PanKmer: *k*-mer based and reference-free pangenome analysis

Anthony Aylward*[1], Semar Petrus*[1], Allen Mamerto[1], Nolan Hartwick[1], Todd P. Michael**[1]

[1]The Plant Molecular and Cellular Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA

*Authors contributed equally to this work
**Corresponding author

Abstract

Summary

Pangenomes are replacing single reference genomes as the definitive representation of DNA sequence within a species or clade. Pangenome analysis predominantly leverages graph-based methods that require computationally intensive multiple genome alignments, do not scale to highly complex eukaryotic genomes, limit their scope to identifying structural variants (SVs), or incur bias by relying on a reference genome. Here, we present PanKmer, a toolkit designed for reference-free analysis of pangenome datasets consisting of dozens to thousands of individual genomes. PanKmer decomposes a set of input genomes into a table of observed *k*-mers and their presence-absence values in each genome. These are stored in an efficient *k*-mer index data format that encodes all forms of variation within the pangenome, including SNPs, INDELs, and SVs. It also includes functions for downstream analysis of the *k*-mer index, such as calculating sequence similarity statistics between individuals at whole-genome or local scales. For example, *k*-mers can be "anchored" in any individual genome to quantify sequence variability or conservation at a specific locus. This facilitates workflows with various biological applications, e.g. identifying cases of hybridization between plant species. PanKmer provides researchers with a valuable and convenient means to explore the full scope of genetic variation in a population, without reference bias.

Availability and implementation

PanKmer is implemented as a Python package, released under a BSD license. The source code is available from the Python Package Index (PyPI) at https://pypi.org/project/pankmer/ as well as Gitlab at https://gitlab.com/salk-tm/pankmer. Full documentation is available at https://salk-tm.gitlab.io/pankmer/.

Contact

tmichael@salk.edu

Supplementary information

Supplementary data are available at *Bioinformatics* online.

Introduction

Pangenomes consolidate genomic sequence data from multiple individual organisms into a single data structure representing the genomes of a population, species, or clade. Pangenomics was first applied to microbes, but has grown to include a wide variety of eukaryotes[1–3]. The advent of plant and animal pangenomes coincided with the reduced cost of NGS sequencing technologies. Crop plants were an early subject of pangenome studies, including soybean, brassica, and wheat species [4–6]. These pangenomes focused on genic sequence only, and they were built by collecting short-read whole-genome sequencing datasets and mapping to a reference, allowing relatively simple if biased assembly of multiple individual genomes. Improvements in long-read sequencing technology and assembly methods have enabled the construction of dozens or hundreds of high-quality genomes for a single plant or animal species. This has facilitated a wave of pangenome studies based on collections of entire assembled genomes [7–13].

The transition from representation and analysis of single genomes to pangenomes presents significant challenges. Most studied populations include extensive structural variation (SV), which means only a fraction of genes or intergenic sequences are present in all individuals. This fraction is referred to as the "core" genome, while the remainder is variously called "dispensable," "variable," or "accessory." [2,14,15] Furthermore, the core genome cannot be defined by a simple linear coordinate system. A useful pangenomic dataset must identify the core genome and facilitate analysis of the variable regions.

Currently the dominant methodology is pangenome graphs, which generally consist of nodes representing segments of genomic sequence connected by edges which allow any individual genome to be traced as a path through the graph [7,16–19]. These graphs have replaced “iterative assembly” methods to advance our understanding of genomic diversity, especially of SV, and demonstrated utility for crop breeding [7,8,11,20]. However, their construction is far from a solved problem and generally relies either on computationally expensive multiple genome alignment or on biased alignment to a single reference. Their application to highly complex eukaryotic genomes, such as those of plants, is limited[21–23]. Such graphs may be limited to as few as a dozen input genomes[12,24]. Other methods gain efficiency by focusing on specific categories of variants[17,19].

