Alignment-free methods for polyploid genomes: Quick and reliable genetic distance estimation

Acer VanWallendael | Mariano Alvarez

1Department of Plant Biology, Michigan State University, East Lansing, MI, USA
2Biology Department, Wesleyan University, Middletown, CT, USA

Correspondence
Acer VanWallendael, Department of Plant Biology, Michigan State University, East Lansing, MI, USA.
Email: vanwaax07@gmail.com

Abstract
Polyploid genomes pose several inherent challenges to population genetic analyses. While alignment-based methods are fundamentally limited in their applicability to polyploids, alignment-free methods bypass most of these limits. We investigated the use of Mash, a k-mer analysis tool that uses the MinHash method to reduce complexity in large genomic data sets, for basic population genetic analyses of polyploid sequences. We measured the degree to which Mash correctly estimated pairwise genetic distance in simulated haploid and polyploid short-read sequences with various levels of missing data. Mash-based estimates of genetic distance were comparable to alignment-based estimates, and were less impacted by missing data. We also used Mash to analyse publicly available short-read data for three polyploid and one diploid species, then compared Mash results to published results. For both simulated and real data, Mash accurately estimated pairwise genetic differences for polyploids as well as diploids as much as 476 times faster than alignment-based methods, though we found that Mash genetic distance estimates could be biased by per-sample read depth. Mash may be a particularly useful addition to the toolkit of polyploid geneticists for rapid confirmation of alignment-based results and for basic population genetics in reference-free systems or those with only poor-quality sequence data available.

KEYWORDS
alignment-free, polyploid, population genetics, restriction-associated digest

1 | INTRODUCTION

By some estimates, 80% of living plants are polyploid, including several essential crop species such as wheat, potato and cotton (Kyriakidou et al., 2018). However, most population genetic methods and tools are designed for diploid markers, prompting researchers in polyploid plants to adopt novel analysis pipelines or force polyploid genomes into diploid analyses. The in silico reduction of polyploid genotypes to diploid ones leads to disposing of informative variation. In addition, it commonly leads to biases and systematic errors in the processing of polyploid genomes (Blischak et al., 2018). While there has been some progress in methods for the analysis of polyploid genomes (Flagel et al., 2019; Meirmans et al., 2018), many methods remain cumbersome, particularly for the analysis of high-throughput sequencing data.

One of the primary goals of population genetic analyses is to estimate population structure by determining the relatedness between individuals. In single nucleotide polymorphism (SNP)-based studies, this involves three basic steps: aligning sequencing reads, calling SNPs and estimating relatedness. Polyploid genomes present challenges to each of these steps. At the alignment step, the first issue is depth of coverage. Researchers frequently sequence genomic DNA to high depths of coverage in order to ensure full representation of the genome. However, a tetraploid genome will require twice the depth of coverage as a diploid, so will require a greater cost to achieve sufficient depth, or will result in fewer useful SNPs.
A second, related issue arises when calling SNPs, termed “allele dosage uncertainty” (Blishchak et al., 2018; Dufresne et al., 2014). Next-generation sequencing platforms, particularly commonly used short read sequencing technologies, assemble many overlapping reads per locus, which gives confidence in determining when a single base pair difference is a true SNP rather than a sequencing error. Diploid genomes can have a relatively high minor allele frequency (MAF) cut-off to distinguish heterozygotes from sequencing errors, but in polyploids low MAFs often represent informative variation on particular rare alleles. The random nature of sequencing depth means that it is virtually impossible to distinguish between categories of allele frequencies. For instance, an octoploid locus with alleles A and B could have a genotype ABBBBBBB, which should have an MAF of 0.125. The genotype AAABBBBB should have an MAF of 0.25. If a sequenced locus’s MAF were 0.18, for instance, it would be challenging to determine which is the true genotype. This challenge is compounded by the fact that fewer high-quality genomes have been assembled for polyploids, particularly for higher ploidy levels.

