Electronic Supplementary Material (ESI) for Molecular BioSystems.

Supplementary Text T1

In Hi-C [1], chromatin is cross linked and then digested into fragments with a restriction enzyme of choice. The fragments are ligated back and marked for detection. Consequently, the cutting preference of the enzyme determines the fragments population. The restriction enzyme can be thought of as a sampler with a sampling rate that depends on the sequence composition (the distribution of recognition sites). Thus, GC composition is likely to play some role in Hi-C coverage. Indeed, a recent analysis [2] showed that GC as well as the fragment length and the uniqueness of the sequence (mappability) are the main sources for bias in Hi-C data. However, this analysis addressed GC in close proximity to restriction sites. Here, we have hypothesized that GC heterogeneity at different scales leads to a multi-scaled bias in Hi-C coverage.

In order to test our hypothesis we have analyzed Hi-C coverage and GC, both at the level of 1 Megabase pairs (Mb) windows and at the level of isochores – long homogenous GC domains [3-5], providing enough data for a comprehensive statistical analysis. We have first analysed HindIII and NcoI Hi-C replicates (see Data Preparation section below), at 1Mb windows, generated for GM06990 (lymphoblastoids) cell line [1]. For this analysis we have calculated the following properties (depicted in Fig. T1.1) for 1Mb regions in the human genome (reference genome hg18):

1. Contained Restriction Fragments (CRF). The number of restriction fragments that are contained within the given region.

2. Total Interaction Frequency (TIF). The total number of contacts observed for the region within itself (*self* TIF), with other regions in the same chromosome (*cis* TIF) and with other regions in other chromosomes (*trans* TIF). The *overall* TIF refers to the sum of the 3 TIF properties.

3. GC Percentage (GCP). The percentage of GC in a given region.

Since the coverage of Hi-C is limited by the mappability of the fragments and by the ability to determine their sequence, some regions are completely unrepresented in all replicates (i.e. regions with zero overall TIF). These regions were discarded here from downstream analysis (see Data Preparation section below).

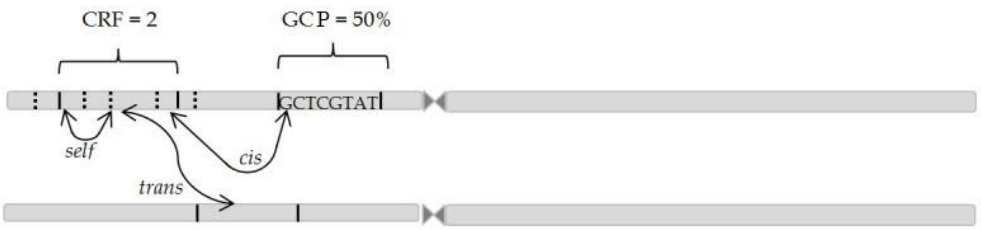

Fig T1.1. Properties calculated for 1 Mb genomic regions. A region (bounded with bold lines) may contain zero or more restriction sites (dashed lines). The CRF is the number of restriction fragments that are fully contained within the region. Based on its genomic content, the GCP is the ratio between the GC content and the total content of the region. The TIF is the sum of total interaction frequencies observed for a given region, within itself (*self*), with other regions on the same chromosomes (*cis*) and with other regions in other chromosomes (*trans*). CRF, Contained Restriction Fragments; GCP, Guanine-Cytosine Percentage; TIF, Total Interaction Frequency.

We have first tested the correlation between coverage (TIF) and fragment occupancy (CRF) and found that CRF of 1Mb regions is strongly positively (Spearman) correlated with TIF for both enzymes, in *cis* and in *trans* (*cis* - HindIII: $\rho = 0.72$, NcoI: $\rho = 0.83$; trans - HindIII: $\rho = 0.81$, NcoI: $\rho = 0.88$; $p < 2.2e\text{-}16$ in both cases), and across chromosomes for overall TIF (average $\rho > 0.83$, for both enzymes, HindIII: maximal $p < 3.8e\text{-}07$, NcoI: maximal $p < 3.3e\text{-}14$, across all chromosomes). We have used here the Spearman correlation as the relationship are not expected to be strictly linear; however, the Pearson correlation also provided evidences for a strong correlation (*cis* - HindIII: $r = 0.73$, NcoI: $r = 0.77$; *trans* - HindIII: $r = 0.51$, NcoI: $r = 0.38$; $p < 2.2e\text{-}16$ in all cases). We have next tested whether the cutting preference of the enzymes (represented by the CRF) is correlated with the GC percentage (GCP). We found that the GCP is positively correlated (Spearman) with the CRF for NcoI ($\rho = 0.60$, $p < 2.2e\text{-}16$) and negatively correlated with the CRF for HindIII ($\rho = -0.55$, $p < 2.2e\text{-}16$), consistent with the GCP of their recognition sites (67% versus 33% for NcoI and HindIII correspondingly). These results suggest that at the 1Mb level, Hi-C coverage is dependent on the cutting preference of the restriction enzyme, which is dependent on the GC content. In order to test whether this trend is multi-scaled we have next analysed the data at the level of isochores.

