SYSTEMS BIOLOGY

# ChIA-PIPE: A fully automated pipeline for comprehensive ChIA-PET data analysis and visualization

**Byoungkoo Lee[1], Jiahui Wang[1], Liuyang Cai[1], Minji Kim[1], Sandeep Namburi[1], Harianto Tjong[1], Yuliang Feng[1], Ping Wang[1], Zhonghui Tang[1], Ahmed Abbas[1], Chia-Lin Wei[1,2]*, Yijun Ruan[1,2,3]*, Sheng Li[1,2,3,4]***



**ChIA-PET (chromatin interaction analysis with paired-end tags) enables genome-wide discovery of chromatin interactions involving specific protein factors, with base pair resolution. Interpretation of ChIA-PET data requires a robust analytic pipeline. Here, we introduce ChIA-PIPE, a fully automated pipeline for ChIA-PET data processing, quality assessment, visualization, and analysis. ChIA-PIPE performs linker filtering, read mapping, peak calling, and loop calling and automates quality control assessment for each dataset. To enable visualization, ChIA-PIPE generates input files for two-dimensional contact map viewing with Juicebox and HiGlass and provides a new dockerized visualization tool for high-resolution, browser-based exploration of peaks and loops. To enable structural interpretation, ChIA-PIPE calls chromatin contact domains, resolves allele-specific peaks and loops, and annotates enhancer-promoter loops. ChIA-PIPE also supports the analysis of other related chromatin-mapping data types.**

## INTRODUCTION

ChIA-PET (chromatin interaction analysis with paired-end tags) enables genome-wide discovery of chromatin interactions involving a specific protein of interest (*1*). The first version of the ChIA-PET protocol extracted short [2 × 21 base pairs (bp)] paired tags for two contacting chromatin sites (*2*). Subsequently, an improved version for longer-read (2 × 150 bp) ChIA-PET was developed, which showed increased mapping accuracy and robustness (*3*). ChIA-PET has been applied to study long-range chromatin interactions in human and mouse cells and has provided key insights into the roles of a number of chromatin architecture proteins and transcriptional factors, including CCCTC-binding factor (CTCF) and RNA polymerase II (RNAPII), in human three-dimensional (3D) genome organization (*4*–*8*). Briefly, long-read ChIA-PET is performed as follows. First, chromatin interactions are stabilized via dual cross-linking by formaldehyde and ethylene glycol bis(succinimidyl succinate) (EGS), and then, the chromatin sample is fragmented by sonication. Immunoprecipitation is performed to enrich for chromatin complexes containing a specific protein of interest. Pairs of DNA fragments in each chromatin complex are joined via bridge linker ligation (proximity ligation). The purified ligation products are then digested by Tn5 transposase (tagmentation) for DNA library construction and high-throughput PET sequencing (*3*).

By design, the sequencing data derived from each ChIA-PET experiment contain three sets of genomic information (*6*): (i) the genome-wide binding profile by the protein factor of interest in the study, analogous to chromatin immunoprecipitation (ChIP)–PET (*9*) and ChIP sequencing (ChIP-seq) data (*10*); (ii) the chromatin interactions between the binding sites involving the protein factor; and (iii) generic chromatin interactions that are not ChIP enriched, analogous to Hi-C data (*11*). Specifically, ChIP-enriched chromatin interactions can be distinguished on the basis of their higher frequency of contact between two given interacting loci as measured by overlapping PET counts over the nonenriched (nonspecific contacts as mostly "singletons") background.

Therefore, starting from the raw sequencing data of a ChIA-PET experiment, an effective computational pipeline must, as a baseline, (i) categorize read pairs for genuine bridge linker sequence between the two genomic tags, (ii) align tags to a reference genome to detect tag mapping locations and deduplicate the mapped tags, (iii) merge overlapping PETs to establish quantitation (PET counts) of potential chromatin contact ("looping") frequency between two chromatin interaction anchor loci, (iv) perform peak calling to identify binding peaks of the protein, (v) overlap protein-binding peaks with chromatin interaction anchors to identify specific protein–involving chromatin interactions, (vi) generate output of ChIA-PET data statistics and quality assessment (QA) metrics, and (vii) support 1D and 2D visualization.

Computational pipelines have previously been developed for short-read ChIA-PET data (*2*, *12*) and for relatively longer-read ChIA-PET data (*3*, *13*) that meet these baseline requirements. However, because of rapid advances in the field, current ChIA-PET computational pipelines for ChIA-PET lack the capacity and flexibility for effective, high-throughput processing of the large volumes of diverse data. Several features are needed. First, a single pipeline with the flexibility to process either short-read (2 × 21 bp) or long-read (2 × 150 bp) data is optimal. Second, with falling sequencing costs, it is now common to generate more than hundreds of millions PET reads per ChIA-PET library. An effective pipeline now requires greater robustness to process these massive datasets. Third, variants of the ChIA-PET method that enable the study of protein-associated chromatin interactions such as HiChIP (*14*) and proximity ligation–assisted ChIP-seq (PLAC-seq) (*15*) have recently been reported. Although HiChIP and PLAC-seq are similar to ChIA-PET, the

[1]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. [2]The Jackson Laboratory Cancer Center, Bar Harbor, ME, USA. [3]Department of Genetics and Genome Sciences, University of Connecticut Health Center, Farmington, CT, USA. [4]Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA.
*Corresponding author. Email: chia-lin.wei@jax.org (C.-L.W.); yijun.ruan@jax.org (Y.R.); sheng.li@jax.org (S.L.)

junction sequences between the two tags of a PET sequence are different among them. Therefore, an effective pipeline should be capable of processing both ChIA-PET, HiChIP, and PLAC-seq data. Fourth, performing peak calling on very large datasets without an input-control sample, as is standard in older pipelines, results in an abundance of false-positive peaks. An optimal pipeline would, by default, perform peak calling with an input-control sample and with stringent parameter settings.

Furthermore, expanded data visualization tools are needed. In the past year, there have been major advances in web-based tools for 2D visualization of chromatin interaction maps, e.g., Juicebox.js (*16*), HiGlass (*17*), WashU Epigenome Browser (*18*), and 3D Genome Browser (*19*). A pipeline that can generate the appropriate input visualization files (.hic file and .cool file) and that can be configured to render the files immediately visible to these tools upon completion of data processing would substantially advance research efforts. In addition, a browser-based visualization tool that provides a high-resolution, clear, and intuitive understanding of chromatin interaction intensity, genomic location, and genomic distance would substantially benefit both computational and experimental biologists.

In addition, new tools for structural interpretation are needed for conversion of large volumes of chromatin interaction data into knowledge. Enhanced tools are needed in three primary areas. (i) Because longer read lengths increase the coverage of heterozygous single-nucleotide polymorphisms (SNPs), an optimal pipeline would have the capacity to test for allele-specific chromatin interactions genome-wide. (ii) It has recently been demonstrated that CTCF-defined chromatin contact domains (CCDs) can be effectively called from ChIA-PET data (*6*). Furthermore, the boundaries of CCDs called from ChIA-PET data have been shown to correspond well to the boundaries of topologically associating domains called from Hi-C data (*6*). Thus, a pipeline that automatically calls CCDs in addition to loops and peaks would substantially expand the scope of biological knowledge gained from data. (iii) The major advantage of ChIA-PET is the ability to detect high-resolution chromatin interactions associated with general (RNAPII) and specific transcription factors. A pipeline that can detect enhancer-promoter ("E-P") interactions could facilitate functional characterization of distal regulatory elements.

We present our new analysis pipeline, ChIA-PIPE, which integrates multiple components to fill in the aforementioned gaps (Table 1). The pipeline is seamlessly integrated and runs from a single launch command, while also having the modularity to allow rerunning of data analysis at any step (Fig. 1 and fig. S1A). The pipeline is parallelized, open source, and can be applied to a ChIA-PET library simply by modifying a configuration (config) file. The pipeline also has a single-command installation script to enable an automated local installation of all dependencies. In addition, ChIA-PIPE generates

**Table 1. Comparison of ChIA-PIPE with existing software tools.**

| Category | Description | CPT | Mango | ChIA-PET2 | ChIA-PIPE |
|---|---|---|---|---|---|
| Experiment types | ChIA-PET | Y | Y | Y | Y |
| | HiChIP | | | Y | Y |
| | PLAC-seq | | | Y | Y |
| Data processing | Seamless execution | | | Y | Y |
| | Linker filtering | Y | Y | Y | Y |
| | Read mapping | Y | Y | Y | Y |
| | Peak calling with MACS2 | Y | Y | Y | Y |
| | Peak calling with SPP (accurate peak calls using input control) | | | | **Y** |
| | Loop calling | Y | Y | Y | Y |
| QA | QA table | Y | Y | Y | Y |
| | QA plots | | | Y | Y |
| Visualization tool compatibility | Juicebox 2D contact maps supported | | | | **Y** |
| | HiGlass 2D contact maps supported | | | | **Y** |
| | WashU Epigenome Browser supported | | | | **Y** |
| New visualization tool | BASIC Browser dockerized, released, and supported | | | | **Y** |
| Structural interpretation | Allele-specific peaks and loops | | | Y | Y |
| | CCD calling and 1D and 2D visualization | | | | **Y** |
| | E-P loop annotation | | | | **Y** |

