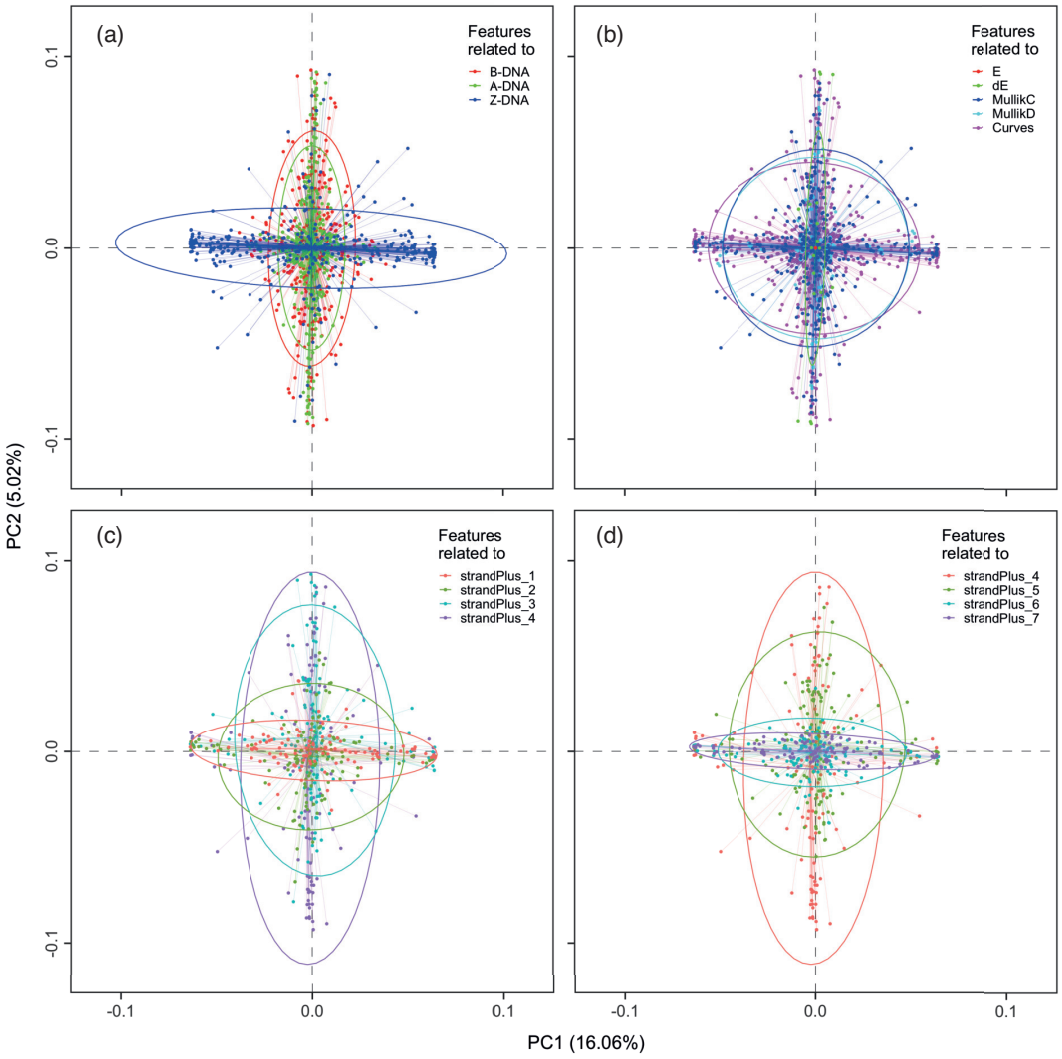


**Figure S1.** Loading plots from principal component analysis (PCA) of heat of formation energy (E), energy differences (dE), Mulliken charges (MullikC), Mulliken population (MullikD), and geometric (Curves) features. Features are grouped by (a) conformation, (b) feature type, and (c) the positions of nucleotide units at the first half and (d) the second half. The axes represent the first (PC1) and the second (PC2) principal components.

