Arima-HiC+/Arima-High Coverage HiC Kit

Arima-HiC Bioinformatics User Guide

Material Part Number: A160600 Document
Part **Number:** A160600 v00
Release Date: 07/19/2021

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$^+$ and Arima-High Coverage HiC Kits, PN **A4110110**, 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$^+$ and Arima-High Coverage HiC kits. 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: A160600 Document Part Number: A160600	July 2021	Initial Release

Table of Contents

1. Introduction

1.1 Arima-HiC Workflow Overview

Hi-C is an adaptation of the location-specific chromosome conformation capture (3C) workflow that achieves genome-wide interrogation of higher order chromatin structure. HiC combines in situ crosslinking and restriction digestion with biotin incorporation, ligation, random fragmentation, and streptavidin precipitation of ligated fragments that previously were in close three-dimensional proximity. The Arima-HiC workflows improved upon the original single restriction enzyme-based digestion by incorporating 2 restriction enzymes, and in the case of the Arima High Coverage HiC, a 4-enzyme cocktail. Both Arima-HiC workflows were extensively optimized for the analysis of fresh frozen or blood tissues as well as cultured or FACS sorted cells/nuclei from vertebrate, insect or plant samples. Typical input for the Arima-HiC protocol includes 500K mammalian cells, 50mg frozen animal tissue, 1 insect, 2ml of blood (50-100ul if nucleated) or 125-250mg of plant tissue, with lower input protocols also available. Another protocol improvement was the development of 2 QC assays evaluating the success at intermediary steps before submitting the sample library for next generation sequencing. This bioinformatics user guide will not focus on the QC steps incorporated into the Arima HiC workflows and the reader is referred to the Arima-HiC user guide for more information. This user guide will, however, discuss how to evaluate next generation sequencing data to understand the quality of the HiC library.

1.2 Experimental Planning

Deciding on the number of replicate samples per condition and the depth of sequencing per sample depends on several factors. If the primary limiting factor is the availability of biological replicates, we recommend deep interrogation of at least 2 samples to a minimum of 600M read pairs. It is recommended to evaluate library complexity (estimated number of uniquely mapping DNA fragments) and the abundance of reads representing long range intra-chromosomal interactions prior to deep sequencing through any low-cost sequencing approach that achieves 0.5-2M read pairs per sample. A complex Arima-HiC library generated from recommended amounts of a high-quality sample will identify 3D loops at 5-10kb resolution from 600M read pairs. Lower starting input or sample quality will reduce the efficiency of this process and might require deeper interrogation if the library complexity is insufficient, or the aggregation of data from additional replicate samples. Statistical analysis of 3D interaction maps has primarily focused on detecting enriched signal over background contacts (within sample analysis), and reproducibility across 2 replicate samples has primarily utilized concordance analysis. One benchmark for reproducibility requires 2 replicates to achieve better than 70% concordance in loops identified to be considered reproducible.

1.3. Arima-HiC Analysis Overview

With the development of the Arima High Coverage HiC product update, it has become much more feasible to expand the applications of this workflow towards whole genome variant analysis, variant phasing or scaffolding of de novo assembled contigs. This bioinformatics user guide will focus on the interrogation of the three-dimensional chromatin structure, contact map generation and the annotation and visualization of these results, and will only provide a rudimentary overview of other recommended analysis tools. For analysis applications that require an Arima-HiC specific mapping pipeline we would like to refer the user to the Arima mapping pipeline (https://github.com/ArimaGenomics/mapping_pipeline), which can be utilized prior to scaffolding a de novo assembly via SALSA2 (Ghurye et al., 2019). We have also successfully used the BWA-MEM mapping tool (Li, 2013) prior to identifying genomic variants via the GATK pipeline (DePristo et al., 2011; Van der Auwera et al., 2013), HapCut2 (Edge eta al., 2016) for variant phasing and, in addition to SALSA2 (Ghurye et al., 2019), the use of 3D-DNA (Dudchenko et al., 2017) for scaffolding de novo assemblies. It has been demonstrated that Arima High Coverage HiC is ideally suited to phase PacBio HiFi reads to generate chromosome-spanning phased de novo assemblies through the use of the DipAsm pipeline (Garg et al., 2020).

For a comprehensive overview of the available tools for HiC analysis we would like to refer the reader to this recent publication: Han and We, 2017. The data analysis section of the 4D nucleome consortium (https://www.4dnucleome.org/software.html) provides a good overview as well.

A significant portion of the 3D genome structure field has adopted the *Juicer* pipeline for data pre-processing and *JuicerTools* for structural feature discovery (e.g. compartments, TADs, and chromatin loops) (Durand and Shamim et al., 2016). In addition, *Juicebox* was developed for data visualization (Durand and Robinson et al., 2016) by the same team. Since *Juicer* produces a comprehensive set of QC metrics for HiC data without having to specify or identify any re-ligation motive(s), Arima Genomics has decided to exclusively utilize the Juicer pipeline to assess Arima-HiC data quality. The landscape for HiC analysis tools is still evolving, so we want to point out a very recent addition that appears to improve the interpretation of chromosome contact maps and the identification of chromosome loops, HiSIF (Zhou et al., 2020).

