Europe PMC Funders Author Manuscripts

Comparison of computational methods for the analysis of Hi-C data

Mattia Forcato[1], **Chiara Nicoletti**[1], **Koustav Pal**[2], **Carmen Maria Livi**[2], **Francesco Ferrari**[#2,3,*], and **Silvio Bicciato**[#1,*]

[1]Dept. of Life Sciences, Center for Genome Research, University of Modena and Reggio Emilia, Modena, Italy

[2]IFOM - The FIRC Institute of Molecular Oncology, Milan, Italy

[3]Institute of Molecular Genetics, National Research Council, Pavia, Italy

[#] These authors contributed equally to this work.

Abstract

Hi-C is a genome-wide sequencing technique to investigate the 3D chromatin conformation inside the nucleus. The most studied structures that can be identified from Hi-C - chromatin interactions and topologically associating domains (TADs) - require computational methods to analyze genome-wide contact probability maps. We quantitatively compared the performances of 13 algorithms for the analysis of Hi-C data from 6 landmark studies and simulations. The comparison revealed clear differences in the performances of methods to identify chromatin interactions and more comparable results of algorithms for TAD detection.

The identification of the three dimensional structure of chromatin inside the nucleus is crucial to decipher how the spatial organization of DNA affects genome functionality and transcription. Methods based on Chromosome Conformation Capture (3C)1 such as Hi-C combine proximity-based DNA ligation with high-throughput sequencing to assess spatial proximity of potentially any pair of genomic loci2. These techniques investigate chromatin structures, as interactions and topologically associating domains (TADs)3. Chromatin

---

[*]To whom correspondence should be addressed, Silvio Bicciato, Dept. of Life Sciences, Center for Genome Research, University of Modena and Reggio Emilia, Via G. Campi, 287 – 41125 Modena, Italy, Tel: +39 059 2055 219, Fax: +39 059 2055 410, silvio.bicciato@unimore.it; Francesco Ferrari, IFOM - The FIRC Institute of Molecular Oncology, Via Adamello, 16 – 20139 Milan, Italy, Tel: +39 02 57430 3830, francesco.ferrari@ifom.eu.

Author Contributions

M.F., C.N. and K.P. collected the experimental data, and implemented the computational pipelines. M.F., C.N., K.P. and C.M.L. analyzed the Hi-C datasets. M.F. and C.N. compiled the list of interaction evidences. F.F. generated the simulated data. M.F., F.F. and S.B. designed the experiments and analyzed the results. M.F., C.N., F.F., and S.B. wrote the manuscript.

Competing Financial Interests

The authors declare no competing financial interests.

Data Availability

The Hi-C experimental data that support the findings of this study are available at the Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra) under the accession numbers listed in Supplementary Table 1. The Hi-C simulated data are available at https://bitbucket.org/mforcato/hictoolscompare. All data used to generate main and supplementary figures are provided as source data files. Any other data supporting the findings of this study is available from the corresponding author upon request.

interactions are contacts between regions far from each other on the linear DNA sequence, but close in the 3D space4. TADs are structural domains consisting of highly self-interacting chromatin regions, with limited interaction with regions in other domains5–7.

Hi-C produces hundreds of millions of read-pairs that are used to generate genome-wide maps containing millions of contacts between genomic loci pairs8–10. The analysis of this enormous amount of genomic data required the development of *ad-hoc* algorithms and computational procedures. Different bioinformatics tools have been recently implemented to efficiently preprocess sequence reads (quality control, alignment, and filtering), remove biases (normalization of contact matrices), and infer chromatin structures10,11. To ensure the reproducibility of results it would be desirable to assess how the various tools perform relative to one another, as algorithmic choices severely impact the identification of chromatin structures and most approaches require heuristic selection of parameters9,12,13.

We quantitatively compared the performances of Hi-C data analysis methods for the identification of chromatin interactions9,14–19 and topological domains5,9,14,20–24 using experimental and simulated data. We also addressed tool usability including running time and computational requirements. In general we see that, depending on the tool, identified structures vary in terms of quantity and characteristics and are more reproducible for TADs than for interactions.

Results

Tools and data preprocessing

We compared thirteen methods for the analysis of Hi-C data (Table 1; Supplementary Notes 1 and 2), using experimental and simulated data. Experimental data have been obtained from 6 landmark studies2,5,7–9,25 selecting 9 datasets with 41 samples covering multiple protocol variations, data resolutions, and cell types (Table 2 and Supplementary Table 1). We generated simulated data with a modified version of the model proposed by Lun and Smyth19 (Supplementary Note 3). The various methods preprocess Hi-C data using different alignment and filtering strategies (Fig. 1a and Supplementary Table 2). Most interaction callers do not include an alignment step and we used Bowtie26, a full-read approach, for read mapping. Instead, HIPPIE, HiCCUPS, and diffHic use chimeric alignment that allows mapping also reads spanning the ligation junction. Each interaction caller adopts a specific filtering method, with the exception of Fit-Hi-C for which we used GOTHiC filtering. Most TAD callers require, as input, a fully preprocessed interaction matrix and thus they do not provide specific approaches for alignment and filtering - TADbit and Arrowhead are the two exceptions. Thus, to maximize comparability, we applied a uniform preprocessing procedure (i.e., Bowtie for alignment and hicpipe for filtering) to create the interaction matrix for TAD identification.

Methods implementing chimeric alignment aligned on average 18.4% (chimeric STAR27 in HIPPIE), 27.4% (chimeric BWA28 in HiCCUPS), and 40.1% (chimeric Bowtie229 in diffHic) more reads than Bowtie. The difference in alignment rate between chimeric and full-read became more evident as the read length increased, ranging from 30.9% (at 36bp) to 55.4% (at 101bp) of additionally aligned reads (chimeric Bowtie2, Fig. 1b).

After the filtering step, HiCCUPS retained the largest number of aligned reads (Fig. 1c), although it is worth noting that it filters only PCR duplicates without discarding other potential artifact reads. diffHic generally filtered the highest proportion of aligned reads (from 27% to 94% depending on the dataset), but, given its higher alignment rate, still retained a large number of reads (Supplementary Table 3). The different experimental protocols severely affected the percentage of filtered reads, with *in situ* Hi-C resulting in more reads passing the filtering step (>76%; Fig. 1c). The smaller fraction of retained reads observed in data generated with the simplified Hi-C protocol was mostly due to a larger amount of PCR duplicates (Supplementary Table 3).

Hi-C read counts are usually summarized at the level of genomic bins with a fixed width larger than the size of individual restriction fragments. For each dataset, we used the same bin size (resolution) of the original publication to call interactions, whereas we used bins of at least 40kb for TADs calling (Table 2).

When a method required a normalization step, we used its original normalization procedure, while we applied hicpipe to normalize the matrices for DomainCaller, InsulationScore, Arrowhead, Armatus, and TADtree (Fig. 1a). In all cases, we did not evaluate the effect of different normalization strategies as thorough comparisons of normalization methods have already been addressed30–32.

Identification of chromatin interactions

On experimental data, the total number of interactions called by each method increased with the number of reads retained by the filtering step, for all tools at any resolution, although the rate of increase varied from tool to tool (Fig. 2a). Consistent with the expectation that 3D interactions mostly occur within chromosomes (*cis*) rather than between chromosomes (*trans*), all methods detected more *cis* than *trans* interactions. In most datasets, GOTHiC called the highest number of *cis* interactions (Supplementary Fig. 1a) and, in general, diffHic found the largest number of *trans* interactions (Supplementary Fig. 1b). For all tools, the rate of increase of the number of interactions with the number of retained reads was higher for *cis* than for *trans* interactions (Supplementary Fig. 1c). HiCCUPS, aggregating nearby peaks into a single interaction, identified fewer interactions than all other tools.

