DOI 10.18699/WIBSB-2018-15

Quantifying genome sequence repeatability by repeater

M. Dai[1, 2], C. Feng[1], M. Chen[1]*

[1] *Department of Bioinformatics, the State Key Laboratory of Plant Physiology and Biochemistry, College of Life Sciences, Zhejiang University, Hangzhou, China*

[2] *Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China*

* *e-mail: mchen@zju.edu.cn*

Key words: DNA repeats, repeatability map

Motivation and Aim: DNA repeats are abundant in eukaryotic genomes and illustrated to play a crucial part in genome evolution and regulation. To identify various repeats and classify them into different families are the key aspects of previously related researches. And a large number of approaches are proposed to detect and annotate repeat elements in the genome. While a few de novo repeats identification methods are able to generate quantitative repetitiveness map of genome sequences based on word counting algorithms. However, characteristics of such repetitiveness pattern are not well recognized and the applications are still limited. Therefore we developed a software named Repeater that quantifies the sequence repeatability, which is defined as a genome property to reflect the sequence repetitive level.

Methods and Algorithms: Repeater mainly consists of three parts. In the first part, we presented min tree, a modified suffix tree data structure, to compute all the exact repeats in the genome. The min tree is realized in Java and contains less redundant information than the suffix tree and able to run in linear time and space. Next, we transformed the fragmented sequence data originated from repeater-core to repeatability map. Both Single Nucleotide Repeatability (SNRP) and Sequence Average Repeatability (SARP) can be generated to represent the repeatability at single base and sequence region level. After we got the genome sequence repeatability map, MACS2 was utilized to capture the peaks which represent the highly repetitive regions.

Results: The analysis shows that Repeater performed well at identifying the highly repetitive sequences in the genome by determining the peak regions of repeatability map. And we found that repeatability map is complementary to the mappability track, so it may also useful in reads mapping check. Combined with ChIP-seq and DNase-seq data, highly repetitive regions identified by our tool are related to epigenetic modifications and chromatin accessibility. That indicates the potential capability of repeatability to serve as a genome feature for further structure and function predictions. The software and an online testing program are freely available at http://bis.zju.edu.cn/repeater. And we also provide repeatability track of common model species implemented with JBrowse on the website.

Conclusion: We developed a software Repeater to quantify the genome sequence repeatability. Highly repetitive regions in the genome can be efficiently detected on the basis of repeatability map. And we demonstrated that repeatability may be useful for reads mapping check and chromatin accessibility prediction.

Acknowledgements: We are grateful to the members of Ming Chen's laboratory for helpful discussions and comments.