Application Note

Agilent
Trusted Answers

Utilizing Agilent SureSelect XT HS2 Target Enrichment and Arima-HiC to Improve Resolution of High Throughput Chromatin Conformation Capture

Authors

Anthony Schmitt,
Arima Genomics, Inc.

Justin Lenhart,
Agilent Technologies, Inc.

Abstract

High throughput chromatin conformation capture (HiC) is a technique used to generate genome-wide chromatin architecture maps. These maps, when acquired with high resolution, can reveal unique details of chromatin organization and dynamics. However, to achieve the increased resolution required to visualize individual chromatin loops, a significant increase (hundreds of millions of reads, or more) in sequencing depth is required. The required depth for whole-genome HiC renders this approach cost-prohibitive, which may lead to poor resolution and/or exclusion of valuable experimental conditions or controls. To address this problem, Agilent has partnered with Arima Genomics and generated a streamlined target enrichment workflow for HiC (capture HiC). This approach allows researchers to focus HiC reads on targeted areas of the genome, reducing sequencing read requirements (and, hence, cost) for high-resolution HiC workflows. Here we present data using the Arima-HiC kit upstream of the Agilent SureSelect XT HS2 target enrichment platform that shows resolution of targeted regions down to 500 bp.

Introduction

Chromosomes within the eukaryotic nucleus occupy discrete chromosomal territories. Within these territories, chromosomes adopt highly complex architectures that maintain a compact structure, yet still support dynamic cellular processes such as transcription, replication, and DNA damage repair. Driven by the rapid development of new genomic tools[1], our understanding of this architecture and its impact on these processes is growing rapidly.

One such technology is high throughput chromatin conformation capture (HiC), which probes the 3-dimensional structure of chromatin via proximity ligation followed by massive parallel sequencing. A typical HiC workflow begins with crosslinking the cells or tissues with formaldehyde to preserve the integrity of the nuclear architecture. The crosslinked chromatin is then fragmented via restriction digestion, the resulting 5′overhangs are filled in with a biotinylated nucleotide, and proximal ligation is performed[2]. This process produces a library of chimeric DNA molecules that possess information regarding the proximity of the initial DNA fragments. The proximal-ligated biotinylated junctions are enriched using streptavidin beads and an NGS sequencing library is constructed. After sequencing, the resulting reads are mapped to the genome to create contact matrices, revealing the chromatin conformation within the cell population.

The power of HiC lies in the ability to capture and reveal the unique topological characteristics of the genome. At low resolutions, HiC experiments reveal A/B compartments as well as topologically associating domains (TADs), while individual chromatin loops can be seen at the highest resolutions. These individual structures, when combined with epigenetic data, provide valuable insight into how chromosomal architecture can impact cellular processes (e.g., enhancer activation of a distal promoter). To achieve the increased resolution required to reveal these interactions, a significant increase in sequencing reads is needed. An effective increase in the resolution (factor of x) requires a quadratic increase (x^2) in the number of sequencing reads[2]. Because a whole-genome HiC experiment designed to interrogate individual chromatin loops requires hundreds of millions to billions of reads, high-resolution HiC can be cost-prohibitive.

One practical approach to increase resolution is to use target enrichment to reduce the necessary number of sequencing reads (and hence sequencing costs). This is done by enriching biotinylated RNA baits that hybridize to pre-determined genomic regions of interest. This sampling of the genome enables these regions to be sequenced to far greater depths for the same or reduced sequencing costs and increases the resolution of the contact matrices in these regions. Here we outline how our SureSelect target enrichment technology, when combined with the Arima-HiC kit (Arima Genomics), produces a simplified capture HiC workflow (Figure 1).

