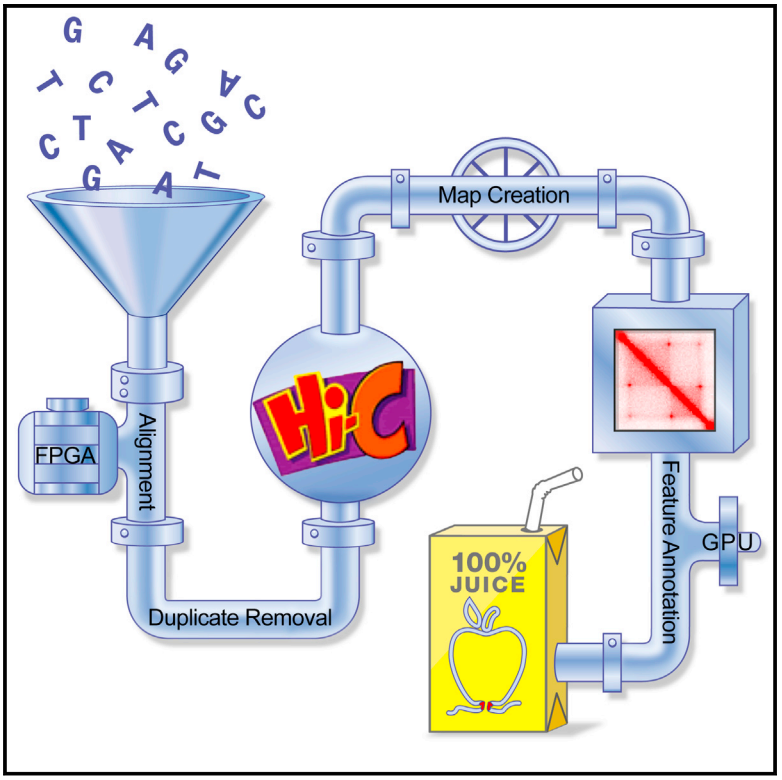
Tool

Cell Systems

Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments

Graphical Abstract



Authors

Neva C. Durand,
Muhammad S. Shamim, Ido Machol,
Suhas S.P. Rao, Miriam H. Huntley,
Eric S. Lander, Erez Lieberman Aiden

Correspondence

erez@erez.com

In Brief

Durand et al. introduce Juicer, a one-click, end-to-end pipeline for analyzing data from Hi-C and other contact mapping experiments. Juicer generates contact matrices at multiple resolutions and identifies features including contact domains and loops.

Highlights

- Juicer enables users to process terabase scale Hi-C datasets with a single click
- Juicer automatically annotates loops and contact domains
- Juicer is available as open source software
- Juicer is compatible with multiple cluster operating systems and with Amazon Web Services

Durand et al., 2016, Cell Systems *3*, 95–98
July 27, 2016
http://dx.doi.org/10.1016/j.cels.2016.07.002

CellPress

Cell Systems
Tool

Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments

Neva C. Durand,[1,2,3,4,10] Muhammad S. Shamim,[1,2,3,10] Ido Machol,[1,2,3] Suhas S.P. Rao,[1,2,3,5] Miriam H. Huntley,[1,2,3,6] Eric S. Lander,[4,7,8] and Erez Lieberman Aiden[1,2,3,4,9,*]
[1]The Center for Genome Architecture, Baylor College of Medicine, Houston, TX 77030, USA
[2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA
[3]Department of Computer Science and Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005, USA
[4]Broad Institute of Harvard and Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA
[5]School of Medicine, Stanford University, Stanford, CA 94305, USA
[6]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA
[7]Department of Biology, MIT, Cambridge, MA 02139, USA
[8]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
[9]Center for Theoretical Biological Physics, Rice University, Houston, TX 77030, USA
[10]Co-first author
*Correspondence: erez@erez.com
http://dx.doi.org/10.1016/j.cels.2016.07.002

SUMMARY

Hi-C experiments explore the 3D structure of the genome, generating terabases of data to create high-resolution contact maps. Here, we introduce Juicer, an open-source tool for analyzing terabase-scale Hi-C datasets. Juicer allows users without a computational background to transform raw sequence data into normalized contact maps with one click. Juicer produces a *hic* file containing compressed contact matrices at many resolutions, facilitating visualization and analysis at multiple scales. Structural features, such as loops and domains, are automatically annotated. Juicer is available as open source software at http://aidenlab.org/juicer/.