When considering the distance between the interacting points in *cis*, GOTHiC found interactions at shorter mean distance, at both 5 and 40kb resolutions (Fig. 2b and Supplementary Fig. 2). At 5kb, Fit-Hi-C called interactions at an average distance of more than 10Mb, as expected being designed to call mid-range interactions. At a resolution of 1Mb, with the exception of HIPPIE, all tools detected interactions with an average distance comprised between 10 (HiCCUPS and GOTHiC) and 53 (diffHic) Mb (Supplementary Fig. 2).

The differences in the number of interactions and in the distance between the interacting points identified by the various methods are immediately evident in the visual representation of the contact matrices (Fig. 2c).

To compare the reproducibility of interactions called in different replicates, we calculated the similarity coefficient of Jaccard (Jaccard Index, JI), as a measure of the overlap between sets of interactions. In general, the reproducibility among replicates of the same data set (intra-dataset) was low at all resolutions (Fig. 2d and Supplementary Fig. 3a), yet significantly higher than random sets of interactions (p-values≤0.001; Supplementary Fig. 3b). Surprisingly, the concordance was higher for *trans* (median JI of 0.19) than for *cis* interactions (median JI<0.03). At low resolution GOTHiC had the highest concordance, most likely due to the fact that it called a large number of short-range interactions in every sample replicate. Conversely, in almost all datasets at high resolution, the interactions found by HiCCUPS were the most conserved among replicates. The quantification of the Jaccard Index considering only the top 1,000 *cis* interactions (called by each method in each replicate of Rao IMR90) resulted, with the exception of Fit-Hi-C, in no overall significant improvement of the concordance (q-value>0.05 in a one-tail Wilcoxon test with Benjamini-Hochberg correction; Supplementary Fig. 4a). Instead, when grouping samples based on increasing number of reads, the reproducibility increased with the number of reads especially for HiCCUPS and GOTHiC (Supplementary Fig. 4b). The interactions identified by HiCCUPS and GOTHiC were the most reproducible also when using the overlap coefficient, a similarity measure more robust to imbalanced number of interactions between the compared replicates (Supplementary Fig. 4c).

The intra-dataset reproducibility remained similar when comparing replicates of the same cell line processed using different restriction enzymes (Supplementary Fig. 5). Instead, the inter-dataset reproducibility, i.e., the concordance between interactions called in samples of the same cell line in different datasets (using different protocols and enzymes), was much lower (median JI<4×10^{-4}; Supplementary Fig. 6).

We then evaluated the performance of each tool in detecting interactions associated to chromatin states related to transcriptional regulation. In particular, for each dataset and cell type, we classified interactions based on the respective chromatin states at their anchoring points33,34. Considering all methods and the data at 5kb resolution, on average 16% of all detected *cis* interactions were classified as promoter-enhancer, 23% as interactions connecting heterochromatin or quiescent states, and 3% as biologically less expected, i.e., connecting promoter or enhancer to heterochromatin or quiescent states (Fig. 2e). At this resolution, HiCCUPS and HOMER called the highest proportion of promoter-enhancer interactions, although not the highest absolute number (Supplementary Fig. 7a). In datasets at 40kb resolution, all methods detected larger proportions of promoter-enhancer interactions due to the higher probability for larger bins to contain an enhancer or a promoter (Supplementary Fig. 7b). On the contrary, the proportion of *trans* interactions, classified as promoter-enhancer, was very low for all tools in almost all datasets (Supplementary Table 5). diffHic returned the highest quantity and percentage of interactions connecting heterochromatin or quiescent states, even though, in some datasets, the proportion of this type of interaction was extremely high for all tools. Irrespective of the method and of the resolution, less than 8% of all *cis* interactions were classified as biologically less expected. For all tools, the enrichment of the number of promoter-enhancer interactions over random expectation tends to be higher in datasets at higher resolution (p-value≤0.01 in a hypergeometric test for most datasets at 5kb; Supplementary Table 6).

All methods identified large proportions of convergent orientation of CTFC motifs, a distinctive feature of specific type of interactions9, among interactions with a single CTCF-binding motif in each of the two interacting bins (Supplementary Note 4).

When comparing the power to recall validated *cis* interaction evidences (Supplementary Table 7), GOTHiC recovered the largest amount of true-positive interactions. HOMER and Fit-Hi-C performed comparably to GOTHiC, although calling a smaller number of total interactions (Fig. 2f). In high-resolution datasets, the best performance was achieved by diffHic although HOMER identified more true-positives than any other tool, at comparable numbers of called interactions (Supplementary Fig. 7c). All tools recalled low proportions of true negatives in almost all datasets, albeit GOTHiC resulted more prone to false positives in datasets at 40kb (Supplementary Fig. 7d).

To assess sensitivity and precision of the methods, we modified the model of Lun and Smyth19 to generate simulated interaction matrices and analyzed the simulated data with HiCCUPS, HOMER, diffHic, and Fit-Hi-C, the only tools that can take as input the sole interaction matrix. For a set of 40 samples, at 8 levels of base interaction strength, all tools called a much larger number of interactions than the 1,000 true interactions (Supplementary Fig. 8a). As for experimental data, Fit-Hi-C called interactions at larger mean distance (Supplementary Fig. 8b-c). The highest sensitivity was achieved by Fit-Hi-C, although all tools displayed an extremely high FDR (i.e., a low precision) (Supplementary Fig. 8d-e).

Identification of Topologically Associating Domains

For TAD calling, we analyzed all experimental data at a resolution of 40kb, with the exception of Lieberman-Aiden for which we used the original 1Mb resolution. Differently from interaction callers, the number of TADs was not increasing with the number of reads retained after filtering for all tools, with the sole exception of Arrowhead (Fig. 3a). The number of identified TADs varied from tool to tool and was, generally, inversely proportional to their size (Fig. 3b). In all datasets at 40kb, on average TADtree called the largest (7638) and Arrowhead the smallest (636) number of TADs. Conversely, at 1Mb, InsulationScore returned the largest number of TADs (Supplementary Table 8). The characteristics of the identified TADs are exemplified in the heatmap representation of the contact matrices (Fig. 3c). Note that some methods partition chromosomes in a continuous set of TADs (HiCseg, TADbit, InsulationScore), whereas the others allow gaps between TADs. Arrowhead and TADtree, which adopt multi-scale approaches, returned nested TADs.

To compare TADs reproducibility, we calculated the Jaccard Index as a measure of the overlap between TAD boundaries across biological replicates. At all resolutions, HiCseg had the highest reproducibility among replicates of the same data set (intra-dataset; Fig. 3d and Supplementary Fig. 9a). In general, the reproducibility of TAD boundaries was higher (median JI of 0.25) than what observed for chromatin interactions. The reproducibility increased with the number of reads for all methods when grouping samples based on increasing number of reads (Supplementary Fig. 9b). TADs identified by HiCseg were the most reproducible also when using the overlap coefficient (Supplementary Fig. 9c).

The intra-dataset reproducibility remained similar for most tools when using different restriction enzymes for the same cell line (Supplementary Fig. 10). Instead, the inter-dataset concordance (i.e., between TAD boundaries called in replicates of the same cell line in different datasets obtained using different protocols and enzymes) was lower than the intra-dataset reproducibility, with TADtree showing the highest and Arrowhead the lowest inter-dataset concordance (Supplementary Fig. 11).