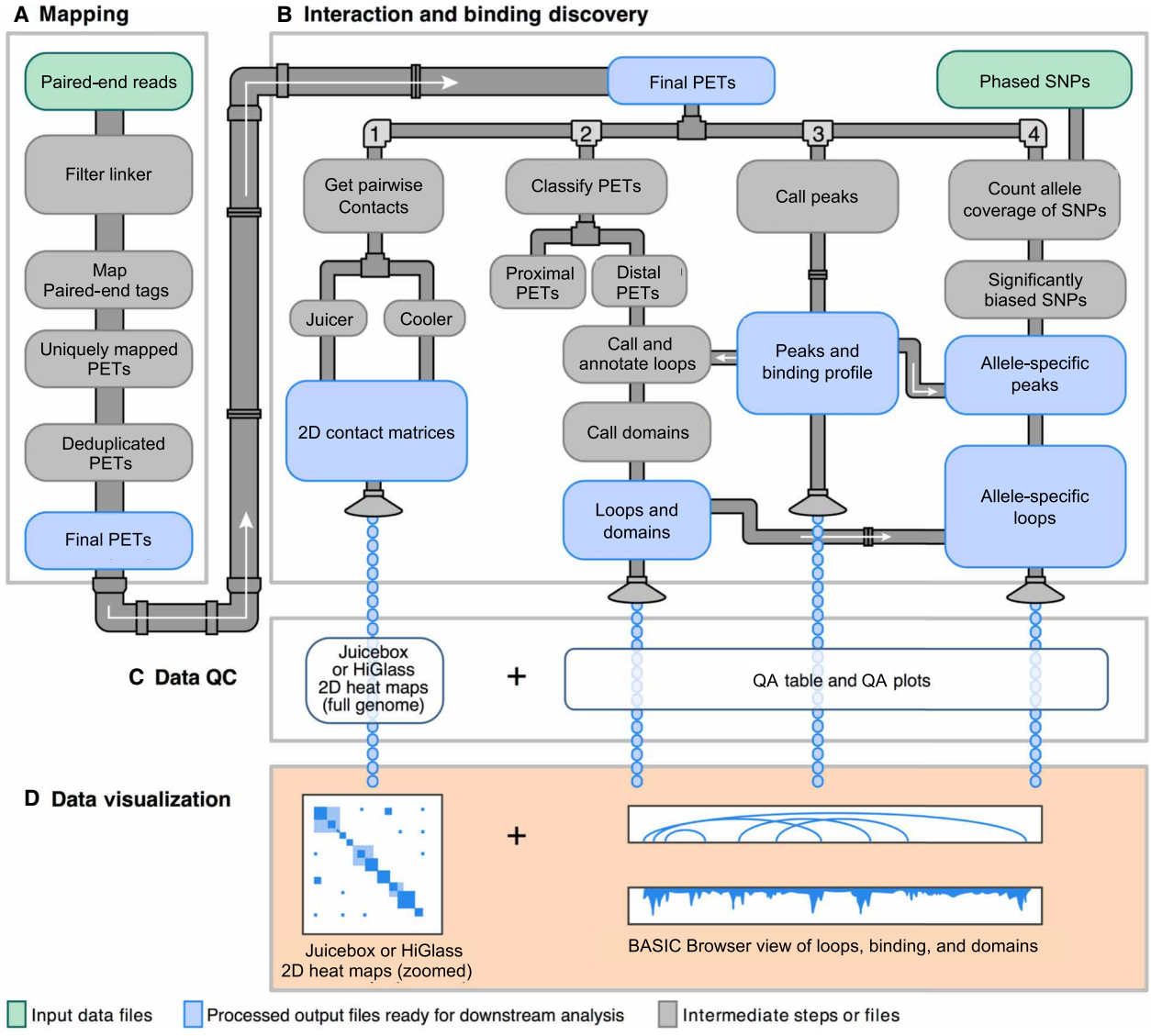



**Fig. 1. ChIA-PIPE architecture.** ChIA-PIPE takes a configuration file and two FASTQ files (R1 reads and R2 reads). (**A**) First, read pairs are scanned for the bridge-linker sequence and partitioned into categories: read pairs with (i) no linker, (ii) one linker and one usable genomic tag, or (iii) a linker and PETs. PETs are aligned to a reference genome. The analysis-ready BAM file contains uniquely mapped, nonredundant PETs. (**B**) 1: 2D chromatin interaction maps are generated using standardized file formats. 2: Using interligation PETs (span ≥8 kb), loops are called and then annotated with peak support. Peak-supported loops are then used to call CCDs. 3: Binding peaks of the protein of interest are identified using SPP (*22*) [or optionally, MACS2 (*23*)]. Before peak calling, there is a recovery step to incorporate all uniquely mapped, nonredundant tags. 4: If phased SNP information is available for the cell type of interest, allele-specific peaks and loops are identified. First, SNPs with significant bias in allele coverage are identified. Then, biased SNPs are used to annotate peaks and loop anchors. (**C**) The pipeline collates QA (quality assessment) metrics from every step into a succinct QA table and generates extensive QA plots. (**D**) The output files from the pipeline are compatible with interactive, high-resolution visualization tools. The 2D chromatin interaction maps can be viewed in Juicebox.js (*16*) or HiGlass (*17*). The loop file, peak file, coverage file, and domain file can be viewed in BASIC Browser.

output files for data visualization in 2D contact maps (Juicebox and HiGlass) and browser-based views. More specifically, we introduce a new loop browser, named BASIC Browser (Browser for Applications in Sequencing and Integrated Comparisons), for interactive, high-resolution visualization of loops, domains, and binding coverage. ChIA-PIPE also enables structural interpretation by implementing the CCD calling based on our prior work (*6*) and annotating E-P loops. ChIA-PIPE is currently used to process all ChIA-PET data for the Encyclopedia of DNA Elements (ENCODE) (*20*) and 4D Nucleome (4DN) consortia (*21*).

## RESULTS

### ChIA-PIPE data processing and analysis

ChIA-PIPE enables seamless execution of high-throughput chromatin interaction data analysis through the combined functions of its four modules (Fig. 1): (i) mapping, (ii) chromatin interactions and protein-binding discovery, (iii) data QA, and (iv) data visualization. For massive datasets with hundreds of millions or billions of sequencing reads, such as the more than 100 ChIA-PET libraries generated by the ENCODE and 4DN consortia, almost every step of the ChIA-PIPE pipeline supports multithreading. In addition, the

config file allows users to customize the number of threads and the memory. The input for ChIA-PIPE is a configuration file and two FASTQ files (one file for R1 reads and another file for R2 reads). The first step (Fig. 1A) is to scan the read pairs for the linker sequence from the proximity ligation step of ChIA-PET and to partition the read pairs into categories: read pairs with (i) no linker sequence, (ii) a linker sequence and one usable genomic tag, or (iii) a linker sequence and PETs. ChIA-PIPE also supports the processing of HiChIP (*14*) and PLAC-seq (*15*) by treating the repaired and ligated restriction site from HiChIP and PLAC-seq as a "pseudolinker" in this first step (see fig. S1B and Methods).

Next, the read pairs in each category are separately aligned to a reference genome (see Methods), and only uniquely mapped and nonredundant tags are retained. For each category, the BAM file of PETs is the final output file that is used for all downstream analyses of chromatin interactions (Fig. 1A). Once the BAM file of analysis-ready tags is generated, ChIA-PIPE performs interaction and binding discovery in several workflows. First, the BAM is converted into a standardized file format for interaction pairs, which is then converted into 2D contact map files for visualization of chromatin interaction heat maps using Juicebox.js (*16*) or HiGlass (*17*) (Fig. 1B.1).

