OXFORD

Briefings in Bioinformatics, 20(2), 2019, 426–435

doi: 10.1093/bib/bbx067
Advance Access Publication Date: 30 June 2017
Paper

Alignment-free inference of hierarchical and reticulate phylogenomic relationships

Guillaume Bernard, Cheong Xin Chan, Yao-ban Chan, Xin-Yi Chua, Yingnan Cong, James M. Hogan, Stefan R. Maetschke and Mark A. Ragan

Corresponding author: Mark A. Ragan, Institute for Molecular Bioscience, the University of Queensland, 306 Carmody Rd, St Lucia, Brisbane, Queensland 4072, Australia. Tel.: +61-7-3346-2616; E-mail: m.ragan@uq.edu.au

Abstract

We are amidst an ongoing flood of sequence data arising from the application of high-throughput technologies, and a concomitant fundamental revision in our understanding of how genomes evolve individually and within the biosphere. Workflows for phylogenomic inference must accommodate data that are not only much larger than before, but often more error prone and perhaps misassembled, or not assembled in the first place. Moreover, genomes of microbes, viruses and plasmids evolve not only by tree-like descent with modification but also by incorporating stretches of exogenous DNA. Thus, next-generation phylogenomics must address computational scalability while rethinking the nature of orthogroups, the alignment of multiple sequences and the inference and comparison of trees. New phylogenomic workflows have begun to take shape based on so-called alignment-free (AF) approaches. Here, we review the conceptual foundations of AF phylogenetics for the hierarchical (vertical) and reticulate (lateral) components of genome evolution, focusing on methods based on *k*-mers. We reflect on what seems to be successful, and on where further development is needed.

Key words: alignment-free; phylogenomics; lateral genetic transfer; k-mer, D2 statistics; TF–IDF

Guillaume Bernard is a PhD candidate at the Institute for Molecular Bioscience. His research focuses on the development of alignment-free approaches in microbial phylogenomics.
Cheong Xin Chan is a senior research officer and affiliate lecturer at the University of Queensland. His research interests include microbial genomics and evolution, and scalable approaches in phylogenomics.
Yao-ban Chan is a lecturer at the University of Melbourne. His research interests include the development and application of mathematical and statistical methods to problems in evolutionary molecular biology.
Xin-Yi Chua is the head of Informatics at QFAB Bioinformatics. She has a keen interest in applying comparative genomics, machine learning and visualization methods to better understand the biology of microbial pathogenesis.
Yingnan Cong is a recent PhD graduate from the University of Queensland. His research interests are in the analysis of genomes using approaches based on natural language processing, data mining and artificial intelligence.
James M. Hogan is an associate professor of computer science at Queensland University of Technology. His research interests include *k*-mer methods for sequence analysis and visualization in bioinformatics.
Stefan R. Maetschke is a computer scientist at IBM Research Australia. The main focus of his work has been on machine learning methods to infer and analyse protein–protein interaction, gene-regulatory and lateral transfer networks.
Mark A. Ragan is a professor of computational genomics at the Institute for Molecular Bioscience. He has a long-standing interest in scalable methods for phylogenomics.

Submitted: 8 February 2017; **Received (in revised form):** 4 May 2017

Introduction

Phylogenomics refers to an important body of theory, methodology and tools applicable to the comparative analysis of genome-scale data within an evolutionary context [1–4]. The field builds on molecular phylogenetics, which since the early 1960s has been developed to elucidate genealogical relationships and evolutionary processes within families of genes or proteins. As the first area of molecular bioscience to develop an explicitly algorithmic approach, and drawing richly on statistics and computational science, phylogenetics is considered a major area within bioinformatics.

By definition, the inference of genealogical relationships must be based on homologous elements. Even before molecules could be fully sequenced, it was known that certain oligopeptides were common to representatives (in different biological species) of individual proteins, e.g. α-haemoglobin or insulin; similarly, 16S ribosomal RNAs in different species shared sets of short oligonucleotides [5]. Indeed, the presence of shared identity in sequences beyond the extent required to deliver conserved function was taken as evidence for homology [6]. As full-length sequences became available, it made sense to discover and display these conserved regions in a multiple sequence alignment (MSA) [7, 8]. Molecular phylogenetics is thus endowed with a richness and precision rarely seen with phenetic characters: homology is no longer 'overall' and subjective, but can be evidenced column by column along a set of aligned sequences. Thus, an MSA is an explicit position-by-position hypothesis of homology. Not coincidentally, an MSA matrix serves as a convenient input to software programs that calculate pairwise dissimilarities (for distance methods) or compute a tree that best explains the patterns in the aligned columns, given a model of sequence change over time (in e.g. parsimony or likelihood methods).

With the advent and spread of genomics, large data sets have become available for phylogenetic inference, with sequences longer (genomes rather than single genes or proteins) and much more numerous. It is no longer unusual to encounter data sets with thousands of genome or (concatenated) exome sequences. Phylogenomics at this scale requires organized data management, significant computational power and large memory. However, phylogenomics can present challenges other than those arising purely from size and scale: the data may be of low quality, assembly may be poor or nonexistent, the sequences may not be collinear over their entire length, different models of sequence change probably apply (e.g. to protein-coding and noncoding regions) and sequence regions may have different origins and evolutionary histories [9]. We consider these factors in turn.

