bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

21.12.16

THE GENOMIC CODE: ISOCHORES ENCODE AND MOLD CHROMATIN DOMAINS

Giorgio Bernardi

gbernardi@uniroma3.it

Science Department, Roma Tre University, Viale Marconi 446, 00146 Rome, Italy, and

Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy.

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

The mechanisms of formation of LADs, the lamina associated domains, and TADs, the topologically associating domains of chromatin, are not yet understood, even if several interesting models have been proposed so far. This problem was approached here by using as a starting point the observation that chromatin architecture has an isochore framework and by having a closer look at both the isochore structure and the isochore/chromatin domain connection. This approach showed that LADs essentially correspond to isochores from the very GC-poor, compositionally very homogeneous L1 family and from the "low-heterogeneity" L2 (or L2$^-$) sub-family; LADs are compositionally flat, flexible chromatin structures that attach themselves to the nuclear lamina in self-interacting clusters. In contrast, TADs (neglecting those that coincide with LADs) comprise the increasingly GC-richer isochores from the "high-heterogeneity" L2 (or L2+) sub-family and from the H1, H2 and H3 families; these isochores were found to be in the form of single or (more frequently) multiple GC peaks that underlie individual chromatin loops or chromatin loop ensembles, respectively. TADs appear to be constrained into loops by the properties that accompany the increasing GC levels of isochore peaks, namely increasing bendability, increasing nuclease accessibility, decreasing supercoiling and decreasing nucleosome density. In conclusion, a "genomic code" underlies the encoding and shaping of chromatin architecture by the compositional features of isochores.

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Isochores, the long (>200Kb), compositionally fairly homogeneous DNA stretches of eukaryotic genomes, were discovered forty years ago (1) during an effort to better understand the compositional compartmentalization of the mammalian genome (2), a genome organization which was detected by fractionating DNA by Cs_2SO_4 gradient ultracentrifugation in the presence of sequence-specific ligands (3). Isochores have been studied since, leading to a number of biologically important insights (4-6).

As far as the correlation of isochore composition with chromatin structure is concerned, it was concluded already some thirty years ago, that "the different GC levels of isochores, their different ratios of CpG to GpC, and the accompanying differences in potential methylation sites are bound to be associated with differences in DNA and chromatin structure and, possibly, with differences in the regulation of gene expression" (7). Many significant advances along this line were made since, for instance by finding strong correlations between isochore composition and all the structural and functional properties of DNA and chromatin that could be tested (4-6).

More specifically, a recent analysis (5,6) revealed the existence of correlations between the composition of isochore families and the properties of 1) LADs, the lamina associated domains (~0.5Mb median size), that are scattered over all chromosomes and correspond to GC-poor sequences (8-10); and 2) TADs, the topologically associating domains (0.2-2 Mb in size), a system of GC-rich loops and boundaries that can be resolved into contact domains (0.185 Mb median size) and that are largely conserved in mammals (11-14). Very recently, a comparison of maps of chromatin domains with maps of isochores from mouse and human chromosomes showed that isochores are the genomic units that underlie chromatin domains (15).

In spite of the recent, impressive advances in our understanding of chromatin structure (see refs. 16-22, for reviews), the problem of the formation mechanism(s) of LADs and TADs is still unsolved. Interesting models have been proposed (23-31), but no satisfactory solution has been reached so far. A

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

critical review (17) discusses two models for the formation of TADs that can be summarized as follows: 1) in the "handcuff model" (32), the two ends of a loop are brought together by architectural proteins such as the sequence-specific, zinc-finger insulation protein CTCF (the CCCTC-binding factor) that bind to each loop boundary and recruit the cohesin complex; a major problem concerns, however, how the boundary of a given loop finds the other boundary in the very complex nuclear space; 2) in the "extrusion model" (24, 26-28), a loop is generated dynamically by a pair of tethered CTCF units that attach to the chromatin fiber and travel along in opposite directions until they reach converging CTCF-binding sites; in this case, the main difficulty concerns the source of energy required by the process. Here, the problem of chromatin domain formation was approached by taking into account the observation that isochores make up the framework of chromatin domains (15) and by having a closer look at the isochore structure and at the isochore/chromatin domain connection.

The choice of chromosome 21 (long arm) as a model human chromosome in the present work was due to the fact that this chromosome: 1) is the smallest human chromosome, allowing a more expanded graphical presentation of the compositional profile; 2) comprises all isochore families; and 3) is a good representative of human chromosomes, as shown by a comparison with all other chromosomes (see ref. 15), even if the relative contributions of different isochore families are different in different chromosomes (33; in agreement with the correlations between isochores and chromosomal bands; 34). As far as nomenclature is concerned, it should be stressed that, although TADs comprise, by definition, all the topologically associating domains, in the context of this article TADs will indicate the topologically associating domains that do not coincide with LADs, the lamina associated domains.

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ISOCHORES: A CLOSER LOOK

A closer look at the isochores of human chromosome 21 will be presented here starting with a bar plot of GC levels of 100Kb DNA segments (Fig. 1A) which shows: 1) a compositionally flat region (region 2) that corresponds to an isochore from the GC-poorest L1 family; 2) several peaks (**a** to **f** in regions 1 and 3) that correspond to a H1 isochore (**a**) and to L2 isochores (**b** to **f**); and 3) a series of GC-richer, sharper peaks (in regions 4 to 6) that correspond to H1, H2 and H3 isochores, respectively.

Both sliding window (35, 36; Fig. 1B) and non-overlapping window (34, 37; Fig. 1C) programs were developed to merge compositionally similar DNA stretches in order to obtain isochore profiles of chromosomes, to assess the relative amounts of isochores that belong to different compositional families and to investigate their changes in evolution (see, for example, ref. 38 for the case of vertebrates).

A new approach used here to investigate isochores consisted in plotting the GC level of each 100Kb block as a point (Fig. 1D) with the aim to look in more detail at the correspondence of isochores with chromatin domains reported by ref. (15). This approach expectedly showed the compositionally flat region 2 and the H1 and L2 peaks (**a** to **f**), that were already evident in Figs. 1A,B,C. It also led, however, to the discovery that the sequences of isochores from H1 (region 4), H2 (region 5) and H3 (region 6) families were not simply fluctuating within the compositional borders of the corresponding isochore families (see Supplementary Table S1), but consisted, in fact, of a series of GC peaks, that will be discussed in connection with chromatin structure in a following section.