The various tools called TADs with consistent enrichment of insulators (e.g. CTCF or BEAF32; Supplementary Table 9) at the TAD boundaries. In almost all datasets, more than 50% of TAD borders overlapped CTCF peaks (Supplementary Table 10). Moreover, all tools identified TADs with an enrichment of CTCF peaks at the TAD borders with Armatus and TADtree returning domains with a stronger CTCF enrichment at their borders (Supplementary Fig. 12a). In Sexton dataset, most tools returned TADs with a clear enrichment, at TAD borders, of BEAF32, an architectural protein reported to be more enriched than CTCF at TAD boundaries in Drosophila7 (Supplementary Fig. 12b).

When using synthetic data, DomainCaller, TADbit and InsulationScore identified a number of TADs comparable to the number of simulated not overlapping TADs, irrespectively of the noise (Supplementary Fig. 13a). As with experimental data, HiCseg called a small number of large TADs, whereas TADtree identified a large number of small domains (Supplementary Fig. 13b). The ability of both methods to identify the correct structures was strongly affected by the noise present in the data (Supplementary Fig. 13c-d). TADbit and Armatus had the highest sensitivity in recovering TAD boundaries, although TADbit displayed a higher precision (low FDR) at all noise levels. These results hold similar when simulating a hierarchy of nested TADs, while the precision of TADtree, specifically designed to identify nested domains, ameliorated in the latter case (Supplementary Fig. 13e-g).

Other analyses

In additional analyses, we compared the performances of interaction callers using a common preprocessing procedure (Supplementary Note 5 and Supplementary Fig. 14) and the computational requirements, running time, and usability of all tools (Supplementary Note 6 and Supplementary Fig. 15).

Discussion

The performances of algorithms for the identification of chromatin interactions and Topologically Associating Domains from Hi-C data have been, in most cases, compared using semi-quantitative approaches19,20,23,24. Indeed, a robust quantification of performance in terms of specificity and sensitivity is hindered by the lack of ground truth positive and negative controls for chromatin architecture and by conceptual difficulties in designing simulators of Hi-C data. To overcome these limitations, we adopted a framework that uses a large set of experimental and synthetic data and exploits various metrics to quantitatively compare the performance of several tools currently available for the analysis of Hi-C data.

Based on this comparison framework, our results indicate that there is no algorithm that can be considered the gold standard to identify chromatin interactions. Independently of the data resolution, the choice of the method impacts the quantity and characteristics of the identified interactions.

Here, to quantitatively assess the concordance of identified interactions, we kept replicates separated while Hi-C replicates are commonly pooled before the analysis to generate a unique sample with higher number of reads. Surprisingly, interactions called in one replicate were poorly conserved in other replicates from the same cell type of the same study. The overall low reproducibility may be partly explained by the fact that biological replicates, being an ensemble of cells in different states and phases of the cell cycle, are not necessarily identical in terms of chromatin contacts, as hypothesized when quantifying reproducibility in terms of the co-occurrence of the same point interaction. Notwithstanding the limited reproducibility, all methods detected comparable, statistically significant proportions of *cis* promoter-enhancer looping interactions and a very small quantity of interactions classified as biologically less plausible.

In agreement with what recently reported by Dali and Blanchette35, TAD callers returned different numbers of TADs with different mean size. However, predicted TADs were more comparable than loops among replicates and were characterized by enrichment in binding sites of known architectural proteins.

Overall, this comparison suggests that, although no single method outperforms others in all situations, TAD callers are methodologically more mature than interaction callers. Among TAD callers, TADbit, Armatus, and TADtree had balanced performances for most metrics in experimental and simulated data. For interaction callers, HOMER and HiCCUPS yielded the highest proportion of interactions with a potential biological significance, although HiCCUPS potentialities (e.g., in terms of absolute number of called interactions) could be fully exploited only in the analysis of very high-resolution datasets.

We observed a difficulty in reconciling the results obtained from experimental and synthetic data, especially for interaction callers. This can be most likely ascribed to the complexity of designing sound strategies to simulate Hi-C datasets with predefined features that represent well-defined and unambiguous true positives and negatives. Although several promising approaches are available from the biophysics of polymer folding modeling36, no algorithm has been proposed so far to generate reads that fully mimic the distribution and biases observed in real Hi-C data. The availability of synthetic data will be essential to rationally tune any algorithm parameter, thus limiting the heuristics currently inherent in the choice of the best setting.

The various tools greatly differed in terms of usability, interoperability, stability of the implementation, and computing resources required to complete the analysis. Considering the pace of data production, priorities for developers should be the deployment of methods able to analyze larger and higher resolution datasets with reasonable amounts of computational resources and the adoption of common data formats to easily exchange inputs and outputs among the various tools37.

Europe PMC Funders Author Manuscripts

Online Methods

Hi-C data analysis tools

We chose algorithms that (i) were specifically designed for the identification of chromatin interactions and TADs and (ii) had a publicly available implementation at the time of our survey (July 2016). An extended description of the methods is provided in Supplementary Notes 1 and 2.

Among the tools to identify chromatin interactions, Fit-Hi-C15 uses spline models to estimate the expected contact probabilities as a function of distance. Statistical significance of interactions is calculated using a binomial distribution and p-values corrected for multiple testing. Fit-Hi-C requires as input a raw count interaction file and a bias file calculated with an implementation of ICE, the iterative correction from Imakaev *et al.*31. In output, Fit-Hi-C returns only *cis* interactions characterized by contact count, p-value, and FDR. Significant interactions have been selected based on the FDR.

In GOTHiC16 significant chromatin interactions are identified using a binomial test followed by Benjamini-Hochberg multiple testing correction. GOTHiC takes aligned reads as input and perform read-pair level filtering and square root of vanilla coverage normalization (a type of implicit normalization). For all interactions (*cis* and *trans*), the algorithm outputs the log2 ratio of observed to expected interactions, p-value, FDR, and the number of supporting read pairs. Here, we used FDR and contact counts to identify significant interactions38.

HOMER17 performs a binomial test to find significant interactions. The input file is in the form of aligned reads; filtering is at read and read-pair level; the implicit normalization method is based on region coverage and distance between regions. All interactions (*cis* and *trans*) are characterized in terms of p-value, FDR, number of supporting read pairs (both observed and expected), and interaction distance. Significant interactions are called setting a threshold on the p-value.

HIPPIE18 implements an approach similar to the one presented in Jin *et al.*8 to call interactions. Significant interactions are detected by fitting a negative binomial distribution, where the expected random contact frequency (mean) is estimated from GC content, mappability, fragment length, and distance, and the overdispersion parameter is fixed and derived from Jin *et al.*8. HIPPIE starts from sequencing reads and performs chimeric alignment, read, read-pair and fragment level filtering, and explicit normalization without binning. The output is a set of restriction fragment-based interactions (inter- and intra-chromosomal) with an associated p-value. Significant interactions have been selected setting a threshold on the p-value.

diffHic19 takes raw sequencing data as input and performs chimeric alignment, read and read-pair level filtering. Significant interactions (*cis* and *trans*) are identified from the raw contact matrix using a *local* approach, *i.e.* searching for bin pairs that have substantially more reads than their neighbors, an approach conceptually similar to HiCCUPS9,14. The enrichment value for each interaction is calculated as the log-fold change between the

abundance (number of read pairs) of the target bin pair and the region of the neighborhood with the largest abundance. Here, we set thresholds on the enrichment, on the number of supporting reads, and on the distance from the diagonal to call interactions. When calling interactions on individual samples, no statistical test is performed and no significance value is returned.

