ARIMA GENOMICS

diagenode
A Hologic Company

Unlock Low-Input 3D Genome Analysis with the Arima-HiC Kit

1. Introduction

The three-dimensional (3D) genome conformation has a profound impact on gene regulation, DNA replication, and DNA damage repair. Recent years have seen a rapid expansion of chromatin conformation capture techniques, including Hi-C[1,2], a sequencing-based assay that interrogates the 3D organization of the genome at unprecedented resolution.

However, previous HiC methods required high input cell numbers of 2-5 million cells, precluding many samples from 3D genome conformation analysis[2]. The Arima-HiC kit initially reduced input requirements down to 500,000 cells per experiment for many sample types and Arima innovation has reduced input requirements even further. In an effort to unlock HiC for previously inaccessible samples types, we have developed a series of low input protocols. The new Arima-HiC, coupled with a custom low input DNA library technology, drastically decrease the input requirements while maintaining high quality data.

By ensuring a higher percentage of recovered cells, our new crosslinking protocol has been optimized for those samples and sample types which are difficult to grow or isolate. This crosslinked input is then used with the standard fast and easy Arima-HiC protocol, along with our low input library prep protocol using a custom method.

2. Materials & Methods

2.1 DNA samples

The Arima-HiC low input sample prep was evaluated using GM12878 cells, which were diluted form 1M to 1K cells into new tubes at half-log order increments, pelleted and then subject to Arima-HiC in triplicate.

Highlights

Reproducible, High Quality Data Regardless of Input Amount

- High percentage of long-range is due to Arima-HiC chemistry
- No drop in long-range percentage at lower cell counts

High Resolution Features with Low Input Amounts

- Loop calling from as few as 50,000 cells
- TAD calling from as few as 10,000 cells
- A/B compartment calling from as few as 5,000 cells

Robust Workflow for Challenging Sample Types

- Capture maximum information from limited samples such as precious clinical samples, sorted cells, plants, and animal tissues
- Chromatin shearing with the Diagenode Bioruptor® Pico assures optimal fragmentation for high quality library preparation and sequencing

