Roche

Application Note

Simplifying the analysis of chromatin conformation

Authors

Adriana Geldart
Sample Prep Support Manager

Rachel W. Kasinskas
Director of Support & Applications

Roche Sequencing & Life Science
Wilmington, MA, USA

Anthony Schmitt
Vice President, Research and Development

Siddarth Selvaraj
Founder, President, and CEO

Arima Genomics, Inc.
San Diego, CA, USA

High-quality and Reproducible Genome Conformation Analyses Utilizing the Arima-HiC Kit and the KAPA HyperPrep Kit

Recent years have seen a rapid expansion in chromatin conformation technologies, facilitating a broader understanding of epigenetics, gene regulation, and genome structure. To date, genome-wide and targeted Hi-C approaches generate the highest resolution data for chromatin contacts between pairs of genomic loci. However widespread adoption has been limited due to cumbersome protocols and the high number of reads—billions per sample—required to perform high-quality loop-resolution analysis. To overcome these limitations, scientists at Roche and Arima Genomics collaborated to develop the Arima-HiC KAPA HyperPrep workflow. This workflow combines two powerful methods, the Arima-HiC kit and KAPA HyperPrep Kit, into a streamlined method that simplifies the entire Hi-C workflow and dramatically reduces the number of reads required per sample.

Arima-HiC solves the technical challenges of the Hi-C method via a 1-day workflow that requires minimal hands-on time. Following crosslinking, chromatin is digested with a specialized cocktail of restriction enzymes that helps access every loop and topological domain (TAD) in the genome. After biotin incorporation and proximity ligation, the resulting biotin-labeled ligation products represent regions of the genome that were in close spatial proximity in their native conformation. These products are then used as input into the KAPA HyperPrep Kit, a rapid, efficient library prep workflow that retains the complexity of the input samples while reducing preparation time by combining multiple steps. The Arima-HiC KAPA HyperPrep workflow has been evaluated by 18 key opinion leaders (KOLs) in the United States and abroad, using inputs as diverse as plant tissues, animal tissues, blood, and cell lines. We requested that each KOL laboratory perform the following steps: test the workflow for ease-of-use by designating laboratory personnel who are inexperienced with Hi-C to perform the experiments; record the number of PCR cycles required for amplification as a metric for library complexity; assess the quality of the libraries and the resulting data by performing shallow sequencing (~0.1X depth); and, finally, report the fraction of reads representing long-range cis *contacts (>15 Kb apart) as a metric for data quality. The combined results of these evaluations produced high-quality Hi-C data illustrating the technical proficiency, consistency, and ease-of-use of the Arima-HiC KAPA HyperPrep workflow.*

ARIMA GENOMICS

Introduction

The three-dimensional (3D) conformation of genomes has a profound impact on gene regulation, DNA replication, and DNA damage repair, and thus has major implications for disease. Recent years have seen a drastic surge of genome conformation mapping technologies to interrogate the 3D organization of the genome. However, most early methods—such as 3C, 4C, and 5C—required targeted interrogation of specific regions, and thus did not provide an unbiased assessment of chromatin structure. The first unbiased approach to genome-wide chromosome conformation analysis, Hi-C, was not developed until 2009.[1] To date, Hi-C and related targeted approaches such as Capture-Hi-C[2,3] generate the highest resolution of chromatin contacts between pairs of genomic loci. These methods have been used to analyze samples from numerous species, sample types, experimental conditions and disease states.[4,5,6,7]

Standard Hi-C protocols are long and involve numerous steps. Briefly, Hi-C technology relies on formaldehyde crosslinking to induce chemical crosslinks between spatially proximal chromatin segments. After a series of manipulations, DNA fragments that are close in spatial proximity are ligated; this creates ligation products with biotin at the ligation junctions. These DNA fragments are then isolated using biotin:streptavidin affinity purification, converted into sequencing-ready libraries through the ligation of adapters, and sequenced via paired-end sequencing. The 3D conformation of the genome can then be determined from the data generated by measuring the frequency with which pairs of chromatin segments formed ligation products during the Hi-C protocol.

While previous Hi-C protocols[1,4] have been tremendously valuable, widespread adoption has been limited due to the long and labor-intensive workflow, inconsistent results, and the high sequencing costs associated with the required sequencing depth for high-resolution chromatin conformation analyses. To overcome these limitations, Arima Genomics developed the simplified and extremely robust Arima-HiC kit that streamlines the Hi-C protocol via a single-tube chemistry and a 6-hour automation-friendly workflow, with only 1 hour of hands-on time. The Arima-HiC kit efficiently produces ligated, biotin-labeled DNA fragments that can then be prepared as Illumina® sequencing libraries using the KAPA HyperPrep Kit. The KAPA HyperPrep Kit is a streamlined library preparation protocol that combines several enzymatic steps to significantly reduce time spent on library preparation and improve consistency. The single-tube chemistry also offers further improvements to library construction efficiency by dramatically improving library complexity. Together, these products provide an optimized Hi-C method that we have called the Arima-HiC KAPA HyperPrep workflow (Figure 1). This new method consistently generates uniform, high-quality, high-complexity libraries that require significantly reduced sequencing depth for robust and reproducible analysis of chromatin loops and topological domains (TADs).

