ARIMA GENOMICS

The Arima-HiC Kit for Reproducible 3D Genome Conformation Analyses

DECEMBER 2018

ARIMA GENOMICS

APPLICATION NOTE
DECEMBER 2018

The Arima-HiC Kit for Reproducible 3D Genome Conformation Analyses

1. Introduction

The three-dimensional (3D) conformation of genomes has a profound impact on gene regulation, DNA replication, and DNA damage repair. Recent years have seen a drastic expansion of genome conformation mapping technologies, including Hi-C, a sequencing-based assay designed to interrogate the 3D organization of the genome. To date, Hi-C and related approaches such as Capture-Hi-C and HiChIP for targeted analyses generate the highest resolution of contact frequency between pairs of genomic loci, across several sample types, species, and experimental conditions.

While existing Hi-C protocols[1,2] have served as a tremendously valuable tool for some in the community, widespread adoption has been precluded by labor-intensive, cumbersome and complex protocols, prolonged ~4 day duration, inconsistent experimental results, and the demand for billions of sequence reads per sample for high-resolution chromatin looping analyses.

To overcome these technical and economic limitations, we have developed a highly simplified and robust Arima-HiC kit that streamlines the Hi-C protocol via a single-tube chemistry and a 6-hour automation-friendly workflow.

This application note describes the optimized Hi-C protocol (hereafter referred to as Arima-HiC, Fig 1) and demonstrates its ability to generate high-quality and high-complexity libraries. The rapid Arima-HiC protocol substantially reduces experimental noise resulting from prolonged exposure of chromatin to external agents and significantly reduces the biased genomic representation ("biased coverage") in existing Hi-C protocols via the use of multiple 4-base cutting restriction enzymes with motifs GATC and GANTC (RE cocktail) for chromatin digestion. Consequently, Arima-HiC obtains enriched signal-to-noise libraries across different sample types and species. Additionally, Arima-HiC produces high quality data with low cell input amounts (50,000 cells or less) that were previously thought to be inaccessible by the technology. When sequenced via next generation sequencers, Arima-HiC libraries required significantly less sequencing depth for robust and reproducible analyses of chromatin interaction loops and topological domains (TADs), thus providing economic benefit to the researcher.

2. Materials & Methods

2.1 DNA samples

The Arima-HiC libraries and sequencing data were evaluated on a broad range of sample-types and species through a beta testing program involving 18 leading labs in the conformation community across the world in US, Europe, and Asia. The beta testers were provided with an Arima-HiC kit and GM12878 Human Lymphoblastoid cells (LCLs) as control. Additionally, some beta testers chose to evaluate the Arima-HiC kit on a variety of sample types such as human muscle and colon cancer tissue, cancer cell lines, Arabidopsis seedlings, among others.

2.2 Homebrew Hi-C and Arima-HiC sample preparation and next-generation sequencing

Arima-HiC is a 6-hour protocol that results in (proximally) ligated DNA, which can be prepared as an Arima-HiC library using pre-validated commercially available library prep kits (Fig 1). After library

