IsoXpressor: a tool to assess transcriptional activity within isochores

Lorraine A. K. Ayad,*,[1] Athanasia-Maria Dourou,[2] Stilianos Arhondakis,[2] and Solon P. Pissis,*,[3,4]

[1]Department of Informatics, King's College London, London, UK

[2]Bioinformatics and Computational Science (BioCoS), Boniali Str. 11-19, Building C, 73134 Chania, Greece

[3]CWI, Amsterdam, The Netherlands

[4]Vrije Universiteit, Amsterdam, The Netherlands

***Corresponding author:** E-mail: lorraine.ayad@kcl.ac.uk; solon.pissis@cwi.nl

Associate Editor:

Letter

Abstract

Genomes are characterised by large regions of homogeneous base compositions known as isochores. The latter are divided into GC-poor and GC-rich classes linked to distinct functional and structural properties. Several studies have addressed how isochores shape function and structure. To aid in this important subject, we present IsoXpressor, a tool designed for the analysis of the functional property of transcription within isochores. IsoXpressor allows users to process RNA-seq data in relation to the isochores, and it can be employed to investigate any biological question of interest for any species. The results presented herein as proof of concept are focused on the pre-implantation process in *Homo sapiens* (human) and *Macaca mulatta* (rhesus monkey).

Key words: isochores, transcription, pre-implantation, GC level, RNA-seq

Genomes contain regions that are largely GC-based, known as isochores. It is known that a link exists between the functional and structural properties between GC-rich and GC-poor isochores. To aid further in these investigations, we have developed IsoXpressor, a tool which analyses the isochores of any given species to enhance our understanding on how isochores drive its functionality and structure.

Introduction

Genomes are characterised by long DNA segments with a fairly homogeneous base composition known as isochores. The latter are divided into 5 classes: L1, L2 and H1 (GC-poor), and H2 and H3 (GC-rich). Several studies have shown distinct functional (e.g., replication timing, expression activity, methylation, gene ontology) and structural (e.g., chromatin structure/architecture and gene density) properties between GC-poor and GC-rich isochores (Arhondakis *et al.*, 2011, 2020; Bernardi, 2015, 2018; Jabbari and Bernardi, 2017; Jabbari *et al.*, 2019). It is clear that a link between isochores and transcription represents a study of relevant evolutionary importance, associating the appearance of GC-rich isochores to natural selection. To aid in this important subject, we present IsoXpressor, a tool that allows for a rapid and accurate investigation of any biological process (e.g., ageing and development) or condition (e.g., cancer and genomic diseases) of any species in relation to its isochores.

Herein, pre-implantation was selected as proof of concept. The utilisation of compositional strategies to investigate biological processes or conditions offers different advantages. The first is to enhance our understanding on how isochores drive functionality by coordinating and shaping regulation mechanisms, where the latter is known to be correlated to GC levels. The second one is of practical relevance, since our approach identifies isochores, i.e., specific genomic regions, that represent targets for a more detailed investigation regarding the genes they contain and their function (Arhondakis *et al.*, 2020).

Methods

IsoXpressor is implemented in Python and is freely available at `https://github.com/lorrainea/IsoXpressor` under GNU GPLv3. A wiki page to facilitate its usage is also available at the same location.

In this section, the external tools used by the pipeline are discussed as well as how IsoXpressor uses their output to produce analytical results for each isochore.

External Tools

The IsoXpressor pipeline makes use of two external tools to be able to assess the transcriptional activity within the isochores.

REAL

REAL (Frousios *et al.*, 2010) is an efficient short-read aligner for next-generation sequencing data. It takes as input two files, the first is a FASTA file containing a reference genome sequence, and the second is a FASTA or FASTQ file containing short reads. It then aligns the reads onto the reference genome.

There are several parameters used by IsoXpressor to run REAL. These are used to accustom the output of IsoXpressor. These include: `-s`, the maximum number of errors allowed in the seed part of the read; `-e`, the total maximum number of errors; and -ℓ, the length of the seeds.

The output of REAL is a text file containing properties of the aligned reads. These include, but are not limited to: the length of the read R_ℓ; the id of the reference sequence; i.e., the chromosome to which the read was aligned to, which we denote by C_R; as well as the starting position R_s of where the read is aligned to the chromosome.

