User Guide

Arima-HiC+ Kit

Arima-HiChIP Bioinformatics User Guide

Material Part Number: A101020
Document Part Number: A160173 v01
Release Date: December 2020

This product is intended for research use only. This product is not intended for diagnostic purposes.

This document and its contents are proprietary to Arima Genomics, Inc ("Arima Genomics"). Use of this document is intended solely for Arima Genomics customers for use with the Arima-HiC+ Kit, P/N A101020, and for no other purpose. This document and its contents shall not be used, distributed or reproduced in whole or in part and/or otherwise communicated or disclosed without the prior written consent of Arima Genomics.

This user manual must be read in advance of using the product and strictly followed by qualified and properly trained personnel to ensure proper use of the Arima-HiC+ kit. Failure to do so may result in damage to the product, injury to persons, and/or damage to other property. Arima Genomics does not assume any liability resulting from improper use of its products or others referenced herein.

U.S. Patent No. US 9,434,985 and 9,708,648 pertains to the use of this product.

TRADEMARKS
Illumina®, MiSeq®, MiniSeq®, NextSeq®, HiSeq®, and NovaSeq™ are trademarks of Illumina, Inc.

Revision History

Document	Date	Description of Change
Material Part Number: A101020 **Document Part Number:** A160173 v00	March 2020	Initial Release
Material Part Number: A101020 **Document Part Number:** A160173 v01	December 2020	Added support for calling ChIP peaks from HiChIP data using MACS2. Added command line arguments to the MAPS pipeline

Table of Contents

Introduction

1.1 Arima-HiChIP Workflow Overview

Arima-HiChIP is an experimental workflow that captures the structure (three-dimensional conformation) of genomes associated with a protein of interest. HiChIP is a Chromatin Immunoprecipitation of Hi-C interactions that are bound by a protein of interest. Hi-C captures genomic interactions by crosslinking intact nuclei with formaldehyde and then digesting the chromatin with a restriction enzyme. These crosslinked, genomic fragments are labeled with biotinylated dNTP's and then ligated together, preserving the chromatin interactions captured by the crosslinking and enabling them to be enriched by Streptavidin conjugation for library prep and sequencing. Chromatin interactions are immunoprecipitated on to protein A beads, similar to how ChIP-seq is performed. Refer to the HiChIP User Guide for Mammalian Cells for more information regarding the assay portion of the Arima-HiChIP workflow.

To analyze Arima-HiChIP data, we strongly recommend using the publicly available tool MAPS (MAPS; Juric, 2019). ChIP-Seq peaks, used by MAPS for loop calling can be generated using the built in MACS2 functionality or generated from the same cells in a ChIP seq experiment using the same antibody. Benchmarking with MAPS, FitHiChIP, and HiCCUPS against CRISPRi-validated gene regulatory interactions revealed that MAPS has the highest sensitivity (true positive rate) while minimizing computational time. Arima-HiChIP data analyzed with MAPS will yield a set of chromatin loops between two genomic loci bound to the immunoprecipitated protein and which interact with each other more than would be expected by a statistical model. These loops are used for identifying protein associated chromatin interactions and comparing such interactions between experimental conditions.

1.2 MAPS overview

MAPS (Model-based analysis of PLAC-seq and HiChIP) detects chromatin loops based on the number of Hi-C interaction counts between chromatin sites occupied by a specific chromatin protein (e.g. histone, transcription factor). Poisson regression modeling is used to correct the data for biases attributed to GC content, restriction fragment length, sequence mappability, and ChIP enrichment. MAPS only considers Hi-C interactions that have a genomic distance of 1kb – 2Mb and has at least one of the two read-pairs overlapping a bin which contains a ChIP peak. Read-pairs that have only one read end overlapping a known ChIP peak are called "XOR" reads and those that have both read-ends overlapping ChIP peaks are called "AND" reads. AND and XOR reads are both used for loop calling and are modeled separately from one another. This approach is meant to maximize loop discovery while minimizing false positive calls. To call chromatin loops, MAPS requires that there be at least 12 Hi-C counts supporting each loop call, the counts are at least 2-fold enriched compared to the Poisson regression model, and have a False Discovery Rate (FDR) of less than 1%.

1.3 ChIP-seq peak dataset

MAPS requires a set of ChIP-seq peaks to call chromatin loops. The Arima-MAPS pipeline V2.0 can seamlessly call ChIP peaks from the HiChIP data by running MACS2 using the short Valid Interaction Pairs (VIPs) output from the feather alignment. Depending on the type of chromatin factor, narrow or broad, specific MACS2 parameters have been chosen that maximize recall of ground truth ChIP peaks while minimizing the false positive peak calls. Ground truth peaks are defined as ENCODE peaks which are called in two or more experiments. Optionally, ChIP-seq peaks can be generated from the same crosslinked cell batch using the same antibody lot as the HiChIP data. ChIP peaks for various chromatin proteins for various cell lines can be sourced from publicly available sources, such as the ENCODE consortium. If selecting ChIP peaks from the ENCODE consortium, we recommend using ChIP peaks that are reproduced in two biological replicates ("replicated peaks"). Reproducible peaks are likely to have low false positive peak calls and will result in more accurate loop calls.

