Methods 170 (2020) 17–32

Contents lists available at ScienceDirect

Methods

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

4C-seq from beginning to end: A detailed protocol for sample preparation and data analysis

Peter H.L. Krijger, Geert Geeven, Valerio Bianchi, Catharina R.E. Hilvering, Wouter de Laat*

Oncode Institute, Hubrecht Institute-KNAW and University Medical Center Utrecht, Utrecht, the Netherlands

ARTICLE INFO

Keywords:
4C-seq
Chromosome conformation capture (3C)
Chromatin folding
Genome architecture
4D nucleome

ABSTRACT

Chromosome conformation capture (3C) methods measure DNA contact frequencies based on nuclear proximity ligation, to uncover *in vivo* genomic folding patterns. 4C–seq is a derivative 3C method, designed to search the genome for sequences contacting a selected genomic site of interest. 4C-seq employs inverse PCR and next generation sequencing to amplify, identify and quantify its proximity ligated DNA fragments. It generates high-resolution contact profiles for selected genomic sites based on limited amounts of sequencing reads. 4C-seq can be used to study multiple aspects of genome organization. It primarily serves to identify specific long-range DNA contacts between individual regulatory DNA modules, forming for example regulatory chromatin loops between enhancers and promoters, or architectural chromatin loops between cohesin- and CTCF- associated domain boundaries. Additionally, 4C-seq contact profiles can reveal the contours of contact domains and can identify the structural domains that co-occupy the same nuclear compartment. Here, we present an improved step-by-step protocol for sample preparation and the generation of 4C-seq sequencing libraries, including an optimized PCR and 4C template purification strategy. In addition, a data processing pipeline is provided which processes multiplexed 4C-seq reads directly from FASTQ files and generates files compatible with standard genome browsers for visualization and further statistical analysis of the data such as peak calling using peakC. The protocols and the pipeline presented should readily allow anyone to generate, visualize and interpret their own high resolution 4C contact datasets.

1. Introduction

Chromosome conformation capture (3C) based methods [1,2] are key technologies to study the three-dimensional organization of genomes and have helped to uncover the intricate relationship between genome structure and transcription [3]. Based mostly on these methods, we now know that enhancers regulate expression of distal target genes via long range regulatory chromatin loops within the genomic context of topologically associated domains (TADs) [4–7]. TADs are hundreds of kilobases to several megabases (Mb) in length and are thought to restrict the search space of enhancers by the formation of CTCF mediated architectural chromatin loops formed by loop extrusion [8,9]. At a larger scale TADs with similar chromatin properties often cluster together in nuclear space into active (A) and inactive (B) compartments [10–13] by a process called phase separation [14].

3C-based technologies rely on formaldehyde-mediated crosslinking of DNA sequences that are spatially proximal in the cell nucleus, followed by DNA fragmentation using restriction enzymes and *in situ* ligation of the cross-linked fragments (“3C template”) (see Fig. 1A). Subsequently, the ligated fragments are de-cross-linked and purified and the ligation junctions are detected and quantified. The ligation frequencies serve as measures of contact frequencies in a population of cells. Various high-throughput variants of 3C have been developed that quantify interaction frequencies using next generation sequencing [2]. Hi-C determines the pairwise interaction frequencies between all possible regions across the genome [11,12] and is currently being employed to catalog and compare the 3D genome organization of many cell types [15]. While high resolution genome wide contact maps require billions of sequencing reads, targeted 3C variants such as 4C-seq [10], UMI-4C [16] and Capture-C [17] have been developed to determine interaction frequencies of one specific genomic site of interest with all other genomic regions and create high resolution contact profiles for selected genomic sites without the need to sequence so deeply [2]. In 4C-Seq, the 3C template is trimmed using a secondary restriction enzyme, followed by circularization in a second ligation step (Fig. 1B). Next an inverse PCR is performed with primers hybridizing to the selected restriction fragment (“the viewpoint”), to amplify its ligation partners (“captures”). 4C-seq primers carry overhang sequences

* Corresponding author.
E-mail address: w.delaat@hubrecht.eu (W. de Laat).

https://doi.org/10.1016/j.ymeth.2019.07.014
Received 15 April 2019; Accepted 14 July 2019
Available online 26 July 2019

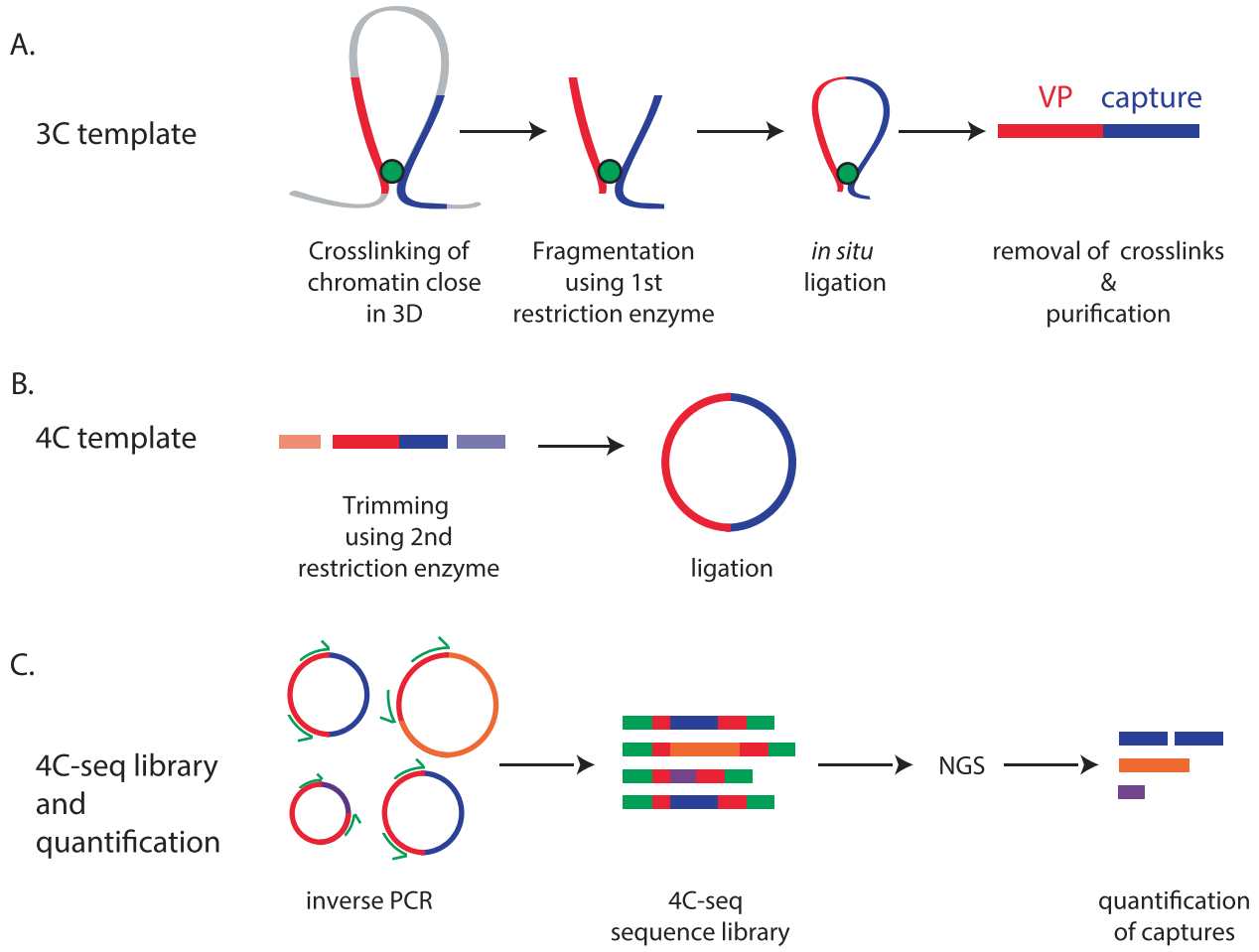

Fig. 1. Outline of the 4C-seq procedure. (A.) Genomic regions that are spatially proximal in the cell nucleus (red and blue) are fixated by formaldehyde induced potein-protein and protein-DNA crosslinks (green) [Section 4.1]. The DNA is fragmented using the primary restriction enzyme [Section 4.3], fragments in close proximity are ligated *in situ* [Section 4.4], after which the crosslinks are removed and the resulting hybrid molecules ("the 3C template") purified [Section 4.5]. (B.) In 4C-seq the 3C template is trimmed using a secondary enzyme [Section 4.6] and circularized in the second ligation step [Section 4.7]. (C.) To identify and quantify fragments that are ligated ("captures", indicated in blue, orange and purple) to the genomic region of interest (the viewpoint, red), an inverse PCR is performed with primers binding outward on the viewpoint [Section 5] and the amplicons are analyzed using next generation sequencing [Section 6]. The number of reads mapped to genomic regions [Section 7] are taken as a measure of the contact frequencies in a population of cells [Section 8]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

such that after PCR the amplified products, but not the 3C template itself, are equipped with the up- and downstream (P5/P7) Illumina sequencing adapters and thus require no further library preparation for sequencing. Following Illumina sequencing, the captures are identified and quantified (Fig. 1C). In this manner, 4C-seq can generate high resolution contact profiles based on the analysis of less than one million sequencing reads. 4C-seq technology is widely used and has been successfully employed to identify regulatory and architectural chromatin loops [18,19], the formation of neo-TADs and the spatial principles of enhancer hijacking in disease [20,21] as well as the clustering of genomic regions with similar chromatin properties such as super enhancers in nuclear space [10,13,22].

In this chapter, a step-by-step protocol is provided for the generation of 4C sequencing libraries, including optimized PCR and purification strategies. In addition, we present a data processing pipeline that can handle multiplexed 4C reads directly from FASTQ files and that outputs files in a range of widely used formats in order to facilitate visualization and further data analysis using standard genome browsers and tools, including our recently developed peak caller for 4C-seq data peakC [23].

2. Materials

2.1. Reagents

Product	Supplier	Catalog number
Fetal bovine serum (FBS)	Sigma Aldrich	F7524
Phosphate buffered saline (PBS) pH 7.3		
Formaldehyde	Sigma Aldrich	1.03999.1000
Glycine	J.T.Baker	0582-05
UltraPure™ Tris Buffer	Invitrogen	15504-020
NaCl	Boom	76028327-5000
EDTA	Invitrogen	15576028
NP-40	Sigma Aldrich	74385
Triton X-100	Sigma Aldrich	T8787
Sodium Dodecyl Sulphate (SDS)	MP biomedicals	811030
cOmplete, EDTA-free Protease Inhibitor	Sigma Aldrich	11873580001
DpnII (10,000 units/ml)	New England Biolabs	R0543
NlaIII (10,000 units/ml)	New England Biolabs	R0125
Csp6I (10,000 units/ml)	Thermo Scientific	ER0211
Proteinase K	Roche	3115801001
T4 DNA Ligase	Roche	10799009001

MgCl2	Sigma Aldrich	M2670
DTT	Sigma Aldrich	43816
ATP	Sigma Aldrich	A2383
Nucleomag 96 PCR Beads (P-Beads)	Macherey Nagel	744100.34
2-Propanol	Boom	76025379
Ethanol	Boom	84028185
Qubit DNA Broad Range	Invitrogen	Q32853
Expand Long Template PCR system	Roche	11759060001
dNTPs	Roche	3622614001
AMPure-XP beads	Beckman Coulter	A63881

2.2. Solutions and buffers

1. 1 M Glycine: Dissolve 7 g of Glycine and complete this to 100 ml with Milli-Q.
2. Isolation buffer (10% (vol/vol) FBS/PBS): Add 4 ml FBS to 36 ml PBS. Prepare fresh and keep at room temperature (RT).
3. Fixation buffer (4% (vol/vol) formaldehyde/Isolation buffer): Add 2 ml 37% (wt/vol) formaldehyde to 16.5 ml Isolation buffer. Formaldehyde is toxic. Handling should take place inside a fume cabinet. Prepare fresh and keep at RT.
4. Cell lysis buffer: 50 mM Tris-HCl pH 7.5, 0.5% (vol/vol) NP-40, 1% (vol/vol) TX-100, 150 mM NaCl, 5 mM EDTA and 1× protease inhibitors. Prepare using Milli-Q. Prepare fresh and keep on ice.
5. 1.2× RE Buffer: Add 120 µl 10× RE Buffer to 880 µl Milli-Q.
6. 10× ligation buffer: 660 mM Tris-HCl, pH 7.5, 50 mM MgCl2, 50 mM DTT, and 10 mM ATP. Prepare using Milli-Q. Aliquots can be stored at −20 °C. Prevent freeze thawing.