IsoSegmenter

The second external tool used by IsoXpressor is IsoSegmenter (Cozzi *et al.*, 2015). It is a tool designed to segment a genomic sequence into isochores. It takes as input a single sequence in FASTA format and splits the sequence into windows of a user specified length. These windows are assigned to different isochore classes (L1, L2, H1, H2, H3) depending on the GC level (percentage of G and C bases) identified in each window.

We make use of the window parameter -w for IsoSegmenter to be able to scan the chromosomes of the input genome.

IsoSegmenter outputs a list of isochores denoting properties such as their GC level, the starting position I_s of the isochore within chromosome C_I, as well as the ending position and the length of the isochore denoted by I_ℓ.

IsoXpressor

Let us describe the main pipeline of the tool. **Input:** A reference genome and a collection of RNA-Seq read datasets, which are partitioned into k processes or conditions.

Output: The number of reads that align within each isochore of the reference genome as well as the corresponding Transcripts Per Kilobase Million (TPM) score (Li and Dewey, 2011) or Reads Per Kilobase Million (RPKM) score (Mortazavi *et al.*, 2008) for each isochore class in each process or condition.

1. The reads are aligned onto the reference genome using the REAL tool.
2. IsoSegmenter extracts the isochores for each chromosome of the reference genome.
3. The count of the aligned reads within each isochore for each chromosome is computed using the output files from Steps 1-2.
4. Using the read counts from Step 3, the TPM and RPKM score of each isochore for each process or condition is computed, representing the transcriptional activity. Note that these scores are adapted to the isochore lengths rather than gene lengths. Details on how TPM and RPKM are computed can be found in the corresponding section below.

Counting Reads

Given the set S of computed isochores and the set T of read sets, IsoXpressor computes the count $r_{i,j}$ of the number of reads identified within each isochore S_i from each read set T_j. For each isochore and read, such that $C_I = C_R$, more than half of the base pair positions $R_s,\dots,R_s+R_\ell-1$ should be the same as as $I_s,\dots,I_s+I_\ell-1$. More formally, as long as one of the two following formulae hold, the count of the number of reads located within this specific isochore is increased.

$$R_s+R_\ell-1-I_s > I_s-R_s \text{ and } R_s+R_\ell \leq I_s+I_\ell$$

or

$$I_s+I_\ell-1-R_s > (R_s+R_\ell-1)-(I_s+I_\ell-1) \text{ and } R_s \geq I_s$$

TPM and RPKM scores of isochores

The TPM of an isochore is computed as follows:

1. Divide the read counts by the length of each isochore in kilobases to obtain the reads per kilobase (RPK).
2. Count up all the RPK values in a sample and divide this number by 1,000,000, to assess the "per million" scaling factor.
3. Divide the RPK values by the "per million" scaling factor to obtain the TPM of each isochore.

The RPKM of an isochore is computed as follows:

1. Count up the total reads in a sample and divide that number by 1,000,000 ("per million" scaling factor).
2. Divide the read counts by the "per million" scaling factor to obtain the reads per million (RPM).
3. Divide the RPM values by the length of the isochore, in kilobases, to obtain the RPKM of each isochore.

IsoXpressor computes the TPM or RPKM expression for each isochore as described in detail below; it uses the same computation for sequencing depth (number of unique reads aligned to the reference genome (Sims *et al.*, 2014)) and isochore length rather than the more commonly used gene length.

Given k conditions or processes, $c_0,\ldots,c_{k-1}$, where c_i is computed using read counts $r_{i,m},\ldots,r_{i,m+n-1}$, where $m,\ldots,m+n-1$ corresponds to the indices of read sets associated to condition c_i within the set of all read sets, the TPM and RPKM scores for each isochore and process or condition are formally computed as follows:

$$A_j = \sum_{i=0}^{|S|-1} \frac{r_{i.j}\times 10^3}{I_{\ell_i}} \qquad \mathrm{TPM}_{i,c_i} = \sum_{j=m}^{m+n-1} \frac{r_{i.j}\times 10^9}{I_{\ell_i}\times A_j\times n}$$

$$B_j = \sum_{i=0}^{|S|-1} r_{i.j} \qquad \mathrm{RPKM}_{i,c_i} = \sum_{j=m}^{m+n-1} \frac{r_{i.j}\times 10^9}{I_{\ell_i}\times B_j\times n}$$

Table 1 contains *synthetic* counts $r_{i,j}$ of each read set T_j which occurs in an isochore S_i where $|S|=3$ and $|T|=4$. Note that random values have been used in this example.