HiCCUPS9,14 is part of the Juicer software suite, a pipeline to process and analyze Hi-C data starting from the raw sequencing files and generating normalized contact matrices at several resolutions. The pipeline aligns raw reads from FASTQ files using Burrows-Wheeler Aligner (BWA) algorithm, pairs the reads, handles chimeras, and merges and sorts the reads to filter out PCR duplicates. Juicer Tools Pre is used to create the normalized Hi-C contact matrix (.hic file) from the filtered read pairs. HiCCUPS takes as input the normalized Hi-C contact matrix to identify chromatin interactions. Specifically, HiCCUPS calls only *cis* interactions detecting pixels enriched with respect to four neighboring areas given the width of the peak and the window size as described in Rao *et al.*9. It returns the centroid of the clusters of significant peaks called using a modified Benjamini-Hochberg FDR.

Since most of the tools to identify Topologically Associating Domains lack the preprocessing steps, to maximize comparability we used a common pipeline based on the scripts of hicpipe30 to align, filter, and normalize the data used in input to the TAD callers.

HiCseg20 performs a 2D-segmentation based on a maximum likelihood approach to partition each chromosome in its constituent TADs directly from raw or normalized contact matrices. Here, we applied HiCseg to the raw Hi-C data.

TADbit21 implements a breakpoint detection algorithm that identifies the optimal segmentation of the chromosome under a Bayesian information criterion (BIC)-penalized likelihood. TADbit requires in input the observed read counts, which are then normalized using a modified implementation of ICE31. Although we used hicpipe for alignment and filtering also for TADbit, this tool contains an alignment module (based on the Genome Multitool (GEM) mapper for iterative alignment) and implements several filters.

DomainCaller5 is a single scale algorithm that identifies TADs using a Hidden Markov Model on the *Directionality Index*. The *Directionality Index* is a score quantifying the bias in downstream, as compared to upstream, contact probabilities for each bin, within a user-defined window of maximum distance. No preprocessing step is directly implemented by DomainCaller, which thus requires an external preprocessing tool to prepare the normalized contact matrix.

The InsulationScore22 is a segmentation algorithm that identifies TAD within normalized Hi-C matrices using a sliding square (insulation square). It combines contact signals inside the square and assigns an *insulation score* to each bin along the diagonal, thus obtaining a one-dimensional insulation vector. TAD boundaries are then identified based on the insulation vector.

Arrowhead9,14 is part of Juicer suite of tools for Hi-C data analysis and visualization. The tool is based on the Arrowhead transformation of Hi-C contact matrix, which results in

translating the patterns of TAD domains from “squares” along the diagonal to “triangles” of high or low signal. For each pair of loci, potential TAD boundaries, the algorithm computes specific scores for the “triangles” designed around the pair of loci, thus exploring the definition of TADs at multiple scales.

As Arrowhead, also TADtree23 can identify nested TADs. It is based on a 1D boundary index similar to the one developed by Sauria *et al.*32. The algorithm is based on the observation that the average enrichment of intra-TAD contacts grows linearly with distance, but when a TAD lies inside another one, its enrichment grows at a faster rate. The best TAD hierarchy is determined using a dynamic programming algorithm. No preprocessing step is directly implemented in TADtree, which thus requires an external preprocessing and normalization pipeline.

Armatus24 adopts a multiscale approach that can identify a consensus set of domains across various resolutions. It is based on a score function that quantifies the quality of a domain based on its local density of interactions. Since Armatus does not directly implement a preprocessing step, it requires a complete preprocessing pipeline to generate the normalized contact matrix.

For each method, we used the default statistical thresholds or the values suggested in the accompanying documentation to identify chromatin interactions or TADs (p-values or FDR). Only in the case of HIPPIE, to guarantee a statistical significance comparable to that of the other tools, we adopted a threshold (p-value<0.01) more conservative than the one suggested in the original publication (p-value<0.1; see Supplementary Note 1).

GOTHiC and HiCseg were run in R-3.1.3 while for diffHic (that requires at least R-3.2.0) we used R-3.2.0. We used version 2.7 for Python.

Experimental Hi-C data

We selected 9 Hi-C datasets from 6 studies obtained with 3 protocols at different resolutions (primarily determined by the restriction enzyme and sequencing depth) in overlapping cell types (n=41 samples; Table 2 and Supplementary Table 1). Data have been generated using *dilution Hi-C*, i.e., the original Hi-C protocol published in Lieberman-Aiden *et al.*2, *simplified Hi-C* introduced in Sexton *et al.*7, and *in situ Hi-C* developed by Rao and colleagues9. Samples comprise human cell lines from various tissues (embryonic stem cells: H1-hESC; fetal lung fibroblasts: IMR90; lymphoblastoid cell lines (LCL): GM12878 and GM06990) and D. melanogaster embryos. All data have been obtained using 6bp or 4bp cutter restriction enzymes. Some replicate samples from Lieberman-Aiden and Rao GM12878 have been processed with both restriction enzymes.

All biological replicates have been analyzed separately. In particular, the Rao GM12878 dataset contained 26 samples obtained with *in situ* protocol and MboI restriction enzyme and divided into a primary (16 technical replicates of 1 sample) and a replicate experiment (10 biological and technical replicates; see Supplementary Table 1 of Rao *et al.*9). Here, we selected the replicate with the highest number of sequenced reads from the primary experiment (i.e., SRR1658572, originally labeled as HIC003 and renamed here as replicate

H) and all the *in situ* samples of the replicate experiment. Moreover, we analyzed as separate samples the technical replicates of the replicate experiment since the authors defined technical replicates also those samples for which cells were cross-linked together but processed independently (Supplementary Table 1 and Supplementary Table 1 of Rao *et al.*9). In the Jin study, it has to be noted that the H1-hESC sample, originally composed of SRR639047, SRR639048, and SRR639049 and here renamed as replicate A, is the same sample of Dixon 2012 H1-hESC, composed of SRR442155, SRR442156, and SRR442157 and here renamed as replicate B (Supplementary Table 1). Both H1-hESC samples from Jin and Dixon 2012 were analyzed with chromatin interaction callers at their original resolutions (5 and 40kb, respectively), while we used only the H1-hESC sample from Dixon 2012 for the TAD analysis, conducted at 40kb for all datasets.

Preprocessing of experimental data

For most of the interaction callers we used the specific preprocessing procedure incorporated in the tool. Instead, with the only exception of TADbit and Arrowhead, all TAD callers require in input a fully preprocessed interaction matrix. For this reason, to maximize the comparability among the various methods, we used the same preprocessing procedure to prepare the data for all tools.

Reads were aligned to the hg19 build of the human genome or dm3 of the fly genome using: i) Bowtie26 (v.1.1.1) in single-end mode with parameters: -m 1 -a --best --strata --chunkmbs 200; ii) Bowtie 229 (v2.2.4) as implemented by diffHic, iii) STAR27 (v2.4.0) as implemented by HIPPIE, and iv) BWA28 (v0.7.15) as implemented by HiCCUPS. Bowtie performs full read alignment whereas diffHic, HIPPIE, and HiCCUPS implement different approaches for chimeric alignment (Supplementary Note 1). Reads aligned with Bowtie were used as input to those interaction callers lacking a specific aligner and to all TAD callers. In particular, for interaction callers, this choice was dictated by constraints in the type of input required by GOTHiC and HOMER that hampered the use of chimeric aligners. After alignment, samples composed of more than one run were merged with SAMtools39.

Most interaction callers implement their own filtering, binning, and normalization strategy (Supplementary Note 1). The filtering step is used to remove low quality reads, reads that may originate from unspecific ligation events or which are not informative. We grouped filters in three major categories: read-level, read-pair level, and fragment-level. Read-level procedures filter reads based on read mapping quality (AQ) and restriction site proximity (RSP). Read-pair level filters remove PCR duplicates (PD), spikes, *i.e.* reads aligning on a region with an abnormally high quantity of reads (S), and read pairs that derive from undigested chromatin (UC). This latter filter can also consider strand orientation to identify potential self-ligation or no ligation events (UC+SLF). Restriction site proximity filter can also be performed at read-pair level. Finally, fragment level filters (FLF) discard fragments based on the restriction site proximity of their reads. Reads have been filtered according to the strategy implemented by each tool. We also filtered out reads aligning on chrY and chrM for hg19 and on chr4, chrY, and all heterochromatic chromosomes for dm3.

