Electronic Preprint for Journal of Information Processing Vol.27

Regular Paper

An Effective Parameter-free Comparison of NGS Short Reads for Phylogeny Reconstruction

Phanucheep Chotnithi[1,a)] Hong Van Le[2,b)] Atsuhiro Takasu[1,2,c)]

Received: March 9, 2019, Accepted: July 1, 2019

Abstract: With the growth of the amount of genomic data generated from high-throughput sequencing, next-generation sequencing (NGS) has become the mainstream format for genome sequence data. NGS presents new challenges for many applications for genome sequence analysis. In sequence comparison applications, traditional multiple-sequence alignment approaches do not provide a solution for analyzing NGS data because of the short-read assembly and computational resource problems. Thus, alignment-free methods are more suitable for NGS data comparisons. Most of the alignment-free methods are based on the k-mer algorithm. However, the characteristics of NGS data make such k-mer-based methods suboptimal because the k parameter is a crucial factor in distance measurement and for the construction of phylogenetic trees. We propose an effective parameter-free comparison of NGS short reads, with the aim of eliminating the dependency on the k parameters. We compared the proposed method with existing methods, and the results show that the proposed method can measure accurate distances for the dataset without requiring any parameter.

Keywords: bioinformatics, phylogeny, NGS read alignment

1. Introduction

Sequencing is a process that transforms data from genome samples into a digitized data sequence. Nowadays, there are a tremendous number of sequencing methods and techniques that are used in the process. Traditional sequencing processes provide long sequences for a DNA sample. However, this sequencing process is only available for a small portion of DNA sequence per sample, such as mitochondrial DNA (mtDNA) or prokaryote DNA. It cannot be used to sequence the whole genome because of the massive length of the DNA sequence. Recently, next-generation sequencing (NGS) [1] has been introduced to achieve high-throughput sequencing as a significant advance over traditional processes. By using a different sequencing technique, NGS provides large numbers of sequence fragments called *reads*, per genome sample, instead of one long sequence of genome data.

The data sequence obtained from the sequencing process is used in sequence comparison and phylogeny reconstruction processes to generate a phylogenetic tree, which is essential for a vast number of studies in the biology field. Typically, sequence comparison algorithms use one long genome sequence, such as 16S rRNA in mtDNA, to calculate the distance between each sequence and to construct a distance matrix [2], [3], [4]. The clustering and classification algorithms are then applied to the distance matrix to construct a phylogenetic tree that shows evolutionary relationships between sequences. To construct an accurate tree, an efficient sequence comparison method is required.

The emergence of NGS short reads data with the new form of genome sequences creates new challenges for traditional methods of sequence comparison [5], [6]. The alignment-based methods such as multiple-sequence alignment (MSA) have trouble dealing with a large proportion of NGS short reads data. Moreover, when NGS was introduced, the differences between NGS short reads data and long sequence data needed to be considered. Assembly, which is a procedure that is used to reconstruct the NGS short reads into the long sequence, is needed when working with NGS data. In the assembly procedure, NGS short reads are mapped onto a template sequence, which involves significant computational cost. However, to assemble the genome without template sequences is very challenging because the reads are mostly short and contain large numbers of repeated genome sequences.

Recently, alignment-free methods for sequence comparison have attracted attention from researchers because of its high processing efficiency compared with alignment-based methods. These methods have an advantage over MSA in the assembly process because they do not require an assembly process; hence, they are scalable to large numbers of NGS short reads. Most alignment-free methods rely on k-mer frequencies to measure the distance [4]. Several alignment-free methods have been proposed to focus specifically on NGS short reads data. CVTree [7], [8], d_2^S [9] and skmer [10] have shown good results for distance measurements and phylogeny reconstruction with both NGS short reads data and long genome sequences. However, these methods depend significantly on a parameter k, and different values of k could lead to different phylogenetic tree results. Hence, it is difficult for researchers to determine which k value would construct a tree that is closest to the natural evolutionary relation between

[1] SOKENDAI, The Graduate University for Advanced Studies, Hayama, Kanagawa 240–0193, Japan
[2] National Institute of Informatics, Chiyoda, Tokyo 101–8430, Japan
[a)] phanucheep@nii.ac.jp
[b)] l-van@nii.ac.jp
[c)] takasu@nii.ac.jp

their input species. Moreover, alignment-free methods remain less accurate than MSA.

The goal of this research is to develop a novel sequence comparison approach that requires no k parameter adjustment while maintaining the accuracy of the result. We utilize the information on short reads alignment for comparison of NGS data. Instead of assembling the NGS short reads then aligning the result with the other sequences to measure their distance, we propose a new method, namely d^{RA}, that is based on the alignment of NGS short reads themselves. The main idea is that if the sequence ends up aligned with others after assembly, their NGS short reads before assembly should also be aligned. By searching for the corresponding NGS short reads between each set and then calculating the distance from their alignment, this method allows the distance to be calculated with no dependency on the k parameter and to maintain the same accuracy as the alignment-based approach. Our method also has no requirement for assembly, like the alignment-free approach.

We have compared our method with alignment-free methods and found that our novel read alignment approach can provide a more accurate distance measurement on three simulated NGS datasets to construct the phylogenetic tree than other alignment-free methods. The phylogenetic trees constructed using the new method are similar to the benchmark tree obtained by other researchers while requiring no parameter adjustment. We also conducted experiments on multiple simulated NGS sets from the same dataset to evaluate the effect on different reads' randomness and coverage. Our approach delivers similar measured distances among each set.

In summary, this paper makes the following contributions:

- We propose a novel sequence comparison approach, namely d^{RA}, which requires no k parameter while maintaining the accuracy of the result. Because d^{RA} is a k-free approach, it can be applied even on NGS sets without benchmark trees, whereas it is difficult to adjust the k parameter for other alignment-free approaches in such NGS sets.
- We utilize the Gaussian mixture model to improve the accuracy of the distance measurement of our approach.
- We conducted experiments on three real datasets to measure the accuracy of our proposed approach in comparison with other alignment-free approaches. Along with the accuracy, we also measured the consistency of pair-wise distance computation to evaluate better the effectiveness of our proposed method. The experimental results indicate that d^{RA} provides higher accuracy while maintaining consistency compared with other baseline methods.
- We also conducted experiments to evaluate the efficiency of the proposed approach. From empirical evidence, d^{RA} mostly outperformed other alignment-free approaches. In some cases, although d^{RA} takes longer processing time, it offers better accuracy and consistency than baseline approaches.

2. Problem Definition

In this research, we focus on the comparison of NGS data sequences for phylogeny reconstruction. With the input as the NGS sets of the species, the phylogenetic tree of those input NGS sets can be reconstructed. This tree shows the evolutionary relationships between the input species according to their genome sequence distances. To construct an accurate phylogenetic tree, a reasonable distance measurement between the NGS sets is required.

Define $A = \{a_1, a_2, \ldots, a_n\}$ as the NGS set with n short reads. Each short read $a_i \in w^*$ is a genome sequence of four nucleotide characters $w = \{A, C, T, G\}$. With several NGS sets as the input, the distance matrix M contains every pair-wise distance between each NGS set. Using the distance matrix M, we can reconstruct the phylogenetic tree result of the input NGS sets.

Many methods have been proposed to measure an accurate distance. Alignment-free approaches are considered as efficient methods for the task. Most of the alignment-free methods are based on the k-mer profile of the sequence. The k-mer profile of the genome sequence s can be defined as all possible substrings in s with length k. The parameter k is crucial for the distance measurement on these k-mer-based methods because it significantly affects the distance measurement result. Therefore, in this paper, we address the problem of k dependency in the k-mer-based methods but still maintain the accuracy of the distance measurement.

3. Related Work

Sequence comparison is a well-studied problem for genome sequence analysis. Many researchers have proposed sequence comparison methods over the past decade. The traditional method is alignment-based MSA. There are several tools that are available for efficient MSA such as the *Clustal* series [11], [12], [13], *T-coffee* [14] and *MUSCLE* [15].

However, with the growth in the number of sequencing techniques, MSA is limited because of its low efficiency for the comparison of large genomes. The alignment-free approaches such as *FFP* [16], *kmacs* [17], *spaced-word* [18], and *kSNP v2* [19] have been introduced to address the problem of MSA. Three k-mer-based alignment-free methods are used as the baselines in this paper; namely, CVTree, d_2^S, and skmer. Both CVTree and d_2^S focus on normalized k-mer frequencies. While skmer is based on k-mer occurrence, CVTree calculates the distance between two genome sequences or NGS short reads sets by using their normalized k-mer frequency vector, called the composite vector (CV). d_2^S is a statistical approach to modify raw distance measures to produce measures that better suit the NGS data.

