SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
PHYLOGENETICS
COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

Received
6 March 2014

Accepted
10 September 2014

Published
30 September 2014

Inferring phylogenies of evolving sequences without multiple sequence alignment

Cheong Xin Chan[1], Guillaume Bernard[1], Olivier Poirion[1]*, James M. Hogan[2] & Mark A. Ragan[1]

[1]Institute for Molecular Bioscience, and ARC Centre of Excellence in Bioinformatics, The University of Queensland, Brisbane, QLD 4072, Australia, [2]School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD 4000, Australia.

Correspondence and requests for materials should be addressed to M.A.R. (m.ragan@uq.edu.au)

* Current address: Laboratoire Ampère, CNRS UMR 5005, École Centrale de Lyon, France.

Alignment-free methods, in which shared properties of sub-sequences (e.g. identity or match length) are extracted and used to compute a distance matrix, have recently been explored for phylogenetic inference. However, the scalability and robustness of these methods to key evolutionary processes remain to be investigated. Here, using simulated sequence sets of various sizes in both nucleotides and amino acids, we systematically assess the accuracy of phylogenetic inference using an alignment-free approach, based on D_2 statistics, under different evolutionary scenarios. We find that compared to a multiple sequence alignment approach, D_2 methods are more robust against among-site rate heterogeneity, compositional biases, genetic rearrangements and insertions/deletions, but are more sensitive to recent sequence divergence and sequence truncation. Across diverse empirical datasets, the alignment-free methods perform well for sequences sharing low divergence, at greater computation speed. Our findings provide strong evidence for the scalability and the potential use of alignment-free methods in large-scale phylogenomics.

Multiple sequence alignment (MSA) has long been a standard stage in phylogenetic workflows[1,2]. In this approach, homologous sequences are first multiply aligned along their full length, yielding positional hypotheses of homology (alignment columns) that are input to maximum parsimony, maximum likelihood (ML) or Bayesian inference, or summarised in a distance matrix and used to compute a tree e.g. by neighbour-joining (NJ). A key assumption of MSA is that in each such set of sequences, homologous positions occur in the same order relative to one another. This is not fully realistic, as genes and genomes are subject to recombination, rearrangement and lateral genetic transfer[3–5]. In sequences so affected, the positional hypothesis of homology generated by MSA will be incomplete or incorrect, diffusing the phylogenetic signal, violating models of the substitution process across sites and branches, and consequently misleading phylogenetic inference[6,7]. These issues can only be intensified by the on-going deluge of sequencing data arising from advances in sequencing technologies[8].

An alternative to MSA in phylogenetic inference is the so-called *alignment-free* approach in which pairwise similarity is computed from sub-sequences, e.g. counts of exact (or inexact) sub-sequences of defined length, or by extension, of conserved sequence patterns[9,10], or alternatively of match lengths[11]. These sub-sequences are known variously as words, *k*-mers or *n*-grams[12]; see refs. 13–15 for recent reviews. A word-count approach for alignment-free sequence comparison uses the D_2 statistic[15–18]. A D_2 score is calculated based on the exact count of shared *k*-mers between any two sequences, thus representing the extent of similarity they share (see Supplementary Note for details). Since the profile of *k*-mers depends on length of the sequence, modifications have been proposed to accommodate this bias, e.g. normalising the D_2 score by the probability of occurrence for each *k*-mer observed in the sequences (D_2^S), or by the mean and variance of *k*-mer occurrences (D_2^*)[17,18]. These studies have demonstrated that D_2^S and D_2^* have greater statistical power than D_2, and that this power increases with sequence length[15,17,18]. These statistics can be easily transformed into a pairwise measure of dissimilarity or distance, which can then be used to compute phylogenetic relationships.

Alignment-free approaches have been adopted in searches of sequence databases[19], clustering of expressed sequence tags[20], and more recently in detecting lateral genetic transfer[11]. By directly computing pairwise dissimilarity or distance using these methods, one can bypass resource-intensive ML or Bayesian approaches in favour of NJ. Some methods implementing approximate ML measures[21,22], although less accurate, are less resource-intens-

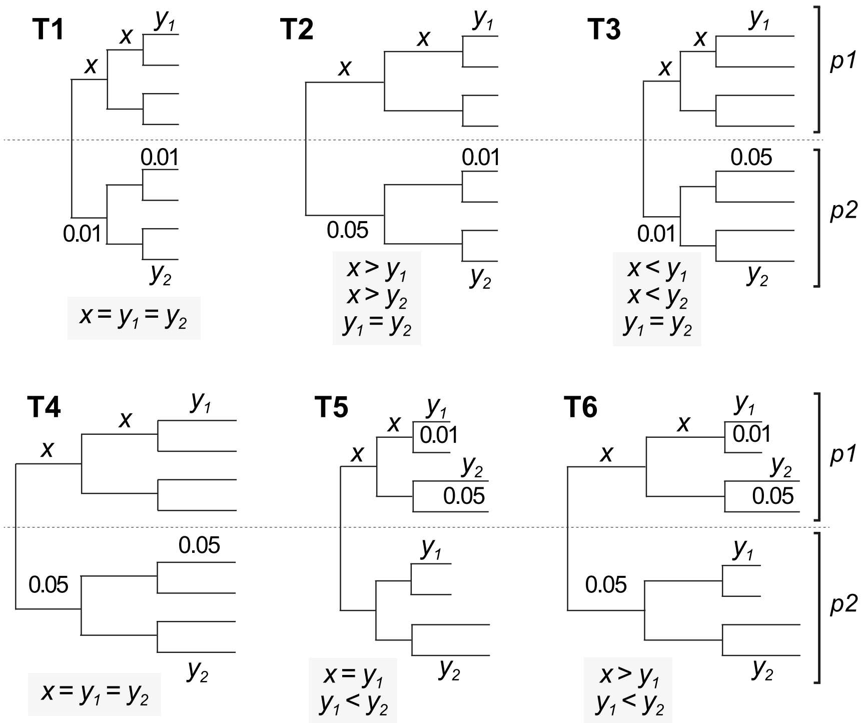

Figure 1 | Trees for simulation of sequence data. Six situations showing distinct combinations of internal (*x*) and terminal (*y*) branches, labelled as T1 through T6, with *y* specified differently between the first (*p1*) and second (*p2*) half of the branches on a tree. The unit of branch lengths is number of substitutions per site. The length of each edge is either 0.01 or 0.05 substitutions per site.

ive. However, the sensitivity of alignment-free methods to different evolutionary scenarios, and the scalability of these methods, have not been systematically investigated.

Here, using both simulated and empirical data we assess the accuracy of alignment-free phylogenetic approaches using D_2 statistics compared to standard MSA-based approaches. Using sets of simulated nucleotide and amino acid sequences, we systematically examine the accuracy and sensitivity of D_2 methods to key molecular evolutionary processes including sequence divergence, among-site rate heterogeneity, biases of G + C content, genetic rearrangements and insertions/deletions, as well as to the technical issue of incomplete sequence data. We demonstrate the scalability and potential of using alignment-free approaches to compute phylogenetic trees quickly and accurately from large-scale DNA or protein data.

Results

For our alignment-free phylogenetic approach, we used D_2 statistics (independently for D_2, D_2^S, D_2^*)[17,18] to generate a score for each possible pair of sequences within a set. Here we also introduce D_2^n, a D_2 statistic that extends each *k*-mer recovered in the sequences to its neighbourhood *n*, i.e. allows *n* number of wildcard residue(s). This simple extension of D_2 is analogous to generation of high-scoring words for the query phase of BLAST[23], and to a published alignment-free measure of sequence similarity[24]; a measure of inexact match has recently been extended to a position-specific context[25]. We denote cases of D_2^n where $n = 1$ as $D_2^{n=1}$ hereinafter. Each of these metrics is described in the Supplementary Note. For each method, we transform the scores *via* logarithmic representation of the geometric mean to estimate evolutionary distances (see Methods). Each resulting distance matrix was then used to calculate phylogenetic relationships using NJ. For comparison, for each sequence set we performed MSA using the popular tool, MUSCLE[26] and inferred a phylogenetic tree using the widely used MrBayes[27]. We use Robinson-Foulds distances[28] to evaluate topological congruence between each of the resulting test trees and a reference tree, normalised to adjust for different tree sizes (see Methods for details). We denote *RF* as the normalised Robinson-Foulds distance. $RF = 0$ indicates that the test tree shows complete topological congruence with the reference, while $RF = 1$ indicates that the test tree has no bipartition in common with the reference. The *RF* for a test tree generated *via* one of the four D_2 methods is denoted as RF_{D2}, RF_{D2S}, RF_{D2*} or RF_{D2n1}, and the equivalent for a test tree generated *via* MSA and MrBayes is denoted as RF_{MSA}.