2.3. Equipment

Product	Supplier	Catalog number
Tube roller	Stuart	SRT6D
Swing-out centrifuge	Eppendorf	5810R
Eppendorf centrifuge	Eppendorf	5424R
Thermomixer	Eppendorf	Compact/F1.5
Water bath	Grand	JBAqua 5
50-ml magnetic separation	Invitrogen	DynaMag™-50
2-ml magnetic separation	Invitrogen	DynaMag™-2
DNA quantifier	Invitrogen	Qubit 2.0
Spectrophotometer	Thermo Scientific	NanoDrop One
PCR machine	Biorad	S1000

2.4. Software

1. A Unix like shell (e.g. Bash v3.2 +)
2. Illumina base-calling software (bcl2fastq) available from http://www.illumina.com/.
3. The 4C-seq pipeline can be downloaded from https://github.com/deLaatLab/.
4. Bowtie2 v2.3 + available from http://bowtie-bio.sourceforge.net/bowtie2/.
5. SAMtools v1.3 + available from http://www.htslib.org/.
6. R v3.5 + available from https://www.r-project.org/.
7. The following R packages available from CRAN:
   - optparse
   - caTools
   - config
8. The following R packages available from Bioconductor:
   - shortRead
   - genomicRanges
   - genomicAlignments
   - BSgenome of interest
9. The peakC package available from https://github.com/deWitLab/peakC/.

2.5. Primers for second round PCR

Primers required for the second round PCR are indicated in table 1 and should be ordered as standard desalted PCR primers without any chemical modifications. The universal forward primer should be mixed in equimolar amounts with each individual index primer and stored as 5 µM aliquots at − 20 °C to prevent freeze thawing and cross-contamination. Primer names correspond to the 6 nucleotide (nt) TruSeq index present in the primer. Indexes that will be pooled within one sequence lane should be carefully selected to have enough sequence complexity as described in Illumina's Index Adapters Pooling Guide.

3. Experimental design

Before starting a 4C-seq experiment several choices need to be made:

3.1. Selection of the cell type of interest

In principle, any cell type can be analyzed using the 4C-seq method, as long as one can obtain sufficient numbers of cells: preferably 5–10 million cells are included, but good results can be obtained with a few hundred thousand cells. Reason to include a large number of cells is that genome topologies are variable between cells and between the individual alleles of single cells and that each diploid cell can only contribute a maximum of two proximity ligated DNA fragments (one for each allele) to the contact profile of a given viewpoint fragment [24,25]. Inverse PCR to amplify the ligation partners of the viewpoint therefore needs to be performed on the equivalent of hundreds of thousands of genomes, in order to generate contact profiles that are reproducible and that expose the most dominant conformations across the cell population.

3.2. Primary restriction enzyme choice

The resolution of a 4C-seq experiment is largely determined by the digestion frequency of the primary restriction enzyme that is used to digest the crosslinked chromatin. While restriction enzymes with 6-bp specificity have been used successfully to detect far-*cis* (> 5 MB from the viewpoint) and inter-chromosomal contacts [10,13,22], the resolution provided by these enzymes (~4 kb) is not suitable for studying near-*cis* contacts such as loops between regulatory elements. In contrast, restriction enzymes recognizing 4-bp motifs enable robust identification of specific interactions between regulatory elements [18] due to the more than tenfold increase in the pool of fragments ends that can be analyzed. As contact domains [20] and far-*cis* contacts can also be detected using 4-bp cutters (See Fig. 6), we recommend using a 4-bp cutter as the primary enzyme. DpnII, NlaIII, Csp6I and MboI are all 4-bp cutters that efficiently cleave crosslinked DNA. The restriction motif distribution around the genomic region of interest usually determines which of these enzymes should be used.

3.3. Viewpoint selection

The restriction fragment that is chosen as a viewpoint (VP) (see Fig. 2A) should overlap or be in close (< 1 kb) linear distance to the genomic site (e.g. promoter, enhancer, single nucleotide variant, genome-edited site) to be analyzed. We preferably select fragments that also contain a recognition site for the secondary restriction enzyme that will be used (the presence of such secondary RE motif makes this a "non-blind" VP). Reason to prefer this configuration is that "blind" VPs are dependent on multiple secondary RE motifs in ligated fragments to trim their circularized ligation products and make subsequent PCR efficient. A further consideration is the length of the VP fragment end, which is the part of the viewpoint fragment between the primary and secondary RE site to which the 4C primers hybridize. This is preferably

Table 1
Primers for the second round PCR. The universal forward primer contains the P5 adapter end that binds to the flowcell (mint green) and the Illumina Truseq read 1 sequencing primer hybridization site (purple). The reverse primers consist of the P7 adapter end that bind to the flowcell (blue), a 6 nt Truseq index (yellow) and the Truseq index sequencing primer hybridization site (green). The sequences that hybridize to the first round amplicons are underlined. The extra two nucleotides (black) present in the reverse primers 13–27 are also present in the original Truseq adapters 13–27, but are not used for indexing.

Primer name	Sequence (5′-3′)
2ndRound_Universal_fw	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
2ndRound_rv_index1	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index2	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index3	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index4	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index5	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index6	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index7	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index8	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index9	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index10	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index11	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index12	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index13	CAAGCAGAAGACGGCATACGAGATTGTTGACTGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index14	CAAGCAGAAGACGGCATACGAGATACGGAACTGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index15	CAAGCAGAAGACGGCATACGAGATTCTGACATGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index16	CAAGCAGAAGACGGCATACGAGATCGGGACGGGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index18	CAAGCAGAAGACGGCATACGAGATGTGCGGACGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index19	CAAGCAGAAGACGGCATACGAGATCGTTTCACGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index20	CAAGCAGAAGACGGCATACGAGATAAGGCCACGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index21	CAAGCAGAAGACGGCATACGAGATTCCGAAACGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index22	CAAGCAGAAGACGGCATACGAGATTACGTACGGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index23	CAAGCAGAAGACGGCATACGAGATATCCACTCGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index25	CAAGCAGAAGACGGCATACGAGATATATCAGTGTGACTGGAGTTCAGACGTGTGCT
2ndRound_rv_index27	CAAGCAGAAGACGGCATACGAGATAAAGGAATGTGACTGGAGTTCAGACGTGTGCT

more than 200 bp to allow efficient circularization during the second ligation step.

3.4. Primer design for two-step PCR strategy

Instead of a one-step PCR procedure [26,27], we nowadays use a two-step PCR strategy to simultaneously amplify the proximity ligated fragments and provide them with adapters for direct sequencing on the common Illumina next generation sequencing platforms (See Fig. 2B and Section 5).

Advantage of this two-step PCR strategy is that it avoids the use of ultralong PCR primers. We frequently noticed that one-step PCR primers which contain adapters that are compatible with the newer generation of Illumina machines (e.g. Nextseq, Miniseq) can affect PCR efficiency. In addition, the two-step PCR strategy allows multiplexed PCR from dozens of viewpoints on the same template (economical use of template) (unpublished data) and enables simultaneous sequencing of indexed experiments performed with the same viewpoint on different 4C templates [9].

The first PCR step is an inverse PCR with VP specific primers that are designed outward on the VP. These primers each carry a small 5′ overhang that represents the 3′end of the TruSeq universal adapter sequence or the 5′end of the TruSeq indexed adapter sequence as well as part of the Illumina sequencing primer hybridization sites. In the second step universal primers are used that hybridize to these overhangs and that introduce the complete Illumina sequencing primer hybridization sites for single and pair-end sequencing, an index and the Illumina adapters required for flow cell binding of the amplicons (P5/P7), such that after this round of PCR the amplicons are ready for Illumina sequencing (See Fig. 2).

Inverse PCR primer pairs are designed using the following recommendations:

1. For standard 4C-seq experiments Single-Read (“Single-End”) sequencing is used to read and identify the fragment ends ligated to the primary restriction site of the VP. As the sequence read starts with the complete primer sequence, the “reading primer” should be designed on the viewpoint as close as possible to its primary restriction site to enable reading sufficiently far into the captured fragment for its identifiability (uniqueness) in the genome (See Section 7). We typically design the primer within 35 nt (including the primer) of the primary restriction site, to have at least 40 nt left (the capture length) for mapping when sequencing 75 nt reads (See Fig. 2C and D).
   The reading primer is designed with a 5′ overhang that serves as the hybridization sequence for the second round primer and optionally a spacer sequence to enable out of phase sequencing of low sequence diversity libraries that otherwise compromise Illumina sequencing (see also Section 6). While spacer sequences can be used as barcodes to separate experiments that analyze the same viewpoint [26,27], we prefer to separate libraries based on the Illumina index present in the second round primers (see Table 1).
   Reading primer: 5′-TACACGACGCTCTTCCGATCT –Spacer sequence (0–5 nt) – VP specific primer sequence – 3′
2. The primer hybridizing closest to the secondary restriction site (the non-reading primer) is designed preferably at a distance of ≤50 bp