ChIA-PIPE also identifies genomic binding peaks of the protein factor (Fig. 1B.3). The peaks called from ChIA-PET data and ChIP-seq data were shown to be comparable previously (*6*). The peak calling step (Fig. 1B.3) uses a merge of all uniquely mapped and nonredundant tags, as even tags that were noninteractions are informative for protein-binding peaks detection. By default, this peak calling is performed using SPP (ChIP-seq processing pipeline) (*22*), which identifies genomic regions of significantly enriched tag density compared with a standard ChIP-seq input-control sample. In practice, this

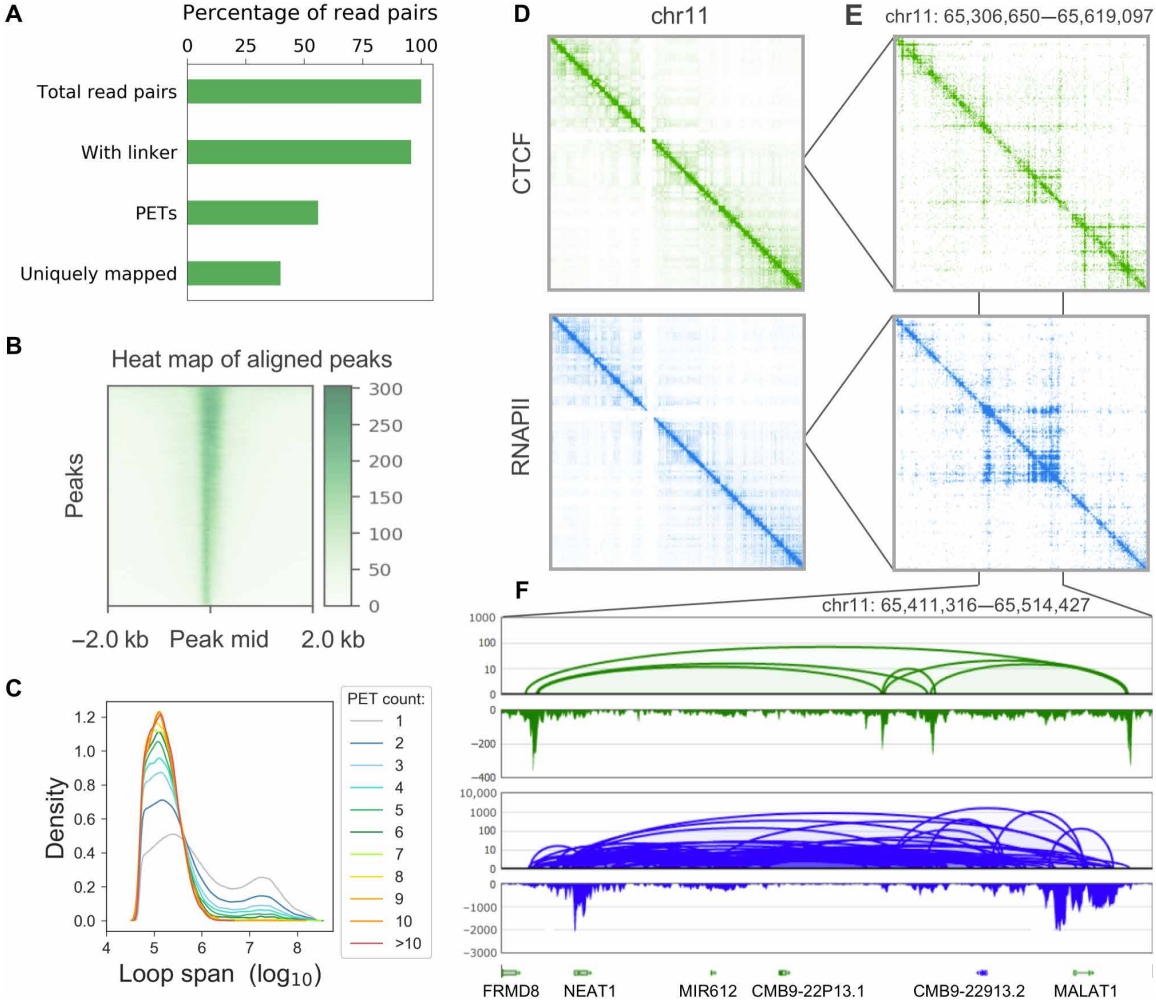



**Fig. 2. ChIA-PIPE QA.** ChIA-PIPE supports QA and visualization for each sequencing library that is processed. (**A**) The percentage bar chart of read pairs during the mapping step in ChIA-PIPE using a HiSeq CTCF ChIA-PET library of HFFc6 cells. The pipeline also generates a QA table with detailed summary statistics (fig. S2B). (**B**) A peak intensity tornado heat map of the read pileup over all peaks (each row) aligned by their midpoints (shown: CTCF in HFFc6). The sharpness of the peak intensity indicates the quality of the binding enrichment. (**C**) The distributions of loop span depending on the PET count of the loops (shown: CTCF in HFFc6). (**D** and **E**) Juicebox.js can be used for QA by visualizing the 2D contact map for full chromosomes or broad chromosomal segments. The enrichment of signal along the diagonal indicates the quality of the library. As examples, we show both CTCF and RNAPII ChIA-PET data in a whole chromosome 11 and a zoomed-in genomic region (chr11: 65,306,650–65,619,097). (**F**) BASIC Browser provides detailed and high-resolution visualization of further zoomed-in chromosomal region (chr11: 65,411,316–65,514,427). BASIC Browser is also used for QA by visualizing a known biological interaction (such as the interaction between NEAT1 and MALAT1) to confirm looping and binding enrichment (shown: CTCF and RNAPII in HFFc6).

approach has exhibited high specificity even with large datasets (fig. S2A) and has become a standard approach in the ENCODE consortium. ChIA-PIPE also has the flexibility for peak calling using MACS2 (*23*) with or without a ChIP-seq input-control sample.

ChIA-PIPE calls and annotates loops. Using "interligation" PETs (ligation between two chromatin fragments as defined by the two tags mapped ≥ 8 kb apart), each mapped tag is extended by 500 bp in its 5′ direction, and the overlapping PETs are merged into "loops," with the frequency of the loop measured by the number of PETs contributing to the loop (see Methods). Each loop is then annotated with the number of its anchors (0, 1, or 2) that are supported by a binding peak. For CTCF ChIA-PET data, CCDs are called using loops with CTCF binding peak support to both anchors (see Methods). If phased genome data and heterozygous SNPs are available for the