3.1 CVTree: CV Alignment-free Method

For a fixed length k, count separately the number of substrings of length $k, k-1, k-2$ on each input sequence. The initial CV is the number of k-mer frequency, which is $N = 4^k$ total dimensions for DNA sequences and $N = 20^k$ for protein sequences in lexicographic order. Calculate the subtraction score for the k-mer a_i:

$$a_i(\alpha_1\alpha_2\cdots\alpha_k) \equiv \frac{f(\alpha_1\alpha_2\cdots\alpha_k) - f^0(\alpha_1\alpha_2\cdots\alpha_k)}{f^0(\alpha_1\alpha_2\cdots\alpha_k)},$$

where $f(\alpha_1\alpha_2\cdots\alpha_k)$ is the frequency of k-mer $\alpha_1\alpha_2\cdots\alpha_k$ and

$f^0(\alpha_1\alpha_2\cdots\alpha_k)$ is the predicted frequency of the k-mer calculated using a $(k-2)$-th Markov assumption.

Let $CV_A = a_1a_2\cdots a_N$ and $CV_B = b_1b_2\cdots b_N$ be the CVs for the species A and B, respectively. Finally, calculate the distance matrix for the modified CV:

$$D(A,B) = (1 - C(CV_A, CV_B))/2,$$

where

$$C(CV_A, CV_B) = \frac{\sum_{i=1}^{N} a_i \times b_i}{\sqrt{\sum_{i=1}^{N} a_i^2 \times \sum_{i=1}^{N} b_i^2}}.$$

3.2 d_2^S k-mer Statistical Alignment-free Method

d_2^S statistics is a modified version of D_2, D_2^*, and D_2^S statistics [14], [15]. They consider the random processes of NGS data in terms of D_2, D_2^*, and D_2^S to model the correct k-mer distribution of NGS data. NGS short reads are small fragments from the original long sequence, which means that the method of sampling those reads will affect the k-mer frequency distribution. Another characteristic of NGS data relevant to d_2^S statistics is that an NGS short read can originate from the forward or reverse strand of the original genome, requiring consideration of not only the k-mer distributions of short-read data themselves but also their complementary sequences.

Suppose that M reads of length β are sampled from a genome of length n. Let X_w and Y_w be the numbers of occurrences of k-mer w in the M pairs of reads from the first genome and the second genome, respectively. We define $\widetilde{X}_w^2 = X_w - M(b-k+1)(p_w + p_{\bar{w}})$ with $\widetilde{Y}_w^2$ being defined analogously. Let $w = w_1w_2\cdots w_k$ and $p_w = p_{w_1}p_{w_2}\cdots p_{w_k}$, with $\bar{w}$ being the complement of word w. Consider two genome sequences taking L letters $(0, 1, \ldots, L-1)$ at each position. For the null model, we assume that the two genomes are independent, and both are generated by models with p_l being the probability of taking state l, $l = 0, 1, \ldots, L-1$. d_2^S can be calculated by:

$$d_2^S = \frac{1}{2}\left(1 - \frac{D_2^S}{\sqrt{\sum_{w\in A^k} \widetilde{X}_w^2/\widetilde{Z}_w^2}\sqrt{\sum_{w\in A^k} \widetilde{Y}_w^2/\widetilde{Z}_w^2}}\right),$$

where

$$D_2^S = \frac{\widetilde{X}_w\widetilde{Y}_w}{\widetilde{Z}_w}$$

and

$$\widetilde{Z}_w = \sqrt{\widetilde{X}_w^2 + \widetilde{Y}_w^2}.$$

3.3 Skmer: Assembly-free and Alignment-free Sample Identification Using Genome Skims

This method is also based on k-mer, like many alignment-free methods. In general, skmer is an improvement of Mash [20], in which the Jaccard index (J), a similarity measure between any two sets (in this case, k-mer occurrence) defined as the size of their intersection divided by the size of their union, is estimated efficiently by using a hashing procedure. The similarity is then used to estimate the genomic distance between two genomes. The problem of Mash is that its similarity is impacted by many factors such as coverage, sequencing error, and data length. Skmer aimed to solve all the effects of these factors with respect to the final similarity. There are two steps in skmer: the first step is using k-mer frequency profiles to estimate the sequencing error and the coverage. Let M_i be the number of k-mer observed i times in the genome-skim. Let $h = argmax_{i\geq 2} M_i$. By defining $\xi = \frac{M_{h+1}}{M_h}(h+1)$, k-mer coverage (λ) and the sequencing error rate (ϵ) can be calculated by the equations:

$$\lambda = \frac{M_1}{M_h}\frac{\xi^h}{h!}e^{-\xi} + \xi(1 - e^{-\xi}),$$

$$\epsilon = 1 - (\xi/\lambda)^{1/k}.$$

In the next step, they use the hashing technique of Mash to compute the Jaccard index J and then compute the final genomic distance using the equation:

$$D = 1 - \left(\frac{2(\zeta_1L_1 + \zeta_2L_2)J}{\eta_1\eta_2(L_1+L_2)(1+J)}\right)^{1/k},$$

where for $i \in \{1, 2\}$, $\eta_i = 1 - e^{-\lambda_i(1-\epsilon_i)^k}$ and $\zeta_i = \eta_i + \lambda_i(1-(1-\epsilon_i)^k)$, and L_i is the estimated genome length.

4. Proposed Method

According to Ref. [21], lack of alignment makes it more difficult to extract all of the possible information about evolutionary distances between species from k-mer-based methods because they only use differences in the presence/absence of k-mers. For example, if k-mer contains multiple substitutions, it is counted as one k-mer difference, which is the same as a k-mer that contains only one substitution. Thus, a lower k is more sensitive to the evolutionary distances than a larger k. However, the lower k causes the homoplasy problem, which is popularly considered as "noise" in the phylogenetic tree reconstruction [21], [22]. Therefore, the parameter k affects distance measurement and needs to be appropriately set.

Moreover, larger k in the k-mer-based methods can deal with the homoplasy problem but is not sensitive to the evolutionary distances because it causes more loss of evolutionary information. Hence, k-mer-based methods require large datasets with vast amounts of data to provide accurate distances and balance the effect of long k-mer [21].

To solve these problems in the k-mer-based methods, we propose a novel approach to eliminate the dependency of the k parameter so that the method works well with not only large datasets but also small datasets. To maintain the accuracy of the method as much as possible, we take advantage of the alignment aspect of MSA because the alignment method evaluates the evolutionary distances based on the mutation that causes the substitution directly. When working with the NGS short reads data, many methods need to use an assembly process, but this is time-consuming and has the problem of lack of suitable reference sequences. Hence, our method focuses on the alignment approach without assembly on NGS data. Then, the problem becomes how to approximate the distance between NGS short reads sets without assembly. To tackle this problem, we propose a method to

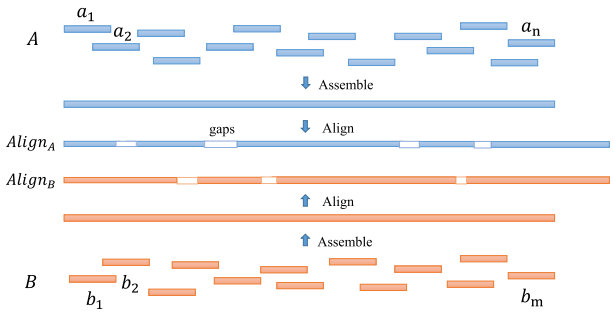

Fig. 1 The traditional method to align two NGS short reads data.

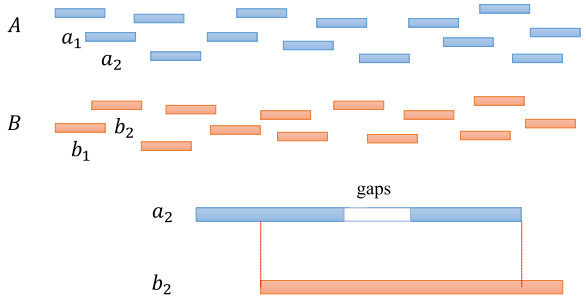

Fig. 2 The relationship of the alignment between two NGS short reads without assembly.

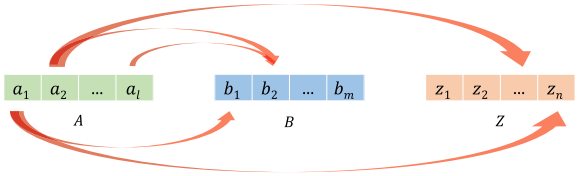

Fig. 3 Alignment pairs searching.

combine the distance of each alignment pair into an accurate pair-wise distance using the Gaussian mixture model. We call our method the *short read alignment approach* or d^{RA}.