K-mer decomposition is an alternative to graph-based pangenomes. In this framework the space of genomic sequences across a population is represented by a set of *k*-mers (sequences of length k). Each individual is represented by a subset: all *k*-mers observed in a single genome. This view does not depend on any coordinate system and therefore sidesteps the difficulties of

multiple genome alignment and graph construction. Analysis of *k*-mers has many applications in genomics and pangenomics, and *k*-mers have recently been used as markers for GWAS[21,25–31].

Here we present PanKmer, a non-graphical *k*-mer decomposition method designed to efficiently represent and analyze the full scope of variation in large pangenomic datasets, with no reliance on a reference genome.

Features & Implementation

k-mer index

The foundational component of PanKmer's pangenome representation is the *k-mer index*. It is constructed from a set G of input genomes by decomposing them into a set K of all unique canonical *k*-mers and noting for each input genome which *k*-mers are present and which are absent (Figure 1a). Each *k*-mer is considered equivalent to its reverse complement and is recorded in canonical form. The *k*-mer index X is then a $|K|$ by $|G|$ table of binary values indicating presence/absence of each canonical *k*-mer in each genome. The index can integrate an arbitrary number of genomes from one or several species, requires no reference genome, and enables a range of downstream analyses.

The index is constructed by scanning all $|G|$ input genomes sequentially, recording newly encountered *k*-mers, and updating presence/absence values with each new genome scanned. To make efficient use of all available CPU's, this process is parallelized across the theoretical *k*-mer space. The set κ of all possible canonical *k*-mers has size $|\kappa| = ½k^4$, which is divided evenly among n segments $\kappa_1 \ldots \kappa_n$. Index construction is then divided into n subprocesses, each of which constructs a sub-index X_i skipping *k*-mers not included in κ_i. To efficiently store and access the *k*-mer index on disk, each *k*-mer is converted to an integer value.

The resultant index is robust to varying contiguity in the input genomes, so chromosome-level assemblies can be directly compared to unscaffolded contigs or unaligned reads.

Adjacency matrix

Once the *k*-mer index is constructed, each input genome g_i is represented by $|K|$ binary values representing presence or absence of each *k*-mer in K. This provides a natural means of calculating pairwise similarity/adjacency values for the input genomes. PanKmer includes a function to calculate the number of shared *k*-mers between all pairs of input genomes and return them as an adjacency matrix. Subsequently, the adjacency values can be used to perform a hierarchical clustering of input genomes and plot adjacency values as a heatmap. The adjacency values may also be converted to Jaccard or Overlap similarities (Fig 1b).

Genome anchoring

While the *k*-mer index does not rely on any specified reference genome, it can be used to contextualize individual sequences. Given a sufficient *k*-mer length (e.g. the default k=31), we

can assume that each k-mer present in an individual genome g_i occurs approximately once in g_i. Therefore, we can quantify variation across any locus in an "anchor" genome g_i by walking along the sequence, checking the *k*-mer that corresponds to each position, and calculating the fraction of genomes in *G* which share that *k*-mer. We refer to this fraction as the "*k*-mer coverage" value at each position. High *k*-mer coverage values indicate core loci which are conserved in many individuals across the pangenome, while low values indicate variable loci which are present only in g_i and a small number of other genomes. Hence, core sequences will have high *k*-mer coverage levels in all target genomes, while variable sequences will have relatively lower levels in each genome that features them.