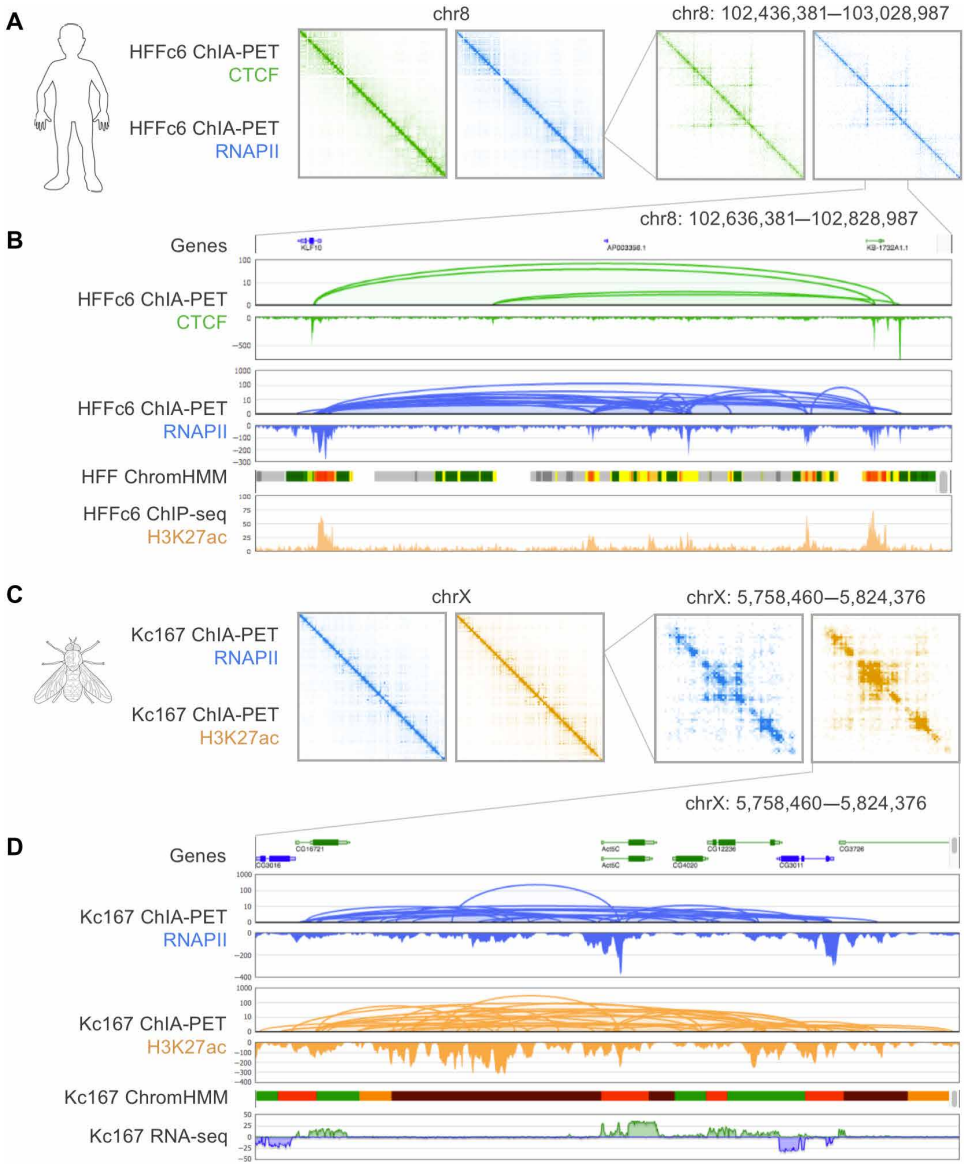



**Fig. 3. ChIA-PIPE visualization.** ChIA-PIPE includes a Docker image of BASIC Browser for interactive, high-resolution ChIA-PET data visualization, also supports 2D contact map visualization using Juicebox and HiGlass. (**A**) Juicebox shows the contact maps of CTCF (green) and RNAPII (blue) ChIA-PET data for the entire chromosome 8 and a specific zoomed-in region in chromosome 8 from HFFc6 cells. (**B**) BASIC Browser shows CTCF loops and peaks (green) as well as RNAPII loops and peaks (blue) from HFFc6 cells. In addition, BASIC Browser supports visualization of other data tracks to facilitate ChIA-PET interpretation. For example, at the top, UCSC Known Genes are shown; and at the bottom, ChromHMM in the same cell type (from the ENCODE portal) and H3K27ac ChIP-seq (orange) in the same cell type (from the 4DN portal) are shown. For the ChromHMM track, red indicates active transcription start sites (TSSs) and yellow and orange indicate enhancers. Thus, BASIC Browser provides a comprehensive genomic view of this region, revealing that a CTCF-mediated loop is encompassing two active genes (KLF10 and KB-1732A1.1) that are interacting with each other and with a set of enhancers between them. (**C**) Juicebox shows the contact maps of RNAPII (blue) and H3K27ac (orange) ChIA-PET for the entire chromosome X and a specific zoomed-in region in chromosome X from *Drosophila* Kc167 cells. (**D**) A more detailed view of the data using BASIC Browser, also displaying UCSC Known Genes, the ChromHMM track, and RNA-seq from Kc167 cells. BASIC Browser also supports strand-specific RNA-seq data visualization to separate genes from sense (green) and antisense strands (blue).

appropriate cell sample, ChIA-PIPE can also identify haplotype-specific peaks and loops with significant allelic bias.

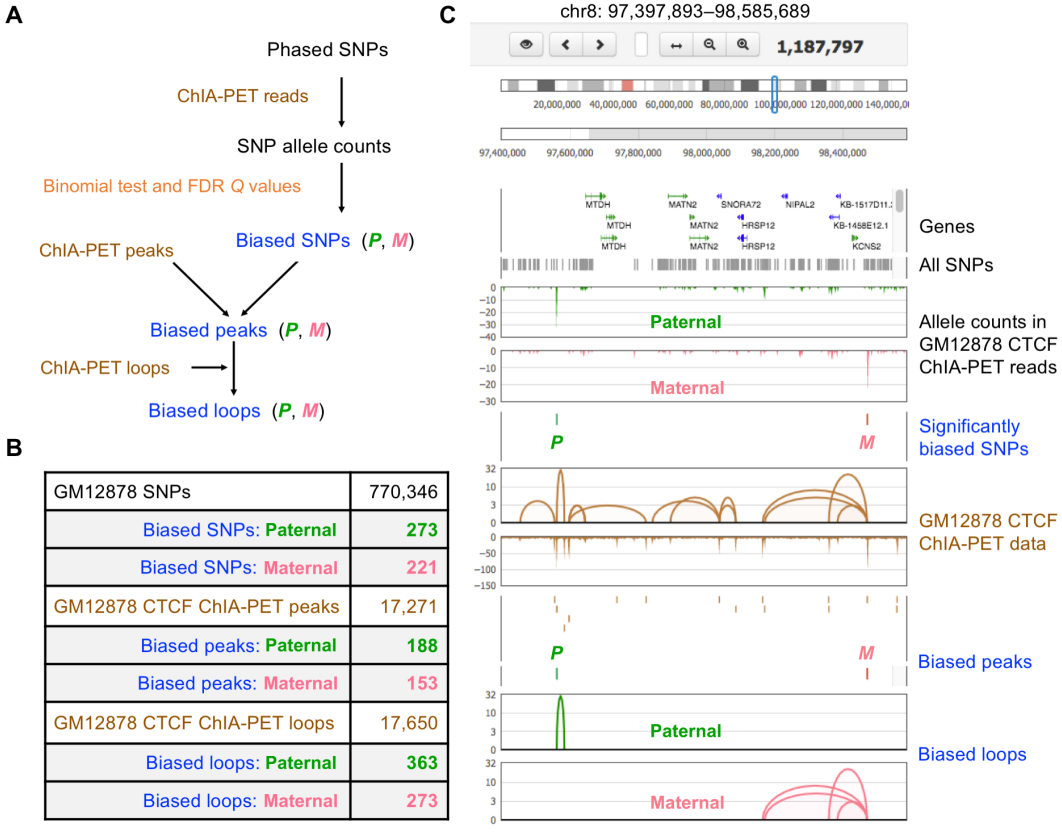
### ChIA-PIPE data QA and visualization

To enable rapid evaluation of library quality, the pipeline reports key data statistics and QA metrics in a CSV file (Fig. 1C and fig. S2B), which can be viewed as an Excel spreadsheet. For example, the majority of the read pairs should contain the linker sequence, from which the majority should be PETs (rather than one-tag read pairs). Further, ChIA-PIPE supports QA visualization for each library. Library-level QA visualizations reveal valuable information about the percentage of read pairs passing each processing step (Fig. 2A), the quality of the binding enrichment (Fig. 2B), and the specificity of the chromatin interactions (Fig. 2C). The percentage bar plot of read pairs passing each processing step is critical for assessing the mapping quality of the library.

ChIA-PIPE automatically generates 2D contact map files for both browser-based tools: a .hic file for Juicebox.js (*16*) and a .cool file for HiGlass (*17*). The 2D contact maps provide informative views of full genomes, each chromosome, or specific chromosomal regions (Fig. 2, D and E; fig. S2, C and D). BASIC Browser provides high-resolution views of smaller chromosomal regions (Fig. 2F).

Once a library has been processed and passes QA, chromatin interactions can be examined by high-resolution, interactive visualization of ChIA-PET data in browsers. The pipeline automatically outputs the appropriate file formats for relevant 2D contact and browser-based visualization tools for ChIA-PET data (Fig. 1D). For browser-based visualization, for example, the CTCF ChIA-PET data, the key tracks of the data are the loop clusters (a BEDPE file), the CCDs (a BED file), and the binding-intensity profile of the CTCF protein factor (a bedgraph file). The binding-intensity track can be readily viewed in two publicly available genome browsers: the University of California Santa Cruz (UCSC) Genome Browser (*24*) and the WashU Epigenome Browser (*18*). However, only the latter can readily display chromatin interaction loops, and even then, the display is not highly intuitive because the loop height (*y* axis) does not scale with the PET count.

To allow improved visualization of chromatin interaction loops, ChIA-PIPE includes its own loop browser—"BASIC Browser"—for interactive, high-resolution ChIA-PET data visualization (Fig. 3). In BASIC Browser, the chromatin interaction tracks display the chromatin loops for their contact frequency (*y* axis, the height of a loop reflects the intensity of connectivity) and interacting anchor distances, along with the protein-binding intensity tracks, which are intuitive and easily interpretable. The BASIC Browser tracks are publication quality ready and can reveal previously unknown biological findings (Fig. 3 and fig. S3). Furthermore, it also supports




| | |
|---|---|
| GM12878 SNPs | 770,346 |
| Biased SNPs: Paternal | 273 |
| Biased SNPs: Maternal | 221 |
| GM12878 CTCF ChIA-PET peaks | 17,271 |
| Biased peaks: Paternal | 188 |
| Biased peaks: Maternal | 153 |
| GM12878 CTCF ChIA-PET loops | 17,650 |
| Biased loops: Paternal | 363 |
| Biased loops: Maternal | 273 |

**Fig. 4. ChIA-PIPE processes allele-specific chromatin interactions.** ChIA-PIPE can resolve allele-specific peaks and loops if phased SNP information is available for the cell type of interest. (**A**) Allele-specific peak and loop calling diagram. First, the allele counts in the ChIA-PET reads are determined for each phased SNP. Then, SNPs with significantly biased allele counts are determined using the binomial test and Benjamini-Hochberg multiple hypothesis testing correction. ChIA-PET peaks that overlap significantly biased SNPs are considered biased peaks. ChIA-PET loops that overlap biased peaks are considered biased loops. (**B**) Processed paternal and maternal counts of biased SNPs, peaks, and loops from GM12878 CTCF ChIA-PET MiSeq data. (**C**) BASIC Browser view of a genomic region with one paternally biased SNP and one maternally biased SNP, and the corresponding biased ChIA-PET peaks and loops.

visualization of gene expression data for functional interpretation of chromatin interactions, including strand-specific gene expression. For easy use, BASIC Browser is available as a Docker image.

### ChIA-PIPE supports downstream analysis for structural interpretation

In addition to greatly improved support for popular visualization tools and introducing a new visualization tool, ChIA-PIPE incorporates several unique data analysis functions that provide substantially increased capacity for structural interpretation of the data, compared to the capacities of the existing published alternative pipelines (Table 1). Specifically, ChIA-PIPE resolves allele-specific peaks and loops, calls and visualizes CCDs, and annotates E-P loops.