To define the method, we consider the relationship between the alignment of assembled sequences and the NGS short reads without assembly. With two NGS sets $A = \{a_1, a_2, \ldots, a_n\}$ and $B = \{b_1, b_2, \ldots, b_m\}$, let S_A and S_B be the sequences assembled from NGS sets A and B, respectively. S_A and S_B are aligned into $Align_A$ and $Align_B$ by inserting some gaps. The distance between these two sequences can be calculated from those aligned sequences, as shown in **Fig. 1**.

To replicate the alignment of the $Align_A$ and $Align_B$ on the NGS short reads without assembly, we assume that, for some NGS short read, $a \in A$, a could be considered to inexactly match (the matching process which allows some mismatch and gaps) with the substring of $Align_B$ because the gaps are allowed in the alignment. Given that $Align_A$ and $Align_B$ are aligned with each other, a is also an inexact match with the substring of $Align_B$. In other words, some NGS short reads $a \in A$ are an inexact match with some NGS short reads $b \in B$. With this information, we could establish the relationship of the alignment with NGS short reads directly without assembly.

For a pair of strings x and y, let $d(x, y)$ denote the normalized unit cost edit distance, i.e., $d(x, y)$ is calculated by the edit distance with all costs of operation being equal to 1 between x and y divided by $max(|x|, |y|)$. Consider two aligned sequences $Align_A = align_{A1} \ldots align_{An}$ and $Align_B = align_{B1} \ldots align_{Bn}$ with the distance between them equal to $d(AlignA, AlignB)$. Assume that the probability that the substitution, insertion, and deletion occur is independent and uniform in $Align_A$ and $Align_B$. Therefore, with any corresponding substrings $s_a = align_{Ai} \ldots align_{Aj}$ and $s_b = align_{Bi} \ldots align_{Bj}$ where $1 \le i, j \le n$, the normalized unit cost edit distance $d(s_a, s_b) \approx d(Align_A, Align_B)$.

Because some NGS short read $a \in A$ can be an inexact match or alignment with some NGS read $b \in B$ when A and B are the NGS sets, we could consider the alignment part between a and b as the corresponding substring of $Align_A$ and $Align_B$. For example, in **Fig. 2**, the NGS short read $a_2 \in A$ is the alignment pair of $b_2 \in B$ with the alignment part shown as the region between the red lines. However, only one alignment pair is not enough to approximate an accurate distance $d(Align_A, Align_B)$. With the collection of the alignment pairs between NGS short reads of A and B, the concatenation of the alignment parts that represent the longer corresponding substrings of $Align_A$ and $Align_B$ would provide more accurate distance approximation of the $d(Align_A, Align_B)$.

4.1 Alignment Pair Searching

At this step, we search for the alignment pair from each NGS sets for every NGS short reads that are required in the next step. The set of alignment pairs between A and B is denoted by $P(A, B)$ as follows:

$$P(A, B) = \bigcup_{i=1}^{n} \underset{(a_i, b) \in \{a_i\} \times B}{argmin}\ d(a_i, b).$$

The alignment pairs in $P(A, B)$ are the pairs of NGS short reads $a \in A$ and $b \in B$ with minimum edit distance. **Figure 3** shows an example of the alignment pair searching. According to the figure, there are several NGS sets A, B to Z. Each set consists of the NGS short reads $A = \{a_1, a_2, \ldots, a_l\}$, $B = \{b_1, b_2, \ldots, b_m\}$ until $Z = \{z_1, z_2, \ldots, z_n\}$ when l, m, and n are the size of sets A, B, and Z, respectively. We search for the alignment pair of each NGS short reads of all other sets. For example, with $a_1 \in A$, we first search the alignment pair of a_1 in B and find b_1. Then, we continue searching in the other sets until the last set Z. In Z, we find the alignment pair is z_n as shown in the figure. Given that the size of each set is not the same, some short reads might be aligned to more than one read. If $|A| > |B|$ then some short read in A could be aligned with the same short read in B. For example, short read a_2 and a_i are paired with b_2.

4.2 Pair-wise Distance Measurement

After retrieving a collection of alignment pairs from the alignment pair searching step, we use the distance of alignment pairs to calculate the final pair-wise distance between NGS sets. For the distance measurement, we consider each alignment pair as a part of the overall alignment between two NGS sets. Hence, we could estimate the distance between any NGS sets by combining the alignment pairs corresponding to those sets with the following equation:

$$d^{RA} = (D(A, B) + D(B, A))/2, \tag{1}$$

$$D(A, B) = \sum_{(a,b) \in P(A,B)} \left(-\frac{3}{4} ln\left(1 - \frac{4}{3} d(a, b)\right)\right) * w_{s(a,b)}. \tag{2}$$

We define $D(A, B)$ as our pair-wise distance between two NGS sets A and B (Eq. (1)). So $D(A, B)$ can be calculated by the summation of the Jukes–Cantor distances of any corresponding alignment pair in $P(A, B)$ with the weight of $w_{s(a,b)}$. The Jukes–Cantor model estimates the evolutionary distance between DNA sequences by considering the mutation rate of the nucleotide. The model assumes that all four nucleotides A, C, T, and G have the same probability of appearing in the sequence and the same mutation rate. Given that the alignment set $P(A, B)$ is not equal to $P(B, A)$, $D(A, B)$ is asymmetric. So, we define our distance measurement d^{RA} as an average of $D(A, B)$ and $D(B, A)$.

The weight $w_{s(a,b)}$ is from the assumption that each individual alignment pair distance should not contribute to the final pair-wise distance equally. The significance of the alignment pairs increases exponentially to the similarity of the alignment pair [23]. Hence, the relationship between the weight $w_{s(a,b)}$ and the similarity $s(a, b) = 1 - d(a, b)$ is defined as follows:

$$w_{s(a,b)} = \frac{\exp(s(a,b))}{\sum_{(a,b)\in P(A,B)} \exp(s(a,b))}. \quad (3)$$

However, there are some cases in which the alignment pairs retrieved in the searching step are not the corresponding alignment pairs from the alignment of the assembly sequences. These non-corresponding alignment pairs should contribute to the pair-wise distance significantly less than the corresponding pairs.

We assume that the distribution of frequency of the all alignment pairs $(a, b) \in P(A, B)$ according to their similarity is a bimodal distribution. The bimodal distribution consists of two modes (peaks). In this case, the distribution of the first mode with less similarity is referred to as *noncorresponding* alignment pairs, and the second mode with more similarity as *corresponding* alignment pairs. The alignment pairs with high similarity have more probability of being the corresponding pairs.

For the alignment pair $(a, b) \in P(A, B)$, let $Prob(a, b)$ denote the probability that the pair (a, b) is the corresponding pair. $prob(a, b)$ can be calculated by learning the bimodal distribution using the Gaussian mixture model with an expectation maximization algorithm [24], [25]. To reduce the significance of noncorresponding alignment pairs, weight $w_{s(a,b)}$ was redefined according to the $Prob(a, b)$ by the following equation:

$$w_{s(a,b)} = \frac{\exp(s(a,b)) * Prob(a,b)}{\sum_{(a,b)\in P(a,b)} \exp(s(a,b)) * Prob(a,b)}. \quad (4)$$

5. Experiment and Evaluation

5.1 Experiment Setup

5.1.1 Datasets

To evaluate our proposed method d^{RA}, we use three datasets, 29 *mammalian mtDNA* sequences [26], [27], 29 *Escherichia/Shigella* [28], and 18 *Drosophila* genomes [10]. The 29 *mammalian mtDNA* dataset consists of the mitrocondrial DNA sequence within 29 mammal species. The 29 *Escherichia/Shigella* dataset consists of the entire genome sequences of 29 species of bacteria in the family of *Escherichia* and *Shigella*. The last dataset is the 18 species of fly (insect) or *Drosophila*. The statistics of all three datasets are shown in **Table 1**.

Table 1 Size and the total sequence lengths of the three datasets.

	Size (MB)	Total sequence lengths
29 *mammalian mtDNA*	0.5	482,127
29 *Escherichia/Shigella*	144	141,962,164
18 *Drosophila*	3,110	3,109,816,396

5.1.2 Experiment Procedure

Given that all three datasets were initially long sequences, we used a tool called *ART* [29] to simulate NGS short reads from the long genome sequences. We used two error models; namely, *454* and *Illumina*, to simulate the NGS high-throughput sequencing results from two different NGS platforms. These methods produced the actual samples as NGS short reads data. The *454* model produces various lengths of NGS short reads and has a high chance of sequencing errors on homopolymer sequences, which include multiple consecutive duplicate characters. Meanwhile, the *Illumina* model provides fixed-lengths of NGS short reads and has no problem with the homopolymer sequences.

