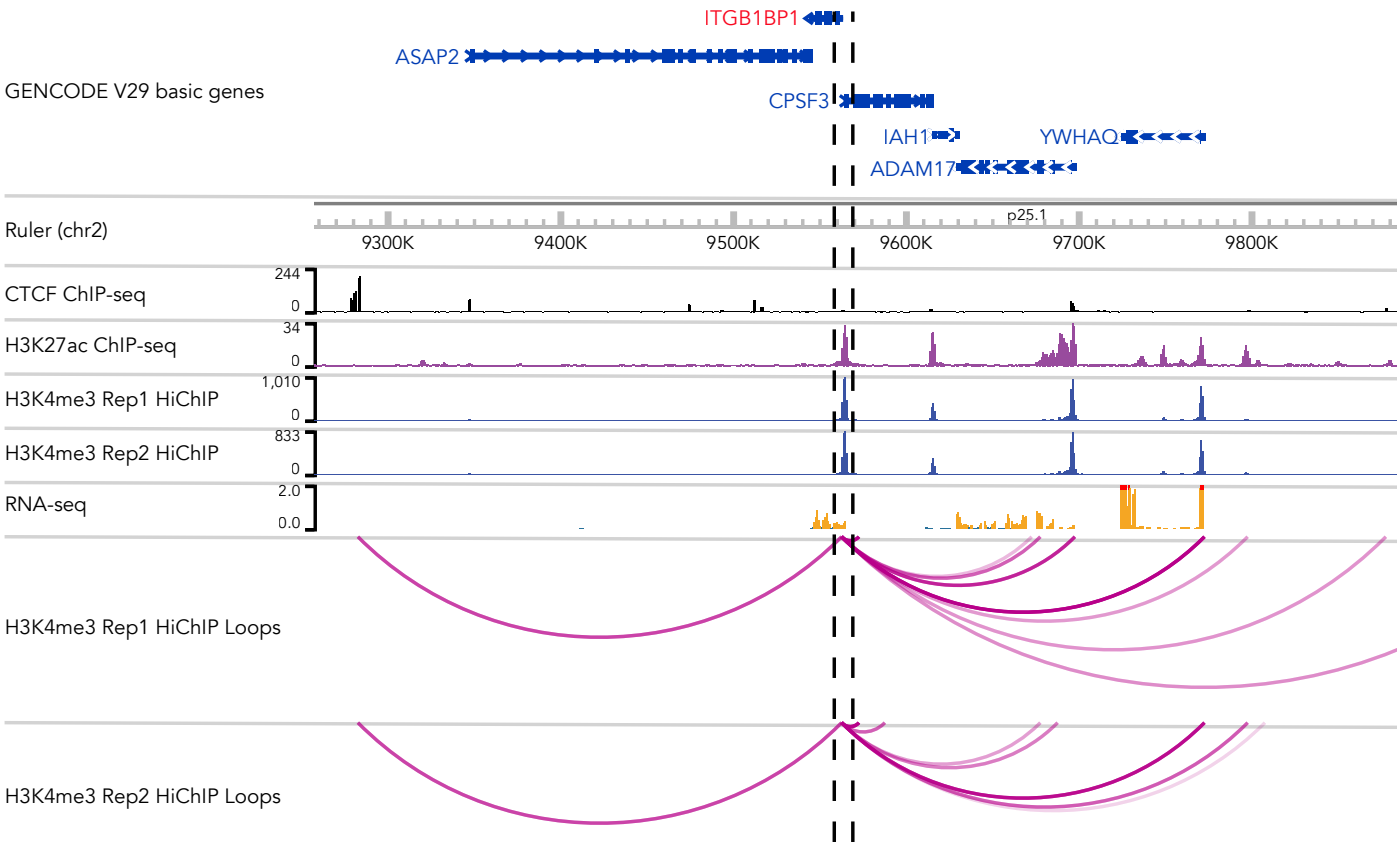
ARIMA GENOMICS

Arima-HiChIP
Accurate and cost-effective discovery of active gene regulatory interactions