Classical (alignment-based) phylogenetics

In the early days of genomics, much effort went into 'finishing and polishing'—joining contigs and resolving conflicts to recover full-length chromosomes nearly free of ambiguities or errors. Thanks to this body of earlier work, key information (e.g. presence or absence of a gene or pathway) can now often be obtained simply by deep sequencing followed by a rough assembly. Depending on one's scientific goals, the breadth-versus-depth trade-off can be pushed dramatically towards breadth, at the expense of data quality. Survey projects now target tens of thousands of bacterial genomes, few of which will assemble into a single contig, while large eukaryote genomes can be approached through transcriptomics, with consequences for MSA including the need to deal with truncated sequences and alternative splice forms. Indeed, there is optimism that phylogenetic trees might be inferred entirely without assembly [10].

Basic MSA requires sequences to be collinear, i.e. to preserve a common ancestral order of elements. Depending on the sequences being compared, these elements may be, for example, nucleotides, codons, amino acids, domains, exons or genes. Non-collinearity may arise because of poor sequence quality or misassembly, but could also be real, particularly at whole-genome scale. Across bacterial genomes, gene order tends to be poorly conserved except among close relatives; exceptions include ribosomal RNA operons and some genes encoding ribosomal proteins. Thus, even with correctly assembled genomes, a separate bioinformatic step is required to match putatively homologous regions before, or as part of, MSA. Like MSA software, whole-genome aligners take different algorithmic approaches and implement different assumptions and trade-offs [11, 12] but in general are CPU- and memory-intensive, with considerable scope for error and ambiguity arising, e.g. from families of repetitive elements, low-complexity regions and paralogs.

Most approaches to phylogenetic inference require a statistical model of sequence evolution [13]. It is not difficult to imagine that different classes of sequence (e.g. those encoding a protein, a functional RNA or no product at all) are best described by different models. Even within a single gene, inference quality may be improved by applying different rate classes or steady-state assumptions, e.g. for DNA regions that encode highly structured versus unstructured regions of proteins, or stems versus loops of ribosomal RNAs. It is computationally onerous to identify, delineate and group these regions, match each to the best model and optimize parameter values. Scalability to genome-scale data would be facilitated by simplifying these models, using a single generic model or, if possible, eliminating them altogether.

In the standard Darwinian model, genomes are inherited vertically from one generation to the next within lineages. To a first approximation, this adequately describes the evolution of nuclear genomes of morphologically complex organisms including animals and plants. However, genomes of bacteria, archaea, protists, viruses and plasmids often contain stretches of DNA acquired laterally from unrelated organisms, or from the environment. Many studies indicate that 10–40% of the genes in some bacterial genomes, and essentially all gene families in bacteria, have been affected by lateral (horizontal) genetic transfer (LGT or HGT) ([14, 15] and references therein). For such genomes, a phylogenomic workflow must distinguish vertical from lateral signal, and treat each separately. To complicate matters, neither genes nor domons (genomic regions corresponding to protein domains) are privileged units of LGT [16, 17]; new lateral events can overwrite older ones; regions of lateral origin may ameliorate, i.e. evolve to become indistinguishable from their new host genome [18]; and older lateral regions will themselves be inherited vertically within subtrees [15, 19]. In MSA-based phylogenomics, these issues are addressed by adding further (computationally demanding) steps to the workflow, e.g. inferring an organismal reference tree and comparing features of its topology with those of individual gene or protein family trees [14, 20]. Opportunities abound for complications to arise from cryptic paralogy, or inappropriate delineation of the units of analysis.

Based on the above, we might summarize our wish list for next-generation phylogenomics [9]: it must be based on homologous signal (a different subset of signal for each evolutionary origin), while avoiding the assumptions inherent in MSA (predefined fixed units of analysis; colinearity). It should incorporate a generic, computationally simple substitution model; be highly scalable to large data, yet robust to low data quality; and perhaps support phylogenetic inference from unassembled sequence reads. Alignment-free (AF) methods offer considerable promise against each of these goals [9].

Alignment-free methods and *k*-mers

AF methods underpin key algorithms in diverse areas of bioinformatics including database searching [21, 22], sequence clustering [23], error correction in sequencing reads [24], genome assembly [25], discriminative prediction of regulatory variants [26, 27] and testing for genetic recombination [28]. In phylogenetics and phylogenomics, AF methods offer alternatives to the assumptions and computational demands of MSA identified above. Following Haubold [29], AF methods can be classified broadly into those based on word (k-mer) count, and those based on match length. Certain other AF methods may fit uncomfortably into these classes, or lie outside them altogether [30]. In the present context, the motivating concept is the same: substrings (perhaps defined by k-mers) that meet certain criteria, and are shared by a set of sequences, can be considered as capturing part of the homology signal and are thus potentially informative on phylogeny. Here, we focus primarily on k-mer count methods.

Substrings (sub-sequences) of defined length are variously known as words, k-mers or n-grams, with k or n denoting the substring length. By disallowing mismatch, degeneracy and indels, k-mer statistics become simpler and the computation more efficient. These strictures may be slightly relaxed (e.g. by allowing limited mismatch to deal with noise) or avoided in part (by recoding into a reduced alphabet), albeit at the risk of crossing into the realm of pattern or motif analysis, for which different and computationally less-favourable methods are required.

Any molecular sequence can be represented as the set of its constituent k-mers (Figure 1). These k-mers are typically allowed to overlap with stride = 1; a larger stride $\leq k$ could be used to reduce the computational effort. Whereas in MSA the linear order of sequence elements is fundamental to recognizing conserved (homologous) positions and identifying conservation profiles, the analogous concept in AF is an order-less matching of k-mers, i.e. an intersection of k-mer sets. For sufficiently large k, any given k-mer is approximately unique to a sequence [31], so in the absence of extenuating circumstances (e.g. strong mutational bias or low-complexity regions), shared instances of that k-mer are likely to be homologous. As sequences progressively diverge on a tree, they share fewer k-mers in common, and the longest k-mer they share tends to be shorter. As we discuss below, these measures can be used to estimate a pairwise distance. For this, the count or frequency of shared k-mers seems to be sufficient, i.e. it is not necessary to keep track of positional information [32] unless we wish to map specific k-mers (e.g. those inferred to have a lateral origin) to genes, structures or functions [33–35].