1.4 ChIP-seq peaks called from HiChIP data

ChIP peaks can now be called from the HiChIP data, using paired-end reads which the two reads in the pair map to within 1kb of each other, using MACS2. ChIP peaks can be called for broad and narrow peak chromatin factors, based on user input. The Arima-MAPS v2.0 pipeline calls "broad" peaks using MACS2 with a "broad-cutoff" of 0.2 when calling peaks for both broad and narrow chromatin factors. Additionally, broad peaks are called with the "--nolambda" flag. Using the "--broad" peak flag in MACS2 for both narrow and broad chromatin factors, along with the other setting mentioned, optimizes the sensitivity and positive predictive value of peak calls for HiChIP data compared to ground truth peak calls.

2.1 Cloning MAPS From GitHub and Installing Dependencies

To clone the latest version of MAPS to the directory of your choice use the command "git clone https://github.com/ijuric/MAPS.git". This will download MAPS and all associated files. The Arima-MAPS_v2.0.sh wrapper script will be downloaded along with MAPS and placed in the "MAPS/Arima_Genomics/" directory. This is the master script that aligns the reads with Feather, calls ChIP peaks with MACS2, performs loop calling with MAPS, and generates Arima QC metrics. It also generates metaplots for visually assessing ChIP enrichment, and arc plots of the statistically significant chromatin loops for data visualization in the WashU Epigenome Browser. Additionally, the Arima-HiChIP specific genomic features files will also be downloaded and will be located in the directory "MAPS/Arima_Genomics/genomic_features/". The genomic features are specific to the two-restriction enzyme Arima-HiC chemistry and are used for calibrating the MAPS loop calling algorithm for this chemistry.

Detailed instructions for downloading and installing Arima-MAPS dependencies are on the Arima-MAPS GitHub page (Arima-MAPS). These dependencies include:

- Python 3.4 (or later)
  - Anaconda3
  - deeptools (v3.3.0)
  - pandas (v0.20.3)
  - numpy (v1.13.1)
  - pysam (v0.15.2)
  - pybedtools (v0.8.0)
  - itertools (v3.2)
  - MACS2 (v2.2.7)
- R (v3.4.3)
  - argparse (v2.0.1)
  - VGAM (v1.1-2)
  - data.table (v1.12.8)
- bedtools (v2.27.1)
- Htslib (v1.10.2)
- samtools (v1.10)
- bcftools (v1.10.2).
- bwa (v0.7.17)

2.2 Overview of Command line Arguments

Table 1. Overview of Required Command Line Arguments.

Argument	Required	Default Value	Description
-I	Yes	None	Absolute path and file prefix of the fastq files, up to "_R1" or ".R1". Example: "/path/to/sample1" for the fastq file "/path/to/sample1_R1.fastq.gz"
-O	Yes	None	Output Directory. directory which the feather and MAPS outputs will be placed.
-o	Yes	None	Organism of the genomic feature file to be used, options: "mm9", "mm10", "hg19", and "hg38". If additional genomic feature files are needed then please contact technical support at: techsupport@arimagenomics.com.
-b	Yes	None	BWA index. Absolute path to the reference genome sequence (.fa). BWA index is expected to be in the same directory as the reference genome. Ex: "/home/reference_sequence/hg19.fa"
-t	Yes	None	Number of threads. Use 4-8 for shallow sequencing and 12-20 for deep sequencing.
-C	Yes	0	Call ChIP peaks from HiChIP data. 0 to use ChIP peak file provided by "-m" option, 1 to call peaks using MACS2
-p	Yes, if "-C 1"	None	Broadness of chromatin factor. Options are: "broad" (histone) or "narrow" (TF). Required if calling peaks from the HiChIP data ("-C 1"). Must Both choices result in broad peaks being called by MACS2.
-m	Yes if "-C 0".	None	MACS2 file path. Path to the ChIP peak file that will be used for loop calling with MAPS. Required if "-C 0".

Table 2. Overview of Optional Command Line Arguments.

Argument	Required	Default Value	Description
-F	No	1	Run feather. 1 to run feather, 0 to skip.
-M	No	1	Run MAPS. 1 to run MAPS on data processed with feather, 0 to skip
-P	No	1	Generate QC table, heatmaps, and arc plots. 1 to generate plots, 0 to skip.
-f	No	None	Patterned flowcell. Use 1 for data from patterned flowcells and 0 for non-patterned flowcells. This is used for calculating optical duplicates and PCR duplicate rates.
-H	No	1	Generate .hic. 1 to generate a .hic file for visualization with Juicer, 0 to skip
-s	No	5000	Bin Size. Resolution of the loops called in bp.
-r	No	2000000	Binning Range. Maximum distance for loop calling in bp.
-d	No	2	False Discovery Rate threshold: 1 = 0.1, 2 = 0.01, 3 = 0.001, ect... Recommended not to change.
-Q	No	30	MAPQ threshold. Phred scaled mapping quality threshold.
-l	No	1000	Minimum genomic distance for loop calling.
-h	No	None	Print help and exit.

2.3 Arima-HiChIP Data QC and Sequencing Recommendations

Prior to deep sequencing Arima-HiChIP libraries (defined here as >50M reads), we recommend performing a shallow sequencing run using a low throughput platform such as on the Illumina® MiniSeq® or MiSeq® to obtain approximately 0.5 – 2M read-pairs for QC assessment of the libraries and Arima-HiChIP data. The shallow sequencing QC metrics output by the Arima-MAPS pipeline are important for assessing library quality for the degree of ChIP enrichment and capture of long-range chromatin interactions, which in turn is used to estimate the required deep sequencing depth needed for robust and reproducible chromatin loop discovery. The Arima-MAPS pipeline is executed the same for either shallow or deep sequencing data, with the exception of the computational resources required. See Section 2.4 below for further discussion on computational resources.

