The ABC's of TADs: how to analyze them, and why they matter

Application Note

1. Introduction

Measuring up to several meters when stretched end to end, eukaryotic genomes are extensively packaged in the cell nucleus, employing a variety of methods to achieve the level of chromosomal packaging required (Iyer *et al.,* 2011). This hierarchical three-dimensional (3D) topology, referred to as genome conformation, heavily influences chromosomal structure and function (Bonev and Gavalli, 2016). A key feature of this topology that has received much recent attention are topologically associating domains (TADs), conformational features multiple kilobases in size that allow for DNA to be organized into frequently interacting regions (Dixon *et al*., 2012). TADs influence gene regulation by controlling the physical interactions between regulatory elements with their target genes. Consequentially, TAD disruption can result in altered gene expression, often leading to changes in phenotype (Nora *et al.,* 2012; Iben-Salem *et al*., 2014; Lupianiez *et al*., 2014).

Genome conformation is a highly dynamic process that changes during cell differentiation and cell cycle stage. Despite this observation, TADs appear to be well conserved across many eukaryotes, including drosophila, zebrafish, frogs, chickens, mammals, and some plants, supporting the hypothesis that they are key regulators of genome function (Krefting *et al.,* 2018; Sotelo-Silveria *et al.,* 2018). Other key regulatory features of eukaryotic genomes include CCCTC-binding factor (CTCF) and isochores. CTCF binds to the genome at specific motifs and interacts with cohesion to form genomic loops (Bonev and Gavalli, 2016), which are a crucial mechanism to bring regulatory elements, like enhancers, into proximity to the promoters of target genes. Isochores are known to play a role in gene regulation, as isochore enrichment is associated with gene deserts, regions lacking coding DNA (Arhondakis *et al.,* 2011).

At Dovetail Genomics, we have commercialized chromatin capture methods, such as the Dovetail™ Hi-C Proximity Ligation Assay, to generate near-chromosome length scaffolds for genome assembly. Our Hi-C approaches leverage this natural 3D genome topology to generate long-range paired-read information used to order and orient chromosomally clustered contigs, thereby efficiently creating a comprehensive genomic scaffold (Putman *et al.,* 2016). Through an extensive investigation, we have extended our best-in-class genome assembly services to include a TAD analysis pipeline. Our new service leverages the final assembly generated by our HiRise™ informatics pipeline from Hi-C data to make TAD calls. Dovetail Genomics' TAD analysis

Figure 1 – The Dovetail Genomics TAD analysis workflow.