Finally, even if SNPs have been properly genotyped, calculating population-level statistics and individual relatedness offers additional challenges. Many of the common expectations of population genetic theory are built on the implicit assumption of diploidy, so many widely used statistics are not particularly useful for polyploids. For example, F-statistics are a cornerstone of population genetics. Owing to a greater number of potential heterozygous states, polyploids typically have high heterozygosity ($H_s$). This results in systematically underestimating population differentiation when using $F_{st}$ (Meirmans et al., 2018). Even where commonly used statistics may apply to polyploid data sets, there are practical concerns as well. Most bioinformatic tools require biallelic SNPs, so polyploid data analysis must produce custom analysis scripts that are challenging for reviewers to evaluate or constrain analyses to the few tools that accommodate polyploids.

Solutions are being continuously developed to overcome the challenges of polyploid genome analysis. Alternatives to F-statistics have been developed that are less impacted by ploidy. One of the most useful is $\varrho$ (Ronfort et al., 1998), which is equivalent to the average relatedness of individuals within populations vs. the average relatedness between populations. $\varrho$ is comparable between ploidy levels (Meirmans et al., 2018), but has not been applied widely in empirical studies. For certain applications, unsupervised machine learning methods have shown great promise in identifying population genetic parameters using methods that are applicable to polyploids (Flagel et al., 2019). These methods are agnostic to many of the assumptions of other population genetic models, but the requirement of extensive training data mean they will not be applicable for all systems.

Currently, there are few resources that are both robust to the large degree of variation inherent to polyploid genomes, and repeatable enough to provide useful benchmarks between population genetic analyses. Since many of the challenges with polyploids come from the alignment and SNP-calling steps, an ideal solution would completely bypass both of these. Many alignment-free sequence analysis techniques have been developed in recent years, mostly using “words” or k-mers, subsamples of sequences with length k. Analyses that use k-mers add flexibility to sequence analysis and can better make use of computing power than alignment-based methods (Leimeister et al., 2014; Vinga & Almeida, 2003). The k-mer methods have already been applied to polyploid analyses with success (Ranallo-Benavidez et al., 2020), and so are a promising tool for population genetics in polyploid systems. There are many ways to compare k-mer decompositions of sequence data, but one of the most promising is the use of the MinHash algorithm (Broder, 1997). MinHash is a type of locality-sensitive hashing, an algorithmic technique to collapse similar input items into short strings, or hashes, and then determine how much they resemble each other (Broder, 1997). MinHash typically uses a Jaccard similarity coefficient to quickly assess how similar sets of hashes are. For sequence data, raw sequence reads for a sample can be decomposed into k-mers tables, which are then hashed and can be quickly compared between individuals. In practice, this process is many times faster than sequence alignment—researchers were able to cluster all of the organisms in the NCBI RefSeq database (54,118 organisms, 618 Gbp; O’Leary et al., 2016) in 35 CPU hr without parallelization (Ondov et al., 2016). The dominant sequence-based MinHash implementation is known as Mash (Ondov et al., 2016). Mash and similar tools have been used widely in microbial metagenomics (Forster et al., 2019; Hendriksen et al., 2019; Ondov et al., 2016), but have not been incorporated into population genetics, despite their flexibility and wide applicability.

In this paper we evaluate the performance of Mash to estimate pairwise sample relationships within and across populations and species in polyploids. First, we simulate haploid and polyploid DNA sequences with known phylogenetic relationships, and show that Mash accurately recovers cladograms for these samples. Next, we remove reads from these simulated data sets to show the impact of missing data on Mash estimation relative to alignment-based analysis, and find that Mash is more robust to missing data than alignment at estimating pairwise relationships. Finally, in several case studies, we use Mash to estimate kinship in reduced-representation sequencing reads for the polyploid genera Capsella and Reynoutria, as well as whole-genome data in polyploid Panicum and diploid Oryza.

2 METHODS

2.1 Simulated data

We first evaluated the performance of Mash using simulated haploid and polyploid sequence data. We simulated phylogenetic trees using toytree (Eaton, 2020) in Python 3.7 for 50 individuals using a tree height of 1e6, then simulated SNP loci following these trees by generating sequences in ipcoal (McKenzie & Eaton, 2020). For haploid data, we simulated 1000 loci with 100 bp each for 50 individuals, diverging with an effective population size of 1e5, and a recombination rate of 1e-9. We simulated polyploid data in an allotetraploid scenario, wherein two ancestrally diverged lineages hybridize, then diversify into population groups. To accomplish this, we
built a phylogenetic tree with a single deep node (7e5 generations), and two symmetrical trees with shallower nodes (Figure 1), which represent the mean locus tree after polyploidization. We simulated 1000 loci for each of 30 individuals using the same parameters listed above, then combined loci from symmetrical branches to create homoeologous polyploid loci. This simulation method does not account for incomplete lineage sorting, wherein individual locus trees do not match the true population genetic history, but effectively simulates the problem of similar homoeologous loci that can confound alignment-based methods. To simulate missing data, we used custom scripts in R to randomly remove sequence data for 5%–75% of reads (R Development Core Team, 1999). We compared Mash and alignment-based analyses for each level of missing data.

2.1.1 Estimating genetic distance

We first calculated the true distance between individuals as the proportion of nucleotide differences ($\pi$) in the simulated sequence data. We then compared these true distances to distances inferred by alignment-based and alignment-free methods. To provide the simplest comparison of methods, we ran a naive pipeline for both alignment-based and Mash methods, using the default parameters throughout. We aligned reads using bwa and samtools (Li & Durbin, 2009; Li et al., 2009), then called SNPs using bcftools (Danecek et al., 2013; Li & Durbin, 2009) to align to the switchgrass version 1.0 reference genome (Lovell et al., 2021), then called SNPs using the stack command in the stacks processing pipeline (Catchen et al., 2013; Li & Durbin, 2009). We calculated the genetic distance between samples using the dist.genpop function in adegenet, then built simple trees of these distances using hclust in base R, with the “complete linkage” setting.

For Mash, we first sketched the simulated reads using mash sketch, which uses MinHashing to produce a reduced-complexity representation of the read set. We then estimated genetic distances between individuals using mash dist. We imported pairwise Mash distances into R and built trees again using hclust. We calculated the true distance between individuals as the proportion of nucleotide differences ($\pi$), and compared this to missing data distance matrices using Mantel tests. We also compared the true tree to each tree inferred from missing data using the comparePhylo function in ape (Paradis et al., 2004).

2.2 Polyploid genomes

To evaluate the performance of Mash on real data, we compared Mash-derived distance matrices to published data in population genetics studies in polyploid species from genera Capsella and Reynoutria (Cornille et al., 2016; VanWallaendaal et al., 2021). Capsella bursa-pastoris is an autotetraploid, while Reynoutria species may be tetraploid, hexaploid or octaploid. We downloaded data for these case studies from the SRA (Capsella: PRJNA299253; Reynoutria: PRJNA574173). Data in both case studies were generated using an Illumina HiSeq 2500. To account for sequencing errors, we removed unique k-mers with counts under 2, used a k-mer length of 21, and increased the number of k-mers per sketch to 1e6. Although we did not test the effect of increasing the minimum number of k-mers, we chose to filter as little as possible to preserve the diversity of possible sequences, as we expect that reduced-representation data would generate many true, but low-copy k-mers. We used the same analysis tools as for simulated data, and visualized sample relationships using principal coordinates analysis in R. Although Mash restricts comparisons between sketches to the size of the smaller sketch, we found that per-sample read count biased Mash analyses, so we randomly removed reads from the raw sequence files to standardize inputs using fastq-tools (https://github.com/djones/fastq-tools).

While these examples are useful for ensuring that Mash can at least match the performance of alignment-based methods, it is not clear in these reference-free studies how well either method estimates true population relationships. To benchmark the performance of Mash against known genetic relationships, we compared whole genome-sequencing data for 16 individuals in the allotetraploid Panicum virgatum (switchgrass; SRA Bioproject PRJNA622568; Lovell et al., 2021) to simulated reduced-representation reads from those same sequences. To generate known relationships, we calculated the genetic distance between individuals using 11 million SNPs. We then simulated reduced-representation sequence data by trimming whole-genome sequences to reads that contain a typical restriction-associated digest (RAD) restriction enzyme cut site (EcoRI). Using the whole-genome SNP data as “true” population relationships, we then evaluated how well Mash and alignment-based methods approximate these relationships. For sequence alignments, we used bwa (Li & Durbin, 2009) to align to the switchgrass version 5 reference genome (Lovell et al., 2021), then called SNPs using the stacks ref_map.pl RAD-processing pipeline (Catchen et al., 2013; Li & Durbin, 2009). We called 783 SNPs from this pipeline, and estimated the genetic distance between aligned samples again using the dist.genpop function in adegenet. For Mash analyses, we calculated...
genetic distances from whole-genome sequences as well as simulated RAD sequences. We used the same parameters as for the case studies: removing k-mers with counts under 2, a k-mer length of 21 and increasing the number of k-mers per sketch to 1e6.

2.3 | Reference diploid genome

To assess the performance of Mash on diploid data, without the implicit biases and complications of polyploid, we also analysed a previously published data set of 52 rice varieties (Zheng et al., 2017; SRA SRP080763). These varieties were chosen to represent the pedigrees of several elite lines, and therefore were accompanied by pedigree information for each line. This data set allowed us to assess the performance of Mash in recapitulating a known relatedness structure. k-mers were sketched and analysed as described above.

2.4 | Computing

All analyses except the reference diploid genome were run on the high-performance computing cluster at the Institute for Cyber-Enabled Research at Michigan State University (two 2.4-GHz 20-core Intel Xeon Gold processors). Compute time estimates were gathered by running analyses using five CPUs and 20 Gb of RAM. The analysis for the referenced diploid genome was run on a personal laptop (MacBook Pro, one 2.2-GHz 4 core Intel i7 processor).

3 | RESULTS

3.1 | Simulated data

For simulated data, we found that Mash correctly reconstructed phylogenetic relationships for all individuals in the populations. The genetic distances were very closely correlated to Mash distance (Mantel test $r = 0.9606$, $p < .001$) for haploid data, and only slightly less tightly correlated for polyploid samples ($r = 0.9332$, $p < .001$). Notably, this concordance was achieved with greatly less compute time; Mash analysis for these samples used 6.2 s of compute time, whereas alignment used 2893.6 s.

Mash analyses and alignment-based analyses differed greatly when we introduced missing data (Figures 1 and 2). Alignment-based analyses were robust to lower levels of missing data. At 0%-5% missing data, alignment-based methods outperform Mash estimation. However, alignment-based distances deviated more greatly from true distances at higher levels (Figure 1). Between 10% and 60% missing data, Mash typically outperformed alignment-based methods, and showed less variation in concordance to true genetic distances (Figure 1). When greater than 60% of data was missing, both methods more poorly predicted genetic distance and showed greater variation.

To better understand how deviations in the pairwise genetic distance estimates can alter group assignment, we compared hierarchical clustering trees between Mash and alignment-based pipelines (Figure 2). While the average Mantel’s $r$ for Mash analyses at 40% missing data was only slightly higher than alignment ($r = 0.954$ vs. $r = 0.942$, Figure 1), the errors in genetic distance had a greater impact on phylogenetic reconstruction. The Mash trees typically showed errors in shallow nodes only, whereas there were rearrangements even in deep nodes for trees reconstructed from the alignment-based method (Figure 2).

3.2 | Reduced-representation polyploid genomics

We reanalysed data from two studies of polyploid species complexes that used genotyping-by-sequencing (Cornille et al., 2016; VanWallendael et al., 2021). Using Mash we were able to reconstruct population groups for both studies (Figure 3a, b). As we discuss further below, there was some bias in Mash analyses in samples with greatly divergent read counts, but we were able to adjust by randomly trimming reads from raw sequence data. For Capsella, Mash showed five population clusters, rather than the three found by ADMIXTURE, but showed high similarity to reconstructed population distributions despite using different methods (Figure 3a). We compared compute time for Reynoutria between Mash and alignment-based methods. To generate distance matrices from raw sequence reads for 25 Gb of data took 0.233 compute hours in Mash and 108 compute hours for alignment. We could not benchmark Capsella alignment-based analyses, but Mash analysed 44.3 Gb of data in 0.410 compute hours. Since sketching is easily multithreaded, most computers can do this in a fraction of the time.

To evaluate the performance of Mash against "known" population relationships in real sequence data, we compared the performance of Mash distances against alignment-based methods for simulated RAD data. We found that Mash effectively reproduced population relationships in Panicum virgatum for both whole-genome and simulated RAD data (Figure 3c). Comparatively, the alignment-based method performed poorly, showing much less difference between known subpopulations than Mash of RAD sequences (Figure 3c; Lovell et al., 2021).

3.3 | Whole-genome diploid analysis

When assessing diploid Oryza sativa individuals, Mash correctly recapitulated the known relationships between rice varieties (Figure 4). Patterns of population structure across the first two multidimensional scaling axes accurately separated HHZ+GC2H individuals from SH527, while maintaining outliers shown in the original plot, demonstrating the accuracy of MinHash methods in capturing patterns of variation in diploids as well as polyploids. As in previous analyses, we noted a minor effect of sequencing depth. Although this issue did not affect the ability of Mash to assess overall
population structure, it indicates that depth correction may improve Mash-derived estimates of population structure. Although we did not test these methods here, other authors have explored the use of additional bias-correction steps, which improve distance calculations (e.g., Tang et al., 2019).

DISCUSSION

This study showed that Mash performs well at assessing genetic relationships between polyploid individuals in a fast, naive and repeatable manner for both simulated and real data. For simulated data, we showed that Mash often outperforms alignment-based methods in distance estimation when missing data is 10% or greater (a reasonable amount of missing data, even in whole genome sequencing projects) and is orders of magnitude faster. For real data, we have shown that Mash can reproduce population structure estimates in both polyploids and diploids, though per-sample read count should be standardized before analysis, as we found a significant bias of read count, or library size, on estimated Mash distances. Mash should therefore be considered a useful tool for researchers investigating population relationships in their polyploid samples.

FIGURE 2 Different types of phylogenetic errors in alignment vs. Mash-based analysis with 40% missing data of simulated polyploid data. The left panel shows the Mash-inferred tree on the left compared to the true tree used to simulate reads on the right. The right panel shows the alignment-inferred tree compared to the true tree on the right.

FIGURE 3 Comparisons of reference-aligned and Mash analyses in three genera: Capsella, Reynoutria and Panicum. The left panels of each section show SNP-based representations of population relationships, while the right panels show re-creations from Mash analyses. (a) Global diversity in Capsella bursa-pastoris; left panel reproduced with permission from Cornille et al. (2016). (b) Diversity in Reynoutria spp., copied from VanWallendael et al. (2021). (c) Whole-genome and simulated RAD data for Panicum virgatum. Data from NCBI SRA PRJNA622568.
(a) *Capsella bursa-pastoris* (Cornille et al. 2016)

(b) *Reynoutria* spp. (VanWallenael et al. 2021)

(c) *Panicum virgatum* (Lovell et al. 2021)
particularly when missing data is high, or when rapid population relationship assessment is necessary.

Removing poorly sequenced samples is a typical step in many bioinformatics pipelines, but variation in per-sample read count in downstream steps has been less studied. In many studies, only loci sequenced to a high depth are considered, so including individual samples with poor sequencing may only result in fewer loci in the final data set. However, in alignment-free methods that consider the entire sequencing set, there may be more pernicious impacts of depth variation. To our knowledge, no other study has reported this source of bias in alignment-free methods. While further study is needed to definitively assess the cause, there are several possibilities. One cause may be that poorly sequenced samples are simply missing unique sequences that are present in deeper-sequenced samples. This reduction in genetic diversity means that poorly sequenced samples appear more similar than they really are. However, in our data, we observed that randomly subsampling individuals to equal library sizes corrects the problem. This should not be the case if the issue is simply missing genetic variants. We suggest that future studies explore the effects of normalization by library size, which has been described extensively in the RNA sequencing literature (Conesa et al., 2016; Love et al., 2014; Robinson & Oshlack, 2010). Possible strategies for normalization include trimmed mean of M-values (TMM) normalization or model-based normalization in subsequent analysis steps.

4.1 Comparison with other alignment-free methods

A number of alignment-free methods using similar methods to perform k-mer hashing were reviewed by Zielezinski et al. (2019). Alternative methods are based on the length of common substrings (Ulitsky et al., 2006), micro-alignments (Leimeister et al., 2014; Ulitsky et al., 2006; Zielezinski et al., 2019), sequence representations based on chaos theory (J. S. Almeida et al., 2001), moments of the positions of the nucleotides (Yau et al., 2008, 2008), Fourier transformations (Yin & Yau, 2015), information theory (Vinga, 2014) and iterated-function systems (Almeida, 2014). Of these, only Mash, AFKS (Alignment-Free-Kmer-Statistics; Luczak et al., 2019; Vinga, 2014) and FFP (Choi & Kim, 2017) were generic enough to perform protein classification, genome-based phylogenies and detection of horizontal gene transfer (Zielezinski et al., 2019). Mash underperformed methods such as andi and phylonium developed for closely related organisms at phylogenetic analysis of bacteria, but was among the top three tools for recovering tree topology in plant species (Zielezinski et al., 2019). For a more thorough review of these methods, see Zielezinski et al. (2019).

Several recent implementations have built on the Mash algorithm for specific uses. Mashtree is a Mash wrapper that automatically builds neighbour-joining trees from Mash outputs using quicktree (Howe et al., 2002; Katz et al., 2019). Kmer-db reports even faster compute times than Mash, largely through an improved k-mer hash and parallel implementation (Deorowicz et al., 2019), sourmash adds some functionality to Mash and implements scaled, rather than thresholded, numbers of retained k-mer hashes (Pierce et al., 2019). This method can enable comparisons of data sets that differ greatly in size, but slows down analysis (Pierce et al., 2019). We checked the performance of these other tools against Mash for the simulated data in this study (data not shown), but neither sourmash nor kmer-db showed a clear advantage over Mash.
4.2 | Potential applications for k-mer hashing-based methods

As the quantity of available sequence data continues to increase exponentially, new tools that can rapidly process these data are needed (Barone et al., 2017). Locality-sensitive hashing has been used for data-intensive jobs such as clustering metagenomes (Rasheed et al., 2013) and genome assembly (Berlin et al., 2015; Rasheed et al., 2013). The use of k-mer methods in molecular ecology has been comparatively limited, perhaps because historically this field has used less data-intensive approaches. However, now that genomics are more feasible for many molecular ecology researchers, k-mer hashing may solve some of the scaling issues that arise. In particular, three main directions will be facilitated by Mash and similar k-mer hashing methods: population genetics for polyploids, reference-free genome-wide association studies (GWAS), and genomics of ancient and otherwise degraded samples.

4.2.1 | Population genetics for polyploids

Historically, population genetics for polyploids has been difficult, due to both the technical difficulty of accurately genotyping loci as well as the assumptions of diploidy in commonly used analyses and tools (Daetwyler & Abbott, 2020; Dufresne et al., 2014; Meirmans et al., 2018). k-mer-based methods are advantageous for this purpose, as they do not rely on accurately assessing genotype likelihood at each locus. Instead, the count data that they produce implicitly takes dosage into account, as different genotypic combinations are represented as different k-mer counts. We also note that because Mash relies on count-based data, it should work in principle for both long- and short-read data, assuming both reach sufficient depth of sequencing. In Mash, these k-mer counts are used to create distances between individuals, which can be used without further assumptions to assess differentiation in polyploid systems, particularly through eigenvector-based visualization such as principal coordinates analysis or multidimensional scaling. Further, k-mer methods represent a new path forward for assessing population differentiation, as many existing methods, including faststructure (Raj et al., 2014) and admixture (Alexander et al., 2009), assume diploidy. Future efforts should investigate the use of clustering algorithms on the distance matrices produced by Mash and similar methods.

Although we did not explicitly test Mash on autoploid data, we expect its performance to be similar. Autoploids (and some allopolyploids) display polysomic inheritance, complicating both genotyping and the estimation of certain population genetic parameters (see Huang et al. for details). However, because Mash does not explicitly genotype a locus and because it appears robust to various demographic histories, we expect that Mash will perform similarly between allo- and autoploid systems. However, we caution the use of k-mer methods in assessing differences between mixed-ploidy populations, as they probably suffer some of the same biases as other distance-based measures (Meirmans et al., 2018).

Assessing population differentiation in polyploid systems is also complicated by the effects of polyploidy on commonly used statistics, such as $F_{ST}$ (Meirmans et al., 2018). Alternative measurements that are useful in polyploid systems (outlined by Meirmans et al., 2018) rely on partitioning variation within and between populations without making assumptions about the nature of individual genotypes. The multivariate analog of $F_{ST}$ is $\Phi_{ST}$, which relies on distance-based assessments of relatedness (Bird et al., 2011). Although we do not describe their use here, the accuracy of k-mer based methods suggests that they may also be useful for generating $\Phi_{ST}$ and other genome-wide estimates between populations, particularly in polyploid systems. We suggest that future efforts be dedicated to assessing the accuracy of k-mers for this task.

To date, we can find no published implementations of Mash or similar methods specifically for polyploid data, despite clear evidence that alignment-based methods pose significant challenges. Solutions to these challenges have mostly focused on generating high-quality polyploid reference genomes (e.g., Edger et al., 2019) or on improving alignment-based methods (Blischak et al., 2018; Clark et al., 2019; Gerard et al., 2018). However, a recent pair of methods, GenomeScope and Smudgeplot, use a k-mer frequency-based method to estimate genome characteristics and visualize broad genome structure of polyploids (Ranallo-Benavidez et al., 2020).

4.2.2 | Genetic and structural variants

Traditionally, population genetics has used a small sample of loci from a genome to infer the history of the organism. A major issue with this method has always been that genetic changes that accompany demographic changes may be missed when using only relatively few neutral loci, particularly when these changes are accompanied by selection (Lowry et al., 2017). Even methods that use many millions of SNPs often miss structural changes in the genome that are increasingly linked to important evolutionary changes (Wellenreuther et al., 2019). While long-read sequencing can improve detection of structural variants, k-mers will also show this variation even in short-read data through presence and absence (Voichek & Weigel, 2020). As molecular ecology studies increasingly use whole-genome coverage to assess population-level relationships, k-mer hashing can allow for vastly reduced compute times and improved detection of both small and structural variation. So far, this method has mostly been applied to population genetics of microbial taxa (Ondov et al., 2016; Kachroo et al. 2019), but has the potential to be useful as more population geneticists adopt whole-genomic methods. In this study, we have shown that Mash can accurately reconstruct population relationships from whole-genome data (Figure 3c), though we did not specifically investigate the importance of structural variants.

4.2.3 | Reference-free GWAS

A breakthrough study recently showed that k-mer hashes can be used in the place of SNPs to perform reference-free genome-wide
analysis (GWA; Voichek & Weigel, 2020). The authors showed that k-mer-based GWAS outperformed SNP-based methods, particularly in confidence of detection of causal variants and in the number of phenotypes linked to structural variants (Voichek & Weigel, 2020). Reference-free GWA is not completely new, but early attempts used computationally inefficient methods that are not scalable to large, complex genomes (Lees et al., 2016). As k-mer counts accurately represent complex polyploid genotypes in an efficient way, they pave the way for large-scale GWAS in polyploid systems. For example, just as k-mer counts replace genotypes in the linear mixed models commonly used for GWAS, k-mer-derived distance matrices can replace the relatedness matrix that is used to control for population structure in the same analyses. Critically, neither of these methods rely on an assumption of diploidy.

4.2.4 | Genomics of ancient and degraded samples

A major finding in this study is that Mash outperforms alignment-based methods when samples have missing data. This leads to similar issues as those posed in polyploid genomes. Since most phylogenetic analyses use Bayesian methods, not distance-based hierarchical clustering, this may not be an issue in that field. Phylogenetic analyses have indeed been found to be largely robust to missing data (Hovmöller et al., 2013), but estimation of specific population genetic parameters can be more impacted (Chattopadhyay et al., 2014; Hovmöller et al., 2013). Ancient DNA (aDNA) studies are particularly prone to problems with missing data, since long-degraded samples often cannot yield full SNP coverage. Recent aDNA studies have included samples with as much as 40% (Marcus et al., 2020) or 50% missing data (Anava et al., 2020), while others removed SNPs with missingness >10% (Sikora et al., 2014), leading to a greatly reduced genomic coverage. Since these studies are particularly challenging to replicate given the barriers to sample collection, it is especially concerning that missing data may interfere with conclusions. k-mer hashing analysis may offer a confirmatory method for aDNA studies.

5 | CONCLUSIONS

Mash and similar alignment-free methods hold great promise for overcoming some of the challenges in both polyploid and diploid genomics. In situations where alignments are particularly fraught, such as in systems with a high level of missing data, large genomic arrangements or data-heavy whole-genome comparisons, the simplicity and speed of these methods may offer at minimum a confirmation of alignment-based results. Other methods will need to be tested against Mash to determine which applies for particular population genetic studies. Similarly, combining machine learning with k-mer-based analyses may hold great promise for fast and reliable population genetic estimation with fewer assumptions than commonly used methods.

ACKNOWLEDGEMENTS

Much thanks to Rafael Gutaker for the initial idea to use Mash for polyploid genomics. We would like to also thank the authors of Cornille et al. (2016), Zheng et al. (2017) and Lovell et al. (2021) for use of their sequences from the SRA.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

A.V. and M.A. designed the study, performed analyses and wrote the manuscript.

OPEN RESEARCH BADGES

This article has earned an Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://doi.org/10.5061/dryad.fqz612jss and https://github.com/avanwallendael/mash_sim.

DATA AVAILABILITY STATEMENT

Data and code for all analyses will be made available on Github: https://github.com/avanwallendael/mash_sim, and simulated loci are available on Dryad: https://doi.org/10.5061/dryad.fqz612jss. External data sets were downloaded from the NCBI SRA Bioprojects PRJNA622568, PRJNA299253 and PRJNA574173.

ORCID

Acer VanWallengael https://orcid.org/0000-0002-3900-2818

REFERENCES

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 19(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Almeida, J. S. (2014). Sequence analysis by iterated maps, a review. Briefings in Bioinformatics, 15(3), 369–375. https://doi.org/10.1093/bib/bbt072

Almeida, J. S., Carrico, J. A., Maretzek, A., Noble, P. A., & Fletcher, M. (2001). Analysis of genomic sequences by Chaos Game Representation. Bioinformatics, 17, 429–437. https://doi.org/10.1093/bioinformatics/17.5.429

Anava, S., Neuhof, M., Gingold, H., Sagy, O. R., Munters, A., Svensson, E. M., Afsinnekoo, E., Danko, D., Fookx, J., Shor, P., Riestra, B., Huchon, D., Mason, C. E., Mizrahi, N., Jakobsson, M., & Rechavi, O. (2020). Illuminating Genetic Mysteries of the Dead Sea Scrolls. Cell, 181(6), 1218–1231. https://doi.org/10.1016/j.cell.2020.04.046

Barone, L., Williams, J., & Micklos, D. (2017). Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. PLoS Computational Biology, 13(10), e1005755. https://doi.org/10.1371/journal.pcbi.1005755

Berlin, K., Koren, S., Chin, C.-S., Drake, J. P., Landolin, J. M., & Phillippy, A. M. (2015). Corrigendum: Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nature Biotechnology, 33(10), 1109.

620
