Briefings in Bioinformatics, 2025, **26(1)**, bbae660
https://doi.org/10.1093/bib/bbae660
Problem Solving Protocol

SatXplor—a comprehensive pipeline for satellite DNA analyses in complex genome assemblies

Marin Volarić, Nevenka Meštrović, Evelin Despot-Slade*

[illegible]

[illegible]

Keywords: SatXplor; pipeline; satellite DNA; genome

Introduction

[illegible]

Received: August 12, 2024. **Revised:** October 31, 2024. **Accepted:** December 4, 2024

[illegible] DNA [illegible] 18. H [illegible] DNA [illegible] DNA [illegible] DNA [illegible] G [illegible] DNA [illegible] T [illegible] DNA [illegible] T [illegible] DNA [illegible] DNA [illegible] DNA [illegible] 19, 21. T [illegible] Tribolium castaneum [illegible] DNA [illegible] DNA [illegible] 5 [illegible] DNA 22 [illegible] 23. O [illegible] DNA [illegible] T. castaneum [illegible] DNA [illegible] 24. O [illegible] DNA [illegible] DNA [illegible] DNA. T [illegible] DNA [illegible]

I [illegible] S X [illegible] T. castaneum [illegible] I [illegible] DNA E [illegible] (S X [illegible]) [illegible] DNA. T [illegible] S X [illegible] (150 350 M [illegible]) Meloidogyne (Meg[illegible] incognita, M. arenaria) Drosophila melanogaster, [illegible] (6.3 G [illegible]) [illegible] Locusta migratoria, [illegible] Arabidopsis thaliana [illegible] W [illegible] S X [illegible] (Homo sapiens) [illegible] (Mus musculus) [illegible] DNA- [illegible] S X [illegible] DNA [illegible] DNA [illegible] T [illegible] DNA [illegible]

Materials and methods

Genome and satDNA data

S X [illegible] T. castaneum, D. melanogaster, L. migratoria, M. incognita, M. arenaria, A. thaliana, H. sapiens, and M. musculus [illegible] S [illegible] T [illegible] 1. I [illegible] DNA [illegible] S [illegible] T [illegible] 2.

SatXplor pipeline overview

Detection and extraction of monomers

T [illegible] NCBI BLAST+ 25, [illegible] P [illegible] DNA [illegible] T [illegible] BLAST [illegible]:

`-evalue 10 -outfmt 6 -max_target_seqs 10000 -task blastn -num_threads 2 -dust no -soft_masking false`

D [illegible] DNA [illegible]

Array creation

A [illegible] DNA [illegible] P [illegible] F [illegible] BLAST [illegible] DNA [illegible] N [illegible] T [illegible] S [illegible] scipy.findpeaks() [illegible] >5% [illegible] 5% [illegible]). A [illegible] U [illegible]

Monomer density profiles

T [illegible] 2D [illegible] BLAST [illegible] P [illegible] F [illegible] 2D [illegible] G [illegible] scipy.stats.gaussian_kde(). F [illegible] NKERNEL_BINS [illegible] 2D [illegible]

K-mer-based profiling of arrays

T [illegible], S X [illegible] DNA [illegible]:

1. E [illegible] DNA [illegible]
2. C [illegible] D [illegible]
3. E [illegible] ([illegible] +/− 500 [illegible]).
4. C [illegible] 32 [illegible]
5. F [illegible] 32 [illegible] 32 [illegible] H [illegible]
6. C [illegible] +/− 5 [illegible]
7. D [illegible] ([illegible] 87% [illegible]).

Alignment and distance calculations

T [illegible] MAFFT 26, [illegible] P [illegible] mafft − −adjustdirection reorder − −threads str(half_threads)

W half_threads CPU ape S R F81

Dimensionality reduction

S PCA UMAP R 27. F MAFFT R, R F81 T F PCA R, PCA() FactoMineR 28. T ggplot2 29. Scree 10 F UMAP R, umap() umap UMAP T UMAP ggplot2

Network and distance graphs

F MAFFT R, F81 0 1. F R F

1. U MAFFT.
2. C F81
3. F N
4. C
5. R
6. C
7. C HTML network3D.

Performance

T S X DNA / A I (R) C (TM) i9 9900 CPU 128GB RAM, >8GB RAM (T 1).