In almost all cases, we set the bin size equal to the highest resolution reported in the original publications. However, due to severe computational requirements, we analyzed Jin dataset

and GM12878 samples of Rao at 5kb with interaction callers and all datasets originally binned at less than 40kb (Jin, Rao, and Dixon 2015) at 40kb with TAD callers.

All tools were run using default or suggested values for preprocessing parameters, filters, and normalization type. In some cases parameters were adjusted according to the adopted resolution, following suggestions from the software documentation or directly from the developers (Supplementary Notes 1 and 2). Some of the steps in the preprocessing workflow have been adapted to the requirements of the specific tools. In particular, since Fit-Hi-C requires in input raw interactions, we used GOTHiC, whose output format can be easily adapted to Fit-Hi-C input, to perform filtering and binning. The binning step was not required for HIPPIE, which calls interactions directly at the restriction fragment level. Whereas when using diffHic for calling interactions in individual samples the normalization step was not performed, since it is not required.

For all TAD callers, we used hicpipe for filtering and binning. hicpipe was also used for normalization in all TAD tools, with the exceptions of TADbit that requires the use of its internal normalization method and HiCseg that was applied to the raw interaction matrix (see Supplementary Note 2).

Simulated Hi-C data

We generated the simulated data using a modification of the procedure proposed by Lun and Smyth19 for a total of 65 samples obtained by varying the level of base interaction strength (for interactions only) and of noise (for TADs only; Supplementary Note 3). The simulated Hi-C count matrices were used as input to the interaction callers (HiCCUPS, HOMER, diffHic, and Fit-Hi-C) and to HiCseg and TADbit that require raw count as input. For all other TAD callers, requiring observed over expected normalized data, the raw count matrices were converted to Vanilla Coverage matrices, as described in Lieberman-Aiden *et al*2.

Performance metrics

To assess the performance of interaction callers, we considered several metrics including: the total number of called interactions; the distance between the interacting points in *cis*; the concordance of results within and between datasets when analyzing different biological replicates; and the type of associated chromatin states. To determine a further basis for comparison, we searched the literature for interactions that had been demonstrated to be present (or absent) in the same cell types of the Hi-C datasets. Namely, we selected interactions validated using other 3C techniques (e.g., 3C, 5C, ChIA-PET) and 3D-FISH, or reported in the literature to be specific of given cell types at a given physiological state (interaction evidences). Moreover, we calculated the sensitivity (true positive rate) and precision of the methods in identifying interactions from simulated data.

To compare TAD callers on experimental data we considered the total number of called TADs, the TAD size, the concordance of TAD boundaries within and between datasets when analyzing biological replicates, and the enrichment at TAD boundaries of known boundary elements (i.e., CTCF and BEAF32).

Comparative analyses

The intersection of the results from different replicates has been generated using the R package ChIPpeakanno.

For both interactions and TAD boundaries, the Jaccard Index of two replicates has been defined as the ratio between the size of the intersection and the size of the union of interactions and TAD boundaries called in the replicates. Jaccard Index empirical p-values were estimated with random permutations of interactions. Namely for each dataset, cell type, and data analysis method, we defined, for each sample, a random set of cis interactions by keeping constant the sample-specific number of interactions and the sample-specific distribution of distances between anchoring points. The first of the two anchoring points for each interaction was randomly selected from the pool of detectable anchoring points, defined as any genomic bin that was called as anchoring point in any sample from the same dataset and cell type. The second anchoring point was randomly defined by sampling from the observed distribution of anchoring point distances. The resulting sets of random interactions were then used to compute random Jaccard Index values in pairwise comparisons. The random sampling of interactions was repeated 1000 times to obtain a null distribution of randomly expected Jaccard Index values for each pairwise comparison. The empirical p-value is estimated as the probability of observing a random Jaccard Index value larger than or equal to the observed one.

Rao GM12878 replicates were divided into 4 groups of samples with increasing number of filtered read pairs. Specifically, replicates B2, B1, A2, A1, G1 constituted the group of samples with less than 40 million reads; A3, D, B, and G2 the group with more than 40 and less than 100 million reads; C2, C1, F, and A the group of samples with a number of filtered reads comprised between 100 and 180 millions; E1 and E2 constituted the group of samples with more than 180 million reads. Replicate H was not included in any of the above groups.

The overlap coefficient of two replicates was defined as the ratio between the size of the intersection and the size of the minimum set of interactions or TAD boundaries called in the replicates.

For interactions and TAD boundaries identified in simulated data, we defined sensitivity as the ratio of correctly identified features to all true features and precision as the ratio of correctly identified features to all called features (1 minus False Discovery Rate).

All comparative analyses were run using R-3.1.3. All box plots have been generated with the R *boxplot* function and default parameters.

Selection of validated interaction evidences

From the literature, we constructed a list of interactions that had been demonstrated to be present (or absent) in the same cell types of the Hi-C datasets using other 3C techniques (e.g., 3C, 5C, ChIA-PET) and 3D-FISH or that are known to exist in specific cell types at a given physiological state (interaction evidences). Altogether, we selected 2439 validated true-positive cell specific interactions, 389 validated true-negatives, 61 true positive evidences, and 138 true negative evidences (Supplementary Table 7). True positive and true

negative interactions were mapped to the bin level (at 40kb and 5kb resolution) and counted only if between not adjacent bins.

Integration with genomic data

Chromatin states for IMR90, H1-hESC and GM12878 (15-states model) were downloaded from Roadmap Epigenomics Consortium33 and chromatin states for fly late embryos (16 states) from modENCODE34 (details in Supplementary Note 7). CTCF and BEAF32 ChIP-seq peaks were retrieved from ENCODE40 and modENCODE34 (Supplementary Table 9). In particular, we considered peaks generated by the uniform analysis pipeline of the ENCODE Analysis Working Group and peaks obtained from combined replicates for modENCODE data.

We used the R package ChIPpeakanno to compare chromatin interactions with chromatin states and TAD boundaries with CTCF and BEAF32 peaks.

Code availability

Examples of how to run each tool, functions to analyze results, calculate general statistics, and performance metrics have been deposited in https://bitbucket.org/mforcato/hictoolscompare.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by AIRC Special Program Molecular Clinical Oncology “5 per mille” and an AIRC PI grant (to S.B.); by AIRC Start-up grant 2015 N.16841 (to F.F.); by Epigenetics Flagship project CNR-MIUR grants (to S.B.). This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 Research and Innovation Program (Grant Agreement No. 670126-DENOVOSTEM to S.B. and M.F.) and from CINECA (ISCRA Class C project HP10CDMGT8 to M.F.). C.M.L. is supported by SIPOD (Structured International Post Doc program of SEMM) a Marie Curie co-funded fellowship. We thank Aaron Lun for sharing the code used to simulate Hi-C data in the diffHic article. We thank Francesca Fanelli and the center for scientific computing of the University of Modena and Reggio Emilia for the use of GPUs. We would also like to thank the authors of all the compared tools for providing support for their methods and for prompt replying to our inquiries. We thank Michelangelo Cordenonsi, Paolo Maiuri, Endre Sebestyen, and Marco Morelli for critical feedback on the manuscript.

