k-mer approaches for biodiversity genomics

Katharine M. Jenike, Lucía Campos-Domínguez, Marilou Boddé, et al.

Genome Res. 2025 35: 219-230 originally published online January 31, 2025
Access the most recent version at doi:10.1101/gr.279452.124

Review

k-mer approaches for biodiversity genomics

Katharine M. Jenike,[1] Lucía Campos-Domínguez,[2] Marilou Boddé,[3] José Cerca,[4,6] Christina N. Hodson,[5] Michael C. Schatz,[1] and Kamil S. Jaron[3]

[1]*Johns Hopkins University, School of Medicine, Baltimore, Maryland 21205, USA;* [2]*Centre for Research in Agricultural Genomics, CRAG (CSIC-IRTA-UAB-UB), Campus UAB, Cerdanyola del Vallès, 08193 Barcelona, Spain;* [3]*Tree of Life, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom;* [4]*Center for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, 0313 Oslo, Norway;* [5]*University College London, UCL Department of Genetics, Evolution & Environment, London, WC1E 6BT, United Kingdom*

The wide array of currently available genomes displays a wonderful diversity in size, composition, and structure and is quickly expanding thanks to several global biodiversity genomics initiatives. However, sequencing of genomes, even with the latest technologies, can still be challenging for both technical (e.g., small physical size, contaminated samples, or access to appropriate sequencing platforms) and biological reasons (e.g., germline-restricted DNA, variable ploidy levels, sex chromosomes, or very large genomes). In recent years, *k*-mer-based techniques have become popular to overcome some of these challenges. They are based on the simple process of dividing the analyzed sequences (e.g., raw reads or genomes) into a set of subsequences of length *k*, called *k*-mers, and then analyzing the frequency or sequences of those *k*-mers. Analyses based on *k*-mers allow for a rapid and intuitive assessment of complex sequencing data sets. Here, we provide a comprehensive review to the theoretical properties and practical applications of *k*-mers in biodiversity genomics with a special focus on genome modeling.

[Supplemental material is available for this article.]

The genomics field has come a long way in the past quarter century. Sequencing and assembling even a partial genome was once a multibillion dollar accomplishment. Now, complete telomere-to-telomere (T2T) assemblies are being produced for an ever-increasing array of species (Nurk et al. 2022; Chen et al. 2023; Zhang et al. 2023), and large global efforts to capture the genomic diversity of life are well underway (Lewin et al. 2022). However, as of this writing, the vast majority of known species do not have an assembly, or even associated sequencing data. In fact, only 1% of known eukaryotes have an associated assembly according to the Genomes on a Tree tracker (Challis et al. 2023). Genome assembly is a complex process that needs to be guided by orthogonal assembly-free approaches. These methods are frequently based on a popular bioinformatic concept of "*k*-mers." *k*-mers have proven to be an efficient and powerful concept for understanding the vast and continually growing sequencing data.

Every genomicist has encountered *k*-mers in one way or another. Although the names have changed substantially over the years (see Box 1), the concept remains the same. *k*-mers are frequently used for a representation of genomic sequences (Marchet et al. 2021), for example, in the back-end of alignment tools such as BLAST (Altschul et al. 1990), Bowtie 2 (Langmead and Salzberg 2012), and minimap2 (Li 2018) and practically all genome assemblers (for a review, see Li and Durbin 2024). More recently, *k*-mers have emerged as a popular framework for analyzing genomic data sets directly in methods such as genome profiling, thanks to the development of user-friendly tools such as GenomeScope (Vurture et al. 2017; Ranallo-Benavidez et al. 2020). Consequently, *k*-mers are now a fundamental part of genomics. To make the concept of *k*-mers more accessible to users and developers of *k*-mer-based software methods, we offer this guide. We define and explain the basic properties of *k*-mers, review the established *k*-mer-based techniques in genomics, and showcase recent creative uses of *k*-mers for complex genomic problems. Together with this paper, we provide online materials for the reader to exercise the knowledge on real biological examples available at GitHub (https://github.com/KamilSJaron/k-mer-approaches-for-biodiversity-genomics).

k-mer basics

In a genomic context, *k*-mers are substrings of nucleotides of length *k* contained within a sequence (e.g., individual reads, reference genome, or any other sequence). *k*-mers are typically used for DNA, but the concept can be applied to RNA and protein sequences as well. Any genomic sequence can be decomposed into a number of consecutive *k*-mers, and this number will depend on both the length of the sequence (L) and *k*-mer length (*k*). For example, in the following sequence: AAGTCCAT (L = 8), there are seven *k*-mers of length 2 (2-mers), six 3-mers, five 4-mers, four 5-mers, three 6-mers, and two 7-mers (Supplemental Fig. S1). The number of *k*-mers in a sequence of length L is equal to $L - k + 1$. This is a general principle that can be applied to any sequence, regardless of the sequence length or composition.

What is the point of further fragmentation of sequencing reads or genomes? At the most fundamental level, they provide a convenient way to break apart a genomic sequence into simpler words. Unlike natural language, which has the benefit of spaces and other punctuation marks, genomes do not intrinsically mark

[6]Present address: Department of Bioinformatics and Genetics, Swedish Museum of Natural History, 114 18 Stockholm, Sweden
Corresponding author: kamil.jaron@sanger.ac.uk
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.279452.124. Freely available online through the *Genome Research* Open Access option.

Box 1. Historical perspective on *k*-mers

The oldest reference to the concept dates back to the legendary work of Claude Shannon (1948), in which he used "**N-grams**" to develop a theory for communication, later to calculate entropy of a natural language. In the mathematically oriented research community, the concept is most often referred to as "***k*-tuples**" (Drmanac et al. 1991; Idury and Waterman 1995), including shorter versions such as "**ktup**" (Lipman and Pearson 1985), or as other prefixes such as "**L-tuple**" (Idury and Waterman 1995) or "**ℓ-tuple**" (Li and Waterman 2003). Perhaps to reach a wider audience, some authors decided to use "***k*-word**" (Lippert et al. 2002; Li and Waterman 2003) or just "**word**" (Reinert et al. 2000). Instead, it was "***k*-mer**" that became the most popular and common expression. Sequencing by hybridization used "**11-mers**" for the oligo templates (Drmanac et al. 1989), although they never used the generalized form with *k*. Although the concept of *k*-mers also appeared in the publication of BLAST as "**w-mers**" (Drmanac et al. 1989), it took nearly a decade before *k*-mer became more commonly used. The use of "*k*-mer" became more common in late 1990s, including within the pioneering work of the whole-genome aligner MUMmer in 1999 (Delcher et al. 1999). In 2000, Liu and Singh (2000) coined "*k*-mer word frequency distribution" and described it as a "signature" of the sequence (Delcher et al. 1999; Liu and Singh 2000). A few years later, Mullikin and Ning (2003) published a "word frequency graph," which is the first record of a *k*-mer spectrum. Publications using the word *k*-mer increased in the following years compared with any other of the terms. The expression "*k*-mer" was solidified as the main way to describe this concept throughout the 2000s with the release of several genome assemblers, read aligners, and specialized software for counting *k*-mers such as Jellyfish (Mullikin and Ning 2003; Marçais and Kingsford 2011). For a few more details, see Supplemental Text S1.

the start or end of words. Nevertheless, *k*-mers impose such a structure with surprisingly powerful outcomes. The fraction of *k*-mers perfectly matching the sequenced template will necessarily be greater than the fraction of reads that are a perfect representation in their whole entirety. For example, one sequencing error in a read means that the read is not a perfect match to the sequenced template, but only a subset of its *k*-mers will inherit the error. A read of length L contains $L-k+1$ *k*-mers, but only *k* of them will contain the incorrect sequence (e.g., for 100 bp reads analyzed with 21-mers, a sequencing error will disrupt 21 of the 80 *k*-mers in the read, except if the error is near the end of the read). The remaining *k*-mers will represent true genomic sequence.

The second benefit is purely computational: Usually we analyze the *k*-mers that are the faithful representation of the sequenced template; therefore, we can use exact matches to count *k*-mers, which is much faster than any imperfect matching algorithm (e.g., using a hash table, or binary search instead of computing a complete alignment) and, for many applications, does not require a reference genome. Most notably, *k*-mers enable de novo genome assembly via de Bruijn graphs used in assemblers such as EULER (Pevzner et al. 2001) or SPAdes (Bankevich et al. 2012). De Bruijn graphs are constructed of *k*-mers found in the reads, followed by a series of sophisticated graph transformations, and are used in such assemblers as Verkko (Rautiainen et al. 2023) or La Jolla Assembler (LJA) (Bankevich et al. 2022), which both use a multiplex de Bruijn graph for assembling long reads (Compeau et al. 2011).

Finally, a third benefit is that *k*-mers can be used to rapidly assess various biological features. For example, they can be used to estimate genomic properties such as genome size, genome repetitiveness, and heterozygosity (Chikhi and Medvedev 2014; Vurture et al. 2017). They can also be used to estimate similarity between genomes without alignment using Simka (Benoit et al. 2016) or even subsets of representative *k*-mers in Mash (Ondov et al. 2016). Before expanding further, however, we need to understand the properties of *k*-mers and how the choice of *k* affects the data set. The direct applications of *k*-mers are the main focus of this paper with emphasis on *k*-mer spectrum and genome profiling.

Essential properties of *k*-mers

The choice of *k* affects two essential properties of *k*-mers: the number of possible distinct *k*-mers in the set and the *k*-mer coverage. With a 4 bp alphabet, the number of possible distinct *k*-mers is 4^k. However, in practice, we usually do not know the strand of the sequencing read, and as a result, the reverse complement sequences are typically counted as the same sequence; for example, CAT and ATG would be counted together. Typically, we select the lexicographically smaller of the reverse complementary *k*-mers (e.g., ATG would represent both ATG and CAT) as the sequence for the pair; the representative *k*-mers are then called "canonical *k*-mers." Consequently, for an odd value of *k*, there are $4^k/2$ possible canonical *k*-mers (for more details, see Wittler 2023). Typically, odd values of *k* are used, because forward and reverse *k*-mers will be unique. In the remainder of this paper, unless stated otherwise, *k*-mers refer to canonical *k*-mers.