*Loading plots for energy and energy difference.* Here, we conducted PCA for only the features related to energy and differences in energy. **Fig. S2** shows loading values of heat of formation (Ehof) terms obtained by PCA. Colours indicate Ehof of the duplex state, differences in Ehof between the duplex and single-strand states (ds_ss), and differences in Ehof terms between the duplex and the central-base-removed states (ds_del1 and ds_del2). We found that the same kind of Ehof features make clusters, that is, clusters of dEhof_ds_del1, dEhof_ds_del2, and so on. Furthermore, features of B-, A-, and Z-DNA are close to each other in these clusters. These may indicate that DNA conformations have less pronounced effects on these features.

*Loading plots for Mulliken charge.* **Fig. S3** shows loading plots obtained by PCA from Mulliken charges only. For clarity, we made separate plots for the base, sugar, and phosphate moieties at $3^{rd} \sim 5^{th}$ nucleotide loci. To clearly show tendencies, features are grouped by the value type for base moieties, and by conformation type at sugar and phosphate moieties. We found that, at the base moiety, the maximum values of B-, A- and Z-DNA are clustered together, while the minimum and mean values make a cluster at the StrandPlus_3 position (see the base moieties in **Fig. S3a**). This tendency is also seen at the StrandPlus_4 and 5 positions (base moieties in **Figs. S3b,c**). This indicates that the prevalent characteristic for forming that cluster is the value type rather than conformation. On the other hand, at the sugar moiety, features of B- and A-DNA lie along the PC2 direction, while those of Z-DNA lie along the PC1 direction (sugar moieties of **Figs. S3a-c**). This tendency can also be seen at phosphate moieties (see the phosphate moieties in **Figs. S3a-c**). This indicates that a conformation of DNA is the prevalent characteristic for forming such a cluster at sugar and phosphate moieties, differing from the base.

*Loading plots for geometric parameters.* **Fig. S4** shows loading plots by PCA of only geometric features of B-DNA. For clarity, we made separate plots for the intra, inter, and backbone parts of geometric features at the $3^{rd} \sim 5^{th}$ nucleotide loci. To clearly show tendencies, features are grouped when they show similar behaviours through the $3^{rd} \sim 5^{th}$

nucleotides. We found that the intra parts of geometric features do not make any clusters although Ax-bend, Stagger, and Opening exhibit somehow a similar tendency (see intra parts in **Figs. S4a-c**). On the other hand, H-Twi, H-Ris, Rise, Slide, and Twist make a cluster in the inter parts (**Figs. S4a-c**). Ampli, Phase, Gamma and Zeta for the backbone parts cluster together. Moreover, these features go back and forth between the plus and minus regions of the PC2 axis, similar to H-Twi, H-Ris, Rise, Slide, and Twist in the inter-parts. This is also applied to Buckle in the intra parts, with a cluster formed by these features. Similarly, in A-DNA, intra (Shear, Buckle, and Ydisp), inter (H-Twi, H-Ris, Rise, and Roll), and backbone (Alpha, Beta, Zeta, Chi, Gamma, Ampli, and Delta) form a cluster (**Fig. S4**). Z-DNA presents two clusters: a cluster formed by intra (Shear), inter (Roll), and backbone (Alpha, Beta, Chi, and Phase), and a cluster formed only by inter (H-Twi, H-Ris, Twist, Rise, and Slide) and backbone (Epsil, Zeta, Ampli, Delta and Gamma) features, as shown in **Fig. S5**.