ChIA-PIPE has a robust workflow for resolving allele-specific peaks and loops when phased SNP data are available for the ChIA-PET data (Fig. 4A). First, heterozygous SNPs with significant allelic bias in the coverage of ChIA-PET reads are identified on the basis of the results of the binomial test and multiple hypothesis testing

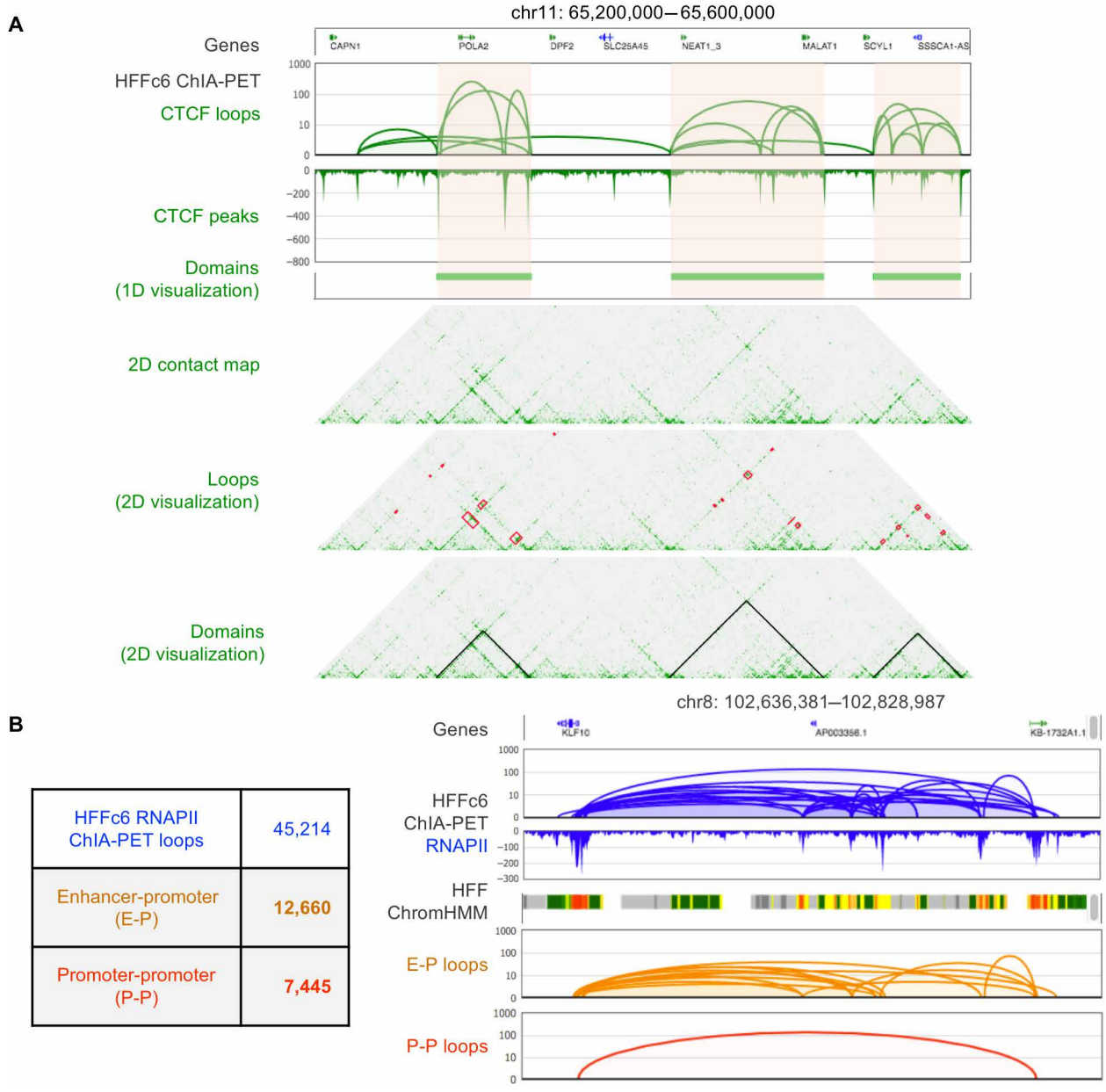


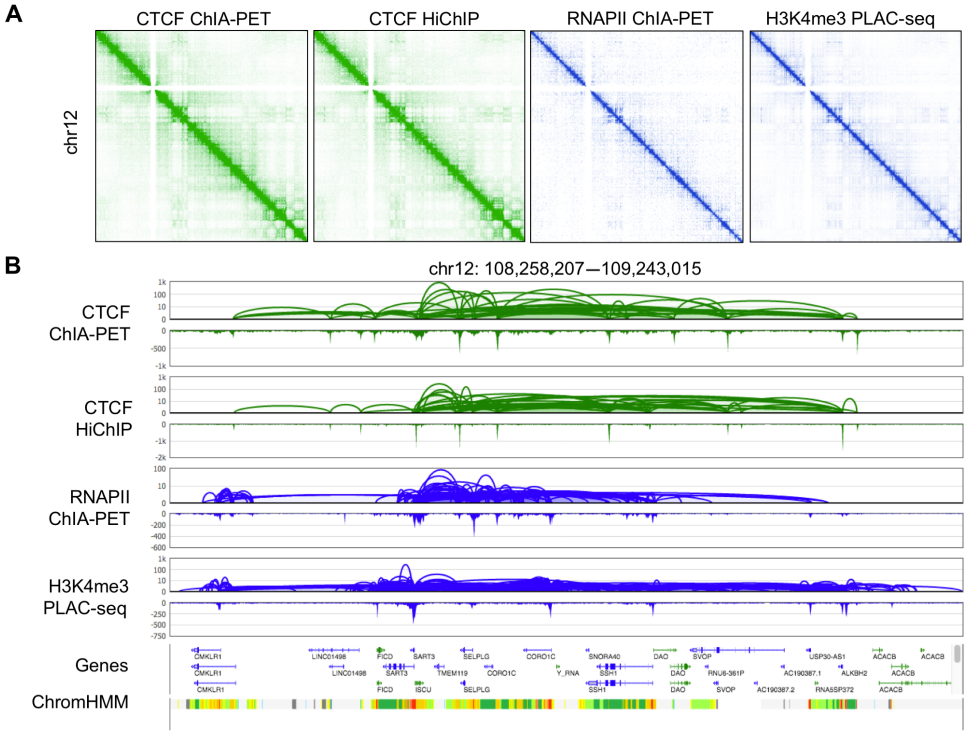

**Fig. 5. ChIA-PIPE processes CCDs and annotates E-P loops.** (**A**) ChIA-PIPE automatically calls CCDs from CTCF ChIA-PET data. The CCDs can be visualized in BASIC Browser together with loops and binding coverage. In addition, ChIA-PIPE automatically generates the input file to visualize loops and CCDs as a 2D annotation in Juicebox contact maps. Red rectangular boxes atop the contact map represent 2D annotation of loops, as corresponding two loop anchors of each loop. Black triangles represent 2D annotation of CCDs. (**B**) ChIA-PIPE automatically annotates enhancer-promoter (E-P) loops and promoter-promoter (P-P) loops from RNAPII ChIA-PET data. The E-P loops and P-P loops can be visualized in BASIC Browser together with the chromatin state (ChromHMM) track. In the ChromHMM track, active promoters are indicated in red and enhancers are indicated in orange and yellow.

correction by the Benjamini-Hochberg procedure (see Methods). Second, peaks that overlap significantly biased SNPs are identified (fig. S4). Third, loop anchors overlapping significantly biased peaks are resolved (see Methods). The ChIA-PIPE report includes paternally biased and maternally biased SNPs, peaks, and loops (Fig. 4B and fig. S4).

ChIA-PIPE automatically calls CCDs from CTCF ChIA-PET datasets. In addition, ChIA-PIPE generates the appropriate output files to support 1D visualization of CCDs in BASIC Browser or 2D visualization of loops and CCDs superimposed atop Juicebox 2D contact maps (Fig. 5A). ChIA-PIPE also automatically annotates E-P loops and promoter-promoter ("P-P") loops from RNAPII ChIA-PET datasets. ChIA-PIPE also generates subset loop files for the different annotations, which enables visualization of each category of loop in a separate track in BASIC Browser. For example, visualization of E-P loops and P-P loops together with gene annotations and ChromHMM chromatin states can be very informative (Fig. 5B), and linking enhancers with their target genes is a highly biologically relevant area that is currently of great interest in the community.

The ENCODE4 and 4DN consortia will generate hundreds of ChIA-PET libraries over the next few years; each library will contain 200 to 400 million next-generation sequencing read pairs. ChIA-PIPE is designed to robustly process the massive datasets by parallelizing with multiple threads when the appropriate hardware is available. For example, ChIA-PIPE processed an HFFc6 CTCF HiSeq library (~300 million read pairs) and detected peaks and chromatin interactions with a run time (wall) of 10 hours (threads, 20; RAM, 60 Gb, cluster operating system, CentOS 6.5; central processing unit, Intel Xeon E5-2670 @ 2.60 GHz).