In contrast, conservation profiles measure local sequence similarity and thus require at least approximate positional information. Classical alignment algorithms such as Smith–Waterman or Needleman–Wunsch use dynamic programming to determine columns and/or blocks of matching residues in a set of sequences. From the number or proportion of conserved residues within a column, a conservation profile can be derived. AF conservation profiles can be constructed by plotting the maximum number of matching k-mers over their mean positions within the set of sequences [36]. While AF conservation profiles do not achieve single-residue resolution, they are faster to compute than classical profiles (linear versus quadratic time with respect to sequence length), yet still allow the identification of conserved regions or domains, even if non-collinear [36].

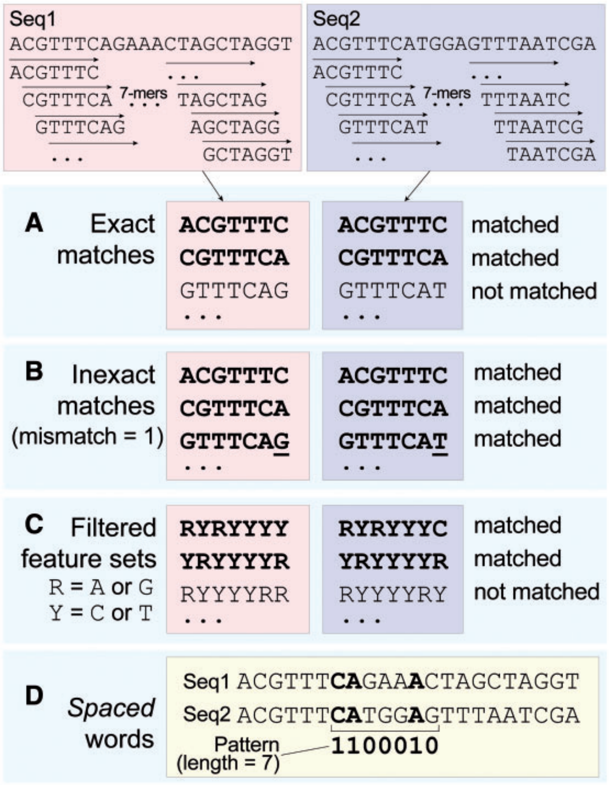


Figure 1. Fundamental concepts and nomenclature of k-mers, illustrated here for overlapping k-mers (k = 7, stride = 1) in two DNA sequences. (**A**) Exact matches, (**B**) inexact matches, (**C**) degenerate bases and (**D**) a binary pattern of match and non-match positions (spaced word matches).

Extraction of k-mer sets from molecular sequences is in principle trivial, accomplished simply by sliding a window of size k over a string representation of the sequence to produce a lexicon of overlapping k-mers (Figure 1). In application to DNA, it is usual to record only canonical k-mers (the lexicographically smaller of a k-mer and its reverse complement), or to work with only the forward strand. Efficient accumulation of k-mer counts requires that we determine the novelty of each k-mer as it appears: previously encountered k-mers must be identified rapidly and their counts incremented, while novel k-mers must be inserted quickly and without impairing the performance of the data store during subsequent queries. Approaches to this task may be categorized broadly into those based on hashing, and those relying more directly on data structures invented for string lookup and spelling correction, notably the suffix tree [37] and the suffix array [38].

Exact hashing methods offer constant-time insertion and lookup, and their use in bioinformatics has a long history. Naïve hashing, however, proves surprisingly slow [39] and requires memory linear in the number of distinct k-mers, which in turn is exponential in the size of the k-mers: $O\left(|A|^k\right)$, where $|A|$ is the size of the alphabet. Jellyfish [39] overcomes many of these problems through careful design of a lock-free multithreaded hash table, using a key encoding and bit packing to ensure far lower memory usage. Tessel, part of the Blue read-correction package [40], markedly reduces memory requirements for sequence reads by excluding singleton k-mers until a second occurrence is observed. Even for modest coverage, genuine k-mers will occur many times, with singletons almost certainly the result of sequencing error. Confirmed k-mers are recorded in optimized, partitioned lock-free hash tables, with a subsequent merge phase to ensure accurate final counts. Melsted and Pritchard [41] address memory usage via a Bloom filter [42], a probabilistic distributed hashing scheme that admits a small chance of a false-positive match. A standard hash table is used to store entries seen twice or more, but the implementation remains single-threaded and is not competitive. Hybrids of this nature have been used recently in the context of de Bruijn graphs [43].

Suffix trees represent a string through its underlying suffixes, each encoded as a path from the root to a leaf, with the start position of the suffix within the string stored in this terminating node. Suffix arrays contain these same start positions, but arranged according to the lexicographical order of the suffixes included. For both, construction time and space are linear in sequence length, hence in the number of k-mers, but the array requires far fewer bits per suffix, perhaps one-fifth to one-third of the footprint of the tree [38]. Lookup is linear in the length of the query, here $O(k)$. Suffix trees have long been applied in substring matching (e.g. in MUMmer 1.0: [44]), while k-mer counters based on suffix arrays have included Meryl (part of the Celera Assembler) [45] and Tallymer [46], the latter enhanced by storing the longest common prefix among suffix groups. While the optimized hashing methods discussed above appear superior for general k-mer counting and retrieval, exact hashing does not preserve locality, limiting its utility for applications based on approximate matching. In contrast, suffix and tree-based approaches may preserve structure common across many k-mer entries, supporting mismatch neighbourhoods and correction of a letter (or longer segment) through replacement by a more strongly weighted alternative. Even so, these tasks may be prohibitively expensive for very large data sets, which may be handled more generally through careful inclusion of disk or solid-state drives.