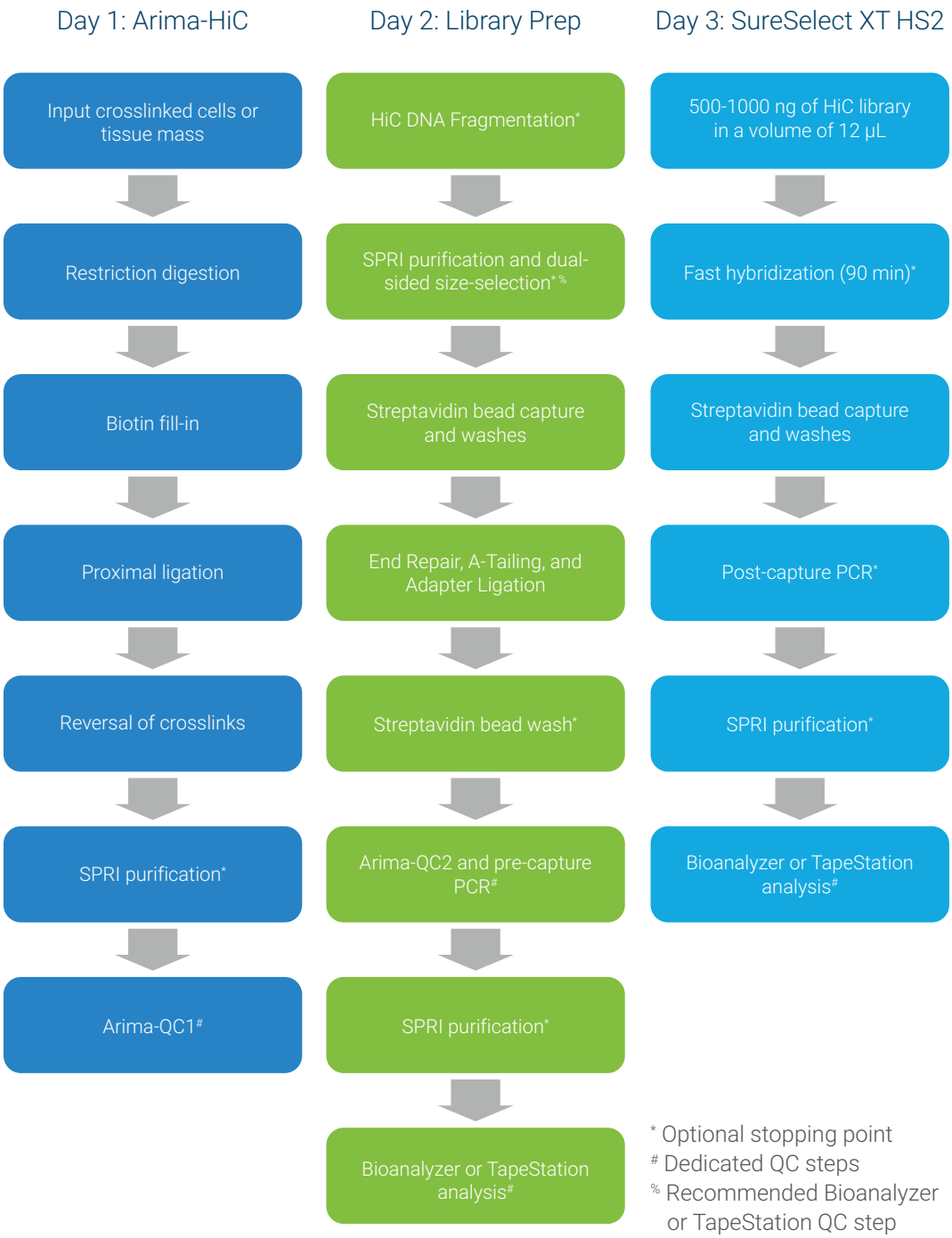


Figure 1. Illustration of the joint Arima / Agilent capture HiC workflow. While the capture HiC workflow is split across three days, there are additional safe stopping points (indicated by *) to provide additional flexibility. Dedicated QC steps are identified by #. Optional (though recommended) Bioanalyzer or TapeStation QC steps are indicated by %. Indexing is supported by direct adapter ligation or during pre-capture PCR. Total time and hands-on time estimates during the protocol is as follows. Day 1: 7.25 hours total time, with 4 hours hands-on time. Day 2: 8 hours total time, with 2.5 hours hands-on time. Day 3: 5.5 hours total time, with 1.5 hours hands-on time. These estimates are based on 4-8 samples processed in parallel with the use of a multi-channel pipette.

Note: crosslinking can be performed at the beginning of Day 1 and will add an hour to the total sample processing time.