Materials and methods

Experimental design and DNA samples

In this study, a new workflow combining the Arima-HiC kit and the KAPA HyperPrep Kit was evaluated by the labs of 18 key opinion leaders (KOLs) in the field of chromatin conformation and genome assembly in Europe, Asia, and the United States.

Each KOL testing lab was provided with GM12878 human lymphoblastoid cells (LCLs), which served as a common sample across all KOL test groups. Individual labs also tested the method on a broad range of other sample types, including human muscle and colon tissue, cancer cell lines, and *Arabidopsis* seedlings.

Sample preparation and sequencing

Samples were processed using the Arima-HiC kit according to supplied user guides, followed by purification using KAPA Pure Beads (Figure 1). After DNA shearing and size selection, biotin-labeled ligation junctions were enriched using streptavidin beads. While the DNA was still bound to the streptavidin beads, the KAPA HyperPrep Kit was used to convert the sheared ligation products into Illumina-compatible sequencing libraries, which were then PCR-amplified. The resulting libraries were quantified using the KAPA Library Quantification Kit and sequenced in paired-end mode via Illumina next-generation sequencers. For results and

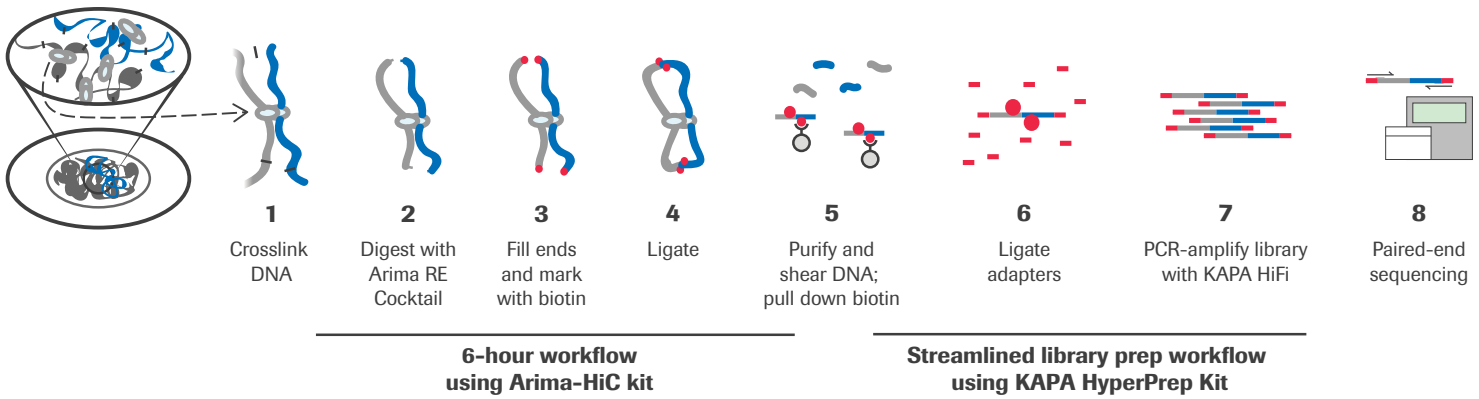

Figure 1: The Arima-HiC KAPA HyperPrep workflow combines the Arima-HiC kit and the KAPA HyperPrep Kit into a streamlined workflow. The Arima-HiC kit yields ligated, biotinylated DNA that serves as input for the KAPA HyperPrep Kit, which then generates sequencing-ready libraries.

Data on file.

discussions in the following sections, the data from Arima-HiC and KAPA HyperPrep Kits were compared to data from homebrew Hi-C protocols that either exactly followed or diverged slightly from existing Hi-C protocols.[1,4]

In addition to the workflow described above, two quality control steps—Arima-QC1 and Arima-QC2—were included at intermediate steps of the workflow. Arima-QC1 measures the success of the Arima-HiC protocol, and Arima-QC2 is an aggregate measure of the efficiency of the complete Arima-HiC KAPA HyperPrep workflow. More information can be found in the User's Guide for the Arima-HiC kit.

Analysis of Arima-HiC KAPA HyperPrep libraries

The Arima-HiC KAPA HyperPrep libraries were evaluated both internally and by KOL testing labs; all testers used multiple metrics to assess quality and complexity.