Using simulated data, we independently assess the sensitivity of D_2 methods to variation in key evolutionary processes: sequence divergence, genetic rearrangement, and insertions/deletions. Because the phylogenetic tree is known for each simulated sequence set, we use that as the reference.

Sequence divergence. We simulated nucleotide sequence sets of various size categories $N = 8$, 32 and 128 (total length, $L = 1500$ nt). For each category, six sequence sets were simulated under an unrooted tree topology across distinct situations of relative branch lengths, with $\alpha = 1$ in an 8-category discrete gamma distribution. Each of these trees (T1 through T6 in Fig. 1; shown for 8-taxon trees) represents a fine-scale scenario of sequence divergence, as determined by different combinations of internal (*x*) and terminal (*y*) branch lengths. In some simulations, we recognise two subsets of *y* (y_1 and y_2) of different length. Sets containing varied divergence levels had different combinations of *x*, y_1 and y_2 as shown in T2, T3, T5 and T6; these are the reference trees for the corresponding sequence sets. For 32- and 128-taxon trees, the topologies were simply expanded for each upper and lower half, as indicated in Fig. 1 (labels *p1* and *p2*). For instance in a 128-taxon tree, the relative lengths (*x*, y_1, y_2) of the first 64 taxa follow pattern *p1*, while the others follow *p2*. For simplicity, *x* and *y* (or y_1 and y_2) were set at either 0.01 or 0.05 (unit in number of substitutions per site). The least-divergent (most-similar) sequence set (T1) was simulated with all branch lengths $x = y_1 = y_2 = 0.01$ (two most dissimilar sequences differ at 0.14 substitutions per site at $N = 128$), whereas the most-divergent (most-dissimilar) set (T4) had $x = y_1 = y_2 = 0.05$ (two

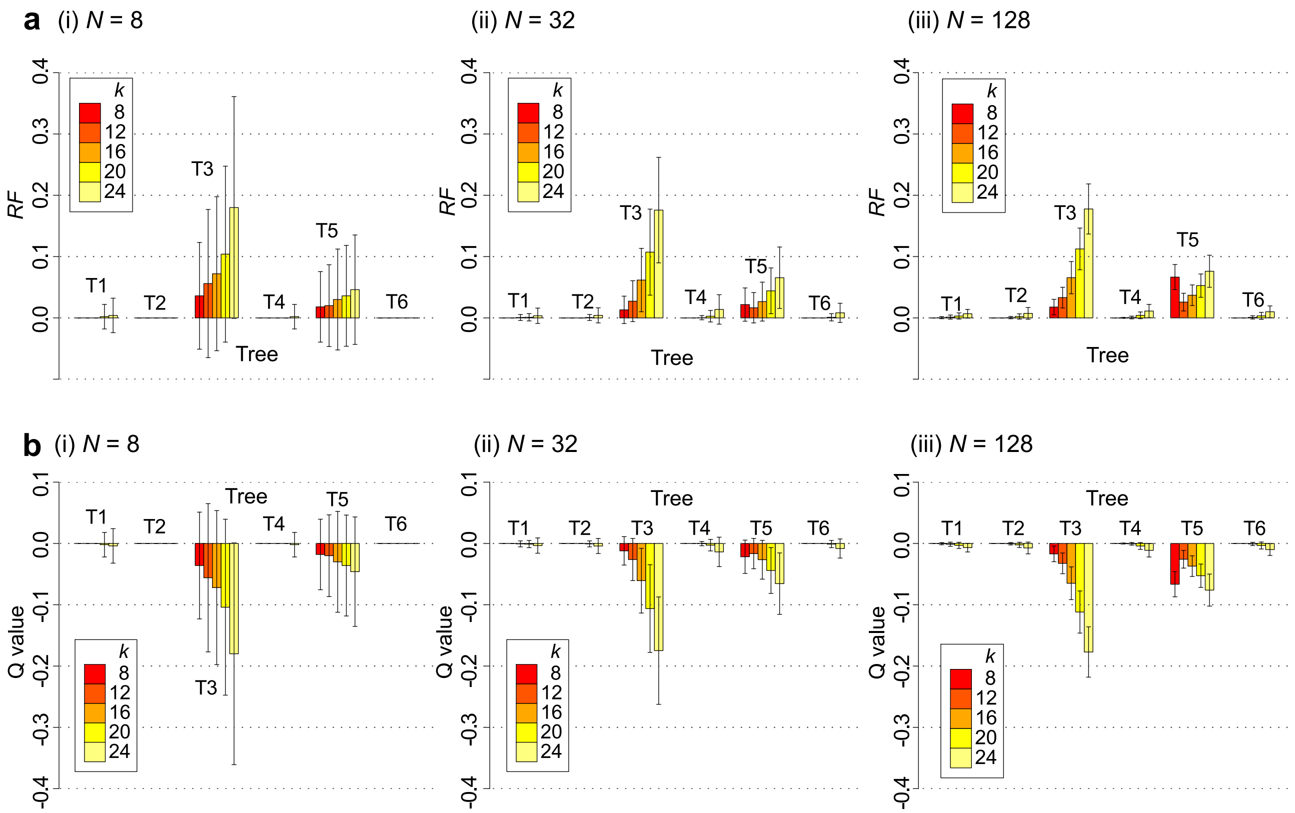

Figure 2 | The accuracy of D_2 methods based on sequence divergence of the nucleotide sequence sets. For each size *N* at (i) 8, (ii) 32 and (iii) 128, mean RF_{D2n1} are shown in (a) across different *k*-mer lengths (shown for *k* = 8, 12, 16, 20, 24), for cases simulated under each of the six trees (T1 through T6 on the *x*-axis). The corresponding Q_{D2n1} for each case is shown in (b). Error bars indicate standard deviation from the mean. See Supplementary Figures S1 through S4 for complete results for all D_2 methods for both nucleotide and protein sequence sets.

most dissimilar sequences differ at 0.70 substitutions per site at $N = 128$). The branch lengths in all these trees are short (two most dissimilar sequences in any set differ at <0.70 substitutions per site), so any MSA-based approaches should have no problem recovering these phylogenies. However, these datasets provide a testable range of sequence divergence to assess the sensitivity of alignment-free methods in recovering the topologies. For each sequence set, we independently derived pairwise distances using D_2, D_2^S, D_2^* and $D_2^{n=1}$, in each case across different *k*-mer lengths (*k* = 4, 8, 12, 16, 20 and 24). Each parameter setting was run with 100 replicates, i.e. 100 × 3 size categories × 6 trees × 4 methods of D_2 statistics × 6 *k*-mer lengths (total of 43200 sequence sets). The same experimental design applies to protein sequences with fixed sequence length of 500 amino acids. See Methods for details.

To compare the performance between MSA-based and the D_2 methods, we denote a relative measure of accuracy $Q_{DX} = RF_{MSA} - RF_{DX}$, where *DX* represents any of the D_2 methods, i.e. Q_{D2} is the *Q* that corresponds to $RF_{MSA} - RF_{D2}$, and so forth. Derived from *RF*, the *Q* values reflect the proportion of bipartitions in a tree, and can be interpreted as the difference between the deviation of each tree from the common reference. The sign of the *Q* value indicates which of the two approaches performs better; if a D_2 method performs better than MSA in recovering the reference tree then $Q > 0$ (i.e. $RF_{MSA} > RF_{DX}$), whereas if a D_2 method performs worse than MSA then $Q < 0$ (i.e. $RF_{MSA} < RF_{DX}$). Where $Q = 0$ (i.e. $RF_{MSA} = RF_{DX}$) the D_2 method performs as well as the MSA-based approach, although the trees could still be incongruent with the reference (i.e. their *RF* could be non-zero).

