ARTICLE IN PRESS

Methods xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

# 5C-ID: Increased resolution Chromosome-Conformation-Capture-Carbon-Copy with *in situ* 3C and double alternating primer design

Ji Hun Kim[a,1], Katelyn R. Titus[a,1], Wanfeng Gong[a], Jonathan A. Beagan[a], Zhendong Cao[b], Jennifer E. Phillips-Cremins[a,b,c,*]

[a] *Department of Bioengineering, University of Pennsylvania, Philadelphia, PA 19104, USA*
[b] *Epigenetics Program, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA*
[c] *Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA*

ARTICLE INFO

*Keywords:*
3D genome folding
Higher-order chromatin organization
Chromosome-Conformation-Capture-Carbon-Copy
5C
Double alternating 5C primer design
*In situ* 3C
Low cell number
Embryonic stem cells

## ABSTRACT

Mammalian genomes are folded in a hierarchy of compartments, topologically associating domains (TADs), subTADs, and looping interactions. Currently, there is a great need to evaluate the link between chromatin topology and genome function across many biological conditions and genetic perturbations. Hi-C can generate genome-wide maps of looping interactions but is intractable for high-throughput comparison of loops across multiple conditions due to the enormous number of reads (> 6 Billion) required per library. Here, we describe 5C-ID, a new version of Chromosome-Conformation-Capture-Carbon-Copy (5C) with restriction digest and ligation performed in the nucleus (*in situ* Chromosome-Conformation-Capture (3C)) and ligation-mediated amplification performed with a double alternating primer design. We demonstrate that 5C-ID produces higher-resolution 3D genome folding maps with reduced spatial noise using markedly lower cell numbers than canonical 5C. 5C-ID enables the creation of high-resolution, high-coverage maps of chromatin loops in up to a 30 Megabase subset of the genome at a fraction of the cost of Hi-C.

## 1. Introduction

Higher-order folding of chromatin in the 3D nucleus has been linked to genome function. Mammalian genomes are arranged in a nested hierarchy of territories [1], compartments [2–4], topologically associating domains [5–8] (TADs), subTADs [3,9], and long-range looping interactions [10,11]. Looping interactions have been linked to at least two mechanistically different modes of control over gene expression. First, enhancers can loop to distal target genes in a highly cell type-specific manner to facilitate their precise spatial-temporal regulation [12–15]. Second, long-range loops anchored by the architectural protein CTCF are often constitutive among cell types and form the structural basis for TADs/subTADs [9]. CTCF-mediated interactions connecting loop domains can create insulated neighborhoods that demarcate the search space of enhancers within the domain [16]. Specifically, CTCF anchored constitutive loops can prevent ectopic enhancer activation of genes outside of the domain or aberrant invasion of nonspecific enhancers into an inappropriate domain [16–20]. Mapping of 3D loops genome-wide across hundreds of cell types, species, and developmental lineages is an active area of intense investigation.

As genome-wide chromatin architecture maps become widely available, a critical emerging goal will be to unravel the cause and effect relationship between looping and gene expression. Indeed, there is a great need in the field to build upon descriptive mapping studies and begin to perturb the 3D genome and evaluate the link between chromatin topology and function. One major limitation preventing progress toward this goal is that Hi-C requires more than six billion reads per replicate to obtain high quality, high resolution, genome-wide looping maps [3,13,21]. The financial and logistical difficulties of obtaining this read depth makes it intractable to conduct studies with multiple perturbations induced by genome editing, differentiation, or drug treatment. Thus, there is a great need for a technology that creates high-resolution 3D genome folding maps at a much lower cost.

Chromosome Conformation Capture Carbon Copy (5C) is proximity ligation technology pioneered by Dekker and colleagues [22,23]. 5C adds a hybrid capture step to the classic Chromosome Conformation Capture (3C) method to facilitate the selection of ligation products that occur only in a subset of the genome [22,24–26]. Loop-resolution maps can be achieved at a fraction of the cost of Hi-C by only querying interactions in a 10–30 Megabase (Mb) subset of the genome [7,9,17,27],

---

[*] Corresponding author at: Department of Bioengineering, University of Pennsylvania, Philadelphia, PA 19104, USA.
*E-mail address:* jcremins@seas.upenn.edu (J.E. Phillips-Cremins).
[1] Authors contributed equally to this work.

https://doi.org/10.1016/j.ymeth.2018.05.005
Received 19 November 2017; Received in revised form 1 March 2018; Accepted 8 May 2018

Please cite this article as: Kim, J.H., Methods (2018), https://doi.org/10.1016/j.ymeth.2018.05.005

allowing many samples and perturbation conditions to be screened in a high-throughput manner. The ability to query a subset of genome contacts is important because genome-editing experiments are often conducted at only one specific location in the genome. Thus, 5C enables the query of the organizing principles governing genome folding at a key subset of loops without requiring the resources to map all loops genome wide.

Despite key advantages in the original 5C technique, it also has key challenges that have held back its widespread use, including: (1) the high number of cells ( > 40 million) required for quality 3C template creation [2,9,27,28], (2) the high amount of spatial noise caused by non-specific ligation products [29,30], and (3) the non-comprehensive nature of the single alternating primer design [7,22–26,31]. Together, these limitations result in a high number of false negatives due to ligation junctions that are not queried and a high number of false positives due to spatial noise due to non-specific ligation. In the present study, we introduce two major modifications to the 5C protocol that lead to increased resolution 3D genome folding maps with reduced spatial noise using markedly lower cell numbers than canonical 5C. We conduct a comparative analysis of *in situ* [3,32,33] vs. canonical dilution 3C [2,28] and a double alternating [17] vs. single alternating primer design [7,22–26,31] and report the downstream effect of these changes on 5C's ability to detect bona fide looping interactions.