We conducted experiments with various values of coverage on each dataset. Coverage is the average times of occurrence of nucleotides at each position in the original sequences that appear in the NGS sets. For example, the coverage value 5x means that the NGS short reads overlap five times according to each position in the original sequences. We could say that the NGS set with 1x coverage is the NGS set with no overlap. The length of NGS short reads was set to 150 bps, with a default parameter for the error distribution for each model.

The size and the total number of original datasets and the simulated NGS sets are summarized in **Table 2**. Because, in practice, researchers usually get low coverage data in the sequencing process, we also conducted experiments on low coverage NGS data in this paper. We simulated four *454* and four *Illumina* NGS sets with 5x coverage in the 29 *mammalian mtDNA* dataset and 1x coverage in the 29 *Escherichia/Shigella* dataset.

In the 18 *Drosophila* dataset, we simulated four *Illumina* NGS sets with 0.1x coverage of the dataset. Because the 18 *Drosophila* dataset is the entire genome sequence dataset, it contains a massive amount of repeated sequences and homopolymer sequences. As noted above, using the *454* model it is possible to have sequencing errors on homopolymer sequences; thus, we did not simulate the *454* NGS sets with this dataset.

With the simulated NGS short reads data, we applied our proposed method d^{RA} to calculate a distance matrix. The phylogenetic tree was then constructed according to the calculated distance matrix.

5.1.3 Baselines Methods

We compared our proposed method with three existing k-mer-based alignment-free methods, CVTree [7], [8], d_2^S [9], and skmer [10]. We used k values in the range from 8 to 31 as suggested by *CVTree*, d_2^S, and *skmer* proponents.

5.1.4 Evaluation Metric

We used the *Clustel Omega* tool [13], followed by the *dnadist* tool in the *PHYLIP* package [30], on aligned sequences from MSA to calculate distance matrices. For each distance matrix, either from MSA or from alignment-free methods, we used the *neighbor* tool in the *PHYLIP* package to construct a phylogenetic

Table 2 Size and the total number of short reads and total sequence lengths of NGS short reads set of all three datasets.

	NGS set	Size (MB)	Total number of short reads	Total sequences length
29 *mammalian mtDNA* (5x)	454_1	5	8,540	1,982,139
	454_2	5	8,571	1,990,340
	454_3	5	8,618	1,989,688
	454_4	5	8,631	1,994,046
	illumina_1	5	16,010	2,401,500
	illumina_2	5	16,010	2,401,500
	illumina_3	5	16,010	2,401,500
	illumina_4	5	16,010	2,401,500
29 *Escherichia/Shigella* (1x)	454_1	260	499,945	111,949,666
	454_2	260	499,782	112,024,738
	454_3	260	499,285	111,918,934
	454_4	260	499,634	111,956,774
	illumina_1	304	946,150	141,922,500
	illumina_2	304	946,169	141,925,350
	illumina_3	304	946,151	141,922,650
	illumina_4	304	946,177	141,926,550
18 *Drosophila* (0.1x)	illumina_1	681	1,908,519	286,277,850
	illumina_2	680	1,907,719	286,157,850
	illumina_3	680	1,907,985	286,197,750
	illumina_4	680	1,908,134	286,220,100

tree by the neighbor-joining method [31].

We used the popular Robinson–Foulds distance (RF) [32] for the evaluation. The RF value was calculated by counting the internal nodes that appear in one tree but not in the others. Let $N = (V, E)$ be a given phylogenetic tree. For any two nodes u, $v \in V$, v is a descendant of u if v is reachable from u in N. For any $v \in V$, define the cluster of v (denoted by $C(v)$) as the set of all leaf nodes that are descendants of v. The cluster collection of N is the multiset $C(N) = \{C(v)|v \in V\}$. The RF distance between two phylogenetic trees N_1 and N_2 is:

$$d_{RF}(N_1, N_2) = (|C(N_1) - C(N_2)| + |C(N_2) - C(N_1)|)/2.$$

A small RF value between two trees means the shapes of the trees are similar. The values for RF range from zero, meaning the two trees are the same, to $2(n-3)$ where n is the number of leaf nodes.

Because MSA is limited by the size of the genome, only the 29 *mammalian mtDNA* dataset is capable of using the tree from MSA as the benchmark tree. The benchmark tree for 29 *Escherichia/Shigella* is the tree studied by the research [28], [33] and 18 *Drosophila* genomes tree is from the phylogenetic tree database *Open Tree of life* [10], [34], [35]. In our implementation, we also used the *USEARCH* tool [36] to search for the alignment pair of any NGS short reads.

To evaluate the consistency of the distance measurement, we utilized the *coefficient of variation* [37]. The coefficient of variation can be calculated by the ratio of the standard deviation σ to the mean μ as follows:

$$CV = \frac{\sigma}{\mu} * 100.$$

5.2 The Accuracy of Phylogenetic Tree Reconstruction

We first estimated phylogenetic trees for 29 *mammalian mtDNA* sequences and 29 *Escherichia/Shigella* genome datasets. For each dataset, we simulated eight NGS short-read sets: four of *454* error model and another four of *Illumina*. For all three alignment-free methods, we set different values of the k parameter to show the effect of this parameter on the phylogenetic tree result.

The results in **Fig. 4** show that d^{RA} provides a beneficial distance measurement, which leads to accurate phylogeny reconstruction in both datasets. The RF distance between phylogenetic trees reconstructed from d^{RA} is the closest to the benchmark tree in most of the NGS short-read sets compared with other methods.

While the k parameter adjustment is required in *CVTree*, d_2^S, and *skmer* to provide the best phylogenetic tree results, d^{RA} does not require such adjustment to provide an accurate result, as shown in **Table 3**. According to Fig. 4, the optimal k for *skmer* in *mammalian mtDNA* sequences dataset is around 13, whereas $k = 31$ provided the best result on the *Escherichia/Shigella* dataset. The phylogenetic tree results by *CVTree* and d_2^S were also affected by the k parameter. In practice, not many NGS sets have benchmark trees, thus adjusting the k parameter to provide the most accurate tree in further analysis is an ambiguous process. d^{RA} is a k-free approach, so it can be applied even on NGS sets without benchmark trees.

Because *Drosophila* has a much larger genome size than *Escherichia/Shigella*, the dataset that includes bacteria data, researchers usually manage to obtain low coverage data of the genome samples by using the NGS process. Therefore, we conducted experiments on 18 *Drosophila* datasets with 0.1x coverage to evaluate the accuracy of our proposed method on low coverage data. As shown in **Fig. 5**, d^{RA} provided a better phylogenetic tree for *Drosophila* in comparison with most of the other baseline approaches. Although *skmer* could also obtain low distances, as found in our approach, it required the k parameter to be tuned to achieve such results. We also observed that *CVTree* and d_2^S could not be used accurately to reconstruct the phylogenetic tree with