Experiment

Materials and methods

HiC Sample Preparation Using the Arima-HiC kit

Ten separate mammalian cell lines (Table 1) were expanded using standard cell culture practices. Cells were grown in manufacturer-recommended media. Adherent cells were passaged upon reaching confluency via enzymatic release with trypsin. Cells were harvested and fixed using 2% formaldehyde and quenched using Stop Solution 2 (provided as part of the Arima-HiC kit (P/N: A510008)). Five hundred thousand crosslinked cells were used as input into the Arima-HiC kit. HiC libraries were generated for each cell line according to the Arima-HiC Mammalian Cell Line User Guide (Doc #A160134). For quality control (QC), the percent of biotin-labeled DNA was quantified using the Arima-QC1 protocol.

Library Fragmentation of Hi-C DNA

The Arima-HiC protocol produces large DNA molecules that must be fragmented prior to library preparation. The Diagenode Bioruptor Pico was used according to manufacturer recommendations. The sample was prepared at a final volume of 100 µL in Elution Buffer (provided as part of the Arima-HiC kit) and then placed into a 0.2 mL microtube (Diagenode Cat # C30010020). The specific settings to fragment the Hi-C libraries to an average size of 400 bp were 4 cycles of 30 seconds ON and 90 seconds OFF. After fragmentation, a double-sided size selection was performed as indicated in the protocol to produce a size distribution between 200 and 600 bp.

Table 1. Ten cell lines that were targeted by the Arima and Agilent Capture-HiC workflow.

Sample	Origin
GM12878	Blood
IMR90	Fibroblast
HCC1395	Blood
HeLa	Cervical*
MCF-7	Breast*
HCT116	Colorectal*
A549	Lung*
SKBR3	Breast*
HCC1395BL	Breast*
K562	Breast*

**Indicates cell line of cancer origin*

Library Preparation

After fragmentation and size selection, biotin-labeled and proximally-ligated DNA was enriched for ligation junctions using Enrichment Beads (provided as part of the Arima-HiC kit). Afterwards, library prep was performed to append Illumina-compatible sequencing adapters while DNA remained bound to Enrichment Beads following a custom Arima-HiC Library preparation user guide (Doc # A160139). The library was PCR amplified utilizing the KAPA HotStart PCR master mix (Cat # KK8500) according to the instructions in that protocol, with 11 cycles of amplification. The libraries were purified post-PCR using AMPure beads (Beckman Coulter Cat # A63880).

Quality Assessment of Input Material and Sequencing Libraries

Nucleic acid sample quality was assessed on either the Agilent 2100 Bioanalyzer (Agilent Technologies, P/N G2939BA) or TapeStation 4200 (P/N G2991AA) system. The High Sensitivity DNA Kit (Agilent Technologies, P/N 5067-4626) or the DNA 1000 kit (Agilent Technologies, P/N 5067-1505) evaluated the quality of the libraries throughout the workflow.

Custom Panel Design and SureSelect Target Enrichment

Target enrichment of Arima-HiC libraries was carried out using a custom design Agilent SureSelect panel. The resulting design (Custom Design ID: S3237574) spanned nine genes or gene clusters and their surrounding regions (Table 2). Internal SureDesign masking tools were utilized to prevent probe placement to repetitive DNA sequences within the targeted regions. Custom probe boosting for the SureSelect XT HS fast hybridization protocol was utilized. Target enrichment was performed using the SureSelect XT HS2 target enrichment system for the Illumina paired-end multiplexed sequencing library (Agilent Technologies, P/N G9987A) according to the protocol (G9983-900000). Ten cycles of post-capture PCR were used.

Sequencing and Data Analysis

Sequencing libraries were analyzed on an Illumina HiSeq X by paired-end sequencing using a 2 × 150 bp read format. Target sequencing depth was 90 to 100 million read pairs per sample or approximately 10 million read pairs per megabase of targeted sequence. GRCh37 (hg19) was used as the reference genome. Using the Juicer pipeline, read pairs were aligned to the genome, removing duplicate reads and multi-mapping reads. After processing the reads, Juicer produced summary statistics, a file containing all Hi-C contacts, and annotated structural features[3]. Specific scripting recommendations are provided below. Data was visualized using Juicebox[4].

Table 2. Targeted genes or gene clusters of interest and region sizes that were targetd by the Agilent design (S3237674).