In purely compositional terms (see Fig. 2 for a larger-scale presentation of the data of Fig. 1D), four different situations were found: 1) a very homogeneous L1 isochore (region 2), in which case the overall GC range of the isochore barely reached 4% GC and all points were within the “fixed” GC borders of isochore family L1 (see Supplementary Table S1); 2) a series of single peaks (regions 1 and 3) corresponding to a H1 isochore (**a**), or to several L2 isochores (**b** to **f**), in which case very few points

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

were slightly beyond the "fixed" isochore family borders, but still within the "extended" borders of Supplementary Table S1 (incidentally, the frequency of L2 peaks in chromosome 21 was much higher than in other chromosomes); 3) two series of GC-richer multi-peak isochores that belonged to H1 (region 4) and H2 (region 5) families in which, again, very few points were slightly beyond the "fixed" isochore family borders, but still within the "extended" family borders; and 4) the compositional very heterogeneous GC peaks of H3 isochores (region 6) in which an overall range of 18% GC was reached and points belonging to the H2 and even to the H1 family were present; for this reason, and in contrast with the other isochore families, H3 isochores did not fit the original definition of "fairly homogeneous" DNA segments (39); it should be mentioned, however, that H3 isochores correspond to an exceptional family, which only represents ~3% of the human genome, is almost absent in the mouse genome and completely absent in the genomes of cold-blooded vertebrates (38).

In conclusion, the results of Figs. 1 and 2 indicate the existence of 1) single-peak isochores such as the sharp peaks from the telomeric H3 isochores and the flatter peaks from H1 (**a**) and L2 (**b-f**) isochores; in both cases (except for peak **d**), the sizes of the peaks are between ~500Kb and ~2,000Kb; and 2) multiple-peak isochores such as the H1 and H2 isochores of regions 4 and 5 of Fig. 1, with average peak sizes above 100Kb and isochore sizes in the range of several hundreds Kb. These size estimates are supported by compositional profiles obtained using windows of 50Kb and 25Kb that provided additional details (in spite of the noise due to repeated sequences and other disturbing factors; see ref. 34) and showed compositional ranges that were expectedly larger than those found at 100Kb (see Supplementary Figure S1 and S2).

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ISOCHORES AND LADs

It is well established that LADs 1) may cover ~35% of the human genome; 2) comprise 1,100-1,400 discrete domains demarcated by CTCF sites and CpG islands; 3) have a median size of ~0.5Mb; 4) are scattered over all chromosomes; 5) can be subdivided into cLADs, *i.e.*, "constitutive" LADs present in the four cell types originally tested and fLADs "facultative" LADs, only present in some cell types (in fact only ~15% of the genome is involved in "stable contacts" present in most cells); 6) are characterized, in the case of cLADs, by conserved positions in syntenic regions of human and mouse chromosomes; 7) show a correspondence of cLADs and ciLADs (the "constitutive inter-LADs") with GC-poor and GC-rich isochores, respectively (8-10), a point confirmed by a very recent comparison of isochore maps with LAD maps (15).

As shown in Figs. 3A and 4, the major LAD of chromosome 21 corresponds to a large L1 isochore. The other LADs correspond to the L1 isochores that separate the L2 peaks (to be described in the following section), and to a weak LAD that corresponds to an L2 isochore representing a "valley" between two H1 isochores. In the case of chromosome 20 (Fig. 3B), the largest LAD corresponds to an L2 isochore while several other L2 isochores are "valley" isochores flanked by H1 isochores. Finally, in the case of the very GC-rich chromosome 19, two LADs correspond to two H1 isochores flanking an H2 isochore and two other LADs correspond to L2 isochores flanking an H1 isochore. These results show that LADs correspond not only to L1 isochores that represent ~19% of the genome (incidentally, not too far from the ~15% involved in "stable contacts"; 10), but also to L2 isochores and even (to a very small extent) to H1 isochores in very GC-rich chromosomes.

Looking at the compositional profiles of L2 isochores (namely of isochores comprised between 37% and 41% GC; see Supplementary Table 1 and Fig. 1), it appears that there are isochores from a "low-heterogeneity" L2 sub-family that may be called $L2^-$ (see Fig. 3A,B,C), showing a flat profile and often present as valleys between GC-richer isochores that contribute ~16% of the genome to LADs

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(19% L1+16% L2 equals the 35% of the genome corresponding to LADs; see above), and isochores from a "high-heterogeneity" L2, or $L2^+$, sub-family that are higher in average GC and in the shape of single peaks (see Figs. 1 and 3A). Now, as shown in Fig. 3, $L2^-$ isochores correspond to LADs, whereas $L2^+$ isochores in fact correspond to TADs and interLADs (see the following section). The remaining L2 isochores are within the "extended" borders of H1 isochores, to which they should be assigned. These isochores plus the $L2^+$ isochores represent 20% of the genome and make, together with the 16% of $L2^-$ isochores, the 36% of L2 isochores (see Supplementary Table S1).

As far as the formation mechanism is concerned, LADs should be visualized as extended chromatin structures corresponding to GC-poor isochores that, after emerging from mitosis, do not fold upon themselves (as TADs do; see the following section), but only twist and bend in order to adapt and attach themselves to (and even embed in) the lamina which is reassembled after mitosis (10). This is indicated by the fact that LADs exhibit self-interactions (see Fig. 4) as well as interactions with other LADs from the same chromosomes (15; see, for example, the two LADs bracketed by black lines in Fig. 4), a finding which indicates that these LADs cluster with each other on the lamina.

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ISOCHORES AND TADs

It should be recalled, as a preliminary remark, that the isochores from the five families (L1, L2, H1, H2 and H3) of the human genome are characterized not only by increasing GC levels and different short-sequence frequencies, but also by increasing compositional heterogeneities, increasing levels of CpG, CpG islands and Alu sequences and decreasing levels of LINE sequences (34, 40-45). Moreover, at the chromatin level, GC increase is correlated with higher bendability (46,47), higher nuclease accessibility (48,49), lower nucleosome density (50) and lower supercoiling (51,52).

The connection of the isochores of chromosome 21 as seen in Figs. 1D and 2 with chromatin loops can be described as follows (see Fig. 4): 1) regions 1 and 3 show a series of H1 (**a**) and L2 (**b** to **f**) isochores in which latter case at least some of their single peaks (**a, b** and **f**) may correspond to single self-interactions (see Fig. 4); 2) region 2 is the GC-poorest L1 isochore already described in the preceding section; 3) the multi-peak H1 isochores of region 4 correspond to a large inter-LAD region and to several self-interactions and are followed by a multi-peak H2 isochore (on the right-hand side of region 5) that corresponds, in contrast, to a single self-interaction (Fig. 4); 4) finally, at the telomeric end, there is a series of H3 isochores (red points comprised between two red arrows) that correspond to a series of self-interactions comprised between the two red lines on the heat map of Fig. 4; this indicates a correspondence of the six H3 isochore peaks with at least three chromatin loops. The two classes of isochores, single-peak and multi-peak, precisely correspond to two classes of TADs, single-loop and multi-loop TADs, respectively (Fig. 4).

Now, one should consider that the local GC gradient within each GC-rich loop is accompanied by the factors already mentioned, namely increased bendability, increased nuclease accessibility, decreased supercoiling and decreased nucleosome density. These factors may constrain the corresponding chromatin to form loops, as shown by the finding that the isochore peaks from chromosome 21 (the isolated L2 isochores, the following H2 isochore and the H3 isochores of the

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

telomeric region) match chromatin loops as well as interLADs (as shown in Fig. 4; see also Fig. 3A and the preceding section), ruling out the possibility that the self-interactions under consideration are due to LADs.