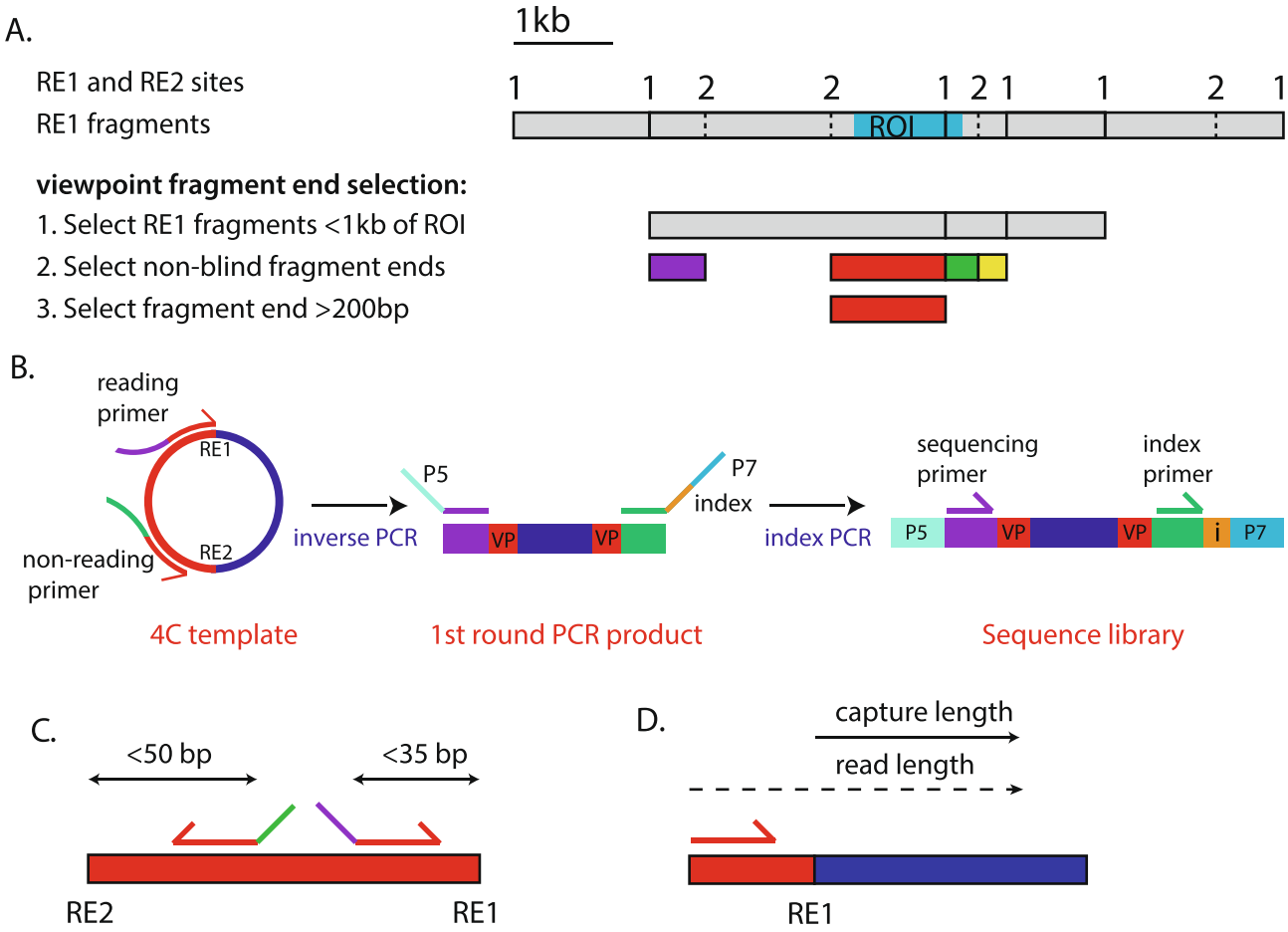

Fig. 2. VP selection and primer design; (A.) Viewpoint fragment end selection criteria. The viewpoint should overlap or be in close linear distance (<1 kb) to the genomic region of interest (ROI, indicated in light blue). We prefer non-blind VPs (fragments containing both the primary (RE1) and secondary enzyme (RE2) motif) with a fragment end that is >200 bp. (B.) The 4C-seq library preparation involves two PCR steps. The first PCR step is an inverse PCR that amplifies the fragments ligated to the VP (dark blue). VP specific primers (red) are designed outward on the VP and carry a 5′overhang (purple and green) that contain part of the Illumina sequencing primer hybridization sites present in TruSeq adapters and that acts as the hybridization sequence for the second round primer. In the second step Illumina adapters (P5 (mint green) and P7 (blue)), an index (yellow) and the complete sequencing primer hybridization sites are added to the amplicons using universal primers to allow direct Illumina sequencing and multiplexing of multiple 4C templates with the same VP primers on the same flowcell. (C.) The reading primer should be designed as close as possible (within 35 nt when using a sequencing read length of 75 nt) of the primary RE site (RE1), while the non-reading primer should preferably hybridize within 50 nt of the secondary RE (RE2). (D.) The capture length is the length of the sequence read part that overlaps with the capture including the RE sites. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

from the secondary restriction site to keep the PCR products as short as possible. The primer is designed with a 5′ overhang that allows hybridization of the 2nd round primer.
Non-reading primer: 5′-ACTGGAGTTCAGACGTGTGCTCTTCCGA TCT- VP specific primer sequence – 3′

3. Primer3 [28] is used to find the optimal primer pair for a given viewpoint fragment, with the following adaptations to the default settings: optimal temperature of 55 °C, minimum of 50 °C and maximum of 65 °C; GC content between 35 and 65%.
4. Primers need to be checked for specificity, for example using NCBI Blast [29].

3.5. Biological replicates

It is recommended to include biological replicates. A given far-*cis* (>5 MB from the viewpoint) or inter-chromosomal contact is usually very rare in the cell population, hence such individual contacts are generally not reproducible between replicate experiments. However, such contacts may reflect the nuclear compartment that is preferentially visited by the TAD that contains the VP, which can be exposed by analyzing capture frequencies across large genomic intervals (>100 kb), and asking whether such genomic regions reproducibly co-localize with the viewpoint [26,27] (See Section 8). In contrast, near-*cis* contacts (<1 Mb from the viewpoint) occur much more frequently, should be quantitatively reproducible and therefore interpretable at much smaller genomic intervals. By quantitatively assessing the average capture frequencies across such small genomic windows one can use 4C-seq to identify preferential chromatin loops such as between promoters and enhancers and between CTCF sites [23] (see note 1).

4. 4C template preparation

4.1. Crosslinking of cells

We start our 4C template preparation preferably with 5–10 million cells, as these amounts best guarantee reproducible DNA yields and allow the analysis of multiple VPs per template. When performing 4C-seq with less than 5 million cells or more than 10 million cells, the protocol should be adapted accordingly.

1. Resuspend cells or tissue of interest as single-cell suspension in isolation buffer at 2×10^6 cells per ml (e.g., 5 ml for 1×10^7 cells) at room temperature. (See note 2).
2. Add an equal volume of freshly prepared 4% fixation buffer (5 ml for

1×10^7 cells) (2% final concentration formaldehyde) and mix by inverting.
3. Incubate at room temperature for 10 min on a tube roller.
4. Immediately add cold 1 M glycine to a final concentration of 0.13 M to quench the reaction (1.5 ml for 1×10^7 cells) and transfer the cells to ice.
5. Immediately centrifuge for 5 min at 500*g* (4 °C) and remove all supernatant.
6. Resuspend pellet in 1 ml of cold PBS, transfer to 1.5 ml Safe-Lock tubes and spin for 5 min at 500*g* (4 °C).
7. Discard supernatant and continue with lysis (see Section 4.2) or flash-freeze pellets in liquid nitrogen. Frozen nuclei can be stored at −80 °C for months.

4.2. Lysis

1. Gently resuspend the pellet in 1 ml of freshly prepared cold cell lysis buffer. Incubate for 20 min on ice (See note 3 and 4).
2. Centrifuge for 5 min at 500*g* (4 °C).
3. Carefully remove the supernatant and resuspend the nuclei in 450 µl 1.2× RE1 buffer.
4. Centrifuge for 5 min at 500*g* (4 °C).
5. Carefully remove supernatant and resuspend the nuclei in 500 µl 1.2× RE1 buffer.
6. Warm up sample to 37°C in a thermomixer and add 15 µl 10% SDS (final, 0.3%).
7. Incubate 1 h at 37 °C while shaking at 750 rpm (See note 5 and 6).
8. Add 75 µl 20% Triton X-100 (final, 2.5%). If nuclear aggregates are present resuspend sample by gentle pipetting until most of the aggregates disappear.
9. Incubate 1 h at 37 °C while shaking at 750 rpm.
10. Take a 5 µl aliquot of the sample and store as the "undigested" control at 4 °C.

4.3. First restriction enzyme digestion

1. Add the primary restriction enzyme (100U for DpnII, NlaIII and Csp6I) and incubate for ~3 h at 37 °C while shaking at 750 rpm.
2. Add a second round of primary restriction enzyme and incubate overnight at 37 °C while shaking at 750 rpm.
3. Take a 5 µl aliquot of the sample as the digested control.
4. Determine the digestion efficiency:
   - 4.1. Add 40 µl 10 mM Tris-HCl pH 7.5 and 5 µl Prot K (10 mg/ml) to both the undigested and digested control.
   - 3.1. Incubate for 1 h at 65 °C while shaking at 500 rpm.
   - 4.2. Add loading dye and load 20 µl of each control alongside each other on a 0.6% (wt/vol) agarose gel.
   - 4.3. The DNA from the undigested control should run as a high molecular weight band (>12 kb). The type of smear of the digested control depends on the genome and RE used. When using Csp6I, DpnII or NlaIII and human or mouse cells most DNA should be below the 10 kb range (with the majority < 5 kb), with DNA digested with NlaIII running lower than DNA digested with DpnII and Csp6I. A high discrete band should be visible when digesting mouse DNA with DpnII due to the absence of recognition sites in the mouse pericentromeric repeats.
   - 4.4. If digestion is of good quality (See Fig. 3), proceed with the first ligation (Section 4.4). Otherwise repeat steps 4.3.2–4.

4.4. First ligation

1. Heat inactivate enzyme as recommended in the manufacturer's instructions. For DpnII, NlaIII and Csp6I inactivate the enzyme by incubating for 20 min at 65 °C.
2. Transfer the samples to a 50-ml centrifugation tube and add 700 µl 10X ligation buffer, Milli-Q to 7 ml and 50 U Ligase and incubate o/n at 16 °C (See note 7 and 8).
3. Determine the ligation efficiency:
   - 3.1. Take a 40 µl aliquot of the sample as the "ligated control".
   - 3.2. Add 5 µl Proteinase K (10 mg/ml) and incubate for 1 h at 65 °C while shaking at 500 rpm.
   - 3.3. Add loading dye and load 20 µl of the digested control and the ligated control alongside each other on a 0.6% (wt/vol) agarose gel.
   - 3.4. A clear upwards shift in molecular weight should be visible (See Fig. 3).
   - 3.5. If ligation is of good quality, proceed with de-crosslinking (Section 4.5). Otherwise add fresh ATP (final, 1 mM) and repeat steps 4.4.2–3.

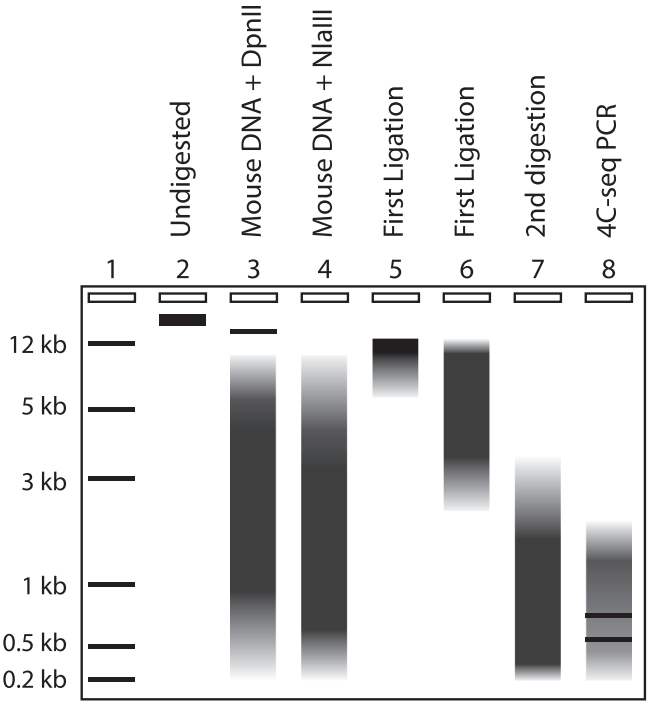

Fig. 3. Quality controls of the intermediate steps of the 4C-seq protocol. Schematic representation of (1.) DNA marker. (2.) The Undigested control. The DNA from the undigested control should run as a high molecular weight band (>12 kb). (3–4.) First digestion with a 4-cutter enzyme. Most DNA should run below the 10 kb range (with the majority <5 kb). A high discrete band should be visible when mouse DNA is digested with DpnII due to the absence of recognition sites in the mouse pericentromeric repeats (lane 3), but not when using NlaIII (lane 4). (5–6.) First Ligation. A clear upwards shift in molecular weight should be visible. A partial ligation (lane 6) can still result in a good 4C-seq result. (7.) Second digestion. The majority of the digested DNA should be below 2.5 kb. (8.) The 4C-seq PCR should result in a DNA smear. The undigested and self-ligation fragments are often strongly amplified in 4C and visible as strong bands.

