 doi:10.20944/preprints202201.0471.v1

Superstructure detection in nucleosome distribution shows common pattern within a chromosome and within the genome

Sujeet Kumar Mishra[1,2], Kunhe Li[1], Simon Brauburger[1], Arnab Bhattacherjee[2], Nestor Norio Oiwa[1,3], Dieter W. Heermann[1*],

1 Institute for Theoretical Physics, Heidelberg University, Heidelberg, Germany
2 School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India
3 Department of Basic Science, Universidade Federal Fluminense, Nova Friburgo, Brazil

* heermann@tphys.uni-heidelberg.de

Abstract

Nucleosome positioning plays an important role in crucial biological processes like replication, transcription, and gene regulation. It has been widely used to predict the genome's function and chromatin organisation. So far, the studies of patterns in nucleosome positioning have been limited to transcription start sites, CTCFs binding sites, and some promoter and loci regions. The genome-wide organisational pattern remains unknown. We have developed a theoretical model to coarse-grain nucleosome positioning data in order to obtain patterns in their distribution. Using hierarchical clustering on the auto-correlation function of this coarse-grained nucleosome positioning data, a genome-wide clustering is obtained for *Candida albicans*. The clustering shows the existence beyond hetero- and eu-chromatin inside the chromosomes.

Keywords: chromatin, nucleosome positioning, nucleosome distribution, heterochromatin, euchromatin, structure classification

Introduction

The genomes of all higher eukaryotes are organised in different structures on multi-length scales (1; 2). Of these organisational structures, the chromosome is the biggest one, being observable under a normal light microscope. The smallest organisational structure, one level above the double helix DNA, is the nucleosome where 147 base pairs (bp) of DNA are wrapped 1.65 times around a histone octamer (3; 4; 5). The arrays of nucleosomes organize to form the chromatin fibre, which folds into two mutually excluded structural domains, namely "heterochromatin" and "euchromatin". The "heterochromatin" regions are enriched with inactive/repressive genes and are usually positioned closer to the periphery of the nucleus. The "euchromatin" regions contain transcriptionally active chromatin (3; 6; 7), genes being located in the interior of the nucleus. The hierarchical packaging of chromatin renders the genome a very compact conformation that provides controlled accessibility of the regulatory DNA sequences (genes) by other DNA binding proteins (DBPs)(8; 9). The chromatin organisation is thus, tightly linked to gene regulation and warrants detailed investigation. Various experimental techniques have been developed to probe

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

the hierarchical chromatin organisation at different length scales. For instance, the "chromatin conformation capture" experiment (e.g. 3C and HiC) (10; 11; 2) captures organisation of chromatin in kbp to Mbp length scale, revealing formation of topologically associated domains (TADs) (12) and chromatin loops (13; 14). Further characterisation of the chromatin fiber at the length scale of genes (∼kbp) is achieved by Micro-C technique that captures the intra-chromatin interactions at a resolution of ∼100bp within an organisation module called chromosomal interaction domains (CIDs) (15; 16). CIDs are much smaller but still similar to TADs. These structural organisations are strongly regulated by the nucleosome positions, length of linker regions, and presence of nucleosome depleted regions (NDR) across the chromosome (17).

The term "nucleosome positioning" refers to the location of nucleosomes along the sequence of genomic DNA. Nucleosome positioning is determined by several factors including DNA sequence, DNA-binding proteins, nucleosome remodelers, RNA polymerases, and more. Although nucleosome positioning is a dynamic process, the sequence-based mapping approach identifies its position only in a cell- and time-averaged manner. The technology of micrococcal nuclease (MNase) digestion combined with high-throughput sequencing (MNase-seq) (18) is a powerful method to map the genome-wide distribution of nucleosome positioning and its occupancy. The resulting occupancy maps are ensemble averages of heterogeneous cell populations. However, it is necessary to retrieve the cell specific features from the population average to reveal the mechanism of nucleosome organisation and its translocation along the genome. Zhang et al. has developed an algorithm called "Nucleosome Positioning from Sequencing" (NPS) to predict accurate nucleosome positioning from the MNase-seq data, which was later improved to iNPS (improved NPS) (19). Furthermore, extensive studies have been performed to recognise nucleosome positioning patterns around CTCFs, transcription start sites (TSSs), exons and introns, promoter and loci regions locally. For instance, a typical nucleosome distribution around TSSs indicates nucleosome depletion, resulting in a nucleosome-free region (NFR) whereas the nucleosomes downstream of TSS are equally spaced (20). A similar observation around CTCF is obtained: An array of well-positioned nucleosomes flank the sites occupied by the insulator binding protein CTCF across the human genome (21). Despite the efforts, the global picture of nucleosome positioning remains elusive until a recent study that has reported three types of nucleosomal arrangement by analyzing the nucleosome spacing and phasing in a genome(22). The evenly spaced nucleosomes in the array are termed as a regular array and irregular otherwise. At a given genomic location in the cell population, nucleosomes may also assume similar positions and are referred to as phased arrays. The phased-regular nucleosome arrays, being most prominent, are the hallmark of chromatin and found to be conserved from yeast to mammals. These phased-regular nucleosome arrays are mostly found near promoter regions of transcribed genes in the yeast genome and near binding sites of high-affinity DBPs in higher eukaryotes. The findings are, however, have limited applicability only at local regions of the chromatin fiber and provide absolutely no information about the nucleosome organisation along a complete chromosome or genome.

We used a theoretical approach to obtain a novel classification of segments across the chromosome based on the similarity in nucleosome patterns. The nucleosome positioning data are used as inputs that are systematically coarse-grained to analyze their auto-correlation function to search for any pattern. The results are processed using hierarchical clustering techniques to investigate if there exists any unique pattern of nucleosome. Our results suggest that the positions and occupancy of nucleosomes in a chromosome are not random, rather they reveal distinct patterns of distribution within a chromosome. Interestingly, the patterns appear to be conserved within the genome as well and are in agreement with the previous study that has reported three distinct nucleosome organisations across the genome. Furthermore, at the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

chromosome level, our approach could capture a few unique patterns in the range of $\sim$ 50 kbp length scale which repeatedly occur throughout the chromosomes, indicating they might play crucial role in regulating gene networks at a more local scale. The study underpins the nucleosome positioning architecture inside a genome that can provide insights into the genome organisation(c.f. Figure 1) not known before.

Data

The technology of micrococcal nuclease (MNase) digestion combined with high-throughput sequencing (MNase-seq) (18) is used to map the distribution of nucleosome occupancy genome-wide. In order to map the MNase-seq data to nucleosome positioning data, several programs were developed, such as NPS (23), nucleR (24) and DANPOS (25). A nucleosome sequencing profile is generated to depict nucleosome distribution in wave-form where nucleosome peaks are detected. The improved nucleosome-positioning algorithm (iNPS) can be applied to identify peaks and correctly detect nucleosome positions (19). One possible output of the iNPS algorithm is in the binary format, with 1's representing a nucleosome being present and 0's for the nucleosome-free regions or linker regions.

The genome-wide study of the species is a challenging task due to its large sequence size which needs theoretical expertise and computational power. For our study, we focused on the species *Candida albicans* due to its comparable smaller genome size (26). It consists of 8 sets of chromosome pairs whose complete genome sequence is available. The raw data of the MNase-seq is available from the Gene Expression Omnibus (GSM1542419) and was measured by Puri et. al. (27). We also accessed the processed iNPS data in the NucMap database by Zhao et al. (28).

Methods

To obtain a consistent classification of the nucleosomal positioning data in genome-wide classes, we perform the following steps (explained in more detail below the list):

1. Each chromosome is divided into segments of 75 kbp of length.
2. For every chromosome, positioning data is coarse-grained.
3. The coarse-grained nucleosome positioning data is used to calculate auto-correlation functions over the different sections.
4. A distance matrix is calculated over all the auto-correlation function data.
5. These segments are clustered. Various distance matrix and clustering algorithms are used to generalize the results.

Genome section classification

In order to extract the global pattern for areas in a genome, the whole genome is separated into sections with equal length. The section length L is an important scale parameter and needs to be properly set. L should not be too large to avoid all features from different areas bounded together. At the same time, L also should not be too small; otherwise, the global structure is flooded by the subtle differences and becomes a pattern for only a single nucleosome. The single nucleosome wrapping length L_n can be used as a lower bound for the choice of L. However, to obtain relevant structure we require that $L >> L_n$. Considering the nucleosome length L_n is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

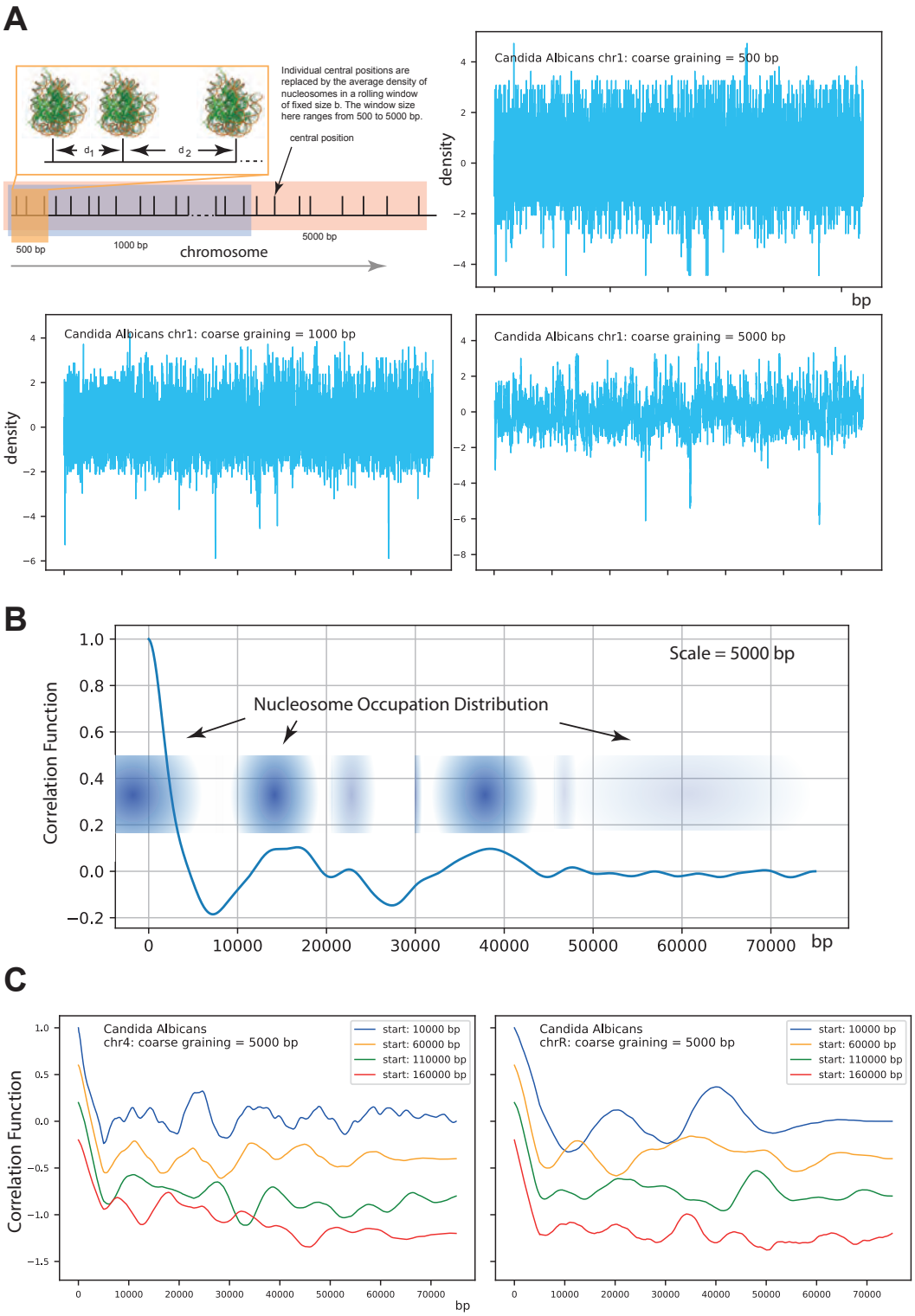

Figure 1: Panel A shows the performed coarse-graining procedure and results for coarse-graining lengths b of 500 bp, 1000 bp and 5000 bp. More structure is visible as b is increased. Going up even further washes out the structure. This is typical for systems with an intrinsic length scale. Panel B shows the correlation among the coarse-grained super nucleosomes. The structure is that of a system exhibiting short range-order that is liquid-like with first and second nearest neighbors. On larger scales larger than 50000 bp there is no order, i.e. there is no correlation. Panel C shows for two chromosomes how the structure differs within as among chromosomes. Nevertheless, common structures are found.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

about 147 bp (3; 4), L is chosen to be 50 kbp. Additionally, to avoid boundary effects, for each section a 12.5 kbp intersection on both sides with its neighbor is added. Hence the total section length L is 75 kbp. This binning is applied to each chromosome. Chr. 2 for example, with a length of $2,231,883$ bp, is separated into 44 sections.

Coarse-graining

The idea of coarse-graining is an established ansatz and tool in physics to describe complex systems on a scale that allows identifying structure. Typically, the structure appears as a collective phenomenon among smaller entities. The idea is to eliminate degrees of freedom, i.e., find a representation of the system on a larger time or space scale, iteratively moving to larger scales without changing the system. Over the last few years, coarse-graining has emerged as a way to model large complex systems and has successfully been applied to other biomolecules like proteins (29).

After the whole genome is separated into sections, coarse-graining is applied for each section. The method we implemented for coarse-graining is the rolling mean method (30). This method takes a window with a certain size (e.g. $b = 5kbp$), computes the averaged value of the nucleosome positioning inside the window, and moves the window to the following location. After this value is computed for each location, coarse-grained data on the scale of the window size is returned. Here, Python pandas.DataFrame.rolling (31) is used to obtain the coarse-graining. To exclude the effect of telomeres, discrete ends of the sections and to incorporate the window size and offset was chosen to be at least

$$\text{offset} \geq \text{window size}/2 \tag{1}$$

Auto-correlation function calculation

An auto-correlation function is a well-known approach in physics and pattern recognition, capturing the inner interaction pattern inside the data (30). Particularly for structures that are liquid-like the auto-correlation function, or in this context the radial distribution function, identifies typical length scales and patterns.

