A critical assessment of topologically associating domain prediction tools

Rola Dali and Mathieu Blanchette*

School of Computer Science and McGill Centre for Bioinformatics, McGill University, Montreal, Canada

ABSTRACT

Topologically associating domains (TADs) have been proposed to be the basic unit of chromosome folding and have been shown to play key roles in genome organization and gene regulation. Several different tools are available for TAD prediction, but their properties have never been thoroughly assessed. In this manuscript, we compare the output of seven different TAD prediction tools on two published Hi-C data sets. TAD predictions varied greatly between tools in number, size distribution and other biological properties. Assessed against a manual annotation of TADs, individual TAD boundary predictions were found to be quite reliable, but their assembly into complete TAD structures was much less so. In addition, many tools were sensitive to sequencing depth and resolution of the interaction frequency matrix. This manuscript provides users and designers of TAD prediction tools with information that will help guide the choice of tools and the interpretation of their predictions.

INTRODUCTION

Mammalian genomes are composed of roughly 3 billion base pairs that code for the instructions needed for cellular function. Stretched end to end, the genome is equivalent to almost two meters in length; yet, it is housed in the cell nucleus whose diameter is at the micrometer scale. To achieve such level of compaction, while still maintaining accessibility to replication and transcription machinery, a multi-layered genome organization structure is essential (1,2). Chromatin proximity ligation techniques, also known as chromosome conformation capture (3C) derivatives, have helped us understand genomic organization by identifying chromatin contacts at an unprecedented scale (3–11). Modified 3C techniques that make use of high-throughput sequencing, such as Hi-C (8), have been recently used to understand cellular genome organization in different cell types, from different organisms including humans and mice (8,9,12–14), plants (15,16), fruit flies (17,18), yeast (19) and bacteria (20), among others (21–24).

Using Hi-C, Lieberman *et al*. reported that the interaction matrices of chromosomes displayed a checkerboard-like pattern suggesting that the genome is divided into two compartments, labelled A (active) and B (inactive). At a higher resolution, compartments are made of megabase-sized domains termed topologically associating domains (TADs) (12,18,25). TADs were originally defined as blocks of chromatin that interact more frequently within themselves than with neighboring regions and that can be identified by visual inspection of the interaction frequency matrix (12). TADs have been identified in several species and have often been found to be conserved among many cell types (12,18,25). TAD boundaries were found to be enriched in CTCF and cohesin binding sites, as well as transcription start sites, histone marks of active chromatin, housekeeping genes and short interspersed nuclear elements (26,27). By compartmentalizing the genome, TADs are fundamental, not only for genome organization and compression but also for genomic function (27). Increasing evidence suggests that TADs play important roles in DNA replication (28) and in modulating transcriptional regulation (29). TAD disruption has also been shown to link to disease (30–33).

The partitioning of the genome in TADs is hierarchical in nature, with TADs often being composed of several sub-TADs (34) and loops (14). TADs also organize in higher-level structures that have been termed meta-TADs (35). There currently is no well-defined biological delineation between these different levels of organization (although see (27) for efforts in that direction). For that reason, and because most of the existing TAD prediction tools (see below) do not distinguish between the different levels of TADs, we use the term 'TAD' to refer to all levels of the TAD hierarchy.

Many different algorithms have been proposed for identifying TADs from Hi-C data (12,14,18,36–43), although not all of their software implementations are readily available. Currently available tools use different algorithms to detect TADs. Some programs allow for hierarchical TAD identification (43) or overlapping TAD boundaries (14), while others do not. To date, these algorithms have not been benchmarked or compared to each other, leaving potential users in the dark about their properties.

---

*To whom correspondence should be addressed. Tel: +1 514 398 5209; Fax: +1 514 398 3883; Email: blanchem@cs.mcgill.ca

Here, we test seven different TAD identification tools and assess a number of biological and statistical properties of their predictions, at different interaction frequency resolutions and sequencing depths. We find that tools produce quite different TAD predictions that vary in TAD size and number but also other biological properties such as relation to CTCF binding sites. Such variability and sensitivity need to be taken into consideration by users in order to properly interpret their data.

MATERIALS AND METHODS

Input data

GM12878 Hi-C data from Rao *et al*. (14) were downloaded from Geo GSE63525 (Experiments HIC001 to HIC029 (replicate 1), produced with the MboI restriction enzyme). H1 ESC Hi-C data from Dixon *et al*. (44) were downloaded from Geo GSM1267196 and GSM1267197 (HindIII restriction enzyme). The downloaded SRA files were converted to fastq using the SRA Toolkit. Paired reads were mapped to the human genome hg19 using HICUP (http://www.bioinformatics.babraham.ac.uk/projects/hicup/) resulting in more than 1.58 billion filtered paired reads. Lower coverage data sets were obtained by downsampling the GM12878 data set to ~100 million and 500 million paired reads, in order to study the effect of sequencing coverage on TAD predictions. Raw contact matrices were generated using the Homer Hi-C pipeline at 25 kb, 50 kb and 100 kb (45) and were used as input to the TAD detection tools. Many tools expected positive integer values (read pair counts) as input, which prevented us from using normalized matrices in our analysis.

GM12878 CTCF ChIP-Seq narrow peak bed file was downloaded from ENCODE (46), under accession code: ENCFF002CDP, last edited in June 2014. H1 ESC CTCF peaks were downloaded from ENCODE (wgEncodeAwgTfbsBroadH1hescCtcfUniPk.narrowPeak).

Tool usage

Tools were downloaded from the providers listed in Table 1. Tools were used based on the parameters recommended by the provider, with the exception of HiCSeg and Armatus. HiCSeg performed better when using the Poisson distribution instead of the recommended Gaussian distribution for our data set. Armatus results were very sensitive to parameters gamma and step size s. Several different *gamma* and *s* values were tested for each condition; gamma was tested from 0.005 to 1 at increments of 0.005 with $s = 0.005$. The value of gamma that produced TADs closest to the number of TADs predicted by other tools was used (see Supplementary Data). TopDom produced TAD annotations that divide genomic regions to 'domain', 'boundaries' and 'gaps'. To compare with other tools, only regions labeled 'domain' were used in this analysis. TADtree produced different result files and a file named proportion_duplicates.txt. The authors recommend choosing the result file where duplication level is 1–2%. For simplicity, and to avoid subjectivity, we used the first TAD file with duplication level > 0. All parameters used with the tools are available in Supplementary Data.

With the exception of the Arrowhead algorithm from Juicer which used as input a specialized .hic format, all other tools used different formats of binned Hi-C interaction matrices. TAD predictions from each tool were visualized using HiCPlotter (47).

Data set coverage

Three different GM12878 data set sequencing depths were used to assess tool performance; the full data set, containing over 1.5 billion interactions (labeled '1.5B'), a randomly downsampled data set with roughly 500 million interactions (labeled '500M') and another with around 100 million interactions (labeled '100M').

Most tools were able to call TADs on all conditions; however, TADtree was very slow at 25 kb that prevented its inclusion in the analysis. Arrowhead gave a 'sparse data set' warning at 100 million and 500 million data set depth. In fact, it performed poorly at 100 million reads, calling less than a total of 12 TADs across all three resolutions. Using the default value of its gamma parameter, Armatus called very different numbers of TADs across the nine different resolution/coverage conditions. Gamma had to be optimized as a function of both resolution and data set depth to result in TAD predictions similar in number to those produced by other tools.