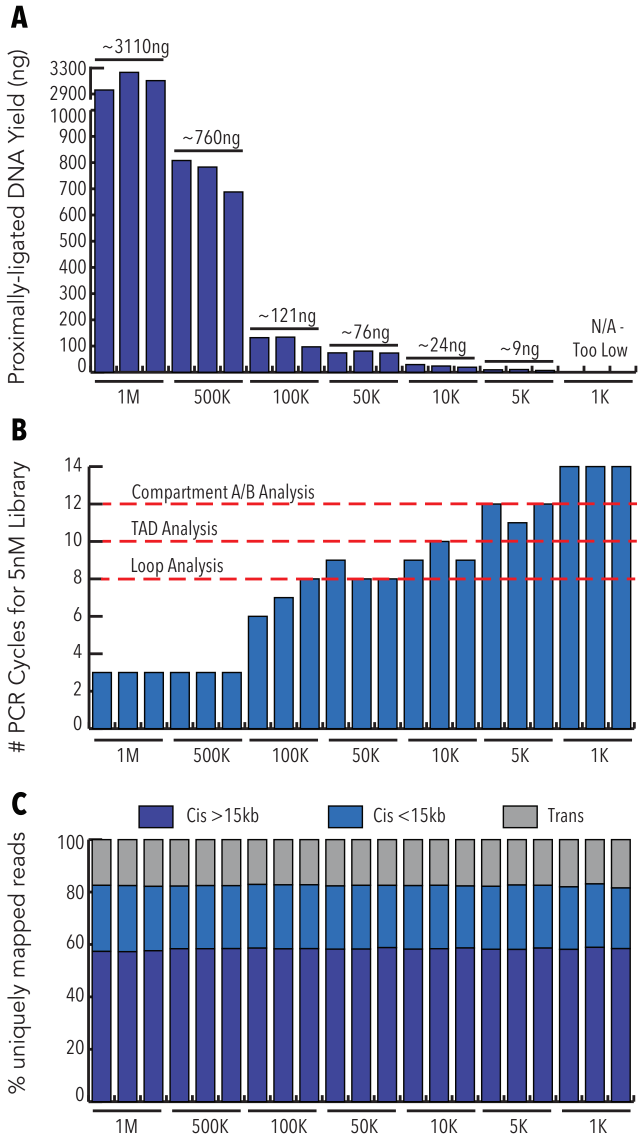

Figure 1: Experimental and sequencing QC metrics for Low Input Arima-HiC on 1K cells to 1M cells. To evaluate Arima-HiC on low input sample material, GM12878 cells were diluted from 1M cells to 1K cells into new tubes at half-log order increments, pelleted, and then subject to Arima-HiC, in triplicate. (A) Barplot showing the total amount of proximally ligated DNA extracted from each sample. For 1K cells, the DNA yield was below the limit of Qubit detection. B) Barplot showing the number of PCR cycles required to amplify each Arima-HiC library to 5nM. Overlaid are suggested PCR cycles cutoffs for various types of Hi-C analysis. For example, Hi-C libraries requiring 8 or fewer PCR cycles are complex enough for chromatin looping analysis. C) Shallow sequencing (~1M read-pairs per library) was performed. Arima-HiC data was mapped to hg19 using BWA-mem, filtered, and paired using the Arima-HiC mapping pipeline (see Github). The final mapped, paired, and monocloncal BAM file was then analyzed for the fraction of long-range cis (>15kb) reads, short-range cis (<15kb) reads, and inter-chromosomal (trans) reads. Shown is a stacked barplot of the fraction of each read type. Libraries with at least 40% cis>15kb represent high quality libraries suitable for deeper sequencing.

2.2 Arima-HiC sample preparation and next-generation sequencing

The Diagenode Bioruptor Pico was used to fragment chromatin to the required optimal fragment range to assure high quality library preparation. Briefly, samples were sheared in 100uL volume, using Diagenode Bioruptor NGS 0.65 ml Diagenode Bioruptor Pico microtubes for DNA shearing (Cat # C30010011). Cycle settings are 30sec ON/90sec OFF, and the cycle number varied depending on input amount.

Arima-HiC is a 6-hour protocol that results in the proximally-ligated DNA, which can be prepared as an Arima-HiC library using a customized library preparation solution. After library prep, the Arima-HiC libraries are sequenced in paired-end mode via Illumina next generation sequencers.

2.3 Analysis of Arima-HiC libraries

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

Hg19 reference genome was used for mapping of Arima-HiC sequence reads for evaluating Arima-HiC library quality, Arima Genomics mapping pipeline[3] was used.

To assess quality, Arima-HiC libraries were sequenced to a low- depth (~1M read-pairs per library). 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].

To evaluate the low input Arima-HiC protocol for robust and reproducible chromatin conformation analysis, Arima-HiC libraries were used for Illumina sequencing (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, topologically associated domains (TADs) and A/B compartments.

3. Results

3.1 Low Input Arima-HiC Method

By following the new low input crosslinking protocol in conjunction with the shearing with the Diagenode Bioruptor for samples of 1 million or fewer cells, users will have significantly higher sample recovery rates. This crosslinked input is then used with the standard fast and easy Arima-HiC protocol, along with our low input library prep solution.

After completing the Arima-HiC workflow, low input users leverage a custom library prep protocol. Using this custom library prep protocol provided by Arima Genomics, Arima-HiC libraries can be generated from as few as 1,000 cells (Figure 1).

3.2 Reproducible, High Quality Data Regardless of Input Amount

At Arima, we believe that the proof is in the data. Low-input samples benefit from the same unique dual restriction enzyme mixture and improved Arima-HiC chemistry as samples with greater input. Ultimately, the chemistry results in greater chromatin accessibility during digestion, manifesting in more annotated structural features when compared to previous methods.

Libraries generated using the Arima-HiC low input workflow contained a large fraction of long-range cis reads, representing high quality libraries. The percentage of long-range cis reads stayed consistent even as cell count decreased, suggesting efficient Arima-HiC chemistry regardless of cell count.