I DNA I/O H

Results

T S X (F . 1), FASTA DNA T DNA BLAST DNA T N DNA I DNA (PCA) U M A P (UMAP) DNA F S X DNA A A L ( ). W S F . 1.

T S X DNA (S T 1) DNA (S T 2). I D. melanogaster 1.688 DNA 30; Meloidogyne 31, 32; L. migratoria 19; DNA A. thaliana 33. F S X (H. sapiens) (HSAT3, HSAT4, HSAT5). F (M. musculus) DNA (M S). T DNA (TRF, TAREAN) (S T 2).

SatDNA exploration

SatDNA detection

F S X NCBI BLAST+ DNA A BLAST DNA M (S T 3). T DNA DNA T DNA

T [illegible] 1. E [illegible], CPU, [illegible]

Genome (size)	Total monomers	Peak memory (GB)	Wall time (seconds)
Arabidopsis thaliana (119 M)	12 107	0.6	230
Drosophila melanogaster (143 M)	426	0.5	120
Meloidogyne incognita (199 M)	10 400	0.8	344
Meloidogyne arenaria (297 M)	83 460	8	626
Tribolium castaneum (225 M)	11 712	1	362
Locusta migratoria (6.2 G)	148 132	25	5238

F [illegible] 1. W [illegible] S X [illegible] DNA [illegible]. S X [illegible] DNA [illegible] DNA [illegible]. T [illegible] S X [illegible]

[illegible]. I [illegible] DNA.

Finding homology between annotated satDNA

T [illegible] S X [illegible] DNA. F [illegible]. U [illegible] 2D [illegible] BLAST [illegible] (S [illegible] T [illegible] 3), [illegible] DNA [illegible] DNA [illegible]. T [illegible] DNA [illegible] (F . 2A). F [illegible] DNA [illegible] (F . 2A, E [illegible] 1). S [illegible] (F . 2A, E [illegible] 2), [illegible] (F . 2A, E [illegible] 3). F [illegible], [illegible] DNA [illegible] (F . 2A, E [illegible] 4). T [illegible] DNA [illegible] DNA [illegible] T [illegible].

Detection of complex satDNA organization

W [illegible] DNA [illegible] (F . 2A); [illegible], [illegible] DNA [illegible]. T [illegible] DNA [illegible] DNA [illegible]. F [illegible] DNA [illegible], S X [illegible] DNA [illegible]. F [illegible] DNA, [illegible] ~0 (F . 2B, E [illegible] 1 [illegible] 2). H [illegible], [illegible] DNA [illegible] (F . 2B, E [illegible] 3 [illegible] 4). A [illegible] DNA [illegible] DNA [illegible].

K-mer-based profiling and edge detection

D [illegible] DNA [illegible]. T [illegible]. F [illegible] 21 [illegible]. T [illegible] (S [illegible] F . 2). T [illegible]. T [illegible]. U [illegible] (F . 2C). T [illegible]

[illegible]

SatDNA analysis

Clustering analysis for uncovering satDNA monomer relationships

[illegible]

[illegible] (S [illegible] F. 3). T [illegible] DNA [illegible] (F. 3A). T [illegible] DNA [illegible] PCA (F. 3A, [illegible]) UMAP (F. 3A, [illegible]). T [illegible] DNA [illegible] T [illegible] DNA [illegible] T [illegible] DNA [illegible] C [illegible] DNA [illegible] (F. 3A, E [illegible] 1), [illegible] F [illegible] DNA [illegible] (F. 3A, E [illegible] 2), [illegible] T [illegible] DNA [illegible] (F. 3A, E [illegible] 3), [illegible] S X [illegible] (PCA) [illegible] (UMAP) [illegible] DNA [illegible]

Unveiling evolutionary trends trough satDNA network analysis

F [illegible], S X [illegible] DNA [illegible] U [illegible] DNA [illegible] E [illegible] T [illegible] DNA [illegible] (F. 3B). A [illegible] (F. 3B, E [illegible] 1). C [illegible] (F. 3B, E [illegible] 2) [illegible] (F. 3B, E [illegible] 3), [illegible] (F. 3B, E [illegible] 4).