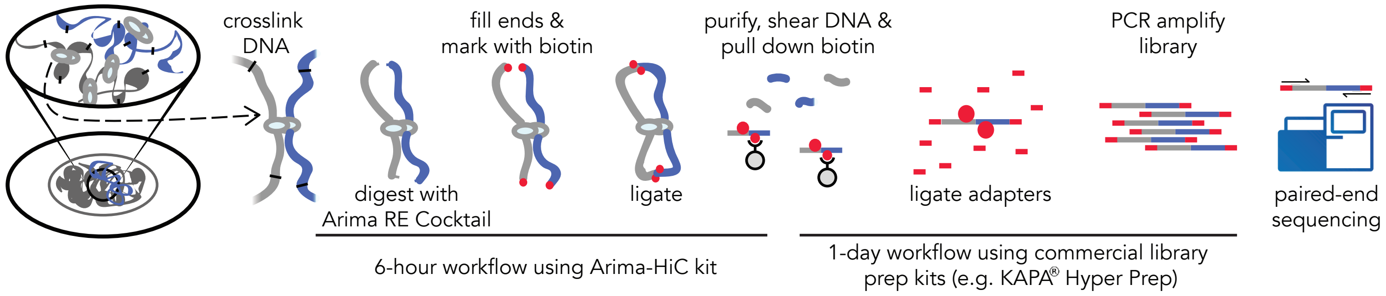

Figure 1: The Arima-HiC workflow results in ligated and biotinylated DNA that is prepared as a library and amplified using pre-validated library prep kits, and then subject to paired-end Illumina sequencing.

prep, the Arima-HiC libraries are sequenced in paired-end mode via Illumina next generation sequencers (Fig.1). For results and discussions in the next sections, the data from Arima-HiC kits were compared to data from homebrew Hi-C protocols that were either exact or slight modifications of existing Hi-C protocols[1,2].

2.3 Analysis of Arima-HiC libraries

The Arima-HiC libraries were evaluated, both internally and by beta testers, using three metrics: quality, complexity, and ease-of-use.

To assess quality, Arima-HiC libraries were sequenced to a low-depth (0.1X). The resulting Arima-HiC sequencing data are mapped to a reference genome and two types of signals are generated: intra-chromosomal cis and inter-chromosomal trans. The cis signal can be further categorized as short-range (<15Kb interactions) and long-range (>15Kb interactions). High percentage of long-range cis interactions is thought to correspond to higher quality of the library, whereas short-range cis and trans interactions usually represent self- and random- ligations often classified as experimental noise[2].

Hg19 and TAIR10 reference genomes were used for human and Arabidopsis, respectively. For mapping of Arima-HiC sequence reads to reference genomes for evaluating Arima-HiC library quality, Arima Genomics mapping pipeline[3] was used.

To assess complexity, Arima-HiC libraries were PCR amplified to obtain a 5nM library. Fewer PCR cycles correspond with higher complexity of the library.

To assess ease-of-use, we requested the beta testers to have an inexperienced staff perform the Arima-HiC protocol and record their ability to generate high-quality and high-complexity Arima-HiC libraries on their first attempt.

2.4 Chromatin loop and topologically associated domains (TADs) analysis

To evaluate the utility of Arima-HiC protocol for robust and reproducible chromatin conformation analyses of interaction loops and TADs, an Arima-HiC library was prepared from GM12878 LCLs and sequenced on the Illumina HiSeqX to generate 1.2B reads (~120X depth, 150bp paired-end reads). The resulting Arima-HiC reads were processed using default parameters via Juicer[4], an open source software to generate normalized contact maps with annotated chromatin loops and TADs. In addition, Juicer[4] also performs aggregate peak analysis (APA) to portray the totality nature of loops and TADs. The raw Arima-HiC reads and processed files can be accessed from ftp://ftp-arimagenomics.sdsc.edu/pub/Conformation/. Finally, to assess reproducibility of analyses, results from Arima-HiC were compared to "Primary", "Replicate" datasets from Rao et al[2].

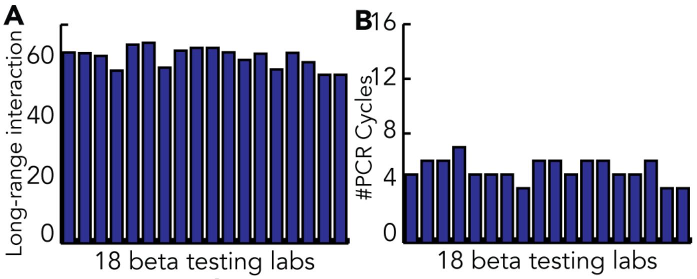

Figure 2: Using the Arima-HiC kits, Beta testers consistently generated high-quality and high-complexity Arima-HiC libraries. (A) Beta testing labs generated high long-range fraction of 55-65%, which typically corresponds to high-quality of the library. (B) Labs required only 4-7 PCR cycles to obtain a 5nM library, indicating the high complexity of the library – i.e. libraries can be sequenced to high depth.