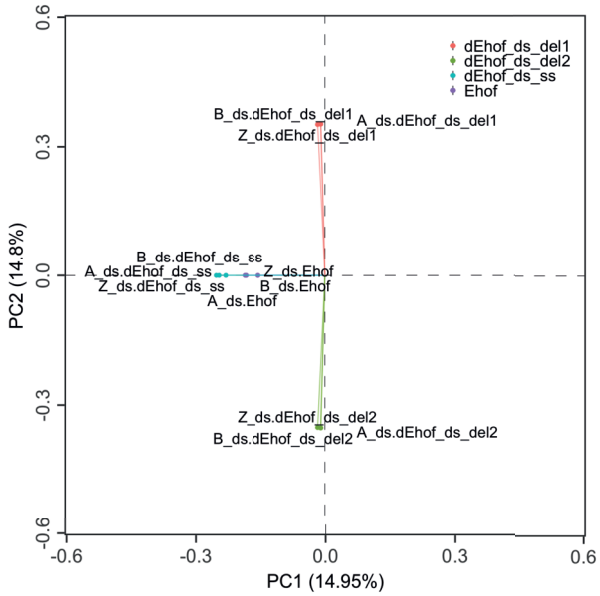


**Figure S2.** A loading plot for heat of formation (Ehof) terms. Colours indicate energy terms of duplex state (ds), differences in energy terms between the duplex and single-strand states (ds_ss), and differences in energy terms between the duplex and the centre base removed states (ds_del1 and ds_del2).