2.4 Compute Resources

For shallow sequencing (0.5 - 2 million raw read-pairs), the Arima-MAPS pipeline requires 12 CPU cores with 48 GB RAM. The shallow sequencing analysis should complete in less than 2 hours, depending on hardware. For deep sequencing (50 – 500 million raw read-pairs), we recommend 16 - 20 CPU cores with at least 64 - 80 GB RAM. Samples with 200 million raw read-pairs will run through the Arima-MAPS pipeline in about 48 hours with the recommended computational resources. Additional resources can be added to decrease the analysis time.

2.5 How to Cite MAPS in Publications

When citing the MAPS pipeline please use: Juric I, Yu M, Abnousi A, Raviram R, Fang R, Zhao Y, et al. (2019) MAPS: Model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. PLoS Comp Biol 15(4): e1006982. https://doi.org/10.1371/journal.pcbi.1006982

Running the Arima-MAPS v2.0 Pipeline

3.1 Overview

This section describes how to run the Arima-MAPS pipeline on a single sample. Before beginning this section, download and install the Arima-MAPS pipeline detailed in Section 2.1. The Arima-MAPS pipeline uses a wrapper script (**Arima-MAPS_v2.0.sh**) to run MAPS, MACS2, and generate additional HiChIP QC metrics and visualization files. The wrapper script runs Feather and MAPS using either the user-specified ChIP peak file or ChIP peaks called from the HiChIP data. The wrapper also generates QC metrics for shallow and deep sequencing data as well as metaplots and arc plots for the statistically significant chromatin loops.

This section walks through how to setup and run MAPS on the test dataset which can be downloaded from our ftp site at: ftp://ftp-arimagenomics.sdsc.edu/pub/MAPS/test_data/. The same process would be used for any other data set. Note: Text surrounded by square brackets, e.g. "[PATH_TO_DATA]" indicates user-supplied information. If copying and pasting from examples in this document, replace the text *and the brackets* with the values that are appropriate for your analysis. For example: /[PATH_TO_DATA]/test/fastq/ would be /path/to/data/test/fastq/ where "/path/to/data/" is the user supplied file path.

Inputs:

a) Path to the folder containing the raw sequencing data in ".fastq" or ".fastq.gz" format..
b) Path to the folder which will contain the outputs.
c) The type of ChIP peaks to call with MACS2 or, optionally, ChIP peaks to be used for loop calling in BED format.
d) Reference genome sequence in FASTA format along with the BWA index files for the reference sequence.
e) The number of CPUs to use for the analysis.

Outputs:

a) Arima-HiChIP shallow and deep sequencing QC tables for use with the Arima-HiChIP_QC_Worksheet.xls.
b) Feather output files.
c) MAPS output files (e.g. chromatin loop calls).
d) ChIP Enrichment metaplots and heatmaps.
e) 1-D HiChIP coverage track in bigWig format.
f) Chromatin loop arc plot in tabix indexed BEDPE format.
g) Optional MACS2 Peaks.

3.2 Arima-MAPS Analysis Workflow

1. Create a directory for the dataset you would like to run MAPS on as well as a sub-directory to contain the datasets.
    a. Example: "mkdir -p /[FASTQ_DIR]/"
2. Place the raw paired-end sequencing data (in .fastq or .fastq.gz format) for this sample into the fastq/ directory.
    a. Example: "/[FASTQ_DIR]/Arima-MAPS-test_R*.fastq.gz"
3. Run the Arima-MAPS pipeline by executing the Arima-MAPS_v2.0.sh script from the command line of a Linux machine with the appropriate computing resources and arguments (discussed in section 2.4). See below for example usage on Arima test data.

```
# Run Arima-MAPS v2.0 with Known ChIP Peaks
sh [PATH_TO_MAPS_INSTALLATION]/bin/Arima-MAPS_v2.0.sh \
-I [FASTQ_DIR]/Arima-MAPS-test \
-O [PATH_TO_OUTPUT_DIR] \
-C 0 \
-m [PEAK_FILE_DIR]/ENCFF247YHM.UW.bed \
-o hg19 \
-b [GENOME DIR]/hg19.fa \
-t 8 \
&> Arima_MAPS_v2.0_Log.txt
```

Figure 1. Example Usage for running Arima-MAPS v2.0 with Known ChIP peaks.

```
# Run Arima-MAPS v2.0 and call ChIP Peaks with MACS2
sh [PATH_TO_MAPS_INSTALLATION]/bin/Arima-MAPS_v2.0.sh \
-I [FASTQ_DIR]/Arima-MAPS-test \
-O [PATH_TO_OUTPUT_DIR] \
-C 1 \
-p broad \
-o hg19 \
-b [GENOME_DIR]/hg19.fa \
-t 8 \
&> Arima_MAPS_v2.0_Log.txt
```

Figure 2. Example Usage for running Arima-MAPS v2.0 and generating ChIP peaks from the HiChIP data.