4.5. Reversal of cross-links and purification

1. Add 30 µl Prot K (10 mg/ml) and incubate o/n at 65 °C.
2. Purify the template using Nucleomag PCR beads ("P-beads") (See note 9).
   - 2.1. Mix the P-bead reagent well until the solution appears homogenous and no pellet is visible.
   - 2.2. Add 1 volume (7 ml for 7 ml ligation sample) isopropanol and 1:100 volume (70 µl for 7 ml ligation sample) P-beads.
   - 2.3. Incubate for at least 30 min on a tube roller at RT.
   - 2.4. Place the tube on a 50-ml magnetic separation rack and remove the supernatant when the solution becomes clear.

2.5. Remove the tube from the magnet and add 20 ml 80% ethanol to resuspend the beads.
2.6. Put the tube on the magnet and remove the supernatant when the solution becomes clear.
2.7. Remove the tube from the magnet and add 20 ml 80% ethanol to resuspend the beads.
2.8. Put the tube on the magnet and remove all but 4 ml of the supernatant when the solution becomes clear.
2.9. Remove the tube from the magnet and resuspend the beads in the remaining 4 ml of ethanol.
2.10. Transfer 2 ml of the ethanol/beads suspension to a 2-ml Safe-Lock tube.
2.11. Place the 2-ml tube in a 2-ml magnetic separation rack and remove the supernatant when the solution becomes clear.
2.12. Transfer the remaining ethanol/beads solution to the 2-ml tube and wait until the solution becomes clear.
2.13. Use 1 ml cleared solution and use it to rinse the 50-ml tube and transfer the remaining beads to the 2-ml tube.
2.14. Remove supernatant when the solution becomes clear.
2.15. Remove the tube from the magnet and centrifuge for a few seconds at 1000*g* to pellet the beads.
2.16. Place the tube back on the magnet and remove any remaining ethanol.
2.17. Air-dry the beads for 10 min at RT on the magnet.
2.18. Resuspend the beads in 300 µl 5 mM Tris-HCl pH 8.0 and incubate at 55 °C for 10 min.
2.19. Add 150 µl Milli-Q to the tube and place the tube back on the magnet for 1 min and transfer supernatant to a new 2-ml tube.
2.20. Place the tube back on the magnet for at least 15 min and transfer supernatant to a new 1.5-ml safe lock tube. Repeat this step until no beads are visible during the transfer step.
2.21. Continue with the second RE digestion (Section 4.6) or store at −20 °C as "3C template".

4.6. Second RE digestion

1. To ~450 µl of the 3C template add 50 µl 10× RE2 buffer and 50 units secondary RE.
2. Incubate o/n at 37 °C while shaking at 500 rpm.
3. Determine digestion efficiency:
   3.1. Take a 2.5 µl aliquot of the sample as the "2nd digestion control" and add 15 µl 10 mM Tris-HCl pH 7.5.
   3.2. Add loading dye and load 20 µl of the 2nd digested control and the ligated control alongside each other on a 0.6% (wt/vol) agarose gel.
   3.3. A clear downward shift in molecular weight should be visible (Fig. 3). The majority of the smear should be < 2.5 kb.
   3.4. If digestion is of good quality, proceed with the second ligation (Section 4.7). Otherwise repeat steps 4.6.1–4.6.3.

4.7. Second Ligation

1. Heat inactivate enzyme as recommended in the manufacturer's instructions. For DpnII, NlaIII and Csp6I inactivate the enzyme by incubating for 20 min at 65 °C.
2. Quantify the amount of template using the Qubit dsDNA BR Assay Kit, following the manufacturer's instructions.
3. Transfer the samples to a 50-ml centrifugation tube and perform a ligation reaction with a final DNA concentration of 5 ng/µl. For 25 µg of template, add 500 µl 10X ligation buffer, 50 U Ligase and Milli-Q up to 5 ml. Incubate overnight at 16 °C (note 10)

4.8. Purification of 4C template

1. Purify the second ligation using P-beads as described in Section 4.5.2.
2. Quantify the amount of template using the Qubit dsDNA BR Assay Kit, following the manufacturer's instructions.
3. Continue with the inverse PCR (Section 5) or store samples at −20 C.

5. 4C-seq PCR

5.1. Primer test

VP specific PCR primers designed according to our recommendations (Sections 3.3 and 3.4) should first be tested using a single PCR to determine their efficiency and to test the 4C template quality (see note 11). Ideally, primers with proven efficiency are included in this test as a positive control for comparison.

1. Perform the PCR using the mix and program indicated below.

   PCR mix:
   2.5 µl 10× PCR buffer I
   0.5 µl dNTP (10 mM)
   2.5 µl reading primer (5 µM)
   2.5 µl non-reading primer (5 µM)
   0.35 µl Expand Long Template Polymerase mix
   100 ng of 4C template
   Milli-Q up to 25 µl.

   PCR program:
   94 °C for 2 min
   30 cycles: 94 °C for 10 s; [primer specific]°C for 1 min; 68 °C for 3 min
   68 °C for 5 min
2. Run 15 µl of the PCR product on a 1.5% (wt/vol) agarose gel (See Fig. 3). If primers give a strong DNA smear (in addition to the prominent "self-ligation" and "undigested" bands (see note 12) and show minimal primer dimers continue with the sequence library preparation (step 5.2).

5.2. Sequencing library preparation

The library preparation approach involves two PCR steps (See Section 3.4 and Fig. 2).

Step 1

The first PCR step is performed to amplify the fragments ligated to the VP. We generally perform 4 PCR reactions with 200 ng 4C template per reaction (See note 13).

1. Perform the first PCR step using the reaction mix and program indicated below.

   Reaction mix:

   5.0 µl 10× PCR buffer I
   1.0 µl dNTP (10 mM)
   5.0 µl reading primer (5 µM)
   5.0 µl non-reading primer (5 µM)
   0.7 µl Expand Long Template Polymerase mix
   200 ng of 4C template
   Milli-Q up to 50 µl.

   PCR program:

   94 °C for 2 min
   16 cycles: 94 °C for 10 s; [primer specific]°C for 1 min; 68 °C for 3 min
   68 °C for 5 min
2. Pool all PCR reactions, mix well and transfer a 50 µl aliquot to a new 1.5-ml tube. Store the remaining PCR product at −20 °C.
3. Perform a 0.8× AMPure XP purification on the 50 µl aliquot:
   3.1. Allow AMPure XP beads to come to RT and vortex prior to use.
   3.2. Add 40 µl of AMPure XP beads to the sample, vortex and spin down tubes briefly.

3.3. Incubate 5 min at room temperature.
3.4. Place sample on a magnetic separation rack and remove the supernatant when the solution becomes clear.
3.5. Gently add 500 µl 80% ethanol without disturbing the beads.
3.6. Remove the supernatant when the solution becomes clear.
3.7. Gently add 500 µl 80% ethanol without disturbing the beads.
3.8. Remove the supernatant and shortly pellet the beads on a tabletop centrifuge (a few seconds at 1000*g*).
3.9. Place the tube back on the magnet rack.
3.10. When the beads have moved from the bottom of the tube, remove the residual ethanol.
3.11. Air-dry sample for 30 sec.
3.12. Resuspend the beads in 50 µl 10 mM Tris-HCl pH 7.5
3.13. Incubate 10 min at room temperature while shaking at 600 rpm.
3.14. Shortly centrifuge the bead solution (a few seconds at 1000*g*) and place back on the magnet rack.
3.15. When the solution is clear, transfer 45 µl of the supernatant to a new 1.5-ml tube.

Step 2.

A second round of PCR is performed on the purified PCR product obtained after round 1. In this second PCR, universal primers are used that contain the Illumina adapters required for flow cell binding of the amplicons, a 6-nt index sequence and the Illumina sequencing primer sequences for single and pair-end sequencing (See Section 2.5 for primer sequences). Select indexes carefully to ensure optimum base calling and demultiplexing by having different bases at each cycle of the index read, as described in Section 6.

4. Perform the second PCR step using the reaction mix and program indicated below.

Reaction mix:

5.0 µl 10× PCR buffer I
1.0 µl dNTP (10 mM)
5.0 µl primer mix with indexed reverse primer of choice (5 µM) (see Section 2.5)
0.7 µl Expand Long Template Polymerase mix
5.0 µl purified round 1 pcr
Milli-Q up to 50 µl

PCR program:

94 °C for 2 min
20 cycles: 94 °C for 10 s; 60 °C for 1 min; 68 °C for 3 min
68 °C for 5 min

5. Clean up using Qiagen PCR purification:
5.1. Add 250 µl Buffer PB to the PCR sample and mix.
5.2. Transfer the sample to a QIAquick column and centrifuge for 30 sec.
5.3. Discard flow-through and place the QIAquick column back into the same tube.
5.4. Add 0.75 ml Buffer PE to the QIAquick column and centrifuge for 30 sec.
5.5. Place the QIAquick column in a new 1.5-ml tube.
5.6. Centrifuge the column for 1 min.
5.7. Place the QIAquick column in a new 1.5-ml tube.
5.8. Add 100 µl elution buffer and wait for 1 min.
5.9. Centrifuge the column for 1 min.
6. Determine the sample quantity and purity using a NanoDrop spectrophotometer. If the A260/A280 absorption ratio is between 1.8 and 2.0 and the A260/230 absorption ratio >2.0 continue with step 7. Otherwise repeat step 5.2.5 and 5.2.6.
7. Run 200 ng of the purified PCR product on a 1.5% agarose gel to determine whether the 4C PCR was successful and no primer dimers are present. Primer dimers can be removed using Ampure XP beads as described in Section 5.2.3.
8. 4C-seq libraries can be stored at −20 °C and can be directly sequenced using an Illumina high-throughput sequencing equipment.

6. Sequencing

As only roughly 1 million reads are required for 4C-seq analysis (see note 14), 4C-seq libraries can be pooled with other sequencing libraries. For simplicity we generally pool 4C-seq libraries based on their Nanodrop values and pool this 4C-seq master library with other sequencing libraries (e.g. ChIP-Seq, Hi-C etc) by pooling equimolar amounts of individual libraries.

Nucleotide diversity is required in the first 4–7 bases of the (first) sequencing read for effective cluster recognition on all Illumina sequencing systems. Nucleotide diversity is also important in the first 25 bases in the (first) sequencing read on all sequencing platforms because this is when phasing/pre-phasing, color matrix corrections, and the pass filter calculations occur. These corrections and calculations are used in base calling and quality score calculations for all cycles in a run for the clusters that pass the filter. As all reads in a 4C-seq library start with the same sequence (the VP primer) and different VP primers might contain the same primary RE sequence at the same nucleotide position, the diversity of the pooled libraries must be determined before sequencing. Low diversity 4C-seq libraries can be sequenced by combining such libraries with high-diversity libraries (ChIP-Seq, Hi-C or a PhiX control). Alternatively, primers can be designed carrying a spacer sequence that forces amplicons to be sequenced out of phase and thereby allows sequencing of low sequence diversity (see Section 3.4).