Target Genes	Target Size (Mb)
MYC	3.00
EGFR	0.87
BCR	0.62
PTEN	1.73
NTRK1	0.86
HOXB	0.91
JUND	1.06
P53	0.80
PDGRFA	3.05
Total	**12.9**

- **juicer.sh -d $DATA_DIR/$SAMPLE -p hg19.chrom.sizes -y hg19_ GATC_GANTC.txt -z hg19.fa -D $JUICER_DIR -t 12 &>$DATA)DIR/$SAMPLE/juicer.log**
- The percent on-target rate was calculated using a few comands leveraging SAMtools, BEDtools, and Linux commands, as described below:
- **Step 1:** Count the total number of lines in the mapped and de-duplicated BAM file:
  samtools view $DATA_DIR/$SAMPLE/$SAMPLE.sorted.nodup.bam | wc -1
- **Step 2:** Divide the above number by 2 to obtain to total number of mapped and de-duplicated read pairs.
- **Step 3:** Extract the on-target read-pairs:
  bedtools intersect -u -bed -a /$SAMPLE.sorted.nodup.bam -b $COVERED_REGIONS.bed>$SAMPLE.on-targeted.bed
- **Step 4:** Count the number of on-target read pairs*
  cut -f4 $SAMPLE.on-targeted.bed | cut -f1 -d"/" | sort | uniq | wc -1
- **Step 5:** To obtain the on-target rate, which we define as the fraction of read pairs that have at least one read end mapping to the targeted region, divide the number of on-target read pairs (**Step 4**) by the number of total mapped and de-duplicated read pairs (**Step 2**).

Results and Discussion

Capture HiC Workflow Optimization

Traditionally, capture HiC based approaches have suffered from labor-intensive protocols, prolonged workflow durations, and high sequencing requirements. To remediate these limitations, we coupled the proven performance of SureSelect XT HS2 target enrichment with the Arima-HiC kit from Arima Genomics to develop a highly simplified and streamlined capture HiC workflow (Figure 1). The Arima-HiC workflow provides high quality HiC libraries from a wide range of sample types and input amounts while offering several important protocol checkpoints designed to monitor HiC performance prior to target enrichment. In addition, the Arima-HiC kit employs a restriction enzyme cocktail used for chromatin digestion, thereby increasing the number of restriction fragments available for proximity ligation, increasing the assay's resolution and library complexity. The SureSelect XT HS2 target enrichment workflow is ideal for this application due to its high performance with challenging sample types. In addition, the 6-hour workflow is driven by a single 90-minute hybridization that reduces the overall workflow from sample to enriched library to as little as three days.

Uncover Chromatin Conformation and Epigenomic Features Around Targeted Genes

To demonstrate performance, we generated a custom SureSelect panel design that targeted the contact domains surrounding nine genes or gene clusters, several of whose functions have been well-studied in cancer research (Table 2). Capture HiC libraries were generated from ten samples representing a wide range of cell types from various tissue origins, including seven cancer and three non-cancer samples (Table 1). All HiC libraries were assessed before hybridization capture using the Arima-QC metrics. For example, the Arima-QC1 metrics indicated that all samples contained 46.3 to 51.2% biotin-labeled ligation products by mass. Libraries were pooled and sequenced on the Illumina HiSeq X platform using paired-end 150 bp reads. Table 3 shows a summary of the quality and capture HiC summary statistics calculated using the open-source Juicer platform. Despite sequencing each library to approximately 2500X coverage, the duplicate rates were on average modest at 25.17% (range: 15.03 to 47.13%). The on-target rate, defined as the fraction of read pairs that have at least one read end mapping to the targeted region, averaged 60.7% (range: 47.9 to 67.8%). Most importantly, about 31% of valid HiC contacts qualified as being both on-target and long-range (>20 kb) cis reads (average 13.7 million reads), indicating the combination of Arima-HiC and SureSelect target enrichment produced high-quality, high-resolution HiC data.

Table 3. Sequencing and quality metrics of the produced libraries from the Capture HiC workflow.