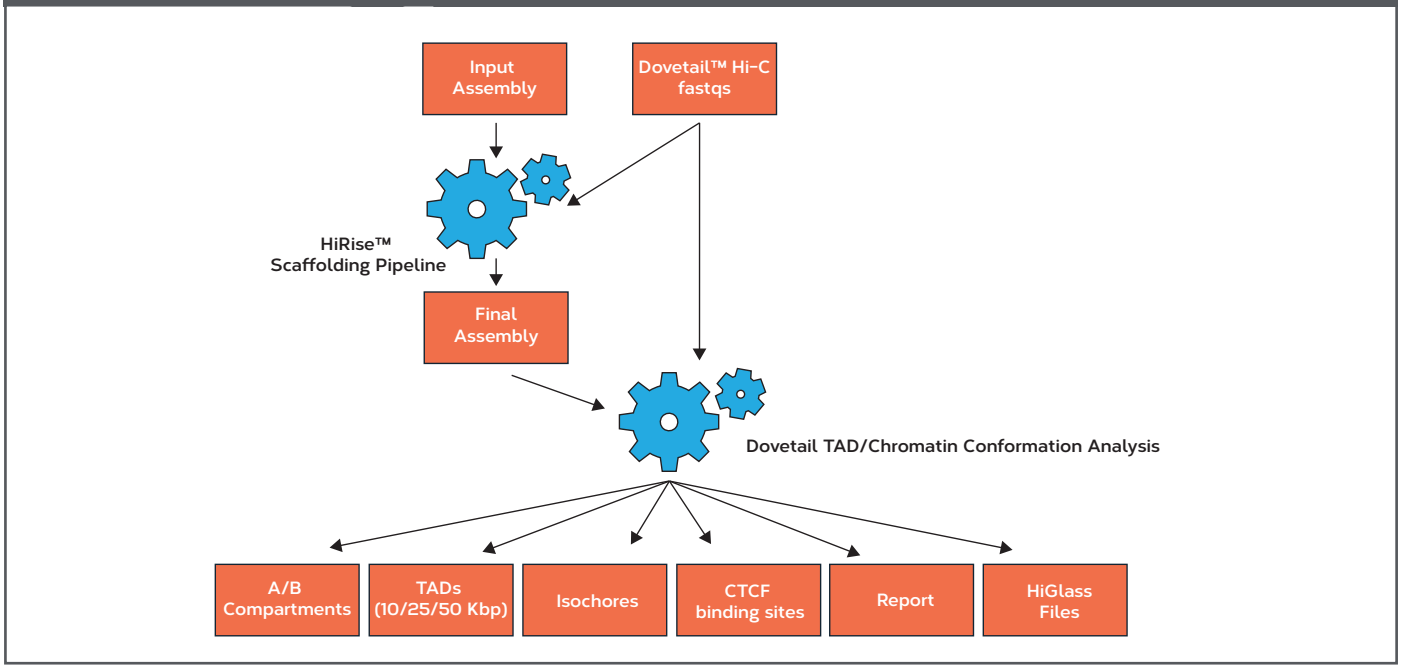


Table 1 – Output files from the Dovetail Genomics TAD Analysis Service including format and brief description of each file type.

File	Format	Description
Manifest	txt	A manifest detailing the contents of each file included in the delivery package
Report.html	html	Summary statistics of the analysis, data processing information, and instructions on HiGlass browser visualization
alignment.bam	bam	File containing sequence alignment data
X.mcool	Mcool	Multiple cooler file containing the Dovetail™ Hi-C matrix of proximity
X.hic	HiC	HiC contact matrix at multiple resolutions in .hic format
X_isochores.bedpe	Bedpe	Output of program which calls isochores – regions of characteristic GC content within a genome
X.multires	multires	Files which can be ingested in HiGlass viewer
Chr_sizes.txt	TSV	Chromosome size file – first column is chromosome name and the second the size of that chromosome
X_AB_compartments.bedpe	Bedpe	A/B compartments from first Eigenvector of contact matrix
X_CTCF_sites.bed	Bed	Predicted CTCF binding sites using Cread
X_TADs_10000.bedpe	Bedpe	Topologically associated domain (TAD) calls using arrowhead at 10,000 bp resolution
X_TADs_25000.bedpe	Bedpe	Topologically associated domain (TAD) calls using arrowhead at 25,000 bp resolution
X_TADs_50000.bedpe	Bedpe	Topologically associated domain (TAD) calls using arrowhead at 50,000 bp resolution

service is ideal for investigating the conformation of your genome of interest **(Figure 1)**.

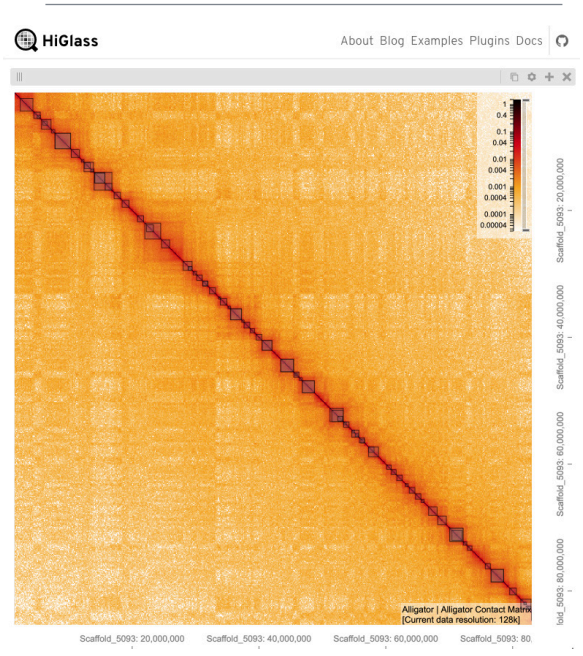
2. Product Description and Deliverables

Our TAD analysis offering is an informatics service that leverages Dovetail™ Hi-C data for a genome-wide TAD analysis **(Figure 1)**. Using the Arrowhead TAD calling pipeline (Rao *et al.*, 2014), our service generates a package of output files that can be used to further explore genomic conformation at a variety of levels **(Table 1)**. The output includes isochore locations (Isofinder, Oliver *et al.*, 2004), CTCF binding sites (Ziebarth *et al.*, 2012), A/B compartment organization, TAD calls at 10 kbp, 25 kbp, and 50 kbp, and a summary report. All files are compatible with HiGlass, an open-source tool used to explore and compare genomic contact matrices, enabling a straightforward view of genomic topological information (Kerpedjiev *et al.*, 2018). The contact matrices are built by using the long-range paired-read property of Hi-C data, where each point in a matrix represents the distance of the proximity-ligation event **(Figure 2)**. Hi-C data, displayed in contact matrices, captures the frequency at which different regions of the genome physically interact, a 2-dimensional view of the hierarchical 3D topology of a genome to be built.