## 2. Results

### 2.1. Overview

A 5C experiment starts with preparation of the 3C template (Fig. 1A and B). Chromatin is fixed within a population of cells with formaldehyde. In canonical dilution 3C [2,28], cellular and nuclear membranes are disrupted and chromatin is digested in solution with a restriction enzyme (Fig. 1A). Ligation is subsequently performed under dilute conditions that promote intra-molecular ligation. By contrast, *in situ* 3C [3,32] involves restriction enzyme digest and ligation within intact nuclei. In both methods, cross-links are reversed and DNA is isolated to create the 3C template, which represents the genome-wide library of possible hybrid ligation junctions across a population cells (Fig. 1B).

The second half of the 5C protocol involves a hybrid capture step based on ligation-mediated amplification to select only a distinct subset of junctions from the genome-wide 3C library (Fig. 1C–F). Canonical 5C [7,22–26,31] is built on an alternating primer design in which every other fragment is represented by either a Forward (FOR) primer binding to the sense strand or a Reverse (REV) primer binding to the antisense strand (Fig. 1C, left). The single alternating design only queries approximately half of all ligation junctions in a target region because only FOR-REV primer ligation events are possible (Fig. 1D–E, left). More recently, Dekker, Lajoie and colleagues created a new double alternating primer design [17] which incorporates two additional "left-oriented' primers, LFOR and LREV (Fig. 1C right). The LFOR primer orientation is designed to the antisense strand on fragments also queried by REV primers, whereas the LREV primer orientation is designed to the sense strand on fragments also queried by FOR primers. Thus, the double alternating 5C primer design, there are now two primers representing each fragment, leading to 4 possible primer ligation orientations (FOR-REV, LFOR-LREV, LFOR-REV, FOR-LREV) and the query of nearly all fragment-fragment ligation events in an a priori selected Megabase (Mb)-scale genomic region (Fig. 1D and E right).

### 2.2. Double alternating primer design achieves increased loop detection sensitivity compared to single alternating design

We hypothesized that by using the double alternating design developed by Dekker and colleagues [17], we could improve canonical 5C's matrix resolution, and the specificity and sensitivity of loop detection. To test this idea, we first started with a canonical dilution 3C template from pluripotent embryonic stem (ES) cells cultured in 2i media (detailed in Materials and Methods) and compared the quality of 5C libraries created at the same genomic region with both single alternating and double alternating primer designs. A tradeoff of the more comprehensive double alternating primer design is the possibility of artifactual 'self-circles' (*i.e.* ligation events between the 5′ and 3′ ends of the same restriction fragment; (5) and (6) in Fig. 1D and E right). We counted the proportion of each possible primer ligation from the double alternating 5C experiment on a dilution 3C template from ES cells. There was an even distribution of ligation events across the four biologically informative primer-primer orientations ((1) FOR-REV: 18.60%, (2) LFOR-LREV: 17.87%, (3) LFOR-REV: 18.11%, (4) FOR-LREV: 18.41%). Importantly, self-circle ligation events ((5) LFOR-REV and (6) FOR-LREV from the same fragment) comprised only < 0.1% of all primer ligations (Fig. 1E), suggesting that the risk of self-ligation is very small.

We visually inspected 4 kb-binned heatmaps of 5C counts in Megabase-scale genomic regions around *Sox2* and *Zfp462* genes after matrix balancing and sequencing depth correction (detailed in Materials and Methods). We observed that the double alternating primer design results in notable improvement in specific, punctate looping signal between known long-range enhancer promoter-interactions compared to the single alternating primer design (Fig. 2A and B). Double alternating 5C maps also showed less missing fragments than single alternating primer maps due to the increased complexity of ligation junctions that are queried and sequenced. In previous 5C studies, a smoothing window at least 5× greater than the bin size was required to reduce the blockiness of maps caused by missing ligation junctions [27,34]. Here, with double alternating design, we can create heatmaps at 4 kb matrix resolution with no smoothing window and still resolve punctate loops. We provide heatmaps at a 4 kb bin resolution and a 12 kb smoothing window for ease of comparison across the technical conditions (Fig. 2A and B).

To further test our qualitative observation of increased looping sensitivity with the double alternating design, we also quantified chromatin looping interactions in each 5C dataset. We modeled binned interactions as a fold-enrichment relative to a background expected model based on distance dependence and local chromatin domain architecture (detailed in Materials and Methods). As previously published [27,30,34], we modeled these Observed/Expected values with a parameterized logistic distribution and subsequently converted p-values to interaction scores (Fig. 2C; detailed in Materials and Methods). After thresholding interaction scores, we clustered adjacent looping pixels into long-range looping interaction clusters (Fig. 2D; detailed in Materials and Methods). Consistent with observations in Fig. 2A and B, the interaction score and loop cluster maps also highlight punctate Sox2 and Zfp462 gene promoter-enhancer looping clusters (Fig. 2C and D). As expected, the chromatin fragments anchoring the base of detected looping interactions contained high signal for H3K27ac, a chromatin modification known to demarcate active non-coding regulatory elements and active transcription start sites. Importantly, we identified key looping interactions between *Zfp462* and distal enhancers with the double alternating primer design that were not present with the single alternating design. The well-established *Sox2*-super enhancer interaction [5,9,27,34–36] was detected by the single alternating design, but significantly more punctate and less blocky/noisy with the double alternating design. Overall, these data indicate that the double alternating primer design allows for more sensitive detection of looping interactions compared to the single alternating 5C primer design.

### 2.3. *In situ* 3C reduces spatial noise in 5C heatmaps compared to dilution 3C

We next assessed the quality of the double alternating 5C experiment using *in situ* 3C and dilution 3C templates. We prepared the *in situ*