Sample	Raw Read-Pairs	Duplicate Rate	Valid HiC Contacts	Long-range Cis (>20 kb) Reads	On-target Rate	On-target Long-range Cis Reads
GM12878	101,016,465	21.57%	50,756,573	25,496,777	55.3%	14,099,718
IMR90	76,981,663	19.76%	40,854,435	22,198,910	56.4%	12,520,185
HCC1395	109,663,060	24.63%	52,759,537	29,101,519	57.8%	16,820,678
HCC1395-BL	83,460,957	25.22%	39,446,603	19,505,755	62.7%	12,230,108
HeLa	84,117,022	15.03%	43,885,667	23,248,838	47.9%	11,136,193
MCF-7	61,646,312	22.63%	32,782,338	15,292,882	67.8%	10,368,574
HCT116	75,433,996	26.27%	37,253,141	19,618,120	65.0%	12,751,778
A549	76,075,806	15.94%	40,919,803	21,245,985	53.7%	11,409,094
SKBR3	158,357,436	47.13%	52,276,171	28,623,726	75.2%	21,525,042
K562	102,585,018	33.54%	46,566,539	21,326,191	65.2%	13,904,677
Averages	**92,933,774**	**25.17%**	**43,750,081**	**22,565,870**	**60.7%**	**13,676,805**

Global contact matrices were generated and visualized in the open-source Juicebox tool. Briefly, these contact matrices possess both an X- and Y-axis representing chromosomal coordinates. Each point within the matrix represents contacts made between the representative loci. The number of successfully captured contacts is visualized using a color spectrum, with increased contacts corresponding to an increase in color intensity. When enriched loci are visualized, high-resolution contact profiles emerge. Figure 2 shows the IMR90 cell line capture HiC data for a select set of captured regions. Each region is binned to 5 kb, which reveals unique details about the topological architecture that spans each respective genetic locus. When combined with publicly available ENCODE data for proteins involved in higher-order chromatin structure (CTCF, the cohesion complex smc3), the H3K27ac active enhancer mark, and RNA-Seq data, the connection between hierarchical architecture and gene regulation becomes apparent. The expected organization of chromatin loops bound by CTCF and cohesin localized to the boundaries is observed, validating the contact maps.