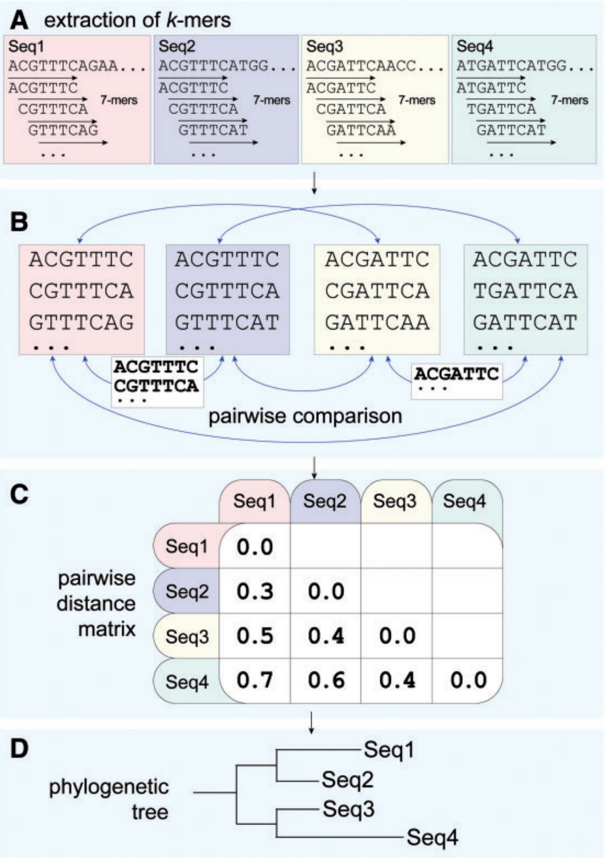
Phylogenetic inference based on k-mers

As we have mentioned, as sequences diverge over time from a common ancestor, they will come to share fewer, and shorter, k-mers. More precisely: given a threshold τ such that k-mers of length $k \geq \tau$ occurring in related sequences can be considered homologous, as these sequences diverge (a) for a fixed $k \geq \tau$, the number of shared k-mers will tend to decrease, (b) over all $k \geq \tau$, the mean length of shared k-mers will tend to decrease and (c) the longest shared k-mer will tend to be shorter. Measures that capture these trends behave as pairwise similarities, and like their classical MSA-based counterparts can be used in distance analysis to generate a tree [47–55]. The optimal τ is likely to be problem- and data-dependent (see below), but could be selected based on the distribution of match lengths in simulated sequences [56], e.g. to maximize the area under the receiver operating characteristic curve or ensure a minimum desired frequency of true positives.

The best-known measure of k-mer distance is based on the D_2 statistic [21, 48, 57–60]. Building on a proposal by Blaisdell [61], D_2 is simply the count of exact k-mer matches between two sequences, summed over all k-mers at a given k. As this count depends on the sequence lengths, D_2 is often normalized by the probability of k-mer occurrence, or by assuming a Poisson distribution [48, 62]. Chan *et al.* [63] introduced a neighbourhood variant. Even so, for D_2-based measures to be applied confidently, particularly in the comparison of closely related sequences, understanding the k-mer structure of actual genomes would be highly desirable [58, 64–67].

Bioinformatic workflows leading to AF trees differ little from their classical counterparts, except that MSA is not required (Figure 2). Putatively homologous sequences (e.g. genomes) are assembled and quality checked, e.g. for illegal characters. K is selected (see below), k-mers are extracted, and distances are computed pairwise (above) and assembled into a triangular matrix that is input into software that implements neighbour joining [68, 69] or another distance-based algorithm. Because distance algorithms build trees by clustering sequences rather than by estimating a measure of changes along internal edges, some authorities consider them non-phylogenetic. Here, we follow Felsenstein ([70]: 145–6) in relegating this distinction to debates over classification, and for the purpose at hand accept distance as a legitimate basis for the statistical inference of phylogeny. Höhl and Ragan [71] pointed out that shared k-mers could be arranged into a (very local) 'alignment' matrix and used as input into likelihood, Bayesian or other (non-distance) algorithms for tree inference, although at the cost of the speed and scalability we hoped to secure by taking an AF approach in the first place.

K is the critical parameter in AF phylogenomics. As we depend on k-mers to capture homology signal, the value we select



	Seq1	Seq2	Seq3	Seq4
Seq1	0.0			
Seq2	0.3	0.0		
Seq3	0.5	0.4	0.0	
Seq4	0.7	0.6	0.4	0.0

Figure 2. An AF phylogenetic workflow in which (**A**) k-mers ($k = 7$, stride $= 1$) are extracted from four sequences (Seq1 through Seq4), (**B**) shared 7-mers are identified by pairwise comparisons, (**C**) a pairwise distance matrix is calculated, from which (**D**) a tree is computed using a distance-based method, e.g. neighbour joining.

for k must be large enough to ensure that few k-mers are present in our analysis purely by chance, but not so large that informative k-mers are arbitrarily excluded and signal unnecessarily attenuated. The factors most important for selecting an optimal k are the alphabet (e.g. nucleotide versus amino acid) and the complexity, divergence and length of the sequences under investigation. Given the complexities of sequence evolution, the performance of AF methods is best assessed by computational simulation rather than analytically. Using *evolver* in PAML [72], we simulated the evolution of sequences on a tree under a general time-reversible model to examine how $G + C$ content, length of terminal or more-basal branches, rearrangement, truncation and among-site rate heterogeneity affect precision and recall. We also simulated indels, and explored trees generated under coalescent and non-ultrametric models [63]. Our results indicate that for k within an optimal range, many AF methods can perform well under basic scenarios [5, 63, 73], indeed better than MSA in the presence of rearrangement or indels [63].