Hi-C experiments probe the 3D structure of DNA and chromatin by ligating and sequencing DNA loci that are spatially proximate to one another (Lieberman-Aiden et al., 2009; Rao et al., 2014). The resulting maps reflect patterns of physical contact between loci, making it possible to deduce how loci are organized in 3D.

Efforts to improve the resolution of 3D maps have caused the amount of DNA sequence produced from Hi-C experiments to skyrocket. Our original maps, derived from 30 million reads and 16 Gb of DNA sequence, described the genome at 1 megabase resolution (Lieberman-Aiden et al., 2009). In contrast, we recently generated 6.5 billion reads and 1.6 Tb of DNA sequence in order to create a single 3D map of the genome at kilobase resolution (Rao et al., 2014).

Although pipelines for Hi-C data analysis exist (Lieberman-Aiden et al., 2009; Schmid et al., 2015; Servant et al., 2015; Sauria et al., 2015), these packages are not designed to process datasets at the terabase scale or to annotate the structural features that these maps reflect. Moreover, when designing tools that require high-performance computation, ensuring reliability and ease-of-use across software platforms and hardware instances becomes a crucial desideratum. Ensuring such compatibility can be a considerable engineering challenge.

Here, we introduce Juicer, an easy-to-use, fully-automated pipeline for the processing and annotation of data from Hi-C and other contact mapping experiments. Juicer is closely based on the algorithms that we recently developed to analyze and annotate our terabase-scale Hi-C experiments (Rao et al., 2014). In order to meet the engineering challenge of handling such massive datasets, Juicer supports the use of parallelization and hardware acceleration whenever possible, including CPU clusters, general-purpose graphics processing units (GPGPUs), and field-programmable gate arrays (FPGAs). Juicer is also compatible with a variety of cloud and cluster architectures.

Juicer comprises three tools, which are designed to be run one-after-another (Figure 1).

First, Juicer transforms raw sequence data into a list of Hi-C contacts (pairs of genomic positions that were adjacent to each other in 3D space during the experiment). To accomplish this, read pairs are aligned to the genome, both duplicates and near-duplicates are removed, and read pairs that align to three or more locations are set aside. When appropriate hardware is available, this procedure can be accelerated either by parallelizing across multiple CPUs or by using an FPGA (see Table 1).

Next, the catalog of contacts is used to create contact matrices. To do so, the linear genome is partitioned into loci of a fixed size, or "resolution," (e.g., 1 Mb or 1 kb). These loci correspond to the rows and columns of a contact matrix; each entry in the matrix reflects the number of contacts observed between the corresponding pair of loci during a Hi-C experiment. Due to factors such as chromatin accessibility, certain loci are observed more frequently in Hi-C experiments. Juicer can adjust for these biases in multiple ways. The options include our original normalization scheme (Lieberman-Aiden et al., 2009), as well as a matrix balancing scheme that ensures that each row and column of the contact matrix sums to the same value (Knight and Ruiz, 2012). A wide array of quality statistics is also calculated, making it possible to assess the success and reliability of a given experiment before the costly deep-sequencing step.

The contact matrices generated in this way are stored efficiently in a compressed format that is designed to facilitate all subsequent computations. For instance, 1 terabyte of raw sequencing data is