3. Results

3.1 Reproducible high-quality and high complexity libraries

To validate the performance of the Arima-HiC chemistry, the Arima-HiC kit was evaluated independently by 18 beta testers with both experienced and inexperienced users. Using Arima-HiC kits, the beta testing labs consistently generated 55-65% long-range interactions and required only 4-7 PCR cycles to obtain 5nM library, demonstrating the high-quality and high-complexity of Arima-HiC libraries (Fig 2A, B).

Historically, Hi-C protocols have lacked quantitative quality control metrics, leading to inconsistent data and failed experiments. To solve this problem, we have developed two simple, quantitative quality-control (QC) steps where the derived QC metrics accurately predict the quality of the library as they correlate strongly with long-range interaction signal (Fig 3).

The data discussed in Fig 3 were generated from 95 experiments conducted internally at Arima Genomics as part of Arima-HiC kit gaurdbanding, where failures were intentionally forced. Arima-QC1 measures the fraction of proximally ligated DNA that has been labeled with biotin while Arima-QC2 is an aggregate measure of proximity ligation and library preparation efficiency. Together, Arima-QC1 and Arima-QC2 measure key aspects of the Arima-HiC workflow using only commonly available equipment. These simple QC steps circumvent the need to run gels and other frequent checks while ensuring the Arima-HiC experiment will result in success.

To further validate the performance of the Arima-HiC chemistry, several beta testers who had prior experience with homebrew Hi-C[1,2] conducted a side-by-side comparison of library quality between homebrew Hi-C protocol[1,2] and Arima-HiC protocol.

As shown in Fig 4A, Arima-HiC generated higher long-range interactions and thus higher quality in every case, regardless of the species and 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, resulting in potentially limited interaction signal for one-third of the genes. Arima-HiC overcomes this limitation by using a RE cocktail and generates near complete accessibility of the genome (Fig 4B). This unique aspect when combined with its rapid duration enables Arima-HiC to generate enriched signal-to-noise libraries (Fig 2,4A).

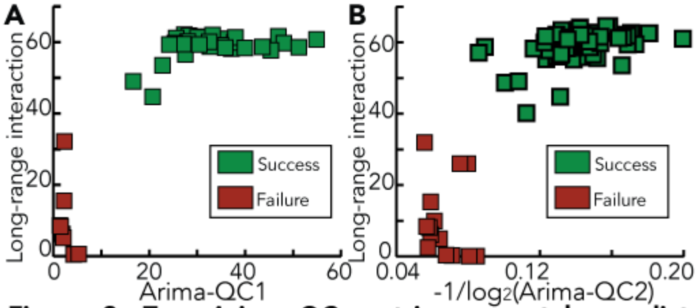

Figure 3: Two Arima QC metrics accurately predict library quality. Arima-QC1 (A) and Arima-QC2 (B) are QC metrics derived from Arima-HiC experimentation that strongly correlate to long-range interaction signal obtained via low-depth sequencing (as discussed in section 2.3 to assess library quality) and, therefore, these QC metrics able to discriminate between successful and failed experiment. Data from 95 internal Arima-HiC kit guardbanding experiments, where failed experiments were intentionally forced.

3.2 High conformation data quality and reproducibility

To evaluate Arima-HiC sequencing data for comprehensive and reproducible chromatin interaction loop and TAD analyses, we compared loops and TADs identified from Arima-HiC library prepared from GM12878 cells sequence to ~1.2B reads (~120X) (see section 2.4 for further details), to those identified from a prior publication[2].