Results

To demonstrate the utility of PanKmer, we constructed a cross-species pangenome of cattail downloaded from NCBI SRA: 10 *Typha domingensis* (TD) and 3 *T. latifolia* (TL) genomes (Figure 1b) (Table S1). Three of the TD genomes (TD01, TD22, TD23) were highly heterozygous (Figure S1) (Table S2). We used *k*-mer profiles to compute two measures of adjacency, the number of shared *k*-mers and Jaccard similarity (Figure 1b). The 3 TL genomes were highly similar to one another, with an average Jaccard similarity value of 0.92 (186M shared k-mers) for TL-TL comparisons, while the average similarity of TD-TD comparisons was 0.60 (151M shared k-mers) (Table S3-4). The average similarity of TL-TD comparisons was 0.19 (64M shared k-mers). We also constructed single-species pangenomes of *T. domingensis* and *T. latifolia* (Figure S2) (Table S5-8). One TD genome, TD01, was an outlier relative to other TD, with Jaccard similarity of only 0.38 (127M shared k-mers) on average. Conversely, TD01 showed relatively high similarity to TL genomes, averaging 0.48 (148M shared k-mers). To inspect TD01 more closely, we calculated average *k*-mer coverage in 100kb bins across its assembled contigs and observed a mixture of TD and TL sequences. A close examination of variability anchored in TD01 revealed the presence of large and small introgressions (Figure 1c, Figure S3). High heterozygosity together with the presence of introgressions suggest TD01 is the descendant of a recent TD-TL hybridization event.

Conclusion

PanKmer enables *k*-mer based analysis of pangenome datasets. It accepts as input a collection of genome assemblies or unaligned reads in FASTA format, and produces a *k*-mer index. The index is agnostic to the contiguity of assemblies. A powerful feature is the ability to anchor the index in any individual genome and identify core or variable loci across the whole *k*-mer index or subsets based on a shared criteria. This allows users to explore the full scope of genetic variation, without incurring bias from the choice of a single reference. PanKmer includes tools for downstream processing of the index which provide biological insights, such as identifying a hybridization event between *T. latifolia* and *T. domingensis* (Figure 1c, Figure S3).

The primary advantage of PanKmer over other pangenome analysis tools is its ability to capture all forms of variation, including SNPs, INDELs, and SVs. Our reference-free and alignment-free algorithm is also more computationally tractable than graph-based methods. On the other hand,

a significant limitation of PanKmer is the loss of spatial relationships between *k*-mers in the index. However, their spatial context can be rescued by projecting the index onto an anchor genome. Currently PanKmer does not have genotyping functions, but these are planned for future releases.

The supplementary material includes a hands-on tutorial, in-depth discussion of the indexing algorithm, and extended tables and figures for the *Typha* hybridization use case.

Tutorials & documentation

Complete tutorials and documentation are available in the supplementary material and online: https://salk-tm.gitlab.io/pankmer/

Author contributions

A.A.: software (main developer, testing), writing (original draft, review and editing) funding acquisition. S.P.: software (original developer), writing (review and editing). A.M.: software (pipeline management, testing), writing (review and editing). N.H.: software (optimization) T.M.: Supervision, funding acquisition, writing (review and editing).

Acknowledgements

We thank Edward S Buckler, Bethany F Econopouly, Zachary R Miller, Peter J Bradbury, and Matt Wiese of the United States Dept. of Agriculture-Agricultural Research Service and Institute for Genomic Diversity at Cornell University for helpful discussion and collaboration on software features and implementation.

Funding

This research was funded by a Bill & Melinda Gates Foundation (BMGF) grant (INV-040541) to T.P.M., and the Tang Genomics Fund.

Conflict of Interest: none declared

Data availability

The raw reads have been deposited in the NCBI short read archive (SRA) under BioProject PRJNA742003. Accession numbers for individual experiments are provided in the supplementary material. The genome assemblies are available for download from Michael lab AWS storage at: https://salk-tm-pub.s3.us-west-2.amazonaws.com/pub-supplementary/PRJNA742003-ASSEMBL.tar