1. Introduction

The three-dimensional (3D) chromatin conformation has a profound impact on numerous biological processes, including but not limited to cis regulation of gene expression. The last decade has witnessed significant expansion of genome conformation mapping technologies, such as Hi-C[1], a genome-wide sequencing-based assay designed to interrogate the 3D chromatin organization of the genome. Despite the remarkable value of Hi-C data demonstrated to date in the mapping of genome structure and analysis of gene regulation across basic and translational research[2-8], genome-wide Hi-C data is not always the most appropriate or cost-effective tool to address hypotheses pertaining to genome folding and gene regulation. To that end, HiChIP[9], also known as PLAC-seq[10], is a HiC-derived protocol that combines Hi-C with chromatin immunoprecipitation (ChIP) to enrich for chromatin interactions associated with the immunoprecipitated protein.

Conventional ChIP analyses measure the 1D localization of chromatin proteins, transcription factors, or histone post-translational modifications. However, gene regulation occurs within the 3D space of the nucleus, and conventional ChIP assays fail to examine the chromatin loops that facilitate gene regulation. HiChIP upgrades conventional ChIP-seq analysis by continuing to measure the 1D localization of chromatin proteins and by simultaneously identifying the long-range chromatin interactions associated with the enriched protein factor of interest.

To date, HiChIP has proven to be a valuable tool to address various research questions. For example, H3K4me3 and K3K27ac HiChIP have been applied to map promoter-enhancer loops at a genome wide scale and integrated with disease risk variants to help prioritize novel genes related to disease pathogenesis[11-13]. HiChIP has also been utilized to map the dynamics of long-range gene regulatory interactions in pluripotency and during differentiation[6]. More broadly, HiChIP has been used as a tool for mapping the CTCF-mediated genome folding and exploring the impact of mis-folding in cancer[14], or to understand the structural basis of promoter-enhancer communication[15].

To facilitate broad and easy access to powerful HiChIP technology, we developed an easy-to-use and reproducible Arima-HiChIP enabled kit that streamlines the HiChIP protocol via a 2-day workflow and is optimized for efficient capture of chromatin looping (3D) and ChIP enrichment (1D) data.

This application note describes the optimized Arima-HiChIP (also known as PLAC-seq) experimental and bioinformatics workflows (Fig.1) and demonstrates its ability to generate high-quality ChIP enrichment and protein-associated chromatin looping data. We demonstrate the reproducible discovery of ~60,000-80,000 chromatin loops associated with H3K27ac or H3K4me3 (Fig.2).

We also demonstrate the high performance and ease of use of the Arima-HiChIP protocol through successful evaluation of our Arima-HiChIP kit from three external beta testing sites (Fig.3). Analysis of these data for key Hi-C and ChIP performance metrics indicates robustness across cell types and protein targets, reproducibility across lab environments, and the ease-of-use and simplicity of the protocol for first-time users. Collectively, these high performing beta testing experiments lead to a significant reduction in sequencing costs compared to genome-wide HiC analyses, ranging from ~100M to ~400M reads depending on the protein target and cell type.

Lastly, the internal and externally generated Arima-HiChIP data enables the differential analysis of protein-associated chromatin loops between cell types, or other experimental and disease context. These differential analyses underscore the multiple modalities of long-range cis-regulatory chromatin looping events that play a role in governing gene expression, and highlight the unique value of adding the 3D looping context to gene regulation studies, where traditional 1D analysis of chromatin and transcription cannot reveal the full picture.

Highlights

Proven performance

- Discovery of active gene regulatory interactions at high resolution
- Demonstrated differential loop calling between sample types

Reproducible and robust

- Reproducible across lab environments, robust across cell types and protein targets

Reduced sequencing costs

- High resolution interactions with reduced sequencing depth

Ease of use

- 2 day workflow with simple protocol for first time-users

Expanding platform

- Potential applications with tissue and transcription factor proteins

"We were very pleased with the informative nature and quality of the Arima-HiChIP data. Not only do we observe intra-chromosomal cis-regulatory interactions across cancer genomes, but also inter-chromosomal ectopic cis-regulatory interactions on rearranged chromosomes. We think these data will help better delineate the relationship between genetic and epigenetic driver events in cancer."