When running the Arima-MAPS v2.0 script, it is required to specify the location and file prefix of the fastq data used for the analysis using the "-I" flag, the location to put the output files with the "-O" flag, the genome name with the "-o" flag, the absolute path to the reference genome and accompanying BWA index with the "-b" flag, and the number of threads for the script to use for the analysis with the "-t" flag (**Fig. 1 and 2**). Along with these inputs, the Arima-MAPS v2.0 pipeline must be run with either a known ChIP peak file or ChIP peaks must be generated from the HiChIP data, see **Fig. 1 and 2**. For example, if the absolute path to the Arima test data is /home/fastq/Arima-MAPS-test_R1.fastq.gz and /home/fastq/Arima-MAPS-test_R2.fastq.gz then the -I flag would be "-I /home/fastq/Arima-MAPS-test". If it is desired for the output files to go into the directory "/home/Arima-MAPS-output/" then set the "-O" flag as "-O /home/Arima-MAPS-output/". To align the data to the hg19 reference genome, first specify the "-o" flag as: "-o hg19", this will use the genomic feature file for the hg19 reference genome, then specify the "-b" flag with the path to the reference genome. The BWA index for the reference genome is assumed to be in the same directory as the reference genome sequence. For instance, if the reference genome is located in "/home/references/hg19.fa" then specify the "-b" as "-b /home/references/hg19.fa". The BWA index will be automatically pulled from the "/home/references/" directory. It is recommended to write the Arima-MAPS v2.0 outputs to a log file ("&> Arima_MAPS_v2.0_log.txt") so that a record of the pipeline's performance can be kept and used for troubleshooting if necessary.

To run the Arima-MAPS v2.0 pipeline with a known ChIP peak (**Fig. 1**), the "-C" flag is set to "0", telling the pipeline not to call peaks from HiChIP data, and the "-m" flag is set to the location of the known ChIP peak file to be used for loop calling. The error "ChIP peak file (-m) is required for loop calling when call_peaks=0!" will be generated if the "-C" flag is set to "0" but there is no ChIP peak file specified with "-m".

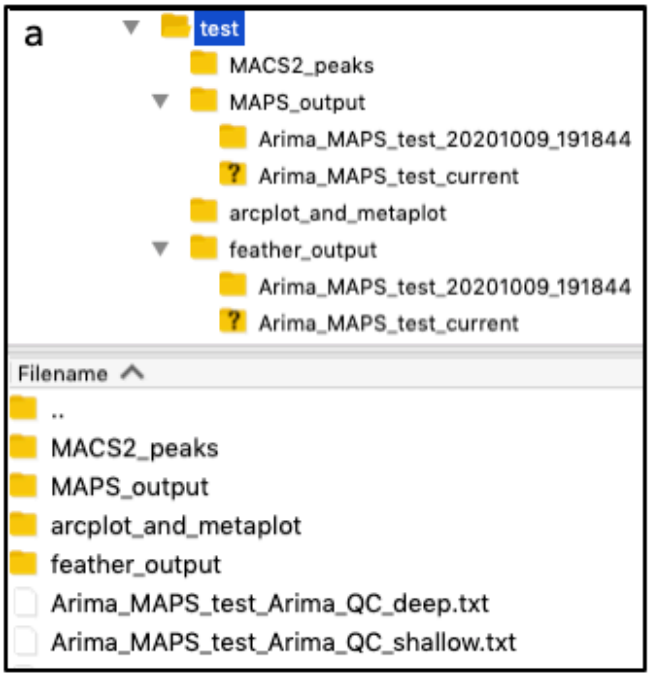
To generate ChIP peaks from the HiChIP data (**Fig. 2**), the "-C" flag should be set to "1" and the broadness of the chromatin factor should be specified with the "-p" flag using either "broad" for broad marks such as histones (H3K27Ac, H3k4me3, ect…) or "narrow" for narrow marks such as transcription factors (CTCF, PolII, ect…). The error message "Broadness of chromatin factor (-p) is required when call_peaks=1!" will be displayed if the "-C" flag is set to "1" but the "-p" flag is not set.

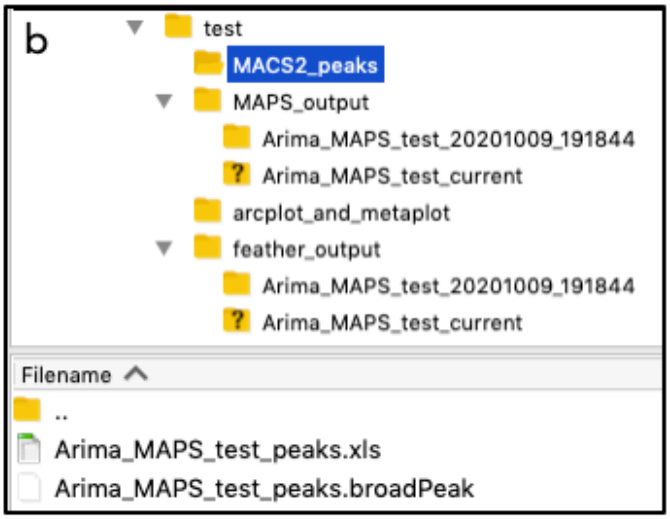
Note: If running Arima-MAPS v 2.0 on a cluster ensure that the correct dependency versions are used by exporting your path variable to your job. Example: 'export PATH=/home/[USER]/bin/:$PATH'. If technical assistance is required please send the log file (Arima_MAPS_v2.0_Log.txt) to **techsupport@arimagenomics.com**