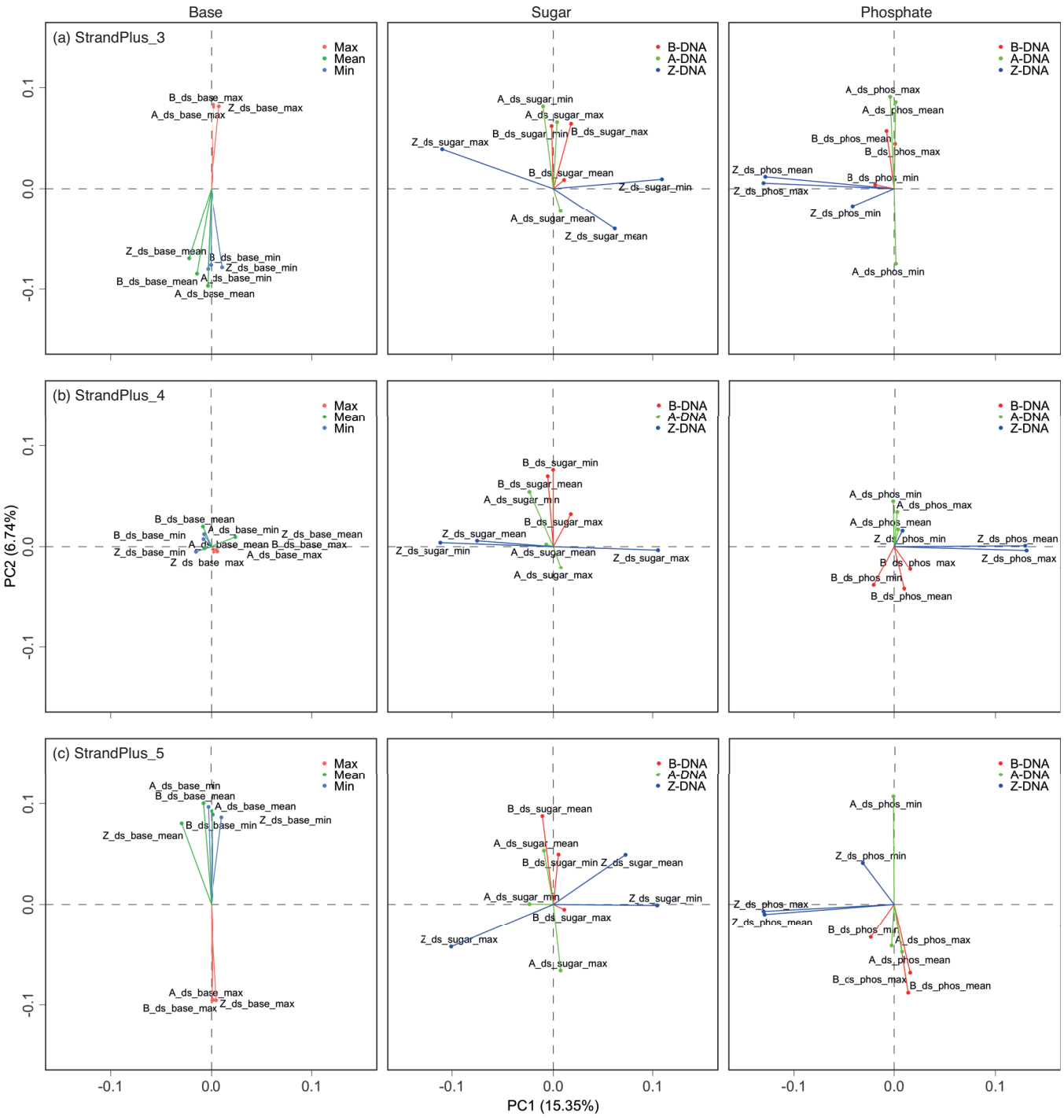


**Figure S3.** Loading plots from PCA based on only Mulliken charges. Each figure corresponds to the base, sugar, and phosphate moieties at the nucleotide units of (a) strandPlue_3, (b) strandPlue_4, and (c) strandPlue_5 sites. To clearly show tendencies, features are grouped by the value types in the base parts, while features are grouped by conformations in the sugar and phosphate parts.