– Peter Scacheri, PhD, Professor, Case Western University School of Medicine

2.Materials and Methods

2.1 Samples

The Arima-HiChIP libraries and sequencing data were evaluated on a range of cell types through internal experiments and a beta testing program involving 3 leading labs studying epigenetic gene regulation. The beta testers were provided with an Arima-HiC+ kit (P/N: A101020), antibodies for either H3K-27ac or H3K4me3 and other reagents required for ChIP, and ~3M crosslinked GM12878 Human Lymphoblastoid cells (LCLs) as control, in replicate. All beta testers also evaluated two of their own samples, in replicate, comprising a variety of cell types including hepatocellular carcinoma cells (HepG2), B-cell lymphoma cells (OCI-Ly7), human fetal fibroblasts (IMR90), and IPSC-derived neurons.

2.2 Arima-HiChIP Library Preparation

Arima-HiChIP is a 2-day protocol that results in proximally-ligated DNA that was associated with an immunoprecipitated protein of interest. This enriched DNA is then prepared as an Arima-HiChIP library in 1-day using a pre-validated commercially available library prep kit. After library prep, the resulting Arima-HiChIP libraries are sequenced in paired-end mode via Illumina next generation sequencers (Fig.1A).

2.3 Bioinformatic Analysis of Arima-HiChIP data

The Arima-HiChIP libraries were evaluated, both internally and by beta testers, using numerous metrics, but with particular emphasis on the qualities of the long-range chromatin conformation and ChIP enrichment signals, the number of raw sequencing reads needed for reproducible H3K4me3- and H3K27ac-associated loops across replicates, and the ease of use of the experimental protocols and bioinformatics pipeline.

To assess the initial quality of the long-range (3D) and ChIP enrichment (1D) signals, Arima-HiChIP libraries were sequenced to a low depth (~1M paired-end reads). The resulting Arima-HiChIP sequencing data were then mapped to a reference genome using a modified version of the open-source MAPS pipeline (Juric et al, 2019) adapted for Arima-HiChIP data (Fig.1B; GitHub Link) and two types of signals were enumerated: (1) long-range *cis* interactions that are captured from proximity ligation during the Hi-C portion of the overall HiChIP workflow, and (2) the fraction of reads enriched at ChIP peaks, produced as a result of the ChIP portion of the overall HiChIP protocol. The product of these two features (3D and 1D) in the HiChIP data were then used to determine the sequencing depth needed for each sample to obtain sufficient long-range *cis* interactions anchored at each ChIP peak for reproducible chromatin looping analyses. All of the aforementioned data features and statistics (plus several more) are tallied in a convenient quality control table output by the MAPS pipeline, and can be copy/pasted into the Arima-HiChIP QC Worksheet for tabulation and analysis.

For deep sequencing analysis, Arima-HiChIP data were again mapped to a reference genome using a modified version of the same open-source MAPS pipeline adapted for Arima-HiChIP data. However, in addition to the quality metrics described above, the deep sequencing data allows the discovery of thousands of protein-associated chromatin loops at a genome-wide scale using MAPS pipeline. These QC metrics, as well as the enumeration of loops, are again reported in a quality control table output by the MAPS pipeline, and can be copy/pasted into the Arima-HiChIP QC Worksheet for tabulation and analysis.