To assess Arima-HiC KAPA HyperPrep data quality, shallow sequencing (~0.1X depth) was performed on an Illumina® MiniSeq™ instrument in paired-end mode with 37 bp read length. The sequence data were mapped to the appropriate reference genome and further processed using the Arima Genomics mapping pipeline.[8] As a proxy for data quality, the final processed dataset was analyzed for the proportion of uniquely mapped monoclonal Hi-C read-pairs that represented long-range intra-chromosomal (*cis*) chromatin interactions (>15 kb interactions). A library was considered high quality when at least 40% of the processed Hi-C reads represented long-range *cis* interactions.

To assess library complexity, the number of PCR cycles required to obtain a 5 nM Arima-HiC KAPA HyperPrep library was recorded. Libraries that required 8 or fewer PCR cycles were considered high complexity, and appropriate for deep sequencing and loop resolution analysis.

Analysis of chromatin looping and topologically associated domains (TADs)

To evaluate the utility of this workflow for robust and reproducible genome conformation analysis of chromatin loops and TADs, an Arima-HiC KAPA HyperPrep library was prepared from GM12878 LCLs and sequenced on the Illumina HiSeqX® to generate 1.2 B reads (~120X depth, 150 bp paired-end reads). The resulting reads were processed using default parameters via the open source Juicer[9] software to generate normalized Hi-C contact maps with annotated chromatin loops and TADs. In addition, Juicer was used to conduct aggregate peak analysis (APA) to portray the totality nature of loops and TADs. The raw reads and processed files can be accessed from ftp://ftp-arimagenomics.sdsc.edu/pub/Conformation/. To assess reproducibility of analyses, results from our analysis was compared to GM12878 "Primary" and "Replicate" datasets from Rao et al.[4] Finally, to assess the sensitivity of loop annotation as a function of sequencing depth, raw reads were randomly sub-sampled and then processed as previously described.

Results and discussion

Reproducible high-quality and high-complexity libraries

To validate the performance of Arima-HiC KAPA HyperPrep, the workflow was evaluated independently by 18 KOL testing labs. All 18 groups generated high-quality, high-complexity libraries on their first attempt, with consistent results across samples. Specifically, the testing labs detected 55% – 65% long-range *cis* interactions and required only 4 to 7 PCR cycles to obtain 5 nM library (Figure 2).

Hi-C protocols have historically lacked quantitative quality control (QC) metrics, leading to inconsistent data and failed experiments. To solve this problem, we have developed two simple, quantitative QC steps that require only commonly available reagents and equipment.

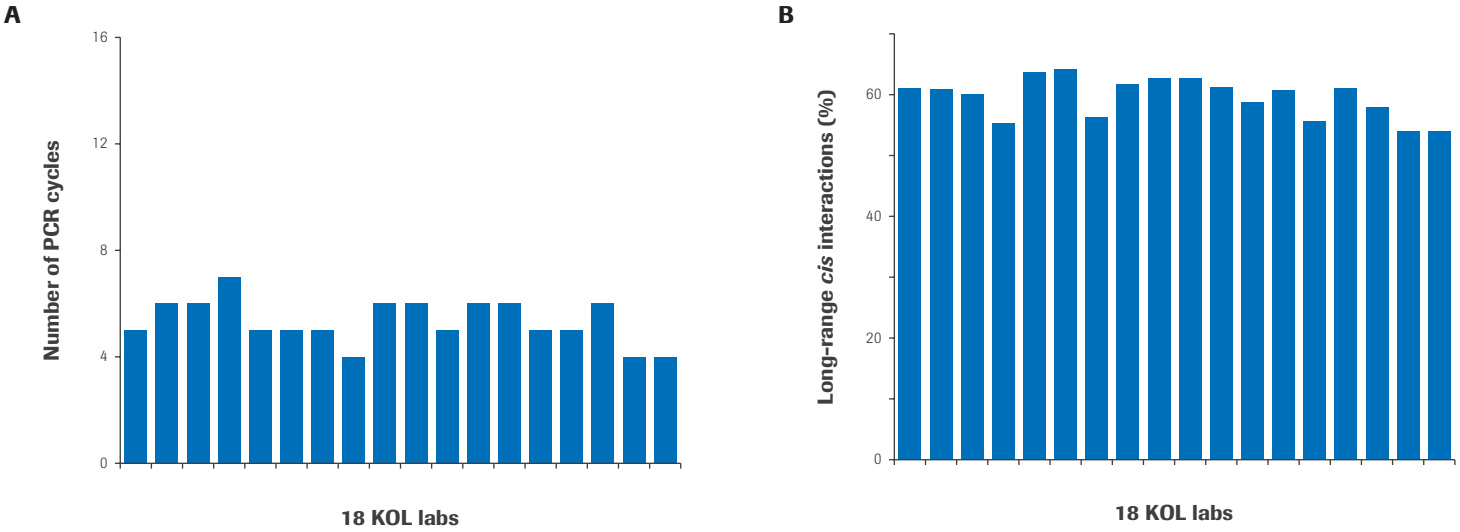

Figure 2: KOL testing labs consistently generated high-complexity, high-quality libraries with the Arima-HiC KAPA HyperPrep workflow. (A) The high complexity of the libraries is indicated by the low number of PCR cycles (4 to 7 cycles) that were required to obtain 5 nM of library. **(B)** The high quality of the libraries is demonstrated by the large fraction of long-range *cis* interactions detected, 55% – 65%.