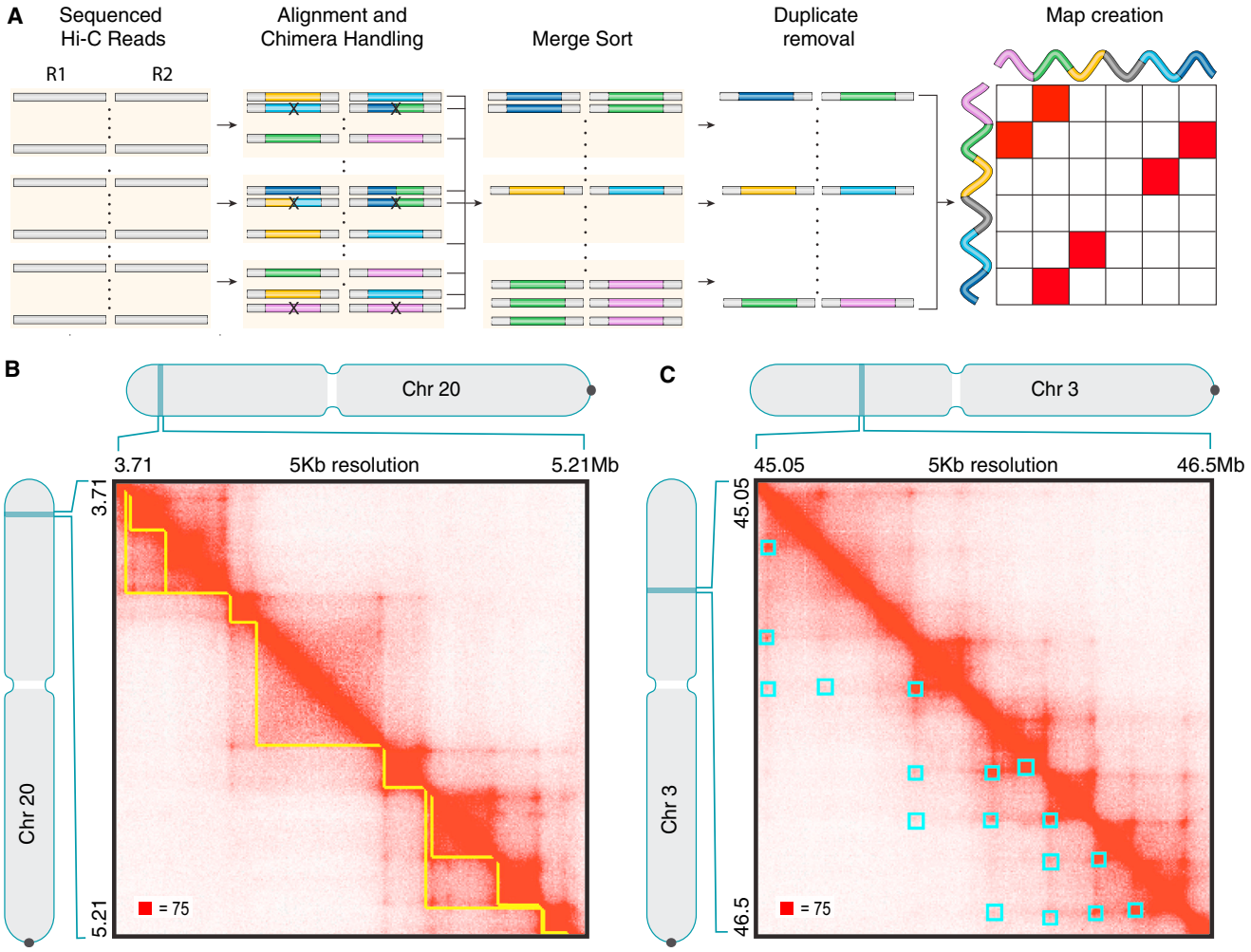


Figure 1. Juicer Analyzes Terabases of Hi-C Data with One Click

(A) Sequenced read pairs (horizontal bars) are aligned to the genome in parallel. Color indicates genomic position. Read pairs aligning to more than two positions are excluded. Those remaining are sorted by position and merged into a single list, at which point duplicate reads are removed. The *hic* file stores contact matrices at many resolutions and can be loaded into Juicebox for visualization. See Table S2.

(B) Contact domains (yellow) are annotated using the Arrowhead algorithm.

(C) Loops (cyan) are annotated using HiCCUPS.

represented as an 80 gigabyte *hic* file containing normalized and non-normalized contact matrices at 18 different resolutions, from 2.5 Mb resolution to single restriction fragment resolution for a 4-cutter restriction enzyme (~400 bp). Contact matrices in the *hic* format can also be visualized using Juicebox, which is described in the accompanying paper (Durand et al., 2016).

Finally, Juicer contains a suite of algorithms that are designed to annotate contact matrices and thus identify features of genome folding. These features include loops, loop anchor motifs, and contact domains.

