Rapid and reproducible haplotyping of complete mitochondrial genomes using split *k*-mers

Douglas S. Stuehler Jr., Liliana M. Cano, Michelle Heck

Abstract

Inconsistent gene selection and trimming are weaknesses in generation of reproducible multilocus sequence type (MLST) phylogenies. As an alternative, we tested split *k*-mer analysis (SKA), which uses short, fixed-length nucleotide sequences, to haplotype mitogenomes of three animal species. SKA accurately haplotyped each species and exceeded MLST sensitivity for the Asian citrus psyllid by identifying haplotypes associated with *Candidatus* Liberibacter asiaticus titer, revealing new information on the genetic basis of vector competence in an economically important insect.

Main text

SKA[1] intuitively adapts *k*-mers by splitting them into three components, a central nucleotide and two flaking sequences of equal length. To ensure specificity, the two flanking sequences must have an exact match in alignments, while to detect SNPs, the central nucleotide(s) is allowed to simultaneously align with mismatches. Mutation-aware split *k*-mers have shown promise for rapid investigations of bacterial outbreaks[2,3], especially in tracking strain divergence independent of reference genomes[4]. SKA is optimally used for prokaryote haploid genomes[1,2].

Eukaryotic mitochondrial genomes, while generally more compact and streamlined than nuclear genomes, share key structural and sequence similarities with bacterial genomes, including their circular organization, gene content, and AT-rich composition, reflecting their evolutionary origin from an ancestral α-proteobacterium[5,6]. We hypothesize that the value of split *k*-mers in mitochondrial haplotyping originates from their ability to create reproducible results, correcting current MLST protocols affected by ambiguous gene selection and alignment trimming (Fig. 1a). Due to unstandardized gene selection, inconsistent gene sets are used for MLST studies of the same species (Fig 1a). Furthermore, due to varying gene lengths and sequence diversity multiple sequence alignments (MSAs) and whole genome alignments (WGAs) must be trimmed to the limiting sequences removing unreliable regions unique to each study (Fig. 1a). The exact coordinates of trimmed sequences are often not reported, making results difficult to reproduce. Programs such as ClipKIT[7] and CIAlign[8] aim to resolve ambiguity by removing poorly aligned or highly gapped regions to enhance phylogenetic accuracy. However, this process is prone to error as excessive trimming may discard valuable phylogenetic information and support incorrect lineages[9]. In another case, such as with Gblocks[10], aggressive trimming can result in entire alignments being removed[7], impacting a dataset's representativeness. Segmented *k*-mers have shown promise in assessing whole mitogenome similarity[11] but have not garnered attention over traditional protocols. To resolve these issues, metagenomic sequencing in conjunction with split *k*-mers can rapidly identify polymorphisms from hundreds to thousands of complete mitogenomes at once. Additionally, sequence conservation of vertically transferred mitochondria allows split *k*-mers

to accurately detect haplotypes across multiple generations and populations. Challenges arise when addressing gene rearrangements in complete mitogenomes; however, split *k*-mers can effectively handle these rearrangements along with sequence inversions commonly seen in tRNAs.

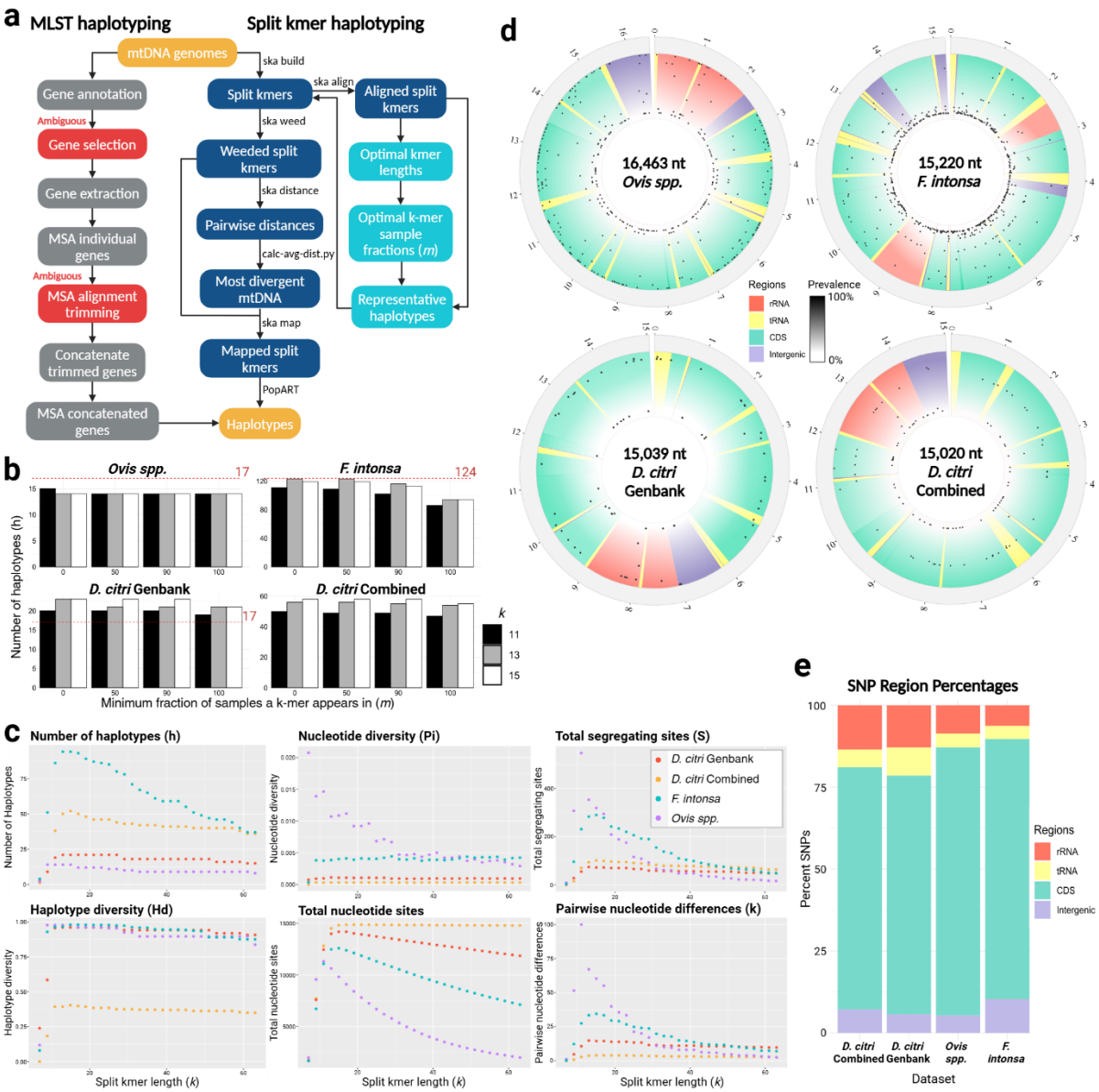

Figure 1. a. Workflow of MLST mtDNA haplotyping compared to haplotyping mtDNA with split *k*-mers. For MLST genes are annotated, ambiguously selected and trimmed, then concatenated to produce CDS-based haplotypes. The split *k*-mer approach includes two major steps: (1) the reference-free *align* algorithm is used to initially examine a dataset to assign optimal *k*-mer length (*k*), *k*-mer sample fraction (*m*), and to examine individual populations with a fraction of samples greater than *m* to assign representative samples, (2) best length *k*-mers are weeded by the optimized *k*-mer sample fraction, mapped to the most divergent mitogenome, and are haplotyped. **b.** The impact of *m* on haplotype resolution using mtDNA *k* values of 11bp, 13bp and 15bp. Red dash lines and red numbers represent the number of haplotypes identified from published analyses using MLST (*F. intonsa* and *D. citri* GenBank)

and WGA (*Ovis spp.)*. **c.** The impact of different *k* values (7 to 63) at *m* = 90 on six statistical measures of population diversity calculated by DnaSP6[12]. **d.** Distribution of SNPs detected in mitochondrial genomes from four datasets generated with BioCircos[13]. SNPs are ranked by percent prevalence. Most divergent mitogenomes were used as references. Reference GenBank accessions are as follows; *Ovis spp.*: JN181255.1, *F. intonsa*: OP546494.1, *D. citri* GenBank: OM181945.1, and *D. citri* Combined: MF614822.1. **e.** The percentage of SNPs residing in four genetic elements of reference mitochondrial genomes. Plots were generated in R[14] and figures were created in BioRender.com.

High-resolution haplotype networks are fundamental to examine populations of individual species while avoiding premature taxonomic grouping from missing data[15]. Haplotyping methods that reduce missing data, detect the greatest number of reliable variants, and result in the accurate reconstruction of population networks provide the greatest haplotype resolution. SKA transparently introduces missing data by assigning ambiguous bases in alignments which are influenced by two easily assessed parameters; the minimum fraction of samples a *k*-mer must be found in (*m*) and split *k*-mer length (*k*). To ensure the greatest detection of SNPs without introducing excess ambiguous bases, the effects of both *k* and *m* should be investigated for each specific dataset as these values are not generalizable. It is important to note that the *align* algorithm is less sensitive in separating haplotypes of highly similar individuals compared to the *map* algorithm[1,2] (Fig. S1).

We investigated the ability of SKA to detect mitochondrial haplotypes in three different species. Haplotype results produced with whole-genome alignment (WGA) and MLST protocols were compared to haplotype results obtained with SKA2[2] as follows. Two complete mitogenome datasets with published MLST haplotype networks from *Diaphorina citri* (n = 31)[16] and *Fraklienella intonsa* (n = 149)[17], and one dataset assessed with WGA from *Ovis spp.* (n = 17)[18] were selected for reanalysis. Lastly, one unpublished dataset of *D. citri* (n = 681) containing citrus grove-collected insect samples from Florida (n = 607), samples from lab colonies from University of California at Davis in California (n = 6), and samples from lab colonies in Ithaca, NY (n = 37) combined with available genomes on GenBank were used in this study. *D. citri* is the insect vector of *Candidatus* Liberibacter asiaticus, the invasive bacterial pathogen responsible for citrus greening disease, which is decimating the Florida citrus industry[19].

Minimum sample fraction filtered k-mers at *m* = 0.9 mapped to the most divergent reference with ska *map*, detected rare SNPs two base pairs apart in two separate *D. citri* samples. Ska *align* executed using split *k*-mers filtered at a *m* = 0.9 threshold by ska *weed* or *align* missed these variants despite detecting other single-sample SNPs. For ska *map*, split *k*-mers are mapped to a reference for the sole purpose of a coordinate system[2], yet this helps to resolve some ambiguous bases called with *align*. We hypothesize the distinction arises from ska *map*'s reference-based algorithm, which uses overlapping *k*-mers to anchor positions and infer variants against a reference sequence, ensuring robust detection even after filtering removes rare *k*-mers. In contrast, ska *align*'s reference-free, consensus-based approach failed to align these SNPs, likely due to disrupted k-mer continuity and inability to create a seamless overlap of chained *k*-mer sequences arising from filtering of *k*-mers and proximity

of the SNPs. The discrepancy of results between *align* and *map* algoithms highlight the critical role of reference anchoring in ska *map*'s sensitivity to such variants. As a result, our analysis supports using *ska align* for initial assessments of *k* and *m* to quickly identify a range of values that continually produce the most haplotypes with sensical network structures. Final assessments and optimizations are best made with *ska map* and the most divergent mtDNA set as the reference (Fig. 1a).

To calculate the most divergent mitogenome, we used ska *distance* with split *k*-mers filtered by both the *m* and *k* values that produced the most haplotypes via initial ska *align* assessments. Lowering the minimum *k*-mer sample fraction filter *m*, to include more sequence data found in fewer samples, increased haplotype resolution as seen with *F. intonsa* and *Ovis spp.* (Fig. 1b). In parallel, a decrease in split *k*-mer length *k* increased haplotype resolution by allowing SNPs to be detected closer together, at the cost of introducing repeated *k*-mers and ambiguous base pairs. Overall, decreasing *k* increased the number of haplotypes, nucleotide diversity, total segregating sites, pairwise nucleotide differences, total nucleotide sites, and haplotype diversity (Fig. 1c). The optimal *k*-mer lengths were dataset specific, as a *k* of 11bp gives the best haplotype resolution for the *Ovis spp.* dataset, a *k* of 13bp is best the *F. intonsa* dataset, and a *k* of 15bp produces the greatest number of haplotypes for both *D. citri* datasets.

Missing data was exacerbated by increased sample diversity and overrepresented populations which both impact *k*-mer filtering and SNP detection. The disproportionate representation of a population resulted in *k*-mers with true SNPs to be filtered out if those *k*-mers are not found in a population, contributing to at least the minimum sample fraction assigned to *m* (Fig. S2). Therefore, sequences from the same species and equal population representation are important for split *k*-mer mtDNA haplotyping. Datasets of individuals from the same genus may also be investigated using split *k*-mers, although for many genera inherent mtDNA sequence diversity will introduce missing data. This was observed with the *Ovis spp.* dataset, where changing either *k* or *m* minimally impacted haplotype resolution and where SKA2 did not outperform classic WGA (Fig. 1b).

SKA *align* will produce an unordered alignment which is used to build phylograms and haplotype networks; however, to learn where SNPs reside within a mitogenome the *map* algorithm must be used. Split *k*-mers mapped to the most divergent mitogenome detected SNPs in rRNA, tRNA, CDS, and intergenic sequences proportional to the total amount of sequence each element harbors in a mitogenome (Fig. 1d, 1e). Furthermore, no bias was found towards recombinant-prone elements, repeats, or low complexity regions (Figure 1d, 1e). Major assumptions underlying phylogenetics are violated if mutations originating from horizonal gene transfer are incorporated into analyses[20]. MLST studies are heavily biased to polymorphisms residing within conserved protein coding sequences to help avoid this issue, despite neglecting polymorphisms in functional intergenic elements and tRNAs. In theory, the use of intergenic sequence for phylogenetics increases the potential introduction of mutations arising from recombination in animal mitogenomes. Albeit, some animal mitochondria, such as higher mammals, rarely experience recombination[21] and in humans the generational impact of recombinant mtDNA is arguably irrelevant[22], making split *k*-mers applicable to a more

diverse sets of mammalian mitogenomes. Conversely, insects experience high rates of mitochondrial gene rearrangement as a method to correct for high rates of nucleotide substitution[23]. Indels produced from these rearrangements are not considered in haplotyping, allowing split *k*-mers to accurately haplotype whole mitogenomes while handling sequence inversions, duplications, and transpositions.

Compared to current methods, we show that split *k*-mers are more sensitive in separating mitochondrial haplotypes of *D. citri* populations with a 29% increase in haplotype resolution compared to a previously published *D. citri* dataset (Fig. 1b, 2a). For the larger and more diverse published datasets of *F. intonsa* and *Ovis spp.*, a decrease in haplotype resolution of 0.8% and 12% was found, respectively. We hypothesize the discrepancy in haplotype resolution is due to the inherent diversity of the datasets causing more SNPs to be spaced closer than the optimal *k*. Regions with a length smaller than *k,* with multiple SNPs, cannot be matched to other *k*-mers, leading to their removal. Thus, highly related individuals are best for split *k*-mer mtDNA haplotyping, such as the *D. citri* datasets, where haplotype resolution increased with split *k*-mers compared to MLST (Fig. 1b, 2c). Even though split *k*-mers were less sensitive compared to traditional MLST for *F. intonsa* and WGA for *Ovis spp.*, split *k*-mers reconstructed each of the five major lineages (L1-L5) of *F. intonsa*, and species-specific populations of *Ovis spp.* in haplotype networks (Figs. 2b, 2c). SKA2 haplotyping of 607 mtDNA collected from Southeast Florida reveals one major haplotype (n = 522) with smaller, more recently emerging haplotypes within the North American haplogroup (Fig. 2d). A lab colony of *D. citri* from Ithaca, NY originating from samples collected in Florida possessed a unique mtDNA haplotype. Of the samples collected from UC-Davis California (n = 6), one sample already published (OM181945.1) represented its own haplotype while the other five grouped with the major North American haplotype (Fig. 2d). Mapping Asian *D. citri* haplotype locations based on recorded latitude and longitude[24] of samples from *D. citri* GenBank-derived mitogenome sequences shows distinct subpopulations of *D. citri* across Southeast Asia (Fig. S3) which further supported the accuracy of split *k*-mer mtDNA haplotyping.

Studies examining psyllid mitochondrial haplotypes and vectoring ability of Liberibacter species are critical to aid in disease epidemiological modeling and to make on-farm management recommendations[25]. Our analysis revealed that there are multiple *D. citri* mtDNA haplotypes in Florida and that there is a correlation between mtDNA haplotypes of *D. citri* and *C*Las titer for samples collected from the same locations (Fig. S4). Two haplotypes represent individuals with low to no *C*Las titer and four haplotypes represented *D. citri* individuals with high *C*Las titer (Fig. S4). MLST analysis of thirteen mtDNA coding sequences did not detect two *D. citri* haplotypes, 43 and 54, which are both associated with *C*Las titer (Fig S3), providing further evidence for the benefit of split *k*-mers to strengthen epidemiological research. A link between vector haplotype and vectored plant pathogen titer has been reported in other psyllid species. Populations of the potato psyllid, *Batericera cockerelli*, also display distinct haplotypes that are either infected or uninfected with *Candidatus* Liberibacter solanacearum[26,27]. Further research using genome-wide association

methods and/or controlled transmission assays with each haplotype-pathogen genotype pair is required to confirm whether these relationships have a genetic or environmental basis.

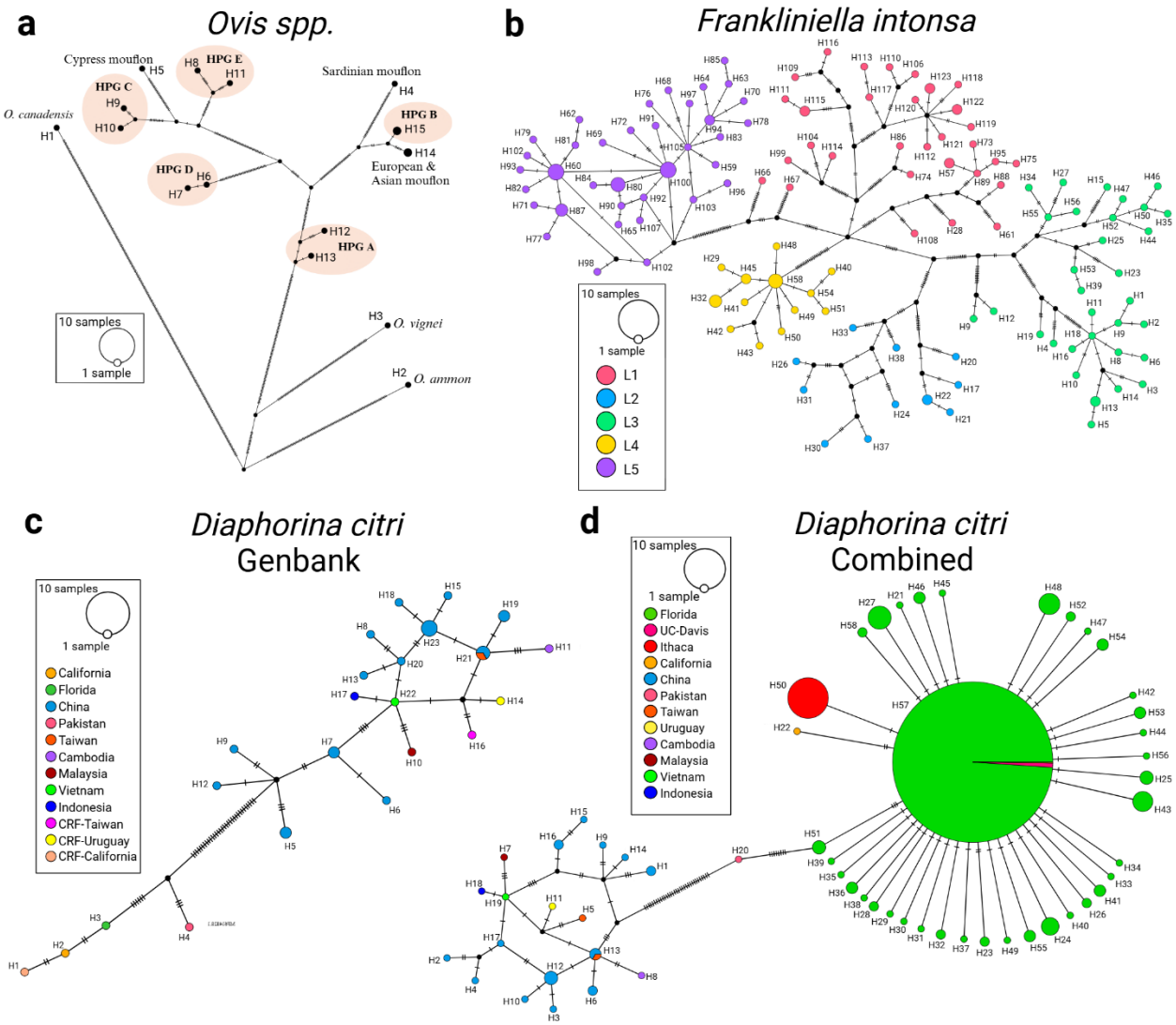

Figure 2. Four haplotype networks obtained using SKA2 *map* and generated in PopART[28] for three different species. Samples within each haplotype can be found in Tables S1-S4. Haplotypes are ranked from most to least divergent. **a.** TCS network of 17 samples using k = 11 and m = 25 from the Ovis Genus labeled by known haplogroups[18]. **b.** Median-joining network using k = 13 and m = 65 of 149 *F. intonsa* mtDNA colored by known taxonomic lineages[17]. **c.** TCS network using k = 15 and m = 90 of 31 *D. citri* mtDNA. **d.** TCS network of 681 *D. citri* mtDNA using k = 15 and m = 90.

The use of *k*-mers provides a foundation for more robust and unambiguous phylogenetic research. Examining large datasets of hundreds of mitogenomes with WGA is memory and time-consuming. Without having to perform read-mapping, gene annotation, multiple sequence alignments, alignment trimming, and gene concatenation split *k*-mers provide a straightforward method to haplotype same species mitochondria identical to or better than

current standard workflows. The ability of split *k*-mers to haplotype mitogenomes greatly simplifies current protocols and empowers researchers to perform rapid, large scale mitogenome haplotyping on small machines with at least 8Gb of memory. While using individuals of the same species, SKA can be a powerful tool. When examining datasets of more diverse individuals, such as with different species, split *k*-mers should be used with caution and will most likely not produce the most resolute haplotype networks. As a general rule, if more than one third of the total nucleotide sites were removed from mtDNA MSAs after aligning or mapping split *k*-mers with the best representative *k* and *m*, the haplotype structure of the dataset should be examined with other methods such as MLST or WGA. We show that for the *Ovis spp.* dataset, WGA significantly increases π and S compared to SKA, but h and Hd differences are not significantly different. For the *F. intonsa* dataset MLST significantly increases π compared to SKA, but h, S, and Hd differences are not significant with n = 149. Lastly, for the *D. citri* Genbank dataset MLST vs. SKA shows no significant differences in h, π, or Hd (S untestable due to unreported value). All statistics and p-values are reported in Table S1. Haplotyping with SKA empowered the first report of an association between *D. citri* haplotype and *C*Las titer, information crucial to disease epidemiology and development of vector management tactics.

Methods

Initial assessment of data and subpopulations was conducted with ska *align* and final parameter optimization was performed with ska *map* (Fig. 1a). All unpublished *D. citri* mitogenomes were assembled using default metaSPAdes[29] after quality filtering reads with FaQCs[30] parameters --mode BWA_plus -q 30 --min_L 50 -n 1 --lc 0.85 --*k*-mer_rarefaction --adapter. Complete mtDNA assemblies were identified through BLASTn to the reference mitogenome *Diaphorina citri* isolate FLpsy mitochondrion KY426015.1. Any mitogenome comprising less than 97% of the average mitogenome assembly length was considered incomplete and removed from this analysis. P-values were calculated for the number of haplotypes (h), nucleotide diversity (π), number of segregating sites (S), and haplotype diversity (hd) to determine if split *k*-mer haplotyping was significantly different from published MLST and WGA methods. For h, we used a two-sample proportion z-test (h/n). For π, we used a Z-test with variance ≈ π / (n(n-1)/2 * L). For S, we used a two-sample proportion z-test (S/L), and for Hd we used a Z-test with variance ≈ Hd(1 - Hd) / n, where n is the number of samples and L is the length of the split *k*-mer alignment.

Optimizing *k, m*, and population distribution

To fully capture the effects of split *k*-mer length (*k*) and the minimal sample fraction (*m*), we examined all lengths of odd *k*-mers from 7 to 63 using the default *m* of 90 and ska *align*[2]. Without having to designate a reference all FASTA alignment files were generated and the impact of *k* was assessed with population diversity statistics such as the number of haplotypes, haplotype diversity, nucleotide diversity, segregating sites, and pairwise nucleotide differences. Candidate *k*-mer lengths detecting the most haplotypes were then used to assess the impact of a lower minimum sample fraction filter (*m*), leading to the inclusion of sequence data found in fewer samples, to review its impact on population diversity statistics.

Each dataset was assessed to find the optimum *k* and *m* values. Population sample counts were normalized within datasets so *k*-mers were not prematurely filtered out by *m*. To reduce the size of overrepresented populations, we examined populations individually with ska *align,* or *map* if the most divergent haplotype could be identified. From the population-specific analysis, one sample from each multi-sample haplotype was selected to represent that haplotype within the whole dataset. This reduced the number of samples representing a population while preserving haplotype resolution.

Examining the impact of sequence diversity

We assessed the impact of diversity by sub setting samples and comparing their haplotype results from WGA or MLST to results of split *k*-mers. Further investigations into the alignment files were performed manually in UGENE[31] to review regions that were labeled as ambiguous, or gaps, and if entire conserved regions such as CDS were gapped the dataset, or samples, were designated as too diverse. An examination of total nucleotide sites available in the FASTA alignment file, compared to original mitogenome assembly lengths, further identified overly diverse datasets.

Distribution of SNPs

To assess the distribution of SNPs for this analysis, we mapped split *k*-mers with optimal values of *k* and *m* for each dataset to the sequence with the longest average predicted genetic distance. The output file in Variant Call Format (VCF) from the ska *map* function was then fed into snippy[32] vcf_to_tab tool along with the reference genome nucleotide files in fasta and gff3 formats to produce a tab-delimited input for Biocircos[13] plotting in R.

Data Availability

All assembled mitogenomes used for this analysis are available on FigShare[33].

Code Availability

Custom scripts and code for this analysis are available on GitHub (https://github.com/dougiestuehler/mtDNA_haplotying).

References

1. Harris, S. R. SKA: Split Kmer Analysis Toolkit for Bacterial Genomic Epidemiology. Preprint at https://doi.org/10.1101/453142 (2018).
2. Derelle, R. *et al.* Seamless, rapid, and accurate analyses of outbreak genomic data using split *k* -mer analysis. *Genome Res.* **34**, 1661–1673 (2024).
3. Maechler, F. *et al.* Split k-mer analysis compared to cgMLST and SNP-based core genome analysis for detecting transmission of vancomycin-resistant enterococci: results from routine outbreak analyses across different hospitals and hospitals networks in Berlin, Germany. *Microb. Genomics* **9**, (2023).
4. Higgins, S. A., Mann, M. & Heck, M. Strain Tracking of ‘ *Candidatus* Liberibacter asiaticus’, the Citrus Greening Pathogen, by High-Resolution Microbiome Analysis of Asian Citrus Psyllids. *Phytopathology®* **112**, 2273–2287 (2022).
5. Gray, M. W., Burger, G. & Lang, B. F. Mitochondrial Evolution. *Science* **283**, 1476–1481 (1999).
6. Wang, Z. & Wu, M. An integrated phylogenomic approach toward pinpointing the origin of mitochondria. *Sci. Rep.* **5**, 7949 (2015).
7. Steenwyk, J. L., Buida, T. J., Li, Y., Shen, X.-X. & Rokas, A. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLOS Biol.* **18**, e3001007 (2020).
8. Tumescheit, C., Firth, A. E. & Brown, K. CIAlign: A highly customisable command line tool to clean, interpret and visualise multiple sequence alignments. *PeerJ* **10**, e12983 (2022).
9. Tan, G. *et al.* Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. *Syst. Biol.* **64**, 778–791 (2015).

10. Castresana, J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).

11. Yu, H.-J. Segmented K-mer and its application on similarity analysis of mitochondrial genome sequences. *Gene* **518**, 419–424 (2013).

12. Rozas, J. *et al.* DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol. Biol. Evol.* **34**, 3299–3302 (2017).

13. Cui, Y. *et al.* BioCircos.js: an interactive Circos JavaScript library for biological data visualization on web applications. *Bioinformatics* **32**, 1740–1742 (2016).

14. R Core Team. A language and environment for statistical computing. *R Found. Stat. Comput. Vienna Austria* (2021).

15. Joly, S., Stevens, M. I. & Van Vuuren, B. J. Haplotype Networks Can Be Misleading in the Presence of Missing Data. *Syst. Biol.* **56**, 857–862 (2007).

16. Carlson, C. R. *et al.* High-quality, chromosome-scale genome assemblies: comparisons of three Diaphorina citri (Asian citrus psyllid) geographic populations. *DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes* **29**, dsac027 (2022).

17. Liu, X. *et al.* Weak genetic structure of flower thrips *Frankliniella intonsa* in China revealed by mitochondrial genomes. *Int. J. Biol. Macromol.* **231**, 123301 (2023).

18. Mereu, P. *et al.* Identification of an ancestral haplotype in the mitochondrial phylogeny of the ovine haplogroup B. *PeerJ* **7**, e7895 (2019).

19. Aregbe, I., Farnsworth, D. & Simnitt, S. Citrus Greening, Hurricanes, and the Decline of the Florida Citrus Industry. *South. Ag Today* (2024).

20. Lapierre, P., Lasek-Nesselquist, E. & Gogarten, J. P. The impact of HGT on phylogenomic reconstruction methods. *Brief. Bioinform.* **15**, 79–90 (2014).

21. Neiman, M. & Taylor, D. R. The causes of mutation accumulation in mitochondrial genomes. *Proc. R. Soc. B Biol. Sci.* **276**, 1201–1209 (2009).

22. Van Den Ameele, J., Li, A. Y. Z., Ma, H. & Chinnery, P. F. Mitochondrial heteroplasmy beyond the oocyte bottleneck. *Semin. Cell Dev. Biol.* **97**, 156–166 (2020).

23. Shao, R. Rates of Gene Rearrangement and Nucleotide Substitution Are Correlated in the Mitochondrial Genomes of Insects. *Mol. Biol. Evol.* **20**, 1612–1619 (2003).

24. Wu, F. *et al.* Population diversity of *DIAPHORINA CITRI* (Hemiptera: Liviidae) in China based on whole mitochondrial genome sequences. *Pest Manag. Sci.* **74**, 2569–2577 (2018).

25. Yao, J. *et al.* Interactions "*Candidatus* Liberibacter solanacearum"—Bactericera cockerelli: Haplotype Effect on Vector Fitness and Gene Expression Analyses. *Front. Cell. Infect. Microbiol.* **6**, (2016).

26. Swisher, K. D., Munyaneza, J. E. & Crosslin, J. M. High Resolution Melting Analysis of the Cytochrome Oxidase I Gene Identifies Three Haplotypes of the Potato Psyllid in the United States. *Environ. Entomol.* **41**, 1019–1028 (2012).

27. Swisher, K. D., Henne, D. C. & Crosslin, J. M. Identification of a fourth haplotype of *Bactericera cockerelli* (Hemiptera: Triozidae) in the United States. *J. Insect Sci. Online* **14**, 161 (2014).

28. Leigh, J. W. & Bryant, D. POPART : full-feature software for haplotype network construction. *Methods Ecol. Evol.* **6**, 1110–1116 (2015).

29. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

30. Lo, C.-C. & Chain, P. S. G. Rapid evaluation and quality control of next generation sequencing data with FaQCs. *BMC Bioinformatics* **15**, 366 (2014).

31. Okonechnikov, K., Golosova, O., Fursov, M., & the UGENE team. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **28**, 1166–1167 (2012).

32. Seemann, T. snippy: fast bacterial variant calling from NGS reads. (2015).

33. Stuehler Jr., D. *Diaphorina citri* Combined dataset SKA mtDNA haplotyping. 10424980 Bytes figshare https://doi.org/10.6084/M9.FIGSHARE.27993887.V1 (2024).

Acknowledgements

We thank Dr. Steven Higgins and Dr. Marina Mann for their assistance in collecting and processing *Diaphorina citri* samples for Illumina short-read sequencing.

This research is supported by U.S. Department of Agriculture (USDA) Agricultural Research Service Project # 8062-22410-007-000-D and the Emergency Citrus Disease Research and Extension Program (ECDRE) projects 2020-70029-33176 and 2022-70029-038503 from the USDA National Institute of Food and Agriculture. This research was also supported in part by an appointment to the Agricultural Research Service Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the USDA. ORISE is managed by ORAU under DOE contract number DE-SC0014664. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of DOE, or ORAU/ORISE.

Author Information

Authors and Affiliations

Oak Ride Institute for Science and Education, Department of Energy, Emerging Pests and Pathogens Research Unit, Ithaca, NY 14853 USA

Douglas S. Stuehler Jr.

Indian River Research and Education Center, Department of Plant Pathology, Institute of Food and Agriculture Sciences, University of Florida, Fort Pierce, FL 34945 USA

Douglas S. Stuehler Jr., Liliana M. Cano

Section of Plant Pathology and Plant Microbe Biology, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853 USA

Michelle Heck

Emerging Pests and Pathogens Research Unit, USDA Agricultural Research Service, Ithaca, NY 14853 USA

Michelle Heck

Contributions

Conceptualization: D.S.S. and M.H. Data curation: D.S.S, M.H., L.M.C. Formal analysis: D.S.S. Funding acquisition: M.H. Investigation: D.S.S., M.H. Methodology: D.S.S., M.H. Project administration: M.H., L.M.C. Resources: M.H., L.M.C. Software: D.S.S. Supervision: M.H., L.M.C. Validation: D.S.S., M.H., L.M.C. Visualization: D.S.S., M.H., L.M.C. Writing - original draft: D.S.S and M.H. Writing - review and editing: D.S.S, M.H., L.M.C.

Corresponding author

Correspondence to Michelle Heck

Ethics declarations

Competing interests

The authors declare no competing interests.