These two steps, Arima-QC1 and Arima-QC2, measure key aspects of the workflow; Arima-QC1 measures the fraction of DNA that has been labeled with biotin after completion of the Arima-HiC portion of the workflow, and Arima-QC2 utilizes the KAPA Library Quantification Kit to provide an aggregate measure of the overall efficiency of the Arima-HiC KAPA HyperPrep workflow.

Figure 3 depicts data from 95 Arima-HiC KAPA HyperPrep experiments conducted internally at Arima Genomics where failures were intentionally forced for some of the samples. The resulting Arima-QC1 and Arima-QC2 metrics accurately predicted the quality of the libraries, as QC metrics strongly correlate with the long-range *cis* interaction signal in the resulting sequencing data. Thus, the incorporation of these two QC steps into the Arima-HiC KAPA HyperPrep workflow saves time and resources by circumventing the need to run gels and perform shallow sequencing while ensuring that only successful experiments will be sequenced.

To further validate the performance of the workflow and the quality of the libraries generated, several KOL testing labs conducted a side-by-side comparison of libraries generated using homebrew Hi-C protocols[1,4] and the Arima-HiC KAPA HyperPrep workflow. As shown in Figure 4A, Arima-HiC KAPA HyperPrep libraries consistently detect a greater proportion of long-range *cis* interactions regardless of sample type.

Computational estimations suggest that about one-third of the genome (estimated from chr1 of the human genome) is inaccessible when Hi-C is performed with a 4-base restriction enzyme (RE) in homebrew methods.[4] The Arima-HiC kit overcomes this limitation by using an RE cocktail that generates nearly complete accessibility of the genome (Figure 4B). The combination of this uniquely powerful RE cocktail and the short duration of the workflow enables the Arima-HiC KAPA HyperPrep workflow to rapidly generate libraries with enriched signal-to-noise ratios. (Figure 2, 4A).

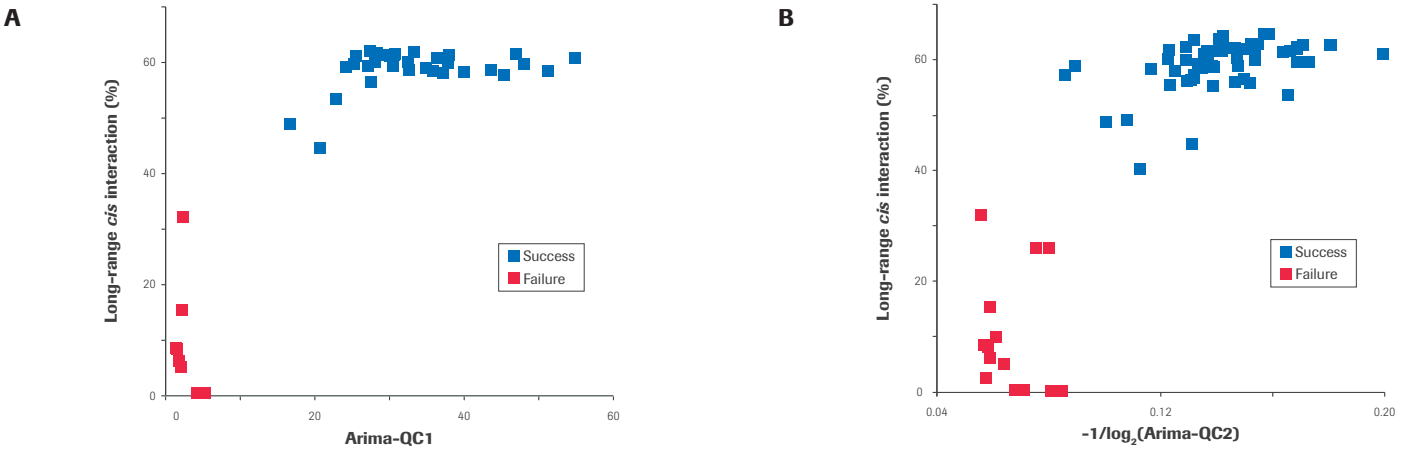

Figure 3: Two quality control metrics, Arima-QC1 and Arima-QC2, accurately reflect library quality.