References

1. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002; 295:1306–1311. [PubMed: 11847345]
2. Lieberman-Aiden E, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009; 326:289–293. [PubMed: 19815776]
3. Pombo A, Dillon N. Three-dimensional genome architecture: players and mechanisms. Nat Rev Mol Cell Biol. 2015; 16:245–257.
4. Cavalli G, Misteli T. Functional implications of genome topology. Nat Struct Mol Biol. 2013; 20:290–9.
5. Dixon JR, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012; 485:376–380. [PubMed: 22495300]

6. Nora EP, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012; 485:381–385.
7. Sexton T, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell. 2012; 148:458–472. [PubMed: 22265598]
8. Jin F, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 2013; 503:290–294. [PubMed: 24141950]
9. Rao SSP, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014; 159:1665–1680. [PubMed: 25497547]
10. Schmitt AD, Hu M, Ren B. Genome-wide mapping and analysis of chromosome architecture. Nat Rev Mol Cell Biol. 2016; 17:743–755. [PubMed: 27580841]
11. Ay F, Noble WS. Analysis methods for studying the 3D architecture of the genome. Genome Biol. 2015; 16:183.
12. Mora A, Sandve GK, Gabrielsen OS, Eskeland R. In the loop: promoter-enhancer interactions and bioinformatics. Brief Bioinform. 2016; 17:980–995. [PubMed: 26586731]
13. Shavit Y, Merelli I, Milanesi L, Lio' P. How computer science can help in understanding the 3D genome architecture. Brief Bioinform. 2016; 17:733–744. [PubMed: 26433013]
14. Durand NC, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 2016; 3:95–98. [PubMed: 27467249]
15. Ay F, Bailey TL, Noble WS. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res. 2014; 24:999–1011. [PubMed: 24501021]
16. Mifsud B, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet. 2015; 47:598–606.
17. Heinz S, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010; 38:576–589. [PubMed: 20513432]
18. Hwang YC, et al. HIPPIE: a high-throughput identification pipeline for promoter interacting enhancer elements. Bioinformatics. 2015; 31:1290–1292. [PubMed: 25480377]
19. Lun ATL, Smyth GK. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. BMC Bioinformatics. 2015; 16:258. [PubMed: 26283514]
20. Lévy-Leduc C, Delattre M, Mary-Huard T, Robin S. Two-dimensional segmentation for analyzing Hi-C data. Bioinformatics. 2014; 30:i386–392.
21. Serra F, Baù D, Filion G, Marti-Renom MA. Structural features of the fly chromatin colors revealed by automatic three-dimensional modeling. bioRxiv. 2016; :036764.doi: 10.1101/036764
22. Crane E, et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature. 2015; 523:240–244. [PubMed: 26030525]
23. Weinreb C, Raphael BJ. Identification of hierarchical chromatin domains. Bioinformatics. 2016; 32:1601–1609. [PubMed: 26315910]
24. Filippova D, Patro R, Duggal G, Kingsford C. Identification of alternative topological domains in chromatin. Algorithms Mol Biol. 2014; 9:14.
25. Dixon JR, et al. Chromatin architecture reorganization during stem cell differentiation. Nature. 2015; 518:331–336.
26. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25. [PubMed: 19261174]
27. Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29:15–21. [PubMed: 23104886]
28. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma Oxf Engl. 2009; 25:1754–1760.
29. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9:357–359.
30. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet. 2011; 43:1059–1065.
31. Imakaev M, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods. 2012; 9:999–1003.

32. Sauria MEG, Phillips-Cremins JE, Corces VG, Taylor J. HiFive: a tool suite for easy and efficient HiC and 5C data analysis. Genome Biol. 2015; 16:237.
33. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518:317–330. [PubMed: 25693563]
34. Ho JWK, et al. Comparative analysis of metazoan chromatin organization. Nature. 2014; 512:449–452. [PubMed: 25164756]
35. Dali R, Blanchette M. A critical assessment of topologically associating domain prediction tools. Nucleic Acids Res. 2017; doi: 10.1093/nar/gkx145
36. Imakaev MV, Fudenberg G, Mirny LA. Modeling chromosomes: Beyond pretty pictures. FEBS Lett. 2015; 589:3031–3036. [PubMed: 26364723]
37. Dekker J, et al. The 4D Nucleome Project. bioRxiv. 2017; 103499. doi: 10.1101/103499
38. Schoenfelder S, et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res. 2015; 25:582–597.
39. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079.
40. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Editorial summary

Six tools to call chromatin loops and seven tools for TAD calling are systematically compared with real and simulated data. The strengths and weaknesses of each tool are discussed.

Europe PMC Funders Author Manuscripts

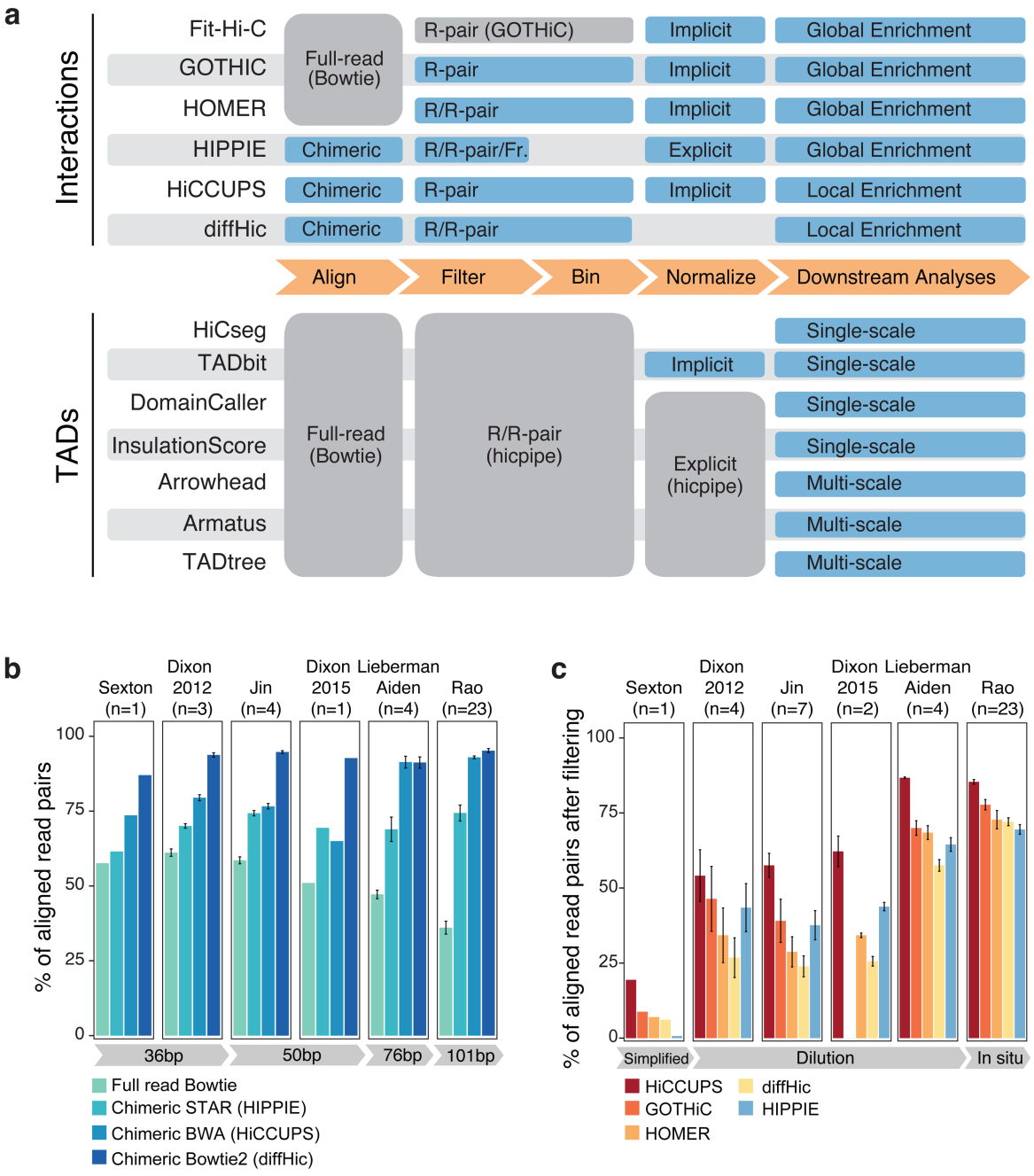

Figure 1. Tools for Hi-C data analysis used in the comparison and performances in data preprocessing.