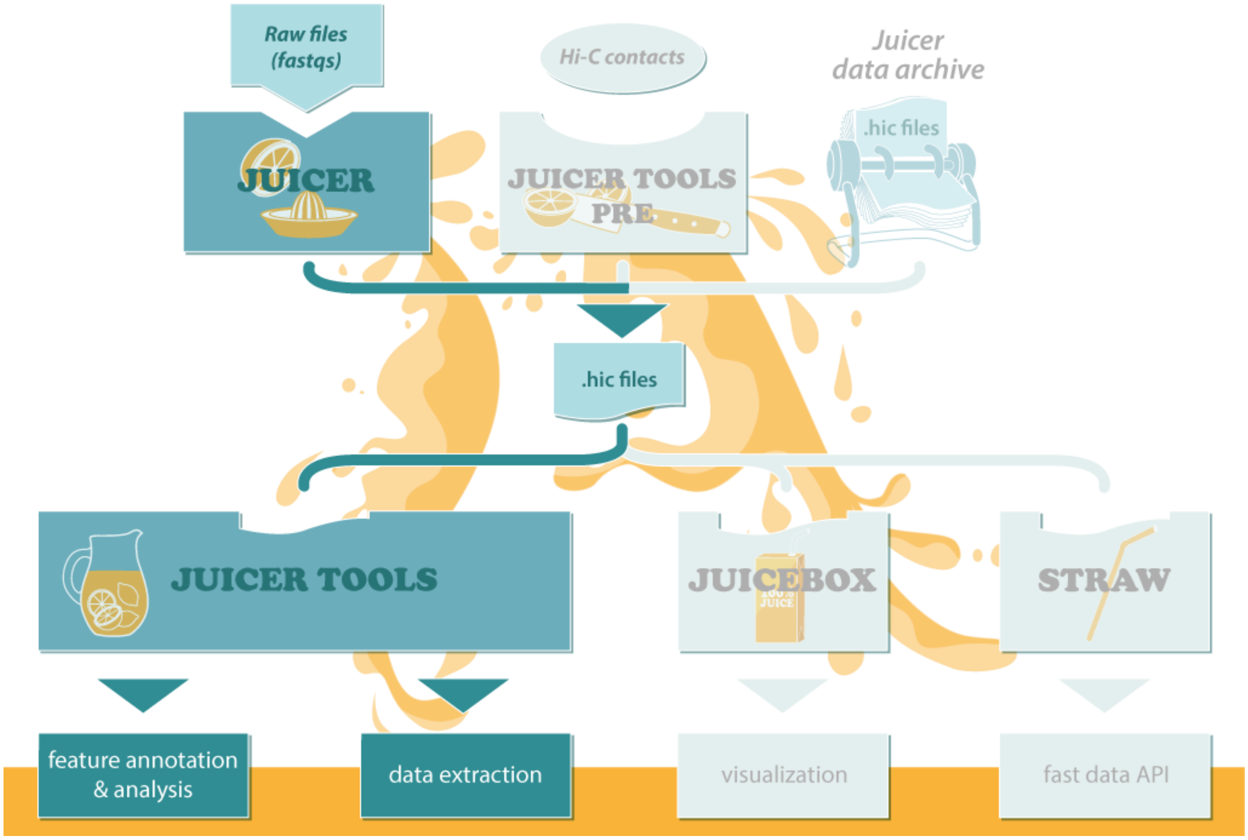


Figure 1. High level overview of the *Juicer, Juicer tools* and *Juicebox* HiC analysis pipeline

2. GETTING STARTED WITH JUICER

The Juicer pipeline is designed to run in a high performance unix compute environment.

For a mammalian genome-wide HiC dataset of 1.2 billion reads/600M read pairs it is suggested to use a single high memory (256GB-512GB) node or server using 20 cores.

Juicer is also available for various resource management softwares: OpenLava, LSF, SLURM and GridEngine.

Due to the varying nature of these deployments, Arima cannot provide support for these on your local system.

Prior to installing Juicer, several dependencies will need to be made available as described at https://github.com/aidenlab/juicer/wiki/Installation#dependencies

- For alignment and creation of the Hi-C pairs file **merged_nodups.txt**:
  - GNU CoreUtils
  - Burrows-Wheeler Aligner (BWA)
- For **.hic** file creation and Juicer tools analysis:
  - Java 1.7 or 1.8 JDK. (Alternative link for Ubuntu/LinuxMint). Minimum system requirements for running Java can be found at http://java.com/en/download/help/sysreq.xml
  - Latest Juicer Tools jar
- For peak calling:
  - CUDA when utilizing a NVIDIA GPU for genome-wide pairwise analysis by HiCCUPS
  - The native libraries included with Juicer are compiled for CUDA 7. Other versions of CUDA can be used, but you will need to download the respective native libraries from JCuda.
  - For best performance, use a dedicated GPU. You may also be able to obtain access to GPU clusters through Amazon Web Services or a local research institution.
  - A modified version of HiCCUPS – CPU-HiCCUPS - is available for analyzing pairwise comparisons within 8MB of each other since "The vast majority of peaks (98%) reflected loops between loci that are <2 Mb apart" (Rao and Huntley et al. 2014)
  - CPU-HiCCUPS

To install Juicer please follow these instructions: https://github.com/aidenlab/juicer/wiki/Installation

Further, Juicer is also available on AWS. For detailed instructions please refer to https://github.com/aidenlab/juicer/wiki/Running-Juicer-on-Amazon-Web-Services

The input file requirements for Juicer are as follow:

- Paired end sequence files ***_R1*.fastq** and ***_R2*.fastq**. We highly recommend 2x150bp read length.
- An in silico digested reference genome to identify read pairs that map within one predicted fragment or use fragment delimited resolutions for loop calling can be generated by the script included with juicer: https://github.com/aidenlab/juicer/blob/master/misc/generate_site_positions.py with the 'Arima' flag. Novel restriction enzyme patterns can be added as additional elements in the scripts pattern section. Contact techsupport@arimagenomics.com for the recognition sequences of the enzyme cocktail for Arima High Coverage HiC.

```
$ python ./generate_site_positions.py Arima hg38 hg38.fa
$ head hg38_Arima.txt
$ chr1 11160 11507 11522 11685…
```

- For human and mouse reference genomes we made the restriction site files for Arima-HiC+ available here: ftp://ftp-arimagenomics.sdsc.edu/pub/JUICER_CUTSITE_FILES
- When not providing a restriction site file for your reference genome to use fragment delimited resolutions, one must run Juicer v1.5 with the -x flag. In Juicer v1.6, fragment maps are no longer included in the .hic file by default but can now be included via the -f flag

  .
- Your reference genome as a `fasta` (.fasta/.fa/.fna) file with the bwa index file in the same location

```
$ bwa index reference.fa
```

- A sizes.genome file containing all the chromosome sizes. It can be generated via samtools or awk

```
$ samtools faidx input.fa
$ cut -f1,2 input.fa.fai > sizes.genome

$ awk 'BEGIN{OFS="\t"}{print $1, $NF}' mygenome_myenzyme.txt >
```

3. RUNNING THE JUICER PIPELINE, OUTPUT FILES AND DATA QC

3.1 Running Juicer

The directions below apply to all systems, including the single node version.

1. Choose your cluster system or single CPU. Juicer is currently available in the cloud on AWS, on LSF, Univa, or SLURM, or on a single CPU.

2. Follow the instructions in the Installation section. Be sure you know how to load the required software on your system; cluster systems might have slightly different names, and you might need to change the master "juicer.sh" script to reflect this.

3. Log into your cluster.

4. Install the appropriate Juicer scripts for your system in a directory; we will assume this directory is `/home/user/juicedir`. For example, if you were using SLURM, you would copy the folder `scripts` underneath SLURM to `/home/user/juicedir/scripts`

5. Under `/home/user/juicedir`, there should be a folder `references` that contains the reference fasta file for your genome and the BWA index files. You can soft-link if necessary, or otherwise download the fasta files from UCSC and run `bwa index` on the fasta file.
   A complete set of frequently used input or reference files is available on the Juicer AWS mirror.

6. Under `/home/user/juicedir`, you should also create a folder `restriction_sites`. This should contain your restriction site file. You can create this file using the generate_site_positions.py Python script, or download already created ones from the Juicer AWS mirror or the Arima Genomics ftp site ftp://ftp-arimagenomics.sdsc.edu/pub/JUICER_CUTSITE_FILES

7. [Optional, only for deep maps]
   Create the motif folder `/home/user/juicedir/references/motif` and populate the two folders "`unique`" and "`inferred`" with data from the Juicer AWS mirror. These folders should contain a combination of RAD21, SMC3, and CTCF ChIP-seq narrow peak BED files generated for GM12878. They should be replaced with appropriate ChIP-Seq files for the relevant sample. Alternative, MotifFinder can be called independently with command line tools at a later time, after the HiCCUPS loop list has been created. For more details, see https://github.com/aidenlab/juicer/wiki/MotifFinder.

8. Create a custom directory (e.g. `mkdir -p /custom/filepath/MyHIC`)

9. Download the test data.

   - Option 1: To see how Juicer runs on a deep sequencing test data set, download the following, consisting of chromosome19 from the Cell 2014 in-situ combined GM12878 map:
     - chr19_R1.fastq.gz
     - chr19_R2.fastq.gz

- Option 2: To run Juicer on a small test data set, download the following MiSeq GM12878 in-situ files:
    - HIC003_R1.fastq.gz
    - HIC003_R2.fastq.gz

10. **For each sample or replicate** one wants to analyze independently, create a fastq directory under the top directory (`cd /custom/filepath/MyHIC; mkdir fastq`). Soft-link or copy your fastq files from shallow and deep sequencing (zipped or unzipped) to that directory.

11. Type `screen` then launch Juicer:

```
$ screen
$ /home/user/juicedir/scripts/juicer.sh [options]
```

Test data:
Running without any options will default to the genome of hg19 and the restriction site of MboI.

For Arima data use the following parameters after CD into the `/custom/filepath/MyHIC`

```
~/juicedir/scripts/juicer/juicer.sh
      -d /path/ to directory containing the /fastq/ sequence reads data folder
      -p /chromosome_sizes/reference_genome.chrom.sizes
      -y /in_silico_digested_genome/reference_genome.digested.txt
      -z /reference/reference_genome.fa
      -D /tools/juicer/ path to juicer   >juicer.log 2>&1
```

The files will be split if necessary and Juicer will launch.

12. Sample output; the "exit code 0" statement means that the split successfully completed.

```
Your job 95416 ("a1439405283split0") has been submitted
Job 95416 exited with exit code 0.
(-: Finished adding all jobs... please wait while processing.
```

13. Single node script will run until it finishes or exits.
14. If there are no jobs left, type `cat debug/finalcheck*;` you should see a "Pipeline successfully completed" message. For some clusters, there might be only one file, e.g. `lsf.out` or `uger.out`; in this case, type `tail lsf.out` to see the message.