Last, ChIA-PIPE is also capable of processing data from related 3D genome mapping methods, including HiChIP and PLAC-seq (Fig. 6 and fig. S5). Two-dimensional heat map visualization and BASIC Browser view show comparable chromatin interaction detected between ChIA-PET, HiChIP, and PLAC-seq detected by ChIA-PIPE. In addition, ChIA-PET loops and domains, identified




C

| GM12878 | CTCF ChIA-PET | CTCF HiChIP | RNAPII ChIA-PET | H3K4me3 PLAC-seq |
|---|---|---|---|---|
| Read pairs | 358,752,218 | 274,443,458 | 132,169,141 | 301,398,362 |
| Uniquely mapped, nonredundant PETs | 106,110,023 | 59,061,469 | 14,780,667 | 28,884,557 |
| Peaks | 28,992 | 23,183 | 51,703 | 25,517 |
| Loops | 1,181,427 | 457,572 | 195,862 | 528,182 |
| Loops with peak support (both anchors) | 57,120 | 33,835 | 113,448 | 417,515 |

**Fig. 6. ChIA-PIPE can process other 3D genome mapping data, including HiChIP and PLAC-seq.** ChIA-PIPE was used to process various comparable datasets, generated from different methods: CTCF ChIA-PET, CTCF HiChIP, RNAPII ChIA-PET, and H3K4me3 PLAC-seq for GM12878 cells. ChIA-PIPE was readily adapted to process HiChIP and PLAC-seq data by treating the ligated restriction enzyme site as a pseudolinker. (**A**) 2D contact maps of the four datasets for the entire chromosome 12 using Juicebox. (**B**) BASIC Browser views of loops and binding coverage from the four datasets aforementioned, alongside gene annotations, and ChromHMM for GM12878 (from the ENCODE portal). (**C**) Quantitative comparison of loops and peak counts of CTCF ChIA-PET, CTCF HiChIP, RNAPII ChIA-PET, and H3K4me3 PLAC-seq datasets processed by ChIA-PIPE.

from ChIA-PIPE, showed comparable loops and domains from Hi-C (fig. S6).

## DISCUSSION

We show that ChIA-PIPE enables seamless execution of comprehensive chromatin interaction data processing, analysis, and visualization. ChIA-PIPE provides unique key functions not available in prior tools, e.g., CPT (ChIA-PET tool), Mango (*12*), and ChIA-PET2 (*13*), shown in Table 1. First, the ChIA-PIPE binding peak–calling module uses input control and optimized parameters to provide high sensitivity and specificity (fig. S2A). Second, ChIA-PIPE provides comprehensive support for 2D contact map visualization tools including HiGlass and Juicebox, which are not supported by any prior ChIA-PET analysis tools. Third, ChIA-PIPE provides a new genome browser, BASIC Browser, built specifically for high-resolution, browser-based exploration of peaks and loops, and strand-specific RNA sequencing (RNA-seq) visualization. BASIC Browser is available as a docker image, with an easy desktop installation. Fourth, ChIA-PIPE uniquely enables useful downstream structural interpretation analysis. These latter functionalities include calling and visualization (1D and 2D) of CCDs, annotation of E-P interactions, and calling of haplotype-specific loops and peaks. There is a critical need for software to effectively identify E-P interactions in gene regulation studies, as the long distance that can separate enhancers and promoters hinders accurate inference of the targets of distal regulatory elements. ChIA-PIPE provides the tools for detection, quantification, and visualization of E-P interactions.

Furthermore, it is critical to assess the quality of the ChIA-PET library before interpreting the biological signals. Therefore, ChIA-PIPE offers the flexibility of multiple alternative supported platforms and comprehensive QAs including (i) key data statistics: linker trimming, read alignment, redundancy, binding peaks, PETs, intra-inter PET ratio, intra-inter loop ratio, and loop numbers at different PET count; (ii) automatic library-level QAs: the percentage of read pairs passing each processing step, the quality of protein factor binding enrichment, and the specificity of chromatin interactions.

Last, ChIA-PIPE is now the production pipeline for ChIA-PET data analysis in the ENCODE and 4DN consortia. We anticipate that ChIA-PIPE will be a valuable resource for the broader research community.

In conclusion, ChIA-PIPE has achieved multiple milestones: (i) full automation as a single-launch command-based pipeline; (ii) robustness to massive datasets; (iii) ChIA-PET sequencing read length flexibility; (iv) accurate peak calling achieved by SPP with an input control sample; (v) automated CCD calling, 1D and 2D visualization; (vi) web-based visualization automation for BASIC Browser, Juicebox.js, and HiGlass; and (vii) adaptability to process data from related protocols (e.g., HiChIP and PLAC-seq). With these achievements, ChIA-PIPE has become a valuable resource used by the ENCODE4 and 4DN consortia and by the broader biological research community.

## METHODS

### Linker filtering

ChIA-PET data from raw compressed FASTQ files—either long-read (fig. S7) or short-read data—that contain bridge linker sequence are processed with ChIA-PET Utilities (CPU; https://github.com/cheehongsg/CPU), a collection of modularized executables developed by our team for performing core ChIA-PET data processing tasks. Paired-end reads with a bridge linker were identified, and the tags flanking the linker were extracted. ChIA-PIPE integrates CPU modules and modules from many other packages into a comprehensive pipeline (fig. S1). Although HiChIP data do not contain a linker sequence, all read pairs with interactions contain a repaired and ligated restriction site (GATCGATC), which can be treated as a pseudolinker in the pipeline.

### Tag alignment

Briefly, PET read sequences are scanned for the bridge linker sequence, and only PETs with the bridge linker are retained for downstream processing. After trimming the linkers, the flanking sequences are mapped to the reference genome of target cells using a hybrid of BWA-ALN (Burrows-Wheeler aligner) and BWA-MEM (Burrows-Wheeler aligner - maximal exact matches) (*25*), which are integrated in CPU module "memaln." The unique alignments [MAPQ (mapping quality) $\geq$ 30] and nonredundant PETs are retained by calling the CPU "dedup" module. The BAM file of PET reads is now ready for further analysis (Fig. 1A).

### Interaction and binding discovery

The BAM file (.bam) of PET read alignment is further processed in four parallel subpipelines for identification of chromatin interactions: (i) 2D contact maps, (ii) loops, (iii) binding peaks, and (iv) haplotype-specific interactions (see Methods).

1. 2D contact maps (Fig. 1B.1). The BAM file of all analysis-ready PETs is processed to generate a 2D contact matrix file (.hic) using Juicer tools (*26*). We generated the contact matrix file for 10 different resolutions (2.5 Mb, 1 Mb, 500 kb, 250 kb, 100 kb, 50 kb, 25 kb, 10 kb, 5 kb, and 1 kb), enabling users to easily adjust multiscale genomic regions from an "all-chromosome" view to view of a specific genomic region, up to 1-kb resolution. Juicebox, described in the data visualization section (see below), is used to access the 2D contact map (*27*). On the basis of the request by 4DN Data Coordination and Integration Center (DCIC), the BAM file of all analysis-ready PETs will also be processed to generate a contact list file (.pairs) that is compatible with the 4DN DCIC's proposed pipeline for Hi-C data analysis.

2. Interaction loops (Fig. 1B.2). This step of the pipeline uses the CPU module termed "cluster." The BAM file of PET reads includes three categories of PET reads: (i) PET reads with no linker sequence detected, (ii) PET reads with a linker sequence detected but with only one usable genomic tag, and (iii) PET reads with a linker sequence detected with both ends having genomic tags. Only PET reads in (iii) are used for detection of interaction loops. Each PET in (iii) is categorized as either a self-ligation PET (two ends of the same DNA fragment) or an interligation PET (two ends from two different DNA fragments in the same chromatin complex) by evaluating the genomic span between the two ends of a PET. PETs with a genomic span of less than or equal to 8 kb are classified as self-ligation PETs and are included as a proxy for ChIP fragments, since they are derived in a manner analogous to derivation of peaks used in ChIP-seq mapping for protein-binding sites. PETs with a genomic span of greater than 8 kb are classified as interligation PETs and represent the long-range interactions of interest. Immunoprecipitation in ChIA-PET protocol is executed at the protein-binding position of each fragmented chromatin complex. Then, the loose DNA ends are ligated and captured for sequencing. Therefore, the 5′ end of each

interligation PET is extended by 500 bp along the reference genome to be more representative of the interacting chromatin fragments (*28*). To reflect the frequency of interaction between two loci, the extended interligation PETs that directly overlap are clustered together as one PET cluster. The PET counts in a PET cluster reflect the relative frequency of interaction between two genomic regions. We observed that many anchors of distinct PET clusters are located within the binding peak region of the same protein factor. It is clear that these binding peak regions represent the real chromatin interaction loci in the nucleus. To streamline the data structure of the PET clusters, we collapsed the individual anchors of all PET clusters with 500-bp extensions to generate merged anchors. For anchors with overlapped binding peaks, we use the summit points as the centers of interacting loci. We refer to the merged PET clusters as chromatin interaction loops. Unclustered individual interligation PETs are referred as PET singletons.