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DISCUSSION

Isochore structure. Plots of GC levels of 100Kb sequences of human chromosome 21, such as those of Figs. 1D and 2, led to the discovery that 1) isochores from the GC-poor L1 family and $L2^-$ sub-family are characterized by compositionally flat and "low-heterogeneity" profiles, respectively; 2) isochores from the GC-richer H1 and H2 families appear as sets of GC peaks; and 3) isochores from the $L2^+$ sub-family and from the GC-richest H3 family are at least largely in the form of individual compositional peaks. These are crucial findings because the picture which emerges is that isochores not only correspond to LADs and TADs, as shown by ref. 15, but they are also characterized by different fine structures that are relevant in terms of the mechanisms of formation of chromatin domains, a point discussed in the following sub-sections.

Isochores and LADs. In the case of LADs, the very low compositional heterogeneity of their DNA backbone, formed by L1 and $L2^-$ isochores, their richness in oligo-A stretches and the corresponding nucleosome depletion, favor their adhesion to the nuclear lamina (which is reassembled at the exit of mitosis), as well as contacts within individual LADs. This implies a bending and twisting of individual LADs in their lamina contacts (and even an embedding in the lamina; see ref. 10). Contacts also take place among LADs from the same chromosomes (interactions decreasing with distance; see ref. 15 and also Fig. 4; in this case, TAD folding helps approaching different LADs from the same chromosomes). It is possible that L1 isochores correspond to "stable (or consistent) contacts" and $L2^-$ isochores to "variable contacts" (as defined in ref. 10). Likewise, LINE sequences (very frequent in L1 isochores; 40) possibly play a role in the L1/lamina contact.

Isochores and TADs. The mechanism of formation of TADs is completely different from that of LADs, because it involves the folding of chromatin domains. On the basis of the results of ref. 15, the formation of chromatin loops may be due to the folding upon themselves of chromatin domains that

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

correspond to isochores. Folding can, however, take place in many ways, as indicated by the models mentioned in the Introduction.

A new model for chromatin folding emerges from the present analysis of isochores. In fact, TADs may be single-loop domains or (more frequently) multi-loop domains, both of these domains corresponding to individual isochores and to TADs. Multi-loop domains can be resolved, however, into contact domains (13), that, in fact, correspond to individual peaks of multi-peak isochores.

The clear correlations of isochores with TADs mean that TADs, whether consisting of single or multiple chromatin loops, are not only encoded but also molded by the DNA sequences of isochores. Indeed, the properties that accompany the GC-increase of isochore peaks constrain the corresponding chromatin to fold in a specific way, in fact with the corresponding CTCF-binding sites matching each other and associating with CTCF proteins (if they dissociated from chromatin during mitosis, which is not sure; see ref. 53). It should be stressed that the term loop is just a conventional definition for the basic structure of TADs, since the folding of chromatin involves supercoiled structures that are increasingly underwound with increasing GC levels (51,52).

A scheme of the models for the formation of LADs and TADs is presented in Fig. 5, which stresses the keypoint of this work, namely the central role played by the compositional properties of DNA sequences in the formation of chromatin domains that range from the highly homogeneous sequences of L1 and $L2^-$ isochores, that correspond to LADs, to the increasingly more heterogeneous sequences of $L2^+$, H1, H2 and H3 isochores that correspond to single- and multiple-loop TADs. As such, the folding model presented here rests on a completely different base, namely isochore sequences, compared to other models that essentially rely on the architectural proteins like CTCF and cohesin (see Introduction). Dixon et al. (17) postulated, however, that the “insulation” observed at TAD boundaries may result from stiffness of the chromatin fiber caused by functional elements (CTCF binding sites, highly active transcription starting sites etc) associated with increased nucleosome

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

density. “Attractive forces” (not better specified, but possibly having to do with supercoiling) within the domain can then confer specific local chromatin interaction, yielding a joint “insulation-attraction” model of TAD formation. Interestingly, this “speculative hypothesis” is compatible with the isochore folding model proposed here.

A general issue. A question concerns whether the present findings on chromosome 21 can be generalized to all human chromosomes. The answer to this question is yes since a comparison of compositional profiles, heatmaps and LAD maps (15) with the “point profiles” (see Figs. 1D and 2) of all human chromosomes (P. Cozzi, L. Milanesi, G. Bernardi, paper in preparation) allows an extension of the results drawn here on human chromosome 21 to all human (and mouse) chromosomes. Moreover: 1) the isochores from the L1 family and the $L2^-$ sub-family correspond to LADs in all human chromosomes; 2) $L2^+$ peaks emerging from an L1 background and also corresponding to inter-LADs are also found, although more rarely, in other human chromosomes; 3) the spikes of the compositional profile of H1 and H2 isochores, as present in the overlapping window plot of Fig. 1B reflect, upon close inspection, the peaks of Fig. 1D; now, these spikes are regularly present in H1 and H2 isochores from all human and mouse chromosomes indicating that these isochores generally are multi-peak isochores; 4) H3 isochores are generally flanked by H2 and H1 isochores as shown by 100Kb “point profiles”.

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

CONCLUSIONS

The main conclusion of this investigation concerns the discovery that isochores not only encode LADs and TADs with their own sequences, but also mold them with the properties (bendability, accessibility etc) that they impose on the corresponding chromatin domains. This general conclusion is supported by a number of observations: 1) LADs and TADs are evolutionarily stable structures, a result which is explained by the evolutionary stability of isochore patterns in the corresponding genomes (38,54); 2) alterations of the architecture of chromatin domains, known to lead to diseases (see the review papers cited in the Introduction) are due to changes in their isochore framework; incidentally, this was predicted by previous investigations (4, 55,56) on "genomic diseases", namely diseases in which sequence alterations do not affect coding or classical regulatory sequences, but only other sequences wrongly considered as "junk DNA" or "non-coding DNA"; 3) TADs show a visible similarity between sperm cells and fibroblasts of mouse, in spite of the replacement of histones with protamines (57); 4) in *Drosophila*, TADs are conserved between polytene bands and interphase chromatin (58); 5) the "mitotic memory", namely the rapid and precise re-establishment of the original interphase chromatin domains at the exit from mitosis, is best explained by the fact that the basic information required for such quick re-establishment is present in the sequences of isochores; needless to say, CTCF (possibly conserved during mitosis; 53) and cohesin also have an important role in the process; 6) TADs are basic units of replication timing (59) and isochores are replication units characterized by all early or all late replicons in GC-rich and GC-poor isochores (60), respectively.