Accompanying the analysis output package is an in-depth guide to loading and viewing your data on HiGlass. Our expected turnaround time for TAD analysis is approximately 2 weeks. Please note that the quality of TAD calling depends on genome features including size, mappability, complexity, and the overall quality and coverage of the assembly.

Figure 2 – Screenshot of a HiGlass view of TADs identified at 50 kbp resolution on the alligator assembly.

The color gradient of the contact map indicates the number of reads supporting each contact point in the matrix, where darker colors indicate greater number of contacts between two sites.



3. Recommendations

The TAD analysis pipeline is engineered to directly plug in to our HiRise™ Genome Assembly Service, thereby requiring only two input files: the genome assembly (fasta) produced by the HiRise™ Assembly Pipeline, and the Hi-C sequencing reads (fastq) that are used to scaffold **(Figure 1)**.

Addition of TAD analysis to an assembly project does not require greater sequencing depth, as our TAD analysis requires 10 million fewer reads than our HiRise™ Genome Assembly to generate 10 kbp bins. Higher resolution is feasible with greater sequencing depth **(Table 2)**. For a detailed description of the topology calling constraints, please refer to **Table 3**. Briefly, the genome size must be less than 10 Gbp and TADs cannot be called efficiently on scaffolds greater than 400 Mbp, due to software limitations. Lastly, at least 40% of the genome must be uniquely mappable in order to provide a complete contact matrix.

Although not the focus of this technical note, we now also offer genome annotations to further enhance downstream biological understandings, an additional feature to our informatics services.

4. Examples

We ran the pipeline on a wide range of species **(Table 4)** to highlight the value and versatility of our new TAD analysis service. In **Figure 3**, we focus on a region of

Table 2 – Sequencing requirements to achieve contact matrices with different resolutions.

The number of read pairs required is calculated for a 1 Gbp genome at 2x150 bp reads. The threshold for number of reads needed to achieve the listed resolution is listed. Note, 80% of the genome must be included in bins with >1,000 single end reads per bin.

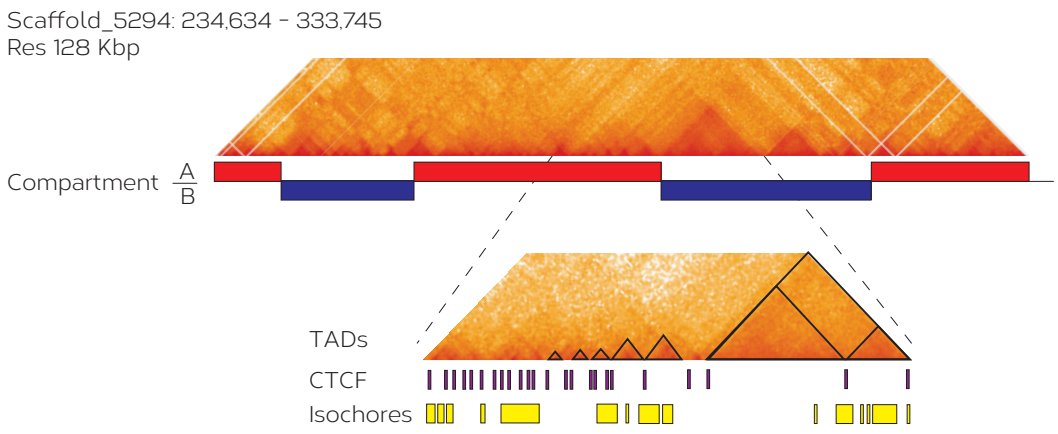
Resolution (Bin Size)	# Read Pairs for 1 Gbp Genome (2x150bp)
500 Bp	1.42 Billion
1 Kbp	760 Million
10 Kbp	90.1 Million(< HiRise™ Requirement of 100M read pairs to scaffold 1 Gb)
100 Kbp	9.53 Million
1 Mbp	972 Thousand
5 Mbp	19.9 Thousand

Table 3 – Assembly metric constraints for the TAD analysis pipeline.