Table 2 contains the TPM results for Table 1 given that $k=2$ where c_0 (Condition 1) is computed using reads $r_{i,0},\ldots,r_{i,1}$ and c_1 (Condition 2) is computed using reads $r_{i,2},\ldots,r_{i,3}$.

Table 3 contains the RPKM results for Table 1 given that $k=2$ where c_0 (Condition 1) is computed using reads $r_{i,0},\ldots,r_{i,1}$ and c_1 (Condition 2) is computed using reads $r_{i,2},\ldots,r_{i,3}$.

IsoXpressor also computes the average TPM and RPKM scores for each chromosome as well as

Table 1. Read counts for 3 isochores and 4 read sets.

Chromosome	Isochore Class	GC Level	Isochore Start	Isochore End	Isochore Size	Reads 1	Reads 2	Reads 3	Reads 4
7	H2	47.41	1	2396	2396	13	5	21	1
2	H2	50.94	673437	873436	200000	725	1007	808	3128
9	H3	56.12	873437	1573436	700000	6666	16489	19444	17242

Table 2. TPM scores for Table 1 where $k = 2$.

Chromosome	Isochore Class	GC Level	Isochore Start	Isochore End	Isochore Size	Condition 1	Condition 2
7	H2	47.41	1	2396	2396	180072.03	113115.77
2	H2	50.94	673437	873436	200000	179648.25	241966.59
9	H3	56.12	873437	1573436	700000	640279.71	644917.64

Table 3. RPKM scores for Table 1 where $k = 2$.

Chromosome	Isochore Class	GC Level	Isochore Start	Isochore End	Isochore Size	Condition 1	Condition 2
7	H2	47.41	1	2396	2396	426.02	226.41
2	H2	50.94	673437	873436	200000	388.65	483.52
9	H3	56.12	873437	1573436	700000	1316.07	1289.65

each isochore class. These are output as CSV files along with the average TPM and RPKM scores for the read counts as well as an output of the raw read counts. It also outputs a visualisation of the chromosome profiles for each chromosome as seen in the Supplement.

Experimental Results

The whole human genome (version GRCh38.p13) was used as input as well as RNA-seq data from 4 pre-implantation stages from the same species: Condition 1 (2-cell/GSM116020-GSM116022); Condition 2 (4-cell/GSM116023-GSM116026); Condition 3 (8-cell/GSM116027-GSM116037); and Condition 4 (Morula/GSM116038-GSM116040) (Xue *et al.*, 2013).

We also used the whole genome of rhesus monkey (version Mmul_10) as input as well as RNA-Seq data from 5 pre-implantation stages from the same species: Condition 1 (2-Cell/GSM2310522); Condition 2 (4-Cell/GSM2310523); Condition 3 (8-Cell/GSM2310524); Condition 4 (Morula/GSM2310525); and Condition 5 (Blastula/GSM2310526); each containing a single RNA-Seq read dataset (Wang *et al.*, 2017).