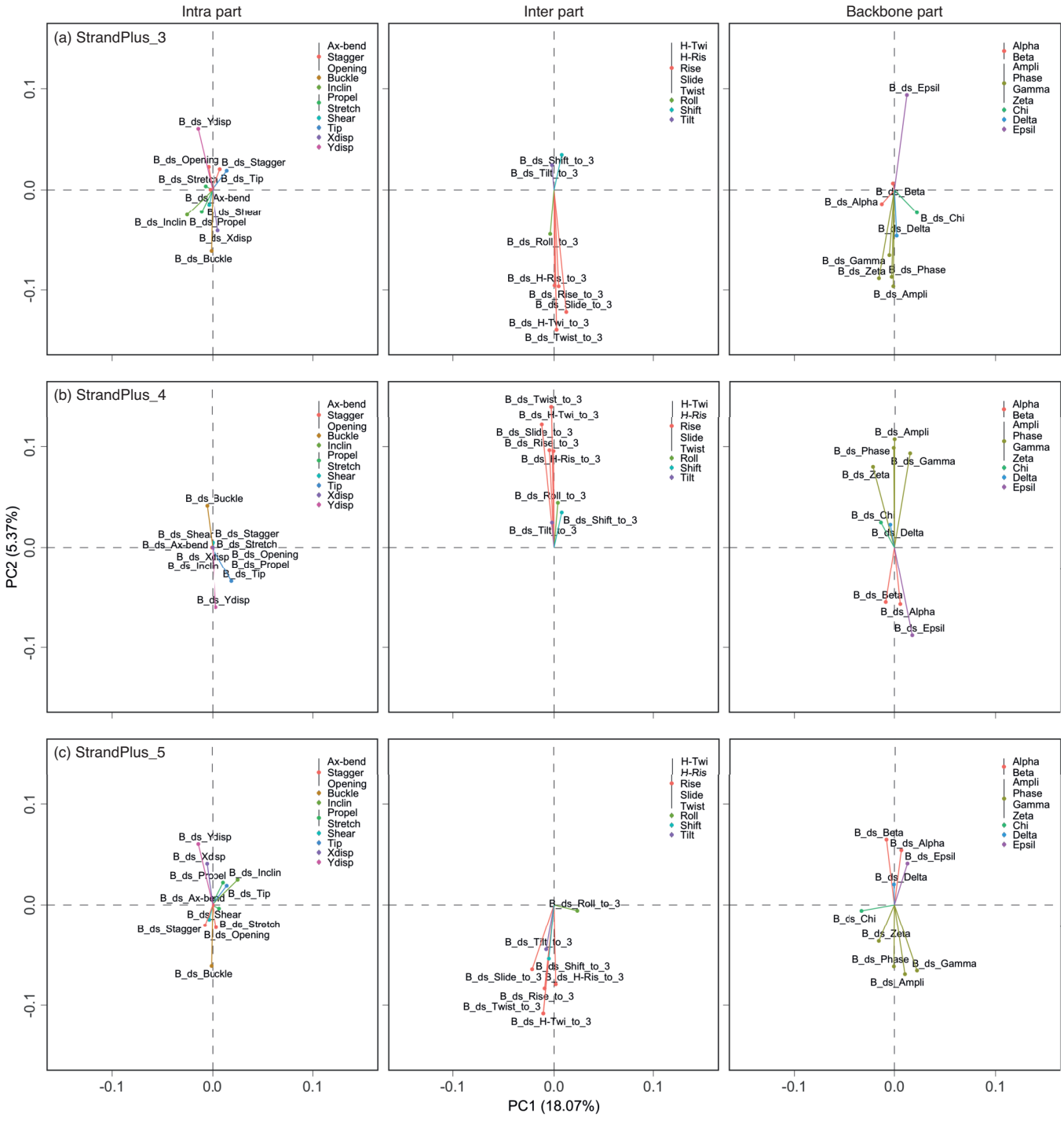


**Figure S4.** Loading plots from PCA based on only geometric features of B-DNA. Each figure corresponds to the intra, inter, and backbone parts at the nucleotide loci of (a) strandPlue_3, (b) strandPlue_4, and (c) strandPlue_5 sites. To clearly show tendencies, features that show similar behaviours through strandPlue_3∼5 are grouped by colours.

*Correlation analysis for B- and A-DNA.* Besides the loading plots, we calculated correlation matrices between features to obtain more insights. **Fig. S5a** shows a heat map for features related to the B-DNA conformation. For clarity, each feature was classified into E&dE, Mulliken Charge, Intra & Inter, and Backbone regions instead of showing all features. Besides the obvious correlations among the features of the same categories (see diagonal parts of **Fig. S5a**), some correlations are also found among features of different categories. For instance, Mulliken charges of base moieties strongly correlate with Zeta and Epsil angles from backbone parameters (**Fig. S5b**), indicating an interesting coupling between the electronic and geometric features in B-DNA. Similarly, we found that Mulliken charges of the phosphate moieties strongly correlate with backbone Alpha in A-DNA (**Fig. S6**).

*Correlation analysis for Z-DNA.* As for the correlation matrix of Z-DNA, a more detailed explanation will be required. **Fig. S7a** shows a heat map for features related to Z-DNA conformation. We found cross correlations all over the plot, which is not observed in B- and A-DNA. Further investigation showed that the Mulliken charge of the phosphate part is strongly correlated with Zeta, Chi, and Delta in backbone parameters. However, strangely, Mulliken charge of phosphate at the $2^{nd}$ nucleotide still correlates with backbone parameters at the $6^{th}$ nucleotide. This demonstrates that parameters away from each other by four nucleotide distance still correlate with each other (a long-range correlation). **Fig. S7c1** shows the mean value of Mulliken charge at the phosphate moieties in Z-DNA. We found that clusters are formed, in which if Mulliken charge is around 0.13, the charge at the next nucleotide is always around 0.16, or vice versa. On the other hand, such a cluster is not observed in B-DNA (**Fig. S7c2**). This indicates that charge values in Z-DNA are not continuous but are seemingly quantised/categorical, i.e. they stay only at around certain values. From these observations, the following can be concluded: The zig-zag backbone structure of Z-DNA interferes with the electronic states in nucleotides, resulting in charge and mechanical values that alternate as small, large, small, large, and so on, values (**Fig.**

**S7c1** right). As a result, long-range order is formed and parameters from nucleotides far away from each other correlate.