Distance mapping of surrounding regions

S DNA [illegible] T [illegible] DNA [illegible] A [illegible] R [illegible] 1.4), [illegible] 500 [illegible] T [illegible] DNA [illegible] A [illegible] A [illegible] DNA [illegible] (F. 3C). O [illegible] DNA [illegible] (F. 3C, E [illegible] 1). T [illegible] DNA [illegible] (F. 3C, E [illegible] 2). I [illegible] DNA [illegible] (F. 3C, E [illegible] 3), [illegible] DNA [illegible] F [illegible] (F. 3C, E [illegible] 4) [illegible] I [illegible]

Microhomology detection in flanking regions of satDNA arrays

A [illegible], [illegible] S X [illegible] DNA [illegible] M [illegible] DNA [illegible] 35 [illegible] DNA [illegible] 21. T [illegible] (F. 3D) [illegible] I [illegible] DNA [illegible] GC [illegible] (F. 3D, E [illegible] 1) [illegible] AT-A [illegible] N (A, T, G, C) [illegible] (F. 3D, E [illegible] 2) [illegible]

Benchmarking

T [illegible] S X [illegible] DNA [illegible] TRASH, TRF, ULTRA, [illegible] S G [illegible] T [illegible] S X [illegible] F [illegible] DNA [illegible] (F. 4A), [illegible] S X [illegible] DNA [illegible] N [illegible] DNA [illegible] (F. 4B). W [illegible] DNA [illegible] S X [illegible] F [illegible] S X [illegible] DNA [illegible] (F. 4C). I [illegible] S X [illegible] DNA [illegible] TRF [illegible] W [illegible] TRASH, TRF, [illegible] ULTRA [illegible] *de novo* [illegible] S [illegible] G [illegible] S X [illegible] DNA [illegible] T [illegible] S X [illegible] (T2T) [illegible] T [illegible]

F [illegible] 3. E [illegible] S X [illegible]. (A) S [illegible] DNA [illegible] PCA [illegible] UMAP. T [illegible]. S [illegible] DNA [illegible] PCA ([illegible]) [illegible] UMAP ([illegible]) (1); [illegible], UMAP (2), [illegible] - [illegible] (3). (B) G [illegible]. G [illegible] DNA [illegible]. E [illegible]. S DNA [illegible] (1), [illegible] (2), [illegible] ([illegible]) (3), [illegible] (4). (C) S DNA [illegible]. P [illegible] 500 [illegible] DNA [illegible]. N [illegible] DNA [illegible] (1) [illegible] ([illegible] DNA), (2) [illegible] (3), [illegible] DNA [illegible] (4). (D) [illegible]. A [illegible] GC [illegible] AT- [illegible] DNA (1) [illegible] ([illegible] T). A [illegible] DNA [illegible] (A D) [illegible] S [illegible] T [illegible] 2.

F [illegible] 4. B [illegible] S [illegible] X [illegible] DNA [illegible] . A [illegible] *T. castaneum* LG3 [illegible] DNA [illegible] 5, 23 [illegible] (A) S [illegible] DNA [illegible] (1) [illegible] (2). (B) A [illegible] DNA [illegible] DNA [illegible] (1) [illegible] (2). (C) Q [illegible] S [illegible] X [illegible] (TRASH [illegible] 16, TRF [illegible] 11, ULTRA [illegible] 12 [illegible] S [illegible] G [illegible] 18) [illegible] DNA [illegible] .

[illegible] DNA [illegible] (F [illegible] . 5). I [illegible] ([illegible] DNA) [illegible] DNA [illegible] (HSAT4, HSAT5 [illegible] HSAT6), [illegible] (F [illegible] . 5A). I [illegible] , S [illegible] X [illegible] DNA (M [illegible] S [illegible]), [illegible] DNA [illegible] (F [illegible] . 5B). T [illegible] S [illegible] X [illegible] DNA [illegible] , [illegible] DNA [illegible] .