For each section j, it is applied on all the coarse-grained data ρ_j. The normalized auto-correlation function $C^j(\tau)$ with respect to distance τ for section j is :

$$C^{\alpha,j}(\tau) = \frac{E[(\rho_i^{\alpha,j} - \mu^{\alpha,j})(\rho_{i+\tau}^{\alpha,j} - \mu^{\alpha,j})]}{(\sigma^{\alpha,j})^2} \tag{2}$$

where $\rho_i^{\alpha,j}$ is the data at position i within the section j of chromosome α. $E(\dots)$ is the mean of everything in the parentheses over all indices i. μ^j is the mean of ρ and σ^j is the variance for the section j. Thus, associated with each section j is the function $C^{\alpha,j}(\tau)$ of chromosome α, hence, at the end we will have N functions $C^{\alpha,j}(\tau)$ where N is the section number for the particular chromosome.

Distance matrix calculation

To classify the functions, a similarity measure is applied and a resulting distance matrix is computed. The distance matrix is a square matrix containing the pairwise distances between all the elements available in the dataset, measuring the proximity between the correlation functions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

Interpreting the functions as high-dimensional vectors, we use the p-norm to define the distance d_p between two functions:

$$d_p(a,b) = \|a-b\|_p = (\sum_{i=1}^{d}(|a_i - b_i|^p)^{1/p} \tag{3}$$

where a and b are the functions in form of vectors. For $p = 2$, the p-norm this corresponds to the Euclidean distance.

Clustering

To identify the unique nucleosome organisation or distribution function, there is a need to cluster the sections together on the basis of similarity among them. We used a clustering approach, i.e., hierarchical clustering (32). This is an unsupervised algorithm that groups similar objects into groups called clusters. It uses a distance matrix to identify the two closest clusters first and then merge the two most similar clusters. This iterative process continues until the clusters are merged to get distinct clusters in a hierarchical manner.

Hierarchical clustering builds a hierarchy of clusters using two methods: agglomerative and divisive algorithms. We used the former, i.e. the Ward method (33) where each observation starts in its own cluster and pairs of clusters are merged moving up the hierarchy.

Statistical Distributions Fitting

Fitting of the distributions was performed using the scipy stats package (34) under Python.

Results

The first indication of non-trivial ordering is given by the distribution of the nucleosome positioning data. The binary nucleosome positioning data for all chromosomes of *Candida albicans* (NucMap database (35)) is subjected to the described coarse-graining and then analyzed (see the histogram of densities in the supplementary information Figures 1.1, 1.1.1 and 1.1.2). The genome-wide normalized nucleosome density shows a non-gaussian behavior with a slight negative skew. Overall, a log-logistic distribution gives the best consistent fit for all chromosomes compared to a normal distribution on the same bin size and rolling average for all chromosomes.

Recall that each chromosome is divided into chunks of 75 kbp with 25 kbp overlapping on each side. The auto-correlation of each chunk is obtained on the coarse-grained nucleosome positioning data. The respective correlation function of each section for all chromosomes are shown in Figure 2 and in detail in supplementary information (Figures 9 to 16). Shown are the correlation functions on the coarse-grained scale as well as a further smoothing to make the features that are common among a class more apparent(see below). The color bar indicates the class. Even though there are variations within a class, certain common features are seen. These features are the first and second peak structure, the height of the peaks, and how long a structure persists. Recall that the zero line indicates that there is no correlation, i.e. there the structure is that of a gas or an unordered behavior. The first peak indicates an increased probability to find a coarse-grained nucleosome at the distance of the peak position, the same applies to the second and additional peaks. If these peaks are of similar height, then there is a stronger long-range ordering. A particular example showing similar heights up to a third peak is in section 12 of chromosome no. 3 (see Figure 11), while section 6 shows a drop in the peak heights. Nevertheless, due to the overall similarity, these fall into the same class.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

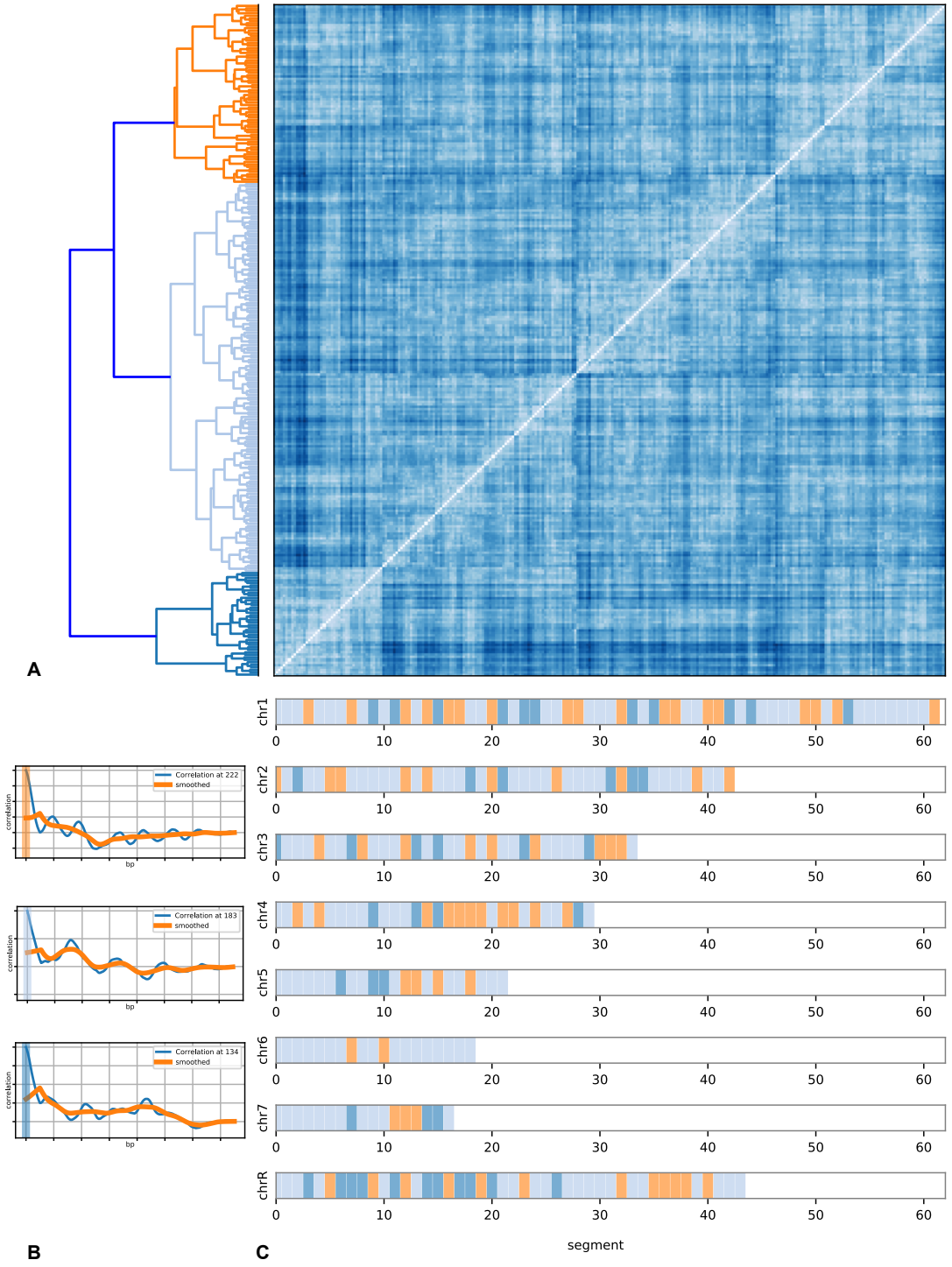

Figure 2: Panel A shows the genome-wide distance matrix between the correlation functions between segments of size 75kbp. Hierarchical clustering was applied to identify common patterns. The matrix was sorted according to the patterns. The left side shows the clustering. Panel B shows the coarse-grained nucleosomal density correlation functions of *Candida albicans* at 5kb coarse-graining. Panel C shows the genome-wide distribution of segments. The pattern classification was done genome-wide to yield three main patterns. These three patterns were assigned colors and the segments of each chromosome corresponding to one of the three patterns are marked.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

With diminishing height the likelihood of the ordering and the strictness of ordering vanishes. Notice that for some of the sections (within one class) many sub-peaks or side-peaks exist, indicating possible sub-orderings. An example on the more extreme side is chromosome 3 and sections such as $3, 5, 16$ etc. Overall, the short-range order is much less pronounced. The orange smoothed line indicates that in this class the salient feature is a smoothly decreasing function indicating a different kind of order than for the class with sections $0, 8, 12, ...$

Even looking at the correlation functions without the indicated class mapping shows that there are universal features beyond fluctuations. Within a class, a more or less pronounced ordering feature is visible. Comparing the different correlation data between the chromosomes, these become apparent.

These observations can be proven more rigorously by applying similarity measures between the correlation functions. Figure 2 shows the resulting distance matrix between all chromosomes and all sections (the individual results are shown in the supplementary information Figure 4 to 6.) Shown is the distance matrix after reordering on the basis of similarity between sections. The color indicates the similarity between the correlation functions. Notice the patterns that emerge from the sorting of the data into classes.

These classes, represented by different colors, are shown in the dendrogram. These classes were obtained by hierarchical clustering. In the lower part of the figure on the left are the typical correlation functions representing the corresponding class with its color code. The orange colored class shows a regular pattern on a short scale whereas the light blue class shows a less stringent regular but still pronounced pattern on a slightly larger scale. The blue colored class shows a rather irregular pattern compared to the other two classes.

These observations are consistent with the typical classification from microscopy data into hetero- and euchromatin. The data shows that the orange and light blue classes can be mapped on heterochromatin. The blue colored class thus is euchromatin. The data also shows that still within any of these classes, the features have many sub-features that we salvaged for the larger patterns to allow a "coarse-grained" view on the ordering of the nucleosomes.

Notice that this partitioning into classes is genome-wide. A consistent classification can be established. This is shown in the mapping of the positions of the section to the chromosomes. Notice that, as expected, not a random mixture of the three colors emerges but rather a clear pattern. The larger chromosomes appear to have more internal structuring compared to the smaller chromosomes that are more homogeneous in their internal structure.

Discussion

The structural organisation of the genome depends on the patterns of nucleosome positioning and their distribution in the genome. At a higher scale, the nucleosome positioning distribution varies across the chromosomes which appear to be conserved along the entire genome. The classification of the chromosomes into segments of the distinct nucleosomal distribution shown here is inline with earlier studies. Although, two major classifications of the chromosomal region as heterochromatin and euchromatin are suggested, we find that their organisations can be further subdivided. Nucleosomes can be well-positioned to form phased and un-phased arrays consisting of regularly spaced nucleosomes or can be fuzzy to form irregular arrays of nucleosomes. The three distinct nucleosome distribution patterns along the genome obtained in our result are in agreement with this study. Moreover, further classification of nucleosomal distribution is obtained along each chromosome. Around five to seven different nucleosome distribution patterns are observed for all chromosomes. However, for the entire genome, three patterns are found to be conserved.

We have analyzed the effect for different $p = 2, 7$ in the p-norm on the outcome of the clustering of similar correlation functions and the outcome comes to be similar for all p. For high p values some of the clusters split into further clusters. In addition, the cosine similarity norm was tested for further verification, yielding similar clustering (see Supplementary Information Section 1.11). This rules out that the clustering is an artifact of the model and its architecture.

Three distinct patterns of nucleosome organisation appear to be conserved in the genome. These kinds of distinct patterns observed in the genome correspond to different gene densities and gene expressions inside the cell. Recent studies by Wiese et. al.(16) suggested that domain formation and genome organisation can be predicted with nucleosome positioning only. So, the distinct patterns obtained from our calculation correspond to different ways of nucleosome positioning and may control domain formation and genome organisation in the cell.

Acknowledgments

This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1 - 390900948 (the Heidelberg STRUCTURES Excellence Cluster). The authors gratefully acknowledge the data storage service SDS@hd supported by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) and the German Research Foundation (DFG) through grant INST 35/1314-1 FUGG and INST 35/1503-1 FUGG. Kunhe Li would like to acknowledge funding by the Chinese Scholarship Council (CSC). Sujeet Kumar Mishra would like to acknowledge funding by the India government Ministry of Science and Technology, Department of Biotechnology (DBT)-Interdisciplinary Research Center for Scientific Computing (IWR) PhD program.

Author Contribution Statement

All authors were involved in the conception, processing of the data, analysis and draftting of the manuscript. D.H. and N.O. supervised the work. All authors discussed the results and commented on the manuscript.

Conflicts of Interest

No conflicts of interest exists.

References

[1] Oluwadare O, Highsmith M, Cheng J. An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data. Biological Procedures Online. 2019;21(1):7. doi:10.1186/s12575-019-0094-0.

[2] Jerkovic I, Cavalli G. Understanding 3D genome organization by multidisciplinary methods. Nature Reviews Molecular Cell Biology. 2021;22(8):511–528. doi:10.1038/s41580-021-00362-w.

[3] Routh A, Sandin S, Rhodes D. Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure. Proceedings of the National Academy of Sciences of the United States of America. 2008;105(26):8872–8877. doi:10.1073/pnas.0802336105.

[4] Bohr J, Olsen K. The size of the nucleosome; 2012. Available from: http://arxiv.org/abs/1102.0761.

[5] Staneva D, Georgieva M, Miloshev G. Kluyveromyces lactis genome harbours a functional linker histone encoding gene. FEMS Yeast Research. 2016;16(4). doi:10.1093/femsyr/fow034.

[6] Bohn M, Diesinger P, Kaufmann R, Weiland Y, Müller P, Gunkel M, et al. Localization Microscopy Reveals Expression-Dependent Parameters of Chromatin Nanostructure. Biophysical journal. 2010;99:1358–67. doi:10.1016/j.bpj.2010.05.043.

[7] Tchasovnikarova IA, Kingston RE. Beyond the Histone Code: A Physical Map of Chromatin States. Molecular Cell. 2018;69(1):5–7. doi:https://doi.org/10.1016/j.molcel.2017.12.018.

[8] Struhl K, Segal E. Determinants of nucleosome positioning. Nat Struct Mol Biol. 2013;20(3):267–273. doi:10.1038/nsmb.2506.

[9] Hilbert L, Sato Y, Kuznetsova K, Bianucci T, Kimura H, Jülicher F, et al. Transcription organizes euchromatin via microphase separation. Nature Communications. 2021;12(1):1360. doi:10.1038/s41467-021-21589-3.