The size of the *k*-mer space (i.e., the number of all possible *k*-mers) increases exponentially with *k* (Supplemental Fig. S2). For $k=3$, there are only 32 possible *k*-mers given a 4 bp alphabet; for $k=7$, there are 8192. For these low *k* values, all possible *k*-mers will appear in a genome many times, and by chance, all these *k*-mers will likely be observed in any species. The relative frequencies of these short *k*-mers ($k \leq 11$), sometimes called genomic signature, carry a phylogenetic signal (Karlin and Burge 1995; Fofanov et al. 2004; for a review, see de la Fuente et al. 2023). Short *k*-mer frequencies are used for alignment-free techniques for species assignment as well as detection of horizontal gene transfer candidates (for review, see Ren et al. 2018).

For most applications we need to select a *k* long enough so that most *k*-mers in the genome will be found only once. The proportion of *k*-mers corresponding to a unique position in the genome increases with a greater *k* (see Fig. 1 of Kurtz et al. 2008, or Schatz et al. 2010). For humans, the smallest *k* with some *k*-mers unique to a specific genomic location is 11, but the vast majority of the genome will be represented by repetitive 11-mers given that the human genome size is much greater than the total number of *k*-mers in the 11-mer space (Supplemental Fig. S3; Aganezov et al. 2022). We estimate the minimum length of *k* by computing the expected number of occurrences of a given *k*-mer in a monoploid genome of length G. This is estimated as $G/4^k$, meaning we need to select *k* to be at least $\log_4(G)$. For the 3 Gbp human genome, this would suggest a minimum length of 16; practically, however, the human genome has many 16 bp repeats so it is still advantageous to select even longer *k*-mers (Whiteford et al. 2005).

With *k* large enough to contain a substantial amount of *k*-mers that are unique to a single position in a genome, we can define the expected *k*-mer coverage (C_k) as the average number of reads that contain a *k*-mer found in a single copy in the genome, also known as monoploid or 1n *k*-mer coverage, in which n denotes the expected copy number (or ploidy) of the *k*-mer in the

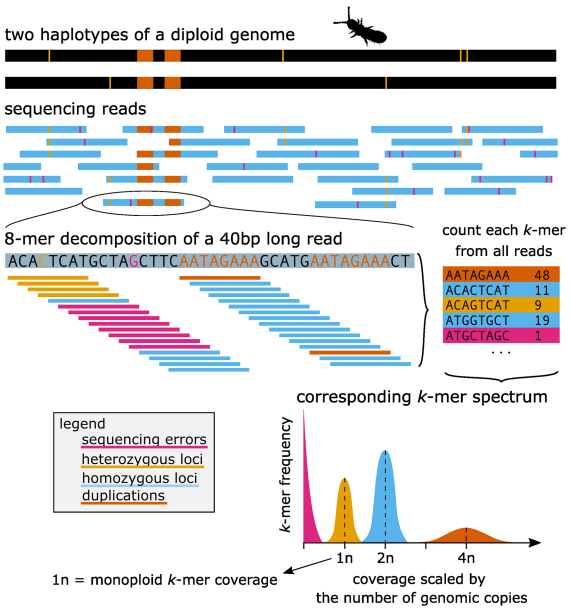

Figure 1. From sequencing reads to *k*-mer spectrum. This figure shows a basic illustrative example of how genomic reads can be translated into a number of *k*-mers that can be counted and represented by a *k*-mer spectrum: The example diploid genome (two haplotypes) has one duplication (orange) and six heterozygous loci (yellow). The sequencing reads contain the corresponding genomic sequence but also sequencing errors (pink). Example of a 40-base long read decomposed to *k*-mers (8-mer). Finally, *k*-mer spectra represent how many different *k*-mers (*y*-axis) show a specific coverage (*x*-axis) in the whole sequencing data set. In the vast majority of real genomic data sets, the peaks in the *k*-mer spectra would overlap, and furthermore, there would be a small number of *k*-mers representing other ploidies too (e.g., 3n for heterozygous duplications, or higher ploidies for more repetitive regions).

genome. Note that the expected *k*-mer coverage (C_k), unlike what per-base genome coverage, is sensitive to choice of *k*, read length, and sequencing error (see Supplemental Text S2).

For a given read set, lower values of *k* give a higher *k*-mer coverage, and the differences in *k*-mer coverage for different values of *k* will be more apparent for short reads (Supplemental Fig. S4). For example, using $k = 51$ for 100 bp reads generates 50 *k*-mers per read, whereas using $k = 100$ is problematic as each read is represented by a single *k*-mer only; therefore, the *k*-mer coverage in the later *k*-mer set would be substantially lower.

Importantly, each wrongly sequenced base owing to a sequencing error will propagate from reads to *k* *k*-mers, which will not accurately represent the sequenced template (except when the error is near the end of the read or there are multiple errors within *k* bp of each other). Therefore, the fraction of affected *k*-mers is again dependent on the chosen value of *k*. Assuming uniformly distributed errors (i.e., each sequenced position within each read has the same probability of being wrong), the probability of a *k*-mer representing the real genomic sequence is then

$$\Pr(\text{error free } k\text{-mer}) = (1 - e)^k,$$

where *e* is the sequencing error rate per base. The lower *k* we chose, the greater the proportion of *k*-mers that will represent the true genome sequence (Supplemental Fig. S4) and therefore be useful for downstream analyses.

Altogether, the choice of *k* is a trade-off between proportion of the genome captured by unique *k*-mers, sequencing coverage, and error rates. Therefore, the "right" value of *k* depends on the application, sequencing depth, sequencing platform, and type of genome sequenced. For example, for short-read assembly, the choice of *k* might optimize the number of recognizable genomic *k*-mers (Chikhi and Medvedev 2014), but for high-fidelity long-read sequencing data, much higher values of *k* can be used, because the *k*-mer coverage will not decrease dramatically for higher values of *k* (see the *k*-mer coverage equation above), allowing for $k > 1000$ for certain applications. This being said, the sequencing error rates still present a limitation on the practical choice of *k*. For short-read data sets, a *k* in the range of 21 to 31 bases is typically chosen as it generates large numbers of *k*-mers from unique positions of a sequenced genome for almost any organism, while still providing high *k*-mer coverages. Furthermore, using $k \leq 32$ is more computationally efficient than longer values of *k*, because these sequences can be represented in 64 bits, making it fast for computers to compare them (i.e., instead of comparing them character by character, they can be compared in a single computer instruction by treating the *k*-mers as 64-bit integers). Finally, for some analyses, it is also beneficial to select odd values for *k*, so that forward and reverse *k*-mers will have distinct sequences; for example, the reverse complement of AT is also AT, but the reverse complement of ACT will be AGT. The largest odd, but computationally efficient, *k* is 31, which makes it a popular choice for a wide range of applications.

Analysis of sequencing libraries using *k*-mer spectra

One of the most common direct applications of *k*-mers in genomics is for characterization of a sequencing data set using the distribution of *k*-mer coverages; this is commonly referred to as *k*-mer spectrum or *k*-mer histogram. A typical *k*-mer spectrum of a moderately heterozygous diploid organism features four apparent coverage peaks (Fig. 1): The first peak represents sequencing errors (those with low coverage; in pink on the figure); the second peak represents unique genomic sequences from heterozygous loci (1n; yellow). This peak will be centered around C_k coverage, sometimes referred as "1n" or "monoploid" *k*-mer coverage. The third peak represents all homozygous loci in the genome centered around $2 * C_k$ coverage (2n; blue), and the fourth, usually a much smaller peak, represents genomic duplications centered around $4 \times C_k$ coverage (4n; orange). In reality, there are frequently other coverage peaks present (e.g., 3n peak of genome duplications in heterozygous state, or sequencing of any off-target genomes such as bacterial contamination), and moreover, the peaks might overlap to the extent that a single coverage threshold might be challenging to clearly separate genomic and error *k*-mers. For these cases such separation is essential, like error correction of sequencing reads, de Brujin graph can be used to recover the true set of genomic *k*-mers (Limasset et al. 2020). However, partially blended peaks allow accurate estimation of genomic properties as well as intuitive interpretation of *k*-mer spectra.

The *k*-mer spectra analysis has become a standard step of genome sequencing (Howe et al. 2021), and the need for efficient tools was recognized by the bioinformatics community. As a result, there has been a significant improvement in performance of *k*-mer counting tools with many additional functionalities developed in *k*-mer toolkits (for an overview of popular *k*-mer counters, see Supplemental Table S1; Marçais and Kingsford 2011; Crusoe et al. 2015; Kokot et al. 2017; Mapleson et al. 2017; Mohamadi et al. 2017; Rhie et al. 2020).

Calculating the *k*-mer spectrum is just the first step. Various models can be fitted to the *k*-mer spectrum to estimate genomic

features such as genome size, heterozygosity, or repetitiveness, collectively called genome modeling or genome profiling.

Fitting a genome profiling model

One of the first genome profiling techniques was designed to estimate *k*-mer coverage and genome size (Li and Waterman 2003). The idea behind genome size estimation is practically the same as calculating *k*-mer coverage. That is because monoploid genome size (G) is the total number of genomic *k*-mers divided by the *k*-mer coverage (C_k) and the ploidy level (p).

$$G = \frac{N\,(L-k+1)}{C_k\,p},$$

where N is the total number of reads in the data set and L is the read length, which together with *k*-mer size k give us the total number of *k*-mers in the data set. The total number of *k*-mers $N\,(L-k+1)$ can be rapidly calculated from a *k*-mer spectrum by $\sum_c cf_c$, where f_c are frequencies of *k*-mers with coverage c. Notably, the equation is robust to any type of genome repetitiveness as long as the coverage C_k is estimated accurately. However, this strategy for genome size estimate includes also the proportion of all the *k*-mers that represent sequencing errors, therefore artificially inflating the estimated genome size. Tools that include an error model subtract the estimated error *k*-mers, which allows for a more precise genome size estimate.