In the case of empirical data, the true tree is unknown, making it impossible to assess performance using measures of precision and recall. To compare AF methods under scenarios of sequence divergence, rearrangement, inversion and LGT, we focused instead on sensitivity to change of parameter value (e.g. k), and accuracy in the sense of recovering accepted subtrees [73]. All nine AF methods we examined were robust against complex genome rearrangements or inversions, and most word-count methods were robust and computationally efficient against moderate levels of LGT. Performance varied with the extent of divergence, with the word-count methods more accurate than match-length methods at higher divergence. The optimal size of k was sensitive to the extent of sequence divergence, but was little affected by the other scenarios we simulated. Thus, for data sets of known divergence, AF methods might be applied without exploratory tuning of k, and could be expected to perform as well or better than MSA-based approaches. However, AF methods have not been rigorously examined under fully realistic scenarios in which different lineages may evolve at different or variable rates, under different models of substitution, and/or with biases that give rise to compositional convergence.

For large data sets of bacterial and archaeal genomes, we inferred biologically realistic AF trees in which many clades familiar from MSA-based studies were recovered. Most differences between the AF and MSA trees involved terminal branches, i.e. the most-closely related genomes. We investigated a multiple-k approach in hopes that longer k might provide better resolution at the termini, while shorter k would be more appropriate for the eroded signal at more-basal bipartitions. In our hands this was unsuccessful, but an adaptive or multiple-k approach might bear more-systematic reinvestigation. Some AF methods can also be used directly on large high-throughput sequencing data, i.e. sets of reads or contigs with only basic assembly, or none at all [10, 54].

Given sets of k-mers from individual sequences, the time required to compute AF distances typically scales linearly with the number of sequences; weighting, normalization or extension to inexact matches will incur additional cost [50, 52, 54, 56, 62, 73–75]. Using D_2^S, we could generate accurate trees for thousands of bacterial genomes in some tens of hours on a moderate-sized cluster [76]. Memory is the main limitation for k-mer-based approaches, but the actual demand depends on the implementation used, and can sometimes be traded off against speed of computation. AF methods with optimal memory consumption are slower than the more memory-greedy methods, with current hashing-based implementations limited to $k = 32$ in most cases [47, 50].

AF methods nonetheless retain certain limitations. In simulations, the D_2-based methods we examined recover the reference topology when applied to sequences of length of 10 000 nt (e.g. small genomes; operons), but are prone to errors at 1500 nt (genes) and fail at 250 nt (domons) [63]. By disregarding singleton k-mers (i.e. erroneous reads), it is possible to improve distance estimates at higher coverage, but this degrades the signal at lower coverage [10]. In the MSA context, distance methods are criticized for reducing the pairwise comparison between sequences to a single number, in the process losing information on patterns of conservation within and among sequences; this is true of k-mer distances as well. Alternative approaches might involve a k-mer substitution model, but this scarcely seems feasible if the substitution matrix would be high dimensional, sparse and dependent on immense data for parameterization. Indeed, we suspect that such an approach would be so computationally expensive that it would negate the advantages of taking an AF approach in the first place. Methods exist for computation with sparse matrices, but to our knowledge have not been explored in a phylogenetic or phylogenomic context.

Alignment-free approaches to lateral genetic transfer

For phylogenomic analysis of genomes potentially affected by LGT, we must also identify and deconvolute vertical and lateral signal. In MSA-based phylogenomics, this is done by appending a filter to the standard workflow: trees inferred for individual gene families are compared with a reference topology, and well-supported but conflicting bipartitions are taken as *prima facie* evidence of LGT [77, 78]. The corresponding gene or protein family might then excluded (to purify the vertical signal) or analysed separately to understand the sources, recipients, processes and impact of LGT.

At first glance, there is much to recommend a similar workflow for AF phylogenomics. An approach built on k-mers might liberate us from having to take genes, or any other predefined features, as the units of analysis. In MSA-based phylogenomics, incongruent signal can be traced back to the underlying gene-family MSA, but this does not tell us which of the aligned gene(s) is/are responsible for the incongruence, the number and quality of alternative signals or the number, quality or location of recombination breakpoints [16, 79]. AF methods might give us fine-scale access to some or all of this information. Indeed, with AF, we have further options. A genomic region might have arisen by LGT if, in the absence of extenuating circumstances, it (a) unexpectedly shares k-mers with a distantly related genome, and hence (b) exhibits an anomalously short D_2 distance to that genome. This is why (c) a distance tree computed for that region will be topologically incongruent with that computed for a vertically inherited regions, or a trusted reference tree. Interestingly, these lines of evidence exactly parallel the three main strategies for LGT detection [80–82].