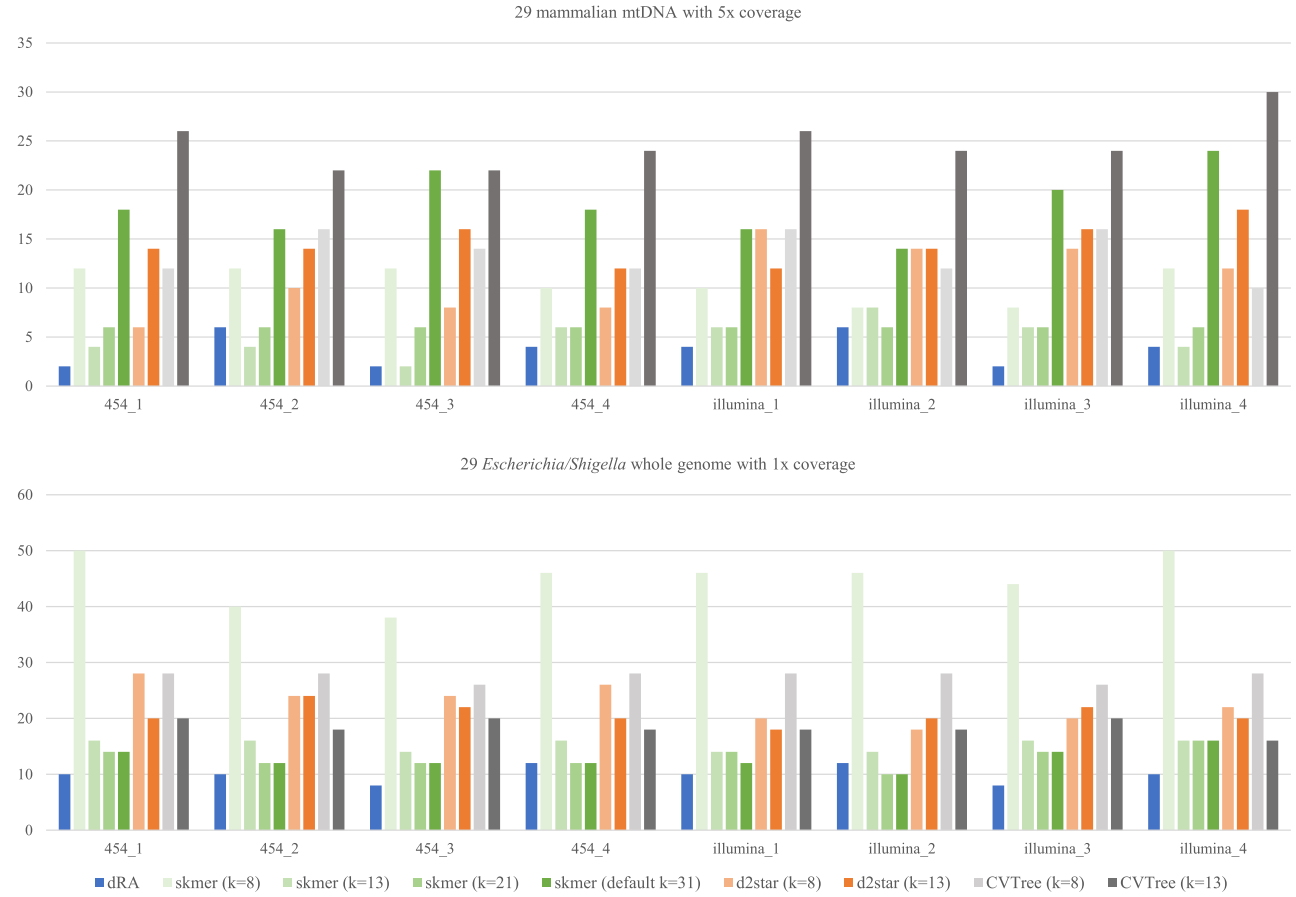

Fig. 4 The RF distance between benchmark tree and phylogenetic trees reconstructed from the distance matrix estimated by our approach (d^{RA}), shown as the blue bar, and other k-mer-based alignment-free methods.

Table 3 Average of RF distance between benchmark tree and phylogenetic trees of all simulated NGS short-read sets.

	mammalian mtDNA (5x)	*Escherichia/Shigella* (1x)
d^{RA}	3.75	10
$Skmer(k=8)$	10.5	45
$Skmer(k=13)$	5	15.25
$Skmer(k=21)$	6	13
$Skmer(k=31)$	18.5	12.75
$d_2^S(k=8)$	11	22.75
$d_2^S(k=13)$	14.5	20.75
$CVTree(k=8)$	13.5	27.5
$CVTree(k=13)$	24.75	18.5

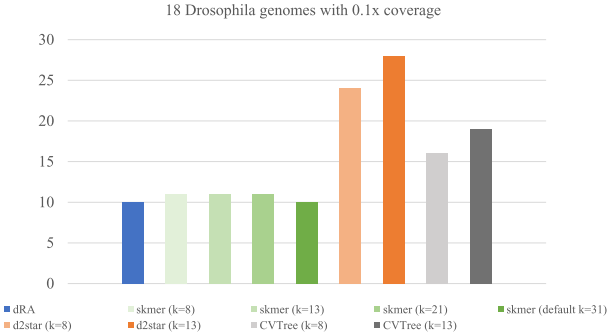

Fig. 5 The average RF distance between benchmark tree and phylogenetic trees reconstructed from the distance matrix estimated using our approach (d^{RA}), shown as the blue bar.

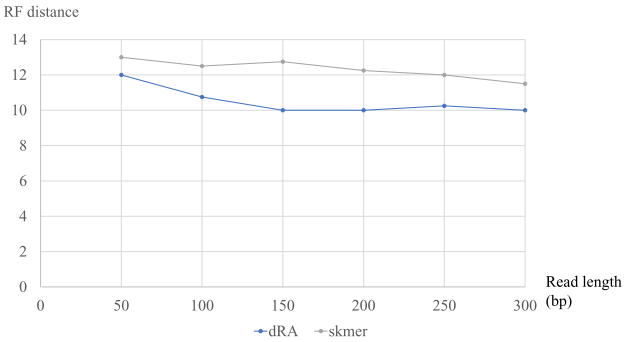

Fig. 6 The RF distance between phylogenetic tree constructed by d^{RA} and the benchmark tree w.r.t. short-read length on the *Escherichia/Shigella* dataset.

this low coverage.

We then evaluated the effect of short-read length on the accuracy of d^{RA}. **Figure 6** summarizes the accuracy with respect to short-read length. According to Fig. 6, d^{RA} does not provide a result on shorter reads (50 bp and 100 bp) that are as accurate as those of the longer reads. Given that the shorter reads contain less information on the alignment between them, the distance calculated from d^{RA} could be less accurate. However, d^{RA} still outperforms *skmer* with respect to accuracy on phylogenetic tree reconstruction.

d^{RA} evaluates the distance between a pair of NGS sets according to the alignment pair of NGS short reads. We conducted ex-

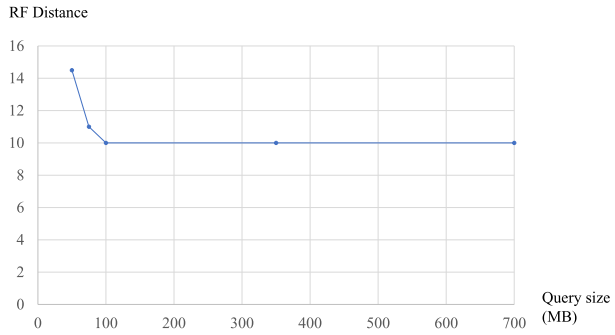

Fig. 7 The RF distance between phylogenetic tree constructed by d^{RA} with varied query size and the benchmark tree on 18 *Drosophila* dataset.

periments (as shown in **Fig. 7**) to examine whether d^{RA} can be used to measure accurate distances with different query sizes. Instead of using all short reads in each set as the query to search for its alignment pair in the other sets, we randomly chose a specific number of short reads as queries. For the 18 *Drosophila* dataset with 0.1x coverage, with the data size of 700 MB, we randomly sampled NGS short reads from each set with an overall size of 50, 70, 100, 350, and 700 MB (all short reads) as a query. The results, summarized in Fig. 7, indicate that d^{RA} can provide tree results close to the benchmark even with low query sizes.

Moreover, as shown in Fig. 5, on all three datasets, the phylogenetic trees of d^{RA} are the closest to the benchmark, regardless of the size and coverage of the datasets. Although the *skmer* can also provide the same results in some cases, we noted that *skmer* is not effective on small datasets such as *mammalian mtDNA* and *Escherichia/Shigella*. Because *skmer* measures the distance based on the k-mer occurrences, small datasets do not provide enough k-mer information for *skmer* to measure accurate distance. *CVTree* and d_2^S also have the same problem. This shows how d^{RA} is a general and effective approach that can be used on any dataset and can be a better option than the other methods.

We also compared the true edit distance and estimated distance given by the following equation:

$$D'(A,B) = \sum_{(a,b)\in P(A,B)} d(a,b) * w_{s(a,b)},$$

where $w_{s(a,b}$ is the weight in Eq. (4).

In this experiment, we used simulated sequences from the *E. coli O157* entire genome sequence. The simulated sequences were set with normalized edit distances equal to 0, 0.01, 0.05, 0.1, 0.15, and 0.2 compared with *E. coli O157* sequence as "true edit distance." For each simulated sequence and also the *E. coli O157* sequence, we generated the corresponding NGS set with varied coverage 0.25x, 0.5x, 1x, 2x, and 4x. Because the distance calculated from d^{RA} already includes the evolutionary distance model in the calculation (the term $-\frac{3}{4}ln\left(1-\frac{4}{3}d(a,b)\right)$ in Eq. (2)), this evolutionary distance model is applied to each alignment pair, not the entire genome sequence. To evaluate the accuracy of the method with true edit distance, we considered using just $d(a,b)$ in Eq. (2) instead of evolutionary distance model term.

