The Arima-HiC Kit and service for Chromosome -spanning Assemblies

PRODUCT DATASHEET

ARIMA GENOMICS

PRODUCT DATASHEET
SEPTEMBER 2018

The Arima-HiC Kit and Service for Chromosome-spanning Assemblies

Introduction

The availability of high-quality references has a profound impact on the understanding of genome function and species evolution. Recent years have seen a rapid expansion of long-read and long-range methods, including Hi-C[1,2], a NGS-based assay that preserves chromosome-range information during the sample prep. The Hi-C sequencing data uses the preserved chromosome-range information to transform contigs to chromosomes. Despite its utility, broad adoption of Hi-C has been plagued by labor-intensive complex protocols, prolonged workflow durations, inconsistent experimental results, excessive sequencing requirements, and expensive Bioinformatics analyses/services.

The Arima-HiC Kit overcomes these technical and economical limitations with the development of a highly simplified and robust protocol that streamlines Hi-C to a 6-hour, 8-step procedure, followed by library prep and NGS (Fig.1). The Arima-HiC sequencing data is then used to scaffold contigs to generate chromosome-spanning assemblies via SALSA[3,4] – collaboratively developed open-source and free software3. These key advancements have persuaded researchers, including the Genome10K consortium5, to use Arima-HiC to generate high-quality assemblies of hundreds of species across plant and animal kingdoms.

"The Arima-HiC kit is a rapid 6-hour protocol that is easy to use, allowing us to generate libraries in challenging samples,"

– David Jiang, Senior Vice President, Global Head of Corporate Development at Novogene Corporation[5,6]

Fast, Easy Workflow

The Arima-HiC workflow was optimized to enable first time Hi-C users to generate high-quality data with ease (Fig.1). The rapid 6-hour protocol limits prolonged exposure of chromatin to external agents, leading to significant enrichment of inter-con-

Highlights

Fast and Easy Workflow

- Simplify processes with flexible input requirements and support for a broad range of species and sample types
- Rapid time to Hi-C libraries with minimal hands-on touch points, 6-hour prep time (1 hour hands-on time)
- Open-source Bioinformatics tool (SALSA) for chromosome-spanning assemblies

Proven Performance

- Obtain high inter-contig Hi-C signal with consistency, regardless of user experience level, via easy workflow
- Generate uniform "spread" of inter-contig Hi-C signal to scaffold contigs regardless of contig gap-sizes via use of RE-cocktail for chromatin digestion
- Sequencing and analysis cost saving via enriched inter-contig signal and open source tools with Arima-HiC

Demonstrated Utility

- Arima-HiC has been selected by the Genome10K consortium to generate high-quality assemblies of 260 species.

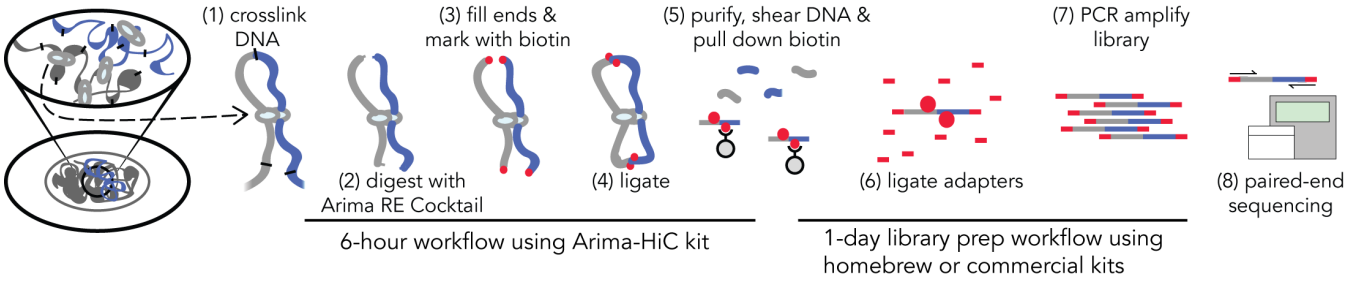

Figure 1: The Arima-HiC workflow results in ligated and biotinylated DNA that is PCR-amplified and prepared as a library using a multitude of library prep kits with appropriate adapters for paired-end Illumina sequencing.

