# Beyond DNA barcoding: The unrealised potential of genome skim data in sample identification

*Running title: Identification using genome skimmed nuDNA*

Kristine Bohmann[1], Siavash Mirarab[2], Vineet Bafna[3], M Thomas P Gilbert[1,4,5]

[1]Section for Evolutionary Genomics, The GLOBE Institute, University of Copenhagen, Øster Farimagsgade 5A, 1352 Copenhagen, Denmark.

[2]Department of Electrical and Computer Engineering, University of California, San Diego, California, USA.

[3]Department of Computer Science and Engineering, University of California, San Diego, California, USA.

[4]Center for Evolutionary Hologenomics, The GLOBE Institute, University of Copenhagen, Øster Farimagsgade 5A, 1352 Copenhagen, Denmark.

[5] NTNU University Museum, N-7491 Trondheim, Norway.

**Author for Correspondence**

M. Thomas P. Gilbert, Tel: +45 23712519, email: tgilbert@sund.ku.dk

**Abstract**

Genetic tools are increasingly used to identify and discriminate between species. One key transition in this process was the recognition of the potential of the ca 658bp fragment of the organelle cytochrome c oxidase I (COI) as a barcode region, which revolutionised animal bioidentification and lead, among others, to the instigation of the Barcode of Life database (BOLD), containing currently barcodes from >7.9 million specimens. Following this discovery, suggestions for other organellar regions and markers, and the primers with which to amplify them, have been continuously proposed. Most recently, the field has taken the leap from PCR based generation of DNA references into shotgun sequencing-based 'genome skimming' alternatives, which the ultimate goal of assembling organellar reference genomes. Unfortunately, in genome skimming approaches, much of the nuclear genome (as much as 99% of the sequence data) is discarded, which is not only wasteful but can also limit the power of discrimination at or below the species level. Here, we advocate that the full shotgun sequence data can be used to assign an identity (that we term for convenience its 'DNA-mark') for both voucher and query samples, without requiring any computationally intensive pretreatment (e.g., assembly) of reads. We argue that if reference databases are populated with such 'DNA-marks', it will enable future DNA-based taxonomic identification to complement, or even replace PCR of barcodes with genome skimming, and we discuss how such methodology ultimately could enable identification to population, or even individual, level.

**Keywords**: Biodiversity, DNA Barcoding, DNA reference databases, Environmental DNA, K-mers, Next Generation Sequencing

## From DNA barcoding to DNA-marking

DNA sequences are increasingly being applied as a tool with which to assign identity to query samples, most famously through the use of so-called ‘DNA barcodes’ (Hebert, Cywinska, Ball, & deWaard, 2003a). Originally conceived as a ca 658bp fragment of the organelle cytochrome c oxidase I (COI) gene to serve as a taxonomic tool for use in animal bioidentification, the idea was elegant. Users would PCR amplify and then Sanger sequence this marker, chosen based on their observations using lepidopterans as a test, to be conserved enough to be targeted with generic (pan-taxa) primer sets, while variable enough to provide variation at the interspecific level (while similarly not varying at the intra-specific level). This elegant idea of a barcoding region with which to tell species across life forms apart quickly caught on, and subsequently a flurry of other organellar regions and markers and associated primer sets were proposed. For example, 16s rRNA was used for animals including mammals (Taylor, 1996), amphibians (Vences, Thomas, van der Meijden, Chiari, & Vieites, 2005) and insects (Clarke, Soubrier, Weyrich, & Cooper, 2014); 12s was proposed for vertebrates (Riaz et al., 2011); Matk (Lahaye et al., 2008) and rbcl (Fazekas et al., 2008) for plants; and ITS for fungi, (Schoch et al., 2012).

As DNA barcoding’s potential became increasingly apparent, it spurred rapid development in a range of associated laboratory and computational techniques to help optimise its performance, through facilitating efficient generation of high quality and economical data. In the laboratory, progress has principally been focused on decreasing the costs for generating single DNA reference and query barcodes - a key step for democratising its use. For example, the state-of-the-art is to use Illumina (Liu, Yang, Zhou, & Zhou, 2017) or PacBio (Hebert et al., 2018) technology to simultaneously sequence multiplexed amplicons derived from voucher specimens, so as to generate tens of thousands of sequence in parallel, thus decreasing sequencing costs to only a few cents per barcode (Hebert et al., 2018). A second avenue of progress relates to the development of computational methods designed to optimise the information potential of barcode data, in particular in light of challenges such as error within

query barcode sequences or incomplete or even erroneous reference databases (e.g. Bridge, Roberts, Spooner, & Panchal, 2003; Briski, Ghabooli, Bailey, & MacIsaac, 2016)). However, perhaps the most important of these developments was the realisation that the power of barcoding is constrained by the quality of reference data against which to compare query sequences, thus the need for comprehensive and curated barcode reference databases based on the sequencing of vouchered information. Hebert and team's BOLD database (Ratnasingham & Hebert, 2007) epitomises this ideal, containing barcode sequences from over >7.9 million specimens (http://www.boldsystems.org/index.php, retrieved February 2020).

Recently, DNA reference databases are increasingly being complemented by shotgun sequencing-based 'genome skimming' alternatives (Coissac, Hollingsworth, Lavergne, & Taberlet, 2016; Nevill et al., 2020; Zeng et al., 2018). In such approaches, while the original barcode loci are sequenced (Liu et al., 2013) with probability depending upon the coverage, the biggest benefit comes from the sequencing and assembly of organellar genomes (Gillett et al., 2014). Unfortunately, much of the nuclear genome (as much as 99% of the sequence) is discarded. Ultimately, this can limit the power of discrimination at or below the species level (Rubinoff & Holland, 2005).

As such, we build on the suggestion first outlined by Coissac and colleagues (Coissac et al., 2016), and advocate that the full shotgun sequence data generated from voucher specimens could also be used to assign an identity (that we term for convenience here its *'DNA-mark'*), without requiring any computationally intensive pretreatment (e.g., assembly) of reads. With such reference information in place, we argue that future studies that aim to assign an identity to query samples could complement, or even replace PCR of barcodes with shotgun sequencing, yielding data that could be matched to information in the reference database using computational methods that treat both the query and reference samples as "bags of

reads" (Sarmashghi, Bohmann, P. Gilbert, Bafna, & Mirarab, 2019). We believe that this methodology ultimately could enable identification to population, or even individual, level.

**The limits of traditional barcoding**

It is impossible to overstate the impact that traditional single-locus DNA barcoding has had over the past 15 years, and it will without doubt continue to represent a fundamental pillar of many future studies. However, after such extensive use, its limitations are also now apparent, raising the obvious question as to whether these can be overcome? Principal among them is the taxonomic resolution at which traditional barcodes can effectively operate - having been chosen with the aim of discriminating at the species level (although even this is not guaranteed), they work sub-optimally as one moves below the species to other units that may interest end users - such as the population, or even individual. This problem is confounded by the 'barcoding gap' challenge, namely that the genetic distance between taxonomic units is not a constant, thus while traditional barcodes may be effective in discriminating between different species in one genus, they may fail to perform on other genera (Shearer & Coffroth, 2008; e.g. Wiemers & Fiedler, 2007). A third limitation inherent to their relatively short length means they rarely can be used to resolve phylogenies with high statistical support, while a fourth challenge relates to the minimum length of intact DNA templates required to successfully PCR amplify a barcode locus. The DNA content of many specimens of interest is often heavily degraded due to age, storage conditions, or chemical treatment, and remaining fragments may simply be too short to allow initial PCR amplification step (Orlando, Gilbert, & Willerslev, 2015). Lastly, heavily degraded samples may also be contaminated with exogenous sources of DNA, which given the sensitivity of PCR, can potentially lead to the co (or even preferential) amplification of the contaminant over the true target (Hofreiter, Serre, Poinar, Kuch, & Pääbo, 2001).

The decreasing cost of sequencing using so-called Next Generation Sequencing (NGS) technologies has provided partial solutions to this problem, in particular thanks to the

131 introduction of the ‘genome skimming’ approaches (Coissac et al., 2016). In their current
132 implementation, DNA extracted from voucher specimens are converted into NGS libraries,
133 shotgun sequenced to relatively low genome coverage, then either original barcode loci such
134 as COI (Liu et al., 2013), or full organellar genomes are reconstructed bioinformatically from
135 this data (Fig. 1) (Gillett et al., 2014). Thanks to library indexing, many samples can be
136 multiplexed before sequencing, meaning that many tens (or even hundreds) of organelle
137 genomes can be sequenced on a single sequencing run (even more, if coupled to target-
138 enrichment (Liu et al., 2016)). This yields a significant increase in information potential. This
139 is further increased by the reduction in DNA preservation requirements when bypassing the
140 conventional PCR step. For genome skimming, DNA fragments as short as 25-30 bp are
141 usable, in stark contrast to the ca 700 bp requirement in traditional barcoding, which can hinder
142 generation of reference sequences from old or badly preserved specimens. In light of these
143 benefits, today several projects have actively chosen to employ genome skimming over
144 traditional PCR to generate barcode-like data, for example the PhyloAlps (phyloalps.org),
145 NORBOL (norbol.org) and DNAmark (dnamark.ku.dk) initiatives, and in doing so are extending
146 the concept of traditional DNA barcode reference databases (Hebert, Cywinska, Ball, &
147 deWaard, 2003b) to encompass organelle genome data. However, while this represents a
148 natural development to traditional barcoding, we highlight that even this approach has its
149 limits. Should sufficient genetic diversity and population structure exist in the target species,
150 organelle genomes might enable us to narrow identification to the sub-species, or even
151 population level; however, unless organelle haplotypes are unique to individual organisms,
152 their resolution stops here. Furthermore, inferences based on single non-recombining loci (no
153 matter how long) are notoriously susceptible to challenges such as Incomplete Lineage
154 Sorting, thus making them suboptimal for assigning identity or inferring evolutionary histories
155 (Funk & Omland, 2003; McKay & Zink, 2010). Lastly and importantly, genome skimming
156 simply seems wasteful as it only exploits a fraction of the generated sequence data. The
157 nuclear DNA component of shotgun sequenced DNA extracts can represent >99% of the

158 reads (Liu et al., 2016), and we argue this holds valuable information that can further the goals
159 of sample identification.
160
161 **Exploiting the power of the nuclear genome**
162 Given that the nuclear genome sequence of any non-clonal organism is a representation of its
163 evolutionary history, it represents the ultimate source of information for those wishing to assign
164 identity to samples. In theory, with enough reference data one could identify every genetically
165 distinct organism on the planet. As such, if one looks to the future, the obvious desirable end
166 goal would be to generate fully assembled nuclear genomes from both query and voucher
167 samples and to do this across the entire Tree of Life, as advocated for example by initiatives
168 such as the Earth Biogenome Project (Lewin et al., 2018) (https://www.earthbiogenome.org/),
169 which are starting to be realised through projects such as the Darwin Tree of Life Project (
170 https://www.sanger.ac.uk/science/collaboration/darwin-tree-life-project). Unfortunately
171 however, while sequencing technology is advancing at a remarkable rate thanks to the
172 increases in accuracy, read length and overall output of platforms such as the PacBio Sequel
173 II, which has allowed generation of largely complete genome assemblies for many organisms,
174 the assembled nuclear genomes come with their own challenges. Firstly, nuclear genomes
175 are expensive to generate as they require sequencing to high depths of coverage. Secondly,
176 the assembly is constrained by depth of sequencing and the repeat structure of the genome.
177 On the one hand, if the depth of sequencing is high, then the computational power needed for
178 the assembly is very high. On the other hand, sequencing depth cannot be too small either as
179 this will be problematic for successful assembly. Typically, a minimum depth of coverage is
180 required that falls in the range of at least 50x for a relatively straightforward diploid organism
181 (Sohn & Nam, 2018). A further challenge is repeat sequences, which when longer than the
182 reads sequenced, can prevent unambiguous assembly. Repeats can be resolved by
183 construction of mate-pair/large insert libraries for short-read technologies or extraction of high
184 molecular weight DNA and long-read sequencing using single molecule sequencing. This in

185 turn limits both which specimens can be used, and complicates the requisite laboratory
186 equipment and skills.
187
188 In summary, the costs of assembling nuclear genomes are high, both with regards to cost,
189 data generation and the computational assembly. This puts nuclear genomes well beyond the
190 budgets and capabilities of most people actively interested in using DNA as a tool for routine
191 taxonomic assignment of many samples. However, given that nuclear genome sequences are
192 unique, regardless of whether they have been assembled into contigs, scaffolds or
193 chromosomes, it follows that even unassembled shotgun sequence data might hold
194 information that could be exploited for taxonomic assignment. And thus given such data is
195 already being generated by current reference database genome skimming and genome
196 projects, we argue that now is the time to explore its potential and develop suitable laboratory
197 and computational tools for its exploitation.
198
199 **Unleashing the full potential of genome skimming using assembly-free methods**
200 How might we best exploit this residual nuclear DNA data? The ideal solution would be an
201 approach that is fast, simple and efficient, and at least in the short term while sequencing costs
202 are still in the range of >10 USD per GB (Rachtman, Balaban, Bafna, & Mirarab, 2020), restrict
203 sequencing effort to a minimum. Our proposed solution is to instead use unassembled reads
204 from the nuclear genome (so-called “bags of reads”) to perform the function currently assigned
205 to barcodes (or organelle genomes), namely populate reference databases against which
206 queries can be matched (Fig. 1 & 2). Critically, such a method would need to be simple and
207 intuitive, and computationally efficient - both with regards to data processing and storage.
208
209 Coissac and colleagues (Coissac et al., 2016) have suggested that assembly-free and
210 mapping-free methods (Blaisdell, 1986; Fan, Ives, Surget-Groba, & Cannon, 2015; Maillet,
211 Collet, Vannier, Lavenier, & Peterlongo, 2014; Song et al., 2013; Vinga & Almeida, 2003)
212 naturally meet many of these criteria. They are typically fast and conceptually simple.

Following this aim, several groups have recently developed methods specifically aimed at handling characteristics specific to genome skimming, including low coverage and sequencing errors (Sarmashghi et al., 2019; Tang, Ren, & Sun, 2019). Indeed, many alignment-free methods are available and their application to genome skims should be explored. We note that accurate analyses of skimming data will require several computational components (Fig. 2). In recent years, a new toolkit of methods for analyzing skimming data has started to emerge. Below, we discuss some of these advances, focusing specifically on analyses based on short oligomers, or k-mers.

**K-mer-based distance calculation.** A collection of k-mers sampled at random from the nuclear genome encodes a remarkable amount of information. For a genome of size $n$, and ignoring repeats, a k-mer of sufficient size ($log_2 n$) will be unique in that genome with high probability. Helpfully, the probability of finding that k-mer in another genome relates directly to the evolutionary distance to the other genome. Modelling two genome-skims simply as sets of k-mers $A$ and $B$, we can define the fraction of shared k-mers by the Jaccard index:

$$J = \frac{|A \cap B|}{|A \cup B|}.$$

$J$ is intimately connected to the genomic distance $D$ between the two organisms (Fan et al., 2015). Assuming all mutations to be equally likely, we can estimate $D$ as

$$D = 1 - \left(\frac{2J}{1+J}\right)^{\frac{1}{k}}.$$

Moreover, $J$ can be computed efficiently, using as few as $10^3$ k-mers using hashing techniques (Ondov 2016). However, this method assumes the coverage is high enough that each k-mer

is sampled at least once. Recently, we developed a method called Skmer that allows for accurate estimation of genomic distance with extremely low (e.g., 0.1X) coverage, even when the coverage is unknown and in the presence of sequencing errors (Sarmashghi et al., 2019). Skmer uses k-mer frequencies to estimate genome length, coverage, and sequencing error and uses the Jaccard index to compute genomic distance using a more complex version of the equation above. Because assembly is not needed, adding new species to the reference set of Skmer requires minimal preprocessing or indexing, and thus, is straightforward. While Skmer has performed well in comparison to other assembly-free methods (Sarmashghi et al., 2019; Zielezinski et al., 2019), our intention here is not to advocate Skmer specifically; our general arguments apply to other assembly-free methods (see Zielezinski et al., 2019).

**Sample identification.** Once the genomic distance is measured, sample identification can follow the standard approach of finding the voucher species with the smallest distance to the query. The tool Skmer has shown high accuracy in this setting. For example, on datasets of *Anopheles* mosquitos with genome skims of size 0.1, 0.5, or 1Gb (corresponding to ~0.5X-7X coverage), Skmer correctly identified the best match to every query skim, even when species close to the query were removed from the reference set (Sarmeshgi, 2018); in more challenging datasets of *Drosophila* and *birds*, Skmer was still correct in 190 out of 210, and 375 out of 460 tests, respectively.

When an exact match to the query species is not available in the reference set, a phylogenetic approach is helpful. Phylogenetic placement can find the best placement of the query on a reference phylogeny of vouchers. Recently developed methods such as APPLES can perform phylogenetic placement using distances alone (Balaban, Sarmashghi, & Mirarab, 2019). Phylogenetic placement can improve accuracy of identification. For example, in a leave-one-out reanalysis of a dataset of 61 lice genome skims (Boyd et al., 2017), APPLES was able to find the correct phylogenetic placement in 97% of cases, whereas simply picking the closest match was accurate in only 54% of the tests (Balaban et al., 2019).

**Read cleanup and filtering.** Before computing distances between DNA-marks, several technical and conceptual issues must be addressed. Standard processing of reads, including adapter removal, deduplication, and merging of paired-end-reads are all needed and can be achieved using standard tools such as BBTools (Bushnell, 2014). A remaining type of preprocessing needed is dealing with extragenic DNA from sources other than the species of interest. While this is a serious issue, we note that it is not unique to a DNA-mark approach, and rather represents an important challenge for the field, and we revisit it later in the article.

**Why haven't genome-wide approaches been adopted yet?**

One valid question is why such approaches have not already been adopted? Firstly, until recently, shotgun sequencing costs per unit sequenced have simply been prohibitively expensive. Nevertheless, as sequencing costs per base continue to drop, the end-to-end costs will be increasingly dominated by processes necessary to the data generation (Fig. 3). This includes for example, the salaries of staff paid to collect voucher samples, extract and generate the DNA data, assemble and run QC on the results, and ultimately upload the data and accessory information into reference databases. Thus, while the difference purely in economic cost of PCR versus shotgun sequencing may at first look significant, the difference in true cost becomes minimal (Fig. 3). Secondly, it might be assumed that the computational burden associated with any NGS-based method is high. However, as already alluded to above, computational burdens for assembly-free methods are considerably reduced. For example, the total running time (using 24 CPU cores) to compute 1081 distances between all pairs of 48 avian genome skims using the Skmer tool took only 33 minutes (Sarmashghi et al., 2019). Thirdly, while map-free, alignment-free methods of comparing genomes (including some based on k-mers) have been known in the Bioinformatics community (Marçais & Kingsford, 2011; Ondov et al., 2016), the power of k-mer analysis for making inference with low-coverage genome-skims was not well understood until recently (Fan et al., 2015; Sarmashghi et al., 2019). Following these advances, user-friendly software programs to

efficiently use the k-mer data are being actively developed, and new methods for improving their accuracy and usability are being designed.

We would argue that the only thing stopping this approach being implemented now is an exploration of its performance and potential, alongside the development of appropriate laboratory methods (such as efficient and cost-effective library build protocols applicable to badly preserved voucher specimens, e.g. Troll et al. (2019)) and development of reference databases with suitable infrastructure.

**Open methodological questions**

As mentioned above, methods for computing genomic distance from genome skims and for phylogenetic analysis of those distances exist. Despite the progress, several unanswered methodological questions need to be further explored by the research community. Some of the questions are computational in nature while others are related to lab techniques and the curation of comprehensive reference libraries. In the following section we briefly discuss what some of these might be.

**Computational questions**

**Coverage:** A natural question is what depth of coverage will be needed for accurate sample identification. The answer is not straightforward and will depend on many factors, including genome length, sequencing errors introduced due to either post mortem DNA degradation (Lindahl, 1993; Pääbo, 1989) or library preparation enzyme and platform sequencing chemistry, and perhaps even the genomic architecture (e.g., the prevalence of repeats and polyploidy). The required depth of coverage is also a function of the genetic similarity between taxa. For example the coverage required to distinguish a human from a chimpanzee sample would be higher than human from gibbon, simply as the former pair share many more k-mers than the latter. Thus, a single number will not be universally applicable to different groups. Moreover, within species diversity is highly variable across the tree of life (Leffler et al., 2012).

Nevertheless, our initial studies show that for species-level identification, 1X coverage may be sufficient in most cases (Sarmashghi et al., 2019), and thus given our aforementioned argument that labor, not sequencing, is the bottleneck, perhaps, using a fixed sequencing effort (say, 2Gb per species) would suffice in most cases. Nevertheless, more research is needed to characterize the exact resolution that can be obtained for a given coverage. This question has to be studied for different types of species with different genomic architectures.

**Population-level characterization.** Related to the question of coverage is the question of resolution: Can a DNA-mark distinguish groups at the subspecies level? Current methods such as Skmer tend to have very high accuracy for distances as low as $10^{-2}$ and reasonable accuracy for distances in the $10^{-3}$ range. For some groups, sub-species identification will require finer resolution. Accurately computing even lower distances despite low coverage (e.g., 1-5X) may be possible with improved methods. We believe increasing the resolution will require a more complex modelling of the genomic structure and in particular the profile of the repeated k-mers across the genome. However, disentangling repeat structure from the k-mer frequency profiles observed due to the random coverage of the genome is not easy and will require new algorithms.

**Mutational models:** Any measure of genomic distance is tightly linked with mutational processes that are modeled. For example, the Skmer method directly models substitutions but not processes such as insertions and deletions, gene duplications and losses, abundant repeats, polyploidy, and horizontal gene transfer. Some of these mutation types (e.g., indels) are arguably modeled by Skmer indirectly. Nevertheless, the robustness of the k-mer based methods needs to be tested and improved in the face of complex mutations such as large-scale duplications. This is especially important for plants and other organisms with complex genomic architecture.

**Sequencing technology:** The exact choice of the sequencing technology will affect not only the lengths of sequences generated and sequencing error rates, but can also introduce biases through preferential sequencing of certain regions over others due to GC content etc (Browne et al., 2020). All of these may impact the accuracy of k-mer-based methods. In practice, it may also be that a reference dataset would be composed of skims sequenced with different technologies. Would query searches against such databases remain unbiased? Since k-mers break down long sequences into short ones anyway, there is reason to hope that they will remain robust to the choice of the sequencing technology. Nevertheless, empirical tests with mixed sequencing technologies currently do not exist.

**Sampling:** If reference databases are not comprehensive, and this goes for any reference database whether traditional barcode, organelle genome or k-mer reference databases, taxonomic assignments of queries can suffer. Besides developing reference libraries with denser sampling, a phylogenetic perspective can also be helpful, as the metagenomics community has learned (Brady & Salzberg, 2009; Janssen et al., 2018; Matsen, 2015; Matsen, Kodner, & Armbrust, 2010; Nguyen, Mirarab, Liu, Pop, & Warnow, 2014). Considering phylogenetic relationships between the query and reference sequences, we can look for the largest taxonomic level (e.g., a genus, family, or class) in which the query can be confidently placed. To this end, we have developed algorithms that combine k-mer-based distances with phylogeny-based placement (Balaban et al., 2019). However, phylogenetic placement of genome skims can further benefit from methods that better characterize placement uncertainty, model rate variations and gene tree discordance across the genome, and incorporate complex substitution models.

**Extragenic DNA:** The most pernicious challenge is the possibility that the generated sequence data derives from more than one source. That is, voucher samples might not only contain DNA from the target species, but also that from other organisms. This could be from naturally impure voucher samples, for example endophytes associated with plants, or the gut

contents of preserved insects, or even simply a result of microbial driven degradation. Alternatively, it could derive from contamination during the laboratory procedures, or even library bleeding during sequencing as has been reported for some Illumina platforms (Kircher, Sawyer, & Meyer, 2012; Sinha et al., 2017) and which may yield impure datasets. While conventional PCR or genome skimming approaches are not immune to contamination, identification and removal of contaminants is a much more straightforward process.

A recent study showed that for assembly-free methods of genome matching, estimates of genomic distance are negatively impacted if contamination are not detected (Rachtman et al., 2020). Using both mathematical modelling and empirical data, the authors elucidated how the amount of contamination and the similarity of the contamination across skims being compared interact with negative impacts of contamination. Contaminating sequence reads can impact k-mer based measures of distance in complex ways. The most damaging scenario is when both the query *and* the reference skims are impure, especially if the impurity of the query skim happens to be similar to that of some reference skims. In a scenario like that, the estimated distance from the query to a reference may be low, not because of the phylogenetic similarity but because of the similarity in contaminants.

One approach to deal with sample impurity is to filter out reads suspected to be contaminants. Existing methods such as BLAST or Kraken (Wood, Lu, & Langmead, 2019) can be used to search reads against databases of known contaminants. For example, if the sample is known to be of an insect, we can match reads against databases of bacteria, fungi, viruses, and mammals. Any strong matches to these can be then eliminated. The analysis by Rachtman et al. (2020) has shown filtering using Kraken-II to be effective in reducing the negative impacts of contamination, but only when the contaminants have relatively close matches to the contaminant reference library (e.g., a match with up to 5-10% genomic distance). This observation leaves us with a methodological gap, namely, efficient yet more effective methods of read matching at higher distances. These search methods should go beyond (near) exact

405 matching to species available in the contaminant database, as those databases will always be
406 incomplete. Instead, they should use the databases as a guide to broadly find reads that have
407 likely originated from organisms other than the clade of interest. An alternative to this
408 “exclusion-filtering” method is inclusion-filtering: designing methods that can identify reads that
409 have, in fact, likely originated from some organism in the clade of interest.
410
411 **Mixture analysis.** The existing methodology for k-mer based analysis of DNA-marks mostly
412 assumes the sample is of one target species (plus contaminants). Akin to metabarcoding, we
413 can imagine a scenario where meta DNA-marks are obtained from samples that include a mix
414 of species of interest. For example, the sample may include a mix of several insects that are
415 hard to physically separate. Or it may be bee-bread, the collection of pollen from several plants
416 and fungi that constitute the food source in a bee nest. Can a DNA-mark from a mixed sample
417 be decomposed into its constituent parts? While designing methods to solve this problem is
418 not trivial, the success of the metagenomic field makes us optimistic that methods for
419 deconvoluting a DNA-marks into their constituent species can be developed in the near future.
420
421 **Sample collection, lab and sequencing developments**
422 As mentioned above, the DNA-mark approach could be complicated by sample impurity. Such
423 can arise at all steps of the workflow but at the very basal step at the point of sample collection.
424 As with other approaches for DNA reference data generation, it is best to collect samples for
425 DNA extraction and sequencing that contain as little DNA from other sources as possible. For
426 instance, avoiding obvious endophytes on plants and avoiding contamination by one’s own
427 DNA and from other sources during collection.
428
429 When generating all types of reference data, DNA-mark reference data included, we need to
430 do it efficiently, cost-effectively and reliably and ensure that it causes minimal destruction to
431 voucher specimens. For generation of DNA-mark reference data, and to some extent all of
432 this is valid for other approaches too, this can be achieved by following validated and

standardised workflows and pipelines. Importantly, these should seek to i) minimise (cross) contamination during lab work, through e.g. working in pre and post PCR laboratories and in clean working environments and by minimising hands-on-labour, e.g. through semi-automated laboratory processing on robots and semi-automated bioinformatic pipelines, ii) simplify DNA extractions so they are pure and relatively universal across sample types, and iii) ensure that protocols for preparation of DNA extracts for sequencing, the so-called library build, are as simple as possible, that they allow low quantities of input DNA and that they account for potential artefacts such as 'library bleeding', which if not taken into account can cause false assignment of sequences to samples and thereby contaminate samples (Kircher et al., 2012; Sinha et al., 2017). With regards to sequencing platforms, these need to be cheap, high-throughput, simple to use and reliable.

**Concluding remarks**

A community effort will be needed if we are to effectively address the challenges associated with using k-mers in general, and in parallel establish the required curated public DNA-mark reference database against which queries can be run. This could, for example, be comprised of both the processed genome skim data and the assembled organellar genomes that can be mined from genome skims. This in turn would ideally be based on both data submitted by those deliberately aiming to contribute to the database, and mined from pre-existing shotgun sequence datasets - as long as sufficient controls are in place to ensure that such data is derived from the taxa it is labelled with (something that has plagued genetic studies, including those based on conventional barcoding, since the introduction of such databases (Mioduchowska, Czyż, Gołdyn, Kur, & Sell, 2018)). Given that such data would naturally complement well established initiatives such as those comprising of either barcode fragments such as the Barcode of Life Database (BOLD), and/or organelle and whole genomes such as in Norbol, Phyloalps and DNAmark and the various initiatives under the Earth BioGenome Project, one desirable strategy might even be to simply embed the framework within one of these resources.

With such initial framework in place, our hope is that this will provide both a valuable tool with which to complement conventional barcoding, and also open up new research questions (Table 1). Obvious potential avenues include exploring whether such approaches might also be used to identify the genetic sources within more complex DNA mixtures, as is currently done using DNA metabarcoding of, for example environmental DNA or DNA extracted from bulk specimen samples (Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012). Other potential avenues could be as a new tool for reconstructing phylogenies, analysing the genetics of populations, and even identifying samples to the individual level.

**Acknowledgements**

The authors would like to thank the Aage V. Jensen Naturfond for their generous funding of the DNAmark project, and Ashot Margaryan for helpful discussion. SM and VB were supported by the National Science Foundation (NSF) grant IIS-1815485.

## References


Balaban, M., Sarmashghi, S., & Mirarab, S. (2019). APPLES: Scalable Distance-Based Phylogenetic Placement with or without Alignments. *Systematic Biology*, syz063.

Blaisdell, B. E. (1986). A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences*, *83*(14), 5155–5159.

Boyd, B. M., Allen, J. M., Nguyen, N.-P., Sweet, A. D., Warnow, T., Shapiro, M. D., … Johnson, K. P. (2017). Phylogenomics using Target-Restricted Assembly Resolves Intrageneric Relationships of Parasitic Lice (Phthiraptera: Columbicola). *Systematic Biology*, *66*(6), 896–911.

Brady, A., & Salzberg, S. L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods*, *6*(9), 673–676.

Bridge, P. D., Roberts, P. J., Spooner, B. M., & Panchal, G. (2003). On the unreliability of published DNA sequences. *The New Phytologist*, *160*(1), 43–48.

Briski, E., Ghabooli, S., Bailey, S. A., & MacIsaac, H. J. (2016). Are genetic databases sufficiently populated to detect non-indigenous species? *Biological Invasions*, *18*(7), 1911–1922.

Browne, P. D., Nielsen, T. K., Kot, W., Aggerholm, A., Gilbert, M. T. P., Puetz, L., … Hansen, L. H. (2020). GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. *GigaScience*, *9*(2). doi: 10.1093/gigascience/giaa008

Bushnell, B. (2014). BBTools software package. *URL Http://sourceforge. Net/projects/bbmap*.

Clarke, L. J., Soubrier, J., Weyrich, L. S., & Cooper, A. (2014). Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. *Molecular Ecology Resources*, *14*(6), 1160–1170.

Coissac, E., Hollingsworth, P. M., Lavergne, S., & Taberlet, P. (2016). From barcodes to genomes: extending the concept of DNA barcoding. *Molecular Ecology*, *25*(7), 1423–1428.

Fan, H., Ives, A. R., Surget-Groba, Y., & Cannon, C. H. (2015). An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC Genomics*, *16*(1), 522.

Fazekas, A. J., Burgess, K. S., Kesanakurti, P. R., Graham, S. W., Newmaster, S. G., Husband, B. C., … Barrett, S. C. H. (2008). Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PloS One*, *3*(7), e2802.

Funk, D. J., & Omland, K. E. (2003). Species-Level Paraphyly and Polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics*, *34*(1), 397–423.

Gillett, C. P. D. T., Crampton-Platt, A., Timmermans, M. J. T. N., Jordal, B. H., Emerson, B. C., & Vogler, A. P. (2014). Bulk De Novo Mitogenome Assembly from Pooled Total DNA Elucidates the Phylogeny of Weevils (Coleoptera: Curculionoidea). *Molecular Biology and Evolution*, *31*(8), 2223–2237.

Hebert, P. D. N., Braukmann, T. W. A., Prosser, S. W. J., Ratnasingham, S., deWaard, J. R., Ivanova, N. V., … Zakharov, E. V. (2018). A Sequel to Sanger: amplicon sequencing that scales. *BMC Genomics*, *19*(1), 219.

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003a). Biological identifications through DNA barcodes. *Proceedings. Biological Sciences / The Royal Society*, *270*(1512), 313–321.

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003b). Biological identifications through DNA barcodes. *Proceedings. Biological Sciences / The Royal Society*, *270*(1512), 313–321.

Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M., & Pääbo, S. (2001). Ancient DNA. *Nature Reviews. Genetics*, *2*(5), 353–359.

Janssen, S., McDonald, D., Gonzalez, A., Navas-Molina, J. A., Jiang, L., Xu, Z. Z., … Knight,

R. (2018). Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information. *mSystems*, *3*(3). doi: 10.1128/mSystems.00021-18

Kircher, M., Sawyer, S., & Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research*, *40*(1), e3.

Lahaye, R., van der Bank, M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G., … Savolainen, V. (2008). DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(8), 2923–2928.

Leffler, E. M., Bullaughey, K., Matute, D. R., Meyer, W. K., Ségurel, L., Venkat, A., … Przeworski, M. (2012). Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biology*, *10*(9), e1001388.

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., … Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(17), 4325–4333.

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature*, *362*(6422), 709–715.

Liu, S., Li, Y., Lu, J., Su, X., Tang, M., Zhang, R., … Zhou, X. (2013). SOAPBarcode: revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons. *Methods in Ecology and Evolution / British Ecological Society*, *4*(12), 1142–1150.

Liu, S., Wang, X., Xie, L., Tan, M., Li, Z., Su, X., … Others. (2016). Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. *Molecular Ecology Resources*, *16*(2), 470–479.

Liu, S., Yang, C., Zhou, C., & Zhou, X. (2017). Filling reference gaps via assembling DNA barcodes using high-throughput sequencing-moving toward barcoding the world. *GigaScience*, *6*(12), 1–8.

Maillet, N., Collet, G., Vannier, T., Lavenier, D., & Peterlongo, P. (2014). Commet: Comparing and combining multiple metagenomic datasets. *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. doi: 10.1109/bibm.2014.6999135

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* , *27*(6), 764–770.

Matsen, F. A., 4th. (2015). Phylogenetics and the human microbiome. *Systematic Biology*, *64*(1), e26–e41.

Matsen, F. A., Kodner, R. B., & Armbrust, E. V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, *11*, 538.

McKay, B. D., & Zink, R. M. (2010). The causes of mitochondrial DNA gene tree paraphyly in birds. *Molecular Phylogenetics and Evolution*, *54*(2), 647–650.

Mioduchowska, M., Czyż, M. J., Gołdyn, B., Kur, J., & Sell, J. (2018). Instances of erroneous DNA barcoding of metazoan invertebrates: Are universal cox1 gene primers too “universal”? *PloS One*, *13*(6), e0199609.

Nevill, P. G., Zhong, X., Tonti-Filippini, J., Byrne, M., Hislop, M., Thiele, K., … Small, I. (2020). Large scale genome skimming from herbarium material for accurate plant identification and phylogenomics. *Plant Methods*, *16*, 1.

Nguyen, N.-P., Mirarab, S., Liu, B., Pop, M., & Warnow, T. (2014). TIPP: taxonomic identification and phylogenetic profiling. *Bioinformatics* , *30*(24), 3548–3555.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1), 132.

Orlando, L., Gilbert, M. T. P., & Willerslev, E. (2015). Reconstructing ancient genomes and epigenomes. *Nature Reviews. Genetics*, *16*(7), 395–408.

Pääbo, S. (1989). Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proceedings of the National Academy of Sciences of the United*

*States of America*, *86*(6), 1939–1943.
Rachtman, E., Balaban, M., Bafna, V., & Mirarab, S. (2020). The impact of contaminants on the accuracy of genome skimming and the effectiveness of exclusion read filters. *Molecular Ecology Resources*. doi: 10.1111/1755-0998.13135
Ratnasingham, S., & Hebert, P. D. N. (2007). bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, *7*(3), 355–364.
Riaz, T., Shehzad, W., Viari, A., Pompanon, F., Taberlet, P., & Coissac, E. (2011). ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research*, *39*(21), e145.
Rubinoff, D., & Holland, B. S. (2005). Between two extremes: mitochondrial DNA is neither the panacea nor the nemesis of phylogenetic and taxonomic inference. *Systematic Biology*, *54*(6), 952–961.
Sarmashghi, S., Bohmann, K., P. Gilbert, M. T., Bafna, V., & Mirarab, S. (2019). Skmer: assembly-free and alignment-free sample identification using genome skims. *Genome Biology*, *20*(1), 34.
Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., … Fungal Barcoding Consortium Author List. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(16), 6241–6246.
Shearer, T. L., & Coffroth, M. A. (2008). DNA BARCODING: Barcoding corals: limited by interspecific divergence, not intraspecific variation. *Molecular Ecology Resources*, *8*(2), 247–255.
Sinha, R., Stanley, G., Gulati, G. S., Ezran, C., Travaglini, K. J., Wei, E., … Weissman, I. L. (2017). Index Switching Causes “Spreading-Of-Signal” Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing (p. 125724). doi: 10.1101/125724
Sohn, J.-I., & Nam, J.-W. (2018). The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics*, *19*(1), 23–40.
Song, K., Ren, J., Zhai, Z., Liu, X., Deng, M., & Sun, F. (2013). Alignment-free sequence comparison based on next-generation sequencing reads. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *20*(2), 64–79.
Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*(8), 2045–2050.
Tang, K., Ren, J., & Sun, F. (2019). Afann: bias adjustment for alignment-free sequence comparison based on sequencing data using neural network regression. *Genome Biology*, *20*(1), 266.
Taylor, P. G. (1996). Reproducibility of ancient DNA sequences from extinct Pleistocene fauna. *Molecular Biology and Evolution*, *13*(1), 283–285.
Troll, C. J., Kapp, J., Rao, V., Harkins, K. M., Cole, C., Naughton, C., … Green, R. E. (2019). A ligation-based single-stranded library preparation method to analyze cell-free DNA and synthetic oligos. *BMC Genomics*, *20*(1), 1023.
Vences, M., Thomas, M., van der Meijden, A., Chiari, Y., & Vieites, D. R. (2005). Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. *Frontiers in Zoology*, *2*(1), 5.
Vinga, S., & Almeida, J. (2003). Alignment-free sequence comparison-a review. *Bioinformatics* , *19*(4), 513–523.
Wiemers, M., & Fiedler, K. (2007). Does the DNA barcoding gap exist? - a case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in Zoology*, *4*, 8.
Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 257.
Zeng, C.-X., Hollingsworth, P. M., Yang, J., He, Z.-S., Zhang, Z.-R., Li, D.-Z., & Yang, J.-B. (2018). Genome skimming herbarium specimens for DNA barcoding and phylogenomics. *Plant Methods*, Vol. 14. doi: 10.1186/s13007-018-0300-0
Zielezinski, A., Girgis, H. Z., Bernard, G., Leimeister, C.-A., Tang, K., Dencker, T., …

Karlowski, W. M. (2019). Benchmarking of alignment-free sequence comparison methods. *Genome Biology*, *20*(1), 144.

**Author Contributions**

KB, SM, VB and MTPG conceived of and co-wrote this opinion equally.

**ORCID**

Kristine Bohmann: 0000-0001-7907-064X
Siavash Mirarab: 0000-0001-5410-1518
Vineet Bafna: 0000-0002-5810-6241
M. Thomas P. Gilbert: 0000-0002-5805-7195

652 **Table 1.** Overview of sample collection, laboratory and sequence processing steps and of
653 applications of DNA-based sample identification methods.
654

| | | **Traditional PCR-based barcoding** | | **Genome skimming* using next generation sequencing** | | **Earth Biogeome Project**** |
|---|---|---|---|---|---|---|
| | | **Sanger sequencing** | **Next generation sequencing** | **Organelle assembly** | **k-mers** | **Whole genome assembly** |
| **Sample collection** | Sampling efforts | Same | Same | Same | Same | Same |
| | Voucher specimen | Same | Same | Same | Same | Same |
| **Lab** | Extraction | Standard | Standard | Standard | Standard | High Molecular Weight |
| | PCR of marker region | Yes | Yes | No | No | No |
| | Library build | No | Yes | Yes | Yes | Yes Multiple types |
| **Sequence read processing** | Initial trimming of sequence reads | Yes (manual) | Yes | Yes | Yes | Yes |
| | Quality check of barcode sequence | Yes (manual) | Yes | Yes | No | Yes |
| | Creating k-mer profile | No | No | No | Yes | No |
| | Assembly of organelle genome | No | No | Yes | Optional | Yes |
| | Assembly of whole genomes | No | No | No | No | Yes |
| **Applications** | Identification at taxonomic species-level | Sometimes | Sometimes | Yes | Yes | Yes |
| | Taxonomic identification of simple samples | Yes | Yes | Yes | Yes | Yes |
| | Taxonomic reconstruction of complex samples | Yes | Yes | Yes unless contains very closely related taxa | Perhaps - remains to be fully explored | No |
| | Population level resolution | Rarely - requires population structure and | Rarely - requires population structure | Sometimes - if characterised by unique | Perhaps - to be fully explored | Yes if sufficient population structure |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | high genetic divergence between populations | and high genetic divergence between populations | organelle haplotypes | | exists |
| | Discerning individual level information | No | No | No | Perhaps | Yes |

*Requires ca. 1 gbp of shotgun sequencing (Coissac et al., 2016). **If funding can be secured, the EBP aims to generate chromosome level genome assemblies for all known Eukaryote species (Lewin et al., 2018).

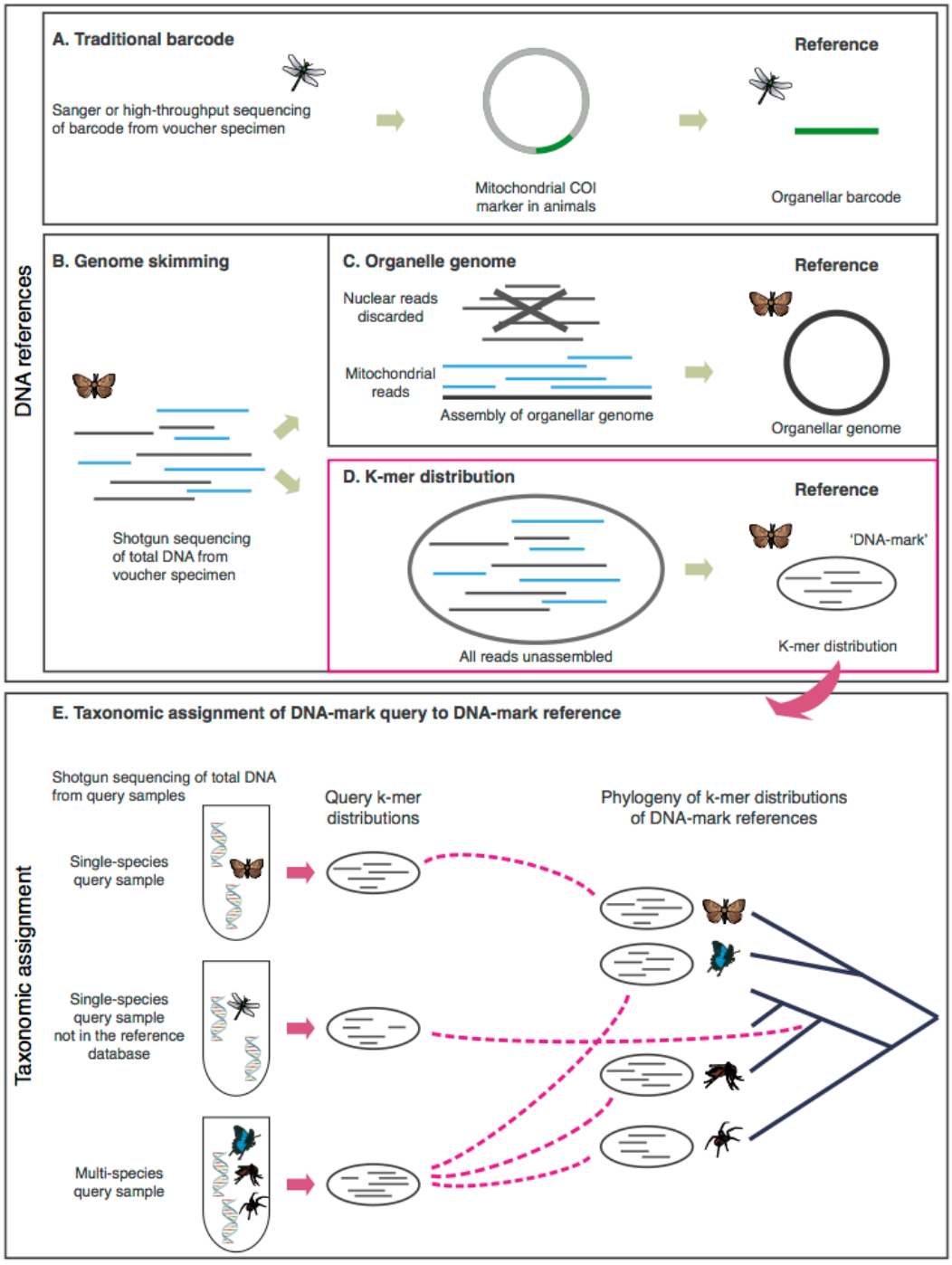


**Figure 1.** Methods to assign a genetic identity to voucher and query samples. (A) Traditional approaches are based on PCR amplification of barcode loci. (B) Increasingly genome-skimming is used to bioinformatically mine the © barcode loci or whole organelle genomes from shotgun sequenced data. (D) We advocate that the remaining data could be used to assign a k-mer profile to the specimen, (E) ultimately enhancing the resolution to which it can be identified (E).

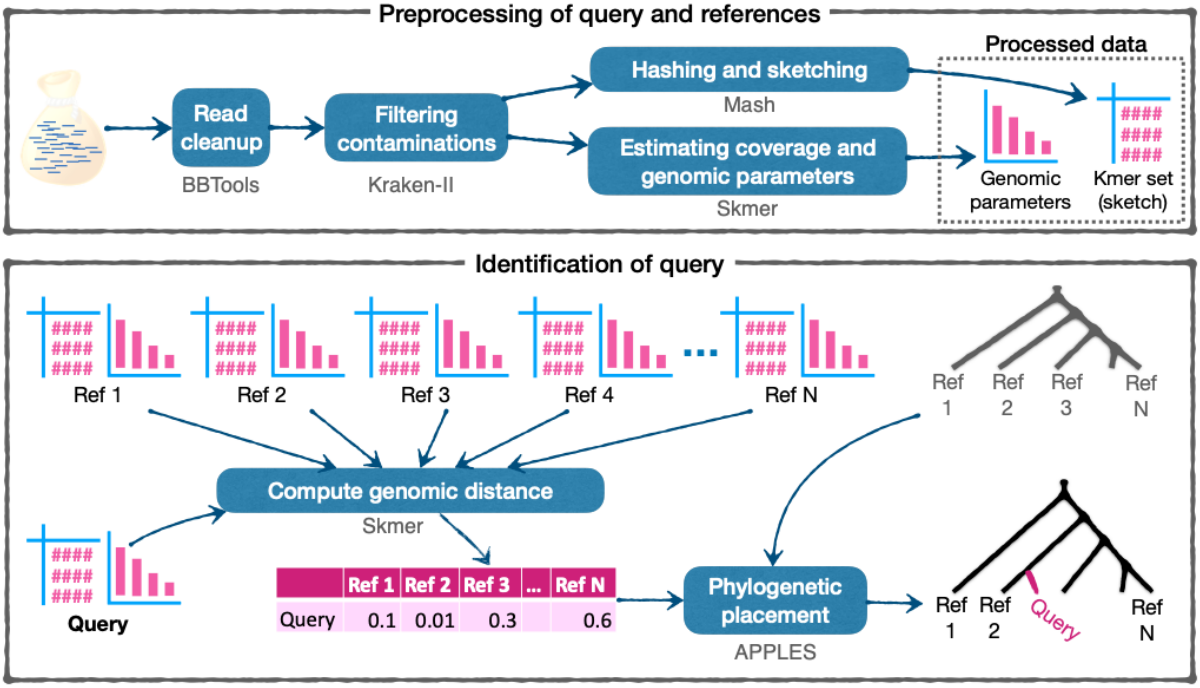


**Figure 2.** Overview of the DNA-mark pipeline. Computational steps are shown in blue boxes, and one example tool that can be used in each step is shown below each box. For each set of reads (whether representing the voucher or the query), the sample has to be first preprocessed in several stages. First, reads are cleaned up to remove adapters, deduplicate reads, and merge paired-end reads. Then, extragenic reads need to be filtered out, typically by matching each read against a database of potential contaminants. The remaining reads need to be represented as k-mers; the set of k-mers need to be hashed and sketched for efficient storage and fast processing. Also, the coverage of the genome skim and properties of the underlying genome (e.g., its size and repeat structure) need to be estimated. Thus, the preprocessing (which needs to happen only once) generates both the k-mer set and the genomic paramters, which are sufficient for sample identification. To identify a new query sample, we need to first compute its distance to the set of reference genome skims. The query can be assigned to the reference with the smallest distance. Alternatively, the query can be placed on a reference phylogenetic tree (which can be computed from the genome skims or can be retrieved from any other source).

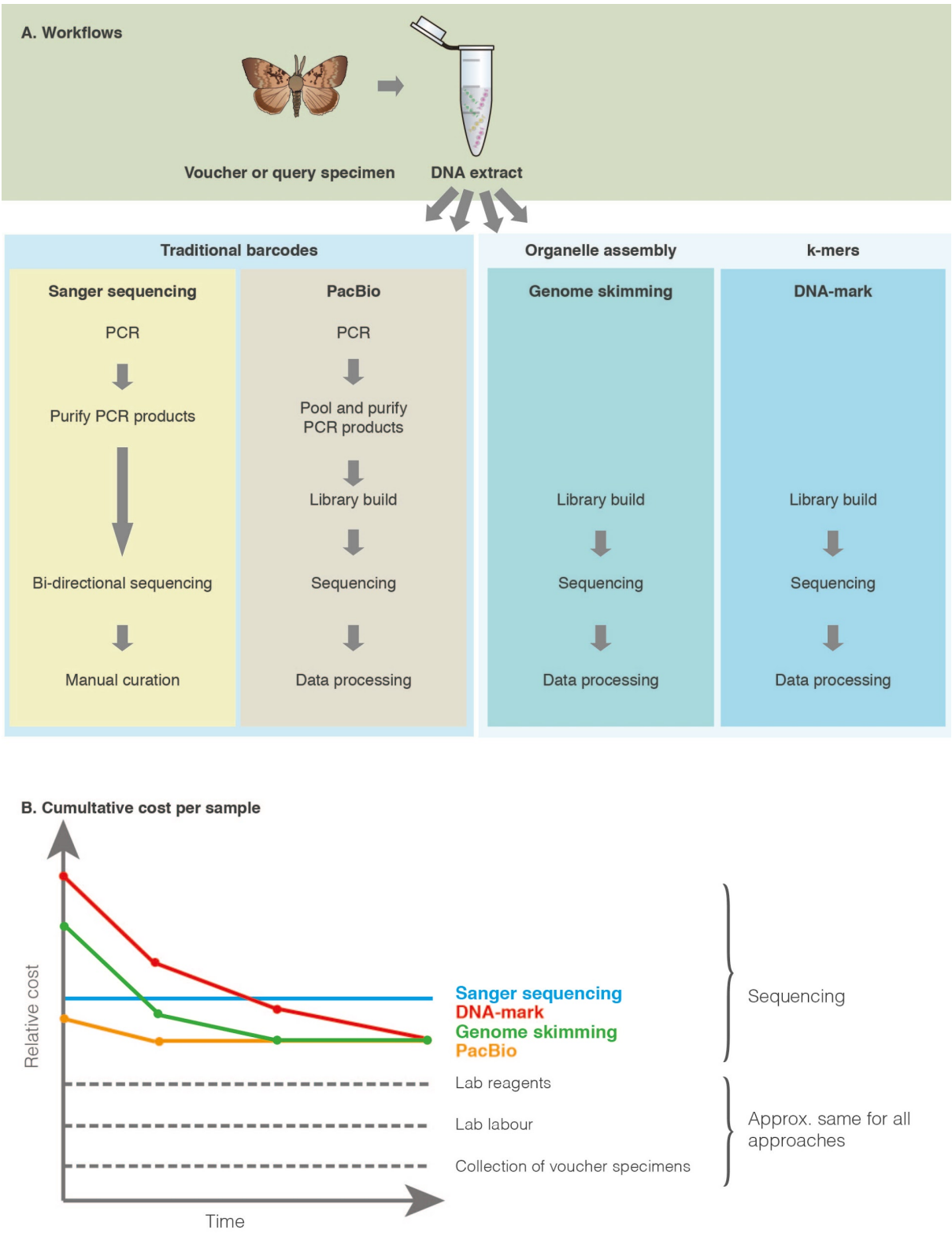


**Figure 3.** (A) Simplified description of the workflow process for generating different types of data that could be used for taxonomic identification. (B) Illustrative example showing that while the cost of the different sequencing techniques is rapidly converging as such methodologies become increasingly economic, the underlying costs of sample collection, vouchering, DNA extraction etc. remains constant. We argue this supports the rationale for exploiting genome skims fully as a tool to complement traditional barcoding.