The comparison henceforth is discussed primarily by visualizing Hi-C data in the form of heatmaps, where each pixel represents the number of reads or contacts connecting two bins, i and j. The higher the number of contacts connecting i and j, the stronger the level of evidence that i and j bins are spatially proximal. Heatmaps are often shown in continuum of colors with white as weaker evidence and red as stronger evidence of spatial proximity. The off-diagonal red pixels in these heatmaps are interesting as they suggest linearly distal but spatially proximal bins, often called as chromatin interaction loops. As an example, Fig 5A shows such a chromatin loop with 74 supporting reads in Rao et al Primary[2] dataset.

Interestingly, Arima-HiC recovered this loop (Fig 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 data to 600M reads and still observe strong loop signal enrichment (Fig 5C), suggesting that Arima-HiC is able to enrich for signal even at much reduced sequencing depth. To analyze if this phenomenon were restricted to few loops or whether it is a genome-wide phenomenon, we performed aggregate peak analysis (APA) which generates APA score that summarizes correlation of the two datasets at the totality of all loop pixels and their neighbors. Arima-HiC generates a higher APA score with Rao et al Primary[2] dataset (Fig 5D) than what the Primary[2] dataset generates with its own Replicate[2] dataset (Fig 5E) – suggesting genome-wide enriched and accurate interaction signal in the Arima-HiC dataset despite being sequenced at significantly reduced depth.

Overall, from the Arima-HiC dataset of 1.2B reads, a total of 15,553 chromatin loops were identified, recalling 6,345 of the 8,058 (79%) of the loops identified in the Rao et al Primary[2] dataset with 3.6B reads (Fig 6A-i). Intrigued by 9,208 loops identified in the Arima-HiC dataset that are not in the Rao et al Primary[2] dataset, we performed APA on these Arima-HiC unique loops. These loops, expectedly, showed high APA score at the loop pixels and neighbors in Arima-HiC dataset (Fig 6A-ii). But, more importantly, these unique Arima-HiC loops showed a moderate signal in the Rao et al Primary[2] dataset (Fig 6A-iii), suggesting that Arima-HiC is capable of identifying thousands of loops not previously identified in the Rao et al Primary[2] dataset. This is likely due to the virtue of higher loop signal enrichment relative to local background of Arima-HiC data along with the added facet of Arima-HiC's additional capability of accessing regions of genome that are inaccessible in homebrew methods as in Rao et al Primary[2] dataset (due to single RE and resulting biased coverage in the homebrew protocol). Notably, additional sequencing depth to Rao et al[2] Primary dataset neither improves loop signal nor enables access to inaccessible regions. Specifically, when Rao et al Primary[2] and Replicate[2] dataset are combined to result in a new dataset that is greater than 4-times the sequencing depth of Arima-HiC, Arima-HiC still recovers 73% of loops and identifies 8,691 unique loops that show moderate signal in the combined[2] dataset (Fig 6B-i,ii,iii). Lastly, when sub-sampled down to 600M reads, Arima-HiC data is still able to detect >13,000 loops (Fig 6C), which is more than both the Rao et al Primary or Combined datasets.

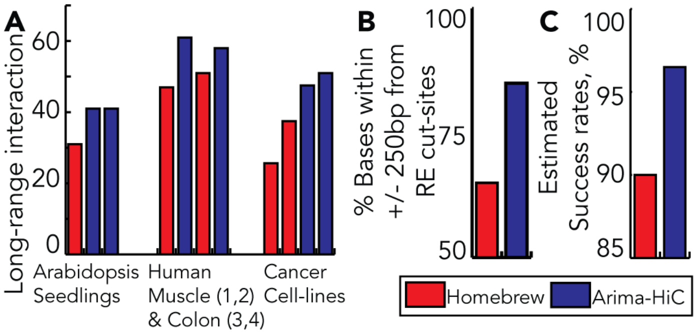

Figure 4: Arima-HiC kit generates higher quality library, more comprehensive genome interaction signal, and greater experimental success rate than homebrew Hi-C protocols. A) Beta testers generated libraries with higher long-range fraction, indicating higher quality, with the Arima-HiC kit across multiple different species and sample types. (B) Arima-HiC produced significantly higher percentage of bases within 250bp from RE cut-sites, indicating near complete accessibility of the genome that leads to more comprehensive interaction signal that was previously inaccessible owing to using a single enzyme in homebrew Hi-C protocols. (C) Arima-HiC demonstrated greater robustness over homebrew Hi-C protocol with a higher success rate in generating high-quality and high-complexity libraries.