tig signal. The use of a unique combination of multiple 4-base cutting restriction enzymes (RE) for chromatin digestion results in greater spread (per-base uniformity) of inter-contig signal to assemble all contigs regardless of gap-sizes. Overall, the Arima-HiC Kit was designed to maximize the ease-of-use, with minimal total time and hands-on steps, compatibility with any downstream library prep kits for Illumina NGS, 96-well plate compatible design, and support for a broad range of sample types and species (Table 1). Upon sequencing of the Arima-HiC libraries, easy-to-use, well-documented and free Bioinformatics tools such as SALSA3,4 can be used to map Arima-HiC data to contigs to scaffold them into accurate chromosome-spanning assemblies. In addition to scaffolding, Arima-HiC data has also been used to polish individual bases of the contigs as Arima-HiC is sequenced on an Illumina NGS3 (Table 1).

Proven Performance

Rigorous external and internal testing resulted in an Arima-HiC Kit with robust performance. When key opinion leaders compared Arima-HiC data with Hi-C data generated by competitors, the Arima-HiC data manifested higher inter-contig signal (Fig.2A). Importantly, the competitor's Hi-C data manifested a rapid drop in inter-contig signal with increase in contig gap-sizes – Arima-HiC data, on the other hand, manifested 2-3 fold higher inter-contig signal regardless of the contig gap-size (Fig.2B). Subsequent SALSA3,4 analyses of Arima-HiC data converted 4Mb Oxford Nanopore contigs to 125Mb NG50 chromosome-spanning human genome assemblies, with as less as 30X Arima-HiC sequencing depth3 (Fig.3). This is an extraordinary performance given that the even the GRCH38 reference assembly generated by cost-prohibitive combination of multiple technologies had an NG50 of only ~140Mb, suggesting that the high quality of Arima-HiC libraries and sequencing data offers substantial technical and economic benefits to the user. These benefits can be leveraged via Arima-HiC kits or services (sample to assembly).

Demonstrated Utility

The ease-of-use, consistency, and proven performances of Arima-HiC workflow, and, the open-source Bioinformatics strategy have made the Arima-HiC kits and services as a highly popular choice for generating chromosome-spanning assemblies. Indeed, Arima-HiC has been selected by the Genome10K consortium to generate high-quality assemblies and base-polishing of >260 vertebrate species5. At the moment, Arima-HiC is being used to assemble >75 species (15 from Genome10K and >60 species from independent customers).

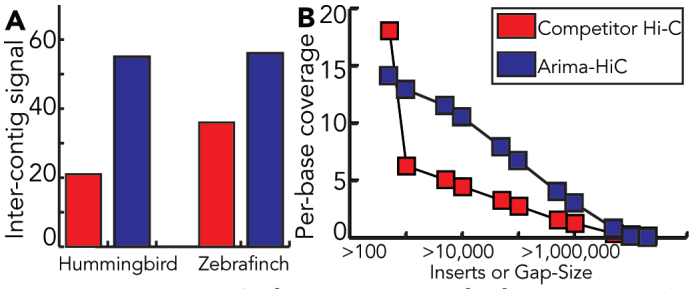

Figure 2: Arima-HiC data generates higher inter-contig signal strength and spread, critical features that enable chromosome-spanning assemblies at reduced sequencing. (A) Arima-HiC data and Competitor's Hi-C (data generated by Dovetail Genomics) is mapped to Hummingbird and Zebrafinch contigs generated by Pacific Biosciences Sequencers. Regardless of the species, Arima-HiC consistently generates higher inter-contig signal. (B) Hummingbird Hi-C datasets from Arima Genomics and Competition analyzed in the context of insert-sizes. That is, when Hi-C reads are categorized by insert-sizes, Arima-HiC maintains high signal (2-3 fold) & coverage regardless of insert-size. In the context of assembly, this "spread" of signal can enable contigs of all gap-sizes to be well-assembled, to generate accurate assemblies at reduced sequencing cost. To enable an unbiased analyses, these were performed with 2M Hi-C reads. Analyses performed by Arang Rhie (NIH). Shared with permission.