Isochores were first identified through DNA ultracentrifugation [3-5] and were defined as "very long stretches (>300kb) of DNA that are homogeneous in base composition" [5]. These regions were shown to be divided into families of high and low GC content suggesting a model of a mosaic genome [5]. Isochores were later shown to correlate with various biological properties such as recombination rates [6], replication timing [7], localization in interphase nucleus [8] and gene density [9-11] where GC rich isochores are replicated earlier, predominantly occupy the nucleus interior and present higher gene density and higher recombination rates. As the genome draft became available, isochores definition was extended to long

regions (not necessarily >300kb) that are fairly homogenous in GC, identified through genome segmentation algorithms, and shown to be consistent with the properties of traditional isochores [12].

For our analysis, we have followed a GC based segmentation (applied to the human genome, reference genome hg18), generated with IsoFinder [12], a popular algorithm for isochore detection. We have compared Hi-C coverage of GC rich and GC poor isochores, across replicates and different cell types, using 5 publicly available Hi-C data sets of 4 human cell lines (Table T1.1): lymphoblastoids (GM06990) - HindIII and NcoI replicates (used for the 1Mb analysis), fibroblasts (IMR90), embryonic stem cells (ES) and myelogenous leukemia cells (K562).

We have assigned each of the 39,383 regions in the IsoFinder segmentation with a class: low (L) or high (H) according to its mean GC content (as provided by the segmentation algorithm). The threshold value (44%) was set according to the mean GC content of isochores in human [14] with close agreement with the threshold separating high and low isochore families (42%) [5]. The relative proportions of classes were similar across chromosomes and overall, with 59% (on average) L regions. L regions were longer on average than H regions with a mean length of 93.7Kb and 41.6Kb correspondingly. Since length differences may act as a biasing factor, we have used the ratio between the total interaction frequency for each chromosome pair (in *cis* and in *trans*) and the region's length, denoted here as the Total Interaction Density (TID).

Table T1.1. Hi-C data sets used in the analysis of GC and Hi-C coverage at the level of isochores. 5 Hi-C data sets of 4 human cell lines were considered: lymphoblastoids (GM06990-HindIII and GM06990-NcoI), fibroblasts (IMR90), Embryonic Stem cells (ES) and myelogenous leukemia cells (K562).

Cell line	Restriction Enzyme	Reference
GM06990	HindIII	[1]
GM06990	NcoI	[1]
IMR90	HindIII	[13]
ES	HindIII	[13]
K562	HindIII	[1]

In order to test whether the Hi-C coverage of L and H regions is consistent with our previous findings at the 1Mb level, we have checked whether L regions have a significantly higher coverage (TID) than H regions, and vice versa, depending on the restriction enzyme and regardless of cell type (a difference was considered significant for Kolmogorov-Smirnov (KS) tests

that achieved $p < 0.05$ for the 2-sided test and $p < 0.01$ for the 1-sided test for the relevant hypothesis and $p > 0.05$ for the 1-sided test for the opposite hypothesis). 1-sided KS tests showed that L regions present a significantly larger TID than H regions in the (healthy) HindIII data sets (Fig. 3). 98.61% (*cis*) and 89.13% (*trans*) of chromosome pairs, on average over the 3 data sets (IMR90, ES and GM06990-HindIII), presented this significant difference ($p < 0.01$). As expected, an opposite trend was observed for the GM06990-NcoI data set (Fig. T1.2), where the TID of H regions was significantly larger than of L regions (100% and 97.83% of chromosome pairs in *cis* and *trans* correspondingly; $p < 0.01$). For the K562 data set the majority of chromosome pairs in *trans* (80.07%) showed a significantly higher TID for L regions (consistent with the HindIII cutting preference), while a mixed signal was observed in *cis* (Fig. T1.2), which can be attributed to an aberrant genome.

We have further verified that the observed relationship between isochores and Hi-C coverage is not a result of a confounding effect of unequal proportions between H and L regions. We have repeated the analysis with a random assignment of classes to regions (while retaining the classification proportions). Under this randomization the described patterns disappeared, even with a less stringent significance threshold ($p < 0.05$). Specifically, only 8.33% (3.48%) of the chromosomes pairs in *cis* (*trans*), presented a significant difference between the TID of L and H regions (on average over the 5 data sets under consideration). Similar (and smaller) percentages were obtained when testing each of the 1-sided hypotheses.

The findings of our analysis can be summarized as follows: (1) The occupancy of restriction sites in a given region is a proxy for its expected Hi-C coverage (2) This occupancy depends on the GC composition of the region (3) The heterogeneity of GC composition at the level of 1 Mb and at the level of isochores leads to a multi-scaled bias in Hi-C data.