Depicted is the correlation between **(A)** Arima-QC1 and **(B)** Arima-QC2 scores for 95 internally produced samples and the long-range interaction signal obtained from low-depth sequencing.

For intentionally failed samples (red), the poor Arima-QC1 and Arima-QC2 scores correlate with detection of few long-range interactions. In contrast, successful samples (blue) show good correlation between Arima-QC1 and Arima-QC2 scores with a high detection of long-range interactions.

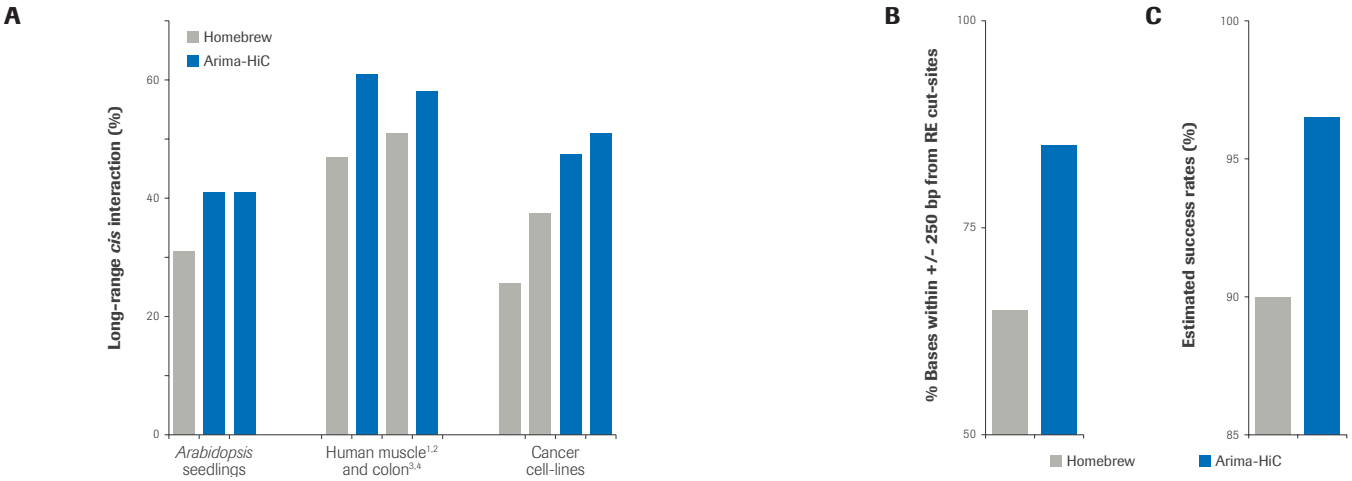

Figure 4: Arima-HiC KAPA HyperPrep libraries outperform homebrew Hi-C libraries in detection of long-range *cis* interactions, genome accessibility, and library success rates in multiple sample types. (A) KOL testing labs compared homebrew and Arima-HiC KAPA HyperPrep libraries created from multiple sample types, and determined that the Arima-HiC KAPA HyperPrep libraries identify more long-range *cis* interactions. **(B)** The RE cocktail used in the Arima-HiC KAPA HyperPrep workflow provides greater genome accessibility than the single 4-bp RE used in homebrew Hi-C workflows, illustrated by the detection of more base pairs within 250 bp of the RE cut sites. **(C)** KOL data for homebrew Hi-C compared to internal Arima Genomics data for 171 Arima-HiC KAPA HyperPrep libraries shows that the Arima-HiC KAPA HyperPrep workflow is more successful.

Data on file.

High-quality, highly reproducible chromatin conformation data

To evaluate whether Arima-HiC KAPA HyperPrep produced a comprehensive, reproducible analysis of chromatin looping and TADs, we compared the Hi-C contacts that were identified in GM12878 cells using Arima-HiC KAPA HyperPrep (sequenced to ~1.2 B reads, or ~120X; see **Materials and methods** for details) to those identified in a previous study conducted by Rao et al.[4] The comparisons are visualized as heatmaps, where the intensity of each pixel represents the number of reads (Hi-C contacts) connecting two bins. The greater the number of contacts connecting the bins, the stronger the level of evidence that they are spatially proximal. As expected, the on-diagonal pixels are red as chromatin segments very close in linear sequence are also very close in 3D space, and thus are more likely to form ligation products during the Hi-C protocol. In contrast, the highly enriched off-diagonal red pixels represent regions that are distal in linear sequence but were spatially proximal in the genome at the time of crosslinking, and thus represent "chromatin loops." Figure 5A shows an example of such a chromatin loop with 74 supporting reads in the Rao et al. Primary[4] dataset.