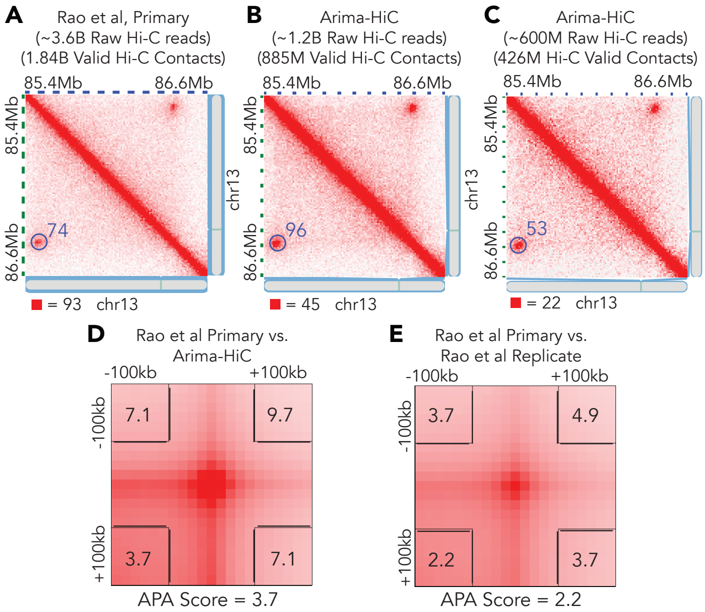

Figure 5: Comparison of signal-enriched Arima-HiC data with previously published Hi-C data (Rao et al) demonstrates the ability of Arima-HiC to recover known chromatin loops and TAD structures even at reduced sequencing depth. (A) Example of a chromatin loop detected in the Rao et al Primary dataset generated from 1.84B valid Hi-C contacts using 2x100bp sequencing. (B) Example of the same chromatin loop detected in the Arima-HiC dataset generated from 885M valid Hi-C contacts using 2x150bp sequencing. (C) Example of the same chromatin loop detected in the Arima-HiC dataset generated from 426M valid Hi-C contacts using 2x150bp 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 datasets, with APA scores between Rao et al Primary dataset and Arima-HiC (D) being even higher than between the (E) Primary and Replicate datasets .

> **"When we tested Arima-HiC kit in our hands, we consistently got very high-quality libraries. Particularly the fraction of inter-chromosomal reads is reduced."**
>
> *- Daan Noordmeer PhD, Leader of I2BC Chromatin Dynamics Group, France*

> **"The Arima-HiC protocol is days quicker than our previous lab's protocol. All together, better quality data for less effort."**
>
> *- Eileen Furlong PhD, Head of EMBL Genome Biology Unit*