Across all D_2 methods used in this study, we found that $D_2^{n=1}$ yielded the smallest *RF* across all categories of size and situations of relative branch length, for both nucleotide (Supplementary Fig. S1) and protein (Supplementary Fig. S2) sequence sets. Figure 2a shows mean RF_{D2n1} at different *k*-mer lengths (shown for $k \geq 8$) in each size category *N* of nucleotide sequence sets, across all trees (T1 through T6; Fig. 1), with the corresponding mean *Q* value shown in Fig. 2b. Across all *N*, $D_2^{n=1}$ recovered the reference topology almost perfectly for sets of sequences simulated under trees T1, T2, T4 and T6 (at $k = 16$, mean $RF_{D2n1} \leq 0.001$ across these sets and all *N*; Fig. 2a), whereas larger RF_{D2n1} distances are observed for cases of T3 and T5 (e.g. for $N = 32$ at $k = 16$, mean $RF_{D2n1} = 0.06$ and 0.03 respectively for T3 and T5; Fig. 2a). The accuracy decreased with increasing *k*, e.g. for $N = 128$ and T3, mean $RF_{D2n1} = 0.01, 0.03, 0.06, 0.11, 0.18$ at $k = 8$, 12, 16, 20 and 24.

While relative performance differed across the simulated scenarios, overall across these sequence sets we find that $D_2^{n=1}$ performed as well as the standard MSA-based approach (e.g. for T1 and T2 at $k = 8$, mean $Q_{D2n1} = 0.00$ in all cases of $N = 8$, 32 and 128; Fig. 2b), with the relative performance *Q* decreasing slightly with increased *k* (e.g. for $N = 32$ at T3, $Q = -0.01, -0.03, -0.06, -0.11, -0.17$ at $k = 8$, 12, 16, 20 and 24). Across all *N* examined here, $D_2^{n=1}$ performed slightly worse than MSA for T3 and T5, e.g. at $k = 8$, $Q_{D2n1} = -0.01$ and -0.02 respectively at $N = 32$; $Q_{D2n1} = -0.02$ and -0.07 respectively at $N = 128$. The bar plots in Fig. 2a almost mirror those in Fig. 2b, suggesting that $RF_{MSA} = 0$ in most cases. Both T3 and T5, the cases problematic for D_2 methods, have short internal branches (*x*) with long terminal branches (*y*: Fig. 1). Our results suggest that D_2 methods are more vulnerable to this situation, while the MSA-based approach performed well across these six cases. *Q* values observed for other D_2 methods across nucleotide and protein sequence sets are shown in Supplementary Fig. S3 and S4 respectively.

To assess the optimal *k*-mer length for use in D_2 methods in deducing phylogenetic relationships from nucleotide and protein

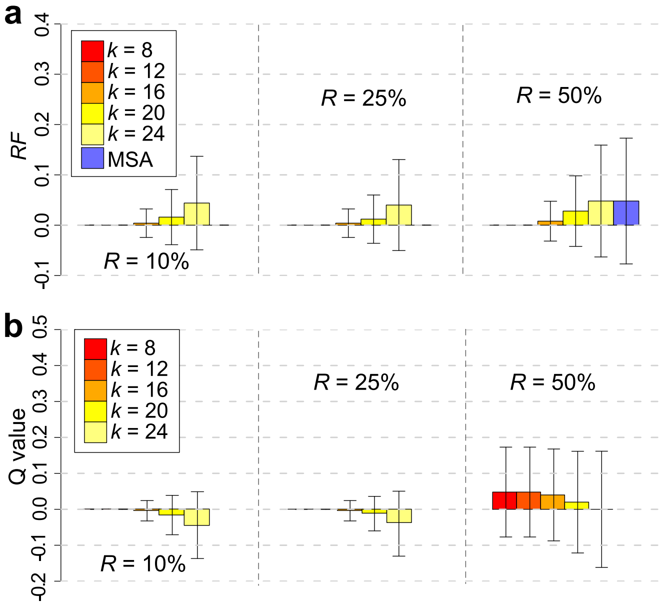

Figure 3 | The accuracy of D_2 methods based on genetic rearrangement. RF_{D2n1} are shown in (a) across different k-mer lengths ($k \geq 8$), as well as that of the standard approach (RF_{MSA}), across different R at 10%, 25% and 50%. The corresponding Q_{D2n1} values are shown in (b). Error bars indicate standard deviation from the mean.

sequences, we compared RF values from all D_2 methods between the two sequence types across $N = 8$, 32 and 128 pooled from all six trees, as shown in Supplementary Fig. S5. For nucleotide sequences, $k = 8$ yielded the lowest RF distances, with $RF = 0$ at $N = 8$ and 32, and $RF < 0.002$ at $N = 128$ across all D_2 methods. For protein sequences, $k = 4$ is the optimal length across all D_2 methods, with $D_2^{n=1}$ yielding the smallest RF distances across all size categories, i.e. $RF_{D2n1} = 0.012$, 0.009 and 0.009 at $N = 8$, 32 and 128. This result supports the notion that optimal k is negatively correlated with alphabet size of the sequence data[9,29,30]. Formal proof appears to be lacking, but might be approached analogously to an earlier study[31].

Two other scenarios relevant to sequence divergence are among-site rate heterogeneity (the presence of fast- *versus* slow-evolving sequence regions), and compositional (G + C content) biases in the sequences. We examined the sensitivity of D_2 methods independently to each these scenarios (see Supplementary Note for detail). Overall, among-site rate variation does not appear to affect drastically the accuracy of either D_2 or MSA-based approaches ($Q = 0$ in most cases at optimal k in Supplementary Fig. S6); the RF values for all analyses of nucleotide and protein sequences are shown respectively in Supplementary Fig. S7 and S8. Interestingly, we note that high G + C proportion (thus low complexity of sequences) plays to the strength of local exact matches, rather than neighbourhood (non-exact) matches as allowed in $D_2^{n=1}$ (Supplementary Fig. S9).

Genetic rearrangement. Here we simulated sequence data to assess the direct impact of genetic rearrangement on the performance of D_2 methods in phylogenetic inference. We define R as the percentage length of a full-length nucleotide sequence that has undergone a non-overlapping rearrangement. We simulated post-hoc rearrangements in half of the sequences in a set of 5000-nt sequences, i.e. at $N = 8$, each of any 4 sequences would have R% of its length rearranged in a non-overlapping manner. Each rearrangement event involves one or more fragments of 250 nt, such that the total rearranged region (i.e. R% of full length) is no longer contiguous (see Methods). Figure 3a shows the average RF_{D2n1} for each k-mer length in nucleotide sequence sets ($N = 8$) across $R = 10$, 25 and 50%, including RF_{MSA} of the MSA-based approach MUSCLE + MrBayes. Across all categories and all k-mer sizes, all methods, alignment-free or not, yielded average $RF < 0.05$ compared to the reference tree. $D_2^{n=1}$ at $k = 8$ or 12 perfectly recovered the reference topologies ($RF_{D2n1} = 0$ in both cases) regardless of R. Figure 3b shows the mean Q values for each of these cases. At $R = 10$% and 25%, we observed $Q = 0$ for $k = 8$ and 12, i.e. $D_2^{n=1}$ performed as well as did the MSA-based approach in recovering the reference topologies. At $R = 50$%, the D_2 methods yielded higher accuracy than did MUSCLE + MrBayes ($Q > 0$ for all k-mer lengths). Compared to MUSCLE (Fig. 3), the use of MAFFT resulted in higher RF and Q values (Supplementary Fig. S10), thus lower accuracy ($p < 2.2 \times 10^{-16}$; see Supplementary Note). Our findings suggest that D_2 methods are more robust to the effect of genetic rearrangement than is the standard approach based on MSA.

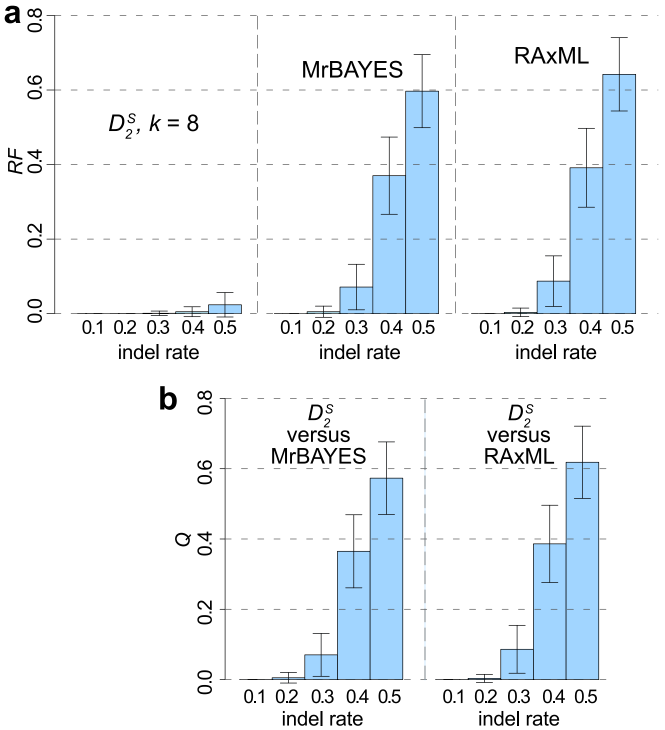

Figure 4 | The accuracy of phylogenetic approaches based on insertions/deletions. RF values are shown in (a) for D_2^S, MUSCLE + MrBayes and MUSCLE + RAxML across different indel rates r. The corresponding Q values for MUSCLE + MrBayes and MUSCLE + RAxML are shown in (b). Error bars indicate standard deviation from the mean.