Interestingly, Arima-HiC KAPA HyperPrep data recovered this loop (Figure 5B) at a slightly higher signal strength (96 supporting reads) despite being sequenced at less than half the sequencing depth. Furthermore, we sub-sampled the Arima-HiC KAPA HyperPrep data to 600 M raw reads and still observed strong loop signal enrichment (Figure 5C), suggesting that this workflow is able to enrich for signal even at reduced sequencing depth.

To determine whether this phenomenon of highly sensitive loop detection was restricted to only a few loops or whether it was a genome-wide phenomenon, we performed aggregate peak analysis (APA). This algorithm generates an APA score that summarizes the correlation between two datasets at totality of all loop pixels and their neighbors; higher scores indicate better correlation. Arima-HiC KAPA HyperPrep generated a higher APA score with the Rao et al. Primary[4] dataset (Figure 5D) than the score generated using the Primary[4] dataset and its own Replicate[4] dataset (Figure 5E), suggesting a genome-wide enriched and accurate chromatin looping signal in the Arima-HiC KAPA HyperPrep dataset despite being sequenced at a significantly reduced depth.

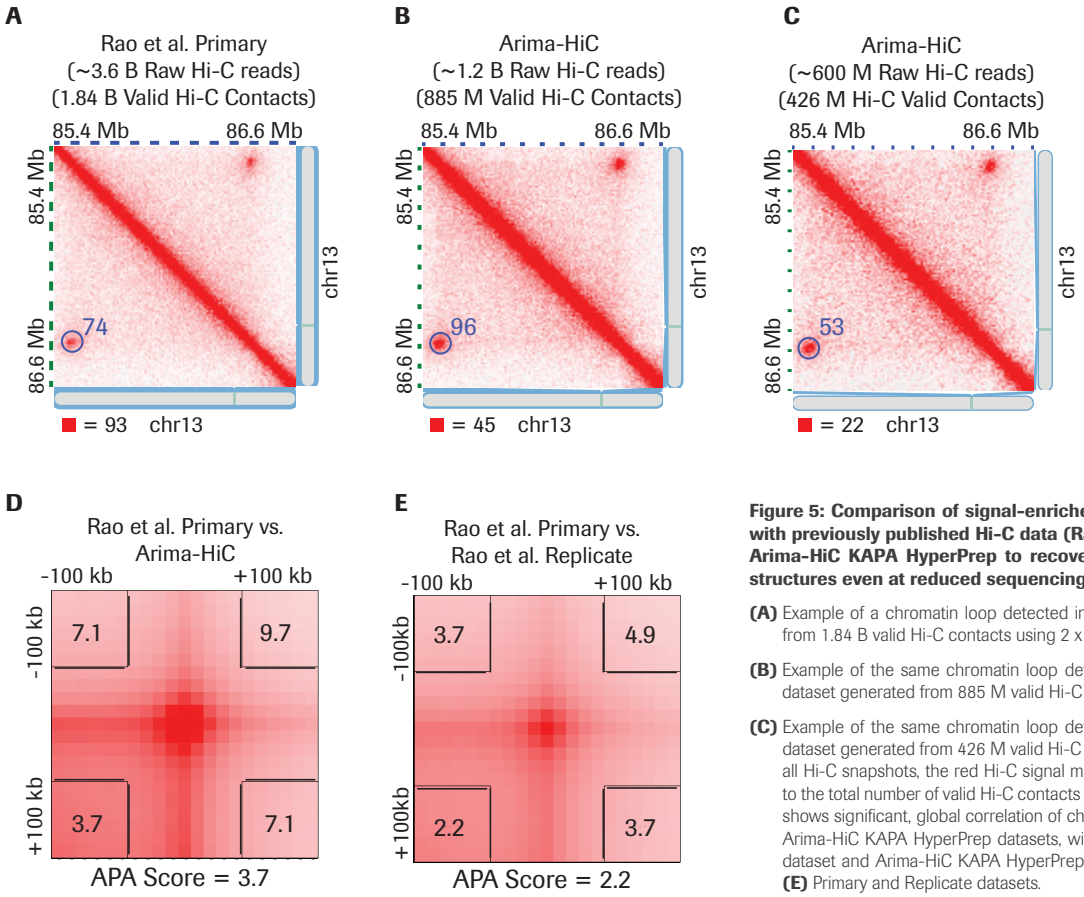

Figure 5: Comparison of signal-enriched Arima-HiC KAPA HyperPrep data with previously published Hi-C data (Rao et al.) demonstrates the ability of Arima-HiC KAPA HyperPrep to recover known chromatin loops and TAD structures even at reduced sequencing depth.

(A) Example of a chromatin loop detected in the Rao et al. Primary dataset generated from 1.84 B valid Hi-C contacts using 2 x 100 bp sequencing.

(B) Example of the same chromatin loop detected in the Arima-HiC KAPA HyperPrep dataset generated from 885 M valid Hi-C contacts using 2 x 150 bp sequencing.