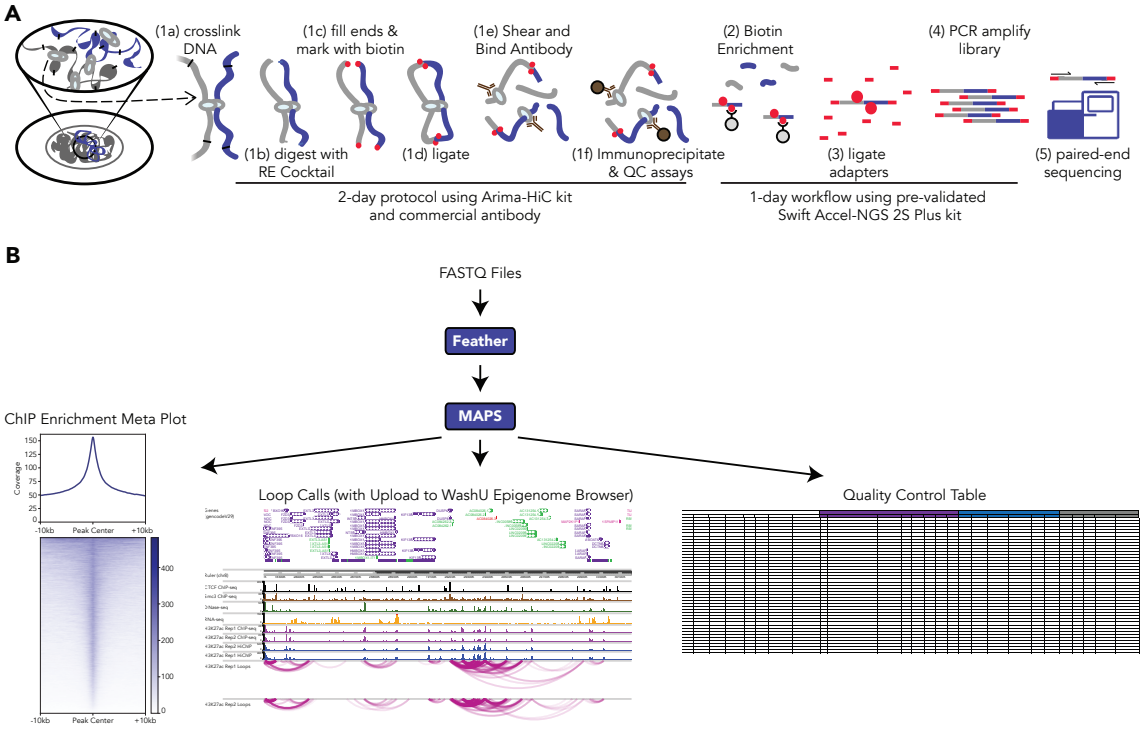


Figure 1. The Arima-HiChIP Experimental and Bioinformatics Workflows. A) The Arima-HiChIP workflow is a streamlined protocol that results in immunoprecipitated biotin-labeld proximally ligated DNA that was associated with the immunoprecipitated protein target. The immunoprecipitated DNA is enriched for biotin and prepared as a library and sequenced in paired-end mode on Illumina sequencing instruments. **B)** Arima-HiChIP data can be analyzed using the Arima-MAPS pipeline, producing ChIP enrichment metaplots and heatmaps around known ChIP peaks (if available), loop calls that can be uploaded to the WashU Epigenome Browser, and a QC table for high level and detailed analysis of data quality.

"Enriching for active chromatin features through H3K27ac or H3K4me4 HiChIP enables our discovery of high-resolution chromatin looping events that facilitate gene control in muscle stem cells during the transition from a state of quiescence to their commitment toward differentiation into regenerating myofibers, during muscle regeneration in healthy conditions, aging or diseases such as muscular dystrophies. The combined analysis of this HiChIP data with our genome-wide Hi-C data is pivotal to our understanding of both the broader compartmentalization and topolological frameworks that influence gene regulation, but also the finer resolution connections between specific cis-regulatory elements."