The results, shown in **Fig. 8**, are the estimated distances from d^{RA} between simulated sequences and *E. coli O157* with different true edit distance and coverage. The x-axis is the true edit distance between simulated sequences and *E. coli O157* sequence.

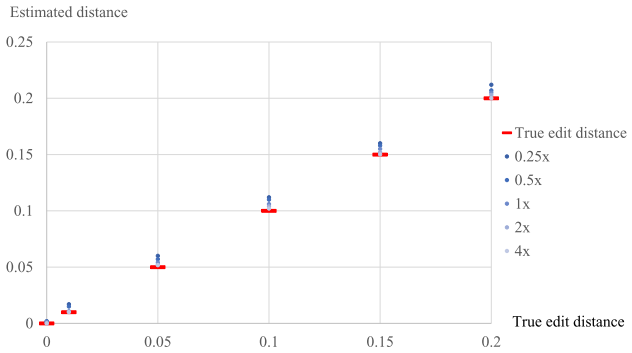

Fig. 8 Comparison of distance calculated by d^{RA} and true edit distance.

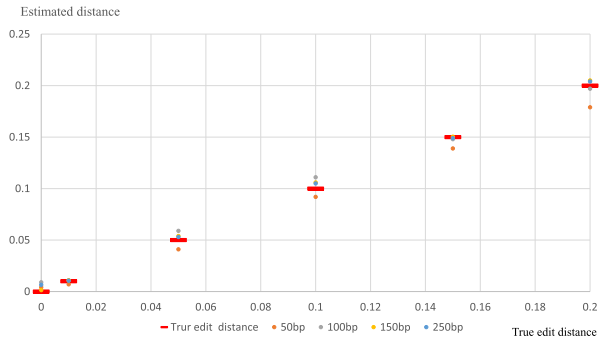

Fig. 9 The comparison of distance calculated by d^{RA} and true edit distance w.r.t. short read length.

Table 4 Average RF distance between benchmark tree and phylogenetic trees constructed from NGS short read sets w.r.t. the edit distance cost.

	mammalian mtDNA (5x)	*Escherichia /Shigella* (1x)	*Drosophila* (0.1x)
d^{RA} : Uniform $(1,1,1)$	3.75	10	10
d^{RA} : Hamming $(1,0,0)$	6	11	10
d^{RA} : $(2,1,1)$	10.25	11.5	10.25
d^{RA} : $(1,2,2)$	8	11.25	10
$Skmer(k=13)$	5	15.25	11
$Skmer(k=31)$	18.5	12.75	10

According to Fig. 8, d^{RA} can be used to estimate accurate distances between NGS sets of simulated sequence and the *E. coli O157* sequence. However, the coverage affects the estimated distance calculated by d^{RA}.

Figure 9 shows that the length of the short reads affect the distance calculation of d^{RA}. In this experiments we also compared the simulated sequences which are set to have the normalized edit distance equal to 0, 0.01, 0.05, 0.1, 0.15 and 0.2 compared with *E.coli O157* sequence as "true edit distance." Figure 9 shows the distance result of calculating distance between NGS sets of simulated sequence and *E.coli O157* sequence with different short reads length. With the short reads length of 50 bp, d^{RA} tends to calculate the distance lower than the true edit distance while the longer length provide the distance close to true edit distance. Although, the estimated distance are less accurate when reads length is short, the phylogenetic tree constructed from those result still provide good tree as shown in Fig. 6.

The proposed method uses the unit cost edit distance to measure the distance d^{RA} among short read sets. However the costs may affect the resultant trees. We evaluated the accuracy of resultant trees with several costs. For the three datasets, **Table 4**

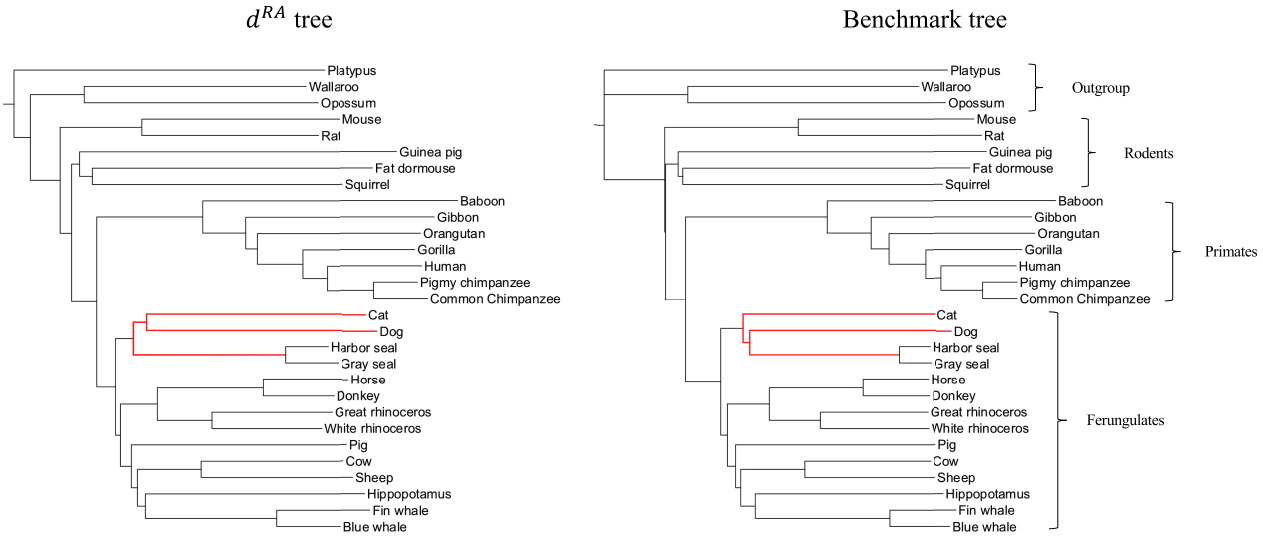

Fig. 10 The comparison of phylogeny tree of 29 *mammalian mtDNA* between d^{RA} tree (left) and the benchmark tree (Right).

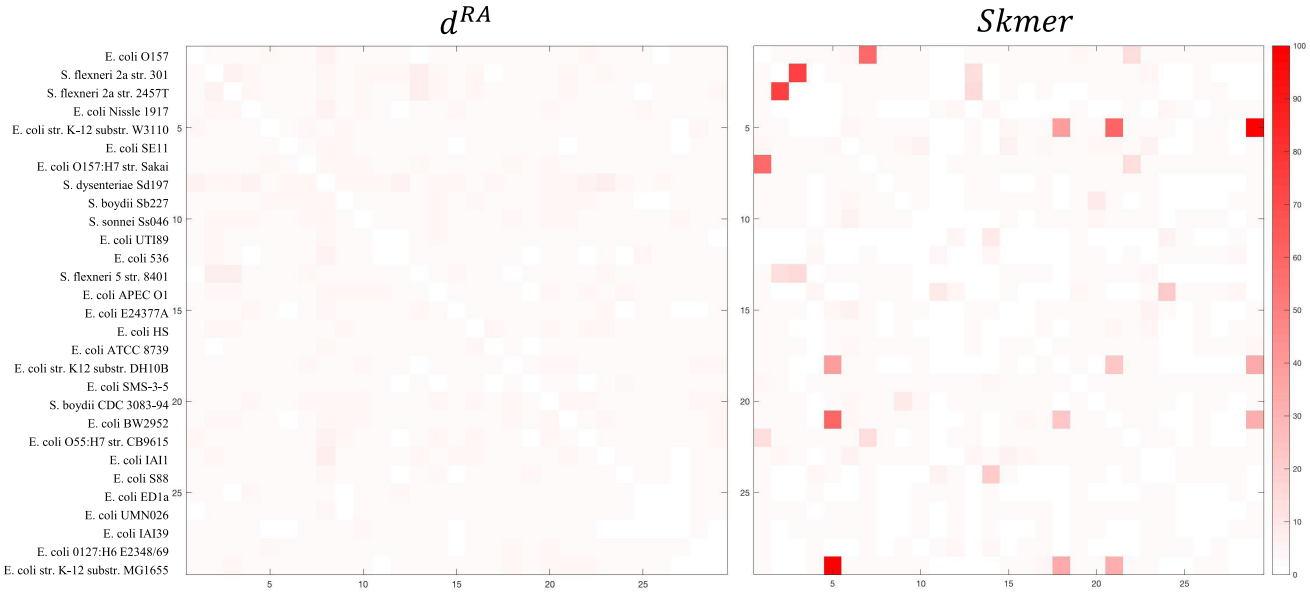

Fig. 11 A heatmap showing the value of the coefficient of variation for each pair-wise distance on multiple NGS sets of the *Escherichia/Shigella* dataset. Red refers to a high coefficient of variation and white is low.

shows the average RF distance between phylogenetic tree and trees constructed by the proposed method d^{RA} and the state-of-the-art *Skmer* with the best parameter k. We examined the accuracy for the costs of the unit cost $(1, 1, 1)$, hamming distance $(1, 0, 0)$, $(2, 1, 1)$, and $(1, 2, 2)$ where first (resp. second and third) component stands for the cost of substitution (resp. addition and deletion).