(C) Example of the same chromatin loop detected in the Arima-HiC KAPA HyperPrep dataset generated from 426 M valid Hi-C contacts using 2 x 150 bp sequencing. For all Hi-C snapshots, the red Hi-C signal maximum threshold is scaled linearly relative to the total number of valid Hi-C contacts in the map. Aggregate peak analysis (APA) shows significant, global correlation of chromatin loop signal between Rao et al. and Arima-HiC KAPA HyperPrep datasets, with APA scores between Rao et al. Primary dataset and Arima-HiC KAPA HyperPrep **(D)** being even higher than between the **(E)** Primary and Replicate datasets.

Overall, a total of 15,553 chromatin loops were identified from the Arima-HiC KAPA HyperPrep dataset of 1.2 B reads. These loops include 6,345 (79%) of the 8,058 the loops identified in the Rao et al. Primary[4] dataset, which contained 3.6 B reads (or, 3X more reads) (Figure 6A – i). In addition, the Arima-HiC KAPA HyperPrep dataset identified 9,208 loops that are not in the Rao et al. Primary[4] dataset. We performed APA on these unique Arima-HiC KAPA HyperPrep loops. As expected, these loops showed high APA scores at the loop pixels and neighbors in the Arima-HiC KAPA HyperPrep dataset (Figure 6A – ii). More importantly, these unique loops also showed a moderate signal in the Rao et al. Primary[4] dataset (Figure 6A – iii), suggesting that Arima-HiC KAPA HyperPrep is capable of identifying thousands of loops that were not previously identified as loops in the Rao et al. Primary[4] dataset. This enhanced level of detection likely results from the higher loop signal enrichment relative to local background in the Arima-HiC KAPA HyperPrep data, which was enabled by the ability of the robust Arima RE mix to access genome regions that are inaccessible in the single-RE homebrew methods used to generate the Rao et al. Primary[4] data.

We performed further analysis of the Rao et al. Primary[4] and Replicate[4] data, combining both their Primary[4] and Replicate[4] data sets to determine whether analysis at greater sequencing depth would improve loop signal or yield information about genome regions that appeared inaccessible at lower depth. The Rao et al. Primary[4] and Replicate[4] datasets were combined to create an expanded dataset (referred to as Combined) with 4 times the sequencing depth of the Arima-HiC KAPA HyperPrep dataset. The Arima-HiC KAPA HyperPrep still recovered 73% of the chromatin loops identified in the Combined dataset, and identified 8,691 unique loops that showed only moderate signal in the Combined dataset (Figure 6B – i, ii, iii), demonstrating the sensitivity of the Arima-HiC workflow. Even when sub-sampled down to 600 M reads, the Arima-HiC data still detected >13,000 loops (Figure 6C), which is more than the number identified in both the Rao et al. Primary[4] or Combined datasets. This finding demonstrates that when the Hi-C library is of very high quality, such as those generated the Arima-HiC KAPA HyperPrep workflow, substantially less sequencing depth is required to recover comprehensive and accurate genome conformation data.

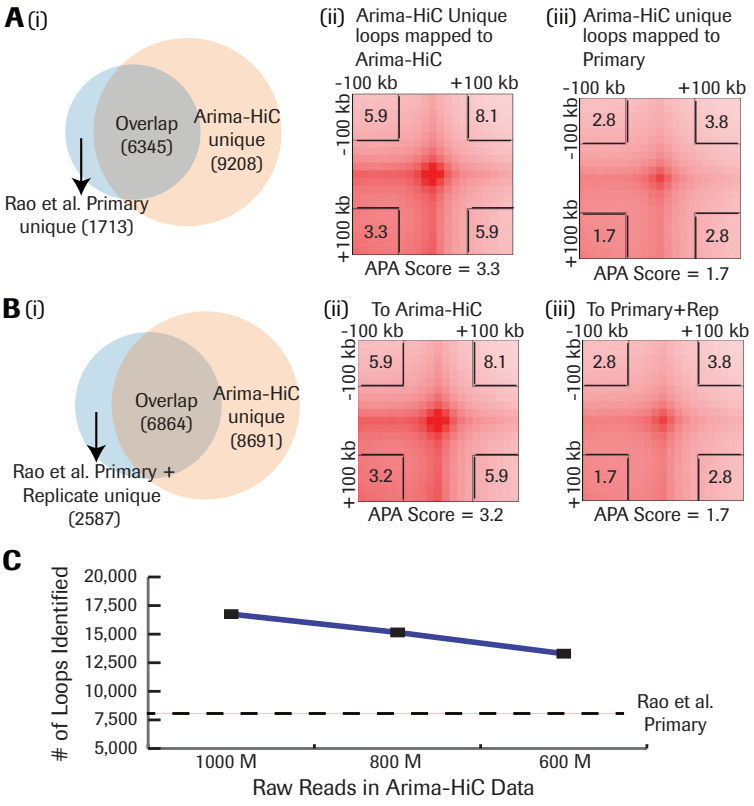

Figure 6: Arima-HiC KAPA HyperPrep demonstrates the ability to discover chromatin loops not identified in a previously published study even at reduced sequencing depth.