[10] Dekker J, Rippe K, Dekker M, Kleckner N. Capturing Chromosome Conformation. Science. 2002;295(5558):1306–1311. doi:10.1126/science.1067799.

[11] van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, et al. Hi-C: a method to study the three-dimensional architecture of genomes. Journal of visualized experiments : JoVE. 2010;6(39):1869. doi:10.3791/1869.

[12] Beagan JA, Phillips-Cremins JE. On the existence and functionality of topologically associating domains. Nature Genetics. 2020;52(1):8–16. doi:10.1038/s41588-019-0561-1.

[13] Ghavi-Helm Y, Jankowski A, Meiers S, Viales RR, Korbel JO, Furlong EEM. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. Nature genetics. 2019;51(8):1272–1282. doi:10.1038/s41588-019-0462-3.

[14] Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012;485(7398):381–385. doi:10.1038/nature11049.

[15] Quentin S, Frédéric B, Giacomo C. Principles of genome folding into topologically associating domains. Science Advances;5(4):eaaw1668. doi:10.1126/sciadv.aaw1668.

[16] Wiese O, Marenduzzo D, Brackley CA. Nucleosome positions alone can be used to predict domains in yeast chromosomes. Proceedings of the National Academy of Sciences. 2019;116(35):17307. doi:10.1073/pnas.1817829116.

[17] Kharerin H, Bai L. Thermodynamic modeling of genome-wide nucleosome depleted regions in yeast. PLOS Computational Biology. 2021;17(1):e1008560–.

[18] Klein DC, Hainer SJ. Genomic methods in profiling DNA accessibility and factor localization. Chromosome Research. 2020;28(1):69–85.

[19] Chen W, Liu Y, Zhu S, Green CD, Wei G, Han JDJ. Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data. Nature communications. 2014;5(1):1–14.

[20] Georgakilas GK, Perdikopanis N, Hatzigeorgiou A. Solving the transcription start site identification problem with ADAPT-CAGE: a Machine Learning algorithm for the analysis of CAGE data. Scientific Reports. 2020;10(1):877. doi:10.1038/s41598-020-57811-3.

[21] Oiwa NN, Cordeiro CE, Heermann DW. The Electronic Behavior of Zinc-Finger Protein Binding Sites in the Context of the DNA Extended Ladder Model. Frontiers in Physics. 2016;4.

[22] Singh AK, Mueller-Planitz F. Nucleosome positioning and spacing: from mechanism to function. Journal of Molecular Biology. 2021; p. 166847.

[23] Schöpflin R, Teif VB, Müller O, Weinberg C, Rippe K, Wedemann G. Modeling nucleosome position distributions from experimental nucleosome positioning maps. Bioinformatics. 2013;29(19):2380–2386.

[24] Flores O, Orozco M. nucleR: a package for non-parametric nucleosome positioning. Bioinformatics. 2011;27(15):2149–2150.

[25] Chen K, Xi Y, Pan X, Li Z, Kaestner K, Tyler J, et al. DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. Genome research. 2013;23(2):341–351.

[26] Price RJ, Weindling E, Berman J, Buscaino A. Chromatin Profiling of the Repetitive and Nonrepetitive Genomes of the Human Fungal Pathogen Candida albicans. mBio. 2019;10(4):e01376–19. doi:10.1128/mBio.01376-19.

[27] Puri S, Lai WKM, Rizzo JM, Buck MJ, Edgerton M. Iron-responsive chromatin remodelling and MAPK signalling enhance adhesion in Candida albicans. Mol Microbiol. 2014;93(2):291–305. doi:10.1111/mmi.12659.

[28] Zhao Y, Wang J, Liang F, Liu Y, Wang Q, Zhang H, et al. NucMap: a database of genome-wide nucleosome positioning map across species. Nucleic acids research. 2019;47(D1):D163–D169.

[29] Kmiecik S, Gront D, Kolinski M, Wieteska L, Dawid AE, Kolinski A. Coarse-Grained Protein Models and Their Applications. Chemical Reviews. 2016;116(14):7898–7936. doi:10.1021/acs.chemrev.6b00163.

[30] Reichl LE. A Modern Course in Statistical Physics, 4th Edition. Wiley; 2016.

[31] pandas development team T. pandas-dev/pandas: Pandas; 2020. Available from: `https://doi.org/10.5281/zenodo.3509134`.

[32] Maimon O, Rokach L. Data Mining and Knowledge Discovery Handbook. 978th ed. Springer-Verlag Berlin Heidelberg; 2005.

[33] B S Everitt SL, Leese M. Cluster Analysis. 4th ed. Oxford University Press; 2001.

[34] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods. 2020;17:261–272. doi:10.1038/s41592-019-0686-2.

[35] Zhao Y, Wang J, Liang F, Liu Y, Wang Q, Zhang H, et al. NucMap: a database of genome-wide nucleosome positioning map across species. Nucleic Acids Research. 2019;47(D1):D163–D169. doi:10.1093/nar/gky980.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

Supporting information

1.1 Nucleosome Density

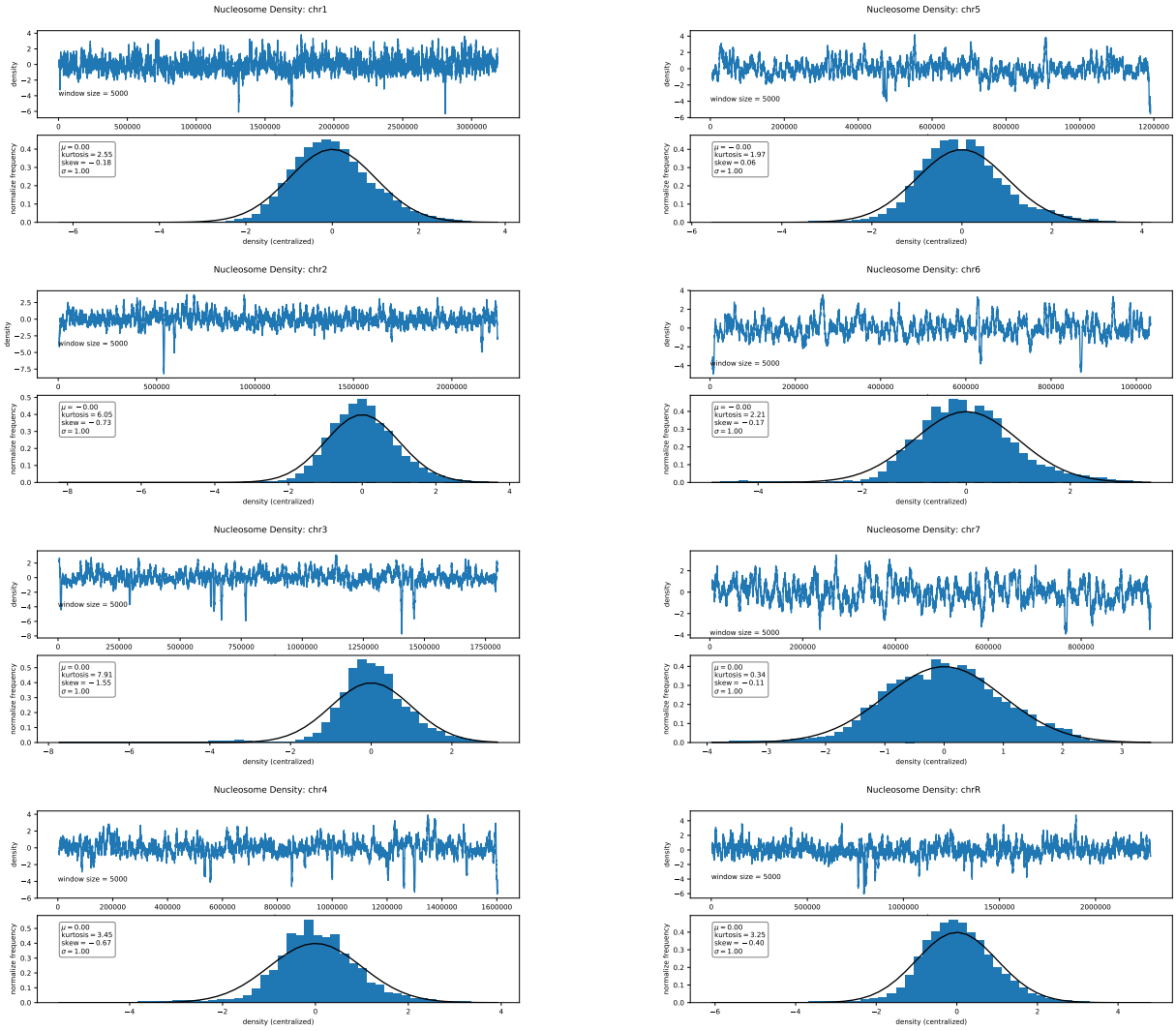

Figure 1: Shown is the nucleosomal density after applying a rolling average with a window size of 5000 of all of the chromosomes (upper panels). The lower panels show the corresponding histogram of the densities with a bin size of 50. The black line is the fit with a gaussian distribution.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

1.1.1 Nucleosome Density at $b = 2500$

Chromosome	Distribution	chi_square	D_statistic
chr1	fisk	1.675740e+05	0.026701
chr1	norm	4.029705e+05	0.047346
chr2	fisk	2.215488e+05	0.034197
chr2	norm	4.888579e+05	0.049294
chr3	fisk	2.315703e+05	0.038608
chr3	norm	1.174085e+06	0.083966
chr4	fisk	1.530916e+05	0.034538
chr4	norm	6.824904e+05	0.070372
chr5	fisk	9.322028e+04	0.030918
chr5	norm	2.783759e+05	0.056306
chr6	fisk	1.280710e+05	0.037654
chr6	norm	2.753396e+05	0.052656
chr7	fisk	5.100021e+04	0.032512
chr7	norm	7.679109e+04	0.031660
chrR	fisk	2.258400e+05	0.033580
chrR	norm	6.023527e+05	0.054608

Table 1: The Fisk distribution, also known as the log-logistic distribution gives the best consistent fit. The fit was done for the bin size of 50 and the rolling average of size 5000. Statistical Kolmogorov-Smirnov test for goodness of fit was done using SciPy.org scipy.stats.kstest function (34). The D statistic is the absolute max distance (supremum) between the CDFs of the two samples. All results show small values D values corresponding to p-values close to 1, the log-logistic distribution may explain the data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

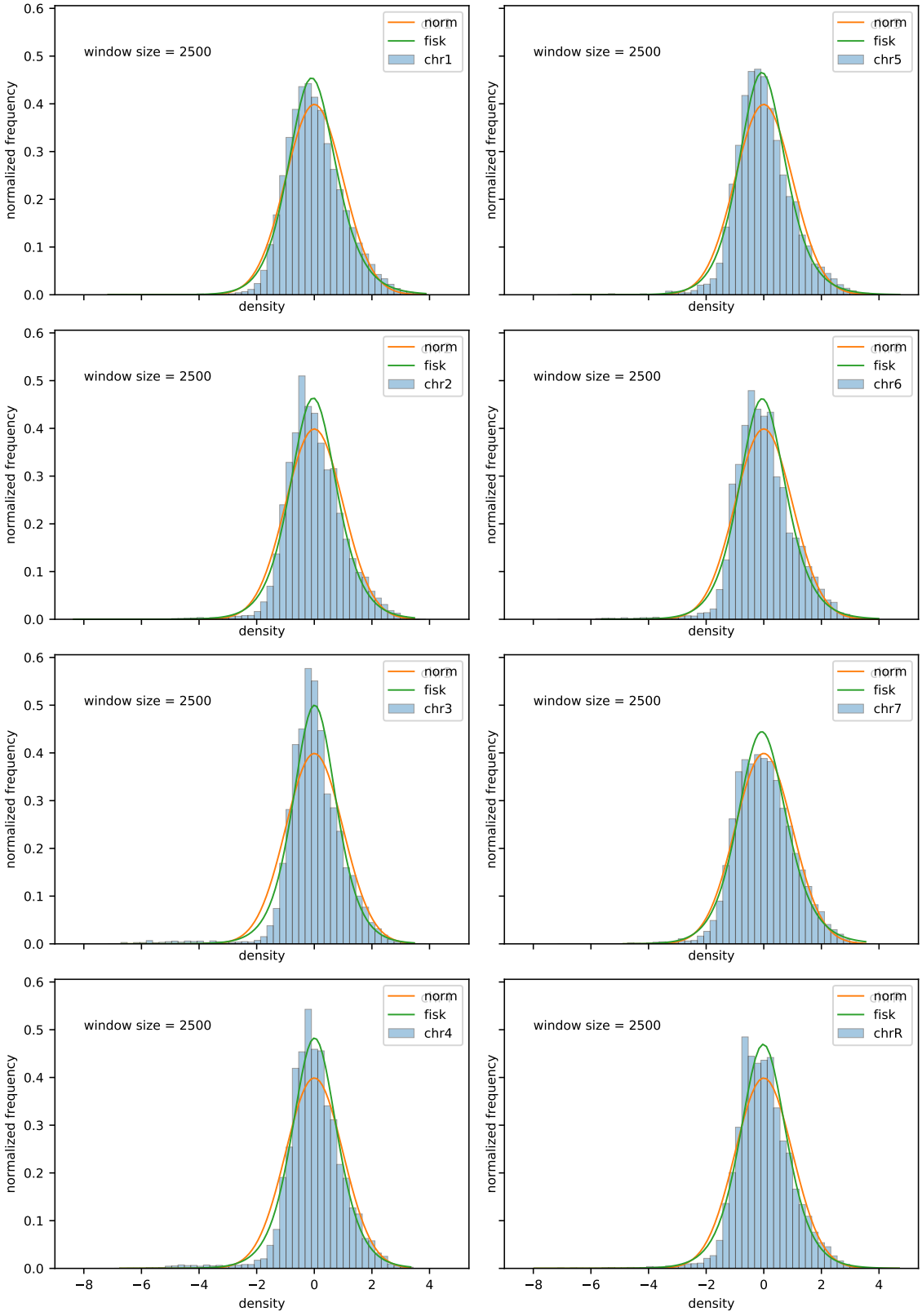

Figure 2: Normalized nucleosome density distributions for all of the chromosomes. The data shows the non-gaussian behavior (red line). For comparison a fit to a Log-logistic distribution is shown yielding a much better consistent fit. The bin size was 50 and the rolling average of size 2500 was used.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