– Pier Lorenzo Puri, M.D., PhD, Professor, Sanford Burnham Presyb

3.Results

3.1 Reproducible discovery of active gene regulatory chromatin loops

To investigate the performance of Arima-HiChIP data, we generated deep sequencing Arima-HiChIP data from replicates of human lymphoblast cells (GM12878), in replicate, when enriching for H3K4me3 and H3K27ac. The data was analyzed using the Arima-MAPS pipeline, and we benchmarked the degree of ChIP enrichment in the Arima-HiChIP data compared to ChIP-seq data generated by ENCODE on the same sample (Fig.2A). For H3K4me3 Arima-HiChIP, we observed a ~6-fold peak to local background enrichment, which is consistent with the matched ChIP-seq data. Similarly, for H3K27ac HiChIP, we observed a ~3-fold peak to local background enrichment, consistent with the matched ChIP-seq data.

The number of raw reads required for reproducible HiChIP loops depends on the quality of the long-range *cis* interactions and ChIP enrichment features of the HiChIP data, and, the number of underlying 1D ChIP peaks that serve as the anchor points for loop discovery. The more 1D peaks, the more total reads are needed to provide sufficient coverage of long-range interactions originating from the ChIP peaks. Based on the quality features of the Arima-HiChIP data described above, we determined the required number of reads for generation of reproducible and robust HiChIP loops (Fig.2B). For comparison sake, plotted alongside the HiChIP data is the recommended number of reads for Arima-HiC data, as well as the ENCODE recommendation for conventional *in situ* HiC data. As expected, the number of reads required for H3K27ac HiChIP exceeds that of H3K4me3, driven by both antibody performance and more underlying 1D peaks in the case of H3K27ac.

To illustrate the type of 1D ChIP enrichment and 3D chromatin looping data obtained from Arima-HiChIP, a snapshot of the MYC locus is provided (Fig.2C). Plotted below the chromatin, DNA accessibility, and transcriptome tracks obtained from ENCODE are the 1D coverage tracks of the Arima-HiChIP data. The H3K27ac HiChIP 1D signal is nearly identical compared to the H3K27ac ChIP-seq signal directly above, suggesting not only the strong signal to noise of the ChIP enrichment in the HiChIP data, but the similar capture of 1D protein localization peaks compared to conventional ChIP-seq.

Moreover, a subset of the ~75,000 H3K27ac-associated chromatin loops are shown at the MYC locus (Fig.2C). These loops appear to connect active cis-regulatory elements, as well as sites not marked by H3K27ac but rather occupied to CTCF, SMC3, or both. To analyze the reproducibility of H3K27ac-associated chromatin loops, we calculated the percentage of chromatin loops from one replicate that are also found in the other replicate (Fig.2D). We observe approximately 85% replicate reproducibility using this metric for H3K27ac HiChIP, and ~74-87% replicate reproducibility for H3K4me3 HiChIP (Fig.2E).