To begin to explore these AF approaches, we simulated the evolution of DNA sequences on a tree using ALF [83] or EvolSimulator [84], and then counted how many 21-mers are shared pairwise within a sliding window of length 60 nucleotides. In the first instance, we did this in the absence of simulated LGT, so as to establish a baseline against which lateral regions could later be detected [85] (Chua, Maetschke and Ragan, unpublished). This is a k-mer variant of approaches long used to find lateral

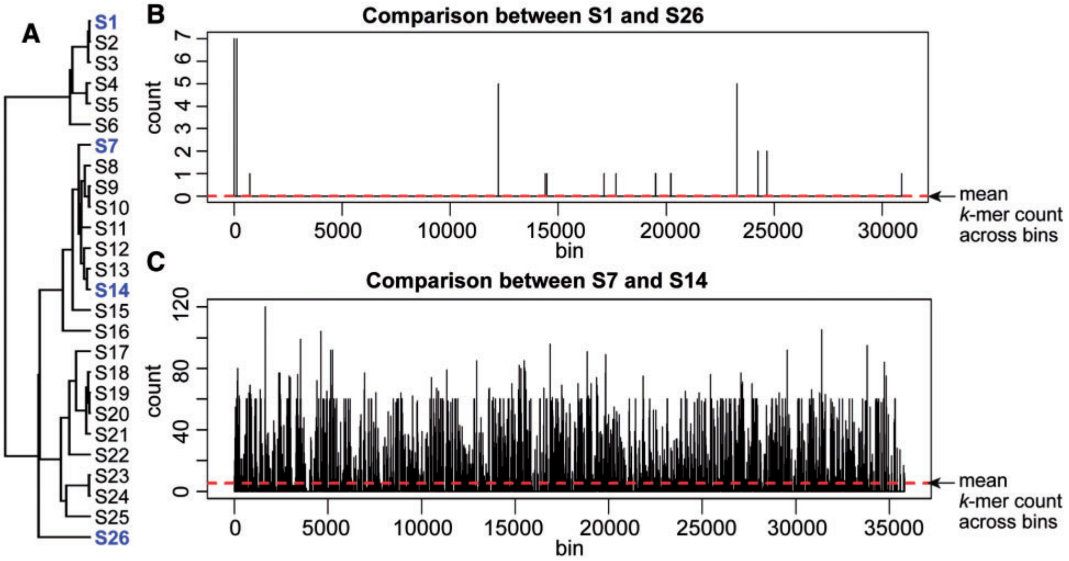


Figure 3. A sliding-window approach of *k*-mer sharing between sequences, illustrated here using a set of 26 sequences simulated [84] on the tree (**A**) depicted at the left. Pairwise comparisons are shown for (**B**) two highly dissimilar sequences, S1 and S26, and (**C**) two similar sequences, S7 and S14. Each plot shows the number of matching 21-mers within a 60-nt window, as it is incremented along S1 or S7, respectively.

regions within sequences based on anomalous G + C content, dinucleotide frequencies or codon usage [18, 81, 82, 86–92]. We found that while the most-divergent sequences shared very few 21-mers (zero in most windows), a few windows shared as many as seven; although these are false positives, we could find no objective statistical criterion by which a hypothesis of LGT could be rejected for them. Conversely, with more-closely related sequence pairs, most windows shared many 21-mers, but the variation was such that it would be impossible to recognize or bound a truly lateral region (Figure 3). Note also that a sliding window approach could work only on assembled genomes or large scaffolds, not on masses of raw reads. The idea seemed promising, but something critical was missing.

Thanks to cross-disciplinary collaboration, we soon discovered what was missing. Document analysis involves concepts that can be identical or analogous to those in molecular phylogenetics [93–96] including the 'contamination' of texts by lateral transfer [97, 98]. A statistic known as term frequency–inverse document frequency (TF–IDF) is widely used to determine the importance of a word in a collection of documents: words that appear frequently in a document, but rarely in the rest of the corpus, carry greater importance for that document. A variant of TF–IDF might be used to detect lateral regions in molecular sequences. *K*-mers can be seen as analogues of words (albeit ones that sometimes overlap each other), groups of similar sequences as documents and a sequence database as a corpus. Unlike in a classical MSA-based workflow, sequences must be arranged into groups, but subsequent steps are AF. Sequence regions that contain *k*-mers infrequent in their own group (TF) but frequent in another group (IDF) are inferred as instances of LGT from the donor group to the recipient sequence [33]. Our unsuccessful idea above (Figure 3) represented IDF without proper TF.

The resulting workflow (Figure 4) differs from AF workflows for purely vertical phylogenetics (e.g. Figure 2) in two main ways: the unit of analysis is not specified up front, and sequences must be arranged into groups. Potential lateral segments are generated by merging *k*-mers that meet the IDF and TF requirements. A parameter *G* specifies the maximum allowable gap between *k*-mers to be merged into a lateral segment; where investigated, the number of LGT detections and total detection length were relatively insensitive to *G*. The resulting segments are typically of different lengths, and may map to intergenic regions, gene fragments, entire and/or multiple genes. In contrast, grouping the sequences in an effective manner proved to be non-trivial, yet critical to performance [33, 34, 99]. TF–IDF performs best when sequences are similar within group but dissimilar between groups; so if our goal is to infer LGT, the best grouping will probably capture hierarchical descent. Even so, it may remain 'difficult to disentangle the effects of group number, size, composition and phylogenetic cohesion' [34].

With simulated data, it was obvious how to delineate groups and TF–IDF performed well, as measured by precision and recall, over a biologically realistic range of sequence lengths and evolutionary distances between and within groups. As expected, greater evolution post-LGT had a deleterious effect on performance, while deletions showed relatively little effect. With empirical data, groups (e.g. γ-proteobacteria) known to engage in LGT were usually prominent in our TF–IDF analyses, while we inferred little or no LGT for groups known to be more quiescent [34]. Because only genes (not arbitrary regions) have gene ontology annotation, to study the functional implications of the inferred LGT, we mapped the inferred lateral segments to genes, using data-dependent length and overlap thresholds [34, 99]. For protein-coding genes, we might alternatively have asked whether the inferred lateral segments overlap regions that encode active sites or SCOP domains.