1.1.2 Nucleosome Density at $b = 5000$

Chromosome	Distribution	chi_square	D_statistic
chr1	fisk	1.678610e+05	0.021809
chr1	norm	4.310806e+05	0.040372
chr2	fisk	1.011539e+05	0.024922
chr2	norm	4.873082e+05	0.048215
chr3	fisk	2.078474e+05	0.038179
chr3	norm	1.362080e+06	0.094966
chr4	fisk	9.270418e+04	0.027728
chr4	norm	6.014198e+05	0.069622
chr5	fisk	4.004712e+04	0.020815
chr5	norm	1.715085e+05	0.048451
chr6	fisk	1.347119e+04	0.020806
chr6	norm	1.609603e+05	0.038205
chr7	fisk	1.682594e+04	0.022172
chr7	norm	1.636277e+04	0.016584
chrR	fisk	9.955245e+04	0.025810
chrR	norm	4.967464e+05	0.052443

Table 2: The Fisk distribution, also known as the log-logistic distribution gives the best consistent fit. The fit was done for the bin size of 50 and the rolling average of size 5000. Statistical Kolmogorov-Smirnov test for goodness of fit was done using SciPy.org scipy.stats.kstest function (34). The D statistic is the absolute max distance (supremum) between the CDFs of the two samples. As the all the results show small values D values corresponding to p-values close to 1, the log-logistic distribution may explain the data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

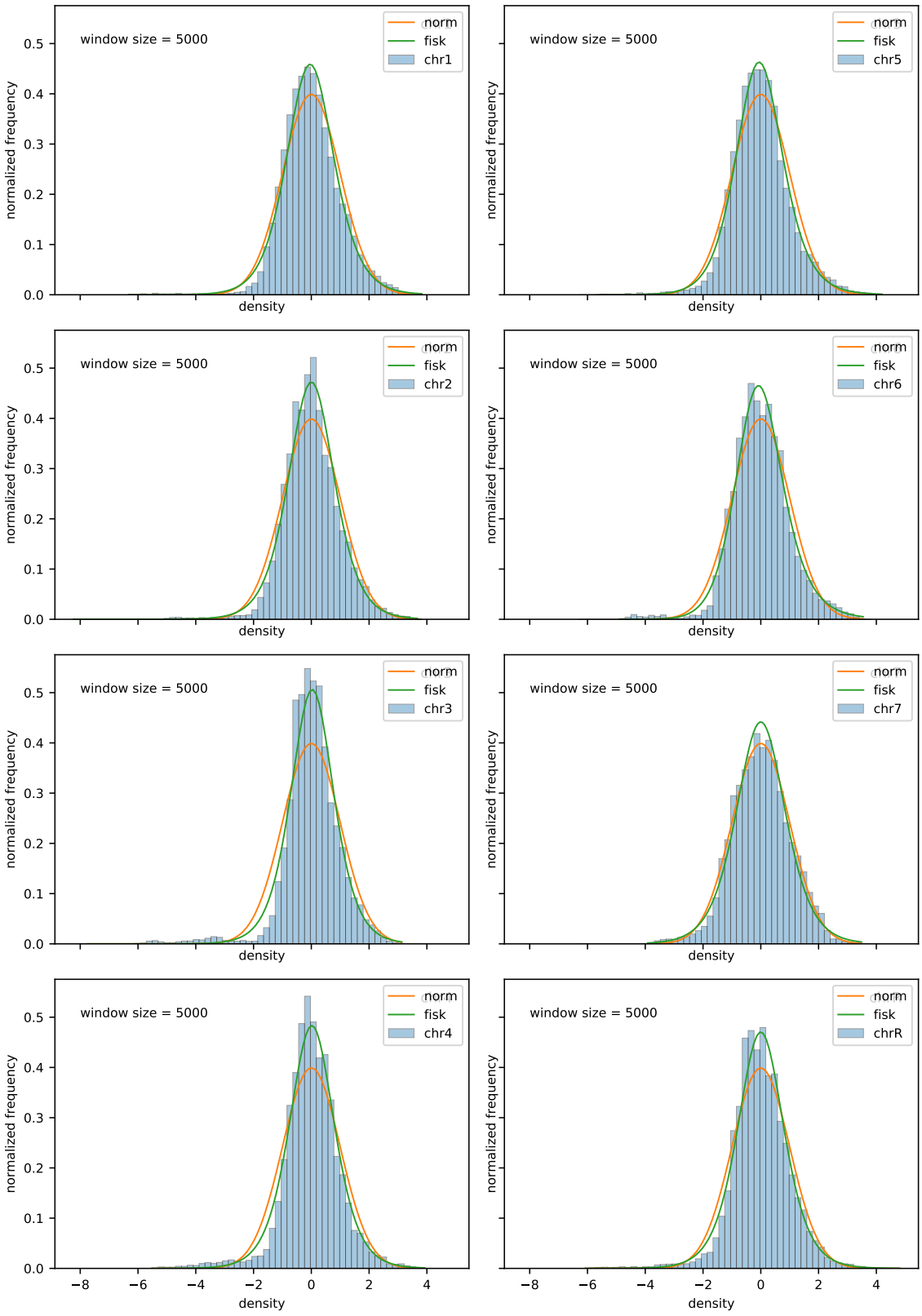

Figure 3: Normalized nucleosome density distributions for all of the chromosomes. The data shows the non-gaussian behavior (red line). For comparison a fit to a Log-logistic distribution is shown yielding a much better consistent fit. The bin size was 50 and the rolling average of size 5000 was used.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

1.2 Distance Matrix for Individual Chromosomes

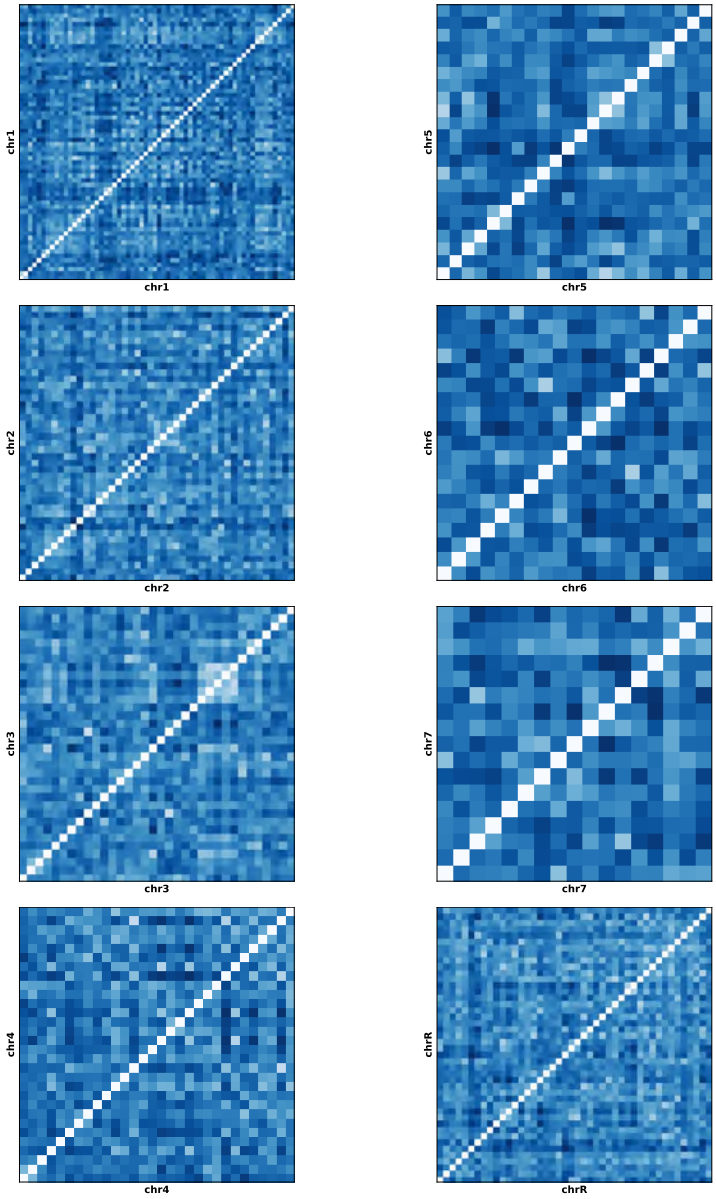

Figure 4: Shown are the distance matrices for all chromosomes. Distance refers to the distance between two correlation functions as measured by the euclidean distance (np.linalg.norm(x-y,ord=norm), with norm = 2. The ordering along the axes corresponds to the coarse-grained sections. The rolling average was of size 5000.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

1.3 Clustering for Individual Chromosomes

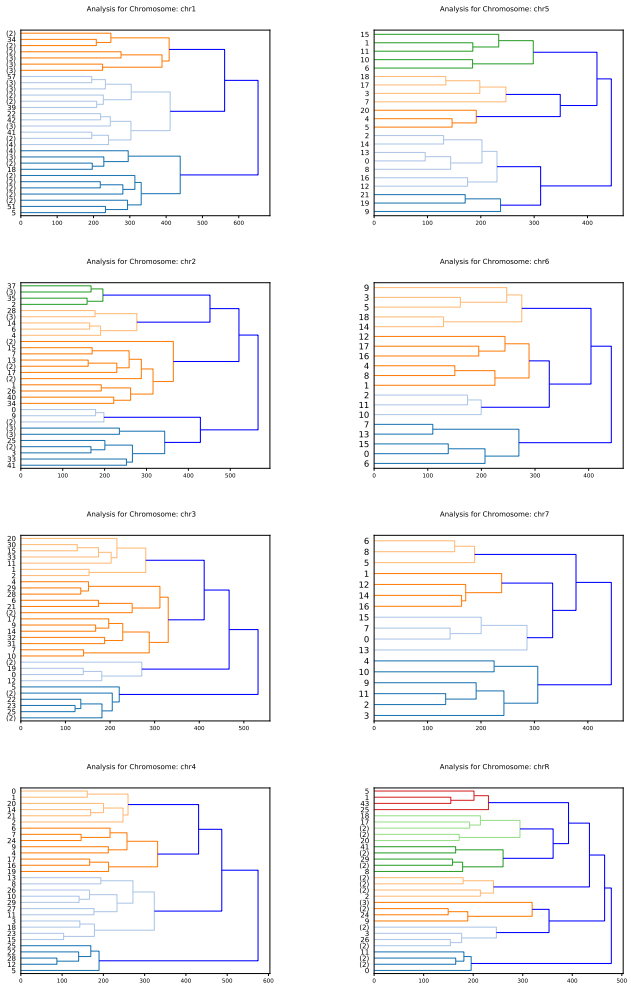

Figure 5: Shown are the dendrograms resulting from the distance matrices for all chromosomes. Results are for the hierarchical clustering on the individual chromosome. The Ward distance was used for the variance minimization algorithm used by SciPy (34). The labels correspond to the distance matrix entries. Labels in parentheses give the number of labels corresponding to the leave. The rolling average was of size 5000. Labels in parentheses give the number of labels corresponding to the leaf.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

1.4 Distance Matrix and Clustering for Individual Chromosomes

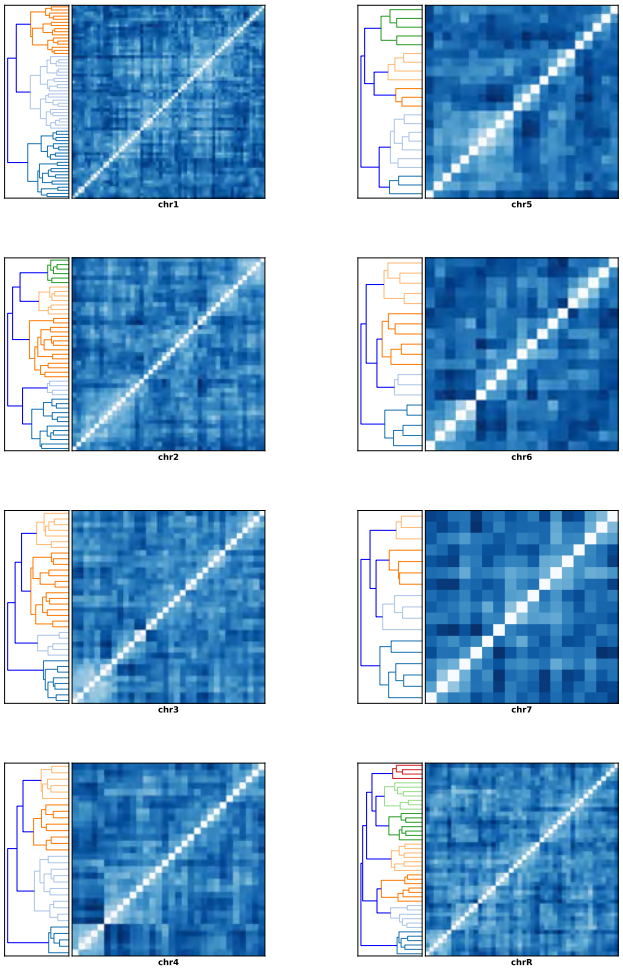

Figure 6: Shown are the distance matrices and corresponding dendrograms for all chromosomes. The matrix entries are sorted to correspond to the identified clusters. The rolling average was of size 5000.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

1.5 Cluster Pattern in Chromosomes

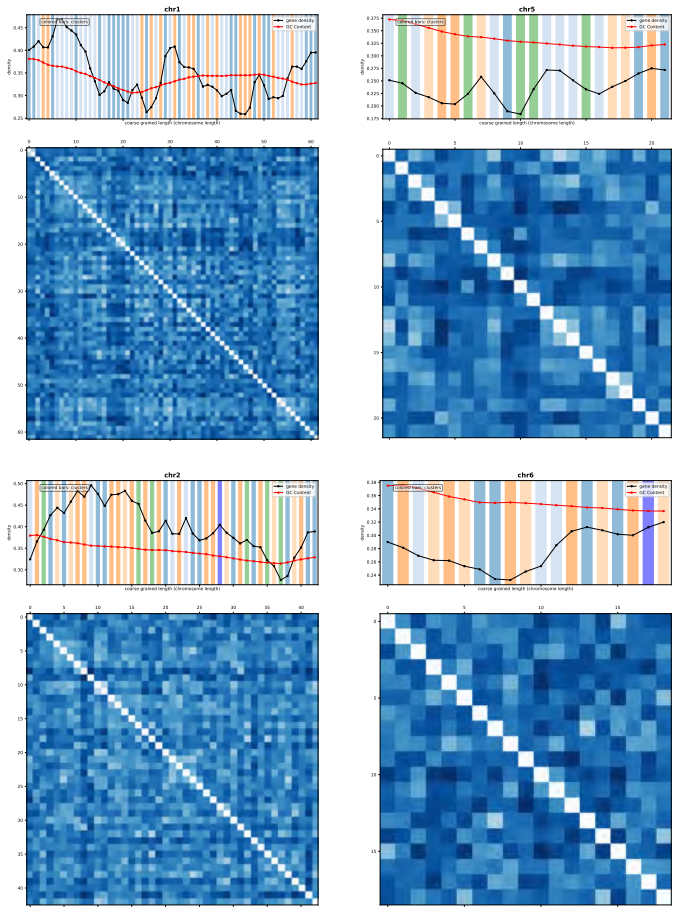

Figure 7: Part 1: Shown are the distance matrices and corresponding mapping of the pattern on the chromosomes. The matrix entries correspond to the positions on the chromosome. The rolling average was of size 5000.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