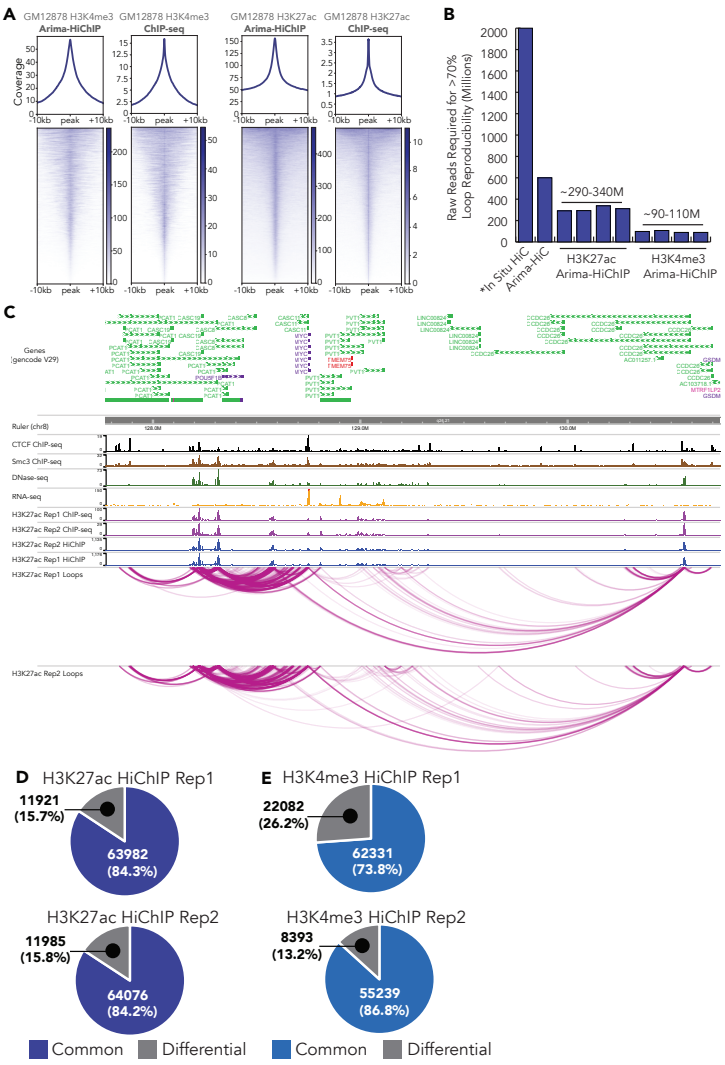


Figure 2. Reproducible discovery of active gene regulatory chromatin loops. **A)** ChIP enrichment metaplots comparing Arima-HiChIP to ChIP-seq. **B)** Estimated number of raw read-pairs required for reproducible chromatin loop discovery for in situ Hi-C, Arima-HiC, and Arima-HiChIP. **C)** WashU Epigenome Browser snapshot of the MYC locus in human lymphoblast cells. The darker the purple arcs, the more stastically significant the loop. **D)** Chromatin looping reproducibility analysis in our H3K27ac Arima-HiChIP data. The pie charts depict the total number of loops from one replicate that were identified (marked as "common") or not identified (marked as "differential") in the other replicate. **E)** Chromatin looping reproducibility analysis in our H3K4me3 Arima-HiChIP data.

3.2 Successful evaluation of Arima-HiChIP by 3 external beta testing sites

To validate the performance of Arima-HiChIP, the Arima-HiChIP workflow was evaluated independently by 3 beta testers (Fig.3). Using Arima-HiC+ kits, the beta testers consistently generated a high percentage of long-range (>15kb) cis interactions (Mean=39% for H3K4me3; Mean=42% for H3K27ac) as well as an accompanying high degree of ChIP enrichment evidenced by a high percentage of HiChIP reads overlapping known ChIP peaks (Mean=80% for H3K4me3; Mean=51% for H3K27ac). Collectively, these data indicate the strong performance of the Arima-HiChIP protocol and kit, and the high-quality of the resulting Arima-HiChIP libraries (Fig.3A).

Of note, the standard deviation in the proportion of long-range cis interactions (1.3% for H3K4me3 and 2.8% for H3K27ac) and reads overlapping known ChIP peaks (1.4% for H3K4me3 and 3.9% for H3K27ac) was minimal across the 3 beta testing labs using control GM12878 cells, underscoring the reproducibility and robustness of the workflow across lab environments and user experience levels.

Based on the quality features of the Arima-HiChIP data described above, we determined the required number of reads for generation of reproducible and robust HiChIP loops (Fig.3B). We observed a mean of 113M raw read-pairs required (STDEV=10M) for H3K4me3 HiChIP loop discovery and 251M (STDEV=52M) raw read-pairs for H3K27ac HiChIP loop discovery. Similar to above and as expected, we observe minimal variance in the sequencing depth requirements across the 3 beta testing labs using control GM12878 cells.