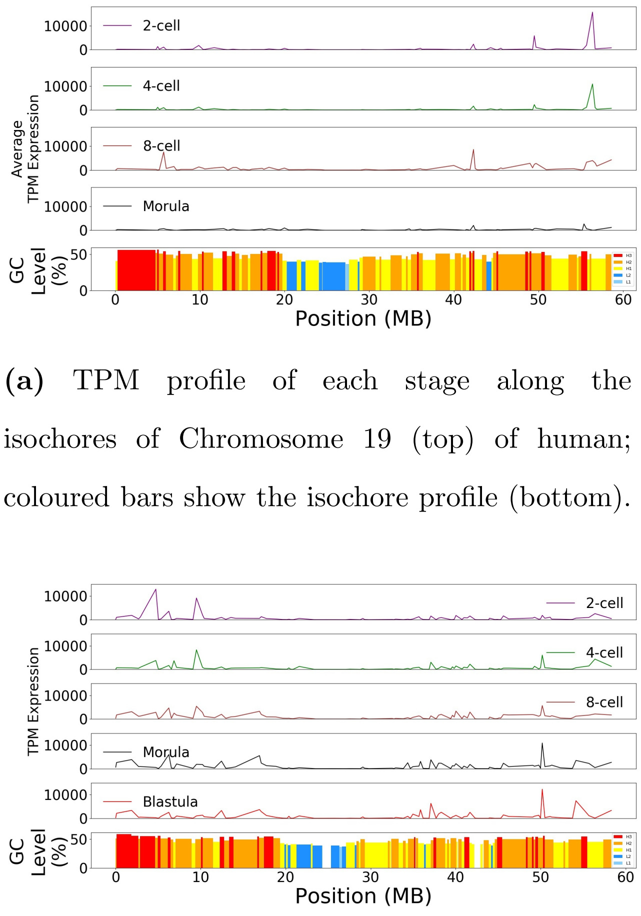
The default parameters of IsoXpressor were used in this experiment ($\mathtt{e} = 5$, $\mathtt{s} = 2$, $\ell = 32$, and $\mathtt{w} = 100000$). The following results show the output produced by IsoXpressor with these inputs and parameters. More experimental results can be found in the Supplement.

Figure 1a shows the profile of GC-rich Chromosome 19 in human and Figure 1b shows likewise of rhesus monkey. Similarly, Figures 2a and 2b show the transcriptional activity of the isochores of Chromosome 14 in human and rhesus monkey, respectively. As observed for Chromosome 19, coordinated transcriptional activity can be assessed for Chromosome 14. From the figures, it is evident that there is an extensive coordinated activation of blocks of adjacent isochores as pre-implantation advances, rather than random activation. Regarding the isochore profiles (bottom) of Chromosome 19 and 14 of human and rhesus monkey as depicted in

Figures 1 and 2, respectively, these agree with those reported for human in (Cozzi *et al.*, 2015), as well as in previous studies (Costantini *et al.*, 2009). In addition, the isochores' distribution across rhesus monkey chromosomes are given for the first time. The results are consistent for all chromosomes (see the Supplement).

Summarising, the expression of the chromosomes' isochores appear to hold a non-random profile during the pre-implantation process. Our findings add evidence of a sequential activation of adjacent blocks of isochores as pre-implantation advances. The latter may reflect regulatory effects such as rearrangements of chromatin structure, supporting the presence of a link between compositional properties to structural, spatial and functional properties at a chromosomal level (Arhondakis *et al.*, 2011; Bernardi, 2018, 2019; Hansen *et al.*, 2018; Jabbari and Bernardi, 2017; Jabbari *et al.*, 2019; Lamolle *et al.*, 2018; Sabbia *et al.*, 2009).

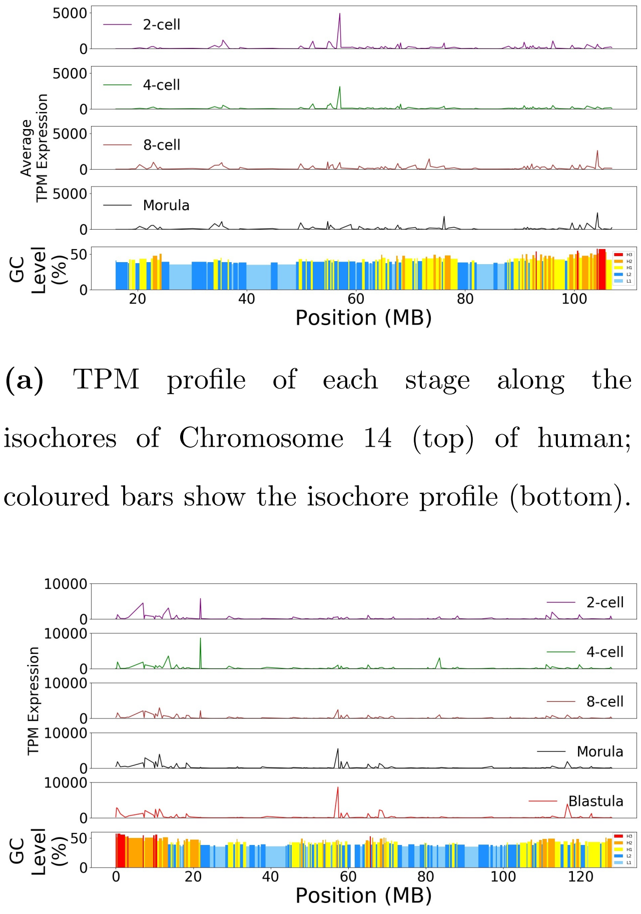
FIG. 1.