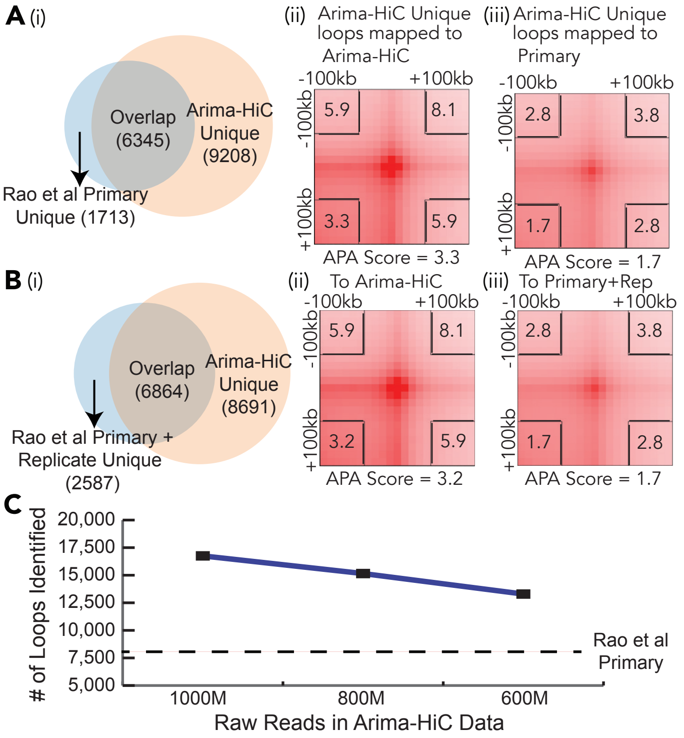

Figure 6: Arima-HiC demonstrates the ability to discover chromatin loops not identified in a previously published study even at reduced sequencing depth. (A-i) Comparison of Arima-HiC data (1.2B raw reads) with Rao et al Primary dataset (3.6B raw reads) showed significant recall of previously identified loops in addition to thousands of previously unidentified loops (Arima-HiC unique loops). (A-ii) High APA score calculated for Arima-HiC unique loops within Arima-HiC data. (A-iii) Unique Arima-HiC 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 maintained significant recall of previously identified loops and Arima-HiC unique loops in a combined Rao et al Primary and Replicate datasets (>6B raw reads) that has 5-times the sequencing depth as Arima-HiC dataset. C) Analysis of the total number of loops identified in Arima-HiC data when sub-sampled down to 600M raw reads, indicating the excellent loop calling sensitivity of the Arima-HiC data.

> **"The ability of Arima-HiC kits to deliver greater insight with less sequencing cost will be critical in leveraging Hi-C technology for understanding gene regulation within the three-dimensional context of the genome."**
>
> *-Changhoon Kim PhD, Chief Technology Officer, Macrogen, Inc.*

This finding demonstrates that when the library is of very high-quality, as in Arima-HiC, less sequencing depth is sufficient to recover comprehensive and accurate aspects of genome conformation.

4. Conclusions

In summary, the Arima HiC protocol, available in the form of Arima-HiC kits, is a streamlined and simplified HiC workflow that produces high quality libraries with consistency. The high quality and enriched signal-to-noise Arima-HiC library enables the identification of both known and previously unknown chromatin interaction loops at much reduced sequencing depths.

For Bioinformatics, we strongly recommend select open-source tools[4] as these tools efficiently handle chimeric Arima-HiC reads from the usage of RE-cocktail. Of note, the utility of RE-cocktail can increase the fraction of chimeric reads, making these tools[4] strongly recommended for Bioinformatics.

For more information on Arima-HiC kits and services, please visit **arimagenomics.com.**

5. Acknowledgements

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

6. References

1. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, 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 326, 289-293 (2009)
2. 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 159, 1665-1680 (2014)
3. Arima Genomics Mapping Pipeline. https://github.com/ArimaGenomics/mapping_pipeline
4. Juicer. https://github.com/theaidenlab/juicer

Emerging Applications – HiChIP/PLAC-Seq and Capture-HiC

Recently our customers have re-purposed the Arima-HiC kit towards HiChIP (also referred to as PLAC-Seq) and Capture-HiC. The motivation in HiChIP is to generate proximally ligated chromatin via Arima-HiC kit and then pull-down DNA that is bound by a pre-decided antibody (e.g. Cohesin, H3K27Ac mark). In this case, the customer purchases Arima-HiC kit from Arima and obtains antibodies from the commercial market. Capture-HiC uses biotinylated probes to enrich for custom regions of interest. Both methods are used to increase Arima-HiC resolution while significantly decreasing sequencing requirements.