In more general terms, this investigation has shown that, by encoding and molding LADs and TADs, DNA plays a newly discovered role which consists in encoding and shaping the basic chromatin architecture and influencing, as a consequence, its functional properties. It should be stressed that this encoding concerns the basic, evolutionarily stable chromatin domains. Indeed, it is well known that, epigenetic modifications and environmental factors may cause changes in chromatin architecture and

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

that, while self-associating domains are stable, chromatin interactions within and between domains change during differentiation (61,62).

The encoding of chromatin domains by genomic units, the isochores, deserves the name of "genomic code". This definition was originally coined (63,64; see also ref. 4) for the compositional correlations that hold among genome sequences (such as those between coding and contiguous non-coding sequences) and that concern the isochore structure of the genome. The original definition was later extended to the correlations between the compositional properties of isochores and all the structural/functional properties of the genome that could be tested (see ref. 4 and Table S2 of ref. 15). Here it is proposed that the definition of "genomic code" be applied to the encoding of chromatin domains by isochores (which, incidentally, explains the correlations of isochore composition with structural/functional properties of the genome).

Acknowledgements:
The author thanks Paolo Ascenzi for hospitality, Giacomo Bernardi, Oliver Clay, and, especially, Kamel Jabbari for critical reading, comments and discussions as well as Caterina Nuvoli for excellent technical help. This research was supported by the Kimura Prize for Molecular Evolution and Evolutionary Genomics, conferred to the author (Tokyo, June 2016).

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References:

1. Macaya G, Thiery JP, Bernardi G (1976) An approach to the organization of eukaryotic genomes at a macromolecular level. *J Mol Biol* 108(1):237–254.

2. Filipski J, Thiery JP, Bernardi G (1973) An analysis of the bovine genome by Cs2SO4-Ag+ density gradient centrifugation. *J Mol Biol* 80: 177-197.

3. Corneo G, Ginelli E, Soave C, Bernardi G (1968) Isolation and characterization of mouse and guinea pig satellite deoxyribonucleic acids. *Biochemistry* 7: 4373-4379.

4. Bernardi G (2004) *Structural and evolutionary genomics: natural selection in genome evolution.* (Elsevier, Amsterdam). This book is freely available at www.giorgiobernardi.eu

5. Bernardi G (2015a) Chromosome Architecture and Genome Organization. *PLoS One* 10(11):e0143739.

6. Bernardi G (2015b) Genome Organization and Chromosome Architecture. *Cold Spring Harb Symp Quant Biol*:27318.

7. Bernardi G, et al. (1985) The mosaic genome of warm-blooded vertebrates. *Science* 228(4702):953–958.

8. Guelen L, et al. (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* 453(7197):948–51.

9. Meuleman W, et al. (2013) Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. *Genome Res* 23(2):270–280.

10. Kind J, et al. (2015) Genome-wide Maps of Nuclear Lamina Interactions in Single Human Cells. *Cell* 163(1):134–147.

11. Dixon JR, et al. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485(7398):376–380.

12. Nora EP, et al. (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485:381–385.

13. Rao SSP, et al. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159(7):1665–1680.

14. Vietri Rudan M, et al. (2015) Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Rep* 10(8):1297–1309.

15. Jabbari K, Bernardi G (2016) An isochore framework underlies chromatin architecture. *Plos One* (in press)

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

16. Bonev B, Cavalli G (2016) Organization and function of the 3D genome. *Nat Rev Genet* 17(11):661–678.

17. Dixon JR, Gorkin DU, Ren B (2016) Chromatin Domains: The Unit of Chromosome Organization. *Mol Cell* 62(5):668–680.

18. Ghirlando R, Felsenfeld G (2016) CTCF: Making the right connections. *Genes Dev* 30(8):881–891.

19. Gonzalez-Sandoval A, Gasser SM (2016) On TADs and LADs: Spatial Control Over Gene Expression. *Trends Genet*. 32 (8):485-495.

20. Merkenschlager M, Nora EP (2016) CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation. *Annu Rev Genomics Hum Genet* 1–27.

21. Pueschel R, et al. (2016) From single genes to entire genomes: the search for a function of nuclear organization. *Development* 143(6):910–23.

22. Solovei I, Thanisch K, Feodorova Y (2016) How to rule the nucleus: divide et impera. *Curr Opin Cell Biol* 40:47–59.

23. Lieberman-Aiden E, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950):289–293.

24. Alipour E, Marko JF (2012) Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res* 40(22):11202–11212.

25. Barbieri M, et al. (2012) Complexity of chromatin folding is captured by the strings and binders switch model. *Proc Natl Acad Sci U S A* 109(40):16173–16178.

26. Fudenberg G, et al. (2016) Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep* 15(9):2038–2049.

27. Sanborn AL, et al. (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci USA* 112(47):E6456–E6465.

28. Dekker J, Mirny L (2016) The 3D Genome as Moderator of Chromosomal Communication. *Cell* 164(6):1110–1121.

29. Di Pierro M, Zhang B, Aiden EL, Wolynes PG, Onuchic JN (2016) Transferable model for chromosome architecture. *Proc Natl Acad Sci*:201613607.

30. Rowley MJ, Corces VG (2016) The three-dimensional genome: Principles and roles of long-distance interactions. *Curr Opin Cell Biol* 40:8–14.

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

31. Ulianov S V, et al. (2016) Active chromatin and transcription play a key role in chromosome partitioning into topologically assoc iating domains. *Genome Res* 26(1):70–84.

32. Vietri Rudan M, Hadjur S (2015) Genetic Tailors: CTCF and Cohesin Shape the Genome During Evolution. *Trends Genet* 31(11):651–660.

33. Costantini M, et al. (2007) Human chromosomal bands: Nested structure, high-definition map and molecular basis. *Chromosoma* 116(1):29–40.

34. Costantini M, Clay O, Auletta F, Bernardi G (2006) An isochore map of human chromosomes. *Genome Res* 16(4):536–541.

35. Pavlícek A, Paces J, Clay O, Bernardi G (2002) A compact view of isochores in the draft human genome sequence. *FEBS Lett* 511(1–3):165–169.

36. Pačes J, et al. (2004) Representing GC variation along eukaryotic chromosomes. *Gene* 333(SUPPL.):135–141.

37. Cozzi P, Milanesi L, Bernardi G (2015) Segmenting the human genome into isochores. *Evol Bioinforma* 11:253–261.

38. Costantini M, Cammarano R, Bernardi G (2009) The evolution of isochore patterns in vertebrate genomes. *BMC Genomics* 10(i):146.

39. Cuny G, Soriano P, Macaya G, Bernardi G (1981) The Major Components of the Mouse and Human Genomes. 1. Preparation, Basic Properties and Compositional Heterogeneity. *Eur J Biochem* 115(2):227–233.

40. Meunier-Rotival M, Soriano P, Cuny G, Strauss F, Bernardi G (1982) Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA. *Proc Natl Acad Sci U S A* 79(2):355–359.

41. Soriano P, Meunier-Rotival M, Bernardi G (1983) The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. *Proc Natl Acad Sci U S A* 80(7):1816–20.