Insertions/deletions. To assess the sensitivity of the alignment-free approach to insertions/deletions (indels) we simulated nucleotide sequence sets ($N = 32$) under tree T4 by incorporating indel events at a predefined rate (r) along the tree branches[32], with the inserted/deleted fragment lengths following a Lavalette distribution[33,34] (maximum length = 100 nt). Figure 4a shows the RF values obtained using D_2^S, two MUSCLE-based methods (MrBayes and the popular ML method RAxML[35,36]) across cases at different values of r; the corresponding Q values for each MSA-based approach are shown in Fig. 4b. At $r = 0.1$, all approaches recovered the reference topology perfectly ($RF = 0$ in all cases). As r increases, observed RF increases proportionately: for trees generated using D_2^S at $r = 0.3$, 0.4 and 0.5, $RF = 0.001$, 0.005 and 0.024. In comparison, the corresponding RF values for MSA-based methods are higher: $RF = 0.071$, 0.370 and 0.597 for MUSCLE + MrBayes and $RF = 0.087$, 0.391, and 0.642 for MUSCLE + RAxML. These results suggest that alignment-free methods are more robust to insertions/deletions ($RF < 0.025$ at $r = 0.5$) than MSA-based approaches ($RF \geq 0.60$ at $r = 0.5$ in both cases), with all observed $Q \geq 0$ (e.g. $Q = 0.07$, 0.37 and

0.57 at $r = 0.3$, 0.4 and 0.5 for MUSCLE + MrBayes: Fig. 4b). Here the use of MAFFT instead of MUSCLE yielded lower *RF* and *Q* values, i.e. a higher accuracy of phylogenetic inference (Supplementary Fig. S11 *versus* Fig. 4; $p < 2.2 \times 10^{-16}$). These findings are consistent with our analysis of other insertion/deletion scenarios including vertically staggered deletions (Supplementary Note and Fig. S12), a (biologically not very realistic) scenario in which MSA is known to perform poorly[37]. Independently, we observed that the accuracy of D_2 methods decreases with increasing extent of sequence truncation, and increases proportionately with sequence length (Supplementary Note and Fig. S13).

Gene family evolution based on coalescence. Here we simulated nucleotide sequence sets under the coalescent model of gene family evolution (within a population)[38,39] across different fixed effective population sizes N_e (see Methods). The N_e parameter affects the overall population structure, thus branching patterns and branch lengths of a tree. Coalescent rate between two lineages is higher within a smaller population[40], thus a smaller N_e yields shorter branch lengths in a tree. All trees are asymmetric, and thus represent a more-realistic biological scenario. We note that the observed performance in this part of our analysis could be affected by one or more scenarios in addition to N_e (and sequence divergence). Figure 5a shows the *RF* values obtained using $D_2^{n=1}$, and by MSA-based approaches using MUSCLE, across cases at varied N_e; the corresponding *Q* values for each MSA-based approach are shown in Fig. 5b. $RF > 0$ was observed across all cases, suggesting that all approaches on average failed to recover known tree topologies perfectly. Observed *RF* values for all approaches increase proportionally with increasing N_e when $N_e \geq 100000$, e.g. for $D_2^{n=1}$, $RF = 0.072$, 0.119, 0.239 and 0.407 at $N_e =$ 100000, 250000, 500000 and 1000000 (Fig. 5a), suggesting an inverse relationship between N_e and the accuracies of these approaches in recovering the known tree topology. At $N_e = 10000$, 100000 and 250000, both $D_2^{n=1}$ and MSA-based approaches yielded almost identical trees (e.g. $Q = -0.007$, -0.010, -0.016 against MUSCLE + RAxML; Fig. 5b), although $D_2^{n=1}$ yielded less-accurate topologies ($Q < 0$). In the extreme cases of $N_e > 250000$, $D_2^{n=1}$ performed substantially worse than any of the two MSA-based methods, e.g. $Q = -0.146$ and -0.279 for MUSCLE + RAxML (Fig. 5b). At the other end of the spectrum, cases of small $N_e = 1000$ also negatively impacted the accuracies of all approaches, i.e. $RF = 0.240$, 0.230 and 0.213 for $D_2^{n=1}$, MUSCLE + MrBayes and MUSCLE + RAxML (Fig. 5a). Results of the corresponding analysis using MAFFT are shown in Supplementary Figure S14 ($p = 0.74$; no significant difference). These findings indicate that in these scenarios, the alignment-free approach yields results similar to those of the MSA-based approaches, regardless of which MSA tool is used, when N_e is reasonably large, but performs substantially worse in extreme cases i.e. when N_e is very small or very large. This observation is plausibly explained by extreme (high/low) sequence divergence (See Supplementary Table S1), although we cannot rule out the impact of other evolutionary scenarios. In an independent analysis across datasets that were simulated under non-ultrametric trees (specifically violating the molecular clock) we observed a similar trend ($RF > 0$; $Q < 0$), with higher *RF* observed for $D_2^{n=1}$ than for MSA-based approaches (Supplementary Fig. S15). This complex scenario is more realistic than ultrametric trees, but we cannot distinguish the effect of clock violation from that of other evolutionary processes.

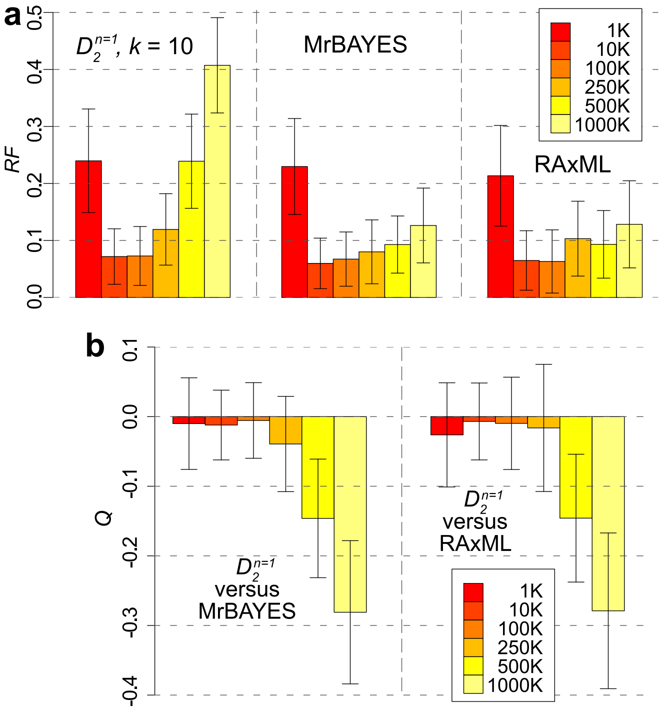

Figure 5 | The accuracy of phylogenetic approaches based on coalescent evolution of gene families. *RF* values are shown in (a) for $D_2^{n=1}$, MUSCLE + MrBayes and MUSCLE + RAxML across different effective population size N_e. The corresponding *Q* values for MUSCLE + MrBayes and MUSCLE + RAxML are shown in (b). Error bars indicate standard deviation from the mean.