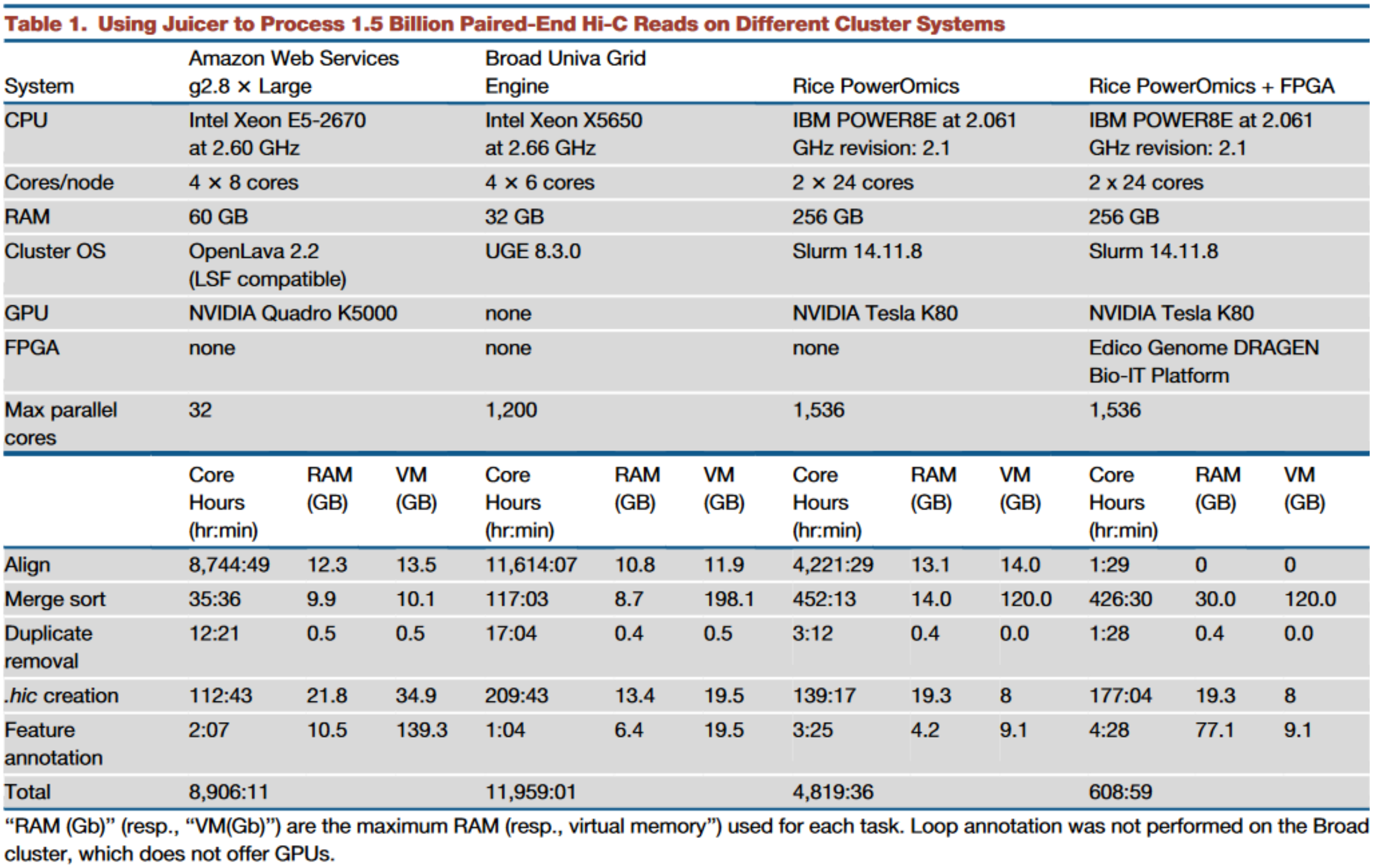


Table 1. Using Juicer to Process 1.5 Billion Paired-End Hi-C Reads on Different Cluster Systems

System	Amazon Web Services g2.8 × Large	Broad Univa Grid Engine	Rice PowerOmics	Rice PowerOmics + FPGA
CPU	Intel Xeon E5-2670 at 2.60 GHz	Intel Xeon X5650 at 2.66 GHz	IBM POWER8E at 2.061 GHz revision: 2.1	IBM POWER8E at 2.061 GHz revision: 2.1
Cores/node	4 × 8 cores	4 × 6 cores	2 × 24 cores	2 x 24 cores
RAM	60 GB	32 GB	256 GB	256 GB
Cluster OS	OpenLava 2.2 (LSF compatible)	UGE 8.3.0	Slurm 14.11.8	Slurm 14.11.8
GPU	NVIDIA Quadro K5000	none	NVIDIA Tesla K80	NVIDIA Tesla K80
FPGA	none	none	none	Edico Genome DRAGEN Bio-IT Platform
Max parallel cores	32	1,200	1,536	1,536

	Core Hours (hr:min)	RAM (GB)	VM (GB)	Core Hours (hr:min)	RAM (GB)	VM (GB)	Core Hours (hr:min)	RAM (GB)	VM (GB)	Core Hours (hr:min)	RAM (GB)	VM (GB)
Align	8,744:49	12.3	13.5	11,614:07	10.8	11.9	4,221:29	13.1	14.0	1:29	0	0
Merge sort	35:36	9.9	10.1	117:03	8.7	198.1	452:13	14.0	120.0	426:30	30.0	120.0
Duplicate removal	12:21	0.5	0.5	17:04	0.4	0.5	3:12	0.4	0.0	1:28	0.4	0.0
.hic creation	112:43	21.8	34.9	209:43	13.4	19.5	139:17	19.3	8	177:04	19.3	8
Feature annotation	2:07	10.5	139.3	1:04	6.4	19.5	3:25	4.2	9.1	4:28	77.1	9.1
Total	8,906:11			11,959:01			4,819:36			608:59		

"RAM (Gb)" (resp., "VM(Gb)") are the maximum RAM (resp., virtual memory") used for each task. Loop annotation was not performed on the Broad cluster, which does not offer GPUs.