Manual annotation

TADs were manually traced on both GM12878 and hESC interaction maps from the full data set at 50 kb resolution for regions 40–45 mb of 10 different, randomly chosen, chromosomes (chr2, chr3, chr4, chr5, chr6, chr7, chr12, chr18, chr20 and chr22). Briefly, interaction maps of the regions of interest were plotted using HiCplotter. In Adobe Illustrator, dotted squares were manually traced around visually identifiable TADs on the interaction map plots. Regions annotated as TADs had the following properties: (i) sharp visual contrast between within and across TAD interaction frequencies, over the entire TAD region; (ii) minimum size of 250 kb. To give all tools an equal chance, we created a dense set of TAD annotations that included any identifiable TAD structure. For example, if two potential TADs were overlapping, both were retained, irrespective of whether one had stronger visual support than the other. TAD boundaries were allowed to overlap or be nested, as long as there is a clearly traceable square along the diagonal. Bed files with TAD ranges were manually created and used for tool comparison.

CTCF enrichment around boundaries

CTCF distribution around predicted tool boundaries was calculated in R using the 'EnrichedHeatmap' Bioconductor package (48). Data were binned at 10 kb.

Computational resources

Data analysis was mainly carried out in R and plots were generated using ggplot2 (49). Tools were tested on the high performance computing McGill University cluster (Linux

Table 1. Properties of selected tools for TAD detection

Tool	Programming language (dependencies)	Input	Functionality	Hierarchical TAD structure	Methodology	Publication and Link to download
Armatus	C++ (C++11, boost)	$n \times n$ interaction matrix -gzipped	Calls TADs on raw and normalized matrices	No	Consensus of multi-resolution TAD analysis obtained from dynamic programming algorithm	(37) https://github.com/ kingsfordgroup/ armatus/releases
Arrowhead	Java (none)	hic format (Not from interaction matrix)	Calls TADs-part of JuiceBox which has several other functionalities	Yes	Identification of domain corners using heuristics	(14) http://www.aidenlab.org/ commandlinetools/
DomainCaller	perl, matlab (none)	$n \times (n+3)$ interaction matrix	Calls TADs	No	Directionality index combined with hidden Markov model	(12) http://chromosome. sdsc.edu/mouse/hi-c/ download.html
HiCSeg	R, C	$n \times n$ interaction matrix	Calls TADs	No	Maximum likelihood segmentation using dynamic programming algorithm	(38) https://cran.r-project. org/web/packages/ HiCseg/index.html
TADbit	Python (Scipy, numpy, matplotlib, imp, mcl, chimera)	$n \times n$ interaction matrix	Mpas reads, normalizes and plots IF matrices. Calls and plots TADs. TAD clustering. 3D modelling	No	BIC-penalized breakpoint detection algorithm based on probabilistic interaction frequency model	bioRxiv 036764 https://github.com/ 3DGenomes/tadbit/
TADtree	Python (Scipy, numpy)	$n \times n$ interaction matrix	Calls TADs	Yes	Hierarchical segmentation based on empirical distributions of interaction frequencies within TAD	(43) http: //compbio.cs.brown.edu/ projects/tadtree/
TopDom	R	$n \times (n + 3)$ interaction matrix	Calls TADs	No	Heuristic for breakpoint detection combined with statistical filtering of false-positives	(41) http://zhoulab.usc. edu/TopDom/

version 2.6.32-504.30.3.el6.x86_64; CentOS release 6.6 (Final)). Scripts for installation and usage have been attached in Supplementary Data for guidance. Commands may vary based on the platform used.

RESULTS

We compared the properties of TAD predictions made by different programs on a Hi-C data set generated by Rao *et al.* (14) from GM12878 cells, which is one of the publicly available data sets with the greatest depth of sequencing coverage. We also repeated our analyses on a second Hi-C data set generated by Dixon *et al.* (44) from human embryonic stem cells, and obtained generally congruent results (with some exceptions; see below) that are presented in Supplementary Data (Supplementary Figure S1). Interaction frequency (IF) files were provided as input to seven different TAD prediction programs (Table 1): Armatus (37), Arrowhead (from the Juicer toolbox) (14), DomainCaller (12), HiCSeg (38), TADbit (bioRxiv 036764), TADTree (43) and TopDom (41). These include the most commonly used tools for which publicly available software was available at the time of our study. While other tools exist (HubPredictor (50), calTADs (Zenodo 59188), HiCdatR (51), Matryoshka (*biorxiv* 032953)), we were unable to include them in this study because of difficulties of installation, execution or interpretation of the output, or because they required external data (e.g. Chip-Seq data) as input. Importantly, the

seven tools we used differ in the specific type of signal being sought, and in assumptions made about the scale and organization of TADs (Table 1). Nonetheless, all produced outputs that can be interpreted as elements of the TAD hierarchy. Default parameters were used for all tools except for HiCSeg and Armatus (see Materials and Methods). The impact of sequencing depth was investigated by analyzing both the complete data set (∼1.5 billion read pairs) and down-sampled data sets of ∼500 million read pairs and 100 million read pairs (Figure 1). To study the impact of the resolution (bin size) of the IF matrices provided as input, we generated IF matrices at 25 kb, 50 kb and 100 kb resolutions. Excessive running time or memory consumption prevented us from performing analyses at higher resolution, as only Arrowhead and DomainCaller could be run to completion at 5 kb resolution on a server with 23 GB of RAM. With increasing resources and more powerful infrastructure, it will become possible to execute TAD analysis at higher resolution.