In addition to the nucleotide diversity in the sequencing read, Illumina recommends pooling indexes with diverse sequences as this optimizes color balance. See Illumina's Index Adapters Pooling Guide for more information.

Depending on the primer design, a Single-Read run with 50-bp (HiSeq) or 75-bp (MiniSeq or NextSeq) read length is generally sufficient for unique identification of captures.

7. Data processing

For the analysis of 4C-seq experiments multiple 4C-seq software packages have been developed in the last decade (see [30,31] for a brief overview). We here describe our current pipeline that processes multiplexed 4C-seq reads directly from FASTQ files. It generates files that enable the direct visualization of results with standard genome browsers and that allow further statistical data analysis, for example to perform peak calling (Fig. 4). In comparison to our previously published pipeline [18], our new pipeline performs read alignment against the complete reference genome in contrast to a reduced genome and is more user friendly (see sup. Fig. 1).

In this section we will explain how to run the pipeline and describe the different modules of the pipeline. We refer the user to the github repository of the pipeline (https://github.com/deLaatLab/pipe4C) for access to the latest version, installation instructions and a step by step tutorial including example files.

7.1. Files required to run the pipeline:

7.1.1. Fastq files

The pipeline requires reads in (compressed) FASTQ format. Illumina Sequencing Systems generate raw data files in binary base call (BCL) format. Illumina offers bcl2fastq conversion software to demultiplex (based on the index used in the non-reading primer) and convert BCL files. If multiple flow cell lanes have been used to sequence the library, a single forward read (read 1) FASTQ file should be created for each index and for each sequence run by combining the FASTQ files for each flow cell lane using a standard "cat" command after BCL conversion or by using the –no-lane-splitting option when running bcl2fastq.

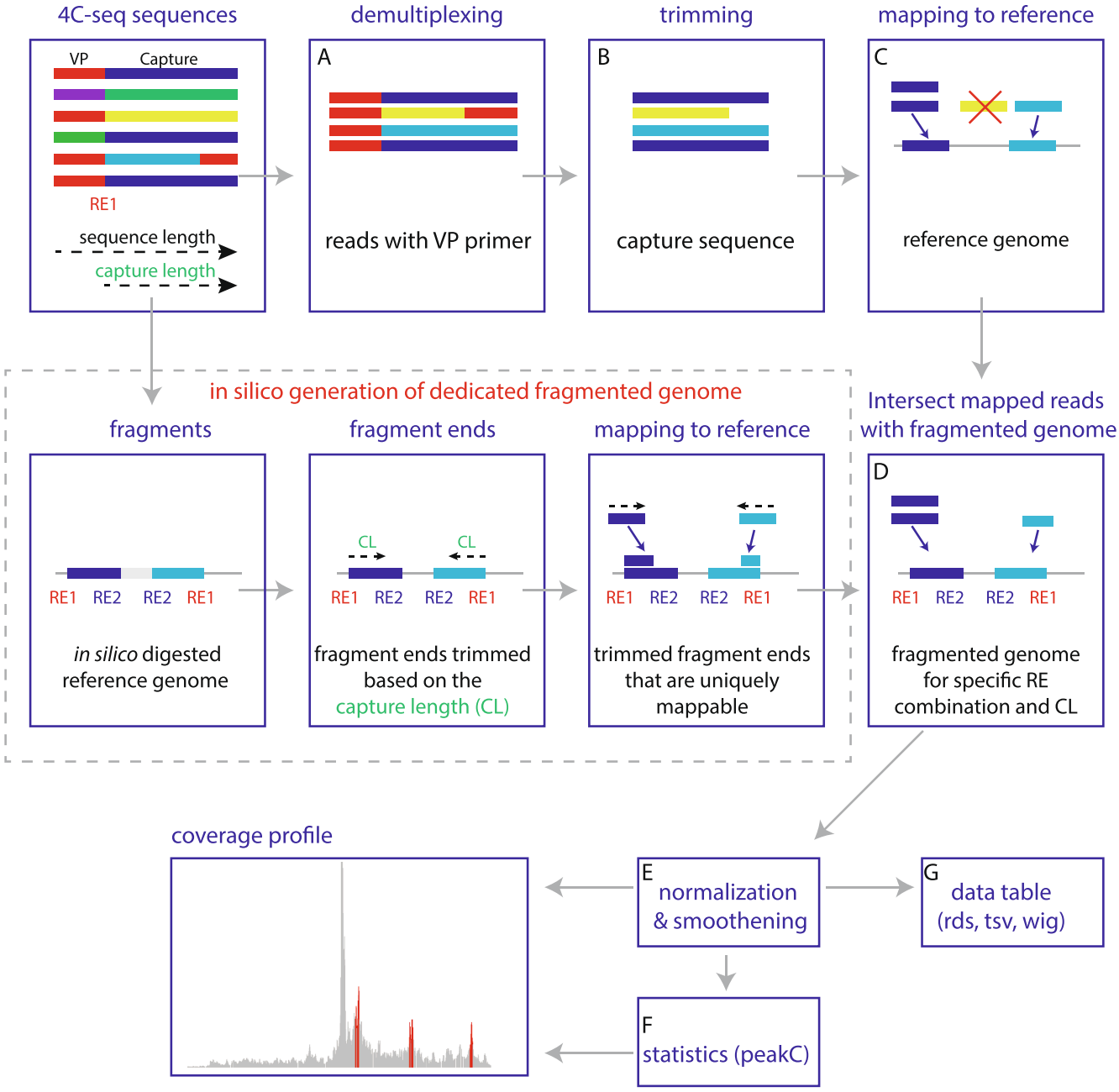

Fig. 4. Schematic overview of the 4C-seq pipeline workflow. The 4C-seq pipeline consists of the following modules: (A.) Demultiplexing of reads; only reads that contain the reading primer (red) are kept. (B.) Trimming of reads; the sequence before RE1 (see text for further explanation) and after RE2 (or a second RE1 in case of a blind fragment) site are trimmed to extract the capture sequence including the RE motifs. (C.) Read mapping to the reference genome using Bowtie2. Reads that do not map (yellow) are removed. (D.) Read counting per restriction fragment end. Reads are only counted if they are mapped to restriction fragment ends that can be mapped uniquely to the reference genome with the capture length (length of the read without the VP sequence) of the analyzed experiment. To identify the unique fragment ends for each enzyme combination and capture length, the reference genome is digested *in silico* and the fragment ends are mapped against the reference genome. (E.) Read count normalization and smoothing. (F–G.) The pipeline generate files compatible with standard genome browsers for visualization and further statistical analysis of the data such as peak calling using peakC. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

7.1.2. Configuration file

Global and system specific parameters (e.g. paths and genome assemblies installed) that are likely to remain constant across different runs of the pipeline are defined in the global configuration file (conf.yml). In each run the pipeline initially loads the parameters defined in this global configuration file, and then proceeds to load run specific parameters (see Section 7.2) and the experiment specific data defined in a separate viewpoint file (vpFile) (see Section 7.1.3). The global configuration file can be edited using any standard text editor. The list of parameters that need to be assigned at least once upon installation on a system in the global configuration file is shown in Table 2.

7.1.3. Viewpoint file

Experiment specific parameters for all 4C-seq experiments in a single sequencing run are organized together in a viewpoint file. Parameters in this file are stored in a tab-delimited format, with each

Table 2
Description of parameters that need to be defined in the configuration file.

Name	Description
fragFolder	Path to the folder containing the fragment end libraries of the reference genomes
normalizeFactor	Reads mapped to the 4C fragment end library are normalized to account for sequencing depth according to the normalizeFactor
enzymes	Enzyme names used in the viewpoint file and their corresponding recognition motifs
genomes	Genome names used in the viewpoint file plus corresponding BSgenome packages
bowtie2	Path to corresponding bowtie2 index of reference genome. The reference genome assembly used to generate the index should match to the reference genome that was used to generate the BSgenome
maxY	Maximal Y value in local 4C *cis* plot
plotView	Number of bp to plot around viewpoint in local 4C *cis* plot
xaxisUnit	X-axis unit (Mb, Kb or bp)
plotType	Plots will either be saved as PDF or PNG
binSize	Genome bin size used in the genome plot
qualityCutoff	Q-score. Trim 3′-end of all sequences using a sliding window as soon as 2 out of 5 nucleotides have quality encoding less than the Q-score. 0 = no trimming
trimLength	Trim reads to defined capture length from 3′-end. 0 = no trimming
minAmountReads	Minimum required amount of reads containing the primer sequence. If less reads are identified the experiment will not be further processed
readsQuality	Bowtie2 minimum required mapping quality score for mapped reads
mapUnique	Extract uniquely mapped reads, based on the lack of XS tag
cores	Number of CPU cores for parallelization
wSize	The running mean window size
nTop	Top fragment ends discarded for calculation of normalizeFactor
nonBlind	Only keep non-blind fragments
wig	Create wig files for all samples
plot	Create viewpoint coverage plot for all samples
genomePlot	Create genomeplot for all samples (only possible if analysis is “all” in vpFile)
tsv	Create tab separated value file for all samples
bins	Count reads for binned regions
mismatchMax	The maximum number of mismatches allowed during demultiplexing

Table 3
Example of a viewpoint file in which two experiments are demultiplexed from the same FASTQ file based on their primer sequence.

expname	primer	firstenzyme	secondenzyme	genome	vpchr	vppos	analysis	Fastq
mESC_Sox2	GAGGGTAATTTTAGCCGATC	DpnII	Csp6I	mm9	3	34547661	all	index1.fastq.gz
mESC_Mccc1	TTGCACCCGTCTTCTTGATC	DpnII	Csp6I	mm9	3	35873313	cis	index1.fastq.gz

Table 4
Description of parameters that are required in the viewpoint file for processing a 4C-seq experiment.

Name	Description
expname	Unique experiment name
primer	Primer sequence
firstenzyme	First restriction enzyme name (nearest to reading primer)
secondenzyme	Second restriction enzyme name
genome	Reference genome of interest
vpchr	The chromosome that contains the viewpoint (See note 15)
vppos	Coordinate of viewpoint position. Any bp position within the VP can be used except the RE motifs (see note 15)
analysis	The final output tables will contain all reads (all) or only the reads that have been mapped to the VP chromosome (cis). For most analysis cis is sufficient and the generated output files will be smaller and therefore easier to process on local computers
fastq	Name of the FASTQ file
spacer (optional)	Spacer length. Number of nt included as spacer in the primer to enable out of phase sequencing. Default = 0. The spacer sequence will not be used for demultiplexing. If the spacer sequence is used as a barcode include the sequence in the primer sequence and set the spacer length to 0

row containing information for a separate experiment (See Table 3). The list of parameters that are required for each experiment in the 4C-seq pipeline viewpoint file is shown in Table 4.

7.2. Running the pipeline

The pipeline can be run with:

Rscript <path to pipe4C.R script> [any additional arguments].

A list of both required and optional command line parameters that are recognized by the pipe4C.R script are shown in Table 5. Default values (except vpFile, fqFolder, outFolder and confFile) are stored in the configuration file. For example,

Rscript pipe4C.R –vpFile [path to vpFile] –fqFolder [path to folder containing the FASTQ files] –outFolder [path to output folder] –cores 8 –wig –plot –genomePlot

will run the pipeline using 8 cores and generates a wig file, a viewpoint plot and a genome plot as output, next to the default outputs (see Section 7.3).