The estimation of the *k*-mer coverage (C_k) is typically done by fitting one, or multiple, distributions to the empirically calculated *k*-mer spectra. The simplest distribution, and a natural choice for a sampling process such as genome sequencing, is the Poisson distribution (Li and Waterman 2003; Liu et al. 2013; Sarmashghi et al. 2021). A Poisson distribution is easy to fit because it is characterized by a single parameter, determining both the mean and the variance. However, in practice the observed variance is usually greater than the variance fitted by the Poisson distribution (i.e., the observed distribution is "overdispersed") so that it might be more appropriate to fit a negative binomial distribution (Vurture et al. 2017; Becher et al. 2020; Ranallo-Benavidez et al. 2020). The latter is a generalization of the Poisson distribution and is characterized by two parameters, a mean parameter and a separate variance parameter, but has a similar shape. A normal or Gaussian distribution can be used for high-coverage values but is not appropriate for low coverage (below 10× coverage per haplotype). This is because a Gaussian distribution is symmetric around the mean, so that at low coverage it will estimate coverage values to be less than zero (negative values) that are not sensical. An alternative distribution, which particularly addresses the fact that some sites have a lower probability of being sequenced than others (e.g., heterochromatin), is the skew normal distribution such as used by findGSE (Sun et al. 2018). No distribution fits all purposes, but in general, the more complicated coverage models are better suited for high-coverage data sets, although in the case of low-coverage data, it is preferable to use the least parameterized distribution (Sarmashghi et al. 2021).

The majority of model-fitting approaches expect relatively high sequencing coverage (>10×) so that they can fit the parameters of the distributions with very few assumptions. Authors of RESPECT took a different route, designed for low-coverage (skimming) data, as well as specifically optimized for data sets in the range 0.5×–2× sequencing coverage (Sarmashghi et al. 2021). The model estimates the number of *k*-mers that are found in one, two, three, etc., counts in the genome, whereas using simple error and coverage models (Sarmashghi et al. 2021). To make the model fit possible, the authors constrained some of the parameters; for example, errors are modeled as a proportion of *k*-mers observed only once in reads, but also the individual counts of *k*-mers in the genomes are constrained by empirically observed ratios in several hundred publicly available genomes (Sarmashghi et al. 2021). These empirical constraints were generated using haploid representation of diploid genomes; therefore, we would expect reliable estimates only for haploid species or diploid genomes with low heterozygosity.

Interestingly, the *k*-mer spectrum can also reveal other properties of the genome. The genome-wide level of heterozygosity can be determined through an analysis of the number of *k*-mers in the 1n versus 2n peaks: At low rates of heterozygosity, the 1n peak will be relatively small because most *k*-mers will be homozygous, but the 1n peak grows taller with higher numbers of heterozygous *k*-mers. Only a relatively modest rate of heterozygosity is needed to elevate the 1n peak to match the 2n peak because each heterozygous variant will cause $2 \times k$ heterozygous *k*-mers (assuming the variants are spaced out appropriately). In this scenario, with $k = 21$, a heterozygosity rate of 1.19% is sufficient for the 1n peak to be as tall as the 2n peak.

Three genome profiling approaches also include estimates of genome-wide heterozygosity. First, in the genomes with low overall heterozygosity (<<0.5%), heterozygous sites will generally be more than *k* nucleotides away from any other heterozygous site. Consequently, each heterozygous site will generate 2^*k heterozygous *k*-mers, which has been proposed as a straightforward estimate of heterozygosity (Liu et al. 2013). However, many species, including outbreeding species and hybrids, display substantially higher levels of heterozygosity (Romiguier et al. 2014; Mackintosh et al. 2019). The problem of linked variants (i.e., multiple heterozygous variants less than *k* bp apart) is addressed in GenomeScope. With the assumption that heterozygous loci and duplications are independent and uniformly distributed across the genome, the fractions of heterozygous and homozygous *k*-mers are calculated as multiplication of per-nucleotide heterozygosity estimate (for a detailed explanation with illustrations, see the supplemental materials of Vurture et al. 2017). The third approach, Tetmer, estimates the genetic diversity of a population and is based on an infinite site model and coalescent theory (Lohse et al. 2011, 2016; Becher et al. 2020). For a more detailed comparison of the three methods, see Supplemental Text S3.

The most popular tool, measured by citations to date, and the tool we will use for all the examples below, is GenomeScope 2.0, which also includes support for polyploid genomes and more advanced model-fitting methods (Ranallo-Benavidez et al. 2020). We will use it to demonstrate various uses and problems with fitting genome models, but we encourage readers to also consider other genome-fitting tools, as those might be more suitable for their specific problems (for an overview, see Supplemental Table S2).

Common signatures of *k*-mer spectra

To generate a high-quality reference of a diploid genome, it is recommended to sequence at least 25×–30× coverage of long reads or, more generally, 15× per haplotype (Rhie et al. 2021; Darwin Tree of Life Project Consortium 2022). Even a simple visual inspection of *k*-mer spectra is valuable to quickly assess if this coverage is achieved. Such coverage should generate a *k*-mer spectrum that shows distinct coverage peaks as demonstrated by the European mistletoe *Viscum album* (Fig. 2A). Sequencing data without

sufficient coverage will have poorly defined peaks, because the homozygous and heterozygous genomic peaks will be blended at the left side of the coverage plot. If the peaks are still visible, it might be possible to fit a meaningful genome model, like in the case of the crayfish *Procambarus virginalis* (Fig. 2B; see data from Gutekunst et al. 2018).

Genome profiling techniques have complicated underlying models with no known analytical solutions. They use a heuristic instead, for example, GenomeScope uses nonlinear least squares, which starts from an initial naive estimate and performs an iterative procedure to update the values until convergence. To evaluate whether a model converged well, we frequently use already known information about the species we sequence, in particular ploidy or genome size that was previously assessed via cytogenetic techniques. Confronting prior knowledge with the estimates derived from the *k*-mer spectra is often helpful in identifying potential problems in the data. In the case of the crayfish, there is a nearly perfect match of genome size estimate from the *k*-mer spectra and flow cytometry (Gutekunst et al. 2018), supporting that the model converged well. Very low sequencing coverage or elevated rates of errors lead to blending of peaks; genomic *k*-mers become indistinguishable from error *k*-mers. This is visible in the chive *Allium schoenoprasum* (Fig. 2C), in which the model (black line) does not fit the data (blue histogram) well. In such cases, the estimated values are artifacts of a poor convergence. The predicted genome size is much lower than what we would expect in the *Allium* genus, in which other species have genomes ranging from 8.4–13.4 Gbp (Henniges et al. 2023), and the coverage is much higher than what we would expect from a spectrum of this shape. Coverage problems are usually resolvable with additional sequencing, whereas high error rates may require a different sequencing technology and/or library preparation.

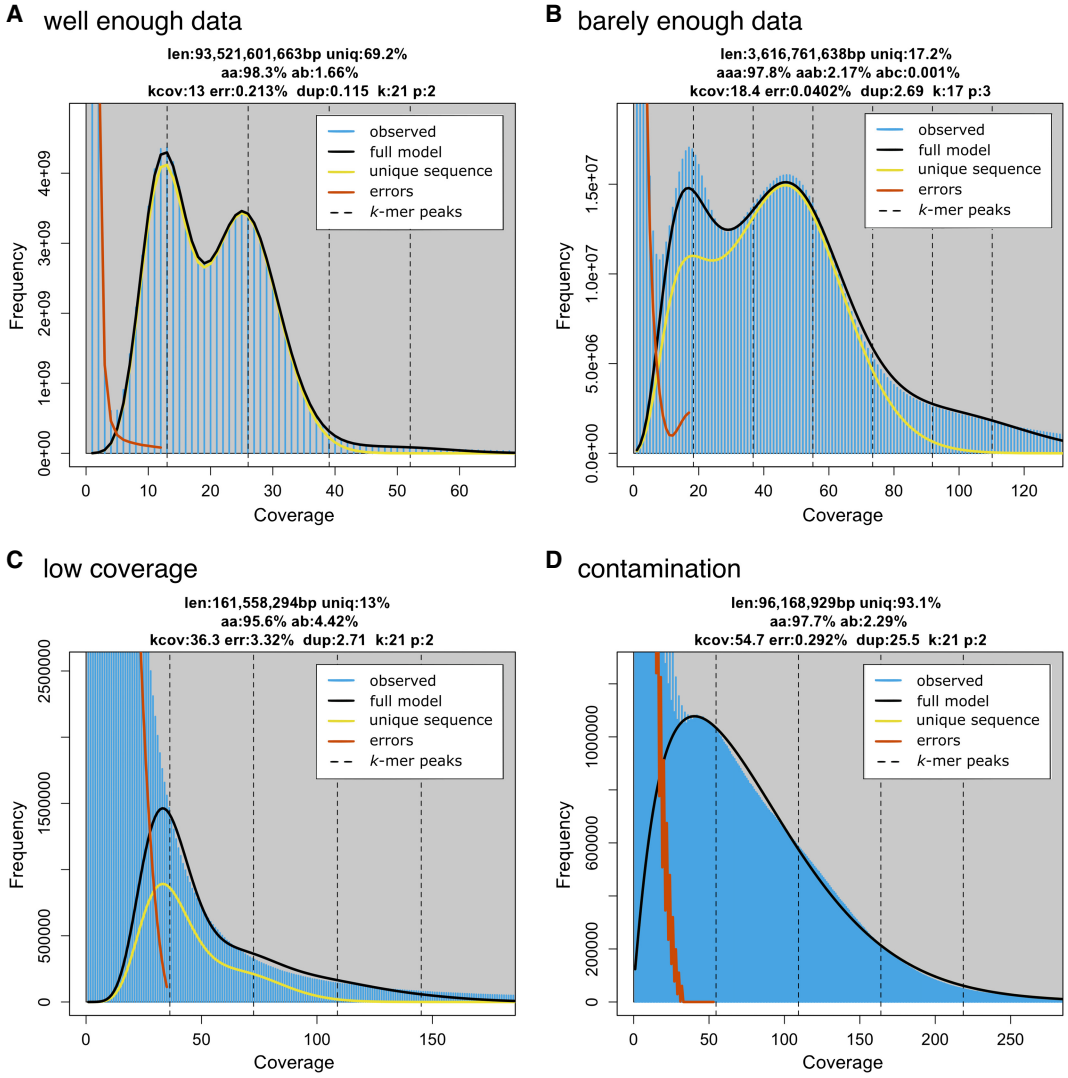

Figure 2. Examples of *k*-mer spectra. (*A*) *Viscum album:* a diploid spectra with enough data to observe two distinct peaks and fit a model that accurately reflects genomic features despite the large size of the genome. (*B*) *Procambarus virginalis: k*-mer spectra of a sample with low coverage, barely sufficient for a model fit. Notably, we used $k = 17$ to increase the *k*-mer coverage and make the model fit possible. (*C*) *Allium schoenoprasum*: The sequencing coverage of this data set is approximately 1×. Error *k*-mers and genome *k*-mers are completely blended; as a consequence, the model did not converge to meaningful estimates. (*D*) *Hypsibius dujardini*: a heavily contaminated sample of a tardigrade.