As we can see in Table 4, the edit cost affects the accuracy, especially for the data containing diverse species like *mammalian mtDNA*. However, the proposed method still constructs better tree than the state-of-the-art *Skmer* with the best k. Therefore we use the unit cost edit distance for measuring the distance of d^{RA}.

Figure 10 shows an example of a phylogenetic tree result for 29 *mammalian mtDNA* dataset. The tree that was reconstructed from the distance matrix calculated by d^{RA} is almost the same as the benchmark tree. According to the dataset, we can categorize

Table 5 The average coefficient of variation.

	mammalian mtDNA (5x)	*Escherichia /Shigella* (1x)	*Drosophila* (0.1x)
d^{RA}	3.62	2.56	2.74
$Skmer(k = 8)$	6.03	11.88	4.65
$Skmer(k = 13)$	4.11	5.01	3.54
$Skmer(k = 21)$	3.38	3.51	2.98
$Skmer(k = 31)$	2.03	3.26	1.76
$d_2^S(k = 8)$	24.60	55.74	2.35
$d_2^S(k = 13)$	4.83	18.89	1.75
$CVTree(k = 8)$	1.59	3.58	1.3
$CVTree(k = 13)$	1.21	1.39	0.82

the input species into four groups: *Primates*, *Ferunguletes*, *Rodents*, and *Outgroup*. d^{RA} was able to separate the 29 species into these four groups effectively. The only difference to the benchmark tree is the branch between cat and dog. In the d^{RA} tree, cat

Table 6 The runtime of each method for all three datasets (seconds).

	d^{RA}	$Skmer(k=8)$	$Skmer(k=31)$	$d_2^S(k=8)$	$d_2^S(k=13)$	$CVTree(k=8)$	$CVTree(k=13)$
mammalian mtDNA(5x:5 MB)	5	7	8	41	4812	3	3
Escherichia/Shigella(1x:300 MB)	78	26	42	87	5439	30	528
Drosophila(0.1x:700 MB)	147	1514	87	100	4153	63	812

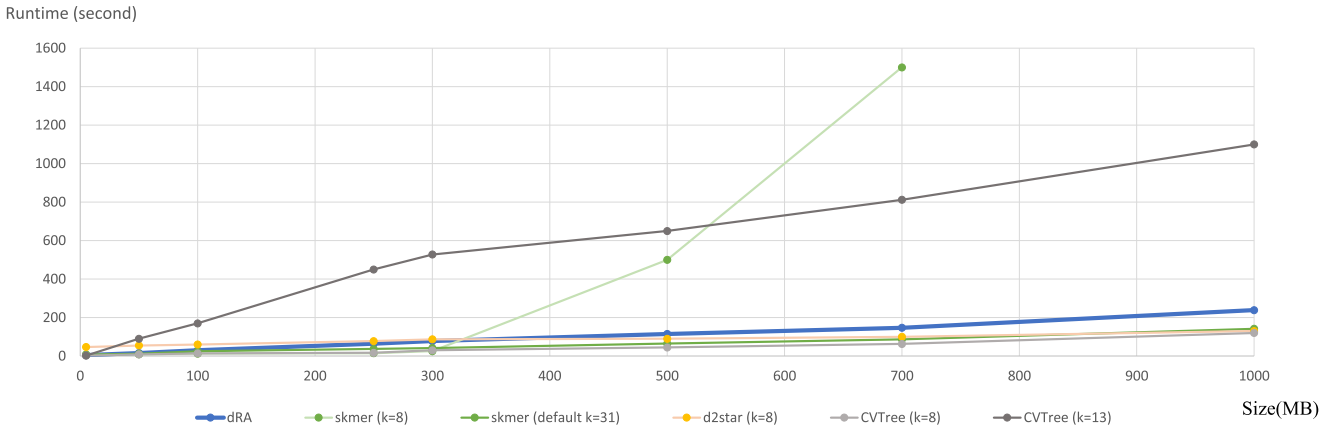

Fig. 12 The runtime of each method w.r.t. data size.

and dog are in a group apart from two seal species. However, in the benchmark tree, the cat is branched out from dog and seals. In this result, the distance measurement from d^{RA} between the cat and the group of dog and seals is not high enough to distinguish them.

5.3 Distance Consistency for Pair-wise Distance

For any dataset, the distance measurement between NGS sets should be almost the same every time, regardless of different NGS short reads. The consistency exposes the difference in distances among multiple NGS sets in the same dataset. Even though the accuracy of the phylogenetic tree reconstruction is an important aspect of evaluating the methods, without consistency, the accuracy is not convincing. Therefore, we also conducted experiments to evaluate the consistency of the distance measurement. We used the coefficient of variation to evaluate the consistency of the methods. **Figure 11** presents a heatmap of the coefficient of variation for each element in the distance matrices calculated from multiple NGS sets in the *Escherichia/Shigella* dataset. In the figure, d^{RA} is compared with the *skmer* ($k = 31$) because it provided RF distance results similar to those of d^{RA} in the accuracy evaluation shown in Table 3. According to Fig. 11, d^{RA} provides a lower coefficient of variation in most of the elements in the distance measurement while *skmer* ($k = 31$) reveals a very high coefficient of variation of distance between some pairs in the *Escherichia/Shigella* dataset despite the good RF distance results.

We evaluated the difference of pair-wise distances computed by the distance matrices in our method d^{RA}, *CVTree*, d_2^S, and *skmer* using different simulated NGS sets of each dataset. **Table 5** shows the average coefficient of variation values of all pairs in the distance matrices. d^{RA} provided a relatively low value of the coefficient of variation compared with the other methods. Thus, it can estimate the distances with not much difference between NGS sets. In this respect, although *CVTree* can calculate the most consistent result, it provided the worst accuracy. When considering the accuracy along with the consistency, d^{RA} reveals the effectiveness of distance measurement with NGS short reads data. d^{RA} provides the closest phylogenetic tree to the benchmark while maintaining the consistency with low coefficient of variation value compared with the other methods.

5.4 Efficiency Evaluation

We compared the runtime of our proposed method with the others using all three datasets with different sizes. The 29 *mammalian mtDNA* with 5x coverage, 29 *Escherichia/Shigella* with 1x coverage, and 18 *Drosophila* with 0.1x coverage have data sizes of 5, 300, and 700 MB, respectively.

For d^{RA}, we ran experiments with a query size of 100 MB for the *Drosophila* dataset. The runtime results are shown in **Table 6**. We observed that d^{RA} could calculate the distance between NGS sets as fast as the alignment-free approaches, although it is based on the alignment among short reads. In k-mer-based methods, the computational time varied by k parameter. The bigger the k value, the longer the time required for the distance calculations. d_2^S showed a huge difference between $k = 8$ and $k = 13$, as did *CVTree*. With the k-free approach, d^{RA} does not require additional calculations to tune the k parameter, thus it provides an accurate phylogenetic tree within a reasonable processing time.

In some cases, d^{RA} runs slower than the other methods. However, in such cases, d^{RA} offers much better phylogenetic tree results. Therefore, it is a worthy trade-off between efficiency and effectiveness. For instance, although d_{RA} is three times slower than *skmer* with $k = 8$ on the *Escherichia/Shigella* dataset, the resultant phylogenetic tree result obtained by *skmer* is more different to the benchmark than d^{RA} for 4 times.

Figure 12 shows how the runtime increases with respect to the data size in comparison with the other methods. Most of the methods showed linear d_{RA} growth according to the data size. However, the k parameter significantly affects the runtime of *skmer*, d_2^S, and *CVTree*. For *skmer*, lower k requires a larger number of k-mers to be considered in the distance calculation. On the other hand, larger k results in a larger dimension k-mer

profile for d_2^S and $CVTree$. This result also shows the advantage of the k-parameter-free method.