Metric	Requirement
Genome Size	< 10 Gbp
Maximum Scaffold Size	< 400 Mbp
Mappability	> 40% of the genome uniquely mappable
Coverage	90 M read pairs per 1 Gbp

Figure 3 – Dovetail TAD analysis on the alligator genome.

The top contact map depicts a scaffold of the alligator genome assembled by the HiRise™ Assembly Pipeline. Below are A (red blocks) and B (blue blocks) compartments. The contact map below is zoomed in on an A to B compartment transition. TADs called from the analysis are outlined in black triangles with tracks for CTCFs (purple bars) and Isochores (yellow bars).



the American alligator (*Alligator mississipiensis*) assembly with several transitions between A Compartments (open chromatin/euchromatin) and B compartments (closed chromatin/heterochromatin), in which TADs of differing size can be observed specifically in a region that falls on an A to B transition. The larger TADs in B compartments feature densely packed chromatin with fewer CTCF bindings sites and areas accessible to transcription, whereas the A compartments are composed of many small TADs, enriched in CTCF binding sites, and often associated with areas of increased transcription, positing that the American alligator genome relies on 3D conformation to regulate gene expression.

5. Final Considerations

TADs consist of multiple, complex loop events, demonstrating the hierarchical nature of genome conformation. It should be kept in mind that TADs may not present in the sample, or TAD boundaries may not be strongly pronounced (Szabo *et al.,* 2019). This can either reflect a biology in which conformation plays less of a role in regulating gene expression (for example, smaller genomes, or during the cell cycle or cellular differentiation) or may be due to technical limitations such as low mappability rates resulting in a sparse contact matrix.

Table 4 – Example statistics from the Dovetail Genomics TAD analysis.

Across all species listed we report genome size (Gbp), the number of TADs at three resolutions, average TAD size and percent of the genome containing TADs at each resolution, and the number of isochores and CTCF binding sites for each genome. While plants do not use CTCF. We report the number of CTCF binding motifs present in plant genomes.

Species	Genome Size	Resolution	Number of TADs	Average TAD Size (Kbp)	% Genome Covered in TADs	# Isochores	# CTCF Sites
Human	3.09 Gbp	10k	1,828	395	22.02%	30,795	12,806
		25k	3,480	803	64.29%		
		50k	2,111	1,603	66.10%		
Alligator	2.13 Gbp	10k	100	819	3.75%	12,507	32.495
		25k	846	856	33.20%		
		50k	1,120	1,289	57.53%		
Snake	1.77 Gbp	10k	139	504	3.83%	7,955	19,889
		25k	459	632	15.92%		
		50k	495	1,124	29.03%		
Dog	2.34 Gbp	10k	1,576	457	29.75%	26,997	9,266
		25k	2,015	615	47.54%		
		50k	1,259	1,192	50.98%		
Watermellon	0.36 Gbp	10k	407	254	25.39%	3,691	757
		25k	299	688	43.28%		
		50k	142	1,325	43.13%		

References

[1]Iyer *et al*. Hierarchies in eukaryotic genome organization: Insights from polymer theory and simulations. BMC Biophysics, 2011.
[2]Putman *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Research, 2016.
[3]Bonev & Cavalli. Organization and function of the 3D genome. Nat Rev Genet, 2016.
[4]Dixon *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 2012.
[5]Ibn-Salem *et al.* Deletions of chromosomal regulatory boundaries are associated with congenital disease. Genome Biol, 2014.
[6]Lupiáñez *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell, 2015.
[7]Krefting *et al.* Evolutionary stability of topologically associating domains is associated with conserved gene regulation. BMC Biology, 2018.
[8]Sotelo-Silveria et al. Entering the Next Dimension: Plant Genomes in 3D. Trends in Plant Science, 2018.
[9]Kerpedjiev *et al*. HiGlass: Web-based visual comparison and exploration of genome interaction maps. Genome Biology, 2018.
[10]Rao *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 2014.
[11]Arhondakis *et al*. Transcriptome map of mouse isochores. BMC Genomics, 2011.
[12]Szabo *et al.,* Principles of genome folding into topologically associating domains. Scientific Advances, 2019.
[13]Oliver *et al.* IsoFinder: computational prediction of isochores in genome sequences. Nucleic acids research, 2004.
[14]Ziebarth *et al*. CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization. Nucleic acids research. 2012.