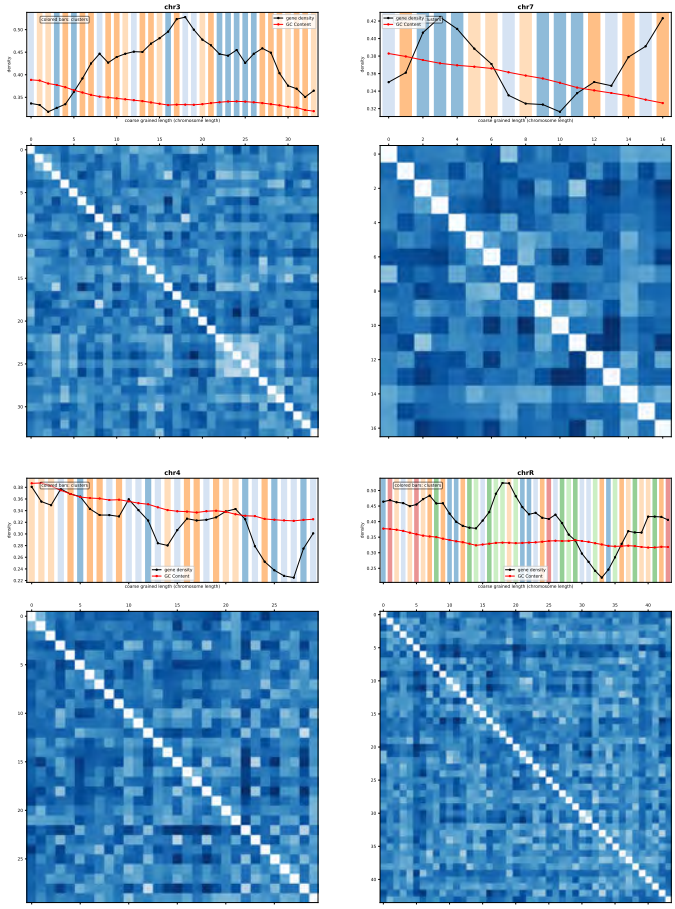

Figure 8: Part 2: Shown are the distance matrices and corresponding mapping of the pattern on the chromosomes. The matrix entries correspond to the positions on the chromosome. The rolling average was of size 5000.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

1.6 Correspondence between Pattern and Correlation Function within individual Chromosome

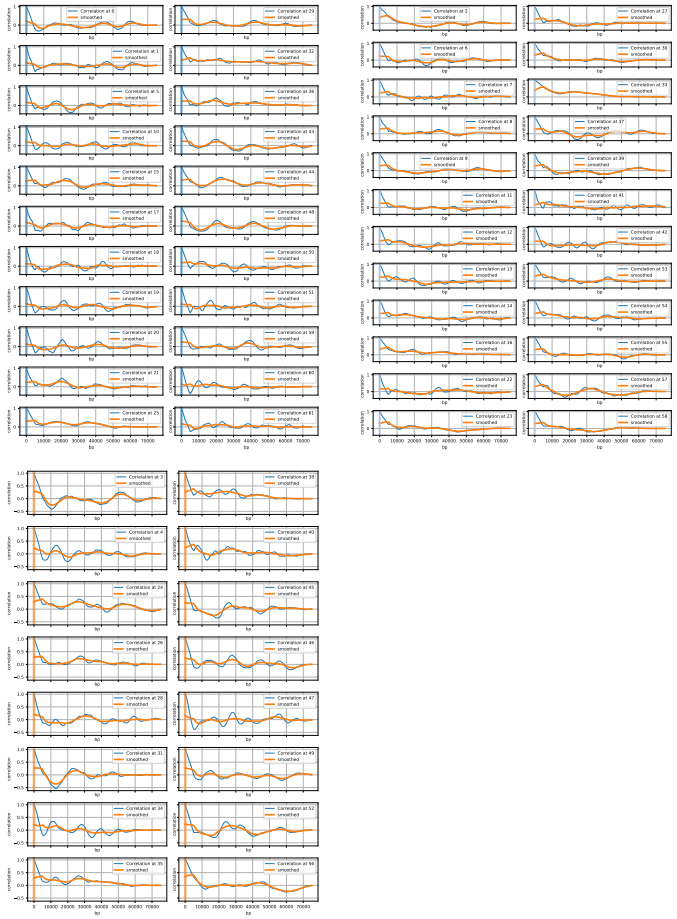

Figure 9: Shown are the correlation functions and the corresponding mapping of the pattern on the chromosome 1. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

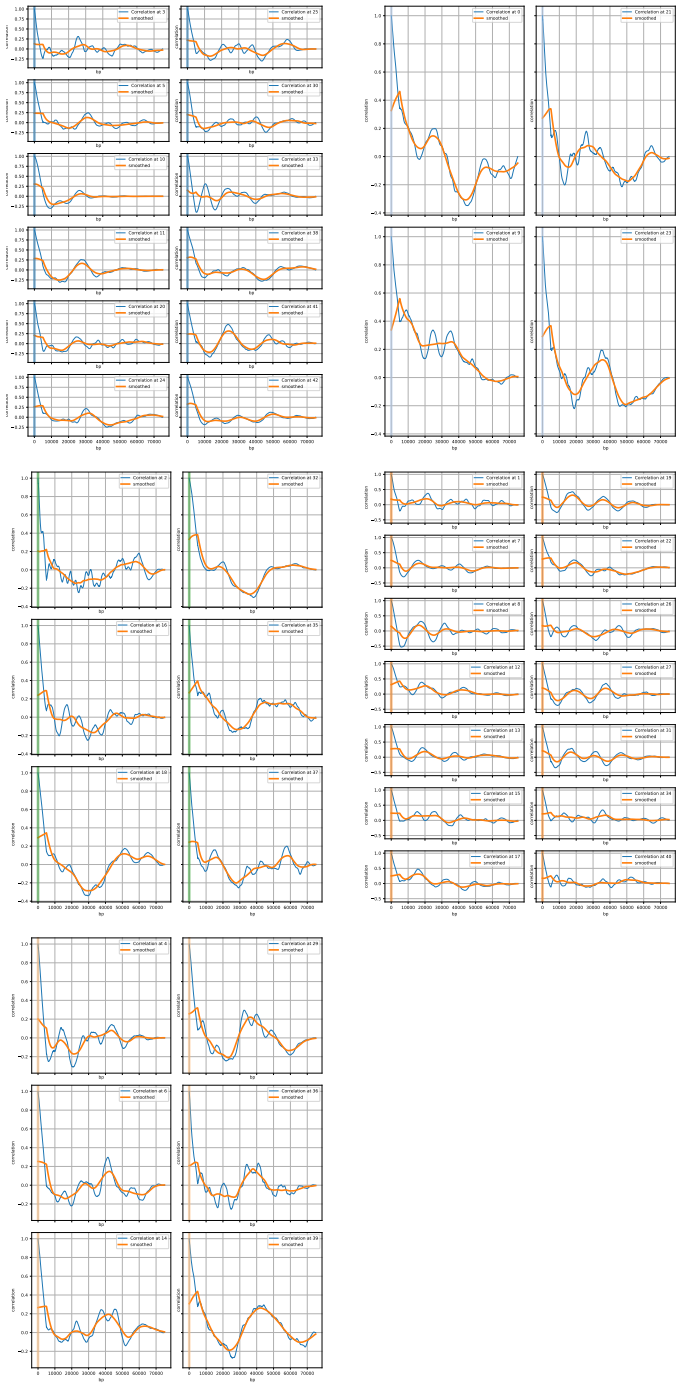

Figure 10: Shown are the correlation functions corresponding mapping of the pattern on the chromosome 2. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

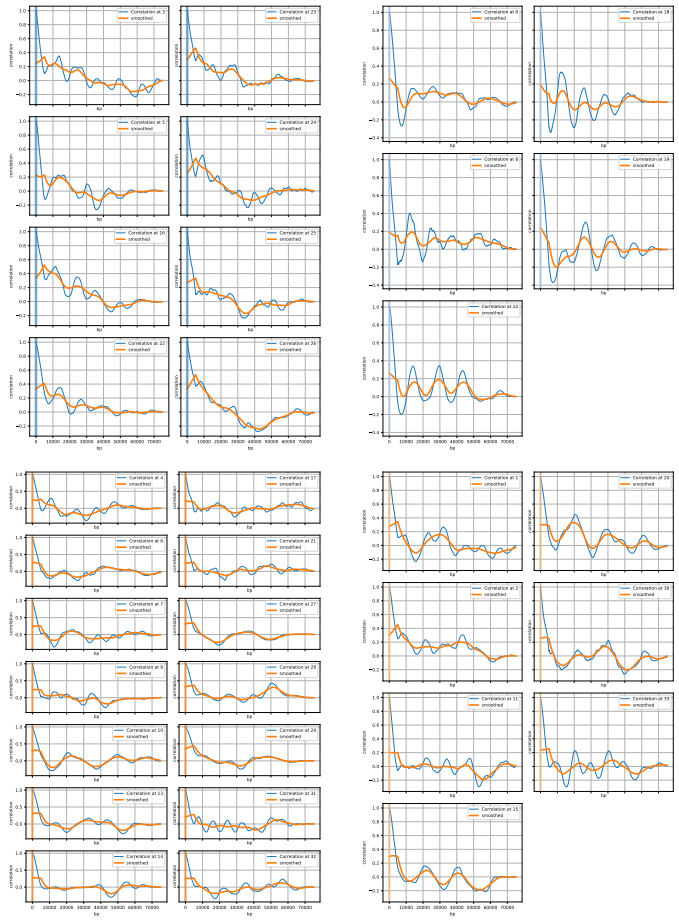

Figure 11: Shown are the correlation functions corresponding mapping of the pattern on the chromosome 3. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

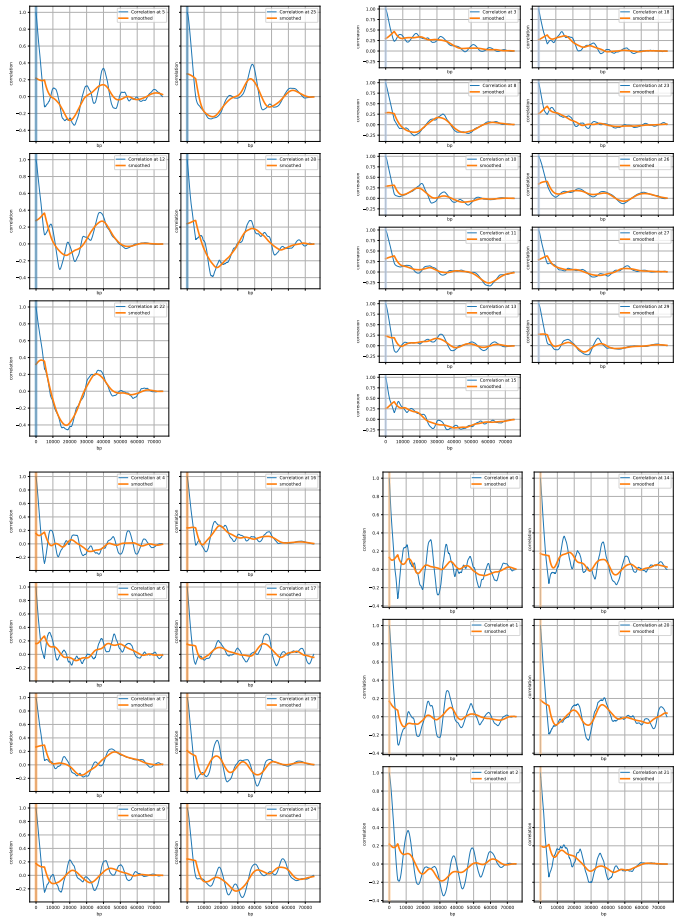

Figure 12: Shown are the correlation functions corresponding mapping of the pattern on the chromosome 4. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

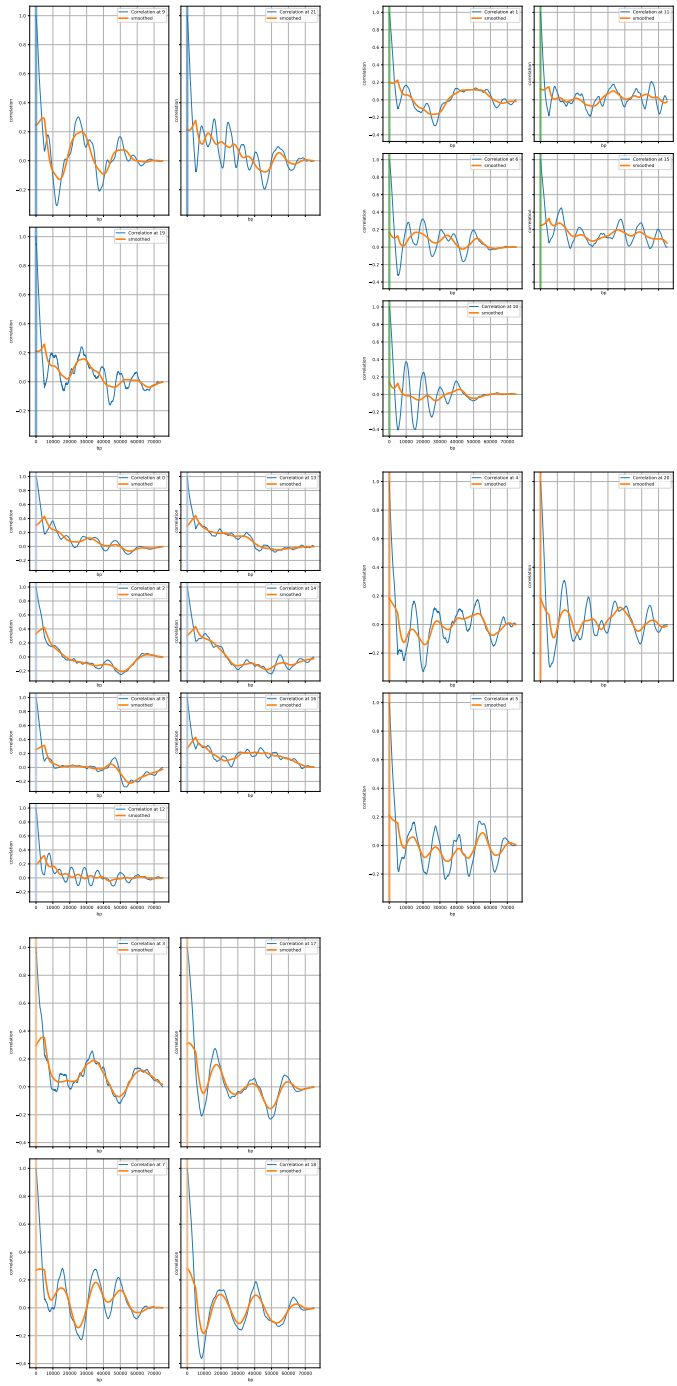

Figure 13: Shown are the correlation functions corresponding mapping of the pattern on the chromosome 5. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