For groups related by a hierarchical tree, it may be possible to extract further information. If a genomic region is inferred to have received genetic material from two or more groups that are topologically adjacent on the tree, we might (depending on details) instead hypothesize that there had been a single transfer from a common ancestor of the donor groups. Imperfect overlap of the inferred lateral regions could be ascribed to the vagaries of subsequent evolution, and/or the IDF threshold being a blunt instrument. On the other hand, transfers from unrelated donor groups would render such a region an

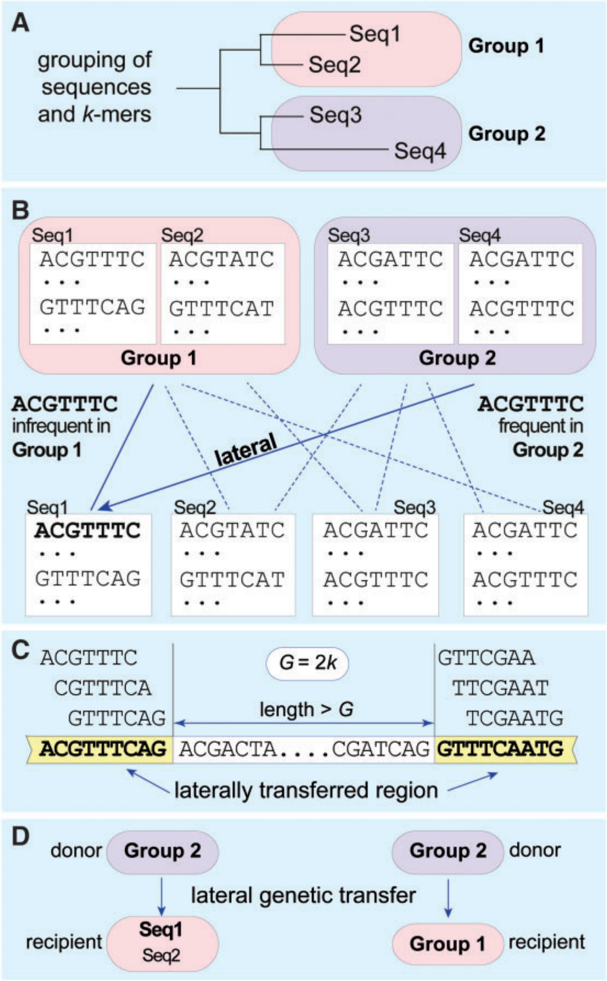


Figure 4. Simplified workflow illustrating the use of TF–IDF to identify lateral genetic transfer. (A) Four sequences (Seq1 through Seq4) are grouped, here into two groups (Group 1 and Group 2) based on a reference tree. (B) All *k*-mers ($k = 7$, stride = 1) from each sequence are compared against the *k*-mers found in each of the two groups. A *k*-mer that is infrequent in the group to which the sequence belongs (TF), but frequent in another group (IDF), illustrated here by ACGTTTC in Seq1 that is infrequent in Group 1 but frequent in Group 2, is inferred to be of lateral origin. (C) Laterally transferred regions are constructed from sets of nearby lateral *k*-mers, where nearby means separated by ≤gap *G*. For representation as a network, recipient sequences are subsumed into their respective groups with the result that transfers inferred from a donor group to a recipient sequence (D, left) are shown as from a donor group to a recipient group (D, right). For clique analysis, edge weight and directionality may further be ignored (see text).

evolutionary mosaic [34]. Although gene loss and LGT among the donor lineages may present further complications, TF–IDF seems to promise a first-ever systematic look at the temporal dynamics of superposed transfers.

Summary information from TF–IDF analysis can be collected in the form of an LGT network. For simplicity of interpretation, recipient sequences are subsumed into their respective groups, and inferred transfer events consolidated as weights on the edges. Given the limitations of most clique-finding algorithms, the weights and directionality of edges are usually ignored [99]. Densely connected regions within an LGT graph—maximum cliques, maximal cliques and paracliques (cliques missing a few edges)—can be extracted using GrAPPA [100]. These structures demarcate genetic exchange communities (GECs), groups of taxa whose members have shared genetic material among themselves by LGT [101]. Taxa that retain membership across biologically reasonable values of *k* (for the examples cited, $20 \le k \le 40$) are considered core nodes of these GECs [99]. Other structures in an LGT graph may also be of biological interest, e.g. bridging nodes that connect cliques [102, 103]. By annotating nodes and edges with metadata, e.g. on environment, genome type or vector, new perspectives may be gained on the genetic structure of the microbial biosphere, and on genetic flow within and across 'independent genetic worlds' [102, 104–107].

The TF–IDF algorithm is scalable, running in $O(nL \cdot \log(nL))$ time where *n* is the number of sequences and *L* their average length. Moreover, as the inferred edges are natively lateral and directional, computationally hard steps involving the generation of a reference topology and comparison with test trees are obviated. However, in its current implementation, TF–IDF is somewhat greedy of memory, preventing its application to very large data sets. Clique finding is computationally demanding, even though information on edge directionality and weight is typically ignored.

An analogous approach could be taken to identify regions of vertical inheritance. Wong and Ragan [108] recognized core regions that find matches in other sequences, extended these regions using a criterion of mutual exclusivity, built a pairwise similarity graph and applied MCL [109] to yield sets of putatively homologous subsequences they called Markov Clusters of Homologous Subsequences (MACHOS). In place of Smith–Waterman, match and extension criteria based on *k*-mers (above) could equally well be used. MACHOS correspond well to known Pfam domain families [108], and offer an AF approach to recognition of orthologs [110]. It has been argued that workflows in which a gene or protein family is, by default, considered to be inherited vertically unless this null hypothesis is specifically rejected gives a conceptually and methodologically unfair advantage to vertical inheritance [111, 112]. Doolittle [111] goes so far as to call this a 'false null'. Parallel AF workflows for lateral and vertical regions could address this objection, inferring 'LGT directly, positively and fairly in large genome-scale datasets' [99].