Analysis of empirical data. To examine the performance of these methods with empirical data, we used 4156 sets of nucleotide sequences and their corresponding phylogenetic trees from TreeBASE (treebase.org)[41]. These sequence sets and trees were obtained from 2471 studies deposited in TreeBASE as of 27 May 2013 (see Supplementary Data for the complete list). As shown in Supplementary Fig. S16, the sizes of these sequence sets range between 6 and 2957 sequences (mean 59.41, median 41 sequences), and within-set sequence similarity has a mean of 90.12% (median 92.37%). For each sequence set, we used each of the D_2 methods (independently for $k = 6$ and 8) to generate a distance matrix, from which we reconstructed a NJ tree. The selection of *k* is based on our observation of an optimal length in the analysis of simulated nucleotide sequence sets (Supplementary Fig. S5). Because the true reference tree is unknown for empirical datasets, we cannot readily assess accuracy. Here we compare each of our resulting test trees inferred using the D_2 methods against the corresponding tree published (and peer-reviewed) in TreeBASE. Because we cannot assume that published trees perfectly reflect true evolutionary relationships, we intentionally do not interpret *RF* as a measure of accuracy here, but instead simply as a measure of (dis)agreement between the trees produced by an alignment-free and an MSA-based approach.

As shown in Supplementary Table S2, the use of $k = 6$ *versus* 8 does not impact *RF* for any D_2 method, with D_2^S yielding the smallest average *RF* (0.438; median 0.409 at $k = 8$). Figure 6 shows the distribution density of *RF* as observed for D_2^S at $k = 8$, based on sizes of the sequence sets *N* (Fig. 6a) and within-set sequence similarity (Fig. 6b). See Supplementary Tables S3 and S4 respectively for the corresponding values. As shown in Fig. 6a and Supplementary Table S3, D_2^S yielded topologies that are more congruent with those generated using the standard MSA approach for small sequence sets (e.g. mean *RF* 0.363, median 0.333 at $N \leq 25$) than for larger sequence sets of $N > 25$ (mean *RF* 0.661, median 0.635 at $N > 500$), and these *RF* distances increase proportionately with increasing *N*. Interestingly, across different categories of within-set sequence similarity (percent identity; *ID*) regardless of *N* (Fig. 6b), density plots of *RF* for cases of $ID > 70\%$ peak at values of *RF* between 0.25 and 0.40, with the

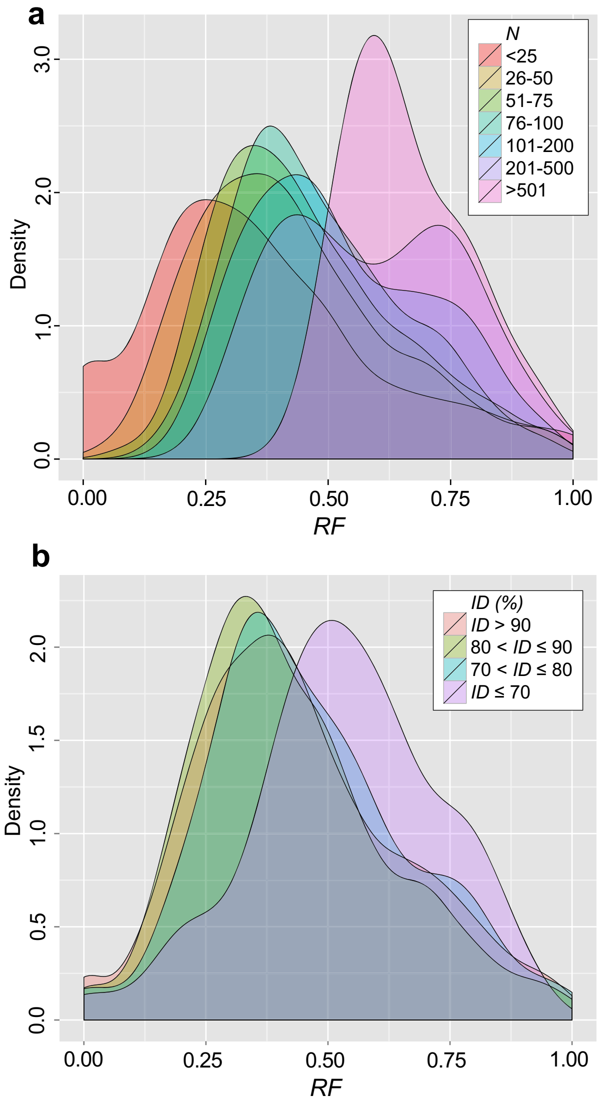

Figure 6 | The accuracy of D_2 methods based on TreeBASE data. The probability density of RF_{D2n1} at $k = 8$ as categorised based on (a) total number of sequences within a set, N (mean and median in Supplementary Table S3), and (b) within-set sequence similarity, ID (mean and median in Supplementary Table S4).

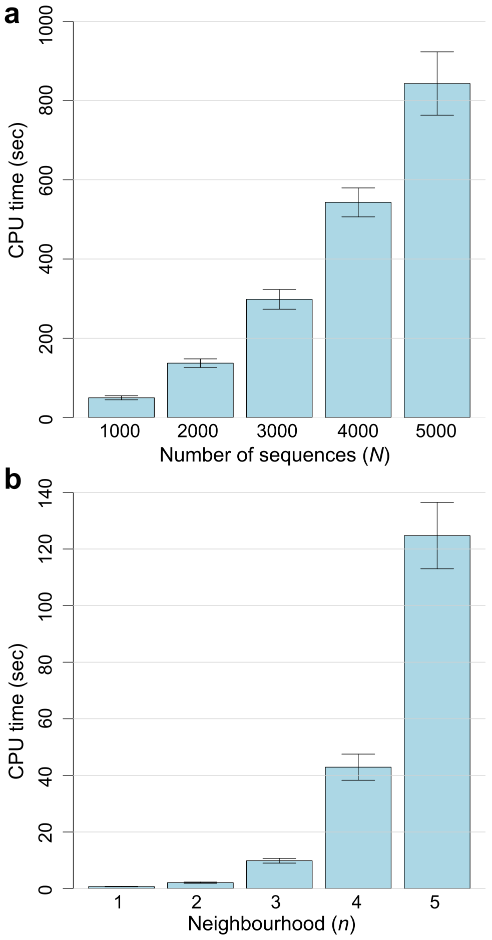

Figure 7 | Computation time of D_2 methods. The computation time in seconds is shown for (a) D_2 method at $k = 8$ across subset of GreenGenes data across datasets of $N = 1000, 2000, 3000, 4000$ and 5000, and for (b) D_2^n analysis across neighbourhood size $n = 1$ through 5, for nucleotide sequence sets of $N = 8$. Error bars indicate standard deviation from the mean.

smallest means observed for highly similar sequence sets (0.424 at ID between 80% and 90%, median 0.392; Supplementary Table S4). RF values increase with decreasing ID, with mean RF 0.533, median 0.528 observed for cases of $ID < 70\%$ (Supplementary Table S4). These findings suggest that the D_2-based approach, across most of these diverse empirical data, yield topologies that are slightly incongruent ($RF < 0.5$ in 2809/4156 trees; D_2^S at $k = 8$) to those arising from the standard MSA-based approach, and that it is rare for both approaches to recover the exact same tree topology ($RF = 0$ recovered by any D_2-based approach in 106/4156 trees).

Computational efficiency and scalability. The computational complexity of various D_2 methods has been described earlier[24] (see also Supplementary Note). Figure 7a shows the computation time required to generate pairwise D_2 distance matrices across large empirical sequence sets ($N = 1000, 2000, 3000, 4000$ and 5000); for the corresponding numerical values see Supplementary Table S5. These large sequence sets are of 16S ribosomal RNA genes sampled from the GreenGenes database (see Methods). Mean computation time increases with N, from 49.77 seconds at $N = 1000$ to 842.98 seconds at $N = 5000$ (17-fold increase). Similarly, memory usage (Supplementary Table S5) increases with N, from 378.24 MB ($N = 1000$) to 2445.31 MB ($N = 5000$; approximately 6-fold increase).

Phylogenetic inference involves details not only of software (e.g. D_2 and *neighbor* in PHYLIP *versus* MUSCLE and MrBayes) but also of parameter settings, implementation (e.g. programming language used, and capacity for multi-threading) and hardware (e.g. machine architecture and its efficiency of memory usage). Therefore, comparing computation time and memory usage between the two approaches is not straightforward. For 50 sets of nucleotide sequence ($N = 8$; $L = 1500$ nt), we observe an average wall time of 1.50, 86.38 and 491.16 seconds for D_2 + *neighbor*, MUSCLE + RAxML and MUSCLE + MrBayes (four-threaded runs; see Methods). For the same analysis across protein sequence sets ($N = 8$; $L = 500$aa), wall times are respectively 1.82, 255.48 and 3047.14 seconds. Here, our alignment-free approach is approximately 140-fold and 1670-fold faster respectively, compared to MUSCLE + RAxML and MUSCLE + MrBayes. These findings suggest that D_2 methods are highly scalable for phylogenetic inference of large-scale sequence data.