42. Jabbari K, Bernardi G (1998) CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. *Gene* 224(1–2):123–128.

43. Costantini M, Bernardi G (2008) The short-sequence designs of isochores from the human genome. *Proc Natl Acad Sci* 105(37):13971–13976.

44. Varriale A, Bernardi G (2010) Distribution of DNA methylation, CpGs, and CpG islands in human isochores. *Genomics* 95(1):25–28.

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

45. Arhondakis S, Auletta F, Bernardi G (2011) Isochores and the regulation of gene expression in the human genome. *Genome Biol Evol* 3:1080–1089.

46. Vinogradov AE (2001) Bendable genes of warm-blooded vertebrates. *Mol Biol Evol* 18:2195–2200.

47. Vinogradov AE (2003) DNA helix: The importance of being GC-rich. *Nucleic Acids Res* 31(7):1838–1844.

48. Di Filippo M, Bernardi G (2008) Mapping DNase-I hypersensitive sites on human isochores. 419(1-2):62–65.

49. Di Filippo M, Bernardi G (2009) The early apoptotic DNA fragmentation targets a small number of specific open chromatin regions. *PLoS One* 4(4):1-7.

50. Fenouil R, et al. (2012) CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Res* 22(12):2399–2408.

51. Naughton C, et al. (2013) Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nat Struct Mol Biol* 20(3):387–395.

52. Benedetti F, Dorier J, Burnier Y, Stasiak A (2014) Models that include supercoiling of topological domains reproduce several known features of interphase chromosomes. *Nucleic Acids Res* 42(5):2848–55.

53. Burke LJ, et al. (2005) CTCF binding and higher order chromatin structure of the H19 locus are maintained in mitotic chromatin. *EMBO J* 24(18):3291–3300.

54. Pavlícek A, Clay O, Jabbari K, Paces J, Bernardi G (2002) Isochore conservation between MHC regions on human chromosome 6 and mouse chromosome 17. *FEBS Lett* 511(1–3):175–177.

55. Bernardi G (2007) The neoselectionist theory of genome evolution. *Proc Natl Acad Sci U S A* 104(20):8385–90.

56. Al Mahmud A, Amore G, Bernardi G (2008) Compositional genome contexts affect gene expression control in sea urchin embryo. *PLoS One* 3(12): 1-8.

57. Battulin N, et al. (2015) Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach. *Genome Biol* 16(1):77.

58. Eagen KP, Hartl TA, Kornberg Correspondence RD, Kornberg RD (2015) Stable Chromosome Condensation Revealed by Chromosome Conformation Capture. *Cell* 163:934–946.

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

59. Pope BD, et al. (2014) Topologically associating domains are stable units of replication-timing regulation. *Nature* 515(7527):402–405.

60. Costantini M, Bernardi G (2008) Replication timing, chromosomal bands, and isochores. *Proc Natl Acad Sci U S A* 105(9):3433–3437.

61. Dixon JR, et al. (2015) Chromatin architecture reorganization during stem cell differentiation. *Nature* 518(7539):331–336.

62. Ji X, et al. (2016) 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. *Cell Stem Cell* 18(2):262–275.

63. Bernardi G (1990) Le genome des vertebrés: organisation, fonction et evolution. *Biofutur* 94 : 43-46.

64. Bernardi G, Bernardi G (1991). Compositional properties of nuclear genes from cold-blooded vertebrates. *J Mol Evol* 33: 57-67.

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure Legends

Figure 1. A: Compositional profile of human chromosome 21 (release hg38) as seen through non-overlapping 100-Kb windows (from Cozzi et al., 2015). The choice of a window size of 100Kb to define the compositional profiles of human chromosomes was due to the fact that 100Kb is a plateau value under which the composition of DNA segments from different isochore families show an increasing variance with decreasing size due to several factors, such as the distribution of interspersed repeated sequences (34). DNA stretches from isochore families L1 to H3 are represented in different colors, deep blue, light blue, yellow, orange, red, respectively. The left-side ordinate values are the minima GC values between the isochore families listed on the right side (see Supplementary Table S1); **a,b,c,d,e,f** correspond to an H1 peak (**a**) and to several $L2^+$ peaks (**b** to **f**; the thin yellow bars in peaks **c**, **e** and **f** correspond to single 100Kb blocks that should be assigned to the H1 family if using "fixed" isochore borders (as shown in the Figure), but to L2 isochores if "extended" borders (see Supplementary Table 1) are used. Six regions, 1 to 6, are identified in the compositional profile by vertical lines and numbers (top and bottom of the Figure).

B. Isochore profile of human chromosome 21 using the matched b37 assembly of Rao et al. (2014) and a sliding 100-Kb window approach (Paces et al., 2004) with "fixed" isochore borders. This procedure expectedly flattens the peaks of Fig. 1A. The color convention is as in Fig. 1A. (From Jabbari and Bernardi, 2016). This profile is slightly more extended on the centromeric side than that of Fig. 1A.

C. Isochore profile of human chromosome 21 (release hg38) using a non-overlapping 100Kb window and the isoPlotter program (Cozzi et al., 2015).

D. GC levels of 100Kb windows of human chromosome 21. This figure shows that individual isochores from the $L2^+$ to H3 families are in the form of peaks. The GC gradients of the peaks do not appear in a clear way in the standard presentation of compositional profiles of chromosomes (Fig. 1A),

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

except for the broad, isolated H1 (**a**) and $L2^+$ peaks **b** to **f**. Upon close inspection, however, the peaks of the H1, H2 and H3 isochores of this Figure show a correspondence with the small peaks of Fig. 1B. Blue and red arrows indicate the regions 2 and 6. Black (as well as blue and red) arrows separate regions 3,4 and 5; horizontal lines correspond to minima GC values between isochore families (see Figs. 1A,B and Supplementary Table 1).

Figure 2A. GC levels of 100Kb windows from L1 and H3 regions of human chromosome 21 comprised between the blue and red arrows (sequences 2 and 6) of Fig. 1D, respectively, are displayed at a higher magnification. The figure accounts for the fact that the approach of Fig. 1A did not show as clearly as Fig. 1D the existence of the large compositional peaks located at the telomere of chromosome 21. Horizontal lines correspond to "fixed" minima GC values between isochore families (see Supplementary Table 1). The abscissa scale is in 100Kb units.

B.C. GC levels of 100Kb windows from the regions 1,3,4 and 5 of human chromosome 21. See the legend of Fig. 2A for other indications.

Figure 3. A. The isochore profile of human chromosome 21 (from ref. 15) is compared with the corresponding (inverted) LAD profile (from ref. 10) to show the correspondence of LADs 1) with L1 isochores (blue lines; two broken blue lines bracket the largest L1 isochore, in which case the LAD seems to be "compressed" relative to the DNA sequence); 2) with one L2 "valley" isochore (blue line, last on the right side); and 3) with four high heterogeneity L2+ "peak" isochores (red lines) that correspond to interLADs, or TADs (see Text). The multicolored bar on the right is the color code for isochore families.