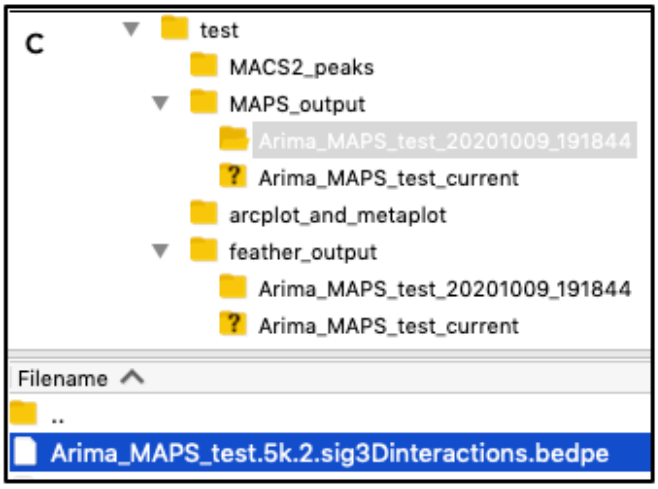
4.1 Introduction

The Arima-MAPS pipeline creates several output files:

1. 2 QC metrics files (one for shallow sequencing and one for deep sequencing).
2. outputs from the Feather pipeline
3. outputs from the MAPS pipeline
4. metaplots for assessing ChIP enrichment relative to known ChIP peaks
5. arc plots for visualizing chromatin loops on the WashU EpiGenome Browser
6. Optional MACS2 ChIP peak calls.







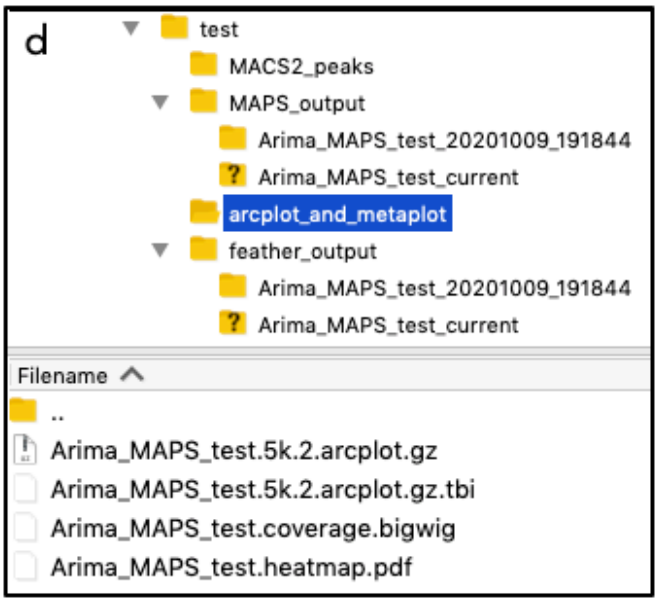


Figure 3. Layout of the output directory structure for the Arima-MAPS pipeline. a) FileZilla screenshot of the top-level directory structure for the test sample "test/". The sub-folders: "MAPS_output", "MACS2_peaks", "arcplot_and_metaplot", and "feather_output" were created automatically by the Arima-MAPS v2.0 pipeline, as well as the Arima Deep Sequencing QC Table and the Arima Shallow Sequencing QC Table. **b)** FileZilla screenshot of the "MACS2_peaks" sub-folder. **c)** FileZilla screenshot of the "MAPS_output" sub-folder. **d)** FileZilla screenshot of the "arcplot_and_metaplot" sub-directory.

4.2 Output Directory Structure

Continuing with the example of sample "Arima-MAPS-test"; after running Arima-MAPS v2.0 in Section 3.2.3, the Pipeline has created four output directories in the top-level directory "test/" and 2 output text files (**Fig. 3a**). The top level "test/" directory was created by the user and was specified using the "-O" option from the command line. The outputs directories are: "MAPS_output/", "arcplot_and_metaplot/", "feather_output/", and optionally "MACS2_peaks" (**Fig. 3a**). Additionally, the Arima Deep Sequencing QC Table (Arima-MAPS-test_Arima_QC_deep.txt) and the Arima Shallow Sequencing QC Table (Arima-MAPS-test_Arima_QC_shallow.txt) are also output (**Fig. 3a**) and are the 2 most critical output files from the Arima-MAPS pipeline, since they contain QC metrics for the Arima-HiChIP_QC_Worksheet.xls. The names of the QC tables are derived from the sample prefix name of the fastq files and are appended with "_Arima_QC_deep.txt" or "_Arima_QC_shallow.txt" depending on the metrics they contain.

The "MACS2_peaks" sub-directory (**Fig. 3b**) contains the ChIP peaks called from the HiChIP data when the user specifies "-C 1". This folder is not created when "-C 0" is specified. This folder contains a .bedfile of the called ChIP peaks "Arima_MAPS_test_peaks.broadPeak" as well as an Excel version of the ChIP peaks "Arima_MAPS_test_peaks.xls".

Within the MAPS_output sub-folder. The Arima-MAPS pipeline automatically created the sub-folder "test/MAPS_output/Arima-MAPS-test_20200315_122347/" based on the dataset name prefix supplied using the "-I" option, the date and the time the analysis was run (**Fig. 3c**). If the Arima-MAPS pipeline is run on the same datasets multiple times, then a new sub-directory will be created for each run with different date and times in the directory name. "Arima-MAPS-test_current" is a symbolic link to the most recently analyzed MAPS_output sub-folder (**Fig. 3c**). There are many intermediate files in the MAPS_output sub-directory, however the most critical is the loop calls file, ending in .sig3Dinteractions.bedpe, and highlighted at the bottom of the panel. The same sub-folder structure is also used for the "feather_output/" sub-directory (**Fig. 3c**). There are many intermediate files in the "feather_output/" directory but all relevant values have been extracted and are in the Arima_QC_deep.txt and Arima_QC_shallow.txt files

The "arcplot_and_metaplot" sub-directory (**Fig. 3d**) has several useful output files for downstream analysis and visualization including arc plots for visualization (.acrplot.gz and .arcplot.gz.tbi), the 1-D ChIP signal from the HiChIP data (.coverage.bigwig), to be used for visualization, and a heatmap (.heatmap.pdf) of the HiChIP enrichment relative to the ChIP peak for the sample.