Loops are identified using the HiCCUPS algorithm (Rao et al., 2014), which searches for clusters of contact matrix entries in which the frequency of contact is enriched relative to the local background. Because there are trillions of pixels in a kilobase-resolution Hi-C map, HiCCUPS is implemented using GPGPUs. Given CTCF and/or cohesin ChIP-seq tracks for the same cell type, HiCCUPS can frequently use FIMO (Grant et al., 2011) to identify the CTCF motif that serves as the anchor for each loop. We recently performed CRISPR experiments disrupting seven different CTCF motifs, each of which was identified by HiCCUPS as the anchor of one or more loops. In each case, disruption of the motif led to disruption of the corresponding loop, thus confirming the accuracy of HiCCUPS loop anchor annotations (Sanborn et al., 2015).

Contact domains are identified using a dynamic programming algorithm that relies on applying the Arrowhead transformation $[Ai,i + d = (M^*_{i,i-d} - M^*_{i,i+d})/(M^*_{i,i-d} + M^*_{i,i+d})]$ to a normalized contact matrix M^* (Rao et al., 2014). Many of these domains are associated with loops and can be disrupted by manipulating the corresponding loop anchors (Sanborn et al., 2015).

It is frequently useful to examine the cumulative signal from a large number of putative features at once, including both loops

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.
See also Table S1.

and domains. To this end, Juicer includes an implementation of Aggregate Peak Analysis (Rao et al., 2014).

Juicer is an open-source project. It is available at https://github.com/theaidenlab/juicer as a series of packages designed for a variety of hardware configurations: either a single machine, or clusters that run LSF, Univa Grid Engine, or SLURM. In addition, Juicer is available on the cloud at Amazon Web Services, and the test data used to review this paper is available at http://dx.doi.org/10.17632/c6bg4cbggn.1. Table 1 displays different performance metrics on each cluster system; the details of each setup are in the supplemental text. Once installed, Juicer can be executed using a single command by users without informatics experience.

EXPERIMENTAL PROCEDURES

All algorithms and data are drawn from Rao et al. (2014), except as described in the Supplemental Information.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2016.07.002.

AUTHOR CONTRIBUTIONS

E.L.A. conceived of this project. N.C.D. created the pipeline. S.S.P.R. created HiCCUPS. M.H.H. created APA. M.H.H. and N.C.D. created Arrowhead. M.S.S. re-implemented all feature annotation algorithms in Java as fully-automated, end-to-end tools. I.M. ported the pipeline to SLURM and AWS. N.C.D., M.S.S., I.M., and E.S.L. contributed to tool development. N.C.D. and E.L.A. prepared the manuscript.

ACKNOWLEDGMENTS

This work was supported by NIH New Innovator Award 1DP2OD008540, NIH 4D Nucleome grant U01HL130010, National Science Foundation (NSF) Physics Frontier Center PHY-1427654, National Human Genome Research Institute (NHGRI) HG006193, Welch Foundation Q-1866, Cancer Prevention Research Institute of Texas Scholar Award R1304, an NVIDIA Research Center Award, an IBM University Challenge Award, a Google Research Award, a McNair Medical Institute Scholar Award, and the President's Early Career Award in Science and Engineering to E.L.A., an NHGRI grant (HG003067) to E.S.L., and a PD Soros Fellowship to S.S.P.R. The Rice PowerOmics cluster was a gift from IBM. The software and test data sets used to review this manuscript are available at http://dx.doi.org/10.17632/c6bg4cbggn.1.

Received: December 2, 2015
Revised: May 2, 2016
Accepted: July 1, 2016
Published: July 27, 2016

REFERENCES

Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., and Aiden, E.L. (2016). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Systems *3*, this issue, 99–101.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018.

Knight, P.A., and Ruiz, D. (2012). A fast algorithm for matrix balancing. IMA J. Numer. Anal. *33*, 1029–1047.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

Sanborn, A.L., Rao, S.S.P., Huang, S.C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. USA *112*, E6456–E6465.

Sauria, M.E., Phillips-Cremins, J.E., Corces, V.G., and Taylor, J. (2015). HiFive: a tool suite for easy and efficient HiC and 5C data analysis. Genome Biol. *16*, 237.

Schmid, M.W., Grob, S., and Grossniklaus, U. (2015). HiCdat: a fast and easy-to-use Hi-C data analysis tool. BMC Bioinformatics *16*, 277.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. *16*, 259.