(A-i) Comparison of Arima-HiC KAPA HyperPrep data (1.2 B raw reads) with the Rao et al. Primary dataset (3.6 B raw reads) showed significant recall of previously identified loops in addition to thousands of previously unidentified loops (Arima-HiC KAPA HyperPrep unique loops).

(A-ii) High APA score calculated for Arima-HiC KAPA HyperPrep unique loops within Arima-HiC KAPA HyperPrep data.

(A-iii) Unique Arima-HiC KAPA HyperPrep loops showed a moderate signal in the Rao et al. Primary dataset, illustrating that these are likely true loops missed by Rao et al.

(B-i, ii, iii) Arima-HiC KAPA HyperPrep maintained significant recall of previously identified loops and Arima-HiC KAPA HyperPrep unique loops in a combined Rao et al. Primary and Replicate dataset (>6 B raw reads; referred to in the text as Combined) that has 5-times the sequencing depth as Arima-HiC KAPA HyperPrep dataset.

(C) Analysis of the total number of loops identified in Arima-HiC KAPA HyperPrep data when sub-sampled down to 600 M raw reads, indicating the excellent loop calling sensitivity of the Arima-HiC KAPA HyperPrep data.

Conclusions

In summary, the Arima-HiC KAPA HyperPrep workflow combines Arima-HiC and the KAPA HyperPrep Kit to create a streamlined and simplified Hi-C workflow that produces high-quality libraries with ease, consistency, and built-in QC checkpoints. The high quality and enriched signal-to-noise Arima-HiC KAPA HyperPrep libraries outperform homebrew methods and enable the identification of both known and previously unknown chromatin interaction loops at reduced sequencing depths.

Acknowledgements

Arima would like to thank the NHGRI for generously funding the Arima-HiC kit development project 1R44HG009584-01.

References

1. Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive Mapping of Long Range Interactions Reveals Folding Principles of the Human Genome. *Science*. 2009 Oct 9; 326(5950), 289 – 293.
2. Dryden NH, Broome LR, Dudbridge F, Johnson N, Orr N, Schoenfelder S, Nagano T, Andrews S, Wingett S, Kozarewa I, Assiotis I, Fenwick K, Maguire SL, Campbell J, Natrajan R, Lambros M, Perrakis E, Ashworth A, Fraser P, Fletcher O. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res*. 2014 Nov;24(11):1854 – 1868.
3. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA, Herman B, Happe S, Higgs A, LeProust E, Follows GA, Fraser P, Luscombe NM, Osborne CS. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet*. 2015 Jun;47(6), 598 – 606.
4. Rao SP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Lieberman-Aiden E. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*. 2014 Dec 18;159,1665 – 1680.
5. Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot JP, Tanay A, Cavalli G. Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell*. 2017 Oct 19;171(3):557 - 572.e24.
6. Dixon JR, Xu J, Dileep V, Zhan Y, Song F, Le VT, Yardımcı GG, Chakraborty A, Bann DV, Wang Y, Clark R, Zhang L, Yang H, Liu T, Iyyanki S, An L, Pool C, Sasaki T, Rivera-Mulia JC, Ozadam H, Lajoie BR, Kaul R, Buckley M, Lee K, Diegel M, Pezic D, Ernst C, Hadjur S, Odom DT, Stamatoyannopoulos JA, Broach JR, Hardison RC, Ay F, Noble WS, Dekker J, Gilbert DM, Yue F. Integrative detection and analysis of structural variation in cancer genomes. *Nat Genet*. 2018 Oct;50(10):1388 – 1398.
7. Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, Cairns J, Wingett SW, Várnai C, Thiecke MJ, Burden F, Farrow S, Cutler AJ, Rehnström K, Downes K, Grassi L, Kostadima M, Freire-Pritchett P, Wang F; BLUEPRINT Consortium, Stunnenberg HG, Todd JA, Zerbino DR, Stegle O, Ouwehand WH, Frontini M, Wallace C, Spivakov M, Fraser P. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoter. *Cell*. 2016 Nov 17;167(5):1369-1384.e19.
8. Arima Genomics Mapping Pipeline. https://github.com/ArimaGenomics/mapping_pipeline
9. Juicer. https://github.com/theaidenlab/juicer

Published by:

Roche Sequencing & Life Science
9115 Hague Road
Indianapolis, IN 46256

sequencing.roche.com

MC-US-03647 APP108004 A434 2/19