Discussion

T [illegible] DNA [illegible] DNA, [illegible] 36. T [illegible] 37. W [illegible]

(T , A) . Sci Rep 2019;9:1 10.

8. S P, T W, S K. et al. H - DNA N W . Sci Rep 2015;5:10315.
9. N S, K S, R A. et al. T . Science 2022;376:44 53.
10. N P, R L, K A. et al. TAREAN: - DNA . Nucleic Acids Res 2017;45: 111.
11. B G. T : DNA . Nucleic Acids Res 1999;27:573 80.
12. O D, W T. ULTRA: A M B T D T R . I : Proc. 2018 ACM Int. Conf. Bioinformatics, Comput. Biol. Heal. Informatics, A C M , N , US, . 37 46, 2018. :// /10.1145/3233547.3233604.
13. K R, B G, K G. M : DNA. Nucleic Acids Res 2003;31: 3672 8.
14. B W, K KK, K O. R ' . Mob DNA 2015;6:4 9.
15. S AFA, H R, G P. RepeatMasker Open-4.0. 2013 2015. ://
16. W P, H M, H IR. TRASH: . Bioinformatics 2023; 39: 308.
17. K I, K E, D M. et al. A N TRF de novo DNA . Plant Theory 2022;11:2103.
18. V MR, K P, P AM. et al. S G : . Bioinformatics 2022;38:2049 51. :// /10.1093/ / 018.
19. R -R FJ, L -L MD, C J. et al. H - DNA . Sci Rep 2016;6:28333.
20. V T, R L, N P. et al. C - - DNA - . Plant J 2020;101:484 500. :// /10.1111/ .14546.
21. S JS, K DE, E DG. et al. D . Mol Biol Evol 2020;37: 2241 56. :// /10.1093/ / 078.
22. G T, D -S E, M N. et al. C H3 T . PLoS Genet 2020;16: 1009115.
23. G T, D M, D -S E. et al. T - - DNA T . Genes (Basel) 2023;14:999.
24. V M, D -S E, V D. et al. L - Tribolium castaneum - DNA - . Genome Res 2024;34:1878 94. :// /10.1101/ .279225.124.
25. C C, C G, A V. et al. BLAST+: . BMC Bioinformatics 2009;10:421.
26. K K, M K, K KI. et al. MAFFT: F . Nucleic Acids Res 2002;30:3059 66.
27. R C T . R: A Language and Environment for Statistical Computing. R F S C , V , A , 2023. :// .R- . /.
28. L S, J J, H F. F M R: R . J Stat Softw 2008;25:1 18.
29. W H. ggplot2: Elegant Graphics for Data Analysis. S - V N . 2016. ISBN 978-3-319-24277-4. :// 2. .
30. L LG, H SL, G JL. O - 1.688 DNA . G3 Genes Genomes Genet 2020;10:4129 46.
31. D -S E, M B, Š S. et al. T - 19- Meloidogyne N - . Mol Biol Evol 2021;38:1943 65. :// /10.1093/ / 336.
32. D -S E, Š S, M B. et al. S DNA . BMC Biol 2022;20:1 19.
33. S CR, G J, M M. et al. C A . Nucleic Acids Res 1988;16:6753 66.
34. R JT, T H, W W. et al. I - . Nat Biotechnol 2011;29:24-6. :// /10.1038/ .1754.
35. L, J R, G T. et al. E DNA: - . Signal Transduct Target Ther 2022;7:342.
36. L S, L M, F D. et al. D DNA 0 0004T (25.)T TMCP/(.)T TM-.0004T (S

46. M N, M J, P N. *et al.* TRT: . *Bioinformatics* 2021;**37**: 731 3. :// . /10.1093/ / 736.

47. E C, E R, T M. *et al.* *RepeatOBserver: tandem repeat visualization and centromere detection.* R 2023.12.30.573697. :// . /10.1101/2023.12.30.573697.

Briefings in Bioinformatics, 2025, **26(1)**, 660
:// . /10.1093/ / 660
P S P