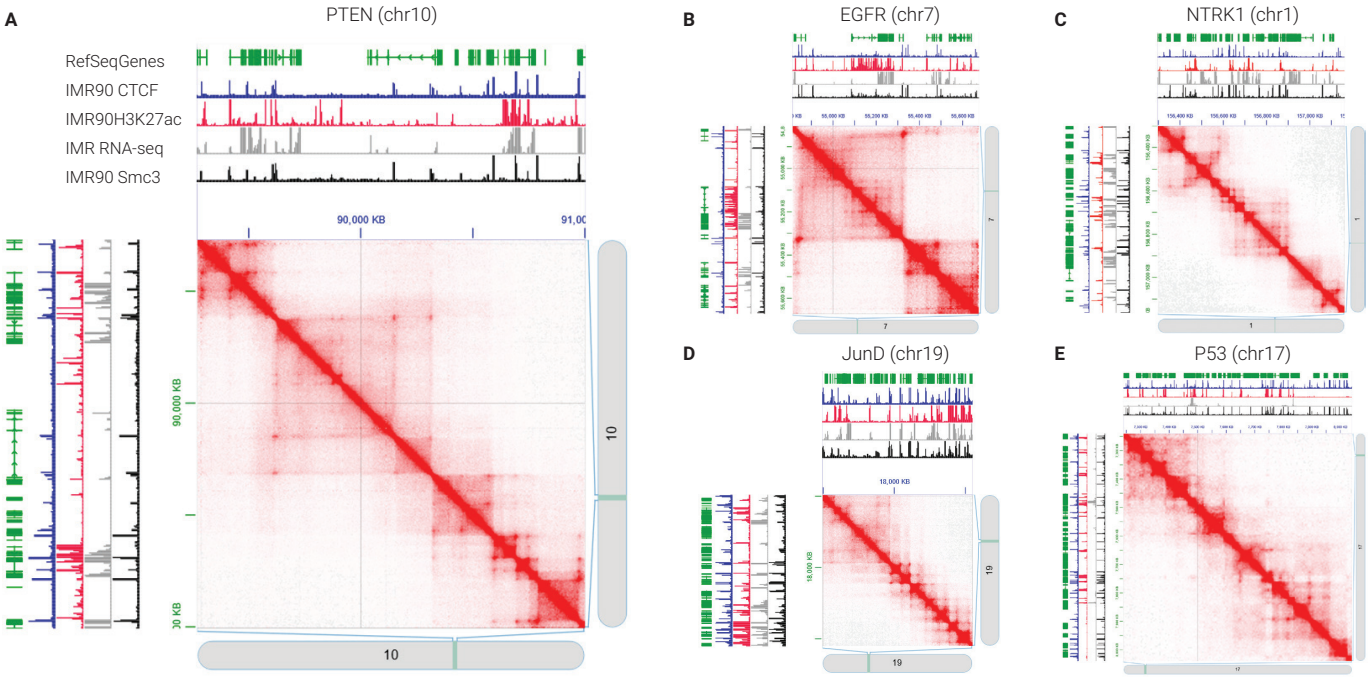
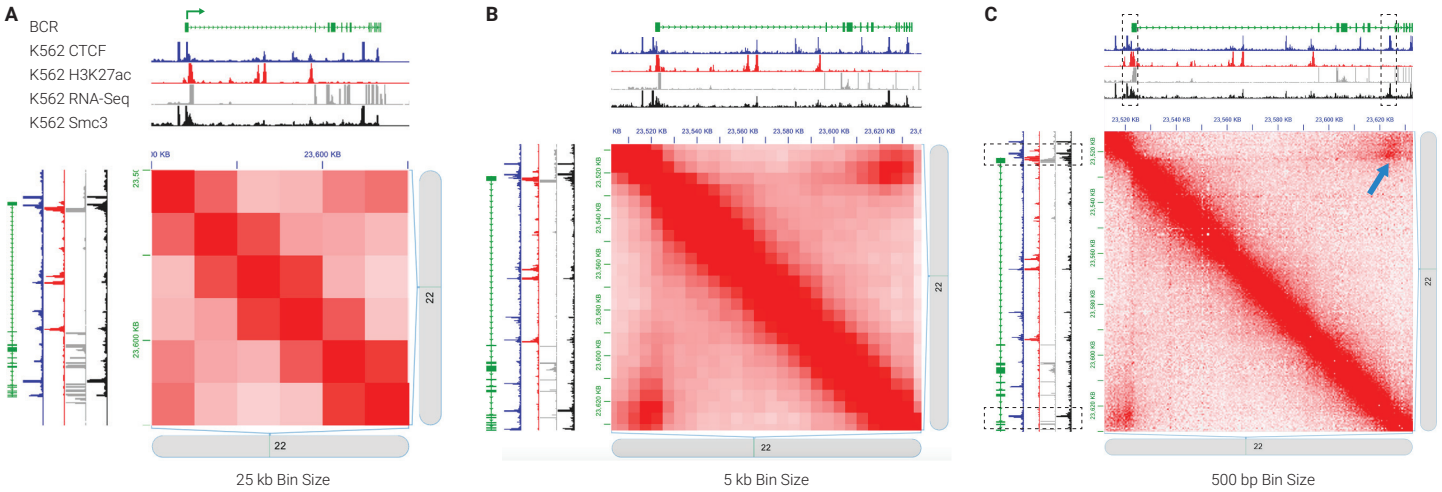


Figure 2. Genome-wide map of capture HiC data in IMR90 cells. When the contact matrix is combined with publicly available ENCODE data (RNA-seq data, H3K27ac mark, CTCF, and Smc3/Rad21 ChIP-seq signal tracks), the connection between the observed hierarchical architecture and gene regulation becomes apparent. Each figure represents a zoomed-in view of capture HiC data for the relevant gene and region. (A) A 1.73 Mb region of chr10 comprising *PTEN*. (B) An 870 kb region on chr7 comprising *EGFR*. (C) An 860 Kb region on chr1 comprising *NTRK1*. (D) A 1.06 Mb region on chr19 comprising *JUND*. (E) An 800 kb region on chr17 comprising *P53*. Data was visualized using Juicebox[4].

Cost-Effective Sub-Kilobase Resolution with Capture HiC

The true power of capture HiC lies in the increase in resolution possible by focusing sequencing reads to regions of interest to concentrate sequencing depth. Figure 3 shows capture HiC data of the *BCR* gene in the K562 cell line binned at progressively higher resolutions of 25 kb (A), 5 kb (B), and 500 bp (C). As the resolution increases, the long-range chromatin contacts become more refined. This increased resolution enables more precise identification of topological landmarks and enables finer-scale observations to be studied.

