

HiCDB: a sensitive and robust method for detecting contact domain boundaries

Fengling Chen^{1,†}, Guipeng Li^{1,2,†}, Michael Q. Zhang^{1,3,4} and Yang Chen^{1,3,*}

¹MOE Key Laboratory of Bioinformatics, Center for Synthetic and Systems Biology, Bioinformatics Division, BNRist, Department of Automation, Tsinghua University, Beijing 100084, China, ²Department of Biology, Medi-X Institute, SUSTech Academy for Advanced Interdisciplinary Studies, Southern University of Science and Technology, Shenzhen 518055, China, ³MOE Key Laboratory of Bioinformatics, Center for Synthetic and Systems Biology, School of Medicine, Tsinghua University, Beijing 100084, China and ⁴Department of Biological Sciences, Center for Systems Biology, The University of Texas, Dallas 800 West Campbell Road, RL11, Richardson, TX 75080-3021, USA

Received March 01, 2018; Revised August 09, 2018; Editorial Decision August 21, 2018; Accepted August 27, 2018

ABSTRACT

Contact domains are closely linked to gene regulation and lineage commitment, while current understanding of contact domains and their boundaries is still limited. Here, we present a novel method HiCDB, which is constructively based on local relative insulation metric and multi-scale aggregation approach to detect contact domain boundaries (CDBs) on Hi-C maps. Compared with other 'state-of-art' methods, HiCDB shows improved sensitivity and specificity in determining CDBs at various Hi-C resolutions. The superiority of HiCDB enabled us to study the epigenetic features of detected CDBs and showed enrichment of architectural proteins and cell-type-specific transcription factor binding sites at CDBs. The further comparison of GM12878 and IMR90 Hi-C datasets suggested that cell-type-specific CDBs are marked by active regulatory signals and correlate with activation of nearby cell identity genes.

INTRODUCTION

Chromatin organization and its functions in both gene regulation and cell identity have drawn great attention in cell biology researches. Recent developments in sequencing and imaging technologies have led to unprecedented progresses toward understanding chromatin organization (1–5). One of the most striking features of chromatin configuration is the squares with enhanced contact frequencies tiling the diagonal of chromatin interaction matrixes observed in Hi-C data (6–9). These squares were originally observed in the 40-kb resolution Hi-C maps and referred as topologically associating domains (TADs) by Dixon *et al.* (7). With increased sequencing depth, Rao *et al.* showed that there are contact domains within the megabase-sized chromatin domains

(8). Phillips-Cremins *et al.* elucidated that cell-type-specific chromatin organization occurs at this sub-megabase scale by looking into the chromosome conformation around six key developmentally regulated genes based on chromosome conformation capture carbon copy (5C) data (10). These cell-type-specific contact domains were also reported in regulation of HoxA genes in limbs development (11). It has also been demonstrated that changes of contact domains are accompanied by alternations in histone modifications and long-term contact pattern (8,12). However, few studies have compared the contact domain boundaries (CDBs) across cell types systemically or uncovered the association between CDBs and genome-wide histone modifications as well as transcription. Herein, sensitive and robust CDB detection methods are of great demand to reveal the function of the CDBs. In particular, deep-sequencing data are preferred for detecting more CDBs, which require the CDB detection methods to be computationally efficient in processing high-resolution Hi-C data.

Several computational methods have been proposed to detect chromatin domains or their boundaries on Hi-C maps (7,8,13–23). These methods can be categorized into 1D statistic-based methods and 2D contact matrix-based methods. The 1D statistic-based methods, such as directionality index (DI), Insulation score and TopDom, calculated a 1D statistic for each bin by averaging interaction frequencies in sliding windows on the original contact matrix (7,15,16). In the DI method, first, a metric called DI was proposed to define the direction preference of each bin in contact with 2 Mb upstream and 2 Mb downstream; then, a hidden Markov model was used to determine the domain boundaries by identifying interaction transitions from the upstream to the downstream (7). The Insulation score method assigned an insulation score to each bin by aggregating interactions of nearby regions. The local minimums of the insulation profile were identified as TAD boundaries (15). As a modification of Insulation score, the TopDom

*To whom correspondence should be addressed. Tel: +86 10 62795578 (Ext. 454); Fax: +86 10 62773552; Email: yc@tsinghua.edu.cn

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.