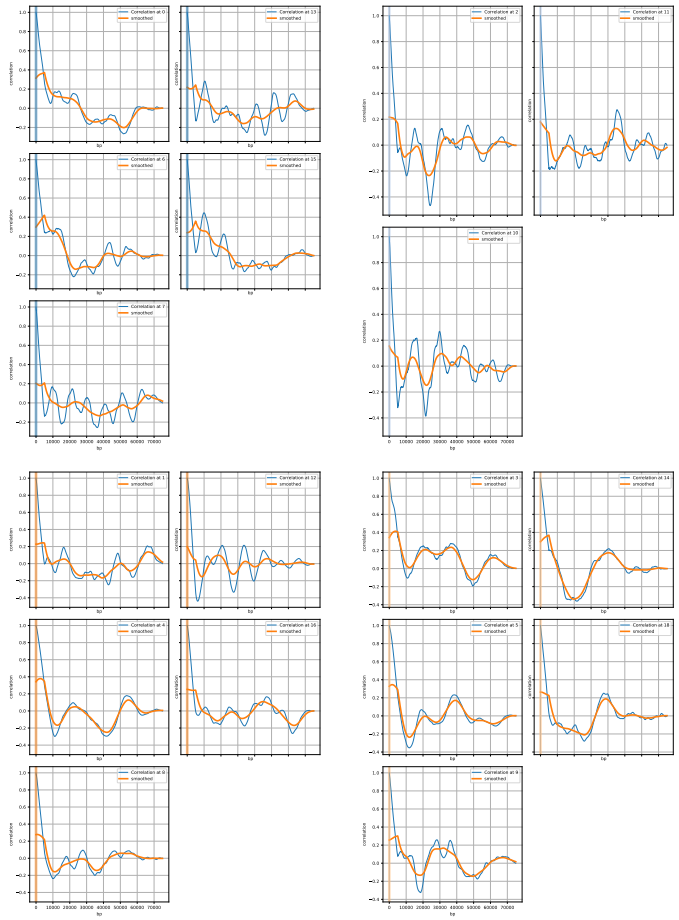

Figure 14: Shown are the correlation functions corresponding mapping of the pattern on the chromosome 6. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

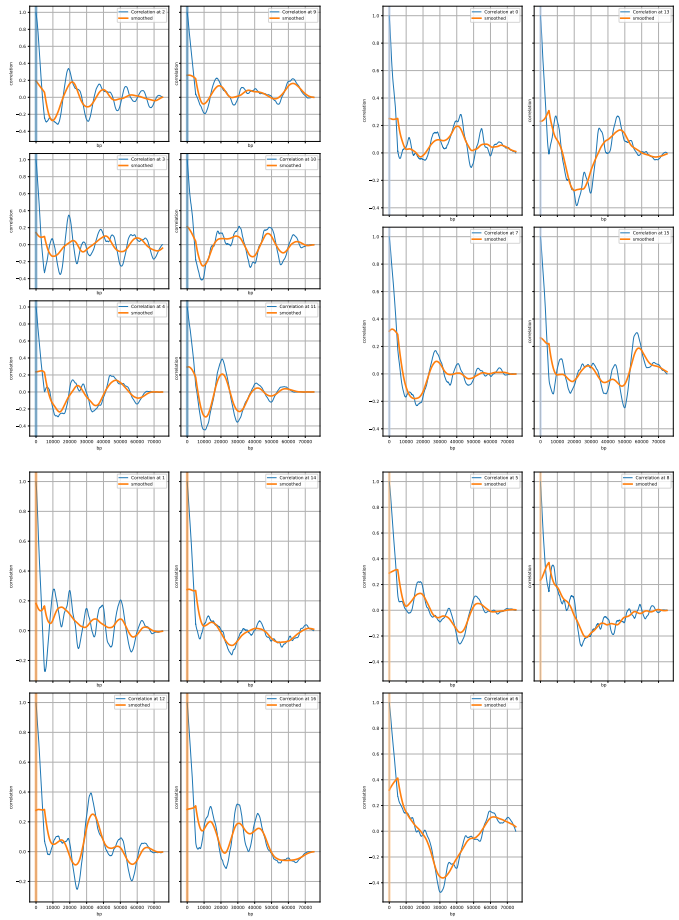

Figure 15: Shown are the correlation functions corresponding mapping of the pattern on the chromosome 7. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

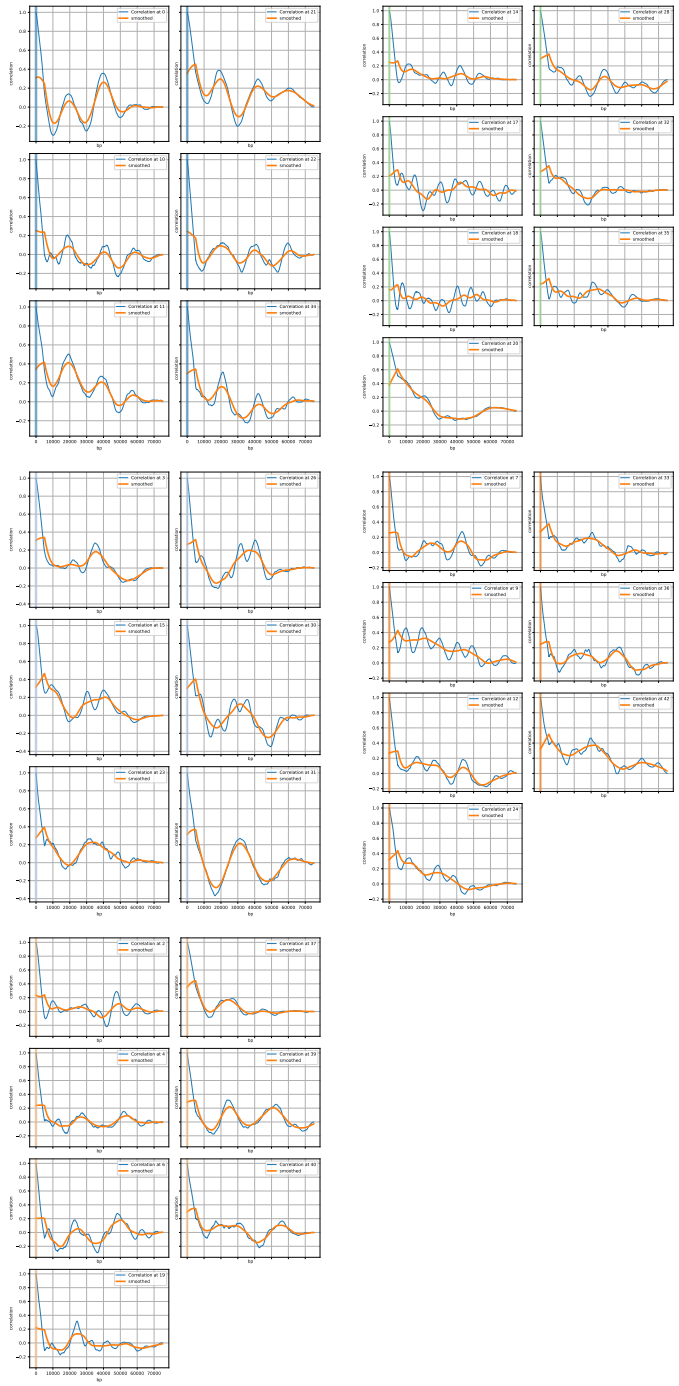

Figure 16: Shown are the correlation functions corresponding mapping of the pattern on the chromosome R. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

1.7 Genome-Wide Distance Matrix

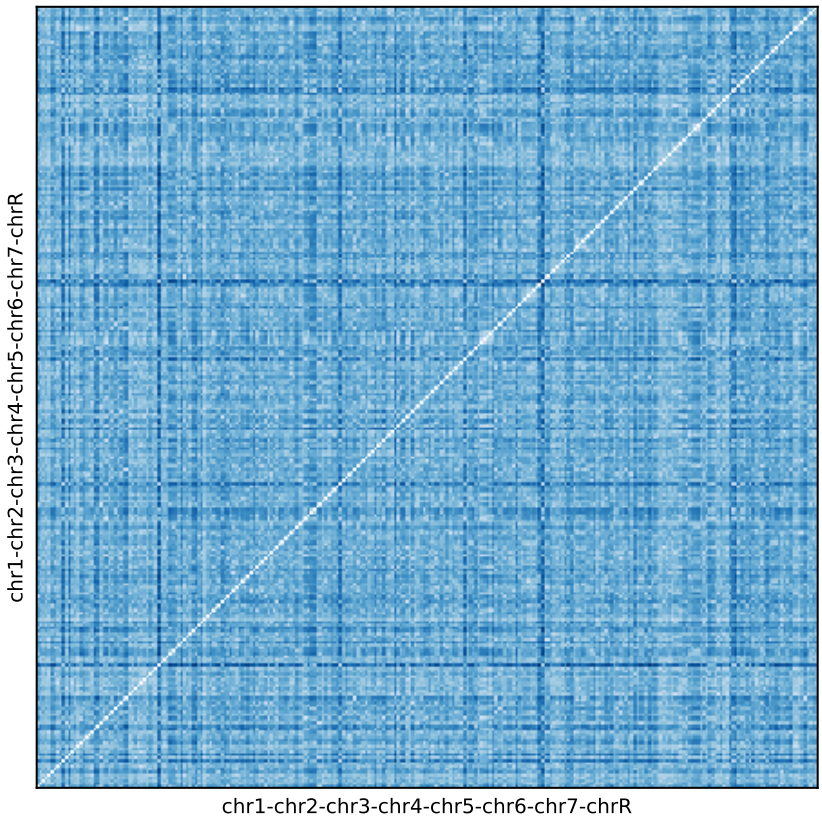

Figure 17: Shown is the genome-wide distance matrice. Distance refers to the distance between two correlation functions as measured by the euclidean distance (np.linalg.norm(x-y,ord=norm), with norm = 2. The ordering along the axes corresponds to the coarse-grained sections. The rolling average was of size 5000.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

1.8 Genome-Wide Clustering

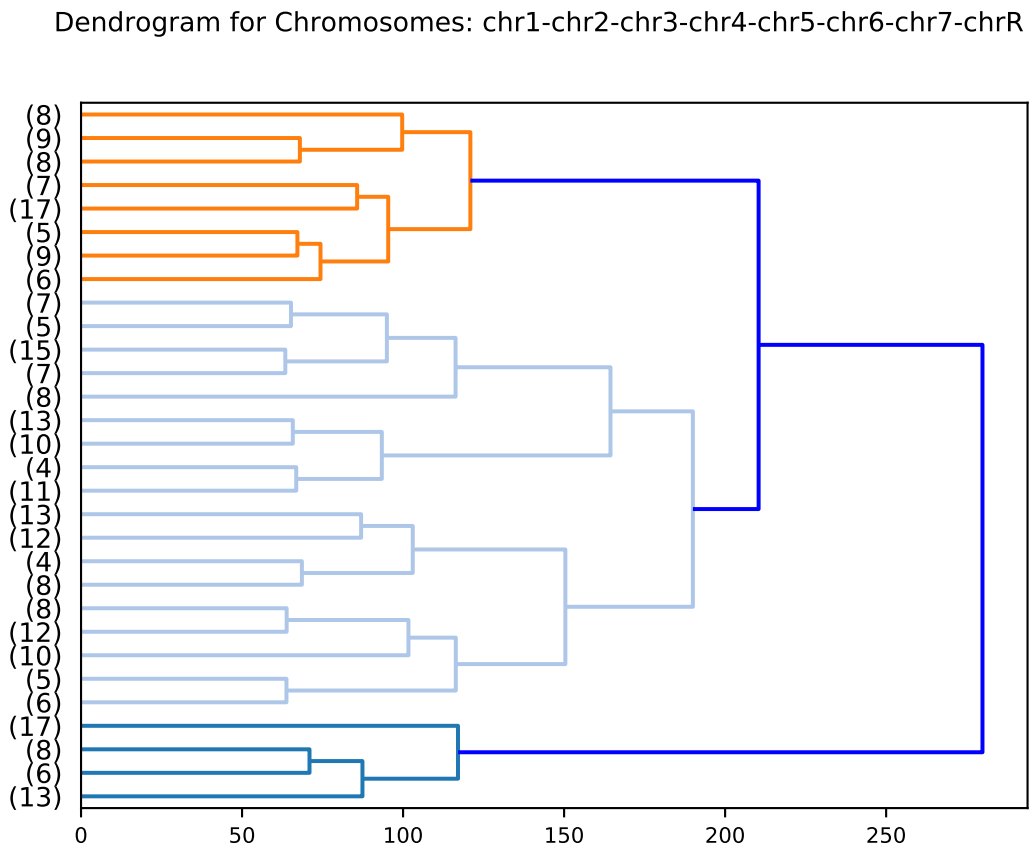

Figure 18: Shown is the dendrograms resulting from the genome-wide distance matrix. Results are for the hierarchical clustering on the individual chromosome. The Ward distance was used for the variance minimization algorithm used by SciPy (34). The labels correspond to the distance matrix entries. Labels in parentheses give the number of labels corresponding to the leave. The rolling average was of size 5000.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