7.3. Description of the pipeline modules and output

7.3.1. Demultiplexing of reads

As 4C-seq libraries can be pooled with other sequencing libraries using the same Illumina TruSeq index, the reads for each 4C-seq library are first extracted by matching the 5′-ends of the reads to the reading primer sequence. The number of mismatches that are allowed can be set with the *mismatchMax* argument. As output a new FASTQ file will be generated (for each experiment defined in the viewpoint file), that contains only sequences for the corresponding experiment.

Table 5
Description of parameters that are recognized by the pipe4C.R script. *are required.

Name	Description
vpFile*	path to the viewpoint file
fqFolder*	path to the folder containing the FASTQ files
outFolder*	path to the output folder
confFile	path to configuration file – default is conf.yml in folder containing the pipeline script
mismatchMax	The maximum number of mismatches allowed during demultiplexing
qualityCutoff	Q-score. Trim 3′-end of all sequences using a sliding window as soon as 2 out of 5 nucleotides has quality encoding less than the Q-score
trimLength	Trim reads to defined capture length from 3′-end
minAmountReads	Minimum required amount of reads containing the primer sequence. If less reads are identified the experiment will not be further processed
readsQuality	Bowtie2 minimum required mapping quality score for mapped reads
mapUnique	Extract uniquely mapped reads, based on the lack of XS tag
cores	Number of CPU cores for parallelization
wSize	The running mean window size
nTop	Top fragment ends discarded for normalization
nonBlind	Only keep non-blind fragments
wig	Create wig files for all samples
plot	Create viewpoint coverage plot for all samples
genomePlot	Create genomeplot for all samples (only possible if analysis is “all” in vpFile)
tsv	Create tab separated value file for all samples
bins	Count reads for binned regions

7.3.2. Trimming of reads

All demultiplexed reads are first scanned from the 5′ side for the presence of the first primary RE motif: we score the percentage of reads having this motif at each given position. Typically, in a good 4C experiment in which there is no a-specific amplification from undesired genomic locations, most reads (typically > 90%) will have this motif at the same given read location: we report this percentage in our report file as “*motifPosperc*”. Typically, this most frequently found location corresponds to the distance spanned between the start of the primer and the start of the primary RE motif as found on the viewpoint fragment. We subsequently trim all reads from their 5′-end until this fixed location, leaving the RE motif intact. In our pipeline we provide the option to also trim the reads from the 3′-end by setting (a) the *qualityCutoff* argument (see Table 5, to trim off poor quality sequence ends based on the base call quality information present in the fastq file) and (b) the *trimLength* argument, which enables trimming all sequences to a fixed length. Generally we don’t use these trimming features, but in case spacer sequences of different lengths are used as barcodes to compare the same viewpoint between conditions [26,27], sequences can be trimmed to a fixed length to minimize variation between experiments. The most frequent read length which is determined after this 5′end (and optionally 3′end) trimming defines the “capture length”, a parameter that we use to select the most optimal read length of the *in silico* restricted genome which we map against (see below). Before mapping, we standardly further trim the sequence reads containing a second primary or secondary RE motif from their 3′-end until this second RE motif. As a consequence of trimming, the reads to be mapped typically have a primary RE recognition motif at their 5′end and are either untrimmed from their 3′end or having a primary or secondary RE recognition motif at their 3′end.

7.3.3. Read mapping to the reference genome

The trimmed reads will be aligned to the reference genome sequence indicated in the viewpoint file using Bowtie2 [32]. The mapping quality (Q-score) and number of CPU cores for parallelization can be set (see Tables 4 And 5). If preferred, unique reads can be extracted with the *mapUnique* argument (see note 16).

7.3.4. Read counting on restriction fragment ends

As restriction fragment ends are the smallest mappable units of 4C experiments, the pipeline counts reads per fragment end. The fragment end alignment module first creates an *in silico* fragment end library, next determines which fragment ends map uniquely to the reference genome with the provided capture length and then counts reads that map to such fragment ends. Reads that do not start at the first RE motif are discarded. The fragment end library only needs to be generated once for each enzyme and capture length combination and can be generated beforehand (to speed up the mapping for future experiments) or on the fly and will be stored for future reference in the fragFolder as defined in the global configuration file. For the generation of the fragment end library the BSgenome packages are used. Bioconductor provides support for many commonly used reference genomes (Hsapiens (e.g. hg19, hg38), Mmusculus (e.g. mm9, mm10), Drosophila melanogaster (e.g. dm3, dm6), Drerio (e.g. danRer10) and many more), which can be downloaded from their repositories. In addition, the BSgenome package facilitates the manual compilation of any genome assembly of interest. Prior to its use by our pipeline, the BSgenome package for a reference genome of interest should be installed and specified in the configuration file.

7.3.5. Read count normalization and smoothing

4C-seq measured ligation events give a semi-quantitative reflection of pairwise interaction frequencies with the viewpoint, experimentally influenced by differences in cross-linking, digestion, ligation and PCR efficiencies. To average out these differences and to minimize the contribution of individual overrepresented fragments, contact frequencies across (small) genomic windows of multiple DNA restriction fragments should first be smoothed using a running mean, rather than interpreting the capture frequencies of individual DNA fragments.

In this module, the read counts per fragment end are normalized to account for depth of sequencing of the 4C-seq library by scaling all reads mapped to the chromosome containing the viewpoint. The scaling will assure that the total sum of mapped reads (excluding the reads mapped to the top 2 fragments – unless stated otherwise in the configuration file) will be 1 million (by default, see normalizeFactor, Table 2). Following normalization, the mapped 4C data are smoothed by averaging using a running mean of a number of fragment ends (the ‘window size’, which can be set by the user). As fragments that lack a restriction site for the secondary enzyme (“blind” fragments) have been reported to give a systematically lower read coverage [18] they can be excluded from the analysis by using the *nonblind* argument when running the pipeline or alternatively be removed afterwards during data analysis (see note 17). Depending on the analysis parameter in the viewpoint file, all reads or only the reads from the viewpoint chromosome are kept. The coordinates, blind/non-blind status, read counts, the normalized read counts and the smoothened normalized read counts are written in R-objects and stored as rds files for quick data analysis using R (see Section 8) and if preferred as tab separated value files (by

Table 6
The list of data quality metrics that are provided in the report file.

Name	Description
vpname	Name of the experiment indicated in the vpFile
nReads	Number of reads in the FASTQ file
motifPosperc	Percentage of reads in the FASTQ file in which the first RE motif position is equal to the most occurring first RE enzyme motif position
readlenperc	Percentage of reads in the FASTQ file in which the read length is equal to the most occurring read length
nMapped	Number of reads mapped to the reference genome
nMappedCis	Number of reads mapped to the chromosome containing the viewpoint
nMappedCisPerc	Percentage of reads mapped to the chromosome containing the viewpoint
fragMapped	Number of reads mapped to unique fragment ends
fragMappedCis	Number of reads mapped to unique fragment ends located on the viewpoint chromosome
fragMappedCisPerc	Percentage of reads mapped to unique fragment ends located on the viewpoint chromosome
fragMappedCisCorr	Number of reads mapped to unique fragment ends located on the viewpoint chromosome excluding the reads mapped to the top fragment ends
fragMappedCisPercCorr	Percentage of reads mapped to unique fragment ends located on the viewpoint chromosome excluding reads mapped to the top fragment ends
nCaptures	Number of fragment ends with at least 1 mapped read
nCisCaptures	Number of fragment ends located on the viewpoint chromosome with at least 1 mapped read
topPct	Percentage of reads mapped to the nTop most captured fragment ends
capt100Kb	Percentage of all unique fragment ends within 100 kb of the viewpoint with at least 1 mapped read
cov100Kb	Percentage of reads mapped to unique fragment ends within 100 kb of the viewpoint
capt1Mb	Percentage of all unique fragment ends within 1 Mb of the viewpoint with at least 1 mapped read
cov1Mb	Percentage of reads mapped to unique fragment ends within 1 Mb of the viewpoint

using the *tsv* argument) for data analysis using any software/tool of choice.

7.3.6. Report file

A report file is generated that contains quality metrics of all the 4C-libraries processed by the pipeline. In addition the quality metrics of each individual experiment are written to the rds file containing the read counts (Table 6).

7.3.7. Viewpoint plot

A coverage plot of the viewpoint region is generated based on the normalized and smoothened data. The genomic region that is visualized, the height of the Y-axis, the unit of the X-axis (Mb, kb, bp) and the file type (PDF or PNG) are defined and can be adjusted in the configuration file. In addition the quality metrics are displayed in this figure for a quick impression of the data.

7.3.8. Genome plot

Reads will be binned in defined genomic regions (The bin size can be set in the configuration file, default 25 kb) and visualized as a coverage plot for all chromosomes. This option is for example useful when performing 4C-seq on samples where high coverage is expected on multiple chromosomes such as samples carrying translocations between chromosomes [33] or when performing 4C-seq using a VP for which the integration site is unknown such as a randomly integrated plasmid (Fig. 5).

7.3.9. Wig file

The normalized smoothened data will be written to a wig file for visualization in genome browsers such as the UCSC Genome Browser [34] and IGV [35].

8. Data analysis

8.1. Quality assessment of the data

Quality assessment of individual 4C-seq experiments can be performed based on the generated report file [27]. First the percentage of reads in the FASTQ file in which the first RE motif position is equal to the most occurring first RE enzyme motif position (motifPosperc) should typically be >90%. A lower percentage might indicate off target binding of the primer to the genome (see note 18) or low sequence quality reads. The percentage of reads in the FASTQ file in which the read length is equal to the most occurring read length (readlenperc) is generally >60%, a lower percentage might indicate the sequencing of primer-dimers or low sequence quality reads. Independent of the genomic location of a viewpoint, a successful 4C-seq experiment should always show a strong enrichment of contacts with DNA fragments flanking the viewpoint on the linear chromosome (in particular, the DNA fragments co-occupying the same TAD). Preferably >60% of the reads should map to the chromosome that contains the viewpoint (the *cis* chromosome), with most reads mapping within 1 Mb of the viewpoint (See note 19). These metrics can be heavily skewed by two abundantly present undesired 3C byproducts: the first being caused by circularization ("self-ligation") of the viewpoint fragment (giving reads that map to the other end of the viewpoint fragment) and the second being the consequence of undigested DNA (giving reads that map to its linear neighboring fragment) [see Figs. 2 and 3. and note 12]. Therefore these metrics should be assessed after exclusion of the top fragments (fragMapped*Cis*Corr and cov1Mb). In addition, in a successful 4C-seq experiment the VP should capture many different fragments. We therefore expect that >40%, but preferably >60%, of the mappable fragment ends within 100 kb up- and downstream of the viewpoint to be captured at least once (capt100Kb > 40%). Low-complexity libraries (capt100Kb < 40%) are often the result of bad template quality rather than insufficient sequencing depth and will often not contain enough information to draw reproducible conclusions [23].

8.2. Visual inspection of the profile

Wig files can be used for visualization in genome browsers such as the UCSC Genome Browser [34] and IGV [35] where 4C-seq coverage tracks can be displayed together with genomic annotation and data from for instance ChIP-Seq, RNA-Seq and other experiments. Visual inspection of coverage profiles close to the VP (<1 MB) reveals interaction domains (where 4C-seq coverage is high on average) that overlap with TADs in Hi-C data (see Fig. 6) [20,36], whereas chromosome wide profiles reveal clustering of genomic regions with a chromatin signature similar to the VP locus [10,13,22] (see Fig. 6). In addition, our pipeline produces R-objects stored as rds files, which contain all mapped 4C-seq reads as well as the mapping statistics and viewpoint information relevant to the experiment. This allows for a more thorough interactive analysis and visualization of the data in R. We refer the user to the github repository at https://github.com/deLaatLab/pipe4C for a short tutorial which highlights the basic functionalities.