This is illustrated by the chromatin loop that is anchored at the promoter and the intronic region preceding exon 9 (indicated by a blue arrow in Figure 3C). Specifically, individual CTCF / cohesin binding sites (dashed black box) can clearly be found to colocalize to the anchor points, arguing for a causative role in the organization. This observation is not possible with lower resolution contact matrices. The extrapolated approximate sequencing depth needed for this quality of high-resolution map of just the *BCR* locus is about 8.3 million read pairs, illustrating how high-quality sequencing data is easily attainable using capture HiC.



D Sequencing Depth (Raw Reads) vs Resolution for BCR Locus

	25 kb Resolution	5 kb Resolution	500 bp Resolution
Capture HIC	165,000	830,000	8,300,000

Figure 3. Increased resolution afforded by capture HiC at a variety of resolutions. Capture HiC data shown at the *BCR* locus in K562 cells at progressively higher resolutions with bin sizes of (A) 25 kb, (B) 5 kb, and (C) 500 bp. The blue arrow in Panel C shows a distal interaction that can clearly be identified with the high-resolution maps afforded by capture HiC. (D) Approximate number of raw read-pairs necessary to obtain either 25 kb, 5 kb, or 500 bp resolution. These calculations assume that only the 620 kb BCR locus is targeted in the capture HiC experiment. All data was processed using Juicer and visualized using Juicebox alongside ENCODE RNA-seq and ChIP-seq data

Improved Resolution Aids the Analyses of Gene Regulation at the PDGFRA Locus

With the increased resolution of capture HiC contact maps, comparisons between different genotypes or experimental conditions become more informative. Within our data set, several cell lines showed no transcription of the *PDGFRA* locus, and correspondingly lacked significant H3K27ac signal as expected (Figure 4, data not shown). However, in the IMR90 cell line, robust transcription of *PDGFRA* occurred and was accompanied by extensive H3K27ac signal localized to the promoter region. The contact matrices binned to 1 kb showed clear looping at a distal enhancer about 300 kb upstream and is anchored by CTCF / Smc3 binding at the promoter and distal enhancer. The CTCF / Smc3 site was individually composed of three different sites, with the promoter site found in the middle. Upon closer inspection, a zoomed-in view revealed that, in the transcriptionally active IMR90 cell line, the promoter CTCF / Smc3 site formed a promoter-enhancer loop (Figure 4C). In contrast, the promoter-enhancer loop was missing in the HeLa cell line (Figure 4D). While it formed long-range interactions with the 2 other CTCF / Smc3 sites, the promoter-enhancer loop did not involve the promoter of *PDGFRA*. Without the improved resolution afforded by capture HiC, this mechanistic detail would be impossible to discern.