1.9 Genome-Wide Distance Matrix and Clustering

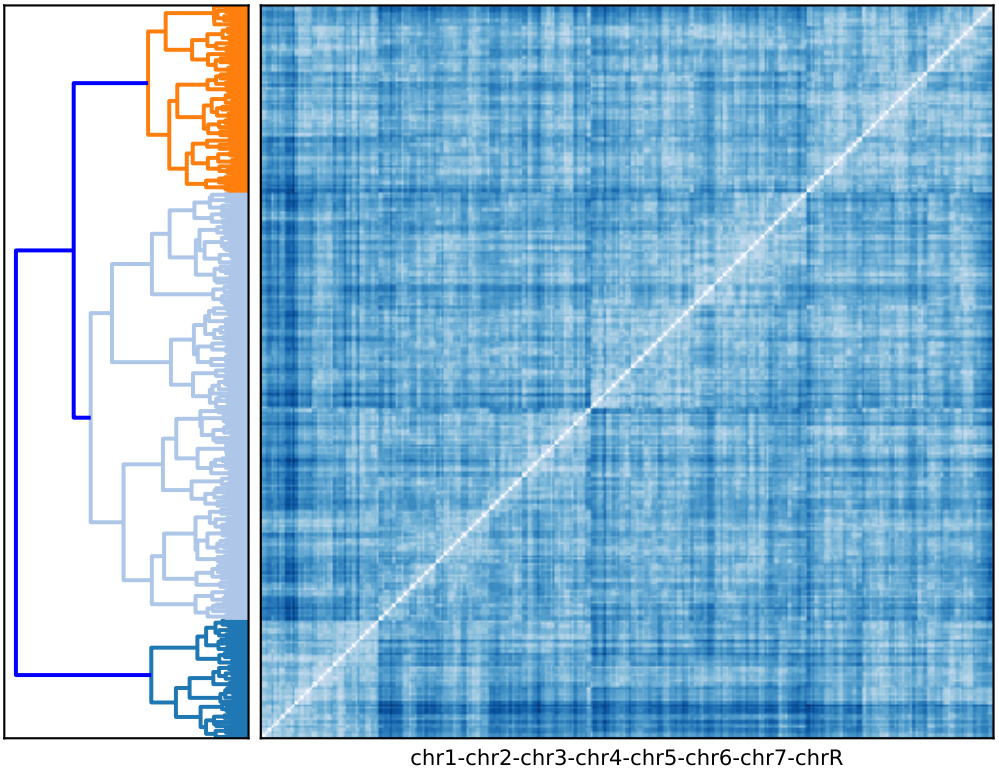

Figure 19: Shown is the genome-wide distance matrix and the corresponding dendrogram. The matrix entries are sorted to correspond to the identified clusters. The rolling average was of size 5000.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

1.10 Correspondence between Pattern and Correlation Function Genome-Wide

Correlation Function corresponding to the Pattern Genome-Wide Pattern No. 1

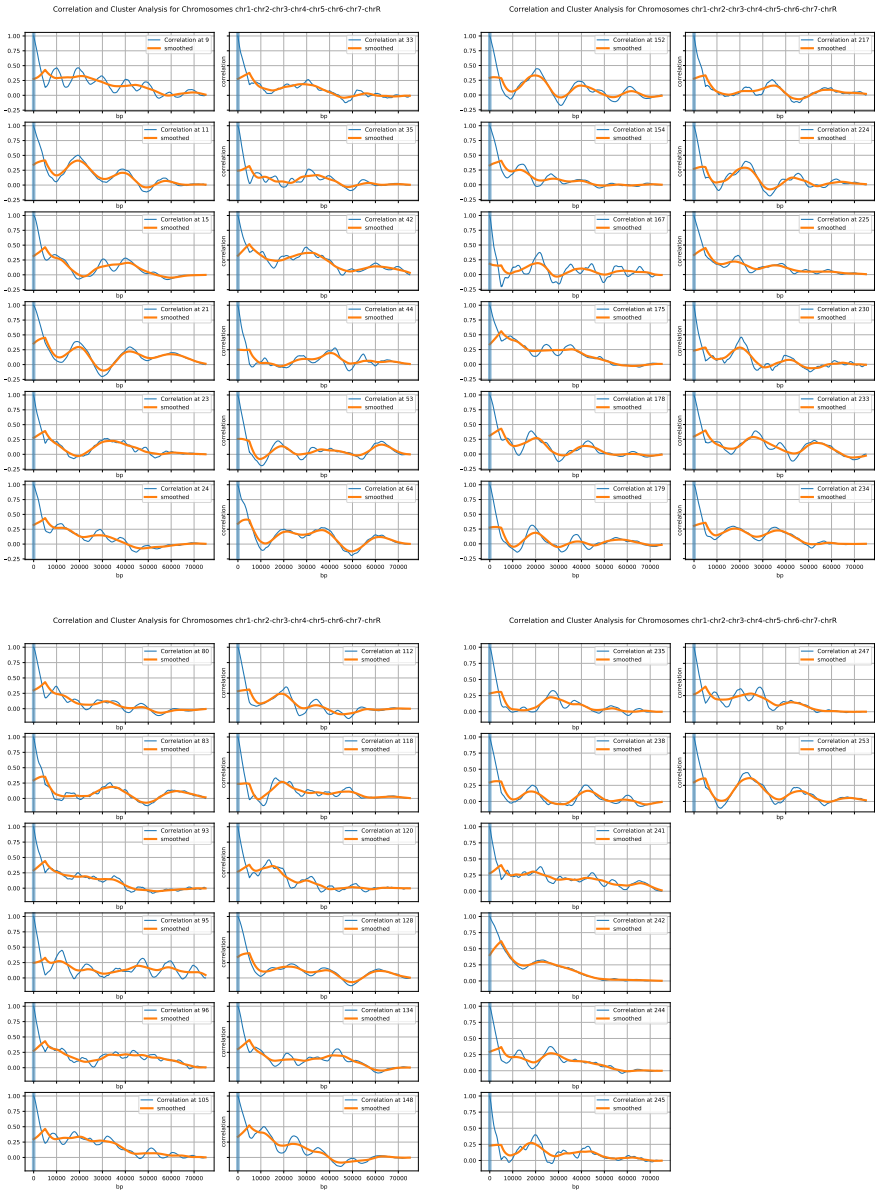

Figure 20: Shown are the correlation functions corresponding mapping of the pattern on the chromosomes. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

Correlation Function corresponding to the Pattern Genome-Wide Pattern No. 2 (1)

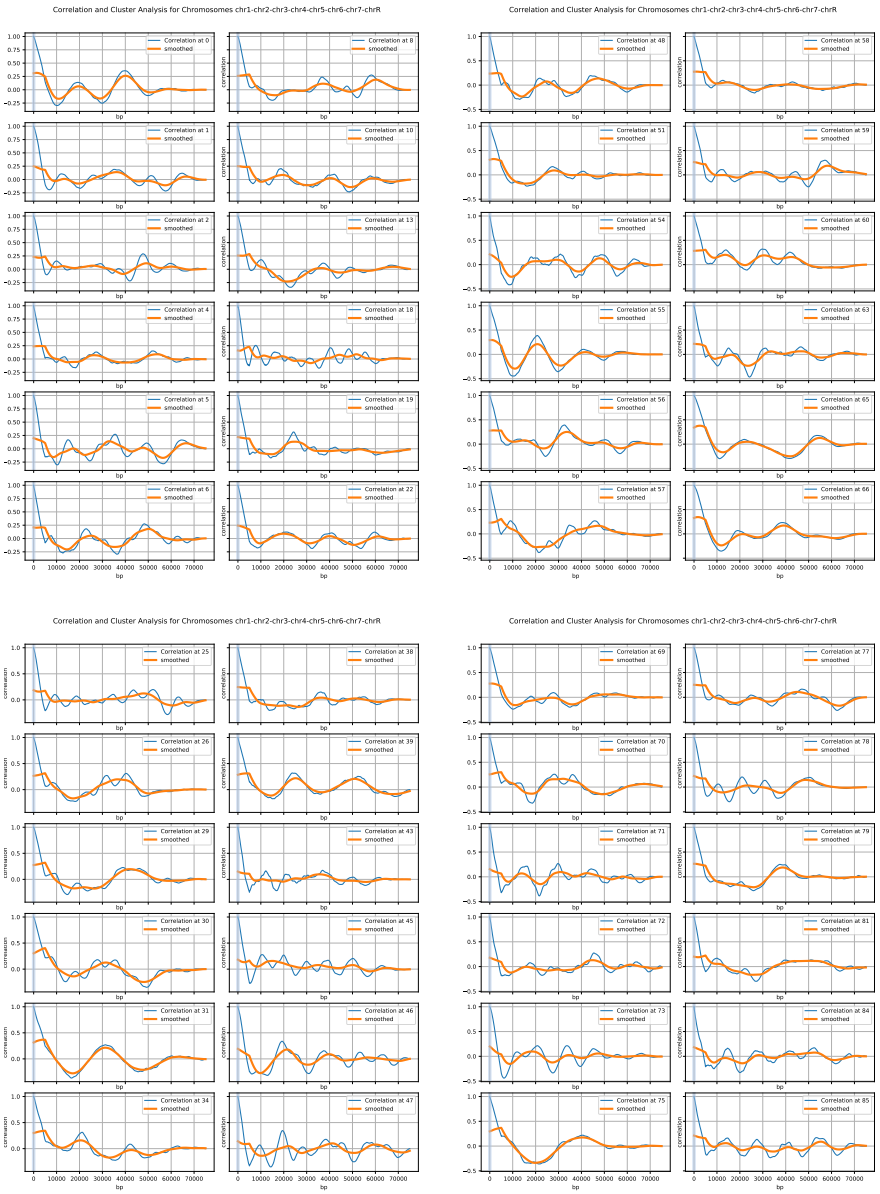

Figure 21: Shown are the correlation functions corresponding mapping of the pattern on the chromosomes. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

Correlation Function corresponding to the Pattern Genome-Wide Pattern No. 2 (2)

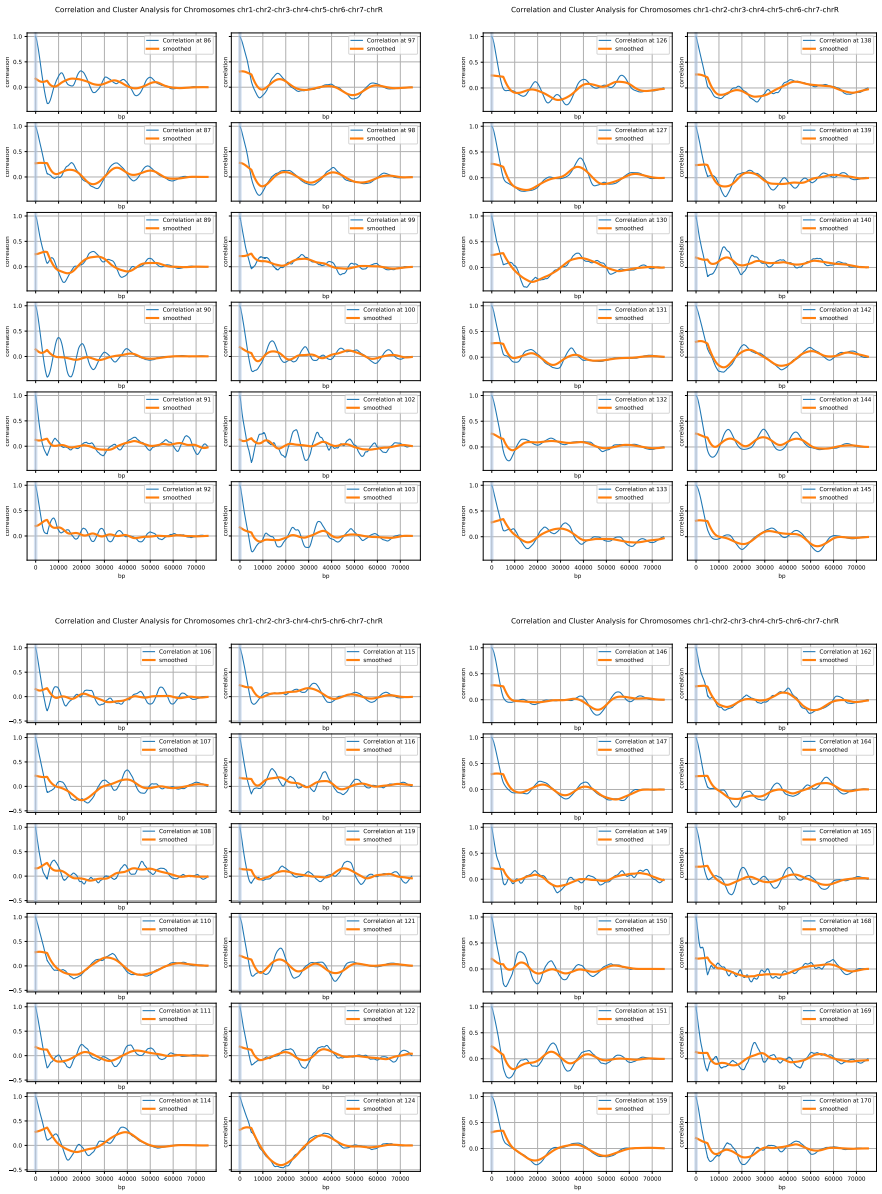

Figure 22: Shown are the correlation functions corresponding mapping of the pattern on the chromosomes. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

Correlation Function corresponding to the Pattern Genome-Wide Pattern No. 2 (3)