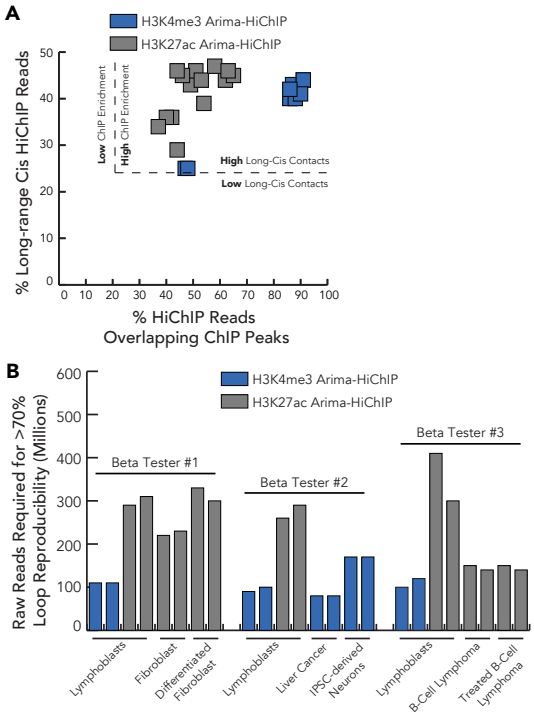


Figure 3. Successful evaluation of Arima-HiChIP by 3 external beta testing sites. A) Scatter plot showing the percentage of Arima-HiChIP reads representating long-range (>15kb) cis contacts (y-axis) and the percentage of HiChIP reads overlapping known ChIP peaks. For all of these experiments, 1D peaks have previously been identified and enabled this form of ChIP enrichment analysis for assay benchmarking. Approximate cutoffs for success are illustrated with dotted lines, where successful libraries fall into the upper right quadrant of the scatter plot. Performance metrics for H3K4me3 Arima-HiChIP are colored in blue and H3K27ac Arima-HiChIP are colored in gray. **B)** Bar plot showing the number of raw read-pairs required for deep sequencing for each of the 24 reactions performed by the 3 beta testing sites on a variety of sample types. Sequencing requirements for H3K4me3 Arima-HiChIP are colored in blue and H3K27ac Arima-HiChIP are colored in gray.

"We have been homebrewing HiChIP because it provides critical insights that were previously obtained from performing separate Hi-C and ChIP-Seq experiments, Arima-HiChIP generated superior data quality and reproducibility even with our most challenging terminally differentiated neuronal samples, where our homebrew version previously failed, thus saving us from substantial repeat costs"

– Yin Shen, PhD, Principal Investigator, UCSF

3.3 Arima-HiChIP uncovers cell-type specific chromatin loops associated with active promoters.

To illustrate the utility of Arima-HiChIP data in exploring differential gene regulation across experimental samples, we identified a significantly different chromatin looping landscape associated with the active promoter of ITGB1BP1 (Fig.4). While this gene is expressed in both cell types as evidenced by the H3K4me3 peak at the promoter region and transcriptional signal across the gene body, the chromatin loops are largely skewed downstream in lymphoblasts towards a series of other active promoters (evidenced by H3K4me3) and putative enhancers (evidenced by H3K27ac ChIP signal but not H3K4me3). Strikingly, the chromatin loops are largely skewed upstream in IPSC-derived neurons, towards a broad range of H3K27ac signal and the active promoter of a neuron-specific gene, ASAP2. These interactions occur upstream, despite several other active promoters and putative enhancers downstream that are also observed in lymphoblast cells. This observation may be described as "enhancer-promoter switching". One possible mechanism facilitating this differential looping landscape could be the CTCF peak at the ITGB1BP1 gene found only in IPSC-derived neurons.

Further analysis of chromatin loops also exemplifies two additional modalities of long-range gene regulation. For example in the H3K4me3-associated loops anchored at ITGB1BP1, it is observed that some promoter-enhancer loops skip over genes, demonstrating the well-known observation that enhancers do not always regulate their nearest gene (REF). It is also observed that promoters significantly interact with other promoters. Taken together, the joint analysis of architectural protein occupancy, chromatin activity, transcription, and chromatin looping provide a more comprehensive view of dynamic gene regulatory mechanisms across cell types.

