DOI 10.18699/SBB-2018-01

CpG islands' clustering uncovers early development genesin the human genome

R.O. Babenko[1], A.V. Tsukanov[1, 2], A.G. Galieva[1], Y.L. Orlov[2], V.N. Babenko[2, 3]
1 *Novosibirsk State University, Novosibirsk, Russia*
2 *Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
** e-mail: rbab@yandex.ru*

Key words: non-coding RNA, transcriptomics, stress response, crop plants

Motivation and Aim: We address the problem of the annotation of CpG islands (CGIs) clusters in the human genome. CpG dinucleotide rich genome regions, also known as CpG islands (CGIs), are important functional elements of vertebrate genomes [1]. In particular, in the majority of vertebrate genes, CpG islands coincide with gene promoter areas. In some cases, the transcription from CpG-island containing promoters is bidirectional, this is related to self-complementarity of CG dinucleotides. CGIs are the key contributors to global methylation landscapes. Degenerate content of CGIs (biased CG frequency) assumes a higher probability of tandem repeats and palindromes inside a CGI.
Methods and Algorithms: We have used own program scripts for tandem repeats and CGI counts. We used a CGI clustering method that is robust relative to the tandem duplication search A set of CGIs was retrieved from the table cpgIslandExt (www.genome.ucsc.edu; version hg19). To identify significant CGI clustering, the human genome was split into 10Kb non overlapping segments (bins) (243 785 bins in total). The number of CGIs per bin (CGI density) was assessed as a total number of CGIs divided by the number of bins. The expected number of CGIs per segment was approximated using a Poisson distribution
Results: Upon analyzing gene content within CGIs clusters, piRNA, tRNA, and miRNA-encoding genes were found as well as CpG-rich homeobox genes reported previously. Chromosome-wide CGI density is positively correlated with replication timing, confirming that CGIs may serve as open chromatin markers. Early embryonic stage expressed KRAB-ZNF genes abundant at chromosome 19 were found to be interlinked with CGI clusters.
Conclusion: We detected that a number of long CGIs and CGI clusters are, in fact, tandem copies with multiple annotated macrosatellites and paralogous genes. This finding implies that tandem expansion of CGIs may serve as a substrate for non-homologous recombination events [1].
Acknowledgements: The research has been supported by RFBR. Computing done at Siberian Supercomputer center SB RAS was supported by budget project 0324-2018-0017.

References

1. Babenko V.N., Bogomolov A.G., Babenko R.O., Galieva E.R., Orlov Y.L. (2018) CpG islands' clustering uncovers early development genes in the human genome. Computer Science Information Systems. 15(2):473-485.