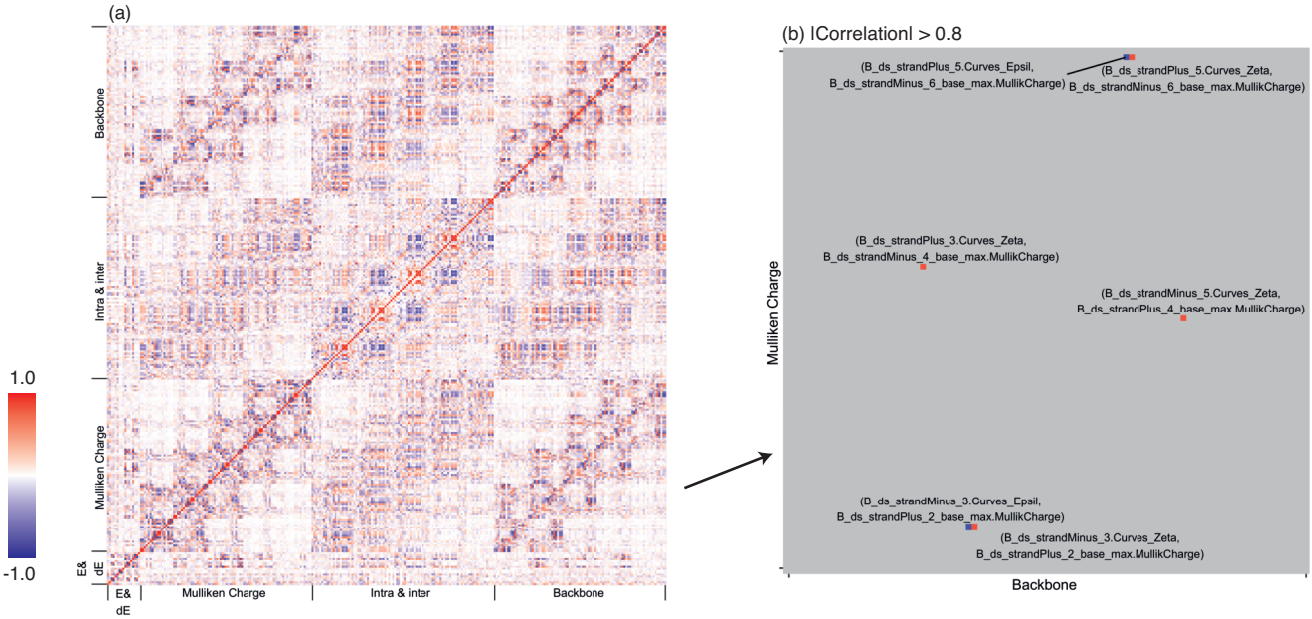


**Figure S5.** (a) A heat map of features related to B-DNA. For clarity, categories of features are shown instead of showing all features. The colour contour indicates a positive and negative correlation between features. (b) A heat map at the Backbone and Mulliken Charge regions with only high correlation points highlighted.

# References


(1) Jolliffe, I. *Principal component analysis*, 2nd ed.; Springer: New York, 2002.

(2) Ringnér, M. What is principal component analysis? *Nat. Biotechnol.* **2008**, *26*, 303–304.

(3) Murat, P.; Marsico, G.; Herdy, B.; Ghanbarian, A.; Portella, G.; Balasubramanian, S. RNA G-quadruplexes at upstream open reading frames cause DHX36- and DHX9-dependent translation of human mRNAs. *Genome Biol.* **2018**, *19*, 229.

(4) Lever, J.; Krzywinski, M.; Altman, N. Principal component analysis. *Nat. Methods* **2017**, *14*, 641–642.