(a) TPM profile of each stage along the isochores of Chromosome 19 (top) of human; coloured bars show the isochore profile (bottom).

(b) TPM profile of each stage along the isochores of Chromosome 19 (top) of the rhesus monkey; coloured bars show the isochore profile (bottom).

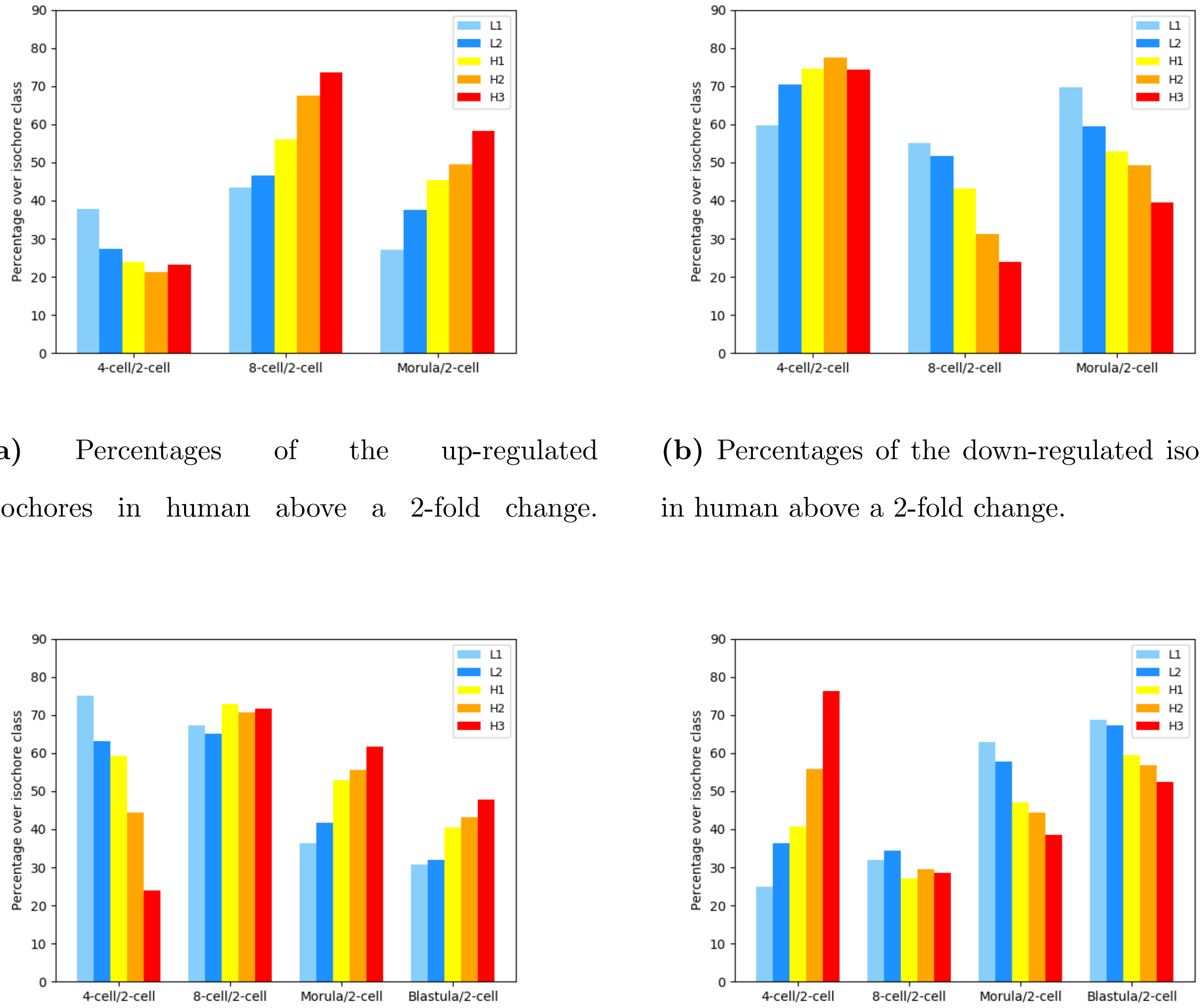
FIG. 2.