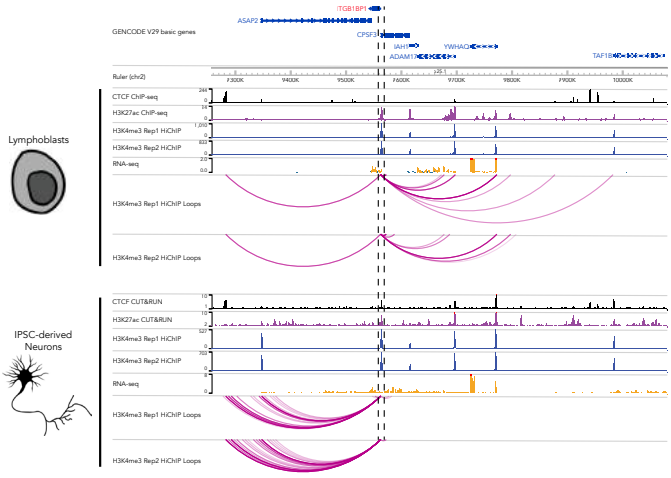


Figure 4. Arima-HiChIP uncovers cell-type specific loops associated with active promoters. WashU Epigenome Browser snapshot of the ITGB1BP1 locus in human lymphoblast cells (top half) and IPSC-derived neurons (bottom half). Shown for each cell type is CTCF occupancy, H3K27ac ChIP-seq, RNA-seq, H3K4me3 HiChIP 1D coverage, and H3K4me3 HiChIP loops. ChIP and RNA-seq data in lympoblasts were obtained from ENCODE, and CUT&RUN and RNA-seq data in IPSC-derived neurons were obtained from the Yin Shen Lab (UCSF). Of note, only loops anchored at the ITGB1BP1 promoter are shown and other loops not associated with ITGB1BP1 promoter region are masked for illustrative and comparative purposes. The darker the purple arcs, the more stastically significant the loop.

4. Conclusions

In summary, the Arima HiChIP protocol, available with the Arima-HiC+ kits, is an easy-to-use and reproducible HiChIP workflow that produces high quality libraries with robustness across cell types. The high quality data is exemplified in internal and external evaluations of the Arima-HiChIP workflow, as evidenced by the strong ChIP enrichment and efficient capture of long-range cis interactions, leading to significantly reduced sequence cost to obtain loops associated with active chromatin marks. These loops, when analyzed in conjunction with other chromatin and transcriptional datasets, provide an integrative topological view of gene regulatory mechanisms within the 3D nucleus and enable the discovery of gene regulatory mechanisms across cell types, disease states, and a variety of other research contexts.

"We have been very pleased with the performance of the Arima Hi-C/HiChIP kit. It has improved reproducibility and overall quality while saving us time and sequencing cost. We particularly value the straightforward quality control steps and the ease of use. The Arima Technical Support Team has also been outstanding. Overall, a great value."

– Tamas Ordog, Professor, Mayo Clinic

5. Acknowledgements

We would like to acknowledge the 3 external beta testing sites for their time in evaluating the Arima-HiChiP workflow and their valuable feedback. We would also like to acknowledge the recipients of the Arima-HiChIP Grant Program for their collaboration and partnership throughout the development of the Arima-HiChIP workflow.

6. References

1. Lieberman-Aiden, Science, 2009
2. Rao, Cell, 2014; Link, Cell, 2018
3. Rajarajan, Science, 2018
4. Hewitt, JBC, 2019
5. Saldana-Meyer, Molecular Cell, 2019
6. Di Giammartino, Nature Cell Biology, 2019
7. Ooi, BMJ, 2019
8. Espeso-Gil, Genome Medicine, 2020
9. Mumbach, Nature Methods, 2016
10. Fang, Cell Research, 2016
11. Mumbach, Nature Genetics, 2017
12. Jeng, J Investigative Dermatology
13. Nott, Science, 2019
14. Flavahan, Nature, 2019
15. Weintraub, Cell, 2017

ARIMA GENOMICS • 6404 Nancy Ridge Drive, San Diego, CA 92121
Office: 858.275.6463 • techsupport@arimagenomics.com • arimagenomics.com