4.3 Arima Shallow and Deep Sequencing QC metrics files

The Arima-MAPS v2.0 pipeline outputs QC tables for use with the **Arima-HiChIP QC Worksheet** to aid in interpretation of the Arima-HiChIP data quality. These two tables with the suffix "_Arima_QC_shallow.txt" and "_Arima_QC_deep.txt", are created in the output directory specified by "-O" from the command line, each time the pipeline is run. The "_Arima_QC_shallow.txt" file includes metrics related to raw sequencing depth, mappability, duplication rate, the target raw reads for deep sequencing, a summary of the reads used for loop calling, a summary of the Hi-C characteristics of the data, summary statistics of the ChIP peak file used for loop calling, and a summary of the ChIP characteristics of the data (e.g. the percent of short Valid Interaction Pairs (VIPs) that overlap the known ChIP peaks and the enrichment score). The "_Arima_QC_deep.txt" QC file has the same QC metrics as the shallow sequencing file with the exception of target sequencing depth. Instead, the "_Arima_QC_deep.txt" file reports the number of loops discovered from the Arima-HiChIP data. These tables should be copied and pasted into the **Arima-HiChIP QC Worksheet** that accompanies the Arima-HiCHIP User Guide for Mammalian Cells. The **Arima-HiChIP QC Worksheet** "Metric Definitions" tab contains detailed descriptions of the metrics in the two QC tables output by the Arima-MAPS pipeline. For shallow sequencing the key metrics are the percent PCR duplicates, number of Long-Range Fragments (LR FRIPS) per ChIP peak bin per million mapped and deduplicated reads, the target raw read depth, the percent intra-chromosomal interactions that are greater than 15kb, and the percent short VIPs overlapping ChIP peaks. A brief description of the key metrics are below:

- **% PCR Dups** - PCR duplicate read-pairs aligned to the genome, where both read-ends have a mapping quality >= 30. This value assesses the complexity of the library. A lower value indicates that a library contains more unique molecules and can tolerate being sequenced to a higher read depth.
- **LR FRIPS per Peak Bin per Million Deduped PE Reads** - The number of AND and XOR reads from MAPS that have a genomic span of 1kb to 2Mb. This value is then normalized by the number of ChIP peaks and then used for estimating the target raw read depth.
- **Target Raw PE Reads** – Estimate of the number of raw PE reads needed to sequence the library for robust and reproducible chromatin loop calls. This metric is based on the number of "LR FRIPS per Peak Bin per Million Deduped PE Reads", the estimated mapping and duplicate rate for deep sequencing datasets, and the LR FRIPS per bin threshold of 382. This value was determined internally to optimize replicate concordance. This metric can be thought of in terms of: "the number of raw PE reads needed to achieve at least 382 FRIPS per bin given the number of FRIPS per bin obtained for every million reads in the shallow sequencing results and given the expected mapping and duplicate rate". This metric is highly dependent on the quality of the antibody and the number of ChIP peaks. Generally, high specificity antibodies will result in libraries needing 90 – 200M reads, medium specificity antibodies will need 200 – 400M reads, and low specificity antibodies will need greater than 400M reads.
- **% INTRA > 15kb pairs** – Percent of intra-chromosomal interactions that span greater than 15kb in linear genomic distance. This value assesses how well the library captured chromatin

interactions between genomic loci. This value should be over 25% in order for the libraries to pass QC, however 40-50% is typically observed in the highest performing libraries.

- **% Short VIPs in Peaks** – percent of VIP's that overlap the ChIP peaks used for loop calling. This metric along with the metaplots and visual inspection of the *coverage.bigwig tracks can be used to assess the specificity of the ChIP. However, this metric is the most important metric for assessing the quality of a HiChIP library as it is directly related to loop call performance and estimated sequencing depth. Good quality libraries have a specificity of 40-90%, with medium quality libraries around 15 – 40%, and poor quality libraries having less than 15%.
- **Enrichment** – ratio of the average peak signal compared to the average signal 10kb upstream. This value can be visualized in the metaplot in the top of the heatmap.pdf as the peak height relative to the left most signal of the graph.

4.4 Metaplots and Arc Plots

The Arima-MAPS Pipeline outputs heatmaps of alignments around known ChIP peaks to aid in Arima-HiChIP data QC. The heatmap is output in the /arcplot_and_metaplot/ directory. The metaplots and heatmaps are generated from all mapped deduplicated reads and show the distribution of those reads around the ChIP peaks used for loop calling. The more enriched the ChIP peak is relative to the background in the metaplot, the better the enrichment is for that library. The heatmap represents the same data as the metaplot but shows individual ChIP peaks on each row. Examples of metaplots and heatmaps from H3K27ac Arima-HiChIP are shown in **Fig. 4a**. Both the metaplot and the heatmap are generated using the .bigwig file located in /arcplot_and_metaplot. This file can be used to examine the coverage of all mapped deduplicated reads from the HiChIP library in a genome browser, such as the WashU EpiGenome Browser.