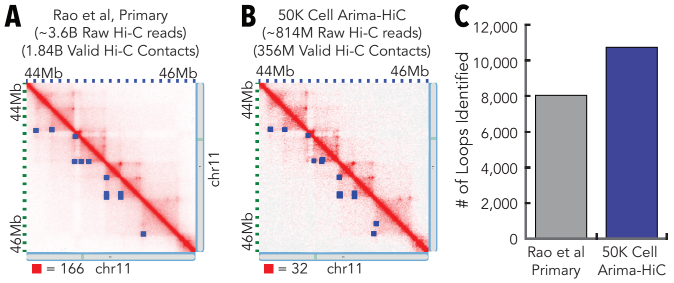

Figure 2: Comparison of Arima-HiC data generated from 50K cells with previously published Hi-C data (Rao et al) demonstrates the ability of Low Input Arima-HiC to identify chromatin loops, even at reduced sequencing depth. (A) Example of chromatin loops detected in the Rao et al Primary dataset generated from 1.84B valid Hi-C contacts using 2x100bp sequencing at chr11:44-46Mb (hg19). (B) Example of chromatin loops detected at the same locus in the 50K cell Arima-HiC dataset generated from 356M valid Hi-C contacts using 2x150bp sequencing. Both plots were generated at 10kb bin resolution. The maximum signal threshold is scaled relative to the number of valid Hi-C contacts in the Hi-C map. (C) Barplot of the total number of chromatin loops identified from Rao et al Primary and 50K cell Arima-HiC data. Loops were called using Juicer (Durand, 2016) with default parameters.

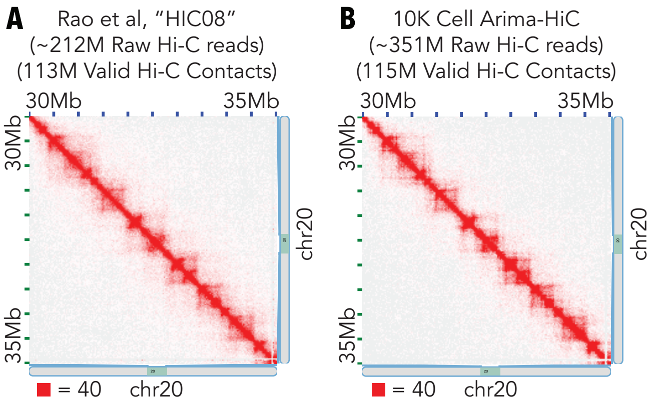

Figure 3: Comparison of Arima-HiC data generated from 10K cells with previously published Hi-C data (Rao et al) demonstrates the ability of Low Input Arima-HiC to identify TADs. (A) Example of TADs observed in the Rao et al "HIC08" dataset generated from ~113M valid Hi-C contacts using 2x100bp sequencing at chr20:30-35Mb (hg19). (B) Example of TADs detected at the same locus in the 10K cell Arima-HiC dataset generated from 115M valid Hi-C contacts using 2x150bp sequencing. Both datasets are binned at 25Kb resolution.

> **"The expanded support for low input is a significant feature as it allows researchers to perform Hi-C across precious samples such as rare cell populations or clinical samples that were previously precluded from Hi-C analysis."**
>
> – Dr. Hiruy Meharena, Massachusets Insitutte of Technology

3.3 High Resolution Features with Low Input Amounts

We then compared Arima-HiC data generated from 50,000 cells and 814 million raw read-pairs to a previously published data set[1] generated with 2-5 million cells and 3.6 billion raw read-pairs. The low input Arima-HiC data recovered known chromatin loops and TADs and identified a significant number of novel looping structures, despite fewer cells and less sequencing depth (Figure 2).

The Arima Low Input HiC methods also significantly reduces the number of cells required to identify TADs. Using our low input sample crosslinking and library prep protocols, we can achieve TAD level resolution from as few as 10,000 cells (Figure 3).

To better understand the positioning of genes between active and inactive compartments, low input Arima-HiC is able to identify A/B compartments from as few as 5,000 cells. This analysis can provide powerful new insight into gene localization from samples which are difficult to obtain or grow (Figure 4).