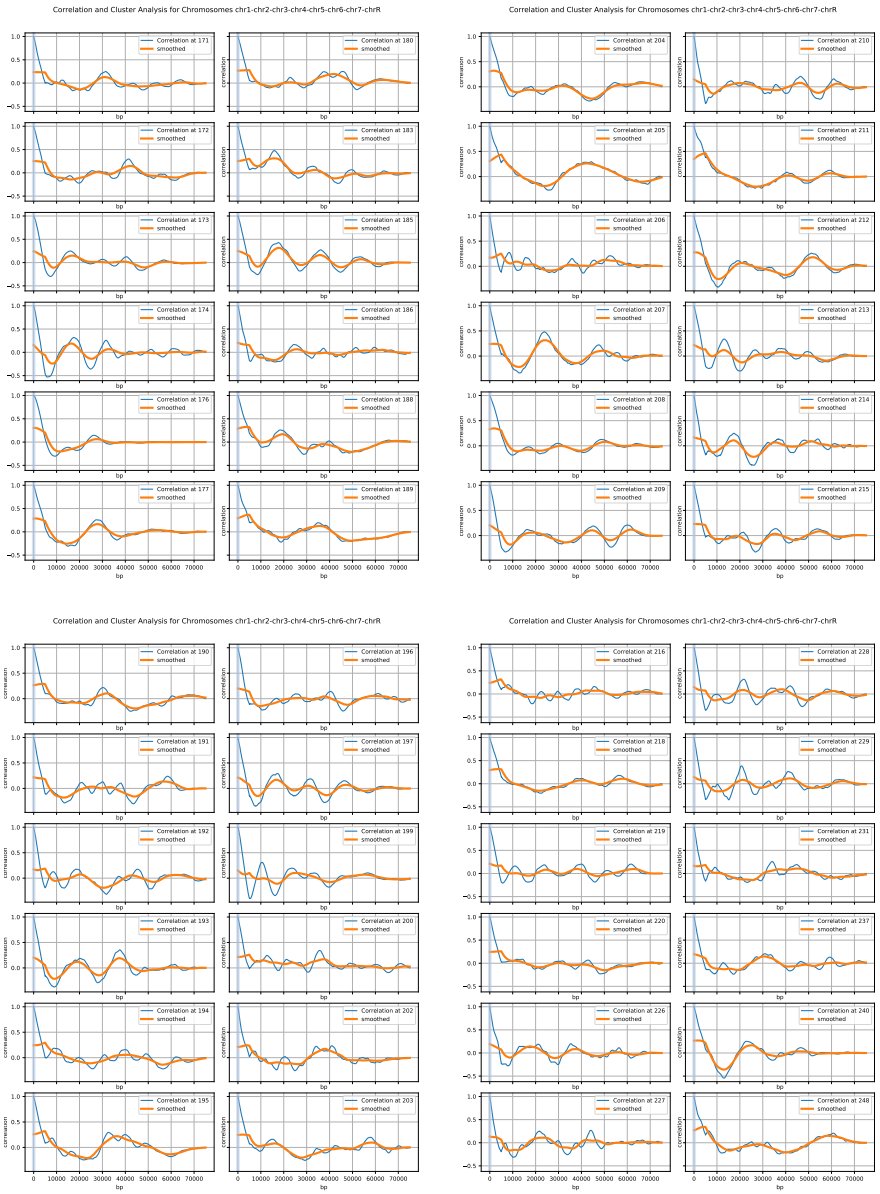

Figure 23: Shown are the correlation functions corresponding mapping of the pattern on the chromosomes. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

Correlation Function corresponding to the Pattern Genome-Wide Pattern No. 2 (4)

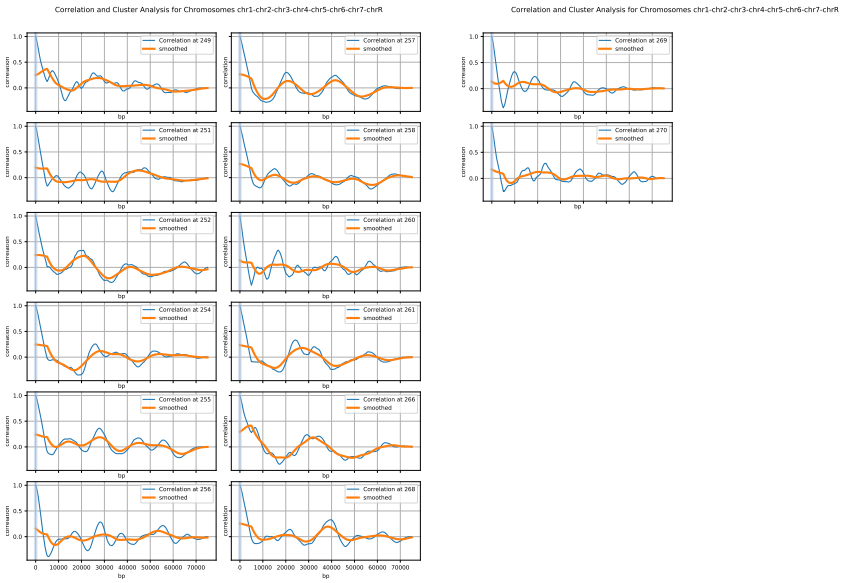

Figure 24: Shown are the correlation functions corresponding mapping of the pattern on the chromosomes. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

Correlation Function corresponding to the Pattern Genome-Wide Pattern No. 3 (1)

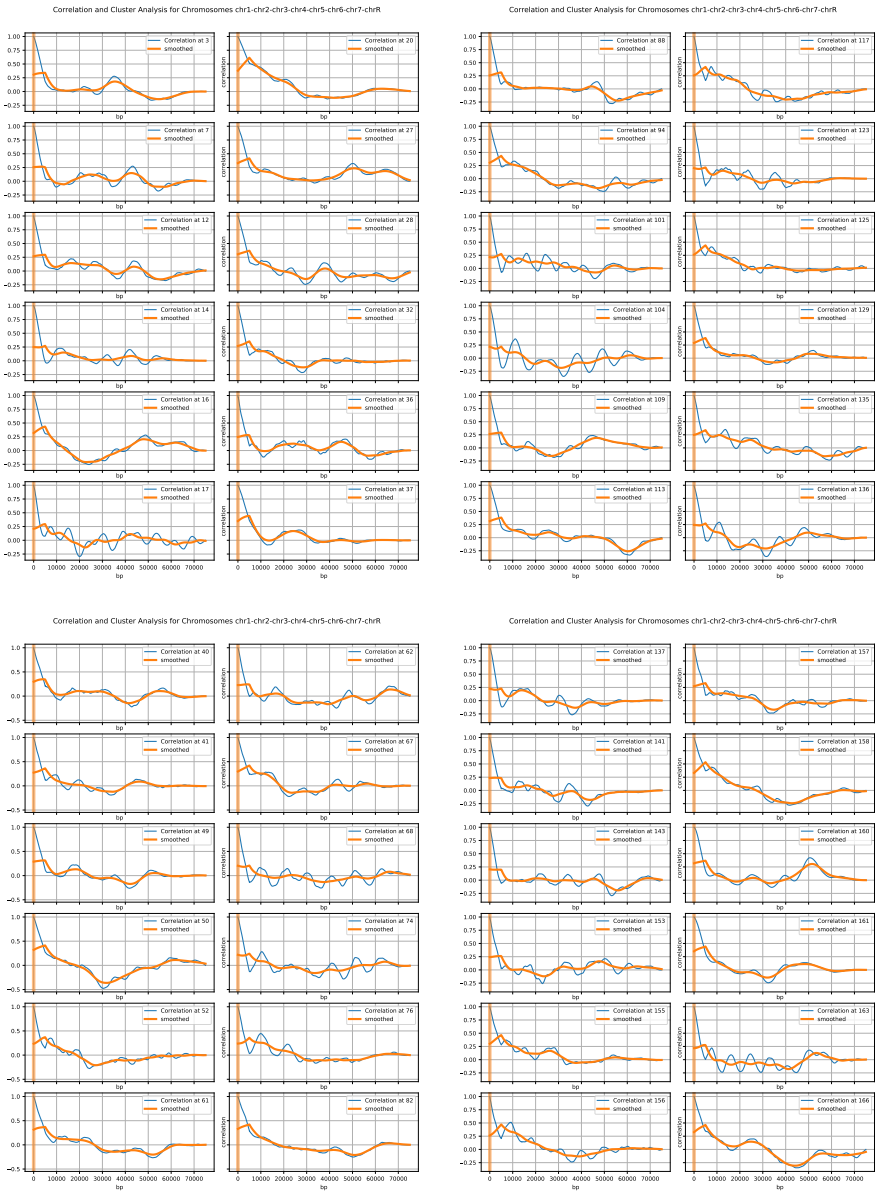

Figure 25: Shown are the correlation functions corresponding mapping of the pattern on the chromosomes. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

Correlation Function corresponding to the Pattern Genome-Wide Pattern No. 3 (2)

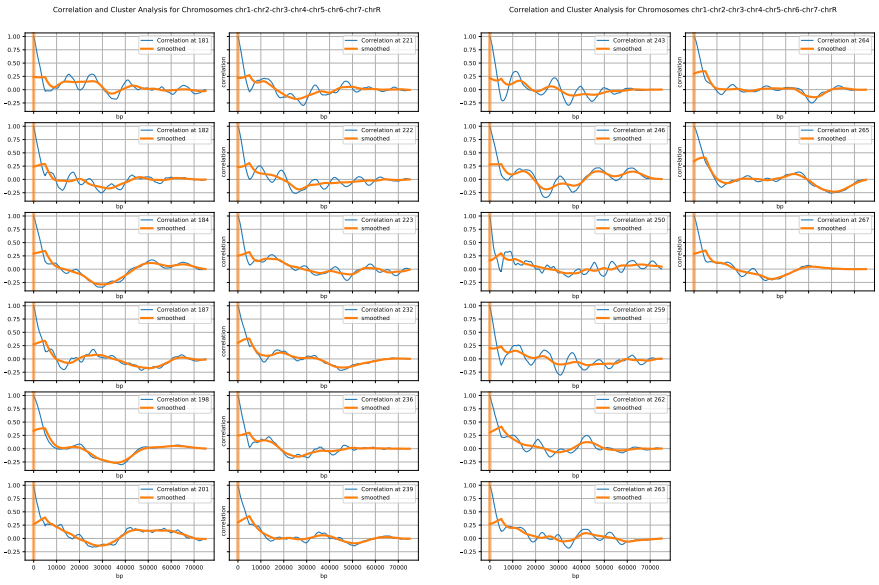

Figure 26: Shown are the correlation functions corresponding mapping of the pattern on the chromosomes. The rolling average was of size 5000. The orange line marked "smoothed" is a smoothed representation of the correlation function (rolling average of size 10000 to highlight the feature commonality between the clustered correlation functions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 January 2022 doi:10.20944/preprints202201.0471.v1

1.11 Comparison of Different Metrics

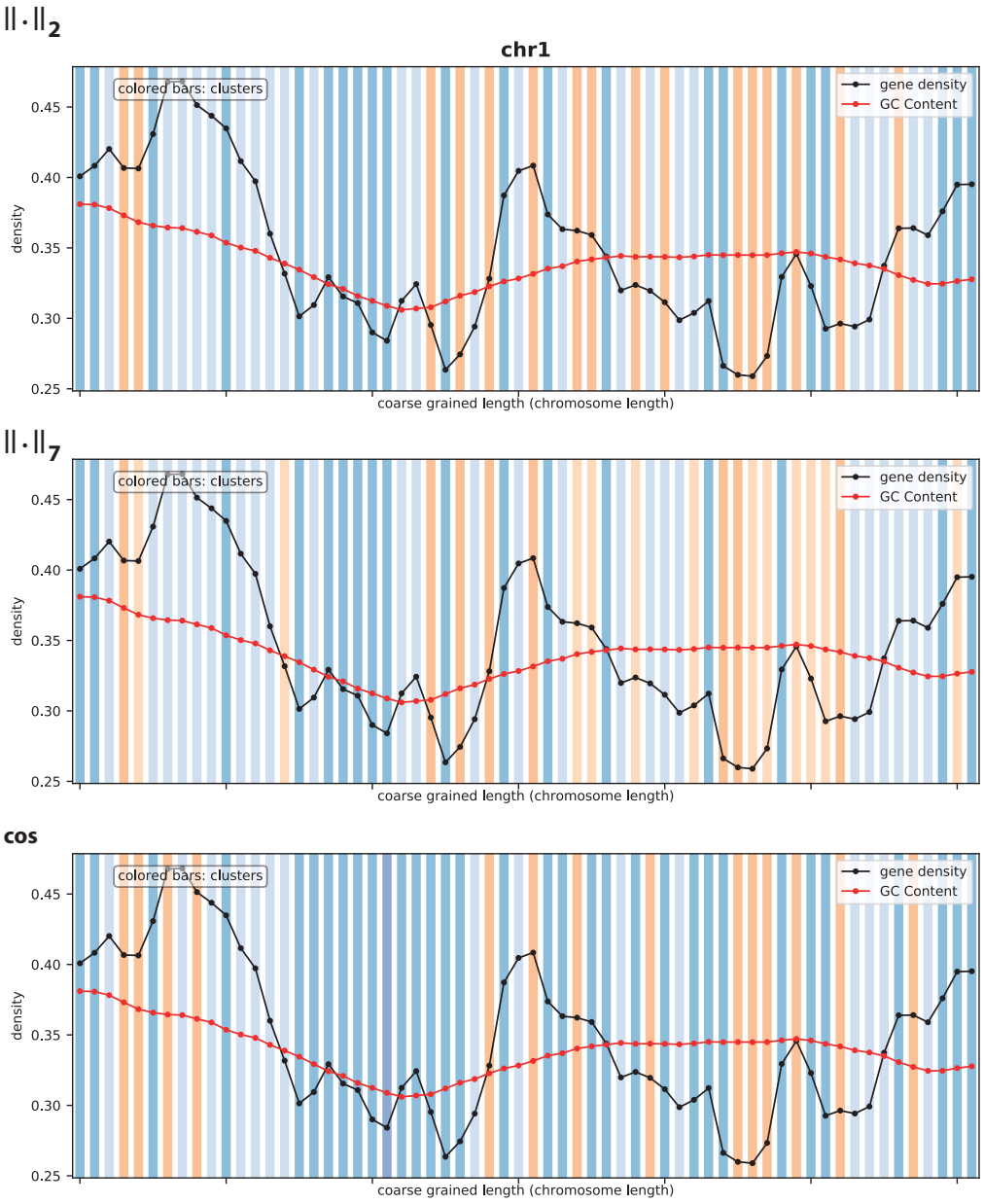

Figure 27: The upper panel shows the classification of the structures with respect to the euclidean distance $\|.\|_2$ while the middle one shows the result for $\|.\|_7$. Note that $\|.\|_7$ shows a further subdivision of the orange colored regions. Otherwise, the structure is stable against the two metrics for the distance between two correlation functions. The black line shows the gene density and the red line the GC content. The lower panel shows the application of the cosine similarity measure. While there are differences between the different metric, overall, a stable pattern is observed. What is remarkable is that the similarity measure shows less variation within certain domains than the other measures.