The Arima-MAPS Pipeline outputs arc plot files for visualizing chromatin loop calls in a genome browser alongside chromatin, transcriptome, or other data types. In **Fig. 4b,** an example genome browser screen shot is shown, in which arc plots are used to investigate looping interactions for two biological replicates in the context of other epigenetic data on the WashU EpiGenome Browser. To view arc plots, follow the instructions below:

1. If MAPS was run on a server, copy both the gzipped arc plot file: [SAMPLE_PREFIX].5k.2.arcplot.gz and its associated tabix index: [SAMPLE_PREFIX].5k.2.arcplot.tbi (Ex: Arima_MAPS_test.5k.2.arcplot.gz and Arima_MAPS_test.5k.2.arcplot.gz.tbi) to a directory on a local computer as this facilitates uploading the arc plot files to the WashU EpiGenome Browser
2. Navigate to the WashU EpiGenome Browser at http://epigenomegateway.wustl.edu/browser/
3. Select the reference genome used for mapping the HiChIP data and select "Go".
4. From the menu bar across the top of the screen, click on the drop down "Tracks" then click on "View Local Tracks"

5. In the dialogue box under “1. Choose track file type:”, select “longrange”
6. In the same dialogue box under “2. Choose track file:”. Select “Browse…”
7. In the browse dialogue box, navigate to the local directory the arc plot files were copied to in step 1.
8. Select both the gzipped arc plot file: [SAMPLE_PREFIX].5k.2.arcplot.gz and its associated tabix index: [SAMPLE_PREFIX].5k.2.arcplot.tbi, then click “open”
9. The text “*track added*” should appear in red text below the browse button.
10. Close the dialogue window by clicking the red “X” in the upper right-hand corner.
11. From the Browser, right click on the track name of the arc plot track that was just loaded.
12. In the popup menu in the “Display mode” field, change “heatmap” to “ARC”.
13. Also, in the popup menu, change the line width and color scale. For the best visual representation of the arc plots we recommend setting the line width to 1 and changing the score to a fixed scale from 0 to 50. The shade of the arcs’ colors correlate with the statistical significance (False Discovery Rate) of each loop call. The darker the color, the lower the FDR value. This will provide a more dynamic range in chromatin loop coloration, which correlates with the statistical significance of the chromatin loop calls and fixes this scale across all samples for better comparison. Adjust the color scale in the popup menu by changing the “Score Scale” field, from “AUTO” to “FIXED”, and then change the “Score max” from “10” to “50”.

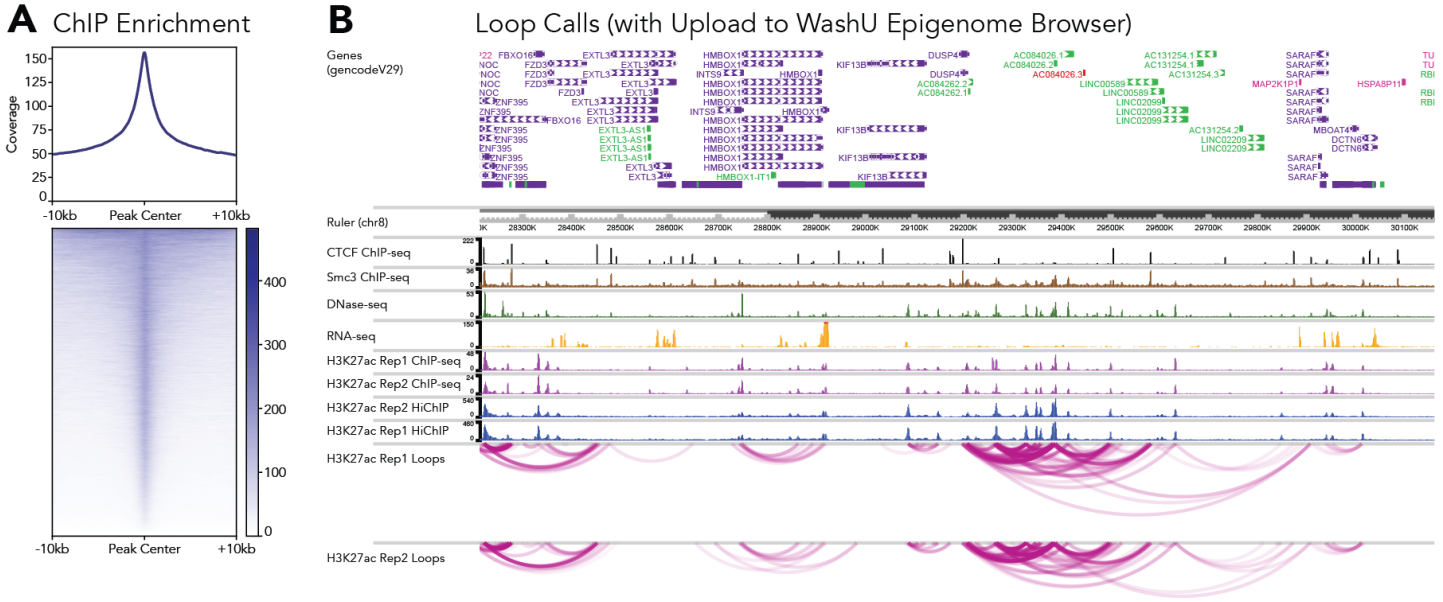


Figure 4. The Arima-HiChiP Experimental and Bioinformatics Workflows. a) Metaplot and heatmap showing the distribution of reads around ChIP peaks. **b)** Arc plots of loop calls from MAPS visualized on the WashU EpiGenome Browser.