a) Tools for the identification of chromatin interactions and TADs from Hi-C data and key analysis steps (orange arrows). Blue boxes detail the strategy used in each analysis step by each tool. A grey box is used when an external tool is required for a preprocessing step. Since most tools perform filtering and binning together, a blue or grey box spanning both steps is used in the schematic workflow. For filtering the following abbreviations are used: read level filtering (R); read-pair level filtering (R-pair); fragment level filtering (Fr.).

b) Percentage of aligned read pairs (alignment rate) for all datasets ordered by read length (grey arrows at the bottom). Data are shown as mean±standard error of the mean. Samples with different or mixed read length were not used when calculating the alignment rate.
c) Percentage of mapped reads retained after filtering (fraction of usable reads) in each dataset, ordered by experimental protocol (grey arrows at the bottom). Data are shown as mean±standard error of the mean. GOTHiC could not be applied to Dixon 2015 since the read-pairing step required an amount of memory larger than 1 TB of RAM.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

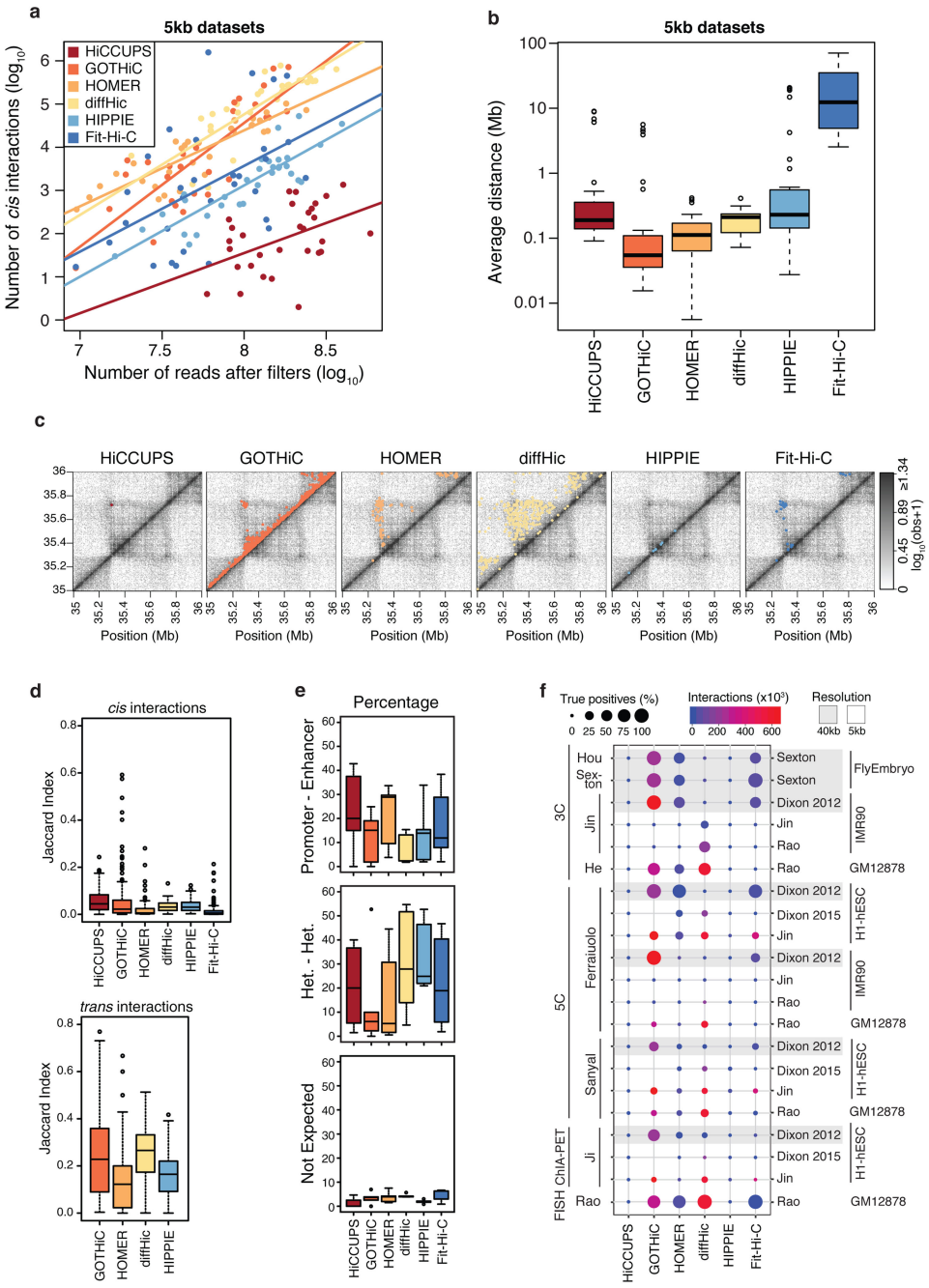

Figure 2. Comparative results of methods for the identification of chromatin interactions.

a) Scatter plot of total number of *cis* interactions called by each method as a function of the number of reads retained by the filtering step in all datasets at 5kb resolution (i.e., Jin H1-hESC, Jin IMR90, Rao GM12878, Rao IMR90, and Dixon 2015 H1-hESC; n= 32). Different points represent sample replicates. Linear interpolation for each method is shown as a solid line.

b) Boxplot of average distances between anchoring points in *cis* interactions (log scale) in sample replicates considering all datasets analyzed at 5kb resolution (n= 32).

c) Heatmap of the contact matrix of Rao GM12878 replicate H (chr21:35,000,000-36,000,000) at 5kb resolution. Identified peaks are marked in different colors for the various methods.
d) Box plots of the Jaccard Index for concordance of *cis* (upper) and *trans* (lower) interaction calls between sample replicates (intra-dataset concordance) for all datasets with at least 2 replicates (n=39; Supplementary Table 1). For Fit-Hi-C and HiCCUPS, the Jaccard Index was calculated only for *cis* interactions since these tools do not return *trans* interactions.
e) Proportion of *cis* interactions classified on the base of the chromatin states at their anchoring points (promoter-enhancer, upper; heterochromatin/quiescent to heterochromatin/quiescent, middle; less expected, lower) in all datasets at 5kb. With the exception of Jin H1-hESC (that contains a single replicate), only *cis* interactions conserved in at least 2 replicates within each dataset were classified using the chromatin states (Supplementary Table 4).
f) Performances in the identification of true positive validated evidences of *cis* interactions. Each row represents the comparison between a list of true positives and the interactions called by each method in each dataset. The dot size is proportional to the percentage of recalled true positives and the dot color accounts for the number of total called interactions. The validation technique and the name of true positive lists are displayed on the left side. The dataset used to call interactions are on the right and shaded in grey if at 40 kb resolution. True-positive interactions were searched among *cis* interactions conserved in at least 2 replicates within each dataset, with the exception of Jin H1-hESC and Sexton (both containing a single replicate). GOTHiC was not applied to Dixon 2015 (see legend of Fig. 1c).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