Contamination is another common cause of blended peaks in a *k*-mer spectrum, especially for small species that are difficult to culture in sterile environments or impossible to culture at all (Cornet and Baurain 2022). In these cases, additional species other than the target are sequenced along with the target sample, resulting in a contaminated genome profile. For example, the true genomic peak in the tardigrade *Hypsibius dujardini* is overlaid with several other cobiont genomes (Fig. 2D; see data from Koutsovoulos et al. 2016). If a contamination consists of a very few species only and the coverage is very high, we might be able to see individual peaks corresponding to the individual sources. Those are, however, usually unevenly spaced (Supplemental Fig. S5). Similar to low-coverage data sets, the values reported by the genome model are most likely inaccurate.

Different sequencing data sets show different variation in sequencing depth, but the general rule is that higher coverage leads to better separation of peaks. Sometimes, we observe blending of peaks, although the coverage is relatively high. This can be caused by various biological or technical reasons. For example, many bony fish sequencing runs show a *k*-mer spectra with a "bridge" blending the error *k*-mers and the 1n peak (Supplemental Fig. S6). This pattern has been attributed to a high proportion of tandem repeats in many fish lineages (Mackintosh et al. 2019; Reinar et al. 2023) and a coverage dropout of low-complexity A/G rich regions of Pacific Biosciences (PacBio) HiFi sequencing (Nurk et al. 2020). Another example is the well-known sequencing depth bias of Illumina short-read sequencing suffers in regard to GC content (Dohm et al. 2008; Nurk et al. 2020). This bias varies between chemistries (Stoler and Nekrutenko 2021) and is reduced with PCR-free strategies (Aird et al. 2011).

The biases above are well described and replicable when using similar samples and sequencing technologies. However, sequencing runs are also affected by sample handling, used preservatives, or occasional manufacturing problems of sequencing flow cells. Even these problems are sometimes visible in *k*-mer spectra. For example, the *k*-mer-based genome size estimate of a cape honeybee sample (142 Mbp) (Supplemental Fig. S7) is much smaller than the published genome size (236 Mbp). This was a case only for one of three samples; the other two samples from the same project (and same population) showed expected genome size. The spectra of the peculiar sample also show much higher blending of peaks, all together indicating this coverage dropout is not driven by biology of the sequenced specimen but rather technical difficulties.

In general, the lack of clearly defined and evenly spaced peaks indicates that there is no single source with sufficient coverage to generate the expected spectra. By keeping in mind the organismal biology of the target species, many of the patterns and potential problems can be anticipated.

Common pitfalls when fitting models

The quality of fit for genome models is largely dependent on the quality and coverage of the data but also on the biological features of the genome. The most common problem of genome models is for the monoploid (1n) *k*-mer coverage to converge on a wrong value. This can happen if the 1n coverage peak is not sufficiently distinct or is significantly higher than the diploid peak. This can be caused by extremely low heterozygosity of the genome (i.e., the 1n signal is very weak) (Supplemental Fig. S8A) or by very low coverage and the 1n peak largely overlapping with the error peak (Fig. 2C).

Regardless of the genome profiling method used, when the 1n coverage is not fitted correctly, none of the estimated values will carry any biological information regarding the genome. Therefore, it is important to inspect fits and make sure the estimates agree with the context of the other known biology. For example, if we sequence a species with low expected heterozygosity (such as diploid selfing plant) and the estimated heterozygosity is >5%, it is extremely likely that the true 1n coverage is about one-half of the estimated one (Supplemental Fig. S8B). This convergence issue affects tools that expect certain levels of heterozygosity, such as GenomeScope. This is not the case for RESPECT, which assumes low or no heterozygosity. Genome profiling models are extremely useful tools, but they require knowledge of assumptions of the fitting algorithm for correct interpretation of the estimates.

Another common issue is that many *k*-mer counting tools stop counting *k*-mer coverages at a user-specified value, with a typical default of 10,000. The last value in these *k*-mer histograms typically represents how many *k*-mers have coverage greater than or equal to the highest counted coverage. This will lose resolution of the frequency of the most common repeats in the genome. For many genomes, it makes little difference; however, some extremely repetitive genomes will have a severely underestimated genome size. When using the default value, the estimated genome size of marbled crayfish is nearly half of the real genome size (Supplemental Fig. S9). When one observes unexpectedly low genome size, it is frequently caused by the truncated *k*-mer spectra. Notably, FastK does not explicitly calculate coverage of every high-coverage *k*-mer; instead, it reports as the last value of the theoretical number of *k*-mers that would span the same total sum of coverage yield as all higher coverage *k*-mers in the genome, ensuring a more accurate genome size estimate (Compeau et al. 2011; https://github.com/thegenemyers/FASTK).

Joint interpretation of the *k*-mer spectra and the assembled genome

Postassembly, estimated genome qualities from *k*-mer models can be used to assess assembly accuracy. One simple yet informative metric is to compare the estimated genome size to the assembly size. This simple comparison is quite powerful because it can quickly indicate misassembly, especially partial duplications caused by heterozygosity. *Ilex aquifolium* (common holly), sequenced by The Darwin Tree of Life Project (Christenhusz et al. 2024) can serve as an example of mismatch between estimated genome size and assembly size. The estimated genome size of *I. aquifolium* based on *k*-mer histogram from PacBio reads is 815.6 Mbp. However, the primary assembly is marginally larger (830.1 Mbp) (Supplemental Fig. S10A), and furthermore, there are 7.2% duplicated BUSCOs in this genome, altogether indicating that there could be some uncollapsed regions in this assembly with sequences from both haplotypes inflating the gene count and (haploid) genome size. It was indeed the case for this genome, and it was resolved by purging the uncollapsed haplotypes using purge_dups (Guan et al. 2020; Christenhusz et al. 2024).

Several tools exist now to evaluate genome assembly accuracy using preassembly *k*-mers. We will specifically discuss Merqury (Rhie et al. 2020); however, this is not the only tool (Mapleson et al. 2017; Mikheenko et al. 2018; Cheng et al. 2021). Essentially, these tools compare the *k*-mers from sequencing reads to the *k*-mers present in the finished genome assembly to determine how complete the assembly is. This operates under the assumption that the finished assembly should contain most of the *k*-mers

that were present in the sequencing reads, excluding low-coverage *k*-mers that are likely owing to sequencing errors, and half of the heterozygous alleles in diploid genomes if we are working with a diploid collapsed assembly (for Merqury plot example on the common holly genome, see Supplemental Fig. S10B). Additionally, Merqury estimates the consensus quality value score, which assumes all *k*-mers present in the finished assembly should be present in the read set. It then estimates the probability that a particular base is an error, which can be reported as the commonly used Phred score (Ewing et al. 1998). This approach is especially useful for species with no reference or a poorly assembled reference. However, it does require high-accuracy reads (such as Illumina or PacBio HiFi) as a reference point, and ideally, these reference reads would be orthogonal to those used for the assembly.

Comparison of sequencing libraries using *k*-mers

k-mers can also be used to identify genomic differences between two sequencing libraries. This analytical approach can be used to estimate their genetic divergence (VanWallendael and Alvarez 2022), differences of repetitive content (Liu et al. 2017; Becher et al. 2022), to identify sex chromosomes if the libraries generated from different sexes or to identify tissue or individual specific chromosomes such as B Chromosomes (Vea et al. 2021). *k*-mer chromosome identification techniques can be used on their own or with well-known techniques already in use such as methods that identify chromosomes based on coverage differences or differences in sequence composition (e.g., SNP differences) between two sequencing libraries (for review, see Palmer et al. 2019). One advantage of *k*-mer techniques over other methods is a reduced reliance on a high-quality reference genome. Therefore, *k*-mer-based techniques may be a better approach in nonmodel organisms with a fragmented or nonexistent reference genome.

The basic approach in analyses using *k*-mers to identify specific chromosomes is to compare *k*-mer frequency in two sequencing libraries that differ in chromosome constitution. These 2D *k*-mer spectra, comparing *k*-mer frequency in two libraries, can be generated using KAT (Mapleson et al. 2017) and allow for the identification of *k*-mers belonging to the chromosome with a different frequency in the two samples. For instance, if identifying a sex chromosome in a species with an XO sex determination system, *k*-mers belonging to the X Chromosome will be at half the frequency in male compared with female sequencing libraries, whereas autosomal *k*-mers will be at the same frequency (Fig. 3A). The *k*-mers belonging to the X Chromosome can then be isolated and either mapped to an assembly to identify scaffolds belonging to this chromosome or mapped to reads to identify reads belonging to the X Chromosome.

Using *k*-mers to identify specific chromosomes is most effective when the region of interest is highly divergent from the other genomic regions. For instance, this approach is difficult to implement to identify the Y Chromosome in a homomorphic XY sex-chromosome system with low differentiation between the X and Y Chromosomes, as few *k*-mers will be restricted to the Y Chromosome. Therefore, situations such as this might require more permissive identification thresholds. The heterozygosity of the samples is also an important consideration. For instance, if identifying a sex chromosome from a male sample and a female sample collected from a natural population, *k*-mers covering heterozygous SNPs will show signatures of *k*-mers specific to sex chromosomes as these will also be present at half the frequency of homozygous autosomal *k*-mers (Fig. 3B). This issue can be somewhat overcome by sequencing more individuals so that individual-specific heterozygous *k*-mers can be distinguished from sex-specific *k*-mers. Finally, an important consideration is the depth of sequencing for each sample. The sequencing coverage needs to be high enough to distinguish peaks in the *k*-mer spectra belonging to chromosomes at different ploidy levels in the samples sequenced and to distinguish from *k*-mers representing sequencing errors, similar to genome profiling.

Some creative use cases of *k*-mers

Up until this point, we have discussed general properties of *k*-mers and how to apply established *k*-mer methods to sequencing data.

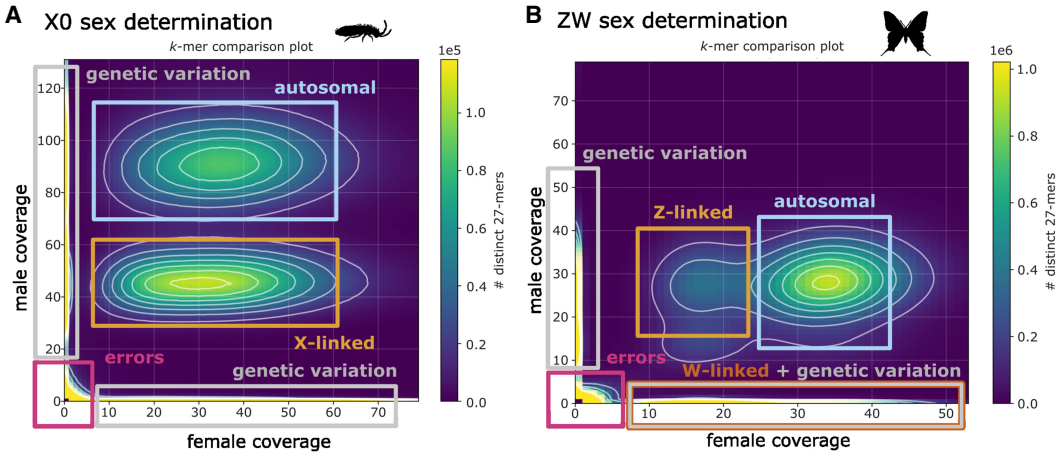

Figure 3. *k*-mer comparison of multiple sequencing libraries. Examples of two 2D *k*-mer spectra for species with the following: (*A*) A XO sex determination system (*Orchesella cincta*; data from Anderson et al. 2022) and (*B*) a ZW (female heterogametic) sex determination system (*Iphiclides podalirius*; data from Ebdon et al. 2024). Both plots show a heatmap of the frequency of *k*-mers in a female (*x*-axis) versus a male (*y*-axis). *k*-mers associated with autosomes or homogametic sex chromosomes (X or Z) can be differentiated in both plots (orange box) as they are at half the frequency in the heterogametic sex compared with autosomal *k*-mers (blue box). Similarly, heterogametic sex-chromosome (W Chromosome) *k*-mers can be identified in panel *B* as those that are absent in males but present in females (brown box). However, it is important to note that the autosomal genetic diversity of the two compared individuals may be mistaken for those belonging to sex chromosomes.

However, there are many problems in biology that are hard to solve with existing tools, which forces us to design novel approaches tailored to the problems and data we have. Here, we showcase the creative use of *k*-mers in three examples of such problems in three different contexts: study of germline DNA, subgenomes of allotetraploid, and species assignment.

Extraction of germline-restricted chromosomes

A specific scenario of comparing libraries was used by Hodson et al. (2022) to identify sequences belonging to the germline-restricted chromosomes in the black-winged fungus gnat *Bradysia coprophila*, in order to explore the evolutionary origins of these unusual chromosomes. In this species, males show a peculiar set of elimination events that generate different karyotypes in different tissues and life stages (for recent review, see Gerbi 2022). Specifically, sperm carries two X Chromosomes, two germline-restricted chromosomes, and one set of autosomes, whereas the soma has a single X Chromosome, no germline-restricted chromosomes (hence the name), and two sets of autosomes. In this case, germline tissue composed mostly of sperm has a different frequency of all chromosome types compared to somatic tissue. To characterize the germline-restricted DNA of this species, approximately 95 testes of unmated males were pooled together in a single library and were sequenced separately from heads of the same individuals to generate a clean somatic-tissue library to compare with the germline.

The comparison of the two libraries using 2D *k*-mer spectra indicated there was indeed an excess of *k*-mers that occurred solely in the testes library belonging to the germline-restricted chromosomes. However, unexpectedly, the coverage of autosomal *k*-mers in the testes was still higher than the coverage of X Chromosome *k*-mers, despite the fact that autosomes are at a lower frequency in sperm (expected 2:1 ratio X Chromosome: autosomal *k*-mers in sperm). This was caused mostly by contamination of the germline library by somatic cells. A rough estimate indicated that ~77% of the germline sequencing library was composed of somatic cells, which caused the X Chromosomes to have lower coverage than autosomes, but not one-half, as one would expect if the sequencing library was just somatic tissue (Supplemental Fig. S11). Therefore, in this case, the fact that the two tissue types had three chromosome types all at different frequencies helped to determine the relative composition of each tissue type in the sequencing libraries.

The chromosome-group specific *k*-mers were matched to contigs, which resulted in near perfect assignment of contigs to chromosomal groups (germline-restricted chromosomes, X Chromosome, autosomes). In this analysis, the quality of the assignment was particularly good for two reasons. First, because the germline-restricted chromosomes happened to be very divergent from homologous regions on the X Chromosome and autosomes (the somatic chromosomes), and second, because this gnat line had extremely low heterozygosity as it was isolated by Charles W. Metz during the 1910s (Gerbi 2022). The *k*-mer-based chromosome sorting allowed for accurate identification of tissue restricted chromosomes from a nonmodel species without a high-quality reference genome (at the time), which facilitated downstream analyses of the origin of the germline-restricted chromosomes (Hodson et al. 2022).

Separating subgenomes of allotetraploids

Polyploid genomes are challenging to assemble owing to their large size, repetitive content that may arise from genomic shock, and the presence of chromosomes with similar composition (homeologs, or chromosomes that originated from speciation and were reunited in the same genome by allopolyploidization) (Kyriakidou et al. 2018; Wang et al. 2023). Despite these challenges, several chromosome-level assemblies of allopolyploids have been published (e.g., Yang et al. 2017; Cerca et al. 2022), in which subgenomes (i.e., the genomes that were reunited as a result of the polyploidization event) are reconstructed. In this context, *k*-mer approaches have been used to separate subgenomes, leading to significant advances in our understanding of genome evolution, including subgenome evolution, such as biases in gene retention, gene function, natural selection, and synteny (Session et al. 2016; Cerca et al. 2022; Session and Rokhsar 2023).

The separation of the two subgenomes is facilitated by the division of a lineages' evolutionary history into three tempos (Supplemental Fig. S12): Tempo 1, the period preceding the speciation event that separated the two ancestral genomes; Tempo 2, the period between the speciation and the polyploidization events, during which ancestral genomes accumulated different transposable elements (TEs) and diverged; and Tempo 3, the period after the polyploidization event, during which both subgenomes coexist in the same nucleus. Now, consider TE accumulation in each of these separate tempos: Tempo 1, TEs accumulating in this tempo are expected to be evenly distributed on both subgenomes; Tempo 2, TEs accumulating after the speciation event and before polyploidization will be distinct on both subgenomes, as the two separate lineages will accumulate their own unique TEs; and Tempo 3, TEs that accumulate after the polyploidization event will be roughly equally distributed between subgenomes (Session et al. 2016; Cerca et al. 2022; Session and Rokhsar 2023).

In this context, *k*-mers serve as a powerful tool in subgenome separation in a chromosome-level assembly. Specifically, because evolutionary signals tend to overwrite each other on the genome, the fragmentation of the sequences offers an effective way to disentangle past processes, which have occurred in the three different tempos (Supplemental Fig. S12; Session et al. 2016; Cerca et al. 2022; Session and Rokhsar 2023). The separation of subgenomes involves two steps: first, the homeologs need to be identified. This can be done using UCEs/COS (ultra conserved elements, conserved ortholog sequence) or alignment-based approaches and synteny. When homeologs are known, *k*-mers allow identifying the accumulation of TEs in the three tempos. Specifically, by performing a *k*-mer spectrum analysis and searching for two signatures, first, we select for *k*-mers that are present in high numbers (i.e., >100×), which should come from repeated areas on the genome, such as TEs; and second, we select differently represented *k*-mers in members of the homeologs (i.e., *k*-mers which are found more than twice as often in one member compared to the other member). The assumption of high numbers targets TEs, whereas the second fishes out differentially represented TEs, thereby effectively focusing on the second tempo, when the ancestral genomes were separated and accumulating their own TEs. By doing a hierarchical clustering based on this *k*-mer separation, the chromosomes group on subgenomes.

Species assignment using short *k*-mers

For specific types of analyses, it can be beneficial to use very short *k*-mers ($k < 10$). For the example discussed here (Boddé et al. 2022), targeted amplicon sequencing was used to analyze haplotypes averaging only 160 bp. The aim of the analysis is to identify the species by comparing the query sequences to a reference panel. As

such, *k*-mers from the reconstructed haplotypes were analyzed instead of the reads directly. Furthermore, because the haplotypes can be oriented by the primers, the analysis uses the full *k*-mer set rather than canonical *k*-mers.

The trade-off in the choice of *k* is between tolerance in sequence variation and captured detail owing to the size of the *k*-mer space. Because the analysis works with reconstructed haplotypes rather than reads, the *k*-mer coverage (C_k) does not play a role in the trade-off. For large *k*, there is little tolerance for variation between the query and the reference, whereas for small *k*, there is a high chance that the same *k*-mer is found in multiple locations in the sequence by chance. For example, in a 149 bp sequence, five evenly spread SNPs result in no 25-mers matching the reference. Conversely, the chance that all 4-mers are unique in a sequence of the same length is incredibly small ($<10^{-22}$). Based on these trade-offs, we selected 8-mers as a reasonable length. With a mean target length of 160 bp, the chance that all 8-mers within a haplotype are unique is 84%.

To perform species assignment, we compute the *k*-mer distance from the query haplotype to each haplotype in the reference panel. The *k*-mer distance quantifies the fraction of matching *k*-mers between query and reference . The nearest neighbor sequence is the reference haplotype that minimizes the *k*-mer distance to the query haplotype. The species label is assigned by identifying the nearest neighbors for all amplicon targets of the query sample and aggregating their contributions to the assignment.

The amplicon panel and the species assignment method were developed to perform species assignment for the entire genus of *Anopheles* mosquitoes. *k*-mers provide an objective way to compare highly diverged sequences, in which multiple sequence alignment or alignment to a single reference genome tends to introduce bias toward better-represented clades in the panel and the reference species, respectively. Moreover, *k*-mers provide a natural way to incorporate small indels in addition to SNPs, which considerably increases the power to distinguish between species when working with <10 kb sequence.

The future of *k*-mers in genomics

As essential as words are to linguistics, *k*-mers are just as essential to bioinformatics. They are one of the most basic ways to represent a biological sequence, yet their utility cannot be overstated, as they are the fundamental data type used for a myriad of applications spanning genomics, transcriptomics, and metagenomics. Building on these successes, several research trends have emerged to make them even more efficient and effective for several important applications.

Within genomics, in addition to profiling increasingly complex biological samples, *k*-mers are now being used to power several genome assembly and analysis applications. One powerful technique has been the rise of "trio binning," in which *k*-mers identified in unassembled reads from parental genomes are used as markers to phase the unassembled reads of their children (Koren et al. 2018). This enables genome-wide phasing and de novo assembly of the individual haplotypes in a sample, which has led to a renaissance in diploid genome assembly with improved contiguity and accuracy over prior approaches (Jarvis et al. 2022), including automating the assembly of complete T2T chromosomes and genomes (Rautiainen et al. 2023; Koren et al. 2024). Another powerful technique has been a focus on "singly unique nucleotide *k*-mers" (or SUNKs) to aid in the assembly of repetitive genomes and repetitive sequences (Sudmant et al. 2010). SUNKs can be identified within unassembled reads and serve as unique sequences to "anchor" reads within a genome assembly with high confidence. They were essential, for example, to resolve segmental duplications in human genomes by identifying localized regions of unique sequence embedded within the complex repeat arrays (Vollger et al. 2019) and were recently used to assemble and validate the centromeres within the human genome (Logsdon et al. 2024). Another clever application has been to use *k*-mers as markers for variation to power genotype-to-phenotype association studies (Voichek and Weigel 2020). This is particularly powerful to tag and analyze structural variations as these are the most difficult class of variation to study from raw reads. Moving forward, we anticipate future advances assembling and analyzing more complex genomes and pangenomes using related *k*-mer-based techniques.

Relatedly, *k*-mers also play a major role in a variety of sequence classification applications. Within metagenomics, the pioneering algorithm kraken (Wood and Salzberg 2014) exploits *k*-mers as signatures of individual species, allowing fast and robust classification of individual reads without alignment. This pushed the development of several newer methods that are further optimized for performance, accuracy, and flexibility (Wood et al. 2019; Shaw and Yu 2024; Ulrich and Renard 2024). Within transcriptomics, the pioneering algorithm sailfish (Patro et al. 2014) demonstrated accurate transcript quantification was possible without alignment by using *k*-mers as markers for individual transcripts. This work led to further developments that are orders of magnitude faster and also even more accurate (Bray et al. 2016; Patro et al. 2017; Srivastava et al. 2020). We expect many future applications for *k*-mers as markers for classification and quantification of diverse samples. We also anticipate future applications in which *k*-mers are embedded into abstract semantic representations for machine learning applications (Ji et al. 2021), analogous to how word2vec (Mikolov et al. 2013), BERT (Devlin et al. 2018), and related approaches (Zaheer et al. 2020) have emerged as cornerstones in natural language processing.

Finally, one of the most important technical developments has been the rise of sampling and sketching techniques, to reduce the computational complexity of *k*-mers (Rowe 2019). The core idea of these approaches is that, for many applications, it is not necessary to exhaustively consider every possible *k*-mer in a sequence. Instead, it is often sufficient to focus on a small representative subset for an analysis, typically representing a few percentages or less of all possible *k*-mers. Because the subset is much smaller than the full list, it is substantially faster to compare the subsets, and it requires much less memory. A MinHash is a type of sketching algorithm that uses a data structure with a vector of hashes. MinHash powers the pioneering mash (Ondov et al. 2016) algorithm to quickly estimate the similarity between pairs of genomes by comparing small lists of representative *k*-mers, enabling thousands of genomes to be compared with each other on a laptop in a few minutes. More recent methods, including Sourmash (Irber et al. 2024), SuperSampler (Rouzé et al. 2023), Dashing (Baker and Langmead 2023), Bindash (Zhao 2019), and Niqki (Agret et al. 2022), have advanced the field with more concise representations, faster processing, and more flexible APIs.

Similar to the MinHash, minimizers can be used to efficiently compare the similarity of two sequences. Minimizers were originally developed to reduce the computing requirements for genome assembly and are now used to accelerate popular aligners like minimap2 (Li 2018) by focusing the algorithm to a smaller list of potential seeds. In this case, the sketch is composed of the minimizers, which are the minimal *k*-mers in a window. The

current research frontier for this work focuses on developing even more advanced schemes for selecting or combining *k*-mers together to ensure the representative subset of *k*-mers is unbiased while being robust to sequencing errors and repetitive elements (for a review, see Das and Schatz 2022). There is also ongoing work to develop highly efficient software libraries and toolkits for the fast and efficient processing of *k*-mers for all these applications, especially to reduce the memory requirements for indexing large sets of *k*-mers (Chikhi and Rizk 2013; Marchet et al. 2021) and indexing *k*-mers across large sets of genomes (Pibiri et al. 2023). Notably, in a recent preprint, Chikhi et al. (2024) were able to index all 50 petabases of the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) using a compressed de Bruijn graph constructed from all of the *k*-mers in all of the reads present.

Overall, *k*-mers continue to be one the central concepts of bioinformatics by continuously opening new avenues for scalable analyses of high-volume genomics data. Whenever a researcher is considering profiling a genome, comparing genomes, or analyzing another alignment, they should ask themselves if it might be accomplished using *k*-mers alone.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This paper drew inspiration from the OH-KNOW workshop; we thank Hugo de Boer and Quentin Mauvisseau from ForBio and other lecturers of the course: Siavash Mirarab, Rishi De-Kayne, Paolo Ribeca, and Hannes Becher. We also thank Giulio Formenti, Konrad Lohse, Shane McCarthy, T. Rhyker Ranallo-Benavidez, Arthur Delcher, and Adam Phillippy for their helpful discussions and feedback on the manuscript. This work is supported, in part, by National Science Foundation awards IOS-1758800, IOS-2216612, and DBI-2419522 (to M.C.S.) and Human Frontiers Scientific Program award RGP0025/2021 (to M.C.S.). K.S.J. was supported by the Wellcome Trust grant number 220540.

References

Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, Taylor DJ, Shafin K, Shumate A, Xiao C, et al. 2022. A complete reference genome improves analysis of human genetic variation. *Science* **376:** eabl3533. doi:10.1126/science.abl3533

Agret C, Cazaux B, Limasset A. 2022. Toward optimal fingerprint indexing for large scale genomics. bioRxiv doi:10.1101/2021.11.04.467355v2

Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12:** R18. doi:10.1186/gb-2011-12-2-r18

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403–410. doi:10.1016/S0022-2836(05)80360-2

Anderson N, Jaron KS, Hodson CN, Couger MB, Ševčík J, Weinstein B, Pirro S, Ross L, Roy SW. 2022. Gene-rich X chromosomes implicate intragenomic conflict in the evolution of bizarre genetic systems. *Proc Natl Acad Sci* **119:** e2122580119. doi:10.1073/pnas.2122580119

Baker DN, Langmead B. 2023. Genomic sketching with multiplicities and locality-sensitive hashing using Dashing 2. *Genome Res* **33:** 1218–1227. doi:10.1101/gr.277655.123

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19:** 455–477. doi:10.1089/cmb.2012.0021

Bankevich A, Bzikadze AV, Kolmogorov M, Antipov D, Pevzner PA. 2022. Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. *Nat Biotechnol* **40:** 1075–1081. doi:10.1038/s41587-022-01220-6

Becher H, Brown MR, Powell G, Metherell C, Riddiford NJ, Twyford AD. 2020. Maintenance of species differences in closely related tetraploid parasitic *Euphrasia* (Orobanchaceae) on an isolated island. *Plant Communications* **1:** 100105. doi:10.1016/j.xplc.2020.100105

Becher H, Sampson J, Twyford AD. 2022. Measuring the invisible: the sequences causal of genome size differences in eyebrights (*Euphrasia*) revealed by *k*-mers. *Front Plant Sci* **13:** 818410. doi:10.3389/fpls.2022.818410

Benoit G, Peterlongo P, Mariadassou M, Drezen E, Schbath S, Lavenier D, Lemaitre C. 2016. Multiple comparative metagenomics using multiset *k*-mer counting. *PeerJ Comput Sci* **2:** e94. doi:10.7717/peerj-cs.94

Boddé M, Makunin A, Ayala D, Bouafou L, Diabaté A, Ekpo UF, Kientega M, Le Goff G, Makanga BK, Ngangue MF, et al. 2022. High-resolution species assignment of mosquitoes using *k*-mer distances on targeted sequences. *eLife* **11:** e78775. doi:10.7554/eLife.78775

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34:** 525–527. doi:10.1038/nbt.3519

Cerca J, Petersen B, Lazaro-Guevara JM, Rivera-Colón A, Birkeland S, Vizueta J, Li S, Li Q, Loureiro J, Kosawang C, et al. 2022. The genomic basis of the plant island syndrome in Darwin's giant daisies. *Nat Commun* **13:** 3729. doi:10.1038/s41467-022-31280-w

Challis R, Kumar S, Sotero-Caio C, Brown M, Blaxter M. 2023. Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic tree of life. *Wellcome Open Res* **8:** 24. doi:10.12688/wellcomeopenres.18658.1

Chen J, Wang Z, Tan K, Huang W, Shi J, Li T, Hu J, Wang K, Wang C, Xin B, et al. 2023. A complete telomere-to-telomere assembly of the maize genome. *Nat Genet* **55:** 1221–1231. doi:10.1038/s41588-023-01419-6

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18:** 170–175. doi:10.1038/s41592-020-01056-5

Chikhi R, Medvedev P. 2014. Informed and automated *k*-mer size selection for genome assembly. *Bioinformatics* **30:** 31–37. doi:10.1093/bioinformatics/btt310

Chikhi R, Rizk G. 2013. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *Algorithms Mol Biol* **8:** 22. doi:10.1186/1748-7188-8-22

Chikhi R, Raffestin B, Korobeynikov A, Edgar R, Babaian A. 2024. Logan: planetary-scale genome assembly surveys life's diversity. bioRxiv doi:10.1101/2024.07.30.605881

Christenhusz MJM, Fay MF, Royal Botanic Gardens Kew Genome Acquisition Lab, Plant Genome Sizing collective, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team, Wellcome Sanger Institute Scientific Operations: Sequencing Operations, Wellcome Sanger Institute Tree of Life Core Informatics team, Tree of Life Core Informatics collective, et al. 2024. The genome sequence of the English holly, *Ilex aquifolium* L. (Aquifoliaceae). *Wellcome Open Res* **9:** 1. doi:10.12688/wellcomeopenres.20748.1

Compeau PEC, Pevzner PA, Tesler G. 2011. How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol* **29:** 987–991. doi:10.1038/nbt.2023

Cornet L, Baurain D. 2022. Contamination detection in genomic data: More is not enough. *Genome Biol* **23:** 60. doi:10.1186/s13059-022-02619-9

Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R, Charbonneau A, Constantinides B, Edvenson G, Fay S, et al. 2015. The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Res* **4:** 900. doi:10.12688/f1000research.6924.1

Darwin Tree of Life Project Consortium. 2022. Sequence locally, think globally: The Darwin Tree of Life Project. *Proc Natl Acad Sci* **119:** e2115642118. doi:10.1073/pnas.2115642118

Das A, Schatz MC. 2022. Sketching and sampling approaches for fast and accurate long read classification. *BMC Bioinformatics* **23:** 452. doi:10.1186/s12859-021-04477-x

de la Fuente R, Díaz-Villanueva W, Arnau V, Moya A. 2023. Genomic signature in evolutionary biology: a review. *Biology (Basel)* **12:** 322. doi:10.3390/biology12020322

Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. 1999. Alignment of whole genomes. *Nucleic Acids Res* **27:** 2369–2376. doi:10.1093/nar/27.11.2369

Devlin J, Chang M-W, Lee K, Toutanova K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [cs.CL]. doi:10.48550/arXiv.1810.04805

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36:** e105. doi:10.1093/nar/gkn425

Drmanac R, Labat I, Brukner I, Crkvenjakov R. 1989. Sequencing of megabase plus DNA by hybridization: theory of the method. *Genomics* **4:** 114–128. doi:10.1016/0888-7543(89)90290-5

Drmanac R, Labat I, Crkvenjakov R. 1991. An algorithm for the DNA sequence generation from k-tuple word contents of the minimal number

of random fragments. *J Biomol Struct Dyn* **8:** 1085–1102. doi:10.1080/07391102.1991.10507867
Ebdon S, Laetsch DR, Vila R, Baird SJE, Lohse K. 2024. Genomic regions of current low hybridisation mark long-term barriers to gene flow in scarce swallowtail butterflies. bioRxiv doi:10.1101/2024.06.03.597101
Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces UsingPhred: I. Accuracy assessment. *Genome Res* **8:** 175–185. doi:10.1101/gr.8.3.175
Fofanov Y, Luo Y, Katili C, Wang J, Belosludtsev Y, Powdrill T, Belapurkar C, Fofanov V, Li T-B, Chumakov S, et al. 2004. How independent are the appearances of *n*-mers in different genomes? *Bioinformatics* **20:** 2421–2428. doi:10.1093/bioinformatics/bth266
Gerbi SA. 2022. Non-random chromosome segregation and chromosome eliminations in the fly *Bradysia* (*Sciara*). *Chromosome Res* **30:** 273–288. doi:10.1007/s10577-022-09701-9
Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. 2020. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36:** 2896–2898. doi:10.1093/bioinformatics/btaa025
Gutekunst J, Andriantsoa R, Falckenhayn C, Hanna K, Stein W, Rasamy J, Lyko F. 2018. Clonal genome evolution and rapid invasive spread of the marbled crayfish. *Nat Ecol Evol* **2:** 567–573. doi:10.1038/s41559-018-0467-9
Henniges MC, Johnston E, Pellicer J, Hidalgo O, Bennett MD, Leitch IJ. 2023. The plant DNA C-values database: a one-stop shop for plant genome size data. *Methods Mol Biol* **2703:** 111–122. doi:10.1007/978-1-0716-3389-2_9
Hodson CN, Jaron KS, Gerbi S, Ross L. 2022. Gene-rich germline-restricted chromosomes in black-winged fungus gnats evolved through hybridization. *PLoS Biol* **20:** e3001559. doi:10.1371/journal.pbio.3001559
Howe K, Chow W, Collins J, Pelan S, Pointon D-L, Sims Y, Torrance J, Tracey A, Wood J. 2021. Significantly improving the quality of genome assemblies through curation. *GigaScience* **10:** giaa153. doi:10.1093/gigascience/giaa153
Idury RM, Waterman MS. 1995. A new algorithm for DNA sequence assembly. *J Comput Biol* **2:** 291–306. doi:10.1089/cmb.1995.2.291
Irber L, Pierce-Ward NT, Abuelanin M, Alexander H, Anant A, Barve K, Baumler C, Botvinnik O, Brooks P, Dsouza D, et al. 2024. Sourmash v4: a multitool to quickly search, compare, and analyze genomic and metagenomic data sets. *J Open Source Softw* **9:** 6830. doi:10.21105/joss.06830
Jarvis ED, Formenti G, Rhie A, Guarracino A, Yang C, Wood J, Tracey A, Thibaud-Nissen F, Vollger MR, Porubsky D, et al. 2022. Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611:** 519–531. doi:10.1038/s41586-022-05325-5
Ji Y, Zhou Z, Liu H, Davuluri RV. 2021. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* **37:** 2112–2120. doi:10.1093/bioinformatics/btab083
Karlin S, Burge C. 1995. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* **11:** 283–290. doi:10.1016/S0168-9525(00)89076-9
Kokot M, Długosz M, Deorowicz S. 2017. KMC 3: counting and manipulating *k*-mer statistics. *Bioinformatics* **33:** 2759–2761. doi:10.1093/bioinformatics/btx304
Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM. 2018. De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol* **36:** 1174–1182. doi:10.1038/nbt.4277
Koren S, Bao Z, Guarracino A, Ou S, Goodwin S, Jenike KM, Lucas J, McNulty B, Park J, Rautiainen M, et al. 2024. Gapless assembly of complete human and plant chromosomes using only nanopore sequencing. *Genome Res* **34:** 1919–1930. doi:10.1101/gr.279334.124
Koutsovoulos G, Kumar S, Laetsch DR, Stevens L, Daub J, Conlon C, Maroon H, Thomas F, Aboobaker AA, Blaxter M. 2016. No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. *Proc Natl Acad Sci* **113:** 5053–5058. doi:10.1073/pnas.1600338113
Kurtz S, Narechania A, Stein JC, Ware D. 2008. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics* **9:** 517. doi:10.1186/1471-2164-9-517
Kyriakidou M, Tai HH, Anglin NL, Ellis D, Strömvik MV. 2018. Current strategies of polyploid plant genome sequence assembly. *Front Plant Sci* **9:** 1660. doi:10.3389/fpls.2018.01660
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359. doi:10.1038/nmeth.1923
Lewin HA, Richards S, Lieberman Aiden E, Allende ML, Archibald JM, Bálint M, Barker KB, Baumgartner B, Belov K, Bertorelle G, et al. 2022. The Earth bioGenome Project 2020: starting the clock. *Proc Natl Acad Sci* **119:** e2115635118. doi:10.1073/pnas.2115635118
Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34:** 3094–3100. doi:10.1093/bioinformatics/bty191
Li H, Durbin R. 2024. Genome assembly in the telomere-to-telomere era. *Nature Reviews Genetics* **25:** 658–670. doi:10.1038/s41576-024-00718-w
Li X, Waterman MS. 2003. Estimating the repeat structure and length of DNA sequences using ℓ-tuples. *Genome Res* **13:** 1916–1922. doi:10.1101/gr.1251803
Limasset A, Flot J-F, Peterlongo P. 2020. Toward perfect reads: self-correction of short reads via mapping on de Bruijn graphs. *Bioinformatics* **36:** 1374–1381. doi:10.1093/bioinformatics/btz102
Lipman DJ, Pearson WR. 1985. Rapid and sensitive protein similarity searches. *Science* **227:** 1435–1441. doi:10.1126/science.2983426
Lippert RA, Huang H, Waterman MS. 2002. Distributional regimes for the number of *k*-word matches between two random sequences. *Proc Natl Acad Sci* **99:** 13980–13989. doi:10.1073/pnas.202468099
Liu D, Singh GB. 2000. Profile based methods for genomic sequence retrieval. *ACM SIGBIO Newsletter* **20:** 6–13. doi:10.1145/370954.370969
Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, Li Z, Chen Y, Mu D, Fan W. 2013. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv:1308.2012 [q-bio.GN]. doi:10.48550/arXiv.1308.2012
Liu S, Zheng J, Migeon P, Ren J, Hu Y, He C, Liu H, Fu J, White FF, Toomajian C, et al. 2017. Unbiased *k*-mer analysis reveals changes in copy number of highly repetitive sequences during maize domestication and improvement. *Sci Rep* **7:** 42444. doi:10.1038/s41598-016-0028-x
Logsdon GA, Rozanski AN, Ryabov F, Potapova T, Shepelev VA, Catacchio CR, Porubsky D, Mao Y, Yoo D, Rautiainen M, et al. 2024. The variation and evolution of complete human centromeres. *Nature* **629:** 136–145. doi:10.1038/s41586-024-07278-3
Lohse K, Harrison RJ, Barton NH. 2011. A general method for calculating likelihoods under the coalescent process. *Genetics* **189:** 977–987. doi:10.1534/genetics.111.129569
Lohse K, Chmelik M, Martin SH, Barton NH. 2016. Efficient strategies for calculating blockwise likelihoods under the coalescent. *Genetics* **202:** 775–786. doi:10.1534/genetics.115.183814
Mackintosh A, Laetsch DR, Hayward A, Charlesworth B, Waterfall M, Vila R, Lohse K. 2019. The determinants of genetic diversity in butterflies. *Nat Commun* **10:** 3466. doi:10.1038/s41467-019-11308-4
Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. 2017. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33:** 574–576. doi:10.1093/bioinformatics/btw663
Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27:** 764–770. doi:10.1093/bioinformatics/btr011
Marchet C, Boucher C, Puglisi SJ, Medvedev P, Salson M, Chikhi R. 2021. Data structures based on *k*-mers for querying large collections of sequencing data sets. *Genome Res* **31:** 1–12. doi:10.1101/gr.260604.119
Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34:** i142–i150. doi:10.1093/bioinformatics/bty266
Mikolov T, Chen K, Corrado G, Dean J. 2013. Efficient estimation of word representations in vector space. arXiv:1301.3781 [cs.CL]. doi:10.48550/arXiv.1301.3781
Mohamadi H, Khan H, Birol I. 2017. ntCard: a streaming algorithm for cardinality estimation in genomics data. *Bioinformatics* **33:** 1324–1330. doi:10.1093/bioinformatics/btw832
Mullikin JC, Ning Z. 2003. The phusion assembler. *Genome Res* **13:** 81–90. doi:10.1101/gr.731003
Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res* **30:** 1291–1305. doi:10.1101/gr.263566.120
Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* **376:** 44–53. doi:10.1126/science.abj6987
Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17:** 132. doi:10.1186/s13059-016-0997-x
Palmer DH, Rogers TF, Dean R, Wright AE. 2019. How to identify sex chromosomes and their turnover. *Mol Ecol* **28:** 4709–4724. doi:10.1111/mec.15245
Patro R, Mount SM, Kingsford C. 2014. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol* **32:** 462–464. doi:10.1038/nbt.2862
Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14:** 417–419. doi:10.1038/nmeth.4197
Pevzner PA, Tang H, Waterman MS. 2001. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci* **98:** 9748–9753. doi:10.1073/pnas.171285098

Pibiri GE, Fan J, Patro R. 2023. Meta-colored compacted de Bruijn graphs. bioRxiv doi:10.1101/2023.07.21.550101v2
Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* **11:** 1432. doi:10.1038/s41467-020-14998-3
Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, Eichler EE, Phillippy AM, Koren S. 2023. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* **41:** 1474–1482. doi:10.1038/s41587-023-01662-6
Reinar WB, Tørresen OK, Nederbragt AJ, Matschiner M, Jentoft S, Jakobsen KS. 2023. Teleost genomic repeat landscapes in light of diversification rates and ecology. *Mob DNA* **14:** 14. doi:10.1186/s13100-023-00302-9
Reinert G, Schbath S, Waterman MS. 2000. Probabilistic and statistical properties of words: an overview. *J Comput Biol* **7:** 1–46. doi:10.1089/10665270050081360
Ren J, Bai X, Lu YY, Tang K, Wang Y, Reinert G, Sun F. 2018. Alignment-free sequence analysis and applications. *Annu Rev Biomed Data Sci* **1:** 93–114. doi:10.1146/annurev-biodatasci-080917-013431
Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21:** 245. doi:10.1186/s13059-020-02134-9
Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592:** 737–746. doi:10.1038/s41586-021-03451-0
Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, Chiari Y, Dernat R, Duret L, Faivre N, et al. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* **515:** 261–263. doi:10.1038/nature13685
Rouzé T, Martayan I, Marchet C, Limasset A. 2023. Fractional hitting sets for efficient and lightweight genomic data sketching. bioRxiv doi:10.1101/2023.06.21.545875v1
Rowe WPM. 2019. When the levee breaks: a practical guide to sketching algorithms for processing the flood of genomic data. *Genome Biol* **20:** 199. doi:10.1186/s13059-019-1809-x
Sarmashghi S, Balaban M, Rachtman E, Touri B, Mirarab S, Bafna V. 2021. Estimating repeat spectra and genome length from low-coverage genome skims with RESPECT. *PLoS Comput Biol* **17:** e1009449. doi:10.1371/journal.pcbi.1009449
Schatz MC, Delcher AL, Salzberg SL. 2010. Assembly of large genomes using second-generation sequencing. *Genome Res* **20:** 1165–1173. doi:10.1101/gr.101360.109
Session AM, Rokhsar DS. 2023. Transposon signatures of allopolyploid genome evolution. *Nat Commun* **14:** 3180. doi:10.1038/s41467-023-38560-z
Session AM, Uno Y, Kwon T, Chapman JA, Toyoda A, Takahashi S, Fukui A, Hikosaka A, Suzuki A, Kondo M, et al. 2016. Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature* **538:** 336–343. doi:10.1038/nature19840
Shannon CE. 1948. A mathematical theory of communication. *Bell Syst Tech J* **27:** 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x
Shaw J, Yu YW. 2024. Rapid species-level metagenome profiling and containment estimation with sylph. *Nat Biotechnol* doi:10.1038/s41587-024-02412-y
Srivastava A, Malik L, Sarkar H, Zakeri M, Almodaresi F, Soneson C, Love MI, Kingsford C, Patro R. 2020. Alignment and mapping methodology influence transcript abundance estimation. *Genome Biol* **21:** 239. doi:10.1186/s13059-020-02151-8
Stoler N, Nekrutenko A. 2021. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom Bioinform* **3:** lqab019. doi:10.1093/nargab/lqab019
Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, 1000 Genomes Project, et al. 2010. Diversity of human copy number variation and multicopy genes. *Science* **330:** 641–646. doi:10.1126/science.1197005
Sun H, Ding J, Piednoël M, Schneeberger K. 2018. *findGSE*: estimating genome size variation within human and *Arabidopsis* using *k*-mer frequencies. *Bioinformatics* **34:** 550–557. doi:10.1093/bioinformatics/btx637
Ulrich J-U, Renard BY. 2024. Fast and space-efficient taxonomic classification of long reads with hierarchical interleaved XOR filters. *Genome Res* **34:** 914–924. doi:10.1101/gr.278623.123
VanWallendael A, Alvarez M. 2022. Alignment-free methods for polyploid genomes: quick and reliable genetic distance estimation. *Mol Ecol Resour* **22:** 612–622. doi:10.1111/1755-0998.13499
Vea IM, de la Filia AG, Jaron KS, Mongue AJ, Ruiz-Ruano FJ, Barlow SEJ, Nelson R, Ross L. 2021. The B chromosome of *Pseudococcus viburni*: a selfish chromosome that exploits whole-genome meiotic drive. bioRxiv doi:10.1101/2021.08.30.458195v1
Voichek Y, Weigel D. 2020. Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nat Genet* **52:** 534–540. doi:10.1038/s41588-020-0612-7
Vollger MR, Dishuck PC, Sorensen M, Welch AE, Dang V, Dougherty ML, Graves-Lindsay TA, Wilson RK, Chaisson MJP, Eichler EE. 2019. Long-read sequence and assembly of segmental duplications. *Nat Methods* **16:** 88–94. doi:10.1038/s41592-018-0236-3
Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33:** 2202–2204. doi:10.1093/bioinformatics/btx153
Wang Y, Yu J, Jiang M, Lei W, Zhang X, Tang H. 2023. Sequencing and assembly of polyploid genomes. *Methods Mol Biol* **2545:** 429–458. doi:10.1007/978-1-0716-2561-3_23
Whiteford N, Haslam N, Weber G, Prügel-Bennett A, Essex JW, Roach PL, Bradley M, Neylon C. 2005. An analysis of the feasibility of short read sequencing. *Nucleic Acids Res* **33:** e171. doi:10.1093/nar/gni170
Wittler R. 2023. General encoding of canonical *k*-mers. *Peer Community J* **3:** e87. doi:10.24072/pcjournal.323
Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **15:** R46. doi:10.1186/gb-2014-15-3-r46
Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol* **20:** 257. doi:10.1186/s13059-019-1891-0
Yang J, Moeinzadeh M-H, Kuhl H, Helmuth J, Xiao P, Haas S, Liu G, Zheng J, Sun Z, Fan W, et al. 2017. Haplotype-resolved sweet potato genome traces back its hexaploidization history. *Nat Plants* **3:** 696–703. doi:10.1038/s41477-017-0002-z
Zaheer M, Guruganesh G, Dubey A, Ainslie J, Alberti C, Ontanon S, Pham P, Ravula A, Wang Q, Yang L, et al. 2020. Big Bird: transformers for longer sequences. arXiv:2007.14062 [cs.LG]. doi:10.48550/arXiv.2007.14062
Zhang C, Xie L, Yu H, Wang J, Chen Q, Wang H. 2023. The T2T genome assembly of soybean cultivar ZH13 and its epigenetic landscapes. *Mol Plant* **16:** 1715–1718. doi:10.1016/j.molp.2023.10.003
Zhao X. 2019. BinDash, software for fast genome distance estimation on a typical personal laptop. *Bioinformatics* **35:** 671–673. doi:10.1093/bioinformatics/bty651

Received April 11, 2024; accepted in revised form January 9, 2025.