Figure from Durant et al, 2016 providing Juicer analysis benchmarks for various compute configurations.

3.2 QC Analysis of Arima-HiC data by Juicer

1. Results are available in the aligned directory. The Hi-C maps are in `inter.hic` (for MAPQ > 0) and `inter_30.hic` (for MAPQ >= 30). The Hi-C maps can be loaded in Juicebox and explored. They can also be used for feature annotation and analysis and to extract matrices at specific resolutions. You can also directly manipulate them with the Straw API

2. These results also include automatic feature annotation. The output files include a genome-wide annotation of loops and, whenever possible, the CTCF motifs that anchor them (identified using the HiCCUPS algorithm). The files also include a genome-wide annotation of contact domains (identified using the Arrowhead algorithm). The formats of these files are described in the Juicebox tutorial online; both files can be loaded into Juicebox as a 2D annotation.

3. When the pipeline has completed successfully, you will see the folders `aligned`, `debug`, and `splits`. The `debug` folder contains logging information for the pipeline. The `splits` folder is a temporary working directory and can be deleted once you are sure the pipeline ran successfully. The `aligned` folder contains the results:
   - **`inter.hic` / `inter_30.hic`**: The .hic files for Hi-C contacts at MAPQ > 0 and at MAPQ >= 30, respectively
   - `merged_nodups.txt`: The Hi-C contacts with duplicates removed. This file is also input to diploid pipelines

- `collisions.txt`: Reads that map to more than two places in the genome
- **`inter.txt`,** `inter_hists.m` **/ `inter_30.txt`,** `inter_30_hists.m`: The statistics and graphs files for Hi-C contacts at MAPQ > 0 and at MAPQ >= 30, respectively. These are also stored within the respective .hic files in the header. The .m files can be loaded into Matlab. The statistics and graphs are displayed under Dataset Metrics when loaded into Juicebox
- `dups.txt`, `opt_dups.txt`: Duplicates and optical duplicates
- `abnormal.sam`, `unmapped.sam`: Abnormal chimeric and unmapped reads
- `merged_sort.txt`: This is a combination of merged_nodups / dups / opt_dups and can be deleted once the pipeline has successfully completed
- `stats_dups.txt` / `stats_dups_hists.m`: Statistics and graphs on the duplicates

4. You should run the script `cleanup.sh` to zip all the text files and delete the unnecessary `splits` directory and `merged_sort.txt` file once you are sure the pipeline has successfully completed.

3.3 Hi-C quality evaluation

As data is processed using the Juicer pipeline, summary statistics are saved in the **`inter.txt`** and **`inter_30.txt`** files, which results in the following output for a high quality Arima-HiC experiment:

```
Sequenced Read Pairs: 1,241,427,115
Normal Paired: 503,620,615 (40.57%)
Chimeric Paired: 584,618,626 (47.09%)
Chimeric Ambiguous: 139,116,434 (11.21%)
Unmapped: 14,071,440 (1.13%)
Ligation Motif Present: 0 (0.00%)
Alignable (Normal+Chimeric Paired): 1,088,239,241 (87.66%)
Unique Reads: 967,863,976 (77.96%)
PCR Duplicates: 112,884,114 (9.09%)
Optical Duplicates: 7,491,151 (0.60%)
Library Complexity Estimate: 4,806,623,533
Intra-fragment Reads: 1,521,030 (0.12% / 0.16%)
Below MAPQ Threshold: 81,372,635 (6.55% / 8.41%)
Hi-C Contacts: 884,970,311 (71.29% / 91.44%)
Ligation Motif Present: 0 (0.00% / 0.00%)
3' Bias (Long Range): 60% - 40%
Pair Type %(L-I-O-R): 25% - 25% - 25% - 25%
Inter-chromosomal: 146,456,197 (11.80% / 15.13%)
Intra-chromosomal: 738,514,114 (59.49% / 76.30%)
Short Range (<20Kb): 298,616,708 (24.05% / 30.85%)
Long Range (>20Kb): 439,895,087 (35.43% / 45.45%)
```

This file provides a wide range of statistics. There are a few we would like to highlight as the most important for assessing your overall Arima-HiC data quality.

Normal Paired: Both paired end reads map uniquely to the reference genome. High utility for HiC, scaffolding and variant calling/phasing applications.

Chimeric Paired: Both 5' ends of a paired end read map to uniquely different locations in the reference genome, but at least one read requires a split in the alignment. An indication that the read overlaps the re-ligation junction. Juicer and its native `bwa` aligner do not utilize ligation junction information for mapping. The uniquely mapped portions of the reads are useful for HiC, scaffolding and variant calling/phasing applications.

Alignable: The sum of the Normal and Chimeric Paired reads. In a successful Arima HiC experiment we expect to align more than 80% of all reads into paired reads for HiC analysis.

Chimeric Ambiguous: At least one of the 5' ends of a paired end read is too short to be mapped uniquely to the reference genome, either due to the proximity of the re-ligation junction or its placement in a repetitive region in the genome. The uniquely mapped portions of a split mapped read are useful for variant calling or phasing applications, but the lack of unambiguous contact information renders these reads ineffective for HiC analysis or scaffolding. Chimeric Ambiguous reads for successful Arima HiC experiments are expected to make up less than 20% of all reads.