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

B. The isochore profile of human chromosome 20 (see also legend of Fig. 3A) is compared with the corresponding (inverted) LAD profile (from ref. 10) to show the correspondence of LADs with a large low-heterogeneity L2 ($L2^-$) isochore (bracketed by broken blue lines; on the left of the panel) and with several L2 "valley" isochores (blue lines), as well as the correspondence of interLADs with GC-rich isochores (red lines).

C. The isochore profile of human chromosome 19 (see legend of Fig. 3A) is compared with the (inverted) LAD profile (from ref. 10) to show that two LADs correspond to H1 isochores (blue lines) flanking a H2 isochore, an interLAD (red line); two other LADs (on the left) correspond to L2 "valley" isochores (blue lines).

Figure 4. The compositional profile of human chromosome 21 (Fig. 1D), is compared with the corresponding LAD profile (an inverted scale is shown on the top right) and the heat map of chromatin interactions; from ref. 15). Two double vertical black lines corresponding to LADs (characterized by self-interactions as well as by intrachromosomal interactions) separate a multi-peak H1 isochore block corresponding to multiple interactions and several $L2^+$ isochores that correspond to inter-LADs. A telomeric block of H3 isochores (defined by red points, arrows and lines) corresponds to several (at least three) self-interactions on the heat map. The contiguous multi-peak H2 isochores corresponds to a single self-interaction.

Figure 5. Formation of LADs and TADs. Three chromatin fibers are taken into consideration: A. a GC-poor chromatin fiber corresponding to an L1 isochore; blue bar) attaches to the lamina, forming a LAD; the wavy profile indicates the physical adaptation by bending and twisting to (and embedding in) the lamina as well as the self-interactions. B. a GC-rich chromatin fiber corresponding to a H3 isochore folds upon itself forms a single-loop TAD (yellow to red color indicate an increasing GC level), and a

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

single contact domain, bounded by CTCF-binding sites (green). C. A GC-rich chromatin fiber corresponding to a H1 (or H2) isochore characterized by two GC peaks folds to form a TAD which comprises two loops (and two contact domains).

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

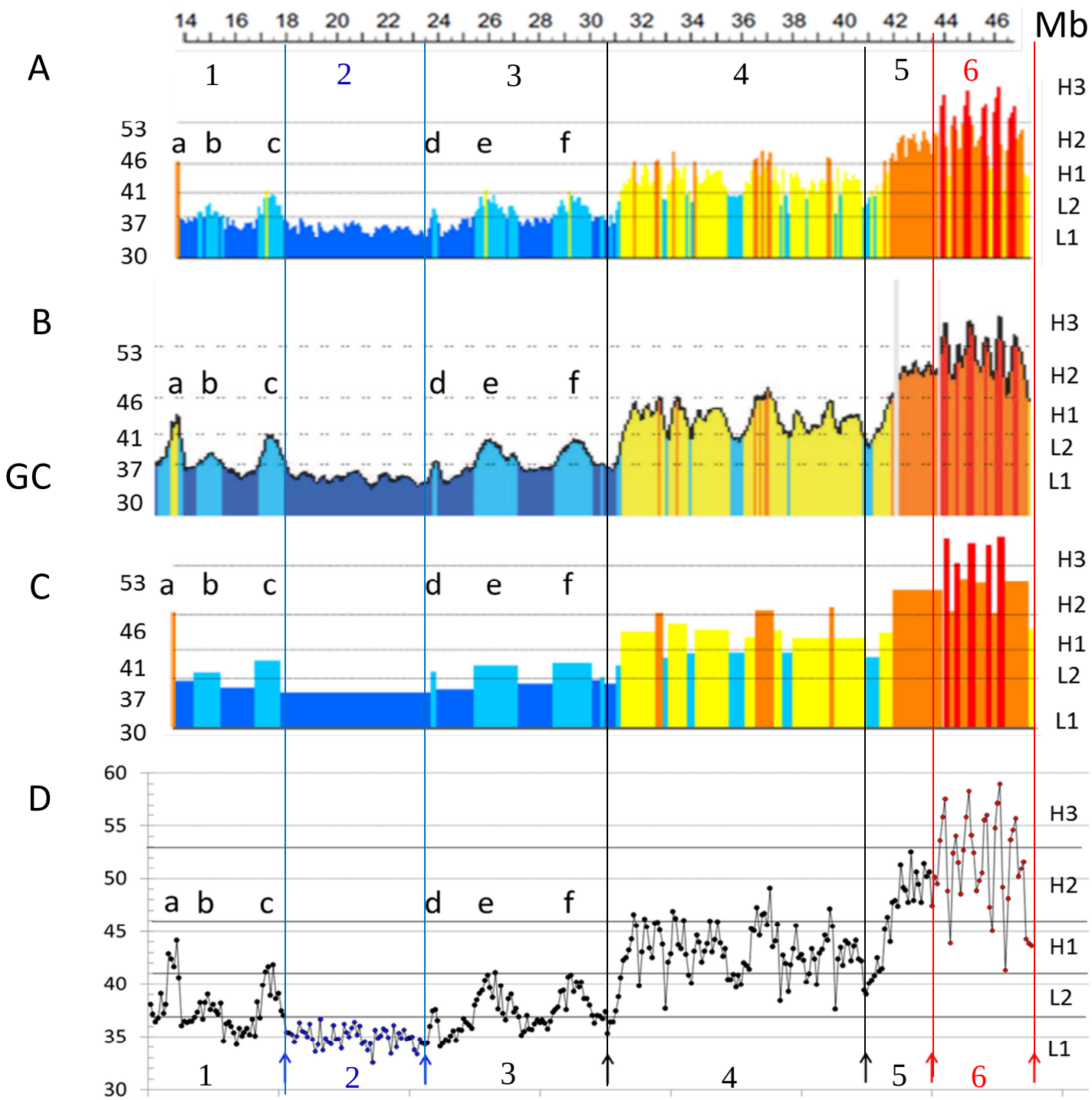
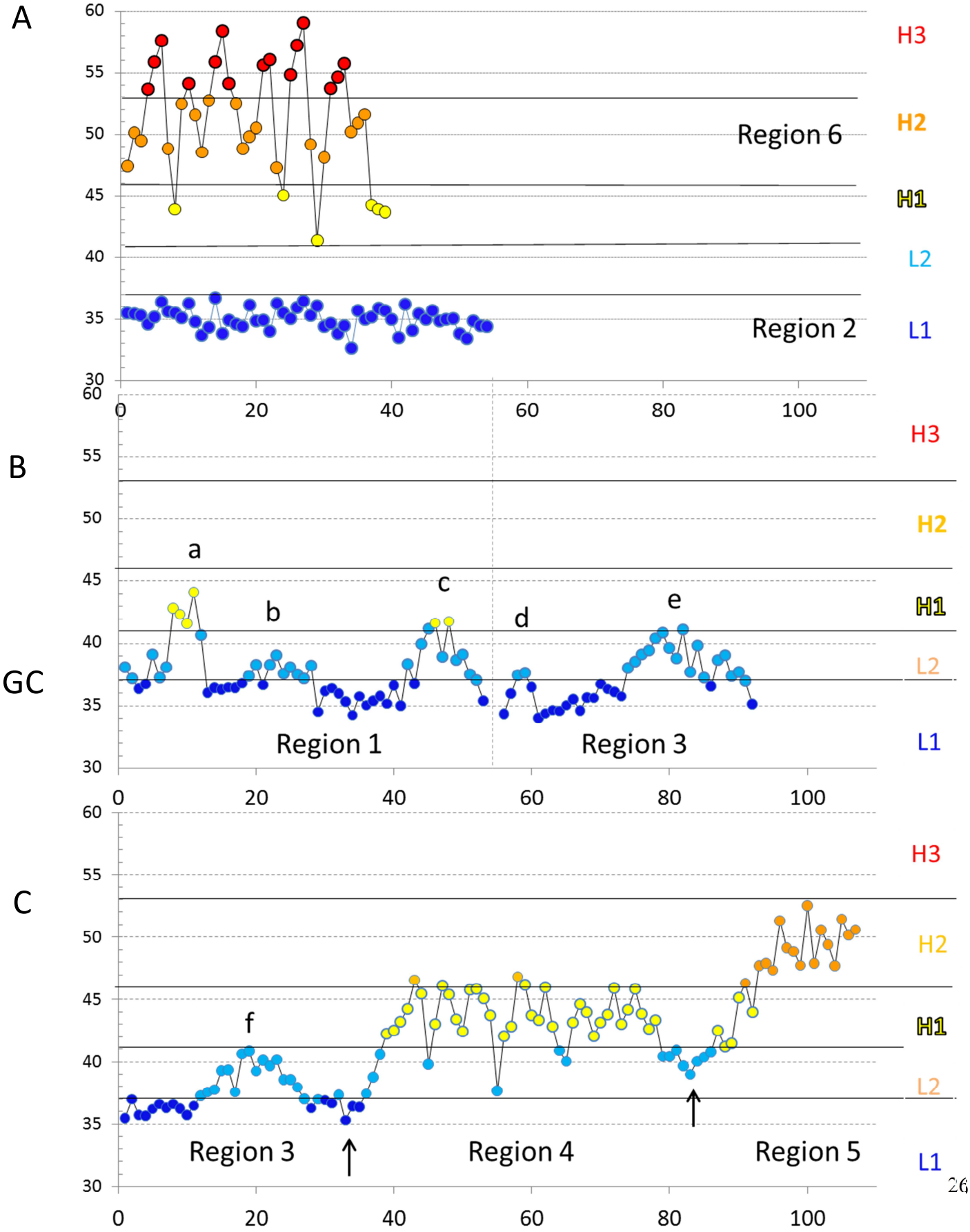