Figure

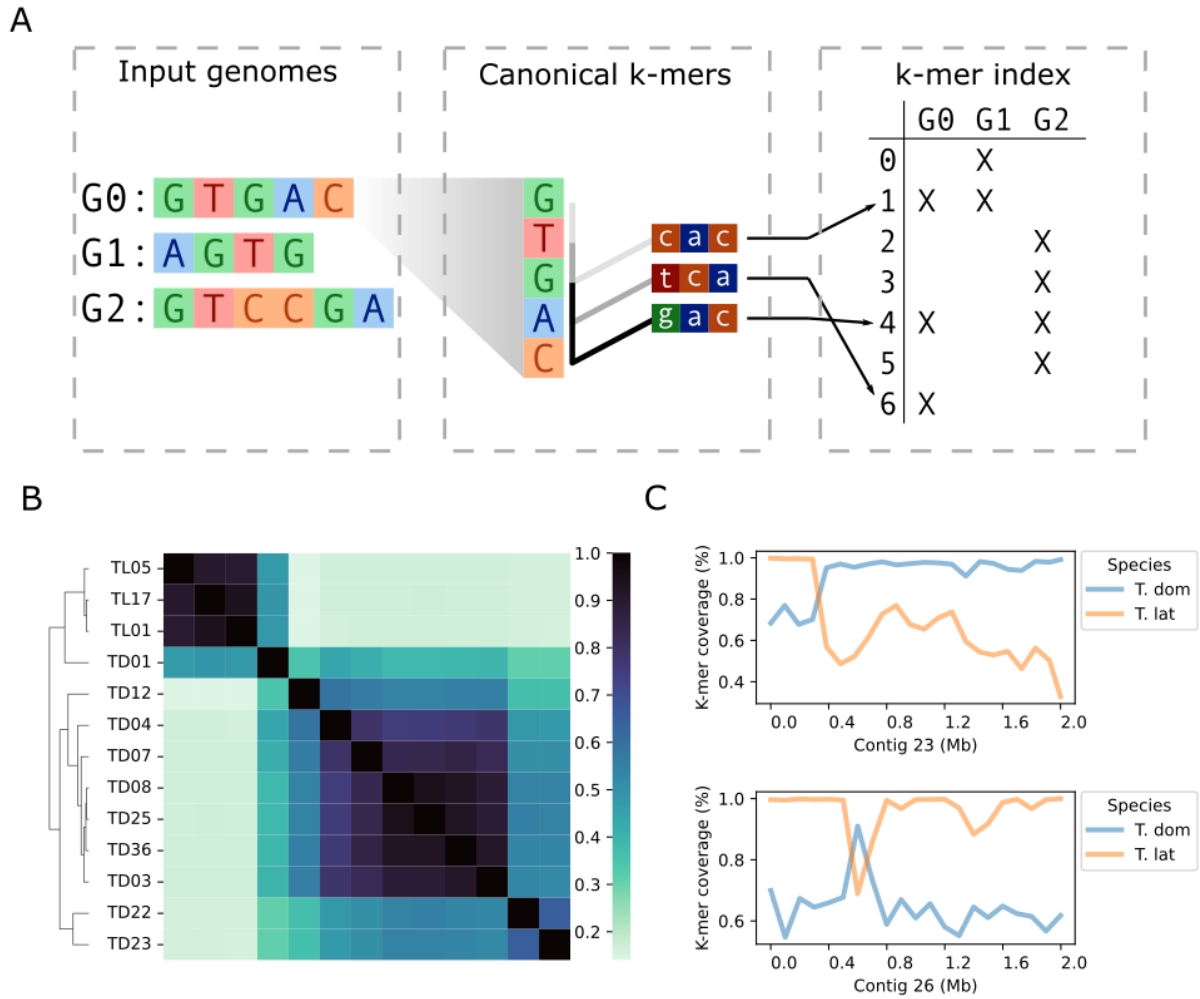


Figure 1: PanKmer enables the rapid estimation of relatedness across the pangenome as well as analysis of specific loci A) Schematic of procedure for constructing the K-mer index. In the example, each genome G0-G2 is decomposed into canonical 3-mers. Each 3-mer is assigned an integer value, and its presence/absence is recorded for each genome in the index. B) Relatedness heatmap of Typha pangenome, Jaccard similarity values shown. C) Genome anchoring plots of representative contigs in TD01. Average *k*-mer coverage of 100-kb bins shown.

References

1. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15**, 589–594 (2005).

2. Golicz, A. A., Bayer, P. E., Bhalla, P. L., Batley, J. & Edwards, D. Pangenomics Comes of

Age: From Bacteria to Plant and Animal Applications. *Trends Genet.* **36**, 132–145 (2020).

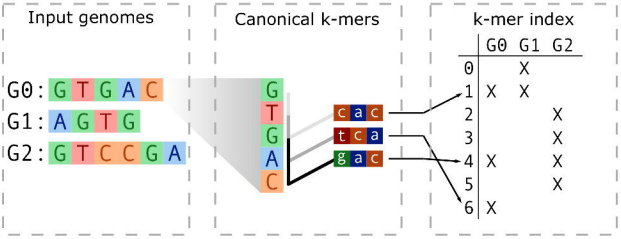
3. Li, W. *et al.* Plant pan-genomics: recent advances, new challenges, and roads ahead. *J. Genet. Genomics* **49**, 833–846 (2022).
4. Li, Y. *et al.* De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32**, 1045–1052 (2014).
5. Golicz, A. A. *et al.* The pangenome of an agronomically important crop plant Brassica oleracea. *Nat. Commun.* **7**, 13390 (2016).
6. Montenegro, J. D. *et al.* The pangenome of hexaploid bread wheat. *Plant J.* **90**, 1007–1013 (2017).
7. Zhou, Y. *et al.* Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* **606**, 527–534 (2022).
8. Shang, L. *et al.* A super pan-genomic landscape of rice. *Cell Res.* **32**, 878–896 (2022).
9. Tong, X. *et al.* High-resolution silkworm pan-genome provides genetic insights into artificial selection and ecological adaptation. *Nat. Commun.* **13**, 5619 (2022).
10. Yang, T. *et al.* Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics. *Nat. Genet.* **54**, 1553–1563 (2022).
11. Gui, S. *et al.* A pan-Zea genome map for enhancing maize improvement. *Genome Biol.* **23**, 178 (2022).
12. Li, H. *et al.* Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. *Nat. Commun.* **13**, 682 (2022).
13. Tang, D. *et al.* Genome evolution and diversity of wild and cultivated potatoes. *Nature* **606**, 535–541 (2022).
14. Aggarwal, S. K. *et al.* Pangenomics in Microbial and Crop Research: Progress, Applications, and Perspectives. *Genes* **13**, 598 (2022).
15. Lei, L. *et al.* Plant Pan-Genomics Comes of Age. *Annu. Rev. Plant Biol.* **72**, 411–435 (2021).

16. Baaijens, J. A. *et al.* Computational graph pangenomics: a tutorial on data structures and their applications. *Nat. Comput.* **21**, 81–108 (2022).
17. Bradbury, P. J. *et al.* The Practical Haplotype Graph, a platform for storing and using pangenomes for imputation. *Bioinformatics* **38**, 3698–3702 (2022).
18. Hickey, G. *et al.* Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* **21**, 35 (2020).
19. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **21**, 265 (2020).
20. Ruperao, P. *et al.* Sorghum Pan-Genome Explores the Functional Utility for Genomic-Assisted Breeding to Accelerate the Genetic Gain. *Front. Plant Sci.* **12**, (2021).
21. Khan, A. W. *et al.* Super-Pangenome by Integrating the Wild Side of a Species for Accelerated Crop Improvement. *Trends Plant Sci.* **25**, 148–158 (2020).
22. Danilevicz, M. F., Tay Fernandez, C. G., Marsh, J. I., Bayer, P. E. & Edwards, D. Plant pangenomics: approaches, applications and advancements. *Curr. Opin. Plant Biol.* **54**, 18–25 (2020).
23. Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J. & Edwards, D. Plant pan-genomes are the new reference. *Nat. Plants* **6**, 914–920 (2020).
24. Zhang, X. *et al.* Pan-genome of Raphanus highlights genetic variation and introgression among domesticated, wild, and weedy radishes. *Mol. Plant* **14**, 2032–2055 (2021).
25. Gupta, P. K. GWAS for genetics of complex quantitative traits: Genome to pangenome and SNPs to SVs and k-mers. *BioEssays* **43**, 2100109 (2021).
26. Jayakodi, M., Schreiber, M., Stein, N. & Mascher, M. Building pan-genome infrastructures for crop plants and their use in association genetics. *DNA Res.* **28**, dsaa030 (2021).
27. Holley, G., Wittler, R. & Stoye, J. Bloom Filter Trie: an alignment-free and reference-free data structure for pan-genome storage. *Algorithms Mol. Biol.* **11**, 3 (2016).
28. Jonkheer, E. M. *et al.* PanTools v3: functional annotation, classification and phylogenomics.

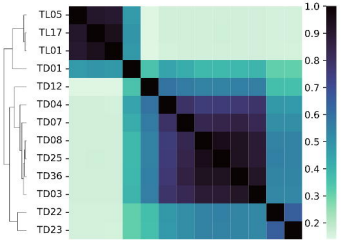
Bioinforma. Oxf. Engl. **38**, 4403–4405 (2022).

29. Aun, E., Brauer, A., Kisand, V., Tenson, T. & Remm, M. A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. *PLoS Comput. Biol.* **14**, e1006434 (2018).
30. Voichek, Y. & Weigel, D. Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nat. Genet.* **52**, 534–540 (2020).
31. Ebler, J. *et al.* Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).

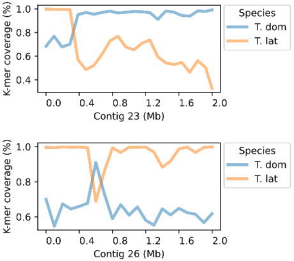
A

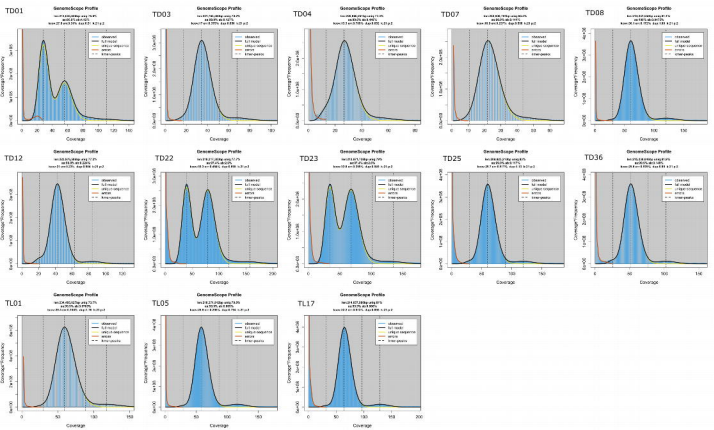

Input genomes
G0: G T G A C
G1: A G T G
G2: G T C C G A
Canonical k-mers
G
T
G
A
C
c a c
t c a
g a c
k-mer index
G0 G1 G2
0 X
1 X X
2 X
3 X
4 X X
5 X
6 X

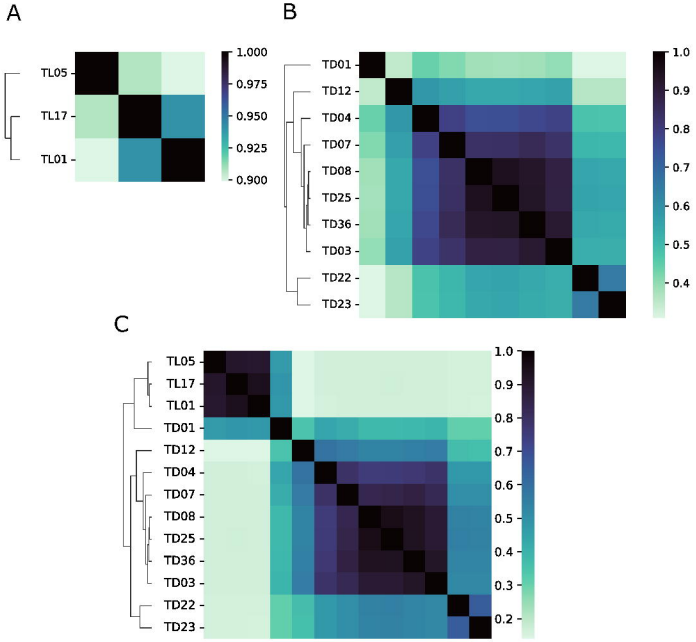
B

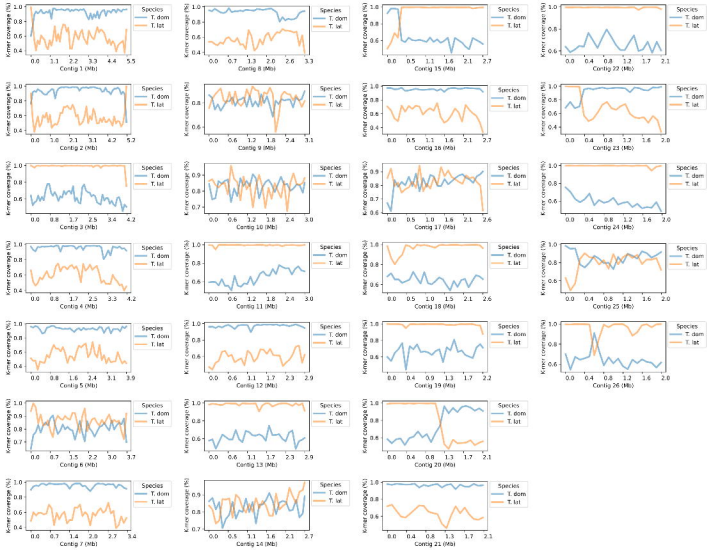

TL05
TL17
TL01
TD01
TD12
TD04
TD07
TD08
TD25
TD36
TD03
TD22
TD23
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

C


K-mer coverage (%)
Contig 23 (Mb)
Species
T. dom
T. lat
K-mer coverage (%)
Contig 26 (Mb)
Species
T. dom
T. lat


TD01
TD03
TD04
TD07
TD08
TD12
TD22
TD23
TD25
TD36
TL01
TL05
TL17
GenomeScope Profile
Coverage
Coverage Frequency


A
TL05
TL17
TL01
1.000
0.975
0.950
0.925
0.900
B
TD01
TD12
TD04
TD07
TD08
TD25
TD36
TD03
TD22
TD23
1.0
0.9
0.8
0.7
0.6
0.5
0.4
C
TL05
TL17
TL01
TD01
TD12
TD04
TD07
TD08
TD25
TD36
TD03
TD22
TD23
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2


Kmer coverage (%)
Species
T. dom
T. lat
Contig 1 (Mb)
Contig 2 (Mb)
Contig 3 (Mb)
Contig 4 (Mb)
Contig 5 (Mb)
Contig 6 (Mb)
Contig 7 (Mb)
Contig 8 (Mb)
Contig 9 (Mb)
Contig 10 (Mb)
Contig 11 (Mb)
Contig 12 (Mb)
Contig 13 (Mb)
Contig 14 (Mb)
Contig 15 (Mb)
Contig 16 (Mb)
Contig 17 (Mb)
Contig 18 (Mb)
Contig 19 (Mb)
Contig 20 (Mb)
Contig 21 (Mb)
Contig 22 (Mb)
Contig 23 (Mb)
Contig 24 (Mb)
Contig 25 (Mb)
Contig 26 (Mb)