Unmapped: Unmapped reads cannot be placed in the reference genome. Factors contributing to elevated percentages of unmapped reads include very low read quality, an incomplete reference genome assembly or contamination with other genomes not represented in the reference. Successful Arima HiC experiments contain 6% or less unmapped reads.

Unique Reads: This statistic provides the total number of alignable reads which mapped uniquely to the reference genome, with PCR and optical duplicates removed. Deduplication involves collapsing reads that map to identical genome locations for both paired end reads into one contact. It also reports the percentage relative to the total number of Sequenced Read Pairs that these unique reads represent. Increasing the sequencing depth for a library lowers the relative abundance of unique reads, which reduces the utility of this metric. Consult the library complexity estimate for a better proxy for the abundance of unique reads.

PCR Duplicates and Library Complexity Estimate: This statistic provides the total number of PCR duplicates in your data set, as well as the percentage relative to the Sequenced Read Pairs that these duplicates represent. The number of PCR duplicates is a function of the complexity of the library (how many unique molecules are estimated to be present in the amplified library) and the depth of sequencing. A complex library is characterized by very low PCR duplicates in shallow sequencing (<1%) which only slowly increases with deeper sequencing. The Juicer pipeline estimates the complexity of the library by extrapolating the duplication rate from the sequencing reads. Shallow sequencing has a larger uncertainty in this estimate compared to deeper sequencing. The estimated diversity of Arima HiC libraries is heavily impacted not only by the quality of the Hi-C reaction, but also by the starting amounts with respect to the number of cells or DNA contained within.

Deciding on a target budget for sequencing depth is heavily dependent on the experimental specifics, but lower quality samples might be better served by sequencing multiple independent samples to a lower depth.

Hi-C Contacts: This statistic provides the total number of reads that contributed to the `.hic` contact matrix in the Juicer pipeline. Often these reads are referred to as the "usable" reads since they make up the final analysis file which is used for the identification of loops, TADs and A/B compartments.
Note: **For this and all subsequent statistics there are two percentages reported; the first is the percentage out of the Sequenced Read Pairs; the second out of the Unique Reads.** We expect a high quality Arima HiC library to produce Hi-C contacts for more than 60% of all alignable reads (statistic not reported in the inter_30.txt file but easily calculated).

Inter-chromosomal: This statistic provides the total number of reads which map to interactions occurring between chromosomes or scaffolds (physically connected contigs). This statistic is also referred to as "trans interactions." Though historically these interactions have been thought of as artifactual, it is possible that this data contains true interactions that may be of interest for the detection of chromosomal translocation events and other structural abnormalities. For the majority of successful HiC experiments, we expect to see ~20% or fewer "trans interactions" out of the total HiC Contacts. Elevated levels of inter-chromosomal contacts could indicate a deterioration of the nuclear matrix integrity prior to formaldehyde crosslinking or an overabundance of burst nuclei prior to completion of the HiC reactions.

Short Range (<20Kb): This statistic provides the total number of reads which map to interactions occurring on the same chromosome, with a distance less than 20kb apart. This statistic is also referred to as "short-cis interactions." Though these interactions can provide useful information for 3D structure, they can also represent short linear distances along DNA which will map along the diagonal of your .hic contact map. This statistic exhibits a wide range of abundances, and since a low percentage of short range interactions could be due to many inter-chromosomal interactions (undesired) or high levels of long range cis interactions (desired), it does not provide a strong indicator for HiC library quality.

Long Range (>20Kb): This statistic provides the total number of reads which map to interactions occurring on the same chromosome or scaffold, with a distance greater than 20kb apart. This statistic is also referred to as "long-cis interactions." These interactions provide the most useful information for studying long range 3D structures as they represent long linear distances along a DNA molecule that are in proximity to each other and potentially interact functionally via loops, TADs or A/B compartmentalization. For good quality data that warrants deep sequencing without reservation, we expect to see >40% "long-cis interactions" out of the total Unique HiC Contacts. At the very least we would expect successful Arima-HiC experiments to produce more long range cis than short range cis interactions. Inadequate abundance of long-range cis interactions might be related to insufficient crosslinking, which would require additional optimization of crosslinking conditions, or sub-optimal efficiency of the Hi-C proximity ligation reaction.

3.4 Feature Annotation by JuicerTools

The output files from the Juicer program (2.2) are used automatically as input for various tools to create HiC contact files. The feature annotation algorithms all operate on a highly compressed **`.hic`** binary file. You can use `java -jar juicebox_tools.jar pre` to create a .hic file from your text data; alternatively, all the algorithms work on URLs as well, so you can use the links in the [Aiden Lab Hi-C Archive] (http://aidenlab.org/data.html) to operate directly on that data without downloading it. For more information on the command line use of the JuicerTools pre command please see here: https://github.com/aidenlab/juicer/wiki/Pre

The default tools for 3D chromatin feature discovery and annotation used by JuicerTools are:

- Arrowhead for finding contact domains
- HiCCUPS for discovering locally enriched contact peaks that mark DNA loop anchors
- Eigenvector for determining A/B compartments
- MotifFinder for finding DNA Motifs for Loops

Arrowhead

Contact domains represent continuous partitions of a genome that exhibit elevated proximity to each other (average 200kb) than would be expected by their linear distance. They are typically represented by squares along the diagonal of a contact map. TADs (topologically associating domains) represent the same concept but at a slightly lower resolution (average size of >800kb) as they were defined by a directionality index across 40kb bins in earlier work (Dixon et al., 2012). Both TADs and contact domains are frequently formed by CTCF-anchored DNA loops in that 2 CTCF motives face in opposite directions to define the contact domain. The original topology-associated domain definition employed a directionality index that switches direction at domain boundaries. The Arrowhead algorithm also measures the directionality preference of a locus that is restricted to contacts at a pre-determined linear distance. The dynamic programming algorithm acts upon a user-defined sliding window size (must be even) and the data is aggregated at a user-defined resolution. The Arrowhead algorithm is described in detail in the Supplemental Experimental Procedures of Rao et al., 2014.