In an independent experiment on nucleotide sequence sets of $N = 8$ (Fig. 7b), we found that computation time for $D_2^{n=1}$ (at $k = 8$) increases exponentially with increasing neighbourhood n, from 0.71 at $n = 1$ to 124.73 seconds at $n = 5$. At greater values of neighbourhood ($n > 2$) i.e. when a higher number of wildcards is considered, the accuracy of D_2^n appears to decrease, more so at larger N (Supplementary Fig. S17; shown for $k = 8$ across nucleotide sequence sets). However, the interplay among n, k and N remains to be investigated systematically.

Discussion

Alignment-free methods yielded similar if not identical tree topologies to those generated using MSA-based approaches across a wide range of data sizes and scenarios. Our findings demonstrate that the accuracy of alignment-free methods, compared to the current standard based on MSA, is more robust against among-site rate heterogeneity, compositional biases, genetic rearrangements and insertions/deletions, but is more sensitive to sequence divergence and the presence of incomplete (truncated) sequence data. The alignment-free methods operated at far greater computation speed (more than 2000 times faster in some cases).

Opposing views have recently been expressed on whether the application of alignment-free methods in phylogenetics reflects a model-free, purely informatic exercise, or alternatively can capture homology signal inherent in evolving sequences[42–44]. Our results support the latter view. The alignment-free approach implemented here appears to have no difficulty, at appropriate parameter settings across our simulated datasets, in capturing homology signal and generating topologies that are very similar or identical to those generated by MSA followed by Bayesian inference, arguably the current standard in phylogenetics (see below). The robustness of alignment-free methods to rearrangements and insertions/deletions represents a critical advantage, since these events are common among microbial genomes[3] and frequently interrupt individual genes[45]. Our findings support the notion that gappy regions tend to be forced into alignment within an MSA framework and thereby bias subsequent phylogenetic inference[37].

Here we used MUSCLE[26] and MrBayes[27] as the standard phylogenetic approach in the analysis of simulated data. Another popular MSA tool is MAFFT[46]; both MUSCLE and MAFFT compare favourably against other MSA tools in a number of benchmark studies[26,47]. A comprehensive analysis of performance across different MSA tools is beyond the scope of this study. Across scenarios of random insertions/deletions, we found little difference in our inference between the use of MUSCLE and MAFFT ($p > 0.5$; Supplementary Table S6), except under the unrealistic scenarios of vertically staggered deletions (Supplementary Fig. S12; $p < 2.2 \times 10^{-16}$) in which MAFFT performed better, lending support to an earlier report[37]. The use of other programs for MSA and phylogenetic inference, or indeed the use of different parameter settings in these programs (e.g. fewer MCMC generations in MrBayes than the 1.5 million used in this study), would inevitably yield somewhat different results. ML is another popular MSA-based method of phylogenetic inference, which estimates goodness-of-fit of sequence data given an underlying evolutionary (substitution) model. ML methods e.g. RAxML[35] are time-consuming, and this has prompted the development of faster though less-accurate implementations e.g. PhyML[21] and/or scalable methods that approximate ML estimates e.g. FastTree[22] (see ref. 48 for a comparative analysis). We generated ML trees for a subset of the simulated sequence data using RAxML and found no or little topological difference between these trees and those generated using MrBayes, as shown by the similar trends of *RF* and *Q* in Figs 4 and 5. In fact, RAxML yielded less-accurate topologies than MrBayes in many cases (larger *RF* observed for RAxML: Fig. 5).

Using extensive simulated data and diverse empirical data (here from the TreeBASE dataset, generated by various programs and phylogenetic inference methods common in the peer-reviewed literature), our results consistently demonstrate the relative accuracy and scalability of alignment-free methods in large-scale phylogenetic inference, regardless of which specific method they were compared against. The empirical datasets used in this study are highly diverse, with various extents of within-set sequence divergence and data sizes. Many of these sequence sets contain partial and/or fragmented sequences (Supplementary Data). As per our analysis of simulated sequence sets, these aspects impact the accuracy of alignment-free methods more than that of MSA-based approach in recovering accurate phylogenies. In addition, we applied $k = 6$ and 8 in our alignment-free approach across these datasets, a decision based on our observation in simulated sets of 1500 nt sequences (Supplementary Fig. S5). In cases where sequences are longer, the representation of distinct k-mers (at $k = 6$ or 8) could be saturated, thus losing the resolution (reducing the distinguishing power of the k-mers) necessary to accurately infer dissimilarity (*vis-à-vis* phylogenetic) relationships among the sequences[9,30]. The correlation between sequence length and k within the context of phylogenetics has been explored to some extent[30,49], e.g. using shortest unique substrings[50], but this issue remains to be systematically investigated. In this study we used NJ to infer phylogenetic trees from the distance matrices generated from D_2 methods; one can imagine using other distance-based approaches, e.g. a weighted least-squares method such as Fitch-Margoliash[51]. In small-scale investigations, we find no topological difference across trees generated using NJ or Fitch-Margoliash.

Conversion of subsequence similarity (profile) scores into a measure that represents the evolutionary relatedness between two full-length sequences remains an active field of research. Here we simply transformed D_2 scores into pairwise distances of sequences using a logarithmic representation of the geometric mean. Other strategies have been proposed to create more-realistic measure of distance or dissimilarity, including the assignment of a p-value for each pairwise score based on a null distribution (hypothesis) of subsequences as observed across the whole dataset[29,52]. Approaches inspired by information retrieval are under consideration.

In general, our results demonstrate the utility and robustness of alignment-free methods across the choice of scoring methods. The non-monotonic relationship between word length and performance, the utility of D_2^S, D_2^* and $D_2^{n=1}$, and the failure of larger mismatch neighbourhoods are broadly consistent with previous reports[18,52]. However, simple D_2 scoring is known to be dominated by single-sequence noise effects as k increases[18]; its good performance here may in part be explained by the normalisation inherent in our distance measure. The one exception to these comments lies in the vulnerability of D_2^*-based approaches to heterogeneous variation, an effect especially pronounced for protein sequences (Supplementary Fig. S6), which may arise from the failure of the variance estimate in the denominator.

Crucially, the computational advantages identified above extend to a broad range of scoring methods and distance transformations. The use of a mismatch neighbourhood has potential to add significantly to both the compute and memory requirements of the process, but these demands are modest for $D_2^{n=1}$ and larger neighbourhoods seem not to improve its performance in phylogenetic inference. Alignment-free methods thus offer computational speed many hundreds or thousands of times faster than the comparable MSA-based approaches, with memory requirements in the hundreds of megabytes, well within the capabilities of even portable commodity devices. To the extent that memory is not an issue, alignment-free methods present an attractive, highly scalable alternative to MSA-based methods in large-scale phylogenetic (and phylogenomic) analyses.

Methods

Simulated sequence data. For all programs, default settings were used unless otherwise specified. We simulated sets of DNA and protein sequences of different sizes ($N = 8$, 16, 32, 128) using *evolver* as implemented in PAML 4.5[53], unless otherwise specified. We used GTR[54] (rate parameters $a = 0.987$, $b = 0.110$, $c = 0.218$, $d = 0.243$, $e = 0.395$)[55] and WAG[56] substitution models respectively for simulation of nucleotide and protein sequences. We detail simulation strategy for each evolutionary scenario below.

Sequence divergence. For each set, sequences of fixed length ($L = 1500$ nt for DNA; 500 amino acids for protein) were simulated on an unrooted symmetrical tree on which the lengths of internal (x) and terminal (y, or y_1 and y_2) branches are set separately, at either 0.01 or 0.05 substitutions per site, to represent six distinct scenarios (Fig. 1; shown for 8-taxon trees). These sequence sets were simulated under a discrete approximation of the gamma distribution (shape parameter $\alpha = 1.0$, 8 categories).

Genetic rearrangement. For each nucleotide sequence set ($N = 8$; $L = 5000$ nt), we relocated one or more region (i.e. individual rearrangement events) of 250 nt within a sequence in a cut-and-paste manner, with no overlaps. We define R as the total percentage length of L that has been relocated. We simulated sequence sets with $R =$ 10, 25 and 50% (each in 50 replicates), such that the total rearranged region is not contiguous. Given the prior expectation that alignment-free methods would be less sensitive to sequence rearrangements, here we simulated sequence sets under tree T3 (Fig. 1), one of the more problematic cases for D_2 methods (as shown in Fig. 2).