6. Conclusion

In this paper, we proposed the k-free approach d^{RA} for NGS data sequence comparison effectively to reconstruct accurate phylogenetic trees and measure the distance between reconstructed trees and benchmark trees. d^{RA} is a novel approach that lies between alignment-based and alignment-free approaches. The d^{RA} distance measurement is based on the collection of alignment between unassembled NGS short reads pairs. While taking advantage of the accuracy aspect of the alignment method, d^{RA} can be performed without an assembly process and can avoid the computational cost associated with assembling and aligning long sequences. The empirical results show that d^{RA} is capable of reconstructing accurate phylogenetic trees without the k parameter even with low coverage data. Although some results obtained at runtime are worse than some other alignment-free methods, there is a fair trade-off with respect to the accuracy without the ambiguous k parameter tuning in the practical use of the method.

References

[1] Metzker, M.L.: Sequencing technologies–the next generation, *Nature reviews. Genetics*, Vol.11, No.1, p.31 (2010).

[2] Waterman, M.S.: *Introduction to computational biology: Maps, sequences and genomes*, CRC Press (1995).

[3] Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G.: *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*, Cambridge University Press (1998).

[4] Vinga, S. and Almeida, J.: Alignment-free sequence comparison — a review, *Bioinformatics*, Vol.19, No.4, pp.513–523 (2003).

[5] Phillips, A., Janies, D. and Wheeler, W.: Multiple sequence alignment in phylogenetic analysis, *Molecular Phylogenetics and Evolution*, Vol.16, No.3, pp.317–330 (2000).

[6] Chan, C.X. and Ragan, M.A.: Next-generation phylogenomics, *Biology Direct*, Vol.8, No.1, p.3 (2013).

[7] Xu, Z. and Hao, B.: CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes, *Nucleic Acids Research*, Vol.37, No.suppl_2, pp.W174–W178 (2009).

[8] Qi, J., Luo, H. and Hao, B.: CVTree: a phylogenetic tree reconstruction tool based on whole genomes, *Nucleic Acids Research*, Vol.32, No.suppl_2, pp.W45–W47 (2004).

[9] Song, K., Ren, J., Zhai, Z., Liu, X., Deng, M. and Sun, F.: Alignment-free sequence comparison based on next-generation sequencing reads, *Journal of Computational Biology*, Vol.20, No.2, pp.64–79 (2013).

[10] Sarmashghi, S., Bohmann, K., Gilbert, M.T.P., Bafna, V. and Mirarab, S.: Skmer: assembly-free and alignment-free sample identification using genome skims, *Genome Biology*, Vol.20, No.1, p.34 (2019).

[11] Higgins, D.G. and Sharp, P.M.: CLUSTAL: A package for performing multiple sequence alignment on a microcomputer, *Gene*, Vol.73, No.1, pp.237–244 (1988).

[12] Thompson, J.D., Higgins, D.G. and Gibson, T.J.: CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Research*, Vol.22, No.22, pp.4673–4680 (1994).

[13] Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al.: Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, *Molecular Systems Biology*, Vol.7, No.1, p.539 (2011).

[14] Notredame, C., Higgins, D.G. and Heringa, J.: T-Coffee: A novel method for fast and accurate multiple sequence alignment, *Journal of Molecular Biology*, Vol.302, No.1, pp.205–217 (2000).

[15] Edgar, R.C.: MUSCLE: Multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Research*, Vol.32, No.5, pp.1792–1797 (2004).

[16] Wang, A. and Ash, G.J.: Whole genome phylogeny of Bacillus by feature frequency profiles (FFP), *Scientific Reports*, Vol.5, p.13644 (2015).

[17] Leimeister, C.-A. and Morgenstern, B.: Kmacs: The k-mismatch average common substring approach to alignment-free sequence comparison, *Bioinformatics*, Vol.30, No.14, pp.2000–2008 (2014).

[18] Leimeister, C.-A., Boden, M., Horwege, S., Lindner, S. and Morgenstern, B.: Fast alignment-free sequence comparison using spaced-word frequencies, *Bioinformatics*, Vol.30, No.14, pp.1991–1999 (2014).

[19] Gardner, S.N. and Hall, B.G.: When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes, *PloS One*, Vol.8, No.12, p.e81760 (2013).

[20] Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S. and Phillippy, A.M.: Mash: Fast genome and metagenome distance estimation using MinHash, *Genome Biology*, Vol.17, No.1, p.132 (2016).

[21] Fan, H., Ives, A.R., Surget-Groba, Y. and Cannon, C.H.: An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data, *BMC Genomics*, Vol.16, No.1, p.522 (2015).

[22] Stuart, G.W., Moffett, K. and Leader, J.J.: A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes, *Molecular Biology and Evolution*, Vol.19, No.4, pp.554–562 (2002).

[23] Miyazawa, S.: A reliable sequence alignment method based on probabilities of residue correspondences, *Protein Engineering, Design and Selection*, Vol.8, No.10, pp.999–1009 (1995).

[24] Dempster, A.P., Laird, N.M. and Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society: Series B* (*Methodological*), Vol.39, No.1, pp.1–22 (1977).

[25] Bilmes, J.A. et al.: A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models, *International Computer Science Institute*, Vol.4, No.510, p.126 (1998).

[26] Otu, H.H. and Sayood, K.: A new sequence distance measure for phylogenetic tree construction, *Bioinformatics*, Vol.19, No.16, pp.2122–2130 (2003).

[27] Cao, Y., Janke, A., Waddell, P.J., Westerman, M., Takenaka, O., Murata, S., Okada, N., Pääbo, S. and Hasegawa, M.: Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders, *Journal of Molecular Evolution*, Vol.47, No.3, pp.307–322 (1998).

[28] Zhou, Z., Li, X., Liu, B., Beutin, L., Xu, J., Ren, Y., Feng, L., Lan, R., Reeves, P.R. and Wang, L.: Derivation of Escherichia coli O157: H7 from its O55: H7 precursor, *PloS One*, Vol.5, No.1, p.e8700 (2010).

[29] Huang, W., Li, L., Myers, J.R. and Marth, G.T.: ART: A next-generation sequencing read simulator, *Bioinformatics*, Vol.28, No.4, pp.593–594 (2011).

[30] Felsenstein, J.: *PHYLIP* (*Phylogeny Inference Package*), *version 3.5 c*, Joseph Felsenstein (1993).

[31] Saitou, N. and Nei, M.: The neighbor-joining method: A new method for reconstructing phylogenetic trees, *Molecular Biology and Evolution*, Vol.4, No.4, pp.406–425 (1987).

[32] Robinson, D.F. and Foulds, L.R.: Comparison of phylogenetic trees, *Mathematical Biosciences*, Vol.53, No.1-2, pp.131–147 (1981).

[33] Tran, N.H. and Chen, X.: Comparison of next-generation sequencing samples using compression-based distances and its application to phylogenetic reconstruction, *BMC Research Notes*, Vol.7, No.1, p.320 (2014).

[34] Hinchliff, C.E., Smith, S.A., Allman, J.F., Burleigh, J.G., Chaudhary, R., Coghill, L.M., Crandall, K.A., Deng, J., Drew, B.T., Gazis, R., Gude, K., Hibbett, D.S., Katz, L.A., Laughinghouse, H.D., McTavish, E.J., Midford, P.E., Owen, C.L., Ree, R.H., Rees, J.A., Soltis, D.E., Williams, T. and Cranston, K.A.: Synthesis of phylogeny and taxonomy into a comprehensive tree of life, *Proc. National Academy of Sciences*, Vol.112, No.41, pp.12764–12769 (online), DOI: 10.1073/pnas.1423041112 (2015).

[35] Sessegolo, C., Burlet, N. and Haudry, A.: Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies, *Biology Letters*, Vol.12, No.8, p.20160407 (2016).

[36] Edgar, R.C.: Search and clustering orders of magnitude faster than BLAST, *Bioinformatics*, Vol.26, No.19, pp.2460–2461 (2010).

[37] Everitt, B. and Skrondal, A.: *The Cambridge Dictionary of Statistics*, Cambridge University Press (2010).

Phanucheep Chotnithi recieved B.S. from Chulalongkorn University, Thailand in 2014. He is graduate student of SOKENDAI (The Graduate University for Advanced Studies), Japan. His research interests is bioinformatics.

Hong Van Le received B.S. from Hanoi University of Technology in 2011 and Dr.Eng. from SOKENDAI (The Graduate University for Advanced Studies) in 2019. She is a researcher of National Institute of Informatics, Japan. Her research interests are key-value stores, spatio-temporal databases, data mining.

Atsuhiro Takasu received B.E., M.E. and Dr.Eng. from the University of Tokyo in 1984, 1986 and 1989, respectively. He is a professor of National Institute of Informatics, Japan. His research interests are data engineering and data mining. He is a member of ACM, IEEE, IEICE, IPSJ and JSAI.

(Editor in Charge: *Yasunori Ishihara*)