HiCCUPS

Many contact domains are defined by long range DNA loops anchored by CTCF but can also contain additional internal loops of varying sizes. When two non-adjacent regions on the same chromosome are actively held together by DNA-binding protein(s), their proximity in three-dimensional space is established by the Hi-C proximity ligation step and leads to the enrichment in the number of contacts on the two-dimensional contact map. The HiCCUPS algorithm (Hi-C Computational Unbiased Peak Search) was designed to identify such locally enriched contact peaks that exclusively represent DNA loops and not simply the edge of a contact domain formed by other processes. Since HiCCUPS calculates multiple contact expectations around each pixel (pairwise bin) instead of one global signal expectation, it exhibits a significantly reduced false positive rate for identifying contact peaks representing loops.

By default, HiCCUPS uses a GPU CUDA node for analysis. If not available, it is possible to run it in CPU mode as described at https://github.com/aidenlab/juicer/wiki/CPU-HiCCUPS. To keep computational burden manageable, the peak finding algorithm of CPU-HiCCUPS restricts the search to 8MB of the diagonal. In practice, most loops (especially CTCF mediated chromatin loops) are within a few megabases of the diagonal.

Eigenvector

When an intra-chromosomal contact matrix was originally converted into an observed/expected matrix at 1MB resolution, it was found that the first principal component (the eigenvector) separated the genome bins into 2 clusters which were then described as compartments. Compartment (A) was enriched for open chromatin marks, while the other cluster (B) was enriched for closed chromatin and it was shown that the A/B compartment structure was highly cell type specific. The Eigenvector algorithm in Juicer recapitulates this coarse genome structure classification, but the utility of a binary classification scheme has been questioned as insufficient to recapitulate the diversity of large-scale chromosomal organization and has been amended to include 6 sub-compartments (A1,A2,B1,B2,B3,B4) by Rao et al., 2014.

MotifFinder (optional)

In order for MotifFinder to search for known conserved binding sites in HiC loop data, one needs to provide cell-type and condition-matched ChIP-Seq peak files and the loop list to the algorithm (see section 2.2.7). Arima Genomics recently released a separate Arima-HiChIP workflow focused on interrogating the three-dimensional interactions associated with various histone modifications and the binding of CTCF or a few select transcription factors. The ability of Arima-HiChIP data to identify ChIP-seq peaks via the MAPS v2.0 pipeline and mine the peak regions for novel enriched sequence motifs in addition to the identification of enriched contact peaks potentially reduces the need for MotifFinder analysis of genome-wide HiC data.

3.5 How to Cite Juicer in Publications

If you use Juicer in your research, please cite: Neva C. Durand, Muhammad S. Shamim, Ido Machol, Suhas S. P. Rao, Miriam H. Huntley, Eric S. Lander, and Erez Lieberman Aiden. "Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments." Cell Systems 3(1), 2016. https://doi.org/10.1016/J.CELS.2016.07.002

4. VISUALIZING ARIMA-HIC RESULTS

4.1 Visualization of Arima Hi-C data via Juicebox.

The native 3D chromatin visualization tool based on the 2D contact matrix that is built into the Juicer analysis workflow is called Juicebox. Installation and use are extensively documented here: https://github.com/aidenlab/Juicebox and recently updated here: https://aidenlab.gitbook.io/juicebox/.

The java jar file can be downloaded here: https://github.com/aidenlab/Juicebox/wiki/Download.

In addition to installing the source code for a local deployment, one can also use a web-based installation here:

https://aidenlab.org/juicebox/

Examples of Juicebox for editing genome assemblies can now be found here: https://www.dnazoo.org/methods

Links to tutorial videos for Juicebox: https://www.youtube.com/watch?v=xjNXyeUSfZM

An introduction to Juicebox by a coauthor of the Juicebox publication, Neva C. Durand:

https://www.youtube.com/watch?v=xA6CLsG_GAs

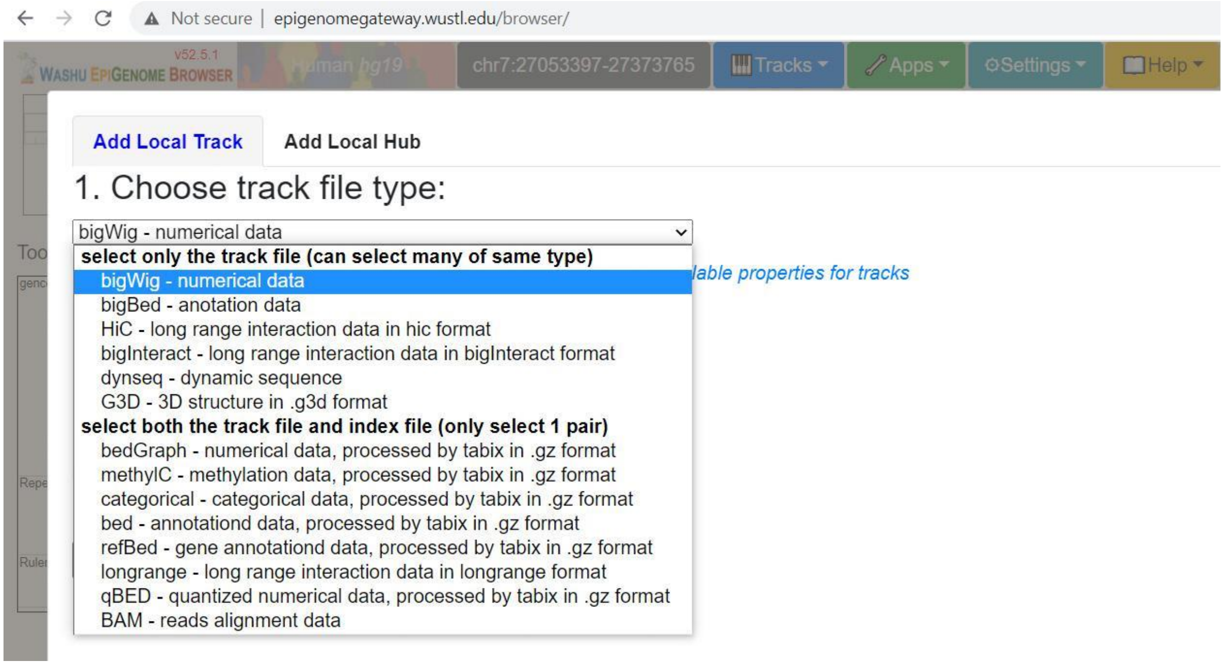
Instructions on how to manually edit a de novo assembly while observing the 2D contact matrix:

https://www.youtube.com/watch?v=Nj7RhQZHM18

4.2 Integrated visualization of Arima-HiC data via the WashU Epigenome Browser

In addition to visualizing 3D contact data via a 2D matrix, it is also possible to visualize individual contacts in linear genome space through ARC plots. Visualizing Hi-C contacts in the context of genomic annotations and overlaid with functional genomics data allows changes in Hi-C loop or domain features to be correlated with these multi-omics structure-function relationships. The most popular open-source tool for that purpose is the WashU Epigenome Browser: http://epigenomegateway.wustl.edu/browser/.

Examples of tracks that can be added to the underlying reference genome can be seen in this intake form:



1. Choose track file type:

bigWig - numerical data

select only the track file (can select many of same type)
- bigWig - numerical data
- bigBed - anotation data
- HiC - long range interaction data in hic format
- bigInteract - long range interaction data in bigInteract format
- dynseq - dynamic sequence
- G3D - 3D structure in .g3d format

select both the track file and index file (only select 1 pair)
- bedGraph - numerical data, processed by tabix in .gz format
- methylC - methylation data, processed by tabix in .gz format
- categorical - categorical data, processed by tabix in .gz format
- bed - annotationd data, processed by tabix in .gz format
- refBed - gene annotationd data, processed by tabix in .gz format
- longrange - long range interaction data in longrange format
- qBED - quantized numerical data, processed by tabix in .gz format
- BAM - reads alignment data

As is evident from the previous figure, Juicer generated .hic files can be directly imported as unique track(s) into the visualization platform.

The use of the WashU Epigenome Browser is extensively documented on the website and the reader is hereby referred to the developer's extensive tutorial section containing documents and video tutorials: http://epigenomegateway.wustl.edu/support/

4.3 How to Cite Juicebox in Publications

If you use Juicebox in your research, please cite: James T. Robinson, Douglass Turner, Neva C. Durand, Helga Thorvaldsdóttir, Jill P. Mesirov, Erez Lieberman Aiden. "Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data." Cell Systems 6(2), 2018

https://doi.org/10.1016/j.cels.2018.01.001

4.4 How to cite the WashU Epigenome Browser in Publications

If you use the WashU Epigenome Browser in your research, please cite:

Xin Zhou, Daofeng Li, Bo Zhang, Rebecca F Lowdon, Nicole B Rockweiler, Renee L Sears, Pamela A F Madden, Ivan Smirnov, Joseph F Costello and Ting Wang, Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser, Nature Biotechnology (2015) doi:10.1038/nbt.3158

Xin Zhou, Daofeng Li, Rebecca F. Lowdon, Joseph F. Costello and Ting Wang, methylC Track: Visual integration of single-base resolution DNA methylation data on the WashU EpiGenome Browser, Bioinformatics 30 (15):2206-2207 (2014)

Xin Zhou, Rebecca F Lowdon, Daofeng Li, Heather A Lawson, Pamela A F Madden, Joseph F Costello, Ting Wang, Exploring long-range genome interactions using the WashU EpiGenome Browser Nature Methods 10, 375-376 (2013)

Xin Zhou, Ting Wang, Using the Wash U Epigenome Browser to Examine Genome-Wide Sequencing Data Current Protocols in Bioinformatics Unit 10.10

Xin Zhou, Brett Maricque, Mingchao Xie, Daofeng Li, Vasavi Sundaram, Eric A Martin, Brian C Koebbe, Cydney Nielsen, Martin Hirst, Peggy Farnham, Robert M Kuhn, Jingchun Zhu, Ivan Smirnov, W James Kent, David Haussler, Pamela A F Madden, Joseph F Costello & Ting Wang, The Human Epigenome Browser at Washington University Nature Methods 8, 989-990 (2011)

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 nonconformance 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