Insertions/deletions. For this analysis, we simulated nucleotide sequence sets of size $N = 32$ ($L = 1500$ nt) using INDELible[32] under tree T4 (Fig. 1), a discrete approximation of the gamma distribution ($\alpha = 1.0$, 8 categories) and GTR model. Indel rates were set at 0.1, 0.2, 0.3, 0.4 and 0.5, with insertion rate = deletion rate; these rates are relative to site substitution rate of 1. Length distribution of inserted/deleted fragments follows a Lavalette distribution[33,34] ($a = 1.1$; maximum indel size 100 nt) as implemented in INDELible[32].

Coalescent model of gene family evolution. We used NetRecodon[57] to simulate gene family evolution under the coalescence model along a tree, each case at a defined effective population size (N_e) of 1000, 10000, 100000, 250000, 500000 and 1000000, with a discrete approximation of the gamma distribution ($\alpha = 0.5$, 8 categories), GTR model and mutation rate $u = 10^{-5}$. Sequence sets of size $N = 32$ ($L = 1500$ nt) were used. Larger N_e values result in longer branch lengths on a tree (see Supplementary Table S1). To simulate violation of molecular clock, relaxed branch lengths were further simulated on these trees using *BranchRelaxer* in GenPhyloData[58], with substitution rates along branches modelled as independent and identically distributed variables in a log-normal scale (IIDLogNormal model: mean 0.0, variance 1.0)[59]. Sequences were then simulated using *evolver* along these new trees as per above.

Empirical sequence data. All 2471 nucleotide datasets in NEXUS format were downloaded from TreeBASE (treebase.org as of 27 May 2013)[41] using a custom script kindly provided by Dr William Piel. For each dataset, one or more nucleotide sequence alignment and their corresponding phylogenetic trees (totalling 4156) were extracted (Supplementary Data). All 406997 unaligned 16S ribosomal RNA gene sequences (sequences_16S_all_gg_2011_1_unaligned.fasta.gz)[60] were downloaded from the GreenGenes database (secondgenome.com/go/2011-greengenes-taxonomy). To assess scalability of D_2 methods on different sizes of sequence sets, these 406997 sequences were randomly selected across set $N = 1000, 2000, 3000, 4000$ and 5000, each in 100 replicates. We follow ref. 61 in defining within-set sequence similarity as the average pairwise similarity between each sequence in a set to the centroid sequence. A centroid sequence within a set is one that yielded the single highest bit score across all pairwise comparisons within the set using BLAST ($e < 10^{-3}$).

Alignment-free phylogenetic approach. For each sequence set, we used D_2 statistics independently for D_2, D_2^S, D_2^*, and $D_2^{n=1}$ to generate a score for each possible pair of sequences within a set (see Supplementary Note for details). These scores were transformed *via* logarithmic representation of the geometric mean to generate a distance. The pairwise distance between sequences a and b, D_{ab} is defined as

$$D_{ab} = \left| \ln\left(\frac{S_{ab}}{\sqrt{S_{aa} \times S_{bb}}} \right) \right|$$

where S_{ab} is the pairwise score between them, and S_{aa} and S_{bb} are the self-matching scores. These transformed pairwise distances closely approximate the angle-based distances in an earlier alignment-free method for inferring protein phylogenies[62]. The resulting distance matrix was used to reconstruct a phylogenetic tree using *neighbor* in PHYLIP v3.69 (evolution.genetics.washington.edu/phylip). Generation of the distance matrix from any of these D_2 methods is implemented in a JAVA program, JIWA, which is freely available at http://bioinformatics.org.au/tools/jiwa/.

Standard phylogenetic approach using multiple sequence alignment. For each sequence set, we used MUSCLE v3.8.31[26] to generate a multiple sequence alignment. For scenarios of genetic rearrangement, insertions/deletions and the coalescent model, we also used MAFFT (mafft-linsi) v7.158b[46]. For other simulated scenarios, alignments were perfectly given during the process of simulation; the use of any MSA tool would not yield any difference in the final alignments. For Bayesian phylogenetic inference, we used MrBayes v3.2.1[27] (MCMC ngen = 1500000 generations, samplefreq = 100, burn-in = 10000 samples, temp = 0.5, nchains = 4; sumt contype = allcompat). We assume the general reversible substitution model (lset Nucmodel = 4by4 Nst = 6) and a mixed amino acid substitution model (prset aamodel = mixed) respectively for nucleotide and protein sequences, under a four-category discrete gamma distribution across all runs (lset rate = gamma ngammacat = 4). In all cases except the insertions/deletions analysis, the standard deviation of split frequencies was <0.01 after 200000 generations. For insertions/deletions analysis, MrBayes was run at larger number of MCMC generations (ngen = 5000000) and burnin (samplefreq = 100, burn-in = 25000 samples), while other parameters remain the same. The standard deviation of split frequencies in most cases was <0.01 after 1000000 generations. For maximum likelihood inference of phylogenetic trees, we used RAxML v8.0.2[36] (-# 100, -t 4, -m GTRGAMMA or PROTGAMMAWAG respectively for nucleotide and protein sequences).

Assessment of accuracy. For each tree generated from a sequence set using D_2 statistics or the standard approach, we compared its topological congruence to a reference tree using the Robinson-Foulds distance[28], as implemented in *treedist* in PHYLIP v3.69 (evolution.genetics.washington.edu/phylip). This distance represents the number of splits (i.e. bipartitions) that are present in only one of the two trees. To facilitate comparison of our results across trees (i.e. sequence sets) of various sizes N, we normalised the distances by the maximum possible distance between two unrooted trees, $2(N - 3)$, following ref. 63. Here we denote RF as the normalised Robinson-Foulds distance, with a value between 0 and 1 that can be interpreted as the proportion of false or missing bipartitions in the test tree topology compared to the reference topology[63]. When $RF = 0$, the test and reference topologies are identical, suggesting high accuracy of the approach. When $RF = 1$, none of the bipartitions in the reference is recovered in the test. In these cases, the trees could have been generated at random, as a pair of randomly generated tree topologies of N taxa has a Robinson-Foulds distance that approximates the denominator for normalisation, $2(N - 3)$[64]. For the simulated data, we used the known tree (under which the sequences were simulated) as the reference. For empirical data from TreeBASE we used the published tree in the database as reference; in these cases, a zero RF does not relate directly to accuracy, but rather reflects the extent to which our method recovers the same topology as the published method based on multiple sequence alignment.

Assessment of computational scalability and runtime. The assessment of computational scalability was carried out using a high-performance distributed-memory computing cluster based on Intel Sandy Bridge 8-core 2.6 GHz processors. Comparative runtime analysis of alignment-free and MSA-based phylogenetic approaches was done on Intel Xeon L5520 8-core 2.26 GHz processors (multi-threaded, four threads). MCMC ngen = 1500000 was used for MrBayes runs.