(a) TPM profile of each stage along the isochores of Chromosome 14 (top) of human; coloured bars show the isochore profile (bottom).

(b) TPM profile of each stage along the isochores of Chromosome 14 (top) of the rhesus monkey; coloured bars show the isochore profile (bottom).

FIG. 3.



(a) Percentages of the up-regulated isochores in human above a 2-fold change.

(b) Percentages of the down-regulated isochores in human above a 2-fold change.

(c) Percentages of the up-regulated isochores in rhesus monkey above a 2-fold change.

(d) Percentages of the down-regulated isochores in rhesus monkey above a 2-fold change.

Figure 3 depicts the percentages of the up- and down-regulated isochores with a fold change above two. The fold-change is computed as the $\log_2$ ratio of the expression (TPM or RPKM) of each stage over that of the 2-cell stage. In this work, up-regulation of an isochore indicates an increase in its transcriptional activity as pre-implantation advances, while down-regulation a decrease (silencing). In other words, the up-regulated isochores here are those that have a 2-fold higher expression (TPM or RPKM) in another condition compared to the 2-cell condition (see position 58MB in Figure 2b). The down-regulated isochores here are those that have a 2-fold lower expression (TPM or RPKM) in another condition compared to the 2-cell (see position 58MB in Figure 2a). The up- and down-regulated percentages are computed by dividing the number of up- or down-regulated isochores of each class over the total number of isochores of the corresponding class.

In human, we observe that after the 4-cell stage there is an increase in the percentage of the GC-rich ones (H2 and H3; Figure 3a) and a decrease of the down-regulated ones (Figure 3b). This can be translated as a tendency of GC-rich isochores to increase their expression as pre-implantation advances. This is similarly observed for H1, although it is less evident than it is for H2 and H3. On the contrary, L1 and L2 display a constant profile as pre-implantation advances. Hence, a lower percentage within up-regulated (Figure 3a) compared to down-regulated isochores (Figure 3b) is observed. The latter indicates that most of the L1 and L2 isochores tend to reduce their expression activity as pre-implantation advances.

In rhesus monkey, profiles are similar to human, with the exception of the H1 class. Indeed, within the up-regulated isochores (Figure 3c) after the first stage (2-cell to 4-cell) the GC-rich (H2 and H3) increase while the GC-poor (L1, L2 and H1) decrease (Figure 3c). It is worth mentioning that L1 and L2 classes reach a decrease of $\approx$50% (Figure 3c). The above indicates that the GC-rich isochores increase their expression (up-regulated) as pre-implantation progresses. Regarding the down-regulated isochores (Figure 3d), during the early stage (between 2 and 4-cell) H2 and H3 classes exhibit a peak compared to the GC-poor classes. This points towards a possible silencing of most of the H2 and H3 isochores at very early stages. Interestingly, percentages of L1 and L2 increase approximately 3 times in the late stages of pre-implantation (Morula and Blastula), further supporting their silencing or down-regulation in these stages.

The observed differences between human (Figure 3a and 3b) and rhesus monkey (Figure 3c and 3d) may partially reflect experimental differences, without excluding potential biological diversification between these species. To conclude, our observations within each species support a preferential activation (up-regulation) of GC-rich isochores at late pre-implantation stages, and a silencing (down-regulation) of the GC-poor ones.

Conclusion

We introduced IsoXpressor, an efficient and accurate tool dedicated to the analysis of isochores' transcriptional activity. Our tool offers the possibility to investigate any biological process or condition and can be applied to any species. To the best of our knowledge, no other automated tool is currently available which is capable of processing RNA-Seq data related to isochores.