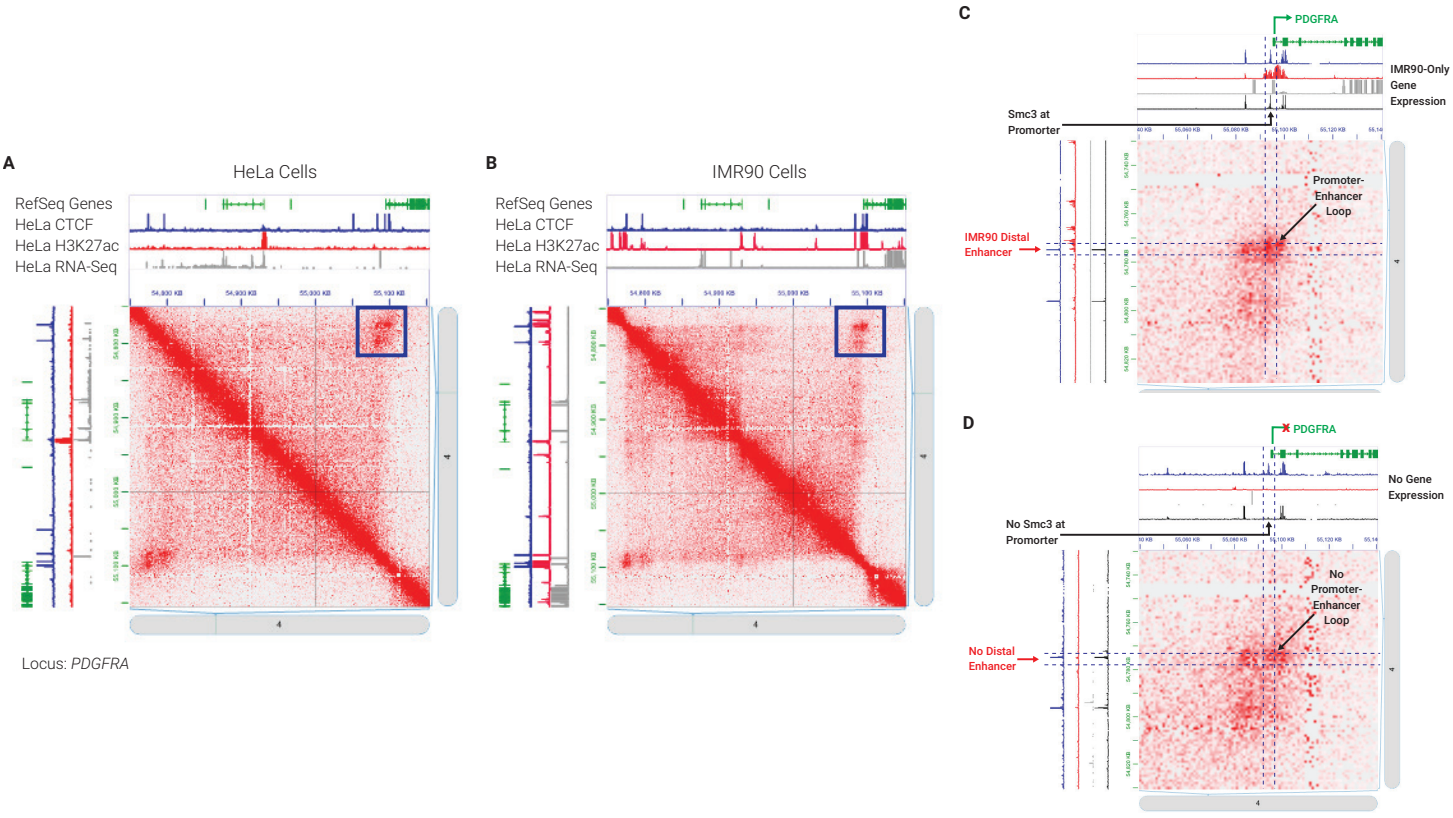


Figure 4. The increased resolution of capture HiC contact maps enables comparisons between different genotypes or experimental conditions and yields valuable epigenetic and mechanistic details. Capture HiC data shown at the *PDGFRA* locus in (A) HeLa and (B) IMR90 cells binned at 1 kb. (C) Zoomed-in view of a chromatin loop in IMR90 cells anchored on the active *PDGFRA* promoter. The loop involves a distal enhancer about 300 kb upstream and is anchored by CTCF/Smc3 binding at the promoter and distal enhancer. (D) Zoomed-in view depicting the absence of a chromatin loop at the inactive *PDGFRA* promoter in HeLa cells. The *PDGFRA* promoter lacks Smc3 binding and the distal region lacks H3K27ac signal.

Conclusion

Despite the unrivaled power of HiC to discern three-dimensional chromosomal structure, adoption of this technology has been slowed in part by the lack of user-friendly workflows and cost-prohibitive sequencing requirements. To solve these problems, Agilent and Arima Genomics have partnered to provide a user-friendly capture HiC workflow that reduces the amount of sequencing needed to produce high-quality, high-resolution targeted HiC maps. The new workflow incorporates an out-of-the-box HiC solution from Arima Genomics with a streamlined target enrichment workflow powered by the Agilent SureSelect XT HS2 kit. Deep sequencing of resulting libraries revealed significant mechanistic detail, illustrating the improved resolution afforded by the combined capture HiC workflow. In summary, thanks to the workflow improvements from Arima and the rapid 90-minute hybridization protocol from Agilent, users can perform what has been an arduous process in as little as three days.

References

1. Dekker, J., Belmont, A., Guttman, M. et al. The 4D nucleome project. *Nature*. **2017**. *549*, 219-226.
2. Schmitt, A., Hu, M., Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nat Rev Mol Cell Biol*. **2016**. *12*, 743-755.
3. Durand, N., Shamim, M., Machol, I., et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems*. **2016**. *3*, 95-98.
4. Durand, N., Robinson, J., Shamim, M., et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems*. **2016**. *3*, 99-101.

www.agilent.com

PR7000-2529

Printed in the USA, July 15, 2020
5994-2218EN

Agilent
Trusted Answers