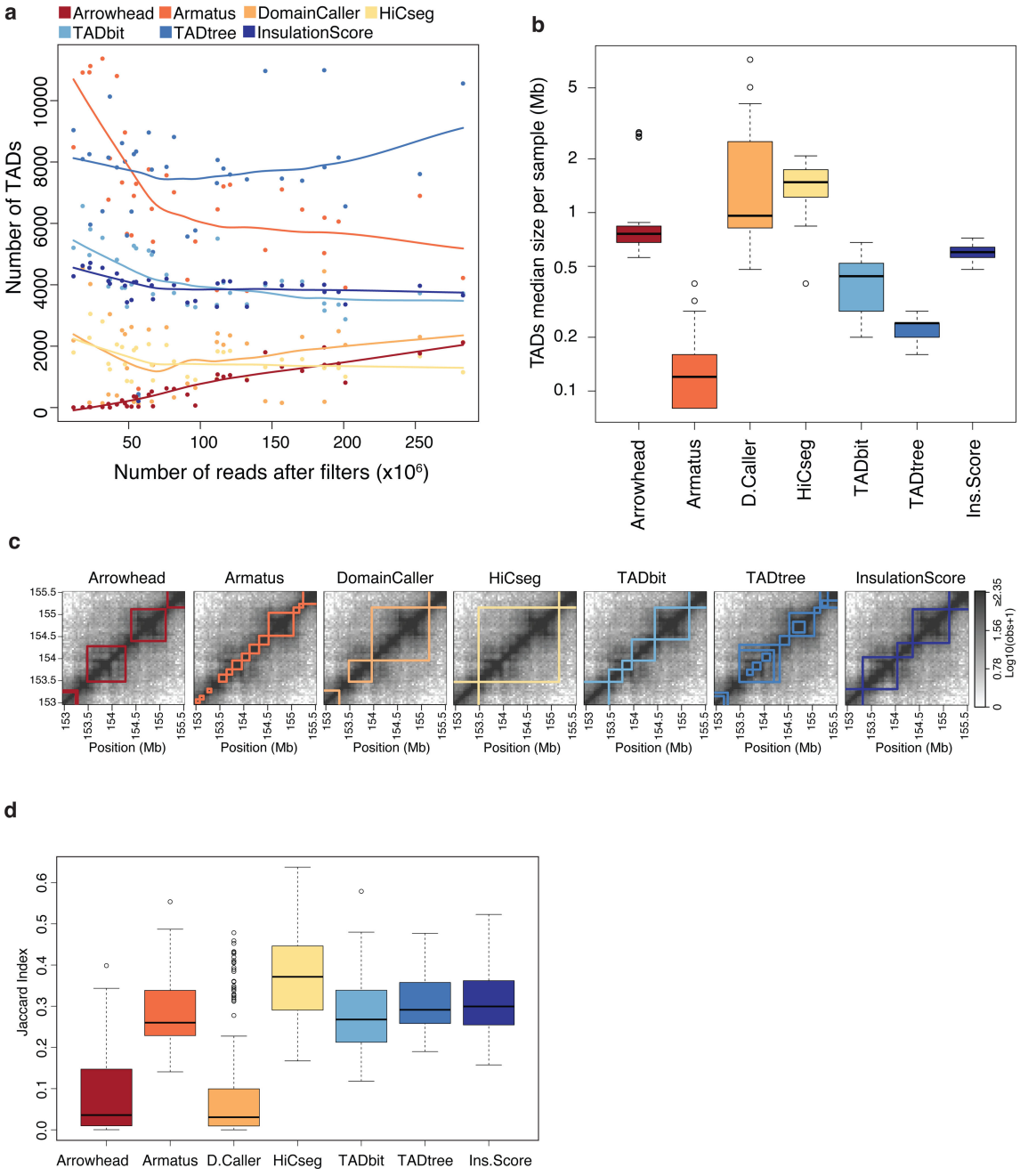

Figure 3. Comparative results of methods for the identification of TADs.

a) Scatter plot of total number of TADs called by each method as a function of the number of reads retained by the filtering step in all datasets except Lieberman-Aiden and Jin H1-hESC (n=36; Supplementary Table 1). Different points represent sample replicates. Loess interpolation for each method is shown as solid line.

b) Boxplot of median TAD size in all replicates of all datasets (analyzed at 40kb) except Lieberman-Aiden and Jin H1-hESC (n=36).

c) Heatmap of the contact matrix of Rao GM12878 replicate H (chr1:153,000,000-155,500,000) at 40kb resolution. Identified TADs are framed in different colors for the various methods.

d) Box plots of the Jaccard Index for concordance of TAD boundaries between sample replicates of all datasets with at least 2 replicates (n=39).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

Methods for Hi-C data analysis used in this comparison.

	Method	Availability	Programming language
Chromatin interactions	Fit-Hi-C15	noble.gs.washington.edu/proj/fit-hi-c	Python
	GOTHiC16	http://bioconductor.org/packages/release/bioc/html/GOTHiC.html	R
	HOMER17	homer.ucsd.edu/homer/download.html	Perl, R
	HIPPIE18	wanglab.pcbi.upenn.edu/hippie	Python, Perl, R
	diffHic19	https://bioconductor.org/packages/release/bioc/html/diffHic.html	R, Python
	HiCCUPS9,14*	github.com/theaidenlab/juicer/wiki/Download	Java
TADs	HiCseg20	https://cran.r-project.org/web/packages/HiCseg/index.html	R
	TADbit21	github.com/3DGenomes/TADbit	Python
	DomainCaller5	http://chromosome.sdsc.edu/mouse/hi-c/download.html	Matlab, Perl
	InsulationScore22	github.com/dekkerlab/crane-nature-2015	Perl
	Arrowhead9,14*	github.com/theaidenlab/juicer/wiki/Download	Java
	TADtree23	compbio.cs.brown.edu/projects/tadtree/	Python
	Armatus24	github.com/kingsfordgroup/armatus	C++

* HiCCUPS and Arrowhead are the algorithms for interaction and TAD calling of the Juicer software suite.

Europe PMC Funders Author Manuscripts

Table 2

Hi-C experimental data.

Study	Cell type				Hi-C Protocol[b]	Restriction Enzyme				Read length (bp)	Median read count (per replicate, in millions)	Resolution (kb)[d]	N° of replicate samples
	LCL[a]	H1-hESC	IMR90	Fly Embryo		HindIII (6bp)	NcoI (6bp)	DpnII (4bp)	MboI (4bp)				
Lieberman-Aiden2	✔				Dilution	✔	✔			76	11	1000	4
Sexton7				✔	Simplified			✔		36	362	40	1
Dixon 20125		✔	✔		Dilution	✔				36-100[c]	328	40	4
Jin8		✔	✔		Dilution	✔				36-50[c]	440	5-40	7
Rao9	✔		✔		In situ			✔	✔	101	240	5-40	23
Dixon 201525		✔			Dilution	✔				36-50[c]	999	5-40	2

[a] LCL: lymphoblastoid cell lines (i.e., GM06990 in Lieberman-Aiden and GM12878 in Rao)

[b] Dilution, simplified, and in-situ refer to the Hi-C protocols presented in Lieberman-Aiden et al., (2009), Sexton et al, (2012), and Rao et al.(2014), respectively

[c] Samples have been sequenced with different read length in the same study

[d] Resolution refers to the resolution used in this comparison. In the case of two values, the first refers to the resolution used for chromatin interactions, the second for TADs.