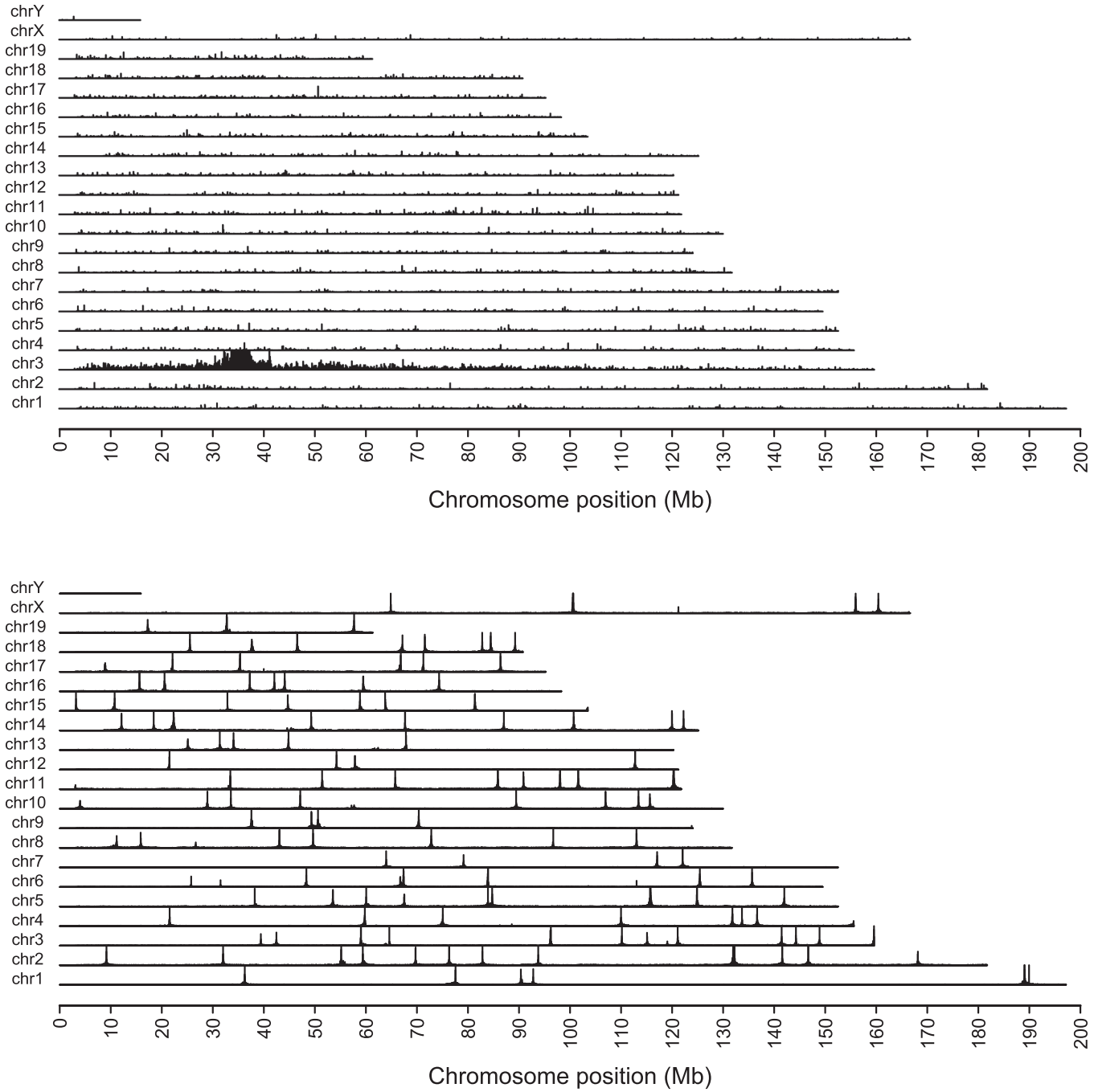

Fig. 5. Genome wide 4C-seq coverage plot (reads per 25-kb genomic bin) of (A.) A 4C-seq experiment using the Mccc1 viewpoint. As expected most reads map to the chromosome containing the VP (chr3). (B.) A 4C-seq experiment in which the viewpoint is located on a plasmid which was randomly integrated many times in the same cell line [38]. Each peak represents an independent integration site.

8.3. Statistical analysis

In order to identify regions that have a higher contact frequency than expected from the surrounding genomic regions we have developed statistical methods for the identification of both near-*cis* and far-*cis* interactions[23,26].

8.3.1. Peak calling for near-cis interactions

To detect preferential chromatin loops such as between CTCF sites [12,19] or between enhancers and promoters [18,23], we have recently developed peakC [23]. PeakC is a method that was designed to identify reproducible peaks in regions close to the viewpoint in 4C-seq data. It computes non-parametric statistics based on ranks of coverage of 4C fragment ends with respect to a background model. This model estimates the background contact frequency of both the region upstream and downstream of the viewpoint independently for each 4C-seq experiment individually and uses a statistical model to identify genomic regions that are significantly contacted.

We provide a function ("doPeakC") in the pipeline to perform peak calling using peakC directly on the rds files produced by the pipeline. While peak identification is possible using single 4C experiments, no significance thresholds can be computed in that case and peak calls will rely on arbitrary thresholds. It is therefore highly recommended to include replicate experiments as this has been shown to yield more reproducible results [23]. Regions identified as peaks can be exported to a bed file for visualization in UCSC and IgV or can be directly visualized

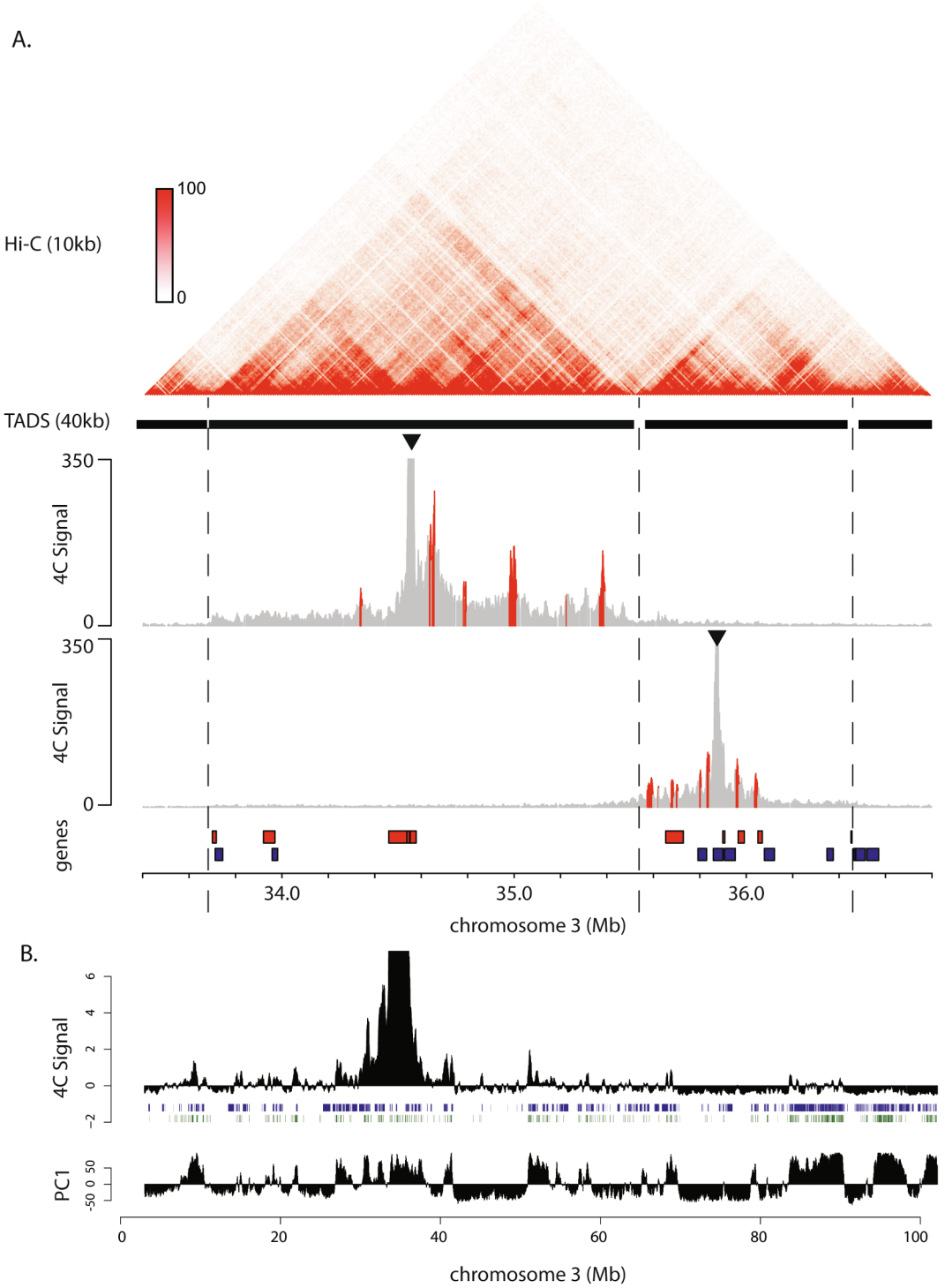

Fig. 6. 4C-seq identifies preferred chromatin loops, contact domains and the clustering of genomic regions with similar chromatin properties. (A.) Visual inspection of coverage profiles close to the VP (< 1 MB) reveal interaction domains that overlap with TADs in Hi-C data. Hi-C contacts maps derived from mouse ESC [29] are shown at 10 kb resolution and have been plotted using the 3D Genome Browser [30]. TADs have been previously called based on the 'Directionality Index' derived from Hi-C data from mouse ESC at 40 kb resolution [31] and are shown as black blocks and borders between them are highlighted with dotted lines. Combined 4C-seq profiles of triplicate experiments [23] are shown for two VPs (Sox2 and Mccc1, indicated as black triangles). Preferential chromatin loops ("peaks") are identified using peakC [17] and are indicated in red. Genes located on the plus strand (red) and minus strand (blue) are shown. (B.) Chromosome wide 4C-seq profiles reveal clustering of genomic regions with a chromatin signature similar to the VP locus. We calculated mean 4C-seq (Sox2 VP) coverage in a running window of 1001 fragment ends and subtracted the 75% quantile of the resulting chromosome-wide 4C-seq coverage vector and plotted this together with genes (blue) and H3K27ac ChIP-seq peaks in mouse ESC (green) [32] (top panel). For comparison, in the bottom panel we plot the first principal component (PC1) obtained after PCA analysis on the correlation matrix of chromosome-wide Hi-C contacts derived from mouse ESC Hi-C data [31]. Positive values indicate the A and negative values the B compartment. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

in R together with the coverage plot (see Fig. 6). We refer the user to the github repository at https://github.com/deLaatLab/pipe4C for an example showing how to use this function and we refer to the peakC manuscript for further explanation of the parameters that can be used.