Conclusions

The power of *k*-mer-based AF approaches relies on proper selection of *k*. The requirement that *k*-mers be approximately unique to a sequence can be satisfied at a much smaller *k* for amino acids (alphabet size 20) than for nucleotides (alphabet size 4). For tree inference, optimal *k* depends on the length and divergence of the sequences, and (more weakly) on the inference method. In our hands, *k* is optimal at about 3–5 for proteins, and 8–10 for genes or RNAs [32, 63, 71]. We set $k = 12$ for a quick assessment of the relative divergence of microbial genome data sets [34, 73], while Greenfield and Roehm [31] used $k > 15$ to identify organisms, genes and functions of interest using unique *k*-mers as tags. For genome trees, optimal *k* ranged from 8 for isolates of the same bacterial species up to 25 across bacteria and archaea [63, 73]. Elhai *et al.* [92] used $k = 8$ to detect genes of recent lateral origin in microbial genomes, but needed to draw on additional lines of evidence to make their approach effective. In TF–IDF, optimal *k* is larger still, for microbial genome data sets in the range 25–40.

AF approaches are beginning to make their mark in phylogenomics and LGT research. Substrings can readily be extracted from sequences, indexed, stored and retrieved. They capture homology signal in evolving sequences, and counts or frequencies of shared *k*-mers can underpin measures of pairwise distance and the computation of distance trees. Distributions of *k*-mers among groups of genomes can reveal donor–recipient relationships in LGT, hence communities of genetic exchange, and may be informative on the temporal dynamics of reticulate evolution. Like their MSA-based counterparts, *k*-mer distance trees can be computed quickly and scale to very large data, without the computational overhead of a complex substitution model (or multiple such models for different sequence regions). Whether this is, on balance, a good thing remains to be seen, even apart from the question of whether it makes sense to infer a genome tree [113]. MSA-based methods have benefitted from more than four decades of development, in the process enriching all their component fields, biological and otherwise. In contrast, AF phylogenomics is still in its infancy. We anticipate that AF methods will mature to provide dependable options in large-scale phylogenomics while stimulating the exploration of other biological questions previously unimaginable within the classical framework.

Key Points

- Molecular sequences can be represented by sets of their constituent *k*-mers. To the extent that these *k*-mer sets capture the signal of homology among these sequences, they can inform on phylogenetic relationships.
- Measures of the intersections of these *k*-mer sets can be used, after normalization, to compute pairwise distances without the need for MSA.
- Trees computed from *k*-mer distances are often biologically reasonable, e.g. recovering recognized taxa, while being robust against evolutionary scenarios that are problematic for alignment-based phylogenetics.
- Using TF–IDF, it is possible to identify regions that have been transferred from a donor group into a recipient sequence. Across a data set, all such pairwise relationships describe an LGT network. Densely connected regions in LGT networks can be interpreted as GECs.
- These AF approaches are computationally fast and scalable to large nucleotide or amino acid data sets.

Acknowledgements

Computational analyses were carried out on high-performance computing systems provided by the National Computational Infrastructure (NCI), Queensland Cyber Infrastructure Foundation (QRIScloud) and the Institute for Molecular Bioscience. Allocation of resources on the NCI National Facility and QRIScloud was supported by the NCI Merit Allocation Scheme (project D85).

Funding

The James S. McDonnell Foundation and The University of Queensland (awards to M.A.R.), and Australian Research Council (grant number DP150101875) (awards to M.A.R. and C.X.C.). The China Scholarship Council (to Y.C.).

References

1. Delsuc F, Brinkmann H, Philippe H. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* 2005;**6**:361–75.
2. Eisen JA, Fraser CM. Phylogenomics: intersection of evolution and genomics. *Science* 2003;**300**:1706–7.
3. Pollock DD, Eisen JA, Doggett NA, *et al*. A case for evolutionary genomics and the comprehensive examination of sequence biodiversity. *Mol Biol Evol* 2000;**17**:1776–88.
4. Sicheritz-Ponten T, Andersson SG. A phylogenomic approach to microbial evolution. *Nucleic Acids Res* 2001;**29**: 545–52.
5. Ragan MA, Bernard G, Chan CX. Molecular phylogenetics before sequences: oligonucleotide catalogs as *k*-mer spectra. *RNA Biol* 2014;**11**:176–85.
6. Margoliash E. Homology: a definition. *Science* 1969;**163**:127.
7. Feng DF, Doolittle RF. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol* 1987;**25**:351–60.
8. Carrillo H, Lipman D. The multiple sequence alignment problem in biology. *SIAM J Appl Math* 1988;**48**:1073–82.
9. Chan CX, Ragan MA. Next-generation phylogenomics. *Biol Direct* 2013;**8**:3.
10. Fan H, Ives AR, Surget-Groba Y, *et al*. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC Genomics* 2015;**16**:522.
11. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 2010;**5**:e11147.
12. Earl D, Nguyen N, Hickey G, *et al*. Alignathon: a competitive assessment of whole-genome alignment methods. *Genome Res* 2014;**24**:2077–89.
13. Tavaré S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect Math Life Sci* 1986;**17**:57–86.
14. Beiko RG, Harlow TJ, Ragan MA. Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci USA* 2005;**102**:14332–7.
15. Gogarten JP, Townsend JP. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* 2005;**3**:679–87.
16. Chan CX, Beiko RG, Darling AE, *et al*. Lateral transfer of genes and gene fragments in prokaryotes. *Genome Biol Evol* 2009;**1**:429–38.
17. Chan CX, Darling AE, Beiko RG, *et al*. Are protein domains modules of lateