1 Introduction and Goals

Anthropogenic pressure and other natural causes have resulted in severe disruption of the global ecosystems in recent years, including climate change with extreme weather events, change (loss) of biodiversity, and invasion of non-native flora and fauna. The deforestation of rainforests and degradation of natural habitats is happening faster than efforts to study and understand the impact of these environmental insults. Beyond undesired changes, recent years have also seen an increase in experimental genetics experiments that could radically change the population distribution in an environment. For example, the Target Malaria project is a $75M effort using CRISPR gene drives to genetically modify, and eliminate mosquitoes (*Anopheles*), and could be deployed in Africa within 2 years[1], but its impact on the local biogeography is completely unknown. Similar gene drives are being organized to eliminate Avian malarial parasite carrying mosquitoes from Hawaii[2], and Rodents from New Zealand[3].

The ability to quickly and inexpensively sample the taxonomic diversity in an environment in real time is critical in this era of rapid climate and biodiversity changes, and for the ethical conduct of these directed evolution experiments in nature[4]. Indeed, such methods are specifically important for organisms such as arthropods[5] and small plants, which are among the most abundant and diverse non-microbial organisms on Earth but lack large-scale descriptions of biogeography and richness. However, large cataloging surveys remain sparse. At least some of this discrepancy can be attributed to the high cost of sorting and identifying samples from large sampling collections.

The molecular technique of choice for measuring biodiversity is (meta)barcoding[6–8], which involves DNA sequencing of taxonomically informative and group-specific markers (e.g., mtDNA COI[6,9] and 12S/16S[10] for animals, plastid genes like trnl and matK[11] for plants, and ITS[12] for fungi) that are variable enough for taxonomic identification, but have flanking regions that are sufficiently conserved to allow for PCR amplification using universal primers. Barcoding is used to taxonomically identify species or in the case of meta-barcoding to deconstruct the species composition of complex samples. Accurate barcoding crucially depends on the coverage of the reference database and the method used to search the database[8]. Computational methods have been developed for finding the closest match in a reference dataset of markers (e.g., TaxI[13]), and for placement of a query into existing marker trees[14,15]. The reference databases for these studies consist of traditional barcode regions, such as COI in the Barcode of Life Data System[16].

The traditional barcoding pipeline has several drawbacks. PCR for marker gene amplification requires relatively high-quality DNA and thus cannot be applied to samples in which the DNA is heavily fragmented. Moreover, since barcodes are short regions, their phylogenetic signal is limited[17]. For example, 896 of the 4,174 species of wasps could not be distinguished from other species using COI barcodes[18]. While low costs have kept PCR-based pipelines attractive, falling costs of shotgun sequencing have now made it possible to shotgun sequence 1-2Gb of a reference specimen sample for $\leq$ $50, inclusive of sample preparation and labor costs. Therefore, researchers have proposed low-pass sequencing (*genome-skimming*) as a viable approach to barcoding[19]. These approaches identify chloroplast/mtDNA marker genes in genomes by mapping to a reference library

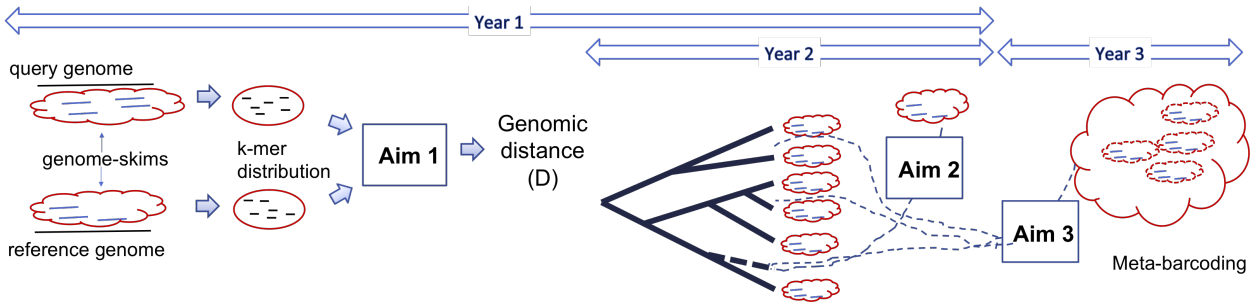


Figure 1: **Overview of the scientific aims.**

or *de novo* assembly[20]. Large databases (e.g., through PhyloAlps[21] and NorBol[22] projects) are using genome-skimming to mine marker genes. Our Danish collaborators (**letter: Gilbert**) have developed a similar genome-skimming project, DNAmark[23]. However, they realized that *accurate* assembly of organelle and marker regions from genome-skims is non-trivial, and the approach discards the vast majority of the reads and is thus wasteful. Therefore, they posed an interesting methodological question: *can the unassembled data be used "as-is" for barcoding, saving on the labor-intensive task of mtDNA assembly and using all available genomic information?*

Motivated by this question, we consider the alternative approach where low coverage genome-skims are used both to create the reference library and the query dataset. We propose the development of computational and statistical tools to address the following key questions (Figure 1):

Aim 1: *Given genome-skims of two organisms, estimate the hamming distance and use the distance to search a library of genome-skims.*

Aim 2: *Use genome-skims for phylogenetic reconstruction, including placement of a query skim onto a taxonomy/phylogeny of the reference organisms.*

Aim 3: (exploratory) *Given a 'meta-barcoding' query (genome-skims of a mix of organisms), identify the constituent organisms and their relative sequence abundance in the samples.*

As this is a *short*, exploratory, proposal budgeting for four graduate student years, we will address Aims 1 (Year 1) and Aim 2 (Years 1 and 2) thoroughly and will perform exploratory empirical and theoretical work on Aim 3 (Year 3) in preparation for a larger, collaborative proposal.

2 Broader Impacts

The proposed activities have broad impact because some of the poorest and under-developed places have most of the remaining bio-diversity in the world. The loss of this bio-diversity can have severe and lasting impact on all people, including in the U.S.A. However, even with dramatically falling costs, genomic sequencing technologies have not penetrated these communities. We will use the grant period to reach out to scientists and build collaborations that start to catalog genome-skims, similar to barcoding efforts such as the Consortium for the Barcode of Life. We will collaborate with Dr. Thomas Gilbert, leader of the center for GeoGenetics at the Natural History Museum of Denmark on developing metabarcoding. We will also initiate a collaboration with Dr. Ethan Bier (UC San Diego, and Tata Institute of Genomics and society) who is developing CRISPR

based technologies to change *Anopheles* population composition in some locations in India with the goal of eliminating malaria. We will use these initial studies to develop inexpensive protocols that will allow anyone in the world to gather genome-skims that can be analyzed by our publicly available tools. Beyond these, we will make outreach efforts: i) train STAR undergraduate students, recruited from both biology and CS; ii) the annual Evolution meeting includes an outreach program promoting the understanding of computational technique; we will participate in Year 3 and use barcoding using genome-skims as an example to promote computational understanding among biology undergraduates; iv) we will hold local seminars to increase awareness of the impact of rapid environmental changes on ecology and the role of computational genomics in alleviating them.

3 Aim 1: Pairwise distance for genome-skims

Our approach to assembly-free and alignment-free barcoding relies on computing a *distance* between a pair of genome-skims. Distances can be used to search a reference library for the closest available match to a query (our main use-case), for phylogenetic analyses (see Aim 2), and for analyzing mixed samples (Aim 3). To minimize cost and human effort, in our genome-skimming pipeline, the reference library and the query are both provided as genome-skims. Therefore, accurate distances need to be estimated despite low and varied coverage, sequencing error, and varying genome length.

Alignment-free measures of distance between sequences have been widely studied[24–27]. Existing methods assume high sequence coverage (an exception from the phylogenetics is discussed in Aim 2). There are two categories of alignment-free methods[25], and the dominant approach is to decompose all reads into fixed length oligomers (denoted *k-mers* with length k). In recent years, researchers have developed fast and accurate tools for computing the k-mer frequencies (JellyFish[28]), for classifying metagenomic reads (Kraken[29]), and for distance computation (Mash[30]).

Preliminary methods: Skmer. We have prototyped a method called Skmer to enable fast and accurate comparison of two genome-skims to compute their hamming distance. Our work builds on Mash (Ondov *et al*[30]) but specifically enables accurate measurement of distance even with low and varied coverage and sequencing errors. Consider an idealized model where two genomes are the outcome of a random process that copies a genome and introduces substitutions at each position with fixed probability d. The hamming distance D between the two genomes has expected value d, which we seek to estimate. The *Jaccard index* J is a similarity measure between two sets (e.g. k-mer collections) defined as the size of their intersection divided by the size of their union. We use the estimate

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

in contrast to the slightly less accurate Mash estimate, $D \approx \frac{1}{k}\ln\left(\frac{J+1}{2J}\right)$. Equation 1 implicitly assumes that each k-mer is sampled at least once with very high probability. This assumption is violated for genome skims in consequential ways for low coverage. As a simple example, suppose that a 21-mer ($k = 21$) is sampled with probability 0.5. Then, even for identical genomes, $J = \frac{1}{3}$, resulting in an estimate of $D \approx 0.032$. We propose to refine these estimates of J to handle genome-skims despite low and uneven coverage, sequencing error, and varying genome-lengths.

We start with a known coverage (but will show how to estimate it below). Each genome of length L is independently sequenced using randomly distributed short reads of length ℓ at coverages c_1 and c_2 to produce two genome-skims with k-mer sets A_1 and A_2, with k large enough to be unique with high probability. The probability of covering each k-mer can be approximated as $\eta_i = 1-e^{-\lambda_i}$ where $\lambda_i = c_i(1 - k/\ell)$. Modeling the sampling of k-mers as independent Bernoulli trials, $|A_i|$ becomes binomially distributed with parameters η_i and L, and $W = |A_1 \cap A_2|$ is binomially distributed with parameters $\eta_1\eta_2(1 - d)^k$ and L. Moreover, $U = |A_1 \cup A_2|$ can be modeled approximately as a Gaussian with mean $(\eta_1 + \eta_2 - \eta_1\eta_2(1 - d)^k)L$. Treating η_1 and η_2 as known and dividing $\frac{W}{L}$ by $\frac{U}{L}$ gives us: $J = \frac{W}{U} = \frac{\eta_1\eta_2(1-D)^k}{\eta_1+\eta_2-\eta_1\eta_2(1-D)^k}$; thus, $D = 1 - \left(\frac{(\eta_1+\eta_2)}{\eta_1\eta_2}\frac{J}{(1+J)}\right)^{\frac{1}{k}}$. Let ϵ denote the base-miscall rate. For large k and small ϵ, the probability that an erroneous k-mer produces a non-novel k-mer is negligible. Therefore, each miscall reduces the number of shared k-mers and increases the total number of observed k-mers. The probability that a k-mers is covered by at least one error-free read is approximately $\eta_i = 1 - e^{-\lambda_i(1-\epsilon)^k}$. Adding up the number of error-free and erroneous k-mers, the total number of k-mers observed from each genome can again be approximately modeled as a Gaussian with mean $\zeta_i L$ where $\zeta_i = \eta_i + \lambda_i(1 - (1 - \epsilon)^k)$. To estimate D, we solve for it using the refined estimate $J = \frac{\eta_1\eta_2(1-D)^k}{\zeta_1+\zeta_2-\eta_1\eta_2(1-D)^k}$. For high coverage (5X), we will use the alternative approach of handling errors by removing singleton k-mers, but we omit details here.

So far, we have assumed identical lengths. On real genomes, with different lengths, one would ideally want to penalize for genome length differences. A rigorous modeling of evolutionary distance for genomes of different length require sophisticated models of gene gain, duplication, and loss (see Aim 2). A simple heuristic, also used by Ondov *et al.*[30], is to use the mean genome length in computing the union (denominator). This ensures that the estimated distance increases as genome lengths becomes successively more different. This leads us to our final estimate of d given by:

$$D = 1 - \left(\frac{2(\zeta_1 L_1 + \zeta_2 L_2)J}{\eta_1\eta_2(L_1 + L_2)(1 + J)}\right)^{1/k} \tag{2}$$

So far we have assumed a perfect knowledge of sequencing depth, but on real genome-skims, we need to estimate the coverage in order to apply our distance correction. First, note that the number of reads covering a k-mer is a Poisson r.v. with mean λ (i.e., the k-mer coverage). Let M denote the total number of k-mers of length k in the genome, and M_i count the number of k-mers covered by i reads. Given M, and for $i \geq 0$, $\mathbb{E}[M_i] = M\frac{\lambda^i}{i!}e^{-\lambda}$. We count the number of times that each k-mer is seen using JellyFish[28]. However, since in a genome-skim, large parts of the genome may not be covered, both M and M_0 are unknown. To deal with this issue, we could take the ratio of consecutive counts to get a series of estimates of λ as $\tilde{\lambda}_i = \frac{M_{i+1}}{M_i}(i + 1)$ for $i = 1, 2, \ldots$. In practice, sequencing errors change the frequency of k-mers which has to be considered when estimating the coverage. Modeling the error further complicates the estimator of the coverage and gives: $\lambda_i = \frac{1}{(1-\epsilon)^k}\frac{\hat{M}_{i+1}}{\hat{M}_i}(i+1)$ (λ_1 has a different formula, which we omit). We will use some heuristics to choose one of the λ_i values as the final coverage estimate (details omitted).

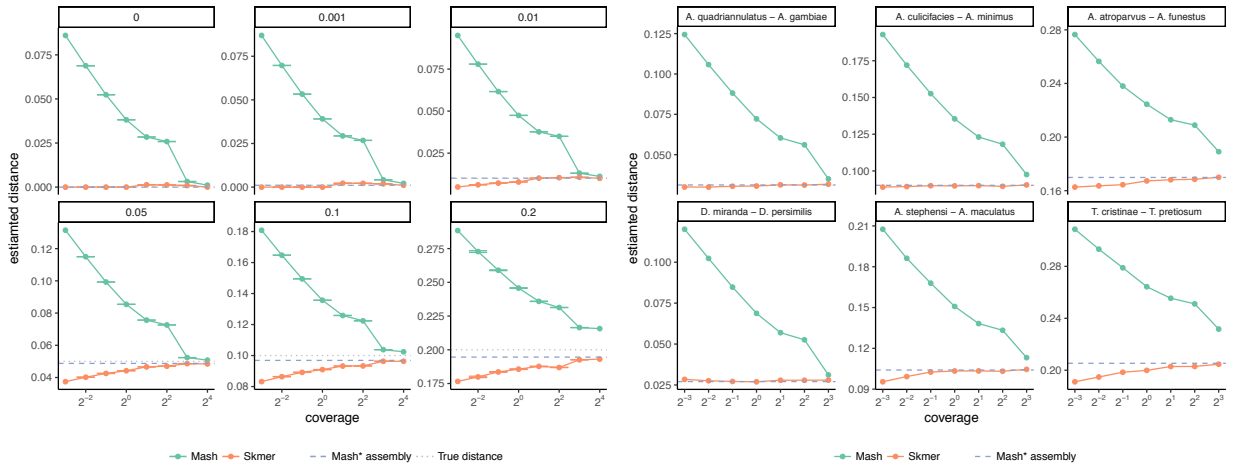


Figure 2: **Mash v.s. Skmer.** Hamming distance between the genome-skims simulated using ART with read length $\ell = 100$, constant base error rate $\epsilon = 0.01$, and varying range of coverage (x-axis). a) Six pairs of genomes simulated by applying substitutions to the assembly of *C. vestalis*. The mean and standard error of distances are shown over 10 repeats. Repeating the process with the *Drosophila melanogaster* genome as the base genome produces similar results. b) Six pairs of insects genomes at different hamming distances. The pairs placed in the top (bottom) row consist of species with similar (different) genome lengths. Ground truth (dashed line) is taken to be Mash run on the assembly (but without the $(1-D)^k \simeq e^{-kD}$ approximation).

Preliminary Results. We compared the accuracy of Mash and a Skmer prototype in estimating the hamming distance between pairs of simulated and real genomes. With high coverage, Mash estimates had high accuracy (Fig. 2a), except for highly diverged genomes (d=0.2). However, the accuracy of Mash quickly degraded when the coverage was reduced to 4X or less. In contrast, Skmer had high accuracy for almost all cases even when the coverage is reduced to $\frac{1}{8}$X. For example, with the true distance set to 0.05 in simulations, Mash estimated the distance as 0.085 with 1X coverage (an overestimation by 70%) while Skmer estimated distance as 0.044 (an underestimation by 12%). The only case where Skmer error was high was with coverage below 1X and $d = 0.2$ (Fig. 2a).

We also tested on several pairs of insect genomes that covered a wide range of mutation distances and genome lengths, and note that natural divergence involved multiple forms of mutations. We used the distance estimated by Mash on *assembled genomes* as the true (or, *optimal*) distance. For all pairs of insects (Fig. 2b), Mash had high errors for coverage below 8X while the Skmer estimates remained extremely close to the optimal distances. For example, the optimal distance between *A. stephensi* (length: 196Mbp) and *A. maculatus* (length: 132Mbp) was 0.104 and it closely matched the Skmer estimate of 0.103 (1% underestimation) with only $\frac{1}{2}$X coverage. The Mash estimate was 0.168 (60% overestimation). Interestingly, on real data, Skmer seemed to have even less error than simulated genomes. We observed similar patterns on 47 avian genomes[31] (data not shown).

Challenges and proposed goals. We build upon the encouraging preliminary data to propose new developments, so as to make Skmer the tool of choice for barcoding using genome-skims.

A1.1 Population-level. Skmer can estimate hamming distances as small as 0.01 accurately from genome-skims with even lower than 1X coverage. What does a distance of 0.01 mean? The answer

will depend on the organisms of interest. For example, two eagles species of the same genus (*H. albicilla* and *H. leucocephalus*) have $D \approx 0.003$ but two Malaria mosquitoes of the same species complex (*A. gambiae* and *A. coluzzii*) have $D \approx 0.018$. Broadly speaking, for eukaryotes, detecting distances in the 10^{-2} order is often enough to distinguish between species. On the other hand, distances in the 10^{-3} order often differentiate between populations or very similar species.

While Skmer has low absolute error across the board, its proportional error can be high if coverage is very low (e.g., 1X or less) *and* the true distance is in the order of 0.001 or lower. This raises several important questions, which we will address. i) Genomes are now available for a large numbers of closely related species and populations (e.g., Darwin finches, flycatchers, rice, mice, and primates/humans). Using these available genomes (and other unpublished genomes available to us by our Danish and UCSD collaborators), we will carefully study the level of resolution that can be obtained using Skmer for various levels of coverage and for various taxonomic groups. We will quantify the level of coverage required with the current approach to distinguish two very closely related species and two populations of the same species. ii) Several parameters of Skmer are currently fixed but can be adjusted for better measuring low distances. For example, we fix $k = 31$. We have evidence that higher k increases resolution for low distances. We can automatically detect the optimal k by finding the threshold at which increasing k does not substantially increase the number of unique k-mers. Thus, we will build a version of Skmer tuned for very low distances.

A1.2 Estimating ϵ. Skmer currently takes the sequencing error rate (ϵ) as input. We have tested the impact of having an incorrect estimate of ϵ and using uneven distributions of error to emulate the Illumina HiSeq2000 platform. Skmer seems robust to mis-specifications of the sequencing error model when the error is underestimated (data not shown). However, over-estimation of error can reduce the accuracy. Fortunately, our analytical framework presents natural ways for computing ϵ, even when coverage c is unknown. Let $\hat{M}_i$ denote the count of k-mers observed i times in the presence of error in a genome with M k-mers. Let $\lambda = c(1 - \frac{k}{\ell})$ and let $\rho = \lambda(1-\epsilon)^k$. We can show that

$$\mathbb{E}[\hat{M}_i] = \begin{cases} M \frac{\rho^i}{i!} e^{-\rho} & i \geq 2 \\ M\left(\rho e^{-\rho} + \lambda - \rho\right) & i = 1 \end{cases} \tag{3}$$

In regimes with sufficient counts of higher multiplicity k-mers, we can use Eqn. 3 and observed values of $\hat{M}_i$ for $i \geq 2$ to estimate ρ then λ using $\hat{M}_1$ (i.e., singleton k-mers). Subsequently, we can estimate c and ϵ. We will improve upon these estimates using k-mers that are linked by reads. While the conceptual framework is clear, many practical choices need to be made. Importantly, these will enable us handle varying error rates of sequencing in the two genomes.

A1.3 Contamination. A major issue that we have so far ignored is the possibility that external DNA originating from parasites, diet, fungi, bacteria, and human contamination may be mixed with a supposedly single-species genome-skim. The presence of such external DNA can negatively impact the estimated distance. We will perform extensive tests to quantify the robustness of Skmer to external DNA by manually injecting fungal, bacterial, and human contamination. Moreover, so far, we have tested Skmer only on simulated genome-skims. We plan to collaborate with our DNAmark colleagues to test Skmer on real genome-skims from known species.

To improve Skmer in the presence of external DNA, several approaches can be considered. For genome-skims in the reference library, we can simply search a database using tools like BLAST[32], USearch[33], or Bowtie[34], or faster k-mer based methods like Kraken[29] to find and eliminate bacterial or fungal contamination. Note that the cost of preprocessing will be amortized over many searches.

For query sequences, even a fast classification tool like Kraken run on all reads *may* prove too slow. But note that we don't need identities of contaminants. *Given a query genome-skim, and a large 'contaminant database' of genome skims from prokaryotes and fungi, we will develop super-fast algorithms for filtering contamination.* We will test membership queries using Bloom filters representing the union of k-mers in all contaminant sequences. Consider a bit-vector B of size m, where each of n k-mers from a contaminant set is hashed using h bits. Then, the probability of a non-contaminant k-mer being hashed to B by chance, $p \leq (1 - e^{-\frac{hn}{m}})^h$. We will denote a read as a contaminant if a fraction θ of its k-mers are hashed to B. For small p, the probability of misclassifying a read as a contaminant is bounded using Chernoff's bound to be $\leq \exp(-\ell(\theta - p)^2)$. Adjusting θ, we can also detect contamination from species *close to* but not identical to those in our database. Thus, Bloom filter may provide a super-fast way of eliminating all contaminant reads.

For even higher speeds, *we will develop algorithmic strategies to compute the true distance without eliminating contaminants.* For a query genome-skim with a mix of k-mers from the real species (A) and those from contaminants (C), and reference skim B, let J' be the Jaccard computed without filtering. We can estimate the fraction of contamination $w = \frac{|A|}{|A \cup C|} \approx \frac{|A|}{|A|+|C|}$ by running Kraken or the Bloom filter on a small subset of reads. We expect contaminants to have very low similarity to the main query species (e.g., fungal DNA in an insect query). Thus, their contribution to the shared k-mer set is negligible (i.e., $|A \cap C| \approx |B \cap C| \approx 0$). Note also that $|A \cup B \cup C|$ and $|A \cup C|$ can be easily computed using JellyFish. Then, we can derive the corrected J:

$$J = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|/|A \cup B \cup C|}{|A \cup B|/|A \cup B \cup C|} = \frac{J'}{1 - \frac{|C|}{|A \cup B \cup C|}} = \frac{J'}{1 - (1-w)\frac{|A \cup C|}{|A \cup B \cup C|}} .$$

Thus, we can use Equation 2 with this contamination-corrected estimate of J. This way of handling contamination is essentially a simpler case of meta-barcoding (Aim 3) because we assume *a priori* that contaminants are distant from the main query species and do not seek their identity.

A1.4 Genome length and repeats. To deal with varied sequence length, we simply used the mean length. We will further study whether varying genome length, abundant repeats, and polyploidy can mislead Skmer by weakening the link between the computed hamming distances and phylogenetic/taxonomical relationship between species. As we will argue (Aim 2), the best way to deal with length differences it to have a model of genome evolution[35]. However, we will also test simpler model-free approaches. For example, eliminating repetitive parts of the genome by removing high frequency k-mers may lead to distances that better reflect the species phylogeny.

A1.5 Software and database development. Skmer is already efficient, thanks to its reliance on efficient Mash[30] and Jellyfish[28]. For example, it took 33 minutes (using 24 cores) to compute distances for $\binom{47}{2}$ pairs of birds. Almost all of that time is spent computing k-mer hash values in Mash and computing k-mer frequency profiles using JellyFish. Both of these can be precomputed for all the reference species. Moreover, searching a query against a library currently requires

computing the Jaccard index versus every species in the library. However, a tree-based approach (see Aim 2) or clustering of reference species could be used to narrow down the search. Many algorithms for this reduced search can be imagined, a topic we will further explore.

To facilitate its adoption in practice, we will develop an improved version of Skmer that is ready to be widely deployed. In our system, users will use a simple command to query a genome-skim, represented simply as a `fastq`/`fasta` file, against a library, also a collection of `fastq` files and associated names. We will make it trivial to add new genomes to the reference set by simply adding a file containing the genome-skim reads to the system. A simple preprocessing step can then be performed to precompute sketches and k-mer profiles. We will thoroughly test the software created and will release preliminary reference databases (although creation of comprehensive reference databases will be beyond our scope). The problem of searching the library is embarrassingly parallel, and we will exploit the parallelism in our released pipeline.

4 Aim 2: Phylogenetic analysis of genome-skim

Phylogenetics can greatly help (meta)barcoding using genome-skims. Take the problem of searching a reference library to find the identify of an unknown sample query. Often, an exact match to the query species is not present in the library. Then, we would like to find which broad group of species the taxon query belongs to (e.g., its genus). An obvious approach is to find the closest match, judged by a measure of distance, and extrapolate from the closest match. Our Skmer method has great accuracy compared to existing methods (Mash and AAF[36], described below) in our preliminary leave-p-out experiments of finding the closest match (Fig 3). But, several issues remain.

First, using the closest match may be misleading when rates of evolution are not constant. For example, in Figure 4a, if q is removed and queried against the remaining species, its closest match will be a, which belongs to a phylogenetically distinct group. A second problem is that the closest match by the the hamming distance may not be the closest match by phylogenetic distance (defined below). Moreover, a phylogenetic framework can help reduce noise in distances. If all the estimated distances were perfectly correct, they would fit uniquely to a tree, a property called *additivity*[37]. But estimated distances often don't fit a unique tree, and the process of fitting a phylogenetic tree to a distance matrix can be considered as noise removal in estimated pairwise distances. Finally, using a phylogeny will free us from the need to use arbitrary groupings of species into taxonomies and enables us to provide answers that reflect the full complexity of the evolutionary history.

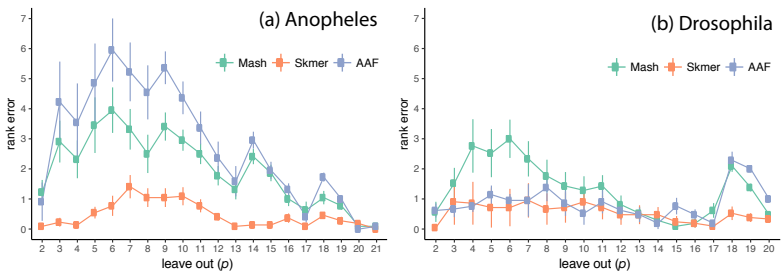
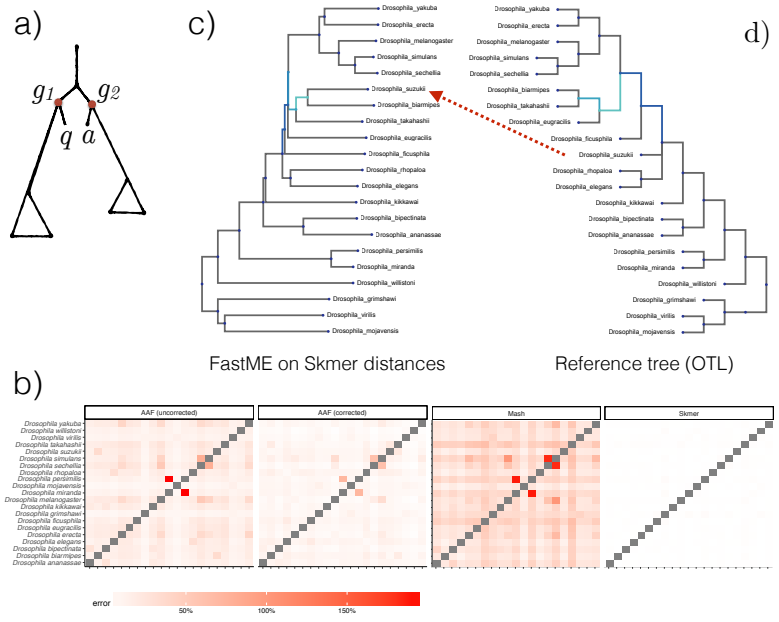


Figure 3: **Leave-p-out search.** For each species, a 0.1–1Gb genome-skim is randomly simulated, its p (x-axis) closest relatives by distances computed from full assemblies are excluded, and the closest remaining match is noted. Each method is used to rank the remaining species. The rank of the true best match minus one is measured as the rank error (y-axis).



d)

function PLACE($t, \mathbf{D}, q$)
  $closest \leftarrow \text{argmin}_{i \neq q}\, \mathbf{D}[i, q]$
  $reroot(t, closest)$
  $u \leftarrow child(closest)$
  while *true* **do**
    if $isLeaf(u)$ **then**
      $n \leftarrow Parent(u)$
      return u
    $(l, r) \leftarrow (left(u), right(u))$
    $(lc, rc) \leftarrow (arbL(lc), arbL(rc))$
    $s \leftarrow fourPoint(\mathbf{D}, q, closest, lc, rc)$
    if $s = closest$ **then**
      return u
    else if $s = lc$ **then**
      $u \leftarrow l$
    else if $betterSide = rc$ **then**
      $u \leftarrow r$
function FOURPOINT($\mathbf{D}, q, a, b, c$)
  $as \leftarrow \mathbf{D}[q, a] + \mathbf{D}[c, b] - \mathbf{D}[q, c] - \mathbf{D}[a, b]$
  $bs \leftarrow \mathbf{D}[q, b] + \mathbf{D}[a, c] - \mathbf{D}[q, a] - \mathbf{D}[b, c]$
  $cs \leftarrow \mathbf{D}[q, c] + \mathbf{D}[b, a] - \mathbf{D}[q, b] - \mathbf{D}[c, a]$
  $m \leftarrow \min(as, bs, cs)$
  return c if $m = c$ else b if $m = b$ else a

Figure 4: **Phylogenetic reconstruction: concepts, results, and algorithms.** **a)** A Felsenstein's zone tree that misleads closest match methods. **b)** Comparing Skmer, Mash, and AAF on all pairs of 21 Drosophila genome-skims (0.1–1Gb of data per genome). Heatmaps show error measured as $|d - d^*|/d^*$ where d is distance computed from the genome-skims and d^* is the distance computed from full genome assemblies. **c)** Results of applying Skmer on Drosophila genome-skims (1Gb each) and using the distance matrix as input to FastME. The Skmer+FastME tree agrees with the current best known tree obtained from Open Tree of Life[38] (OTL); the position of *D. suzukii* is left unresolved in OTL but it is placed as sister to *D. biarmipes* in the Skmer+FastME tree. **d)** Algorithm for distance-based phylogenetic placement of query q to tree t using distance matrix $\mathbf{D}$ using the four point condition. $arbL(x)$: an arbitrary leaf under x.

To make our estimates of distance more useful, we will build a phylogenetic framework around Skmer. This framework will address several problems. A2.1. Compute an asymptotically additive phylogenetic distance for complex substitution models such as Generalized Time Reversible (GTR) from genome-skims. A2.2. Given a set of genome-skims (e.g., the reference library), infer an accurate species phylogeny. A2.3. Incorporate biological causes of *gene tree discordance* in inferring the phylogeny. A2.4. Given a reference tree with genome-skims at the leaves and a query genome-skim, add the query to the reference efficiently and accurately (*phylogenetic placement*).

A2.1. Phylogenetic distance correction

Background. The hamming distance is not additive and thus cannot correspond to phylogenetic distances. However, assuming the simple Jukes-Cantor (JC) model[39], where all DNA letters and all substitutions are equiprobable, the hamming distance d can be phylogenetically "corrected" to produce an additive distance $t = -\frac{3}{4}\ln 1 - \frac{4}{3}d$, measured in units of the expected number of mutations since the LCA. More complex models of sequence evolution also allow the calculation of corrected distances, but not solely based on the hamming distance. For example, the K2P model[40] allows a different rate for transversions and transitions, and to compute distances under this model,

the hamming distance needs to be divided into fraction of transitions and transversions[40]. Under the most general time-reversible Markov model (GTR[41]), distances can be computed given two matrices: $\mathbf{P}$ where the (i, j) element is the column normalized proportion of sites that has letters i and j in the two genomes, and a diagonal matrix $\mathbf{D}$ where the element (i, i) is the portion of letters in both genomes that has letter i. Then, the corrected distance[42] is given by $t = -\operatorname{trace}(\log(\mathbf{P})\mathbf{D})$. An even more general correction is log-det method[43,44]: $t = -\ln\det(\mathbf{F})$ where $\mathbf{F}$ is a non-normalized version of $\mathbf{P}$. Corrected distances converge to additivity as the sequence length increases.

Approach. We intend to compute distances corrected with the GTR model. To do so, we need need substitutions broken down by specific pairs of letters. For $i, j \in \{\text{A,C,G,T}\}$, we need $x_{ij} = x_{ji}$, the number of mutations of the form $i \leftrightarrow j$. However, the paradigm of computing distance by counting k-mers treats all mutations alike. Formally: $d = (x_{\text{AC}} + x_{\text{AG}} + x_{\text{AT}} + x_{\text{CG}} + x_{\text{CT}} + x_{\text{GT}})/L$. We can use the following approach to compute individual x_{ij} for all i, j:

1. Replace G and T with C, and compute distance $d_{\text{A}} = (x_{\text{AC}} + x_{\text{AG}} + x_{\text{AT}})/L$.
2. Replace G and T with A, and compute distance $d_{\text{C}} = (x_{\text{AC}} + x_{\text{CG}} + x_{\text{CT}})/L$.
3. Replace G with T, and compute distance $d_{\text{AC}} = (x_{\text{AC}} + x_{\text{AG}} + x_{\text{AT}} + x_{\text{CG}} + x_{\text{CT}})/L$.

Combining, we get $x_{\text{AC}} = (d_{\text{A}} + d_{\text{C}} - d_{\text{AC}}) \times L$.

A similar procedure can be used to compute all x_{ij}, from which we can easily compute both $\mathbf{P}$ and $\mathbf{D}$ matrices required to compute corrected GTR matrices using $t = -\operatorname{trace}(\log(\mathbf{P})\mathbf{D})$. We refer to the results as the Jaccard-based estimates of the GTR matrix or J-GTR for short.

We will test if J-GTR distances can be computed accurately given genome-skims. These analyses can be done in simulations that follow GTR. Assuming fixed GTR parameters across the genome is unreasonable and therefore, we will test the robustness of the J-GTR distances in simulations that vary GTR parameters across the genome. Using the log-det approach, which also requires x_{ij}'s as input, promises to model variations across the genome. We test if the log-det approach can handle these scenarios better. Several models of rate-across-sites heterogeneity exist and under simplifying assumptions some allow for distance correction. We will also test such models.

Distance correction based on the JC model is a monotonic function and cannot impact the ranking of reference genome-skims with respect to their similarity to a query skim. However, the same is not true for the J-GTR correction; it can rank reference species differently than hamming distance and therefore, can impact results of our main use-case. On simulated and real datasets, we will test if J-GTR correction improves the detection of the closest match to a query sequence.

Challenges. Our method of computing J-GTR distanced reduces the space of possible k-mers to 2^k possibilities instead of 4^k. Therefore, an increased k may be needed. We will study the best choice of k, allowing a different k for steps 1 and 2 versus step 3. Computing all x_{ij} values using our procedure will increase the running time. Since Skmer is very fast, this may not be present a challenge. However, depending on the size of the reference dataset, minimizing computational time may prove important. It is possible that the minHash sketching approach used for computing the Jaccard distance could be changed to compute all d_i and d_{ij} values without actually replacing

letters in genomes. This can perhaps be accomplished by simply using several hash function, each of which ignores some letters. These precomputed sketches will then be used in subsequent steps.

A2.2. Phylogenetic reconstruction.

Background. Given phylogenetically corrected additive or nearly additive distances, phylogenetic reconstruction with high accuracy is doable with several algorithms, including minimum evolution[45], neighbor joining[46], and DCM[47] and efficient implementations exist[48–50]. Moreover, distance-based methods are theoretically as powerful as maximum likelihood[51].

Assembly-free phylogenetics has a rich history, especially using k-mers[36,52–63]. Despite some success stories[64,65], these methods are rarely used in practice, perhaps due to a wide-spread opinion (not shared by all) that alignment-free methods are not as accurate as alignment-based methods. Alignment-free methods are mostly tested on short single-gene datasets and their supposed inferiority to alignment-based methods is mostly concerned with the most difficult phylogenetic problems that include short deep branches. Our use-cases are not greatly impacted by possible shortcomings of alignment-free approaches because we have genome-wide data and because our main objective is not to resolve difficult phylogenetic branches but to accurately barcode new samples.

Most alignment-free methods assume high coverage and ignore sequencing error[27] and therefore are not applicable to genome-skims. For example, a recent result by Allman *et al.*[66] computes the corrected distance between two sequences based on a *k-mer vector* containing frequencies of all possible 4^k possible k-mers; this only works for low k and requires sequence assemblies (but not sequence alignments). The main *assembly-free* method that also seeks to work with low coverage is AAF[67]. It uses the Jaccard index, a translation to hamming (similar to Eq. 1), and a translation based on JC to compute a phylogeny. It then corrects terminal branch lengths to account for moderately low coverages. We have compared our Skmer approach to AAF and have observed that Skmer outperforms AAF dramatically (Figs. 3, 4b). AAF works well at mildly low coverage (e.g., 2X) but not at 1X coverage or lower (data not shown). Moreover, it is slow and hard to use. Finally, AAF only allows *de novo* phylogenetic inference and not placement or database search.

Approach. Our approach first computes *phylogenetic* distances using k-mer profiles (Aim 2.1) and then uses existing distance-based methods for tree inference. We will test neighbor joining[46], minimum evolution[48], and the DCM family of methods[68]. As an example, for 21 Drosophila genomes, we computed distances using Skmer from 1GB genome-skims, corrected distance using the JC transformation, and used FastME to infer a tree; the resulting tree was essentially identical to the best known phylogeny (Fig. 4c).

Challenges. The true phylogeny is never known. We will use the coalescent-based `ms`[69] simulator in addition to the more complex `Evolver`[70,71]. Since both simulators have limitations, we will also use real data. We will compare our reconstructed phylogenies against published genome-wide phylogenies that are believed to be mostly correct, focusing only on branches that have support in the literature. We have successfully used this approach in the past in several publications[72–75].

A2.3. Accounting for genome-wide discordance.

Background. So far, we have assumed the genome follows a single evolutionary history. The true phylogenetic history may change across the genome for several reasons, including Incomplete Lineage Sorting (ILS), duplication and loss (duploss), hybridization, and horizontal gene transfer (HGT)[76,77]. Hybridization and HGT may render a phylogenetic tree meaningless, necessitating a phylogenetic network[78]; however, other cause of discordance (ILS and duploss) are compatible with a single *species tree* and many discordant *gene trees*. Several statistical models for various causes of discordance have been designed[79] and reconstruction methods have been accordingly developed. Summary methods first infer gene trees and then summarize them to build species trees[80–85]. Site-based methods[86,87] combine all aligned loci and directly compute the species tree. Co-estimation methods[88,89] infer gene trees and species trees together and gene tree improvement methods[73,90] infer better gene trees. However, all these methods are alignment-based. To our knowledge, no assembly-free and alignment-free method considers gene tree discordance. However, more recently, theoretical results have started to emerge. The most relevant is the work of Dasarathy *et al.*[91,92] who show that in theory (never tested on data) that pairwise hamming distances can be used to build the species tree despite gene tree discordance. Moreover, Durden and Sullivant show that the species tree is theoretically identifiable from k-mer frequencies despite gene tree discordance[93].

Approach. We will develop provably statistically consistent estimators of the species tree under the multi-species coalescent model[94] of ILS and given genome-skims as input. Skmer can compute hamming distances from genome-skims and there is a deep theoretical connection between the hamming distance and ILS[92]. We will exploit this theoretical connection. Dealing with divergences from a strict molecular clock and divergences from a JC model will both need new theoretical development. We have a successful history of developing state-of-the-art methods of dealing with gene tree discordance[72–75,80,81] and we will use our existing pipelines for testing new methods.

Another major challenge is the prevalence of duplications, losses, and repeats in genomes of many species, such as plants. Birth/death models of duplication and loss and models of polyploidy exist[95–97]. A major question is how the phylogenetic distance should be defined under these models and the literature concerning these questions[98] has assumed available genome assemblies. We will seek to develop new measures of distance between genomes in the presence of duplication, loss, and repeats that can be computed from genome-skims. The existing models may prove too complex for k-mer data. However, simplified models will be considered with the goal of computing Jaccard index under between two genomes that can differ by duplications, losses, and repeats.

A2.4. Phylogenetic placement

Background. Several methods for phylogenetic placement exist[14,15,99]. Placement, unlike *de novo* phylogenetic inference, is fast, rendering it useful for query classification, especially of microbiome[100]. Placement methods first align the query data to a reference marker gene and then use maximum likelihood to find the best position for the query sequence in the reference tree. Thus, both alignment-based and alignment-free existing methods assume assembled marker genes are available for the *reference* set, making them unusable when reference are also genome-skims. Using k-mers to phylogenetically/taxonomically characterize reads has also been studied (e.g., Kraken[29]).

Approach. We will develop new algorithms for placing a query sequence into a phylogenetic tree where both the query and the tree are represented as genome-skims. The most promising approach is to formulate a distance-based optimization problem: given a query sequence q a phylogenetic tree t and a distance matrix $\mathbf{D}$ that includes both the library and the query genome-skims, find the placement of the query into the reference phylogeny that maximally matches $\mathbf{D}$. Fast solutions using the four-point condition can be designed (e.g., our preliminary algorithm in Figure 4d). We will develop more sophisticated forms of the four-point condition method and will also explore algorithms using the minimum evolution principal. Beyond a distance-based formulation, we can seek to find long "marker" k-mers that are unique to a node (internal and terminal) in the phylogenetic tree; preprocessing and finding these markers will then allow a binary search algorithm to find the best placement; this strategy, which has been successfully used in metagenomics[29], may have limitations as the coverage drops; with low coverage, marker k-mers may be missing from the reference, query, or both. Nevertheless, they should be explored.

Challenges. Phylogenetic placement may not be exact, especially given low coverage. Therefore, we will need to have either a way to produce a distribution on possible placements or at least, a measure of support for any given placement. While likelihood-based methods avail themselves naturally to such measures, defining a statistical measure of support based on distances will be challenging. Slow methods such as bootstrapping should ideally be avoided. Perhaps, local measures of support, successfully used by us[101] and others[102,103] for other phylogenetic problems, can be developed. Beyond support, running time will be an issue. Our goal is to enable the placement of a query genome-skim onto a preprocessed reference phylogeny of 10,000 species with less than ten minutes of running time. Any implementation of the placement algorithm should be highly optimized and should exploit the abundant parallelism inherently available in this problem.

5 Aim3: Meta-barcoding

In *meta-barcoding*, we consider the scenario where the sample is not a single organism, but a collection. The goal is to identify the constituent taxa, or to place them in a reference phylogeny. Unlike metagenomics used for microorganisms, here, it is often possible to physically separate out organisms; however, the ability to directly barcode a mixed sample greatly reduces the effort/cost. In meta-barcoding, we can assume that the number of constituent taxa, s, is small (e.g., $m \leq 20$). From a computational perspective, meta-barcoding has many parallels to metagenomics with the difference that we do not expect to see hundreds of species in a sample. However, the nature of the reference library is very different. In our desired meta-barcoding system, both the query sample *and* the reference library of known species are represented by unassembled genome-skims. Most existing tools designed for metagenomics either require a form of assembly for the *reference* library or have explicitly or implicitly assumed that the reference is sampled with high coverage (a very reasonable assumption for microorganisms with small genomes). Even when they do not make such assumptions, they have only been *tested* with high coverage references. Thus, whether they can be used for meta-barcoding against a reference database of genome-skims is unclear.

Solving the meta-barcoding problem accurately and efficiently is going to prove difficult. Given the scope of this project, we will build the foundations required for meta-barcoding, but will leave a thorough solution to future projects. We plan to accomplish two tasks. A3.1. We test existing methods from the metagenomic field to see how they perform for low coverage unassembled reference genomes. A3.2. We lay the theoretical foundations required to solve the problem. As mentioned in Aim 1, methods of handling contamination may also prove useful here.

A3.1. Many metagenomic techniques[104–106] require marker loci and thus cannot be applied to reference genome-skims. Compositional methods (e.g., PhyloPythia[107,108], Phymm[109], and NBC[110]) do not require markers and may work on genome-skims. To understand whether existing methods are sufficient for meta-barcoding of genome-skims, we plan to test them thoroughly. For example, Kraken[29] can place every read of a sample individually on a taxonomy. To do so, it needs to precompute the LCA in the taxonomy of all the nodes that include a k-mer. When the coverage of the reference sequences is low, this approach may not work as lack of coverage may push down LCAs. Moreover, each read is analyzed by Kraken individually. When we expect to have hundreds of species in a sample, as in metagenomics, this approach is perhaps defensible. However, for meta-barcoding, the independent treatment of reads may lead to gross over-estimation of the number of constituent species. As part of this project, we will carefully study existing composition-based metagenomic methods in the context of meta-barcoding with genome-skims.

A3.2. Assuming existing approaches are not sufficient (our conjecture), we will explore better methods. The problem can be formalized in several ways. The most promising approach is to expand phylogenetic placement to mixed samples. Let a meta-barcoding sample Q contain s species and let ν_i, $1 \leq i \leq s$, denote the unknown fraction of k-mers from each species. By definition, $\sum_i \nu_i = 1$. We denote a phylogeny on r reference taxa by an additive distance matrix D. The additive property ensures the existence of a tree in which the tree-distance between taxa i and j is also $T[i,j] = T[j,i]$. Define *placement* of Q as the construction of an augmented additive matrix D' over $r+s$ taxa where $T'[i,j] = T[i,j]$ for all $1 \leq i \leq j \leq r$. For all $1 \leq i \leq r$, $1 \leq j \leq s$, $T'[i, r+j]$ denotes the tree distance between query taxon j and reference taxon i. Also, let H' be the matrix of expected hamming distances computed from the tree T' according to a Markov model of evolution such as GTR. The placement T', in effect, is an answer to the meta-barcoding problem. If desired, a taxonomic classification can be built on top of the phylogenetic placement, as done in the metagenomic literature[105]. Finally, we may also desire estimates of ν_i values.

If the query contains M k-mers, query j contributes $\nu_j M$ k-mers, covering $\eta_j = 1 - e^{-\nu_j M/L}$ fraction of its k-mers (assuming all genomes have the same size but this can be easily relaxed). Assuming no sequencing error, the probability p_i that a k-mer from reference i is covered by the query is $p_i = 1 - \prod_j (1 - \eta_j(1 - H'[i, r+j])^k)$. Then, the Jaccard index between taxon i and the query can be estimated as

$$J_i = \frac{\eta_i p_i}{(\eta_i + p_i - \eta_i p_i)}$$

We will estimate $T', \boldsymbol{\nu}$ using an optimization $\text{argmin}_{T',\boldsymbol{\nu}} \|\boldsymbol{J} - \hat{\boldsymbol{J}}\|_p$, where $\hat{\boldsymbol{J}}$ represent the observed Jaccard indices against each of the reference sequences and $\hat{\boldsymbol{J}}$ is the vector of J_i values. The choice

of norm ($p = 1$ or $p = 2$) will be decided based on empirical data, and the type of optimization, and some sparsity constraints on $\boldsymbol{\nu}$ will be added as s is likely to be small. We will also explore alternative formalizations based on a likelihood score for each solution. We will study these approaches theoretically and empirically against simulated and real data from collaborators.

6 Intellectual merit

The proposed research creates a novel paradigm for measuring biodiversity. If successful, the activities will allow us to sample genomes of multiple organisms for a fraction of the current costs of labor and genome sequencing required for other approaches. The proposal uses a number of novel algorithmic and statistical ideas to correctly estimate genomic distance based on genome-skims. Specifically, it will be the first systematic analysis of how close we can get to computing the genomic distance using only a small and random fraction of an individual's genome. The proposed methods can easily be extended to other vexing problems such as the identification of phylogenetic placement of a previously unknown organism, provenance of a museum animal, sibling analyses, and possibly forensic science. The investigators have a strong history of prior research in related fields, but have complementary expertise. Lead PI Mirarab is an expert in evolution and phylogeny reconstruction of organisms and helped reconstruct the state of the art in avian[111] and plant[112] phylogeny, has developed highly used tools for phylogeny inference[80,81] and metagenomics[99,105]. PI Bafna has expertise in genomics, and the use of novel techniques for analyzing complex structural variation[113] as well as the development of new sequencing technologies and computational tools for analyzing the resulting data[114–116]. He also has expertise in population genetics, and the impact of environment (selection processes) in shaping the population genetic variation landscape[117,118].

7 Results from Prior NSF Support

Bafna: NSF-III (1318386) "Algorithms for decoding complex patterns of genomic variation" ($500K, 9/01/13-8/31/17). **Intellectual Merit:** Multiple publications resulted from this grant[115–136], focused on exploiting the allele frequency spectrum for identifying regions under selection, experimental evolution, haplotype assembly, and decoding of complex structural variation. **Broader Impacts:** include software tools, CLEAR for analyzing experimental evolution[134], InPhaDel[130] and Hapcut2[115] for haplotyping, and PreCIOSS for identifying carriers of a selective sweep[118].
Mirarab: III (1565862) "Using Genomic Context to Understand Evolutionary Histories of Individual Genes" ($175,000, 7/1/2016–6/30/2018). **Intellectual Merit:** Several publications have results from the grant[72,74,137–141]. **Broader Impacts:** include several publicly available software tools: DISTIQUE[137], MV Rooting[138], ASTRAL-3[74], and TreeShrink[140]. All tools are available at `http://eceweb.ucsd.edu/ smirarab/software.html`. We have also created benchmark datasets for the community to use: `https://sites.google.com/eng.ucsd.edu/datasets/`. We also mentored three undergraduate students, one through the STAR program, through this project.