Fig. 1

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

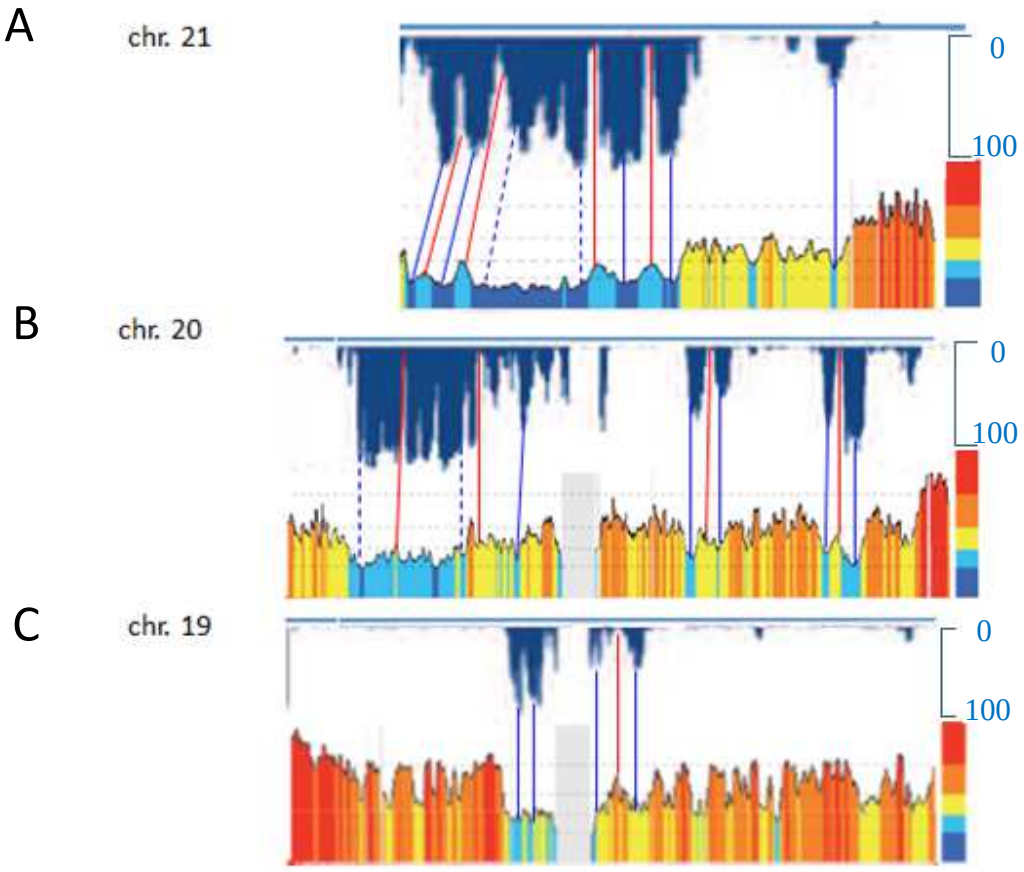


Fig. 3

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

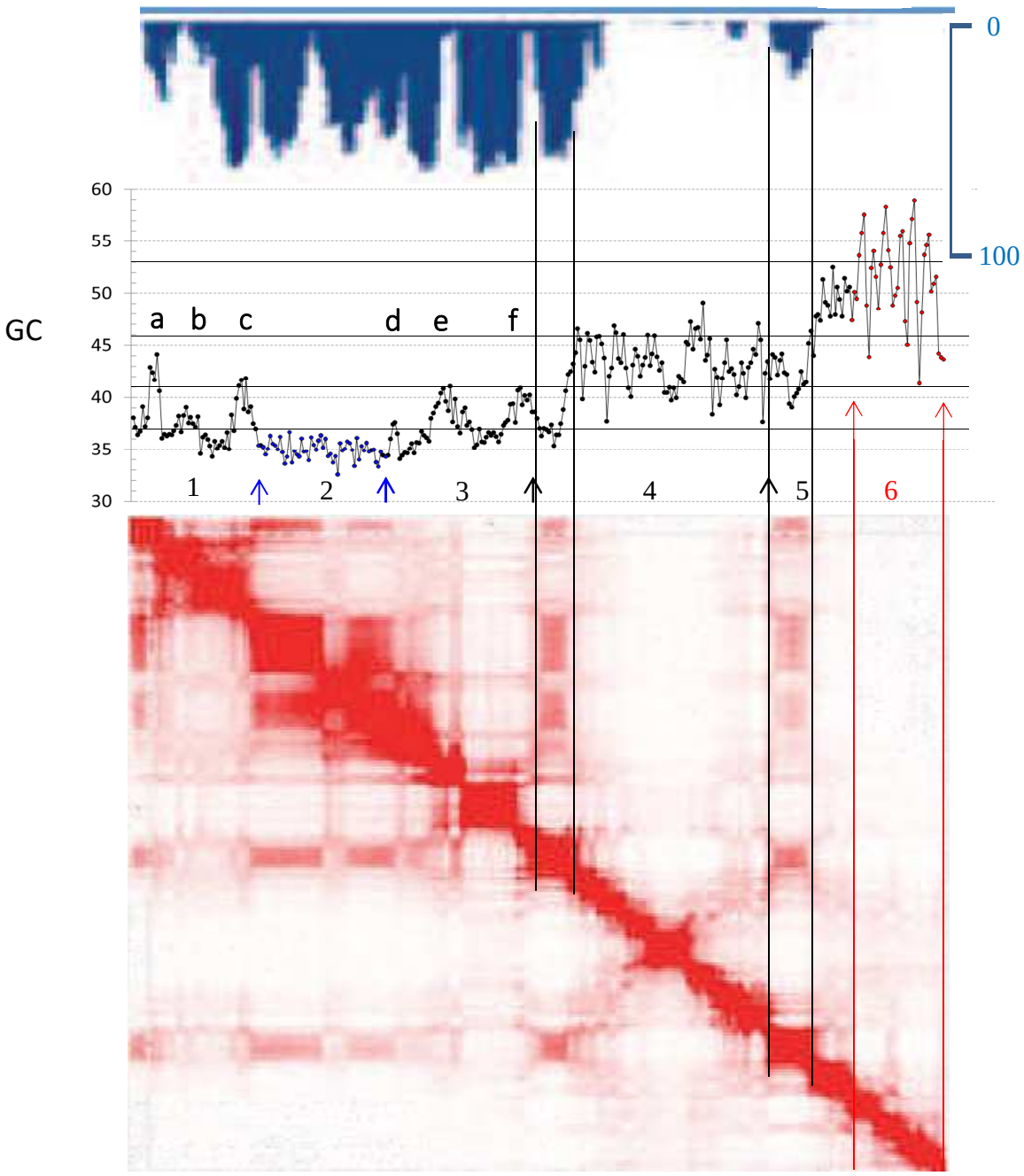