Despite the solidity of the results presented, it would be relevant to use larger datasets in order to gain a more holistic perspective on the case study under investigation. However, we do believe that the findings presented herein offer novel insights on how genome organisation, in terms of isochores, acts upon regulation during the preimplantation process in both human and rhesus monkey e.g., affecting chromatin architecture (Bernardi, 2015, 2019; Du *et al.*, 2017).

We anticipate that IsoXpressor will open new perspectives on investigating important biological questions, where the genome is considered as a functional and structural ensemble, with its most fundamental property, the base composition, shaping regulation.

Data Availability

The data underlying this article are available in the article and in its online supplementary material.

References

Arhondakis, S., Auletta, F., and Bernardi, G. 2011. Isochores and the Regulation of Gene Expression in the Human Genome. *Genome Biology and Evolution*, 3: 1080–1089.

Arhondakis, S., Milanesi, M., Castrignanò, T., Gioiosa, S., Valentini, A., and Chillemi, G. 2020. Evidence of distinct gene functional patterns in GC-poor and GC-rich isochores in bos taurus. *Animal Genetics*, 51(3): 358–368.

Bernardi, G. 2015. Chromosome architecture and genome organization. *PLOS ONE*, 10(11): 1–21.

Bernardi, G. 2018. The formation of chromatin domains involves a primary step based on the 3-d structure of dna. *Scientific Reports*, 8(1): 17821.

Bernardi, G. 2019. The genomic code: A pervasive encoding/molding of chromatin structures and a solution of the "non-coding dna" mystery. *BioEssays*, 41(12): 1900106.

Costantini, M., Cammarano, R., and Bernardi, G. 2009. The evolution of isochore patterns in vertebrate genomes. *BMC Genomics*, 10(1): 146.

Cozzi, P., Milanesi, L., and Bernardi, G. 2015. Segmenting the human genome into isochores. *Evolutionary Bioinformatics*, 11.

Du, Z., Zheng, H., Huang, B., Ma, R., Wu, J., Zhang, X., He, J., Xiang, Y., Wang, Q., Li, Y., Ma, J., Zhang, X., Zhang, K., Wang, Y., Zhang, M. Q., Gao, J., Dixon, J. R., Wang, X., Zeng, J., and Xie, W. 2017. Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature*, 547(7662): 232–235.

Frousios, K., Iliopoulos, C. S., Mouchard, L., Pissis, S. P., and Tischler, G. 2010. REAL: an efficient REad ALigner for next generation sequencing reads. In *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, BCB '10, pages 154–159, New York, NY, USA. ACM.

Hansen, A. S., Cattoglio, C., Darzacq, X., and Tjian, R. 2018. Recent evidence that tads and chromatin loops are dynamic structures. *Nucleus*, 9(1): 20–32.

Jabbari, K. and Bernardi, G. 2017. An isochore framework underlies chromatin architecture. *PLOS ONE*, 12(1): 1–12.

Jabbari, K., Chakraborty, M., and Wiehe, T. 2019. DNA sequence-dependent chromatin architecture and nuclear hubs formation. *Scientific Reports*, 9(1): 14646.

Lamolle, G., Sabbia, V., Musto, H., and Bernardi, G. 2018. The short-sequence design of dna and its involvement in the 3-d structure of the genome. *Scientific Reports*, 8(1): 17820.

Li, B. and Dewey, C. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. bmc bioinformtics 12:323. *BMC Bioinformatics*, 12: 93–99.

Mortazavi, A., Williams, B., Mccue, K., Schaeffer, L., and Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5: 621–8.

Sabbia, V., Romero, H., Musto, H., and Naya, H. 2009. Composition profile of the human genome at the chromosome level. *Journal of Biomolecular Structure and Dynamics*, 27(3): 361–369.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2): 121–132.

Wang, X., Liu, D., He, D., Suo, S., Xia, X., He, X., Han, J.-D. J., and Zheng, P. 2017. Transcriptome analyses of rhesus monkey preimplantation embryos reveal a reduced capacity for dna double-strand break repair in primate oocytes and early embryos. *Genome research*, 27(4): 567–579.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.-y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y. E., Liu, J.-y., Horvath, S., and Fan, G. 2013. Genetic programs in human and mouse early embryos revealed by single-cell rna sequencing. *Nature*, 500(7464): 593–597.