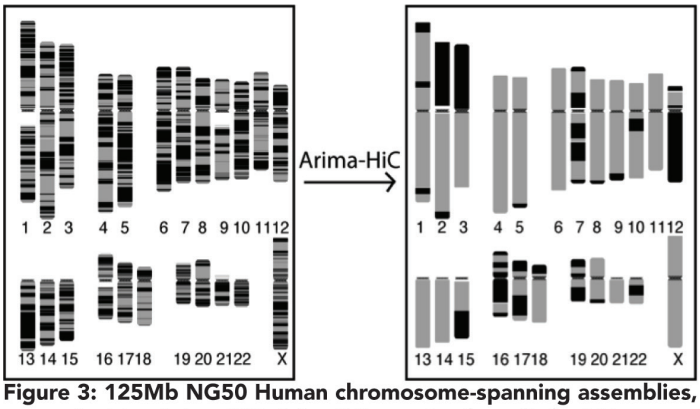

Figure 3: 125Mb NG50 Human chromosome-spanning assemblies, generated by Arima-HiC data 4Mb contigs from Oxford Nanopore converted to 125Mb assemblies via Arima-HiC data, using SALSA open-source tool. Change in color in the ideogram represents a contig gap or error. Post Arima-HiC assemblies have minimal color change, suggesting long (NG50 125Mb and chromosome-spanning) & accurate assemblies. Figure from Ref (3), shared with permission.

"We were consistently impressed by the quality and usefulness of the data Arima generated."

– *Olivier Fedrigo, Director of Vertebrate Genomics Laboratory at the Rockefeller University, Participant Lab, Genome10K Consortium*[5,6]

Table 1: Arima-HiC Specifications

Total Time	6 hours
Hands-on Time	1 hour
Number of Steps	8
Automation Capability	Single–pot, 96-well plate compatible
Restriction Enzymes (RE)	RE cutting at GATC and GANTC
Per-base Genome Uniformity "Spread" (fraction of genome with avg. seq. depth)	~90%, as estimated from Human genome. Similar performance expected in all species.
Sample Types	Seeds, Tissue, blood, cell-lines, whole insects
Sample Storage Conditions	Fresh/frozen, Cross-linked, RNALater, Ethanol
Input Quantity	Amount of sample manifesting 1ug DNA
Species	Plants, Invertebrates, Vertebrates
Library Prep Compatibility	KAPA HyperPrep, Illumina TruSeq, and Swift Accel NGS 2S
NGS Compatibility	Illumina NGS
Library Complexity	1 reaction, 200G
Sequencing Depth (X), recommended	Arima-HiC: 15-60X, depending on the contig NG50
Data Analysis	For kit users, use SALSA tool (Ref 3).
Example Proven Results	4Mb Contig to 120Mb Assembly in Humans3. Currently assembling >75 species.

Additional Details

Please refer to the Genome Conformation Application Note, available by contacting info@arimagenomics.com

Learn more online at www.arimagenomics.com

References

1. Lieberman-Aiden E, et al "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome" Science 326, 289-293 (2009)
2. Rao SP, et al "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping" Cell 159, 1665-1680 (2014)
3. Ghurye J, et al "Integrating Hi-C links with assembly graphs for chromosome-scale assembly" BioRxiv, doi: http://dx.doi.org/10.1101/261149 (2018)
4. SALSA Bioinformatics Tool for Arima-HiC. https://github.com/machinegun/SALSA
5. Genome10K Press Release (1). https://www.businesswire.com/news/home/20180117006561/en
6. Genome10K Press Release (2). https://www.genomeweb.com/genetic-research/vertebrate-genomes-project-plans-combine-technologies-near-gapless-assemblies

Learn more online at

arimagenomics.com

ARIMA GENOMICS