Fig. 4

bioRxiv preprint doi: https://doi.org/10.1101/096487; this version posted December 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

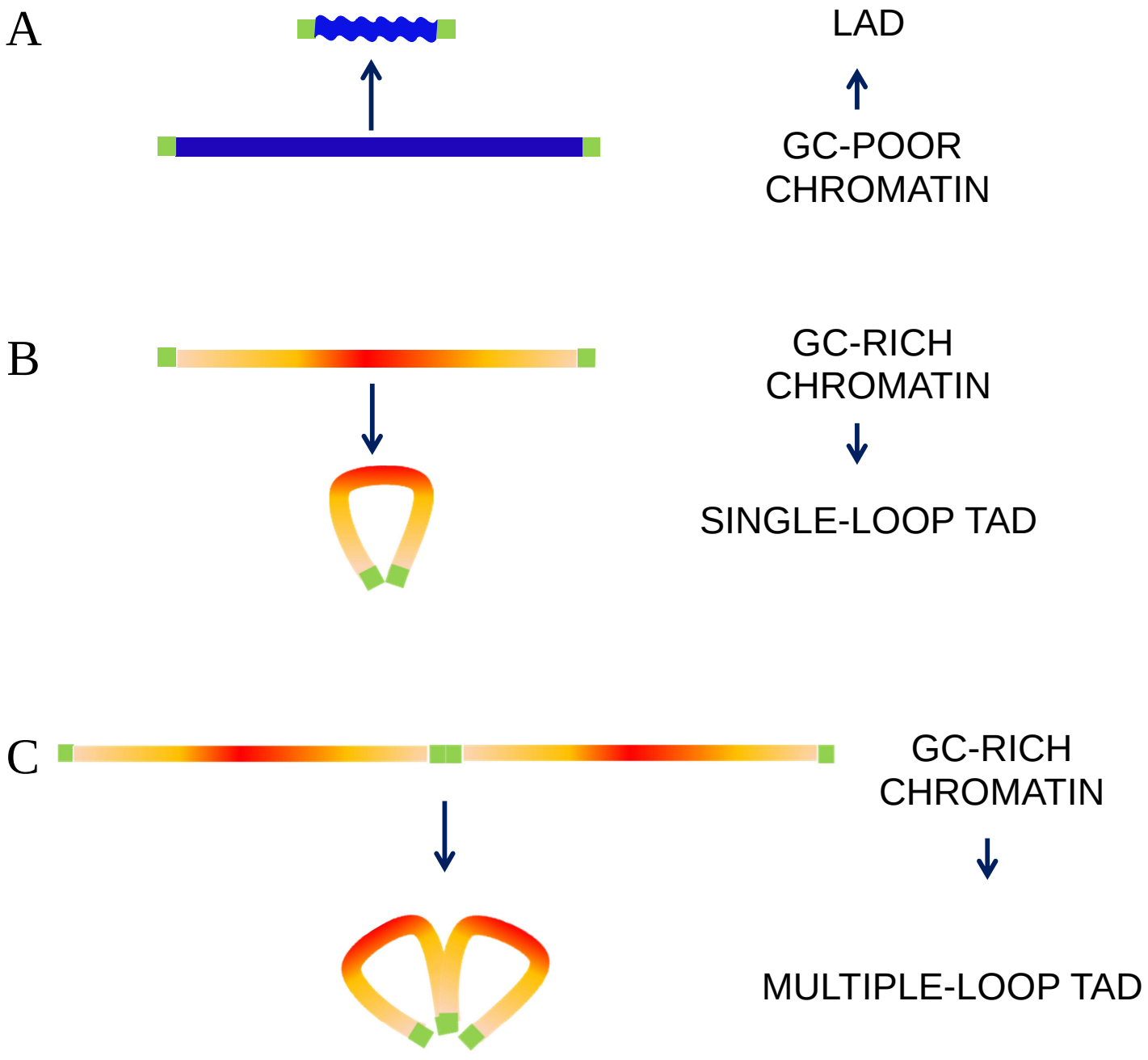


Fig. 5