4. Conclusions

Robust Workflow for Challenging Sample Types

By combining the optimized chemistry of the Arima-HiC kit along with our new low input protocols, the potential applications of powerful Hi-C technology are unlocked. When studying samples that are difficult to obtain or grow, our low input solutions can help you understand genome structure across a new range of low input samples. In addition, the Diagenode Bioruptor Pico assures that chromatin is sheared to optimal fragment lengths.

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 L A, 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, Macholl, 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. https://www.cell.com/fulltext/S2405-4712(16)30219-8

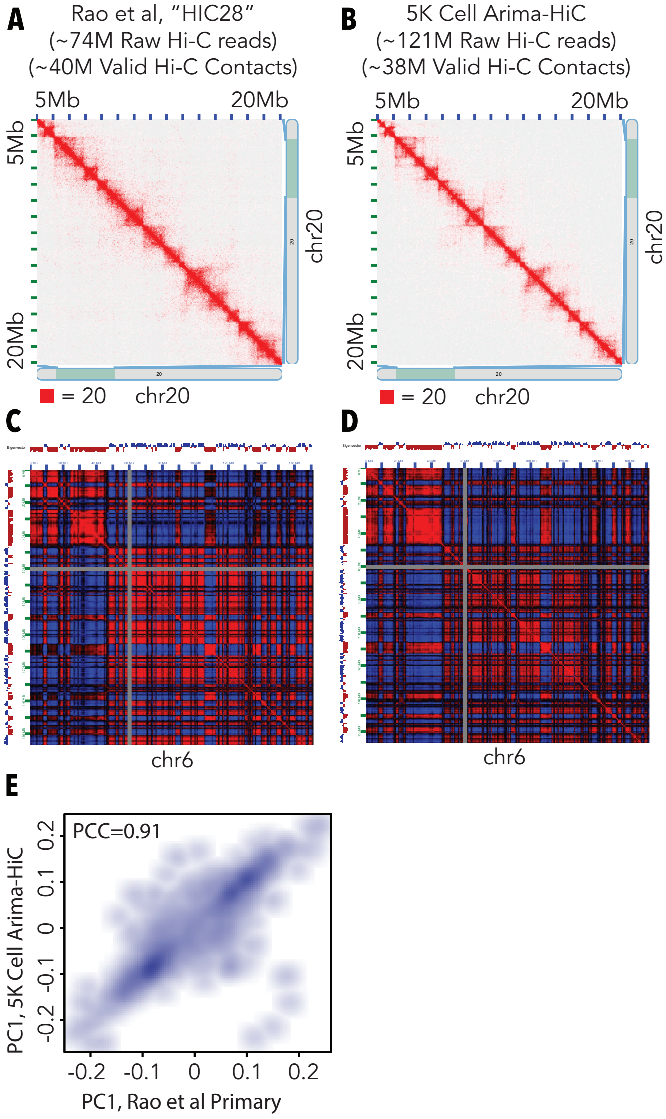

Figure 4: Comparison of Arima-HiC data generated from 5K cells with previously published Hi-C data (Rao et al) demonstrates the ability of Low Input Arima-HiC to identify Compartment A/B patterns. (A) Example of TADs observed in the Rao et al "HIC28" dataset generated from ~40M valid Hi-C contacts using 2x100bp sequencing at chr20:5-20Mb (hg19). (B) Example of TADs detected at the same locus in the 5K cell Arima-HiC dataset generated from ~38M valid Hi-C contacts using 2x150bp sequencing. Both datasets are binned at 50kb resolution. C) Pearson correlation matrix of chr20 in the same Rao et al "HIC28" dataset. D) Pearson correlation matrix of chr20 in the same 5K cell Arima-HiC dataset. In both plots, red corresponds to positive correlation, and blue to anti-correlation. The tracks surrounding each Pearson correlation matrix are the eigenvectors from a PCA analysis. The sign of the eigenvectors is arbitrary, but they reflect the Compartment A/B status. E) Scatterplot showing the genome-wide correlation between eigenvalues at all 1Mb bins.