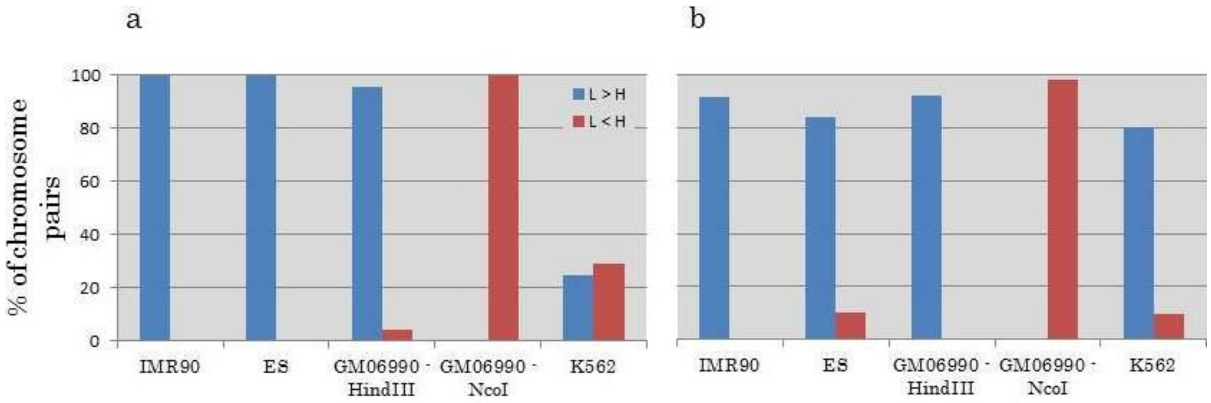

Fig. T1.2. Percentage of chromosome pairs in *cis* (a) and in *trans* (b) in which L regions presented significantly larger (blue) or smaller (red) TID than H regions.

Data preparation

Hi-C replicates (mapped contact positions; accession: GSE18199) generated for a GM06990 cell line [1] were downloaded from the Gene Expression Omnibus (GEO) [15] and transformed to corresponding contact maps (a contact map is a matrix, where the cell M[i,j] in a given contact map M takes the values of the contact frequency between 2 regions i and j, derived from a given segmentation of the genome). The contact maps were combined according to the restriction enzyme used, resulting in 2 contact maps - one for each enzyme. We have then removed all uncovered regions (i.e. those with zero row and column sums). Next, TIF (*self*, *cis* and *trans*) properties were calculated for each region from the contact maps. In order to calculate the CRF property, we have first generated the expected restriction fragments of each enzyme for the hg18 reference genome. Fragments were then matched to regions and the number of contained fragments was calculated. The GCP property was calculated for each domain according to its genome sequence (reference genome hg18).

Isochores boundaries generated by IsoFinder for the human reference genome hg18 were downloaded from: http://bioinfo2.ugr.es/isochores1/. Processed Hi-C data sets (mapped position pairs) generated for 4 human cell lines (Table T1.1) were downloaded from GEO [15]. Given the data sets and the IsoFinder segmentation file we have generated contact maps for each replicate following by addition of contact maps (replicates) for each cell line (unless the replicates were generated by different restriction enzymes as done for the GM06990 data set). The total contact frequency of each region was calculated for each chromosomal pair in *cis* and in *trans* (sum of rows and columns for the contact map of each chromosomal pair).

References

1. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES and Dekker J: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome**. *Science* 2009, **326**:289–93.
2. Yaffe E and Tanay A: **Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture**. *Nat. Genet*. 2011, **43**:1059–65.
3. Macaya G, Thiery JP and Bernardi, G: **An approach to the organization of eukaryotic genomes at a macromolecular level**. *J Mol Biol* 1976, **108**:237–54
4. Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M and Rodier F: **The mosaic genome of warm-blooded vertebrates**. *Science* 1985, **228**: 953–58.
5. Bernardi G: **Isochores and the evolutionary genomics of vertebrates**. *Gene* 2000, **241**:3–17.
6. Fullerton S, Carvalho A, and Clark A: **Local rates of recombination are positively correlated with GC content in the human genome**. *Mol Biol Evol* 2001, **18**:1139–42.

7. Tenzen T, Yamagata T, Fukagawa T, Sugaya K, Ando A, Inoko H, Gojobori T, Fujiyama A, Okumura K, Ikemura T: **Precise switching of DNA replication timing in the GC content transition area in the human major histocompatibility complex**. *Mol Cell Biol* 1997, **17**:4043–4050.

8. Saccone S, Federico C, and Bernardi G. **Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds**. *Gene* 2002, **300:**169–78

9. Mouchiroud D, D'Onofrio G, Aïssani B, Macaya G, Gautier C, Bernardi G: **The distribution of genes in the human genome**. *Gene* 1991, **100**:181–87.

10. Zoubak S, Clay O and Bernardi G: **The gene distribution of the human genome**. *Gene* 1996, 174:95–102.

11. Lander, E.S., *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860–921.

12. Oliver JL, Carpena P, Hackenberg M, and Bernaola-Galva´n P: **IsoFinder: computational prediction of isochores in genome sequences**. *Nucl Acids Res* 2004, **32**:W287–W292.

13. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS and Ren B: **Topological domains in mammalian genomes identified by analysis of chromatin interactions**. *Nature* 2012, **485**:376–80.

14. Oliver JL, Bernaola-Galván P, Carpena P, Román-Roldán R: **Isochore chromosome maps of eukaryotic genomes**. *Gene* 2001, **276:**47–56.

15. Edgar R, Domrachev M and Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucl. Acids Res*. 2002, **30**:207-10.