1. Edgar, R. C. & Batzoglou, S. Multiple sequence alignment. *Curr. Opin. Struct. Biol.* **16**, 368–373 (2006).
2. Notredame, C. Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput. Biol.* **3**, 1405–1408 (2007).
3. Darling, A. E., Miklos, I. & Ragan, M. A. Dynamics of genome rearrangement in bacterial populations. *PLoS Genet.* **4**, e1000128 (2008).
4. Puigbò, P., Wolf, Y. I. & Koonin, E. V. The tree and net components of prokaryote evolution. *Genome Biol. Evol.* **2**, 745–756 (2010).
5. Zhaxybayeva, O. & Doolittle, W. F. Lateral gene transfer. *Curr. Biol.* **21**, R242–246 (2011).
6. Wong, K. M., Suchard, M. A. & Huelsenbeck, J. P. Alignment uncertainty and genomic analysis. *Science* **319**, 473–476 (2008).
7. Wu, M. T., Chatterji, S. & Eisen, J. A. Accounting for alignment uncertainty in phylogenomics. *PLoS ONE* **7**, e30288 (2012).
8. Chan, C. X. & Ragan, M. A. Next-generation phylogenomics. *Biol. Direct* **8**, 3 (2013).
9. Höhl, M. & Ragan, M. A. Is multiple-sequence alignment required for accurate inference of phylogeny? *Syst. Biol.* **56**, 206–221 (2007).
10. Höhl, M., Rigoutsos, I. & Ragan, M. A. Pattern-based phylogenetic distance estimation and tree reconstruction. *Evol Bioinform Online* **2**, 359–375 (2006).
11. Domazet-Lošo, M. & Haubold, B. Alignment-free detection of local similarity among viral and bacterial genomes. *Bioinformatics* **27**, 1466–1472 (2011).
12. Vinga, S. & Almeida, J. Alignment-free sequence comparison - a review. *Bioinformatics* **19**, 513–523 (2003).
13. Bonham-Carter, O., Steele, J. & Bastola, D. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Brief. Bioinform.*, In Press, DOI:10.1093/bib/bbt052 (2013).
14. Haubold, B. Alignment-free phylogenetics and population genetics. *Brief. Bioinform.* **15**, 407–418 (2014).
15. Song, K. *et al.* New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Brief. Bioinform.* **15**, 343–353 (2014).
16. Torney, D. C., Burks, C., Davison, D. & Sirotkin, K. M. in *Computers and DNA - Santa Fe Institute Studies in the Sciences of Complexity, Vol. 7* (eds. Bell, G. & Marr, R.) 109–125 (Addison-Wesley, Reading, MA; 1990).
17. Wan, L., Reinert, G., Sun, F. & Waterman, M. S. Alignment-free sequence comparison (II): theoretical power of comparison statistics. *J Comput. Biol.* **17**, 1467–1490 (2010).
18. Reinert, G., Chew, D., Sun, F. & Waterman, M. S. Alignment-free sequence comparison (I): statistics and power. *J Comput. Biol.* **16**, 1615–1634 (2009).
19. Hide, W., Burke, J. & Davison, D. B. Biological evaluation of d^2, an algorithm for high-performance sequence comparison. *J Comput. Biol.* **1**, 199–215 (1994).

20. Miller, R. T. *et al.* A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res.* **9**, 1143–1155 (1999).
21. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
22. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
23. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
24. Göke, J., Schulz, M. H., Lasserre, J. & Vingron, M. Estimation of pairwise sequence similarity of mammalian enhancers with word neighbourhood counts. *Bioinformatics* **28**, 656–663 (2012).
25. Yi, H. & Jin, L. Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic Acids Res.* **41**, e75 (2013).
26. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
27. Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).
28. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
29. Forêt, S., Wilson, S. R. & Burden, C. J. Empirical distribution of *k*-word matches in biological sequences. *Pattern Recognit.* **42**, 539–548 (2009).
30. Forêt, S., Kantorovitz, M. R. & Burden, C. J. Asymptotic behaviour and optimal word size for exact and approximate word matches between random sequences. *BMC Bioinformatics* **7** **Suppl 5**, S21 (2006).
31. Huffman, D. A. A method for the construction of minimum-redundancy codes. *Proc. IRE* **40**, 1098–1101 (1952).
32. Fletcher, W. & Yang, Z. INDELible: a flexible simulator of biological sequence evolution. *Mol. Biol. Evol.* **26**, 1879–1888 (2009).
33. Lavalette, D. Facteur d'impact: impartialité ou impuissance? (INSERM U350 Institut Curie-Recherche, Bât. 112, Centre Universitaire, Orsay, France; 1996).
34. Popescu, I. I. On a Zipf's Law extension to impact factors. *Glottometrics* **6**, 83–93 (2003).
35. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
36. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
37. Golubchik, T., Wise, M. J., Easteal, S. & Jermiin, L. S. Mind the gaps: evidence of bias in estimates of multiple sequence alignments. *Mol. Biol. Evol.* **24**, 2433–2442 (2007).
38. Kingman, J. F. C. The coalescent. *Stoch. Proc. Appl.* **13**, 235–248 (1982).
39. Tellier, A. & Lemaire, C. Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Mol. Ecol.* **23**, 2637–2652 (2014).
40. Sjödin, P., Kaj, I., Krone, S., Lascoux, M. & Nordborg, M. On the meaning and existence of an effective population size. *Genetics* **169**, 1061–1070 (2005).
41. Piel, W. H., Donoghue, M. J. & Sanderson, M. J. in *To the interoperable "Catalog of Life" with partners Species 2000 Asia Oceania. NIES Research Report*, *Vol. 171* (eds. Shimura, J., Wilson, K. L. & Gordon, D.) 41–47 (National Institute for Environmental Studies, Tsukuba, Japan; 2002).
42. Posada, D. Phylogenetic models of molecular evolution: next-generation data, fit, and performance. *J. Mol. Evol.* **76**, 351–352 (2013).
43. Ragan, M. A. & Chan, C. X. Biological intuition in alignment-free methods: response to Posada. *J. Mol. Evol.* **77**, 1–2 (2013).
44. Ragan, M. A., Bernard, G. & Chan, C. X. Molecular phylogenetics before sequences: Oligonucleotide catalogs as *k*-mer spectra. *RNA Biol.* **11**, 176–185 (2014).
45. Chan, C. X., Darling, A. E., Beiko, R. G. & Ragan, M. A. Are protein domains modules of lateral genetic transfer? *PLoS ONE* **4**, e4524 (2009).
46. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
47. Thompson, J. D., Linard, B., Lecompte, O. & Poch, O. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS ONE* **6**, e18093 (2011).
48. Liu, K., Linder, C. R. & Warnow, T. RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. *PLoS ONE* **6**, e27731 (2011).
49. Gunasinghe, U., Alahakoon, D. & Bedingfield, S. Extraction of high quality *k*-words for alignment-free sequence comparison. *J. Theor. Biol.* **358**, 31–51 (2014).
50. Haubold, B. & Pfaffelhuber, P. Alignment-free population genomics: an efficient estimator of sequence diversity. *G3* **2**, 883–889 (2012).
51. Fitch, W. M. & Margoliash, E. Construction of phylogenetic trees. *Science* **155**, 279–284 (1967).
52. Burden, C. J., Kantorovitz, M. R. & Wilson, S. R. Approximate word matches between two random sequences. *Ann. Appl. Probab.* **18**, 1–21 (2008).
53. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
54. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* **17**, 57–86 (1986).
55. Yang, Z. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39**, 105–111 (1994).
56. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–699 (2001).
57. Arenas, M. & Posada, D. Coalescent simulation of intracodon recombination. *Genetics* **184**, 429–437 (2010).
58. Sjöstrand, J., Arvestad, L., Lagergren, J. & Sennblad, B. GenPhyloData: realistic simulation of gene family evolution. *BMC Bioinformatics* **14**, 209 (2013).
59. Drummond, A. J., Ho, S. Y., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
60. McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6**, 610–618 (2012).
61. Chan, C. X., Mahbob, M. & Ragan, M. A. Clustering evolving proteins into homologous families. *BMC Bioinformatics* **14**, 120 (2013).
62. Stuart, G. W., Moffett, K. & Baker, S. Integrated gene and species phylogenies from unaligned whole genome protein sequences. *Bioinformatics* **18**, 100–108 (2002).
63. Kupczok, A., Schmidt, H. & von Haeseler, A. Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. *Algorithms Mol. Biol.* **5**, 37 (2010).
64. Bryant, D. & Steel, M. Computing the distribution of a tree metric. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **6**, 420–426 (2009).

Acknowledgments

We thank The University of Queensland and James S McDonnell Foundation for financial support. CXC is supported by a University of Queensland Early Career Researcher grant. This work was supported by computational resources of the National Computational Infrastructure (NCI) National Facility systems and NCI Specialised Facility in Bioinformatics through the NCI Merit Allocation Scheme (Project d85). We thank Professor Michael Waterman for helpful suggestions, and Dr Lars Jermiin and Professor David Penny for their constructive feedback on this work.

Author contributions

C.X.C., J.M.H. and M.A.R. conceived the project. C.X.C., G.B. and M.A.R. designed the experiments, C.X.C., G.B. and O.P. implemented the analysis workflow and conducted the experiments, C.X.C., G.B., J.M.H. and M.A.R. analysed and interpreted the results, C.X.C. prepared all figures and tables, C.X.C. and M.A.R. prepared and wrote the manuscript. All authors reviewed, commented on and approved the final manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Chan, C.X., Bernard, G., Poirion, O., Hogan, J.M. & Ragan, M.A. Inferring phylogenies of evolving sequences without multiple sequence alignment. *Sci. Rep.* **4**, 6504; DOI:10.1038/srep06504 (2014).