8.3.2. Far-Cis & trans

An R script to identify far-*cis* and *trans* contacts can be found in Splinter et al. [26]. In contrast to near-*cis* contacts, far-*cis* and *trans* contacts are sparse, are likely to originate from a single ligation event and are therefore binarized (>1 reads per fragment end is set to 1) to avoid possible PCR artifacts to influence results. Enrichment of contacts in a local window of fragment ends *w* is quantified with respect to a larger window *W*. A null model for this enrichment statistic is obtained by permutation of the reads and an FDR threshold is computed from the distribution of z-score normalized values of the statistic in the permuted samples. This normalization assumes that the total number of captures within each window *w* follows a binomial distribution, with success probability *p* determined by the total number of covered fragments in *W*. For far-*cis* analysis, window sizes of w = 100 and W = 3000 are recommended when using HindIII as the primary RE, but should be adjusted when using enzymes recognizing 4 bp motifs to reach similar genomic size bins. For *trans* contacts, a very similar method is used, but since *trans* contacts are even more rare, larger window sizes are required.

9. Notes

1. Because performing replicate experiments is not always possible, for instance due to limited source material, we have also developed a slightly modified version of our peak calling method for single template 4C experiments.
2. When using tissue samples we prepare a single-cell suspension by filtering the cells through a cell strainer. Failing to generate single cells may result in crosslinking differences within the studied population and inefficient lysis.
3. For some cell types, longer incubation is required for digestion to be effective or even may require douncing.
4. When using frozen cell pellets wait until the cells are thawed before resuspension in lysis buffer.
5. The shaking intensity required to prevent cell sedimentation varies per cell type. If sedimentation occurs perform shaking at 900 rpm.
6. Aggregation of nuclei can occur after adding SDS and is cell type dependent. Most of the aggregates will disappear after the addition of Triton X-100.
7. Primary 3C ligations can also be performed in smaller volumes [37], but may require spinning down the nuclei before ligation as described in Rao et al. [12].
8. We use Roche T4 Ligase. Roche and many other companies use Weiss-units. NEB uses Cohesive end ligation unit. 1 CELU = 0,015 weiss units.
9. Alternatively the DNA can be purified using phenol-chloroform or ethanol precipitation as described in [26,30].
10. Unlike the first ligation, the second ligation is performed to circularize the trimmed fragments and therefore the ligation efficiency cannot be determined based on its molecular weight.
11. When performing 4C using primary cells it might be difficult to isolate enough cells to both test the designed primers and perform the experiments. 4C primers can in those cases also be tested using 4C template derived from other cell types, although PCR smears can differ between cell types.
12. Two non-informative 4C circles are frequently generated during template preparation and are caused by circularization ("self-ligation") of the viewpoint fragment or the failure to digest the RE1 of the VP fragment end ("undigested"). These circles often are visible as distinct PCR products on gel and often appear as highly common reads after sequencing. If the size of these fragments are too short for circularization or too long to PCR efficiently these PCR products might not be present.
13. The amount of template input depends on the quality of the template, the VP that is analyzed and the fraction of cells in the population that share contacts. Less input can be used when template material is limited, however using 800 ng of template in our hands gives the most reproducible results for a range of VPs.
14. Subsampling analysis of 4C data shows that roughly 50,000–100,000 *cis*-mapped 4C reads are sufficient to generate reproducible 4C profiles (see sup. Fig. 2). We used the example 4C dataset from mouse ESCs (one single replicate) with a viewpoint in the Sox2 locus for our subsampling analysis. The number of mapped 4C-reads in cis for this experiment was 908426. For a range of different sample sizes (n = 1000, 2000, 5000, 10,000, 20,000, 50,000, 100,000, 200,000 and 500,000), we re-sampled mapped 4C-reads 1000 times with replacement and re-computed the sequence-depth normalized 4C profiles in cis. We then computed spearman's rank correlation coefficient for each possible pair of 're-sampled' 4C profiles around the VP (<1 Mb) (1000 * 999/2 = 499,500 pairs). Note that the number of 'raw' sequence reads required for 4C-seq to obtain these numbers of mapped reads depends on multiple factors such as the template quality and the percentage "self-ligation" and "undigested" reads. For this particular dataset roughly 70% of raw reads remained as mapped in cis after processing with our pipeline. For simplification we routinely aim for ~1 million reads for all our experiments.
15. vpchr and vppos are used to calculate the quality metrics and to draw the viewpoint plot, but do not affect the mapping of the reads. If the genomic position of the VP is unknown, for example when the VP is located on a random integrated plasmid any vpchr and vppos (for example chr1 and 1e6) can be used, and the integration site can be determined by studying genome-wide 4C plots after setting the parameter analysis to "all" in the viewpoint file.
16. We generally run the pipeline with a Q-score of 1 and do not filter for unique reads at this step as using higher Q-scores or filtering for unique reads generally does not improve the 4C result.
17. As blind fragments are captured less frequently and have a lower read coverage compared to non-blind fragments exclusion of blind fragments may improve the local 4C profile.
18. If an off target viewpoint is generated due to nonspecific binding of primers, the primer sequence that is used for demultiplexing can be extended untill the first restriction motif to extract only those reads that start from the viewpoint of interest.
19. When preparing 4C template using (previously frozen) primary tissue, the percentage of the reads mapping to the *cis* chromosome frequently drops to ~50%. While indicating the presence of random ligations in the sample, the local coverage around the VP is often still good enough to draw reproducible conclusions.

Author contributions

P.H.L.K and G.G designed figures and wrote the manuscript. P.H.L.K and C.R.E.H. designed and performed experiments. P.H.L.K, G.G and V.B. designed and performed the computational analysis. W.d.L. conceived and supervised the study and wrote the manuscript.

Acknowledgments

We thank all present and past members of the de Laat lab for their input at various stages of the development of the 4C-seq protocol and mapping pipeline. This work was sponsored by a NWO VICI grant (724.012.003) and a grant from the Fondation Leducq (14CVD01) to W.d.L., as well as by the Oncode Institute which is partly financed by KWF, the Dutch Cancer Society.

Declaration of Competing Interest

We declare competing financial interests: P.H.L.K., and G.G. are shareholders of Cergentis. W.d.L. is founder and shareholder of Cergentis.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ymeth.2019.07.014.

References

[1] J. Dekker, et al., Capturing chromosome conformation, Science 295 (5558) (2002) 1306–1311.
[2] A. Denker, W. de Laat, The second decade of 3C technologies: detailed insights into nuclear organization, Genes Dev. 30 (12) (2016) 1357–1382.
[3] M.W. Vermunt, D. Zhang, G.A. Blobel, The interdependence of gene-regulatory elements and the 3D genome, J. Cell Biol. 218 (1) (2019) 12–26.
[4] J.R. Dixon, D.U. Gorkin, B. Ren, Chromatin domains: the unit of chromosome organization, Mol. Cell 62 (5) (2016) 668–680.
[5] J.R. Dixon, et al., Topological domains in mammalian genomes identified by analysis of chromatin interactions, Nature 485 (7398) (2012) 376–380.
[6] T. Sexton, et al., Three-dimensional folding and functional organization principles of the Drosophila genome, Cell 148 (3) (2012) 458–472.
[7] E.P. Nora, et al., Spatial partitioning of the regulatory landscape of the X-inactivation centre, Nature 485 (7398) (2012) 381–385.
[8] M.J. Rowley, V.G. Corces, Organizational principles of 3D genome architecture, Nat. Rev. Genet. 19 (12) (2018) 789–800.
[9] J.H.I. Haarhuis, et al., e cohesin release factor WAPL restricts chromatin loop extension, Cell 169 (4) (2017) 693–707 e14.
[10] M. Simonis, et al., Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C), Nat. Genet. 38 (11) (2006) 1348–1354.
[11] E. Lieberman-Aiden, et al., Comprehensive mapping of long-range interactions reveals folding principles of the human genome, Science 326 (5950) (2009) 289–293.
[12] S.S. Rao, et al., A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping, Cell 159 (7) (2014) 1665–1680.
[13] P.J. Wijchers, et al., Cause and consequence of tethering a SubTAD to different nuclear compartments, Mol. Cell 61 (3) (2016) 461–473.
[14] J. Nuebler, et al., Chromatin organization by an interplay of loop extrusion and compartmental segregation, Proc. Natl. Acad. Sci. USA 115 (29) (2018) E6697–E6706.
[15] J. Dekker, et al., The 4D nucleome project, Nature 549 (7671) (2017) 219–226.
[16] O. Schwartzman, et al., UMI-4C for quantitative and targeted chromosomal contact profiling, Nat. Meth. 13 (8) (2016) 685–691.
[17] J.R. Hughes, et al., Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment, Nat. Genet. 46 (2) (2014) 205–212.
[18] H.J. van de Werken, et al., Robust 4C-seq data analysis to screen for regulatory DNA interactions, Nat. Meth. 9 (10) (2012) 969–972.
[19] E. de Wit, et al., CTCF binding polarity determines chromatin looping, Mol. Cell 60 (4) (2015) 676–684.
[20] D.G. Lupianez, et al., Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions, Cell 161 (5) (2015) 1012–1025.
[21] S. Groschel, et al., A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia, Cell 157 (2) (2014) 369–381.
[22] E. de Wit, et al., The pluripotent genome in three dimensions is shaped around pluripotency factors, Nature 501 (7466) (2013) 227–231.
[23] G. Geeven, et al., peakC: a flexible, non-parametric peak calling package for 4C and Capture-C data, Nucl. Acids Res. 46 (15) (2018) e91.
[24] P.H. Krijger, W. de Laat, Identical cells with different 3D genomes; cause and consequences? Curr. Opin. Genet. Dev. 23 (2) (2013) 191–196.
[25] E.H. Finn, et al., Extensive heterogeneity and intrinsic variation in spatial genome organization, Cell 176 (6) (2019) 1502–1515 e10.
[26] E. Splinter, et al., Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation, Methods 58 (3) (2012) 221–230.
[27] H.J. van de Werken, et al., 4C technology: protocols and data analysis, Meth. Enzymol. 513 (2012) 89–112.
[28] S. Rozen, H. Skaletsky, Primer3 on the WWW for general users and for biologist programmers, Meth. Mol. Biol. 132 (2000) 365–386.
[29] S.F. Altschul, et al., Basic local alignment search tool, J. Mol. Biol. 215 (3) (1990) 403–410.
[30] R.W. Brouwer, et al., Unbiased interrogation of 3D genome topology using chromosome conformation capture coupled to high-throughput sequencing (4C-Seq), Meth. Mol. Biol. 1507 (2017) 199–220.
[31] C. Walter, et al., Benchmarking of 4C-seq pipelines based on real and simulated data, Bioinformatics (2019) pii: btz426.
[32] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Meth. 9 (4) (2012) 357–359.
[33] M. Simonis, et al., High-resolution identification of balanced and complex chromosomal rearrangements by 4C technology, Nat. Meth. 6 (11) (2009) 837–842.
[34] W.J. Kent, et al., The human genome browser at UCSC, Genome Res. 12 (6) (2002) 996–1006.
[35] J.T. Robinson, et al., Integrative genomics viewer, Nat. Biotechnol. 29 (1) (2011) 24–26.
[36] G. Andrey, et al., A switch between topological domains underlies HoxD genes collinearity in mouse limbs, Science 340 (6137) (2013) 1234167.
[37] A. Allahyar, et al., Enhancer hubs and loop collisions identified from single-allele topologies, Nat. Genet. 50 (8) (2018) 1151–1160.
[38] J. Redolfi, et al., DamC reveals principles of chromatin folding in vivo without crosslinking and ligation, Nat. Struct. Mol. Biol. 26 (6) (2019) 471–480.