TAD predictions vary significantly between tools

The number (Figure 2A and Supplementary Figure S1A) and size distribution (Figure 2B, C and Supplementary Figure S1B) of the TADs predicted by each tool vary significantly, and for many tools are highly dependant on sequencing depth, resolution or both. At 50 kb resolution and 500 million read pairs of sequencing coverage (a typical Hi-
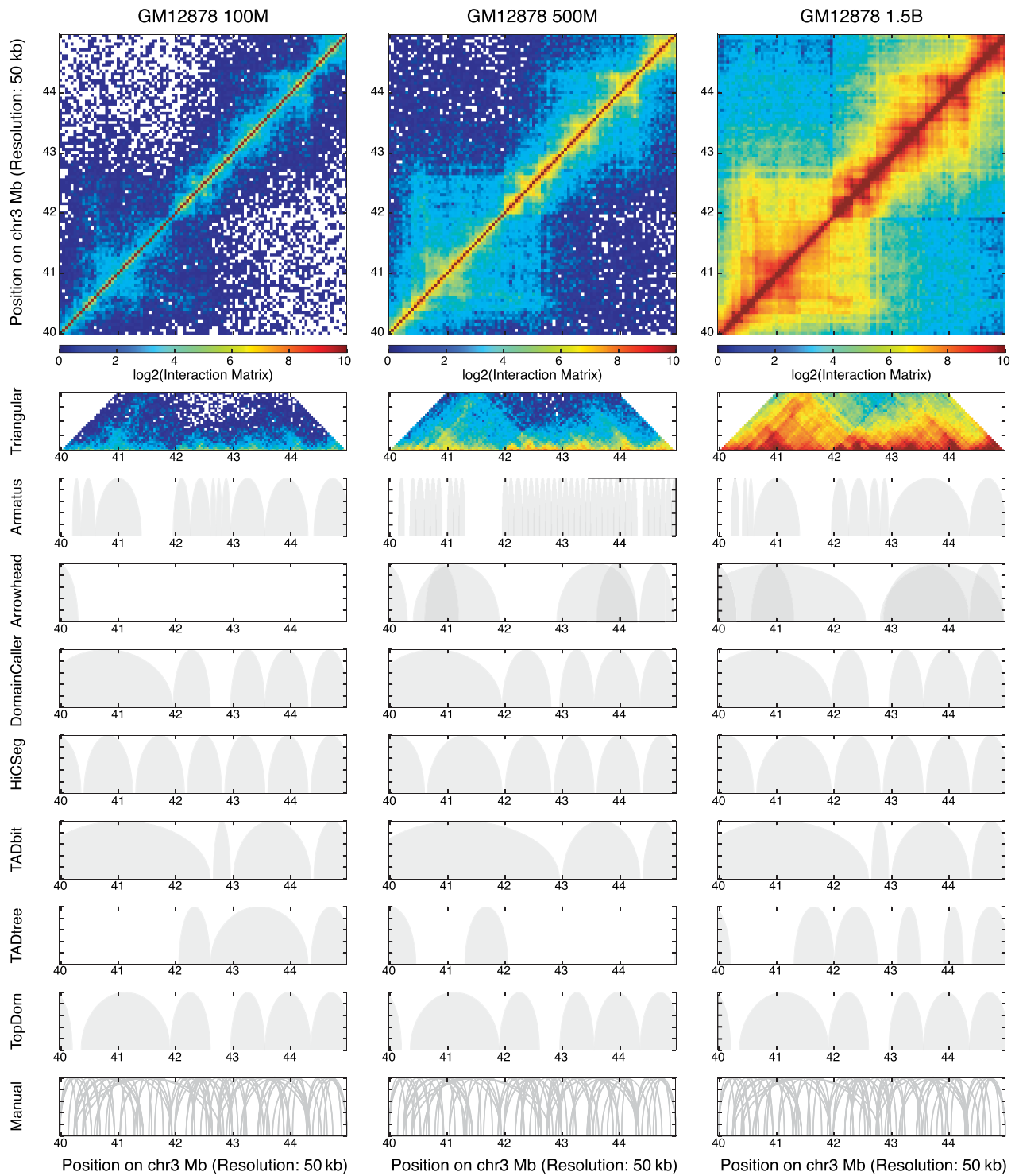
Figure 1. Topologically associating domains (TAD) detection using seven tools at three levels of sequencing depth, from the Rao *et al.* GM12878 Hi-C data set, at 50 kb resolution, for chr3:40–45 Mb.
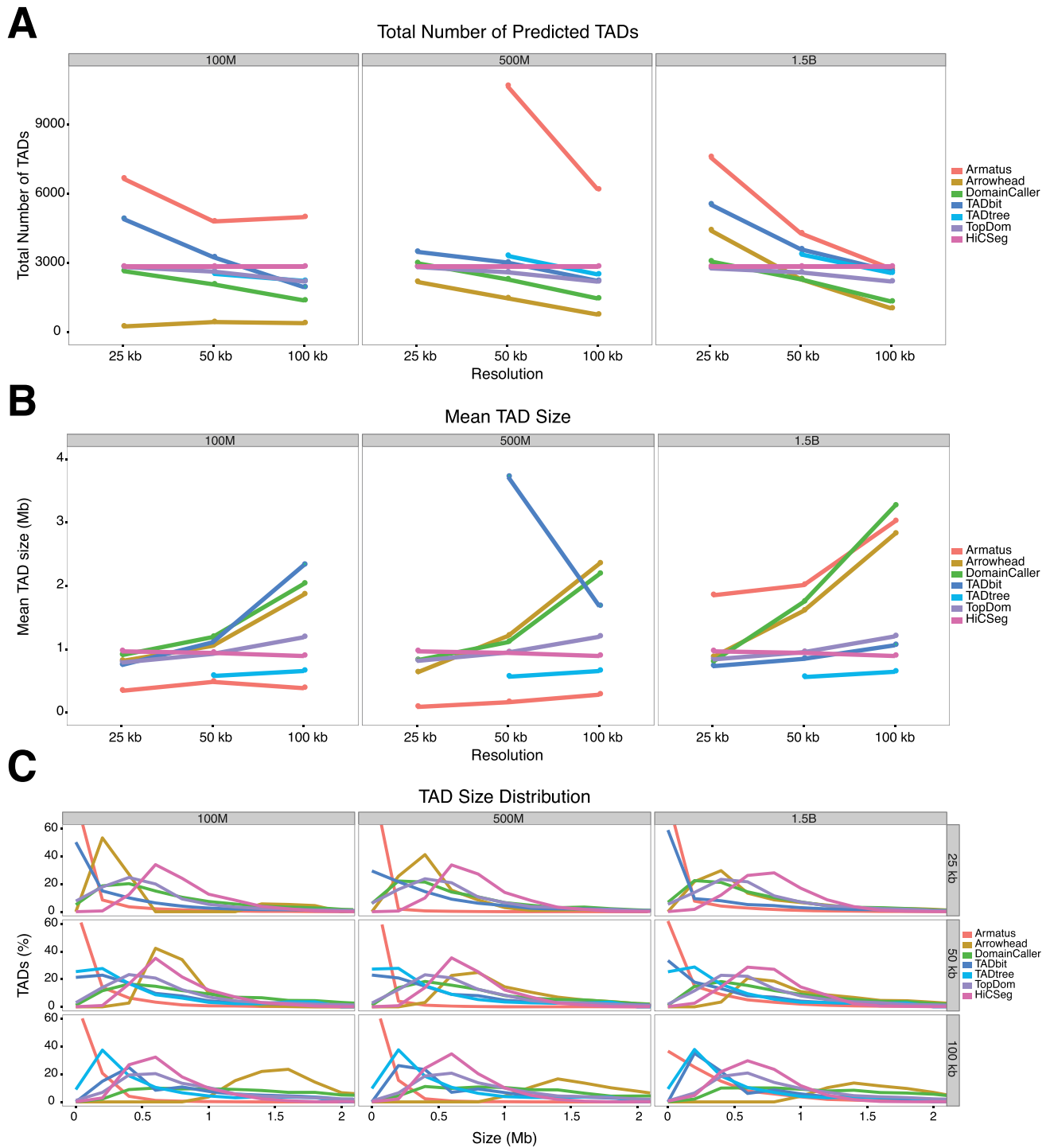
Figure 2. TAD prediction statistics of seven TAD prediction tools as function of sequencing depth and bin resolution. (**A**) Number of predicted TADs at sequencing depth of 100 million, 500 million and over 1.5 billion read pairs at resolutions of 25 kb, 50 kb and 100 kb. (**B**) Mean size of predicted TADs at each condition. (**C**) TAD size distribution for each tool.

C experiment sequencing depth and resolution), the average TAD sizes ranged from 215 kb for Armatus to 1.2 Mb for Arrowhead, with most tools' averages around 1 Mb. In some cases, TADs covered most of the genome (HiCSeg, TADbit: ∼99% genome coverage), while others left much of the genome outside TADs (TADtree: ∼51% genome coverage). Most tools tend to predict fewer but larger TADs at lower resolution, with Arrowhead and DomainCaller more than doubling the mean TAD size when the resolution is increased from 25 kb to 100 kb. The larger number of TAD predictions made by these two tools at 25 kb resolution is not the result of a drastic increase in tiny TAD predictions, but of a more general shift toward a finer partitioning of the genome into TADs. These two tools are also quite sensitive to sequencing depth, with mean TAD size increasing by more than 50% between our low and high coverage data sets. Only HiCSeg, TopDom, and TADtree make predictions that are relatively robust to resolution and sequence depth.

TAD prediction tools have to trade-off high within-TAD IFs for low between-TAD IFs, and they differ in how this trade-off is implemented (Figure 3 and Supplementary Figure S1C). The two tools that tend to predict the smallest TADs, Armatus and TADtree exhibit the highest within-TAD IF values, but also the highest between-TAD IF. Arrowhead, TADbit, TopDom and DomainCaller all exhibit similar within- and between-TAD IF distributions, with the latter obtaining the largest difference between the two, i.e. yielding TADs that are slightly denser and/or better separated than its competitors on this data set (*P*-value $< 2 \times 10^{-16}$, Kolmogorov–Smirnov test).

To check concordance among tools, the number of common TAD boundaries was computed. Some tools had more than 75% of their called boundaries detected by at least one other tool (DomainCaller, TopDom, HiCSeg) while others made more unique predictions (Figure 4).

TAD concordance with manual annotation

Concordance among tools may be confounded by the use of similar algorithms underlying the tools and is, perhaps, not a precise measure of tool accuracy. To compare the tools more empirically, we manually annotated domains (henceforth referred to as TADs, although they include domains at all levels of the TAD hierarchy) in 10 regions of 5 Mb each, chosen from different chromosomes (see Materials and Methods). TADs were manually traced on interaction maps from the data set with the highest sequencing depth, at 50 kb resolution (Figure 5A). The manual annotation included any visually identifiable TAD at least 250 kb in size. Manually annotated TADs display a wide range of sizes from a few hundred kilobases to a couple of megabases, with a mean size of around 650 kb (Figure 5B and Supplementary Figure S1D). Importantly, our manual TAD annotation contains many overlapping TADs, often nested but sometimes not. Because of this, it includes many more TADs (11 TADs/Mb) than most tools predict (0.5 TADs/Mb to 2 TADs/Mb). As such, it should be regarded as a comprehensive list of all TAD predictions supported by visual inspection.

TAD prediction concordance with the manual annotation was evaluated at two levels: (i) individual TAD boundary prediction (Figure 6A, Supplementary Figure S1E), i.e. level of agreement between the sets of positions identified as TAD boundaries, and (ii) complete TAD prediction (Figure 6B, Supplementary Figure S1F), where a TAD prediction is deemed correct if both boundaries agree with the manual annotation (within the appropriate tolerance to accommodate different resolutions). For each tool and each of the two modes, we evaluated the positive predictive value (PPV; fraction of predictions made by a tool that matched the manually annotated calls).

Accurate boundary detection does not necessarily translate into accurate TAD prediction, as correctly pairing boundaries is sometimes challenging (see Discussion). Tools generally performed much better at boundary prediction than at full TAD detection, with boundary prediction PPV exceeding 90% for some tools but TAD prediction PPV rarely exceeding 40% (at 50 kb resolution). TopDom, DomainCaller and HiCSeg produce TAD boundary predictions (Figure 6A) that were generally in excellent agreement with our manual annotation and were robust to variation in both resolution and sequencing coverage. Although one may expect that increased sequencing coverage would in general improve boundary prediction accuracy, this is generally not the case, with this parameter either having no effect or erratic effects on tools' PPV (TADtree, Armatus), or, in the case of TADbit, negatively impacting accuracy.

At the complete TAD level (Figure 6B), the level of agreement with our manual annotation was generally lower for all tools. Front runners are TopDom, DomainCaller and TADbit, as well as TADtree, which seems to make up for its lower accuracy at boundary detection with a seemingly stronger boundary pairing approach. Because of the different levels of tolerance needed to accommodate differences in resolution of the computational predictions and manual annotation, it is difficult to assess the impact of resolution on TAD prediction accuracy. However, we note that DomainCaller's predictions are those that show best agreement with the manual annotation at 25 kb resolution, but they drop in number and agreement at lower resolution. In contrast, TADtree's accuracy ranking improves at lower resolution. Most of these observations were replicated in our manual annotation of TADs in human ESC (Supplementary Figure S1E and F), with the exception of TADbit, which performed significantly worse on the hESC data, and HiCSeq, which fared somewhat better. Overall, the tool that stood out most in terms of its agreement with our manual annotation and its stability with respect to resolution and coverage was TopDom.

CTCF sites are unequally enriched at predicted TAD boundaries

Genomic regions surrounding TAD boundaries have previously been found to be enriched with CTCF binding sites (12). We measured the enrichment of CTCF sites (46) around the predicted TAD boundaries (Figure 7 and Supplementary Figures S1G and S2). Whereas all tools exhibited enrichment of CTCF peaks around their predicted boundaries, this enrichment was strongest for TADtree and
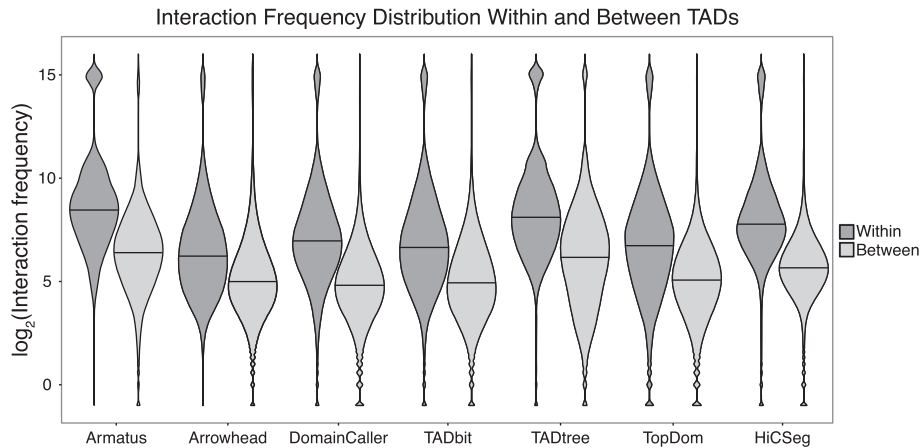
Figure 3. Violin plots of the within-TAD interaction frequency versus between-TAD interaction frequency across different tools (predictions made on the complete GM12878 data set, at 50 kb resolution). Between-TAD interaction frequency corresponds to pairs of bins that are located in adjacent TADs, within a maximum distance of 10 Mb.
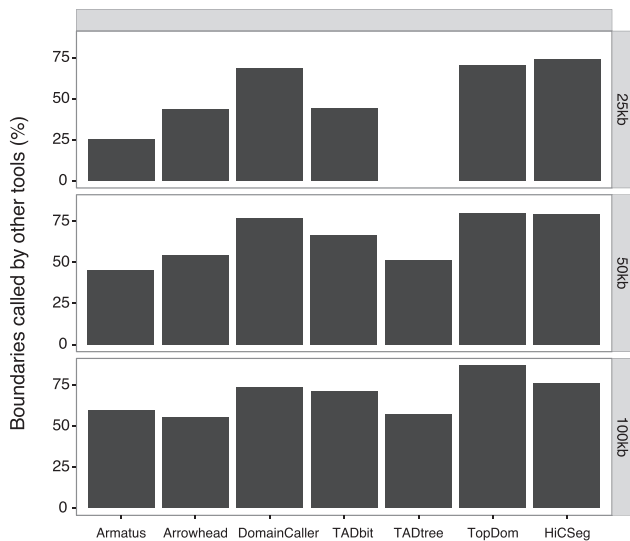

Figure 4. Fraction of TAD boundaries identified by more than one tool (predictions made on the complete GM12878 data set), at three different resolutions.

TopDom (*P*-value < 0.01 (Fisher exact test) over all other tools for the CTCF peak enrichment within 25 kb of a boundary).

Running time and memory consumption

Ease of installation and use, running time and memory requirements vary substantially from tool to tool. Running time ranged from seconds per chromosome (Arrowhead, Armatus, TopDom), to several minutes (DomainCaller, HiCSeg) and even days (TADtree) (Figure 8A). Running time generally scaled linearly with the number of bins per chromosome except for HiCSeg and TADtree, but were largely unaffected by sequencing depth. Memory consumption scaled quadratically with resolution (except for Arrowhead and DomainCaller, for which it was nearly constant). Memory consumption was within 8 GB for most tools for

even the highest resolution analyses (25 kb bins), making them useable on personal computers (Figure 8B). The only significant exception was TADtree, which required in excess of 20 GB of RAM at 25 kb resolution and could not be run to completion. Again, sequencing depth did not significantly affect memory usage of any of the tools.

DISCUSSION

The importance of TADs as a key level of genomic organization is now making consensus in the chromatin structure community, with important roles in gene regulation, genome organization, DNA replication and disease. Although several tools are available to predict TADs from HiC data, there is currently little understanding of the differences among them, and of their sensitivity to various experimental parameters. Here, we compared seven TAD prediction tools, and found that the properties of their predictions differed substantially, potentially affecting conclusions researchers may reach based on them. We argue that the lack of consensus on exactly what a TAD is, combined with the variability in TAD prediction tools, lead to inconsistencies across studies and are an impediment to research in the field. The lack of a unanimously accepted biological or mathematical definition of TADs (although see (27)), combined with the absence of external experimental data that would support or contradict particular TAD predictions, makes a formal benchmarking impossible and even undesirable. As such, this manuscript does not aim to recommend or discourage the use of any particular tool but to describe the current state of the art in TAD detection in order to inform future users and tool developers.

When comparing and assessing TAD predictions made by different tools, it is important to remember that each tool was built based on different assumptions about TADs (size distribution, type of Hi-C signal detected, presence/absence of overlap and/or nesting, etc.), resulting in often fairly discordant predictions. Arguably, tools were created to focus on different aspects of TAD discovery (e.g. Armatus was built to identify 'alternative' domains). One also needs to take into consideration the hierarchical nature of chro-
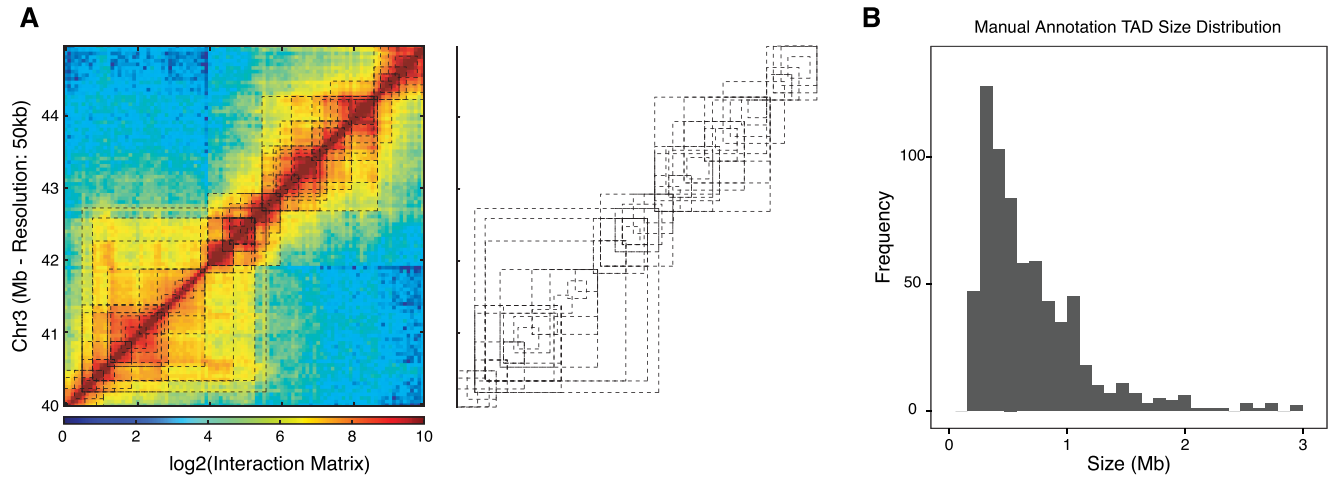
Figure 5. Manual annotation of TADs. TADs were manually traced on interaction maps at 50 kb resolution for 5 Mb regions in the genome (chr2, chr3, chr4, chr5, chr6, chr7, chr12, chr18, chr20 and chr22). (**A**) Example of the manual annotation for chr3. (**B**) TAD size distribution of manually annotated TADs.
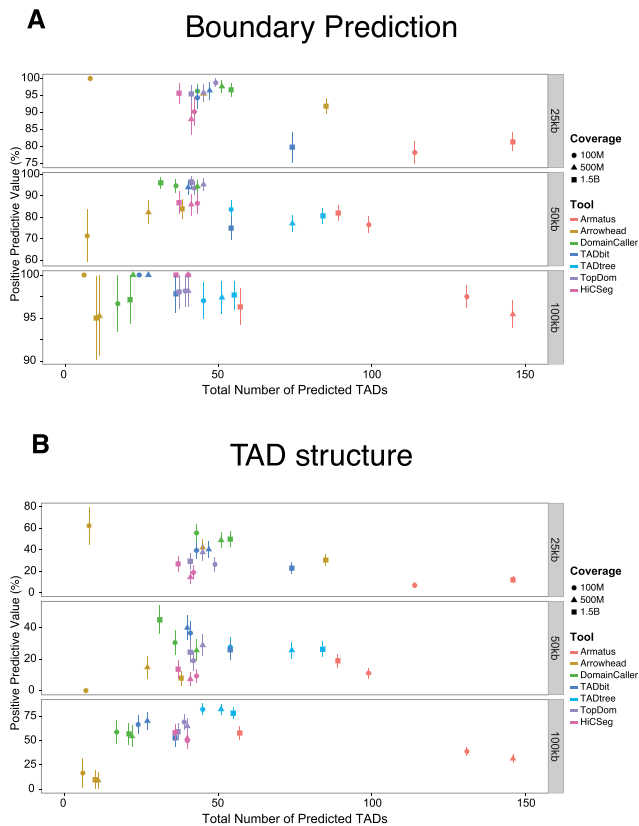

Figure 6. Tool prediction concordance with manual annotation. (**A**) Positive predictive value for each tool for boundary detection as function of predicted TAD numbers at three different resolutions. (**B**) Positive predictive value for each tool for complete TAD structure detection as function of predicted TAD numbers at three different resolutions. Error bars correspond to one standard deviation of the estimates. For predictions made at 25 kb resolution, a 25 kb margin was used to compare against the manual annotation (which was produced at 50 kb resolution). At 50 kb resolution, no tolerance was allowed. At 100 kb resolution, 50 kb tolerance was used. It is worth noting that Arrowhead did not predict many TADs in our annotated regions (see Materials and Methods), resulting in large confidence intervals.
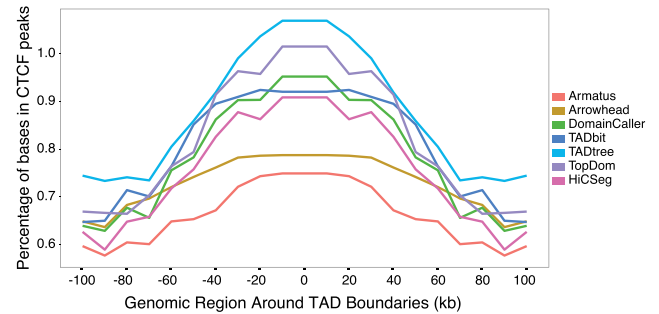

Figure 7. GM12878 CTCF ChIP-Seq peak distribution around predicted boundaries detected at 50 kb resolution. CTCF ChIP-Seq peak enrichment in a 100 kb window around TAD boundaries predicted at 50 kb resolution.

mosome interaction structures, from loops to sub-TADs, TADs, meta-TADs and sub-compartments, none of which are clearly defined. Indeed, few tools clearly specify the type of domains they aim to identify. Most are generically described as TAD predictors, but Arrowhead was reported to identify interaction domains generally smaller than TADs (14), while TADtree predicts entire TAD hierarchies (43). However, the detected domain level of many tools appears to be dependent on the parameters of the both Hi-C data (coverage, resolution and noise) and of the tool itself.

Comparative studies of the type presented here always come with caveats that one should keep in mind. First, each tool has a set of parameters that can be modified to generate different outputs; here, we used the parameters suggested by the authors, with few exceptions (see Materials and Methods). Second, some tools (TADbit and Arrowhead) assign prediction scores to each predicted TAD, which could allow further TAD prediction filtering; those were not used here. Finally, our manual TAD annotation is certainly imperfect and its underlying assumptions do not fit all tools equally well; we nonetheless argue that it sheds valuable light on each tool's predictions' properties.
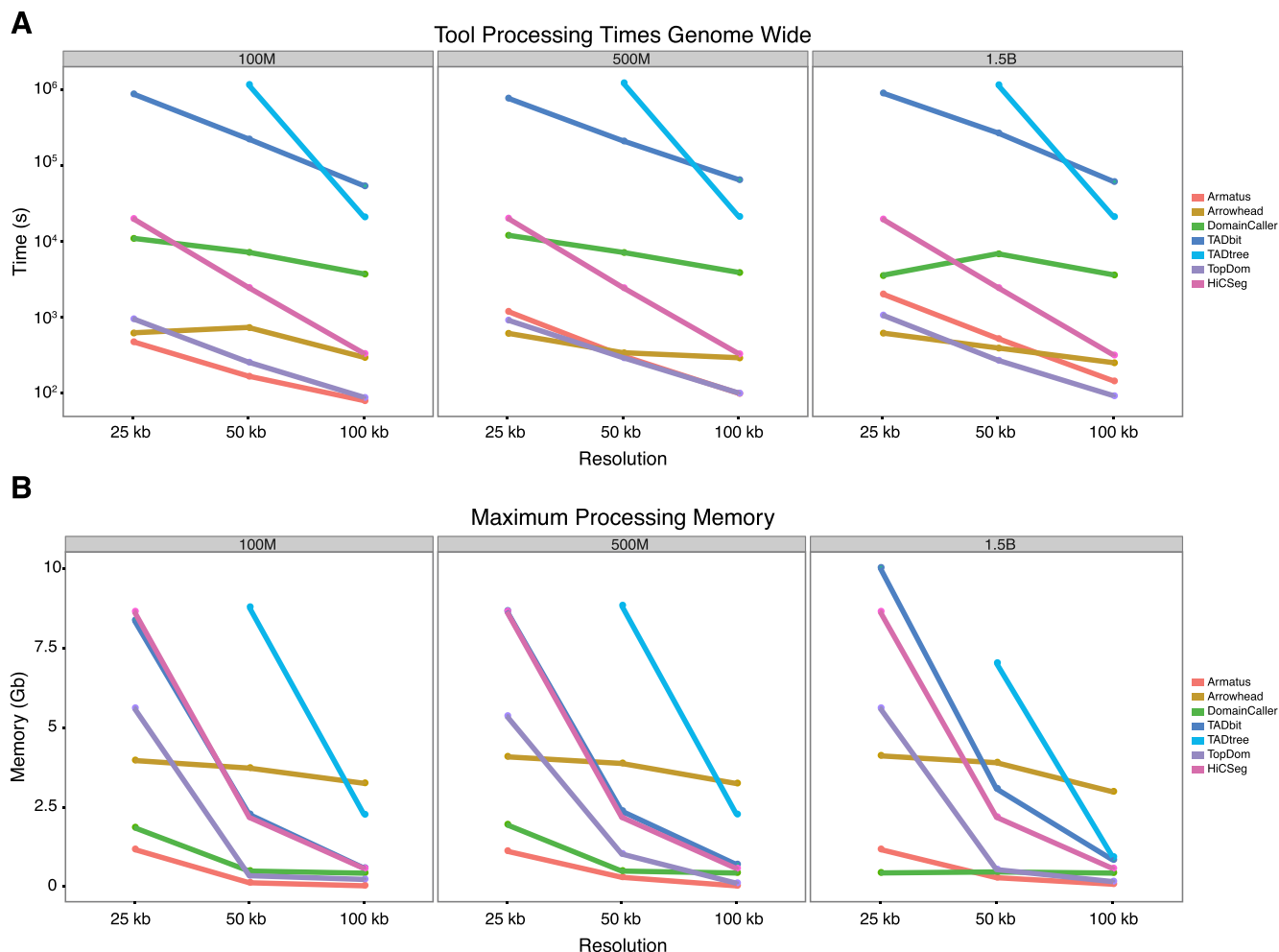
A

B

Figure 8. Running time and memory consumption. (**A**) Processing time required to call TADs at different resolutions and sequencing depths. (**B**) Memory required to call TADs at each condition.

Although rarely discussed, ease of installation and use are important factors affecting tool adoption. Tools varied in the amount of work needed for their installation, ranging from almost none (Arrowhead and TopDom) to the requirement to install various system dependencies (Table 1). All tools and dependencies were freely available with the exception of DomainCaller, which is free but uses Matlab. Tools also differed in the number of parameters that need to be set by the user, ranging from one (Arrowhead and TAD-bit) to six (TADtree). Although all tools provide default values for these parameters, their detailed impact on the predictions and their suitability at different resolutions and sequencing coverage are rarely described. Input format for each tool varied as well (Table 1). While most tools used different forms of interaction matrices, Arrowhead uses a specialized .hic format generated specifically by Juicer (14,52). Standardization of Hi-C interaction frequency matrix formats, as well as standardization of TAD prediction outputs, would make tools easier to use for non-bioinformatics experts. Finally, running time and memory requirements, as well as code stability, were, in some cases, impediments to their utilization, especially for higher-resolution analyses.

Differences among TAD prediction tools

Tools resulted in different numbers of TADs predicted and different mean TAD sizes (Figure 2A, B, Supplementary Figure S1A and B). A closer look at the TAD size distribution predicted by each tool showed TAD predictions peaking at different sizes for each tool (Figure 2C), suggesting that tools specialize in detecting TADs at particular sizes. This creates a restriction in the ability of any particular tool to capture the full range of TADs found in a cell. Furthermore, few tools provide easy to use parameters that can help users control the distribution of TAD sizes, with the exception of Armatus and HiCSeg.

Visual inspection of predicted TAD boundaries showed that no tool detected all TADs perfectly and that the complex nested structure of TADs was in general poorly captured. It is clear though that different tools detect different types of boundaries with different success rates. Crucially, although the prediction of individual boundaries was generally done with good accuracy, the pairing of boundaries to produce TAD structures was often deficient. Looking at the manual annotation, we observed that predicting TAD structures is not as simple as detecting TAD boundaries and

joining consecutive boundaries into TADs. Common TAD structure prediction errors included calling smaller TADs by joining consecutive TAD boundaries that may belong to larger meta-TADs or by calling large TADs by joining two consecutive TADs by missing the dividing boundary; resulting in predicted TAD structures with large variations in interaction intensities within the TAD. TAD tools also sometimes missed TAD boundaries by a bin or two since some tools called a TAD boundary at the first sign of change in interaction intensity (scanning the matrix from left to right) without evaluating whether there is a stronger boundary a few bins away. Another common error was to call a boundary correctly but to label it as a TAD start boundary instead of a TAD end boundary or vice versa.

Tools generally had low sensitivity, often picking up less than 10% of manually annotated TAD structures. In fact, almost 25% of the manually annotated boundaries never got detected by any of the tools. These low detection rates might be due to the assumption most tools make about non-overlapping TADs. Our manual annotation suggests that many TADs overlap by about 100–150 kb, which makes many TADs and TAD boundaries invisible to tools that disallow overlaps. While overlapping TADs should not exist within a single haplotype of a single cell, Hi-C data capture a population average where chromatin can exist in different forms in different cells. Thus, it would be useful for tools to have the flexibility of calling TADs across the genome with more flexible positional restrictions regarding overlap and nesting.

Robustness to resolution and sequencing depth

Most TAD prediction tools take as input an interaction frequency matrix where read counts are binned at a certain fixed resolution (e.g. 50 kb). Resolution and sequencing depth are tightly linked, with higher coverage data sets allowing, in principle, analyses to be performed at higher resolution. While sequencing depth is not under the full control of the researcher due to cost limitations and to the many Hi-C artefacts that get filtered out of the raw sequenced reads, resolution is a parameter that can be controlled more easily. In an ideal world, TAD predictions should be robust to variation in sequencing depth. Being large-scale structures, most TADs are visually detectable even in low-coverage data sets. Increased sequencing depth should result in finer positioning of the TAD boundaries, and perhaps in an increased ability to detect 'weak' or small TADs, but not in a systematic shift in TAD properties. Only a few tools exhibit this type of robustness (DomainCaller, TopDom, HiCSeg). Another desirable property would be the robustness to variation in the resolution of the interaction frequency matrix provided as input. Reducing the resolution tends to increase the contrast between pairs of regions that belong to the same TAD and those that do not, which often helps in TAD identification, at the cost of a loss in terms of accuracy in the positioning of the boundaries and ability to detect very small TADs. Only TopDom and HiCSeg exhibit a good level of robustness with respect to resolution. However, there is really no reason why binning to reduced resolution should be necessary, as binning is simply destroy-

ing information. Future TAD prediction programs should eventually move away from this pre-processing step.

TAD boundaries and CTCF peaks

TAD boundaries have previously been shown to be enriched in CTCF binding sites ([12]). Indeed, all tools tested showed an enrichment of CTCF ChIP-Seq sites around the boundaries they predicted. Over 20% of boundaries predicted by most tools have at least one CTCF site within 5 kb, higher than what would be expected if TAD boundaries were picked at random (all *P*-values $< 2 \times 10^{-16}$). Nonetheless, there were substantial differences between tools at that level, with some tools' boundary predictions being nearly two times more enriched for CTCF sites compared to others. We observed a weak negative correlation between TAD size and CTCF enrichment, suggesting that perhaps larger TADs do not need CTCF binding to stabilize their structure as much as smaller ones do.

It is not surprising that there is much variation in the predictions made by different TAD detection tools. The community's definition of TADs is still rather vague ([27]), which leaves tool developers with a lot of space for creative freedom widening the gap between outputs from different tools. The hierarchical, nested structure of TADs also complicates matters. The scientific community interested in TADs has been using terms like TADs, meta-TADs, sub-TADs and loops without a clear biological or operational definition of the difference between them, which hinders our understanding of the properties of each subtype and the link between the TAD hierarchy and compartments. As these definitions get refined and a consensus emerges, tools will improve and the differences between tool outputs will be reduced. While Dixon *et al.* ([27]) recently attempted to provide a unified definition of TAD-like structures, proposing that the essential feature of TADs is their stability through mitosis and conservation through cell lineages, that definition is difficult to exploit computationally. A definition that is autonomous to a single experiment and independent of time is needed to guide TAD detection and propel the field forward.

CONCLUSION

TAD prediction is not yet a solved problem. TAD prediction tools make many different assumptions regarding overlap, nesting and size of TAD structure that restrict the tools from predicting the full TAD landscape in cells. Next generations of TAD prediction tools will hopefully relax these assumptions to better capture the full range of TADs. This cannot be done successfully until more precise biological and mathematical definitions of TADs are agreed upon by the community. In the meantime, researchers using these tools need to be aware of the biases and properties of each tool in order to properly interpret their results, and are advised to use more than one tool in order to get a more complete picture of the TAD structures in their data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

REFERENCES

1. Fraser,J., Williamson,I., Bickmore,W.A. and Dostie,J. (2015) An overview of genome organization and how we got there: from FISH to Hi-C. *Microbiol. Mol. Biol. Rev.*, **79**, 347–372.
2. Pombo,A. and Dillon,N. (2015) Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.*, **16**, 245–257.
3. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.
4. Simonis,M., Klous,P., Splinter,E., Moshkin,Y., Willemsen,R., de Wit,E., van Steensel,B. and de Laat,W. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.*, **38**, 1348–1354.
5. Zhao,Z., Tavoosidana,G., Sjolinder,M., Gondor,A., Mariano,P., Wang,S., Kanduri,C., Lezcano,M., Singh Sandhu,K., Singh,U. *et al.* (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.*, **38**, 1341–1347.
6. Hagege,H., Klous,P., Braem,C., Splinter,E., Dekker,J., Cathala,G., de Laat,W. and Forne,T. (2007) Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nat. Protoc.*, **2**, 1722–1733.
7. Dostie,J., Richmond,T.A., Arnaout,R.A., Selzer,R.R., Lee,W.L., Honan,T.A., Rubio,E.D., Krumm,A., Lamb,J., Nusbaum,C. *et al.* (2006) Chromosome Conformation Capture Carbon Copy (5C): Aamassively parallel solution for mapping interactions between genomic elements. *Genome Res.*, **16**, 1299–1309.
8. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
9. Kalhor,R., Tjong,H., Jayathilaka,N., Alber,F. and Chen,L. (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotech.*, **30**, 90–98.
10. Ma,W., Ay,F., Lee,C., Gulsoy,G., Deng,X., Cook,S., Hesson,J., Cavanaugh,C., Ware,C.B., Krumm,A. *et al.* (2015) Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nat. Methods*, **12**, 71–78.
11. de Wit,E. and de Laat,W. (2012) A decade of 3C technologies: insights into nuclear organization. *Genes Dev.*, **26**, 11–24.
12. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
13. Nagano,T., Lubling,Y., Stevens,T.J., Schoenfelder,S., Yaffe,E., Dean,W., Laue,E.D., Tanay,A. and Fraser,P. (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, **502**, 59–64.
14. Rao,S.S.P., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S. *et al.* (2014) A 3D map of the human genome at Kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
15. Liu,C. and Weigel,D. (2015) Chromatin in 3D: progress and prospects for plants. *Genome Biol.*, **16**, 170.
16. Wang,C., Liu,C., Roqueiro,D., Grimm,D., Schwab,R., Becker,C., Lanz,C. and Weigel,D. (2015) Genome-wide analysis of local chromatin packing in Arabidopsis thaliana. *Genome Res.*, **25**, 246–256.
17. Hou,C., Li,L., Qin,Z.S. and Corces,V.G. (2012) Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol. Cell*, **48**, 471–484.
18. Sexton,T., Yaffe,E., Kenigsberg,E., Bantignies,F., Leblanc,B., Hoichman,M., Parrinello,H., Tanay,A. and Cavalli,G. (2012) Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, **148**, 458–472.
19. Duan,Z., Andronescu,M., Schutz,K., McIlwain,S., Kim,Y.J., Lee,C., Shendure,J., Fields,S., Blau,C.A. and Noble,W.S. (2010) A three-dimensional model of the yeast genome. *Nature*, **465**, 363–367.
20. Le,T.B., Imakaev,M.V., Mirny,L.A. and Laub,M.T. (2013) High-resolution mapping of the spatial organization of a bacterial chromosome. *Science*, **342**, 731–734.
21. Ay,F., Bunnik,E.M., Varoquaux,N., Bol,S.M., Prudhomme,J., Vert,J.P., Noble,W.S. and Le Roch,K.G. (2014) Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Res.*, **24**, 974–988.
22. Burton,J.N., Liachko,I., Dunham,M.J. and Shendure,J. (2014) Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3 (Bethesda)*, **4**, 1339–1346.
23. Marbouty,M., Cournac,A., Flot,J.F., Marie-Nelly,H., Mozziconacci,J. and Koszul,R. (2014) Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *eLife*, **3**, e03318.
24. Vietri Rudan,M., Barrington,C., Henderson,S., Ernst,C., Odom,D.T., Tanay,A. and Hadjur,S. (2015) Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.*, **10**, 1297–1309.
25. Nora,E.P., Lajoie,B.R., Schulz,E.G., Giorgetti,L., Okamoto,I., Servant,N., Piolot,T., van Berkum,N.L., Meisig,J., Sedat,J. *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**, 381–385.
26. Ea,V., Baudement,M.O., Lesne,A. and Forne,T. (2015) Contribution of topological domains and loop formation to 3D chromatin organization. *Genes*, **6**, 734–750.
27. Dixon,J.R., Gorkin,D.U. and Ren,B. (2016) Chromatin domains: the unit of chromosome organization. *Mol. Cell*, **62**, 668–680.
28. Pope,B.D., Ryba,T., Dileep,V., Yue,F., Wu,W., Denas,O., Vera,D.L., Wang,Y., Hansen,R.S., Canfield,T.K. *et al.* (2014) Topologically associating domains are stable units of replication-timing regulation. *Nature*, **515**, 402–405.
29. Gorkin,D.U., Leung,D. and Ren,B. (2014) The 3D genome in transcriptional regulation and pluripotency. *Cell Stem Cell*, **14**, 762–775.
30. Lupiáñez,D.G., Spielmann,M. and Mundlos,S. (2016) Breaking TADs: how alterations of chromatin domains result in disease. *Trends Genet.*, **32**, 225–237.
31. Lupianez,D.G., Kraft,K., Heinrich,V., Krawitz,P., Brancati,F., Klopocki,E., Horn,D., Kayserili,H., Opitz,J.M., Laxova,R. *et al.* (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, **161**, 1012–1025.
32. Hnisz,D., Weintraub,A.S., Day,D.S., Valton,A.L., Bak,R.O., Li,C.H., Goldmann,J., Lajoie,B.R., Fan,Z.P., Sigova,A.A. *et al.* (2016) Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, **351**, 1454–1458.
33. Flavahan,W.A., Drier,Y., Liau,B.B., Gillespie,S.M., Venteicher,A.S., Stemmer-Rachamimov,A.O., Suva,M.L. and Bernstein,B.E. (2016) Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, **529**, 110–114.

34. Rocha,P.P., Raviram,R., Bonneau,R. and Skok,J.A. (2015) Breaking TADs: insights into hierarchical genome organization. *Epigenomics*, **7**, 523–526.

35. Fraser,J., Ferrai,C., Chiariello,A.M., Schueler,M., Rito,T., Laudanno,G., Barbieri,M., Moore,B.L., Kraemer,D.C., Aitken,S. *et al.* (2015) Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol. Syst. Biol.*, **11**, 852–852.

36. Chen,Y., Wang,Y., Xuan,Z., Chen,M. and Zhang,M.Q. (2016) De novo deciphering three-dimensional chromatin interaction and topological domains by wavelet transformation of epigenetic profiles. *Nucleic Acids Res.*, **44**, e106.

37. Filippova,D., Patro,R., Duggal,G. and Kingsford,C. (2014) Identification of alternative topological domains in chromatin. *Algorithms Mol. Biol.*, **9**, 1–11.

38. Lévy-Leduc,C., Delattre,M., Mary-Huard,T. and Robin,S. (2014) Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics*, **30**, i386–i392.

39. Libbrecht,M.W., Ay,F., Hoffman,M.M., Gilbert,D.M., Bilmes,J.A. and Noble,W.S. (2015) Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell-type-specific expression. *Genome Res.*, **25**, 544–557.

40. Phillips-Cremins,J.E., Sauria,M.E., Sanyal,A., Gerasimova,T.I., Lajoie,B.R., Bell,J.S., Ong,C.T., Hookway,T.A., Guo,C., Sun,Y. *et al.* (2013) Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, **153**, 1281–1295.

41. Shin,H., Shi,Y., Dai,C., Tjong,H., Gong,K., Alber,F. and Zhou,X.J. (2016) TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.*, **44**, e70.

42. Wang,X. (2016) calTADs: A hierarchical domain caller for Hi-C data. *Zenodo*, doi:10.5281/zenodo.59188.

43. Weinreb,C. and Raphael,B.J. (2015) Identification of hierarchical chromatin domains. *Bioinformatics*, **32**, 1601–1609.

44. Dixon,J.R., Jung,I., Selvaraj,S., Shen,Y., Antosiewicz-Bourget,J.E., Lee,A.Y., Ye,Z., Kim,A., Rajagopal,N., Xie,W. *et al.* (2015) Chromatin architecture reorganization during stem cell differentiation. *Nature*, **518**, 331–336.

45. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H., Glass,C.K. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.

46. Consortium,E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

47. Akdemir,K.C. and Chin,L. (2015) HiCPlotter integrates genomic data with interaction matrices. *Genome Biol.*, **16**, 198–198.

48. Gu,Z. (2016) EnrichedHeatmap: Making Enriched Heatmaps. *R Package Version 1.2.2*, https://github.com/jokergoo/EnrichedHeatmap.

49. Wickham,H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. HSpringer-Verlag, NY.

50. Huang,J., Marco,E., Pinello,L. and Yuan,G.-C. (2015) Predicting chromatin organization using histone marks. *Genome Biol.*, **16**, 1–11.

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

52. Durand,N.C., Robinson,J.T., Shamim,M.S., Machol,I., Mesirov,J.P., Lander,E.S. and Aiden,E.L. (2016) Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.*, **3**, 99–101.