4.5 Feather key output files

Feather outputs several useful data files in the /feather_output/[SAMPLE_PREFIX] _current/, 2 of which are listed below. If feather is run multiple times, data analyzed from a particular date can be found in /feather_output/[SAMPLE_PREFIX] _[DATE_TIME]/, with /feather_output/[SAMPLE_PREFIX] _current/ being a symbolic link to the most recent run date.

- [SAMPLE_PREFIX].paired.rmdup.bam – deduplicated and sorted bam file of all reads.
- [SAMPLE_PREFIX].shrt.vip.sort.bed - .bed file of the short VIPs which is useful for intersecting with ChIP peak files to assess enrichment.

4.6 MAPS key output files

Outputs from MAPS can be found here: /MAPS_output/[SAMPLE_PREFIX] _current/. If MAPS is run multiple times, data analyzed from a particular date can be found in /MAPS_output/[SAMPLE_PREFIX] _[DATE_TIME]/, with /MAPS_output/[SAMPLE_PREFIX] _current/ being a link to the most recent run date.

[SAMPLE_PREFIX].5k.2.sig3Dinteractions.bedpe – list of all loops generated by MAPS in a .bedpe format. The key columns in this output are the coordinates of the first loop base: "chr1", "start1", and "end1", the coordinates of the second loop base: "chr2", "start2", and "end2", and the false discovery rate of the looping interaction: "fdr". Consult the MAPS GitHub page (MAPS) for more information on the columns of this file.

Warranty and Contact Info

WARRANTY DISCLAIMERS

THE EXPRESS WARRANTIES AND THE REMEDIES SET FORTH ABOVE ARE IN LIEU OF, AND ARIMA GENOMICS AND ITS LICENSORS, SUPPLIERS AND REPRESENTATIVES HEREBY DISCLAIM, ALL OTHER REMEDIES AND WARRANTIES, EXPRESS, STATUTORY, IMPLIED, OR OTHERWISE, INCLUDING, BUT NOT LIMITED TO, ANY WARRANTIES OF MERCHANTABILITY, SATISFACTORY QUALITY, NONINFRINGEMENT OR FITNESS FOR A PARTICULAR PURPOSE, OR REGARDING RESULTS OBTAINED THROUGH THE USE OF ANY PRODUCT OR SERVICE (INCLUDING, WITHOUT LIMITATION, ANY CLAIM OF INACCURATE, INVALID OR INCOMPLETE RESULTS), IN EACH CASE HOWEVER ARISING, INCLUDING WITHOUT LIMITATION FROM A COURSE OF PERFORMANCE, DEALING OR USAGE OF TRADE, OR OTHERWISE. TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, ARIMA AND ITS LICENSORS, SUPPLIERS AND REPRESENTATIVES SHALL NOT BE LIABLE FOR LOSS OF USE, PROFITS, REVENUE, GOODWILL, BUSINESS OR OTHER FINANCIAL LOSS OR BUSINESS INTERUPTION, OR COSTS OF SUBSTITUTE GOODS OR SERVICES, OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, EXEMPLARY OR INDIRECT DAMAGES FOR BREACH OF WARRANTY.

WARRANTY

All warranties are personal to the Purchaser and may not be transferred or assigned to a third-party, including an affiliate of the Purchaser. The warranty described below excludes any stand-alone third-party goods that may be acquired or used with the Product. Arima Genomics only warrants that the kit reagents will be made and tested in accordance with Arima Genomics manufacturing and quality control processes. Arima Genomics makes no warranty that the reagents provided in this kit will work as intended by the Purchaser or for the Purchaser's intended uses. ARIMA GENOMICS MAKES NO OTHER WARRANTY, EXPRESSED OR IMPLIED. THERE IS NO WARRANTY OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. The warranty provided herein and the data and descriptions of Arima Genomics products appearing in Arima Genomics product literature and website may not be altered except by express written agreement signed by an officer of Arima Genomics. Representations, oral or written, which are inconsistent with this warranty or such publications are not authorized and if given, should not be relied upon.

The foregoing warranties do not apply to the extent a non-conformance is due to (i) abuse, misuse, neglect, negligence, accident, improper storage, or use contrary to the Documentation or Specifications, (ii) use that is an Excluded Use, (iii) improper handling, (iv) unauthorized alterations, (v) natural disasters, or (vi) use with a third-party's good that is not specified in the product documentation. In the event of a breach of the foregoing warranty, customer shall promptly contact Arima Genomics customer support to report the non-conformance and shall cooperate with Arima Genomics in confirming or diagnosing the non-conformance. Additionally, Arima Genomics may request return shipment of the non-conforming product at Arima Genomics cost. Arima Genomics sole obligation shall be to replace the applicable product or part thereof, provided the customer notifies Arima Genomics within 90 days of any such breach. If after exercising reasonable efforts, Arima Genomics is unable to replace the product, then Arima Genomics shall refund to the Purchaser all monies paid for such applicable product.

CONTACT US

Technical Support: techsupport@arimagenomics.com
Order Support: ordersupport@arimagenomics.com