3. Protein factor binding peaks (Fig. 1B.3). All uniquely mapped and nonredundant analysis-ready tags including self-ligation and interligation PETs are used for identifying protein factor binding peaks. In addition, two categories of PET tags that are excluded from chromatin loop detection are recovered and included in peak calling for protein factor binding. Specifically, the recovered read categories are (i) PET reads with no linker sequence detected and (ii) PET reads with a linker sequence detected but with only one usable genomic tag. While these reads are uninformative for chromatin loop detection, they are informative for peak calling of protein factor binding. Peak calling for protein factor binding is then performed using the SPP pipeline (version 1.13) (*22*) with parameters srange = c(200, 5000), bin = 20, window.size = 500, and z.thr = 6. Optionally, the MACS2 pipeline (version 2.1.0) (*23*) can be used for protein factor binding peak identification with default parameters. In addition, bedtools (*29*) is used to generate BedGraph files of the protein factor binding coverage along the chromosomes for browser-based visualization.

## ChIA-PET QA visualizations

For each ChIA-PET library, we first perform a QA test sequencing to generate 5 million to 10 million PET reads from a MiSeq run, which can usually test two to four ChIA-PET libraries. Once a quality library is verified by the test sequencing data, we then generate ~200 million PET reads from a HiSeq "production" run. The test and production sequencing data are both processed using ChIA-PIPE, detailed below. Now, the pipeline provides library-level quality examination, i.e., sequencing alignment quality, peak intensity tornado plot [generated by deepTools2 (*30*)], loop spanning distribution, and loop-level anchor support distribution. The contact map file generated by the pipeline is ready to be examined by Juicebox (.hic) or HiGlass (.cool).

## Interactive ChIA-PET data visualization

### *Interactive visualization of ChIA-PET 2D contact maps using Juicebox.js and HiGlass*

ChIA-PIPE uses Juicebox.js and HiGlass to visualize ChIA-PET data in 2D contact maps. After the duplications are removed and the uniquely mapped PETs are retained, the BAM file can be converted to a merged_nodups.txt file and used as input to the Juicer tool Pre command, creating a .hic file. Loops called by ChIA-PIPE can also be visualized in Juicebox by representing them in the appropriate text file format, which is analyzed using the data processing tool termed Juicer (*26*). Juicer generates a contact matrix file (.hic), which is in a highly compressed binary file format and can be accessed by Juicebox (*27*) visualization software. These tools were initially developed to visualize Hi-C data, and we adopted them to process and visualize our ChIA-PET 2D contact maps. ChIA-PIPE uses Juicer to generate the contact matrix file for 10 different resolutions (2.5 Mb, 1 Mb, 500 kb, 250 kb, 100 kb, 50 kb, 25 kb, 10 kb, 5 kb, and 1 kb), and thus, users can easily adjust multiscale genomic regions from an all-chromosomes view to a view of a specific genomic region up to 1-kb resolution. The chromatin contact matrix is based on the adjustable resolution, from a bin size of 2.5 Mb × 2.5 Mb to 1 kb × 1 kb. The contact matrix contains the total contact count within the specific genomic bin. Contact signal intensity or counts are represented by the intensity of red color in default, from low contact (light red) to high contact (dark red). The default color can be changed by user. Four different normalization methods (none, coverage, coverage_sqrt, and balanced) can be applied to the contact heat map data to remove technical noise or adjust biased signal.

### *BASIC Browser for interactive, high-resolution visualization of ChIA-PET loops, domains, and binding coverage*

ChIA-PIPE uses an in-house–developed genome browser to visualize the binding peaks and chromatin loops (Fig. 3B). The chromatin interaction data file (loop.gz; Fig. 1B.2) is uploaded to the BASIC Browser and displays chromatin contacts as arcs between the two genomic loci for each loop. The length of an arc indicates the linear genomic distance of a loop, and the height of an arc reflects the contact frequency (PET counts) detected in the ChIA-PET data. BASIC Browser enables the visualization of the binding peaks and demarcates the summit of binding sites from the protein-binding profile data (.bedgraph) (Fig. 1B.3). In addition, allele-specific files for binding peaks and chromatin interactions (Fig. 1B.4) are also visualized. The advantages of using browser-based visualization includes detailed bp resolution presentation of chromatin loops, in relation to protein-binding peaks, and the capacity to integrate these data with other genome browser information on other experimental data (e.g., RNA-seq, ChIP-seq, and ATAC-seq).

## ChIA-PIPE downstream analysis

### *Allele-specific peaks and loops*

There are several steps involved in determining haplotype-specific chromatin interactions if the phased genome sequencing information is available for the same cell of ChIA-PET (Figs. 1B.4 and 4A). As an example, we used the GM12878 cell line, which was derived from the 1000 Genome Project human subject NA12878, whose genome-wide SNP phasing information is available (ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg19/). By using this SNP phasing information, uniquely aligned (MAPQ ≥ 30) CTCF and RNAPII ChIA-PET reads were given a haplotype assignment depending on whether they overlapped a phased SNP and, if so, which allele they corresponded to. The possible haplotype assignments are maternal (M), paternal (P), or not determined (N). This analysis was done independently for the two ends of all interligation PETs. Therefore, a PET could have the following possible paired haplotypes at the two ends of intrachromosomal loops: M-M, P-P, M-N, P-N, M-P, or N-N. The allele counts were determined using samtools pileup (*31*).

ChIA-PIPE determines the haplotype specificity for the anchors of each chromatin interactions as follows: (i) Phased SNPs with

biased protein factor binding coverage were identified. The maternal and paternal allele counts of individual phased SNPs were computed and tested for allele bias using a binomial test. SNPs with Benjamini-Hochberg–adjusted *P* values [i.e., FDR (false discovery rate) *Q* values] ≤0.1 (*32*) were considered to have significantly biased protein factor binding coverage. (ii) The interaction anchors that overlapped with biased SNPs were then assigned a haplotype corresponding to the bias direction of the SNP. If an anchor overlapped with multiple biased SNPs and the bias directions of these SNPs were consistent, then the haplotype assignment of this anchor was given accordingly; otherwise, the haplotype of this anchor was considered to be not determined. In addition, if an anchor overlapped with multiple phased SNPs and the SNP with the highest binding coverage showed no allelic bias, then the haplotype of such an anchor was also considered to be not determined. The above procedures were performed in the CTCF and RNAPII ChIA-PET datasets independently. For CTCF ChIA-PET MiSeq data from GM12878 cells, 636 CTCF interaction loops were identified as phased interactions. Allele-specific chromatin loops and anchors can also be used to show chromatin loop specificity between cells with different genetic backgrounds (Fig. 4 and fig. S4).

### *CCD calling*

CCDs are called from CTCF ChIA-PET data using a modified version of our prior algorithm (*6*). First, only loops with both anchors peak supported are retained. Then, the loops are further filtered on the basis of PET counts (interaction frequencies), i.e., the loops are ordered by increasing PET count, and using a percentile cutoff (67th percentile), only the top one-third of loops with respect to PET count are retained. Using this refined collection of loops, CCDs are then defined as a genomic region of at least 25 kb in length that has continuous coverage by loops with no gaps or break points.

### *E-P annotation of loops*

To perform the E-P annotation of loops, the user supplies, via the config file, a BED file of enhancer coordinates and a BED file of promoter coordinates. If ChromHMM chromatin state calls are available for the cell type of interest, then the subset of enhancer and promoter coordinates can easily be extracted from this file. After loops are called (PET count ≥3 and peak support), each loop is assigned a unique ID. The loops are then split into two files: one for the left anchors of the loops and one for the right anchors of the loops. The left anchor BED file is then intersected with the enhancer BED file and separately with the promoter BED file using bedtools (*29*). Similarly, the right anchor BED file is intersected with the enhancer and promoter BED files. Once the intersected output files are available, a custom Python script is used to determine if each loop ID qualifies as an E-P loop, P-P loop, or neither.

## SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/28/eaay2078/DC1

View/request a protocol for this paper from *Bio-protocol*.

## REFERENCES AND NOTES


1. M. J. Fullwood, M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Y. Chew, P. Y. H. Huang, W.-J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. S. A. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. M. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W.-K. Sung, E. T. Liu, C.-L. Wei, E. Cheung, Y. Ruan, An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
2. G. Li, M. J. Fullwood, H. Xu, F. H. Mulawadi, S. Velkov, V. Vega, P. N. Ariyaratne, Y. B. Mohamed, H.-S. Ooi, C. Tennakoon, C.-L. Wei, Y. Ruan, W.-K. Sung, ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol.* **11**, R22 (2010).
3. X. Li, O. J. Luo, P. Wang, M. Zheng, D. Wang, E. Piecuch, J. J. Zhu, S. Z. Tian, Z. Tang, G. Li, Y. Ruan, Long-read ChIA-PET for base-pair-resolution mapping of haplotype-specific chromatin interactions. *Nat. Protoc.* **12**, 899–915 (2017).
4. G. Li, X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C.-L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W.-K. Sung, M. Snyder, Y. Ruan, Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).
5. J. M. Dowen, Z. P. Fan, D. Hnisz, G. Ren, B. J. Abraham, L. N. Zhang, A. S. Weintraub, J. Schuijers, T. I. Lee, K. Zhao, R. A. Young, Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374–387 (2014).
6. Z. Tang, O. J. Luo, X. Li, M. Zheng, J. J. Zhu, P. Szalaj, P. Trzaskoma, A. Magalska, J. Wlodarczyk, B. Ruszczycki, P. Michalski, E. Piecuch, P. Wang, D. Wang, S. Z. Tian, M. Penrad-Mobayed, L. M. Sachs, X. Ruan, C.-L. Wei, E. T. Liu, G. M. Wilczynski, D. Plewczynski, G. Li, Y. Ruan, CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* **163**, 1611–1627 (2015).
7. X. Ji, D. B. Dadon, B. E. Powell, Z. P. Fan, D. Borges-Rivera, S. Shachar, A. S. Weintraub, D. Hnisz, G. Pegoraro, T. I. Lee, T. Misteli, R. Jaenisch, R. A. Young, 3D chromosome regulatory landscape of human pluripotent cells. *Cell Stem Cell* **18**, 262–275 (2016).
8. A. S. Weintraub, C. H. Li, A. V. Zamudio, A. A. Sigova, N. M. Hannett, D. S. Day, B. J. Abraham, M. A. Cohen, B. Nabet, D. L. Buckley, Y. E. Guo, D. Hnisz, R. Jaenisch, J. E. Bradner, N. S. Gray, R. A. Young, YY1 is a structural regulator of enhancer-promoter loops. *Cell* **171**, 1573–1588.e28 (2017).
9. C.-L. Wei, Q. Wu, V. B. Vega, K. P. Chiu, P. Ng, T. Zhang, A. Shahab, H. C. Yong, Y. Fu, Z. Weng, J. Liu, X. D. Zhao, J.-L. Chew, Y. L. Lee, V. A. Kuznetsov, W.-K. Sung, L. D. Miller, B. Lim, E. T. Liu, Q. Yu, H.-H. Ng, Y. Ruan, A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124**, 207–219 (2006).
10. A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, K. Zhao, High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
11. E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, J. Dekker, Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
12. D. H. Phanstiel, A. P. Boyle, N. Heidari, M. P. Snyder, Mango: A bias-correcting ChIA-PET analysis pipeline. *Bioinformatics* **31**, 3092–3098 (2015).
13. G. Li, Y. Chen, M. P. Snyder, M. Q. Zhang, ChIA-PET2: A versatile and flexible pipeline for ChIA-PET data analysis. *Nucleic Acids Res.* **45**, e4 (2017).
14. M. R. Mumbach, A. J. Rubin, R. A. Flynn, C. Dai, P. A. Khavari, W. J. Greenleaf, H. Y. Chang, HiChIP: Efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).
15. R. Fang, M. Yu, G. Li, S. Chee, T. Liu, A. D. Schmitt, B. Ren, Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res.* **26**, 1345–1348 (2016).
16. J. T. Robinson, D. Turner, N. C. Durand, H. Thorvaldsdóttir, J. P. Mesirov, E. L. Aiden, Juicebox.js provides a cloud-based visualization system for Hi-C data. *Cell Syst.* **6**, 256–258.e1 (2018).
17. P. Kerpedjiev, N. Abdennur, F. Lekschas, C. McCallum, K. Dinkla, H. Strobelt, J. M. Luber, S. B. Ouellette, A. Azhir, N. Kumar, J. Hwang, S. Lee, B. H. Alver, H. Pfister, L. A. Mirny, P. J. Park, N. Gehlenborg, HiGlass: Web-based visual exploration and analysis of genome interaction maps. *Genome Biol.* **19**, 125 (2018).
18. X. Zhou, R. F. Lowdon, D. Li, H. A. Lawson, P. A. F. Madden, J. F. Costello, T. Wang, Exploring long-range genome interactions using the WashU Epigenome Browser. *Nat. Methods* **10**, 375–376 (2013).
19. Y. Wang, F. Song, B. Zhang, L. Zhang, J. Xu, D. Kuang, D. Li, M. N. K. Choudhary, Y. Li, M. Hu, R. Hardison, T. Wang, F. Yue, The 3D Genome Browser: A web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol.* **19**, 151 (2018).
20. The ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
21. J. Dekker, A. S. Belmont, M. Guttman, V. O. Leshyk, J. T. Lis, S. Lomvardas, L. A. Mirny, C. C. O'Shea, P. J. Park, B. Ren, J. C. R. Politz, J. Shendure, S. Zhong; 4D Nucleome Network, The 4D nucleome project. *Nature* **549**, 219–226 (2017).
22. P. V. Kharchenko, M. Y. Tolstorukov, P. J. Park, Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* **26**, 1351–1359 (2008).

23. Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nussbaum, R. M. Myers, M. Brown, W. Li, X. S. Liu, Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
24. K. R. Rosenbloom, C. A. Sloan, V. S. Malladi, T. R. Dreszer, K. Learned, V. M. Kirkup, M. C. Wong, M. Maddren, R. Fang, S. G. Heitner, B. T. Lee, G. P. Barber, R. A. Harte, M. Diekhans, J. C. Long, S. P. Wilder, A. S. Zweig, D. Karolchik, R. M. Kuhn, D. Haussler, W. J. Kent, ENCODE data in the UCSC Genome Browser: Year 5 update. *Nucleic Acids Res.* **41**, D56–D63 (2013).
25. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
26. N. C. Durand, M. S. Shamim, I. Machol, S. S. P. Rao, M. H. Huntley, E. S. Lander, E. L. Aiden, Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
27. N. C. Durand, J. T. Robinson, M. S. Shamim, I. Machol, J. P. Mesirov, E. S. Lander, E. L. Aiden, Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
28. D. Capurso, Z. Tang, Y. Ruan, Methods for comparative ChIA-PET and Hi-C data analysis. *Methods* **170**, 69–74 (2020).
29. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
30. F. Ramírez, D. P. Ryan, B. Grüning, V. Bhardwaj, F. Kilpert, A. S. Richter, S. Heyne, F. Dündar, T. Manke, deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
31. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
32. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B. Methodol.* **57**, 289–300 (1995).

**Acknowledgments:** We thank S. Sampson from The Jackson Laboratory for editing this manuscript. We thank the members of the Wei laboratory, Ruan laboratory, and Li laboratory, C.-H. Wong, J. George (The Jackson Laboratory), and I. Gabdank (Stanford University) for the helpful discussions; Z. Reifsnyder for artistic improvement of the figures; and the legacy programmers who contributed to BASIC Browser during its early development. **Funding:** Y.R. was supported by the NIH ENCODE (UM1 HG009409), 4DN (U54 DK107967), and JAX Director's Innovation Fund (JAX-DIF 19000-18-02). S.L. was supported by the NIH National Institute of General Medical Sciences (R35GM133562), Leukemia Research Foundation 2017 Funding Cycle Scientific Research Grant, The Jackson Laboratory Director's Innovation Fund (JAX-DIF 19000-17-13 and 19000-20-05), and The Jackson Laboratory Cancer Center New Investigator Award. Research reported in this publication was partially supported by the National Cancer Institute of the NIH under award number P30CA034196. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.
**Author contributions:** C.-L.W., Y.R., and S.L. designed the overall concept. B.L., J.W., L.C., M.K., S.N., H.T., Z.T., and A.A. developed and tested the pipeline. Y.F. and P.W. produced the wet-lab experiment to produce sequence libraries. All authors contributed to the draft and revised the manuscript. **Competing interests:** The authors declare that they have no competing interests. ChIA-PIPE is implemented in bash, awk, Python, Perl, and R and is freely available under the MIT license at https://github.com/TheJacksonLaboratory/ChIA-PIPE. BASIC Browser is available as a Docker image, which is freely available at https://github.com/TheJacksonLaboratory/basic-browser. Availability of data and materials: The HFFc6 ChIA-PET data used in the current paper have been deposited to 4DN portal and are available to download at https://data.4dnucleome.org/experiment-set-replicates/4DNESCQ7ZD21/. The GM12878 HiChIP data used in this manuscript were downloaded from the Sequence Read Archive (SRR3467175). **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors

Submitted 29 May 2019
Accepted 28 May 2020
Published 10 July 2020
10.1126/sciadv.aay2078

**Citation:** B. Lee, J. Wang, L. Cai, M. Kim, S. Namburi, H. Tjong, Y. Feng, P. Wang, Z. Tang, A. Abbas, C.-L. Wei, Y. Ruan, S. Li, ChIA-PIPE: A fully automated pipeline for comprehensive ChIA-PET data analysis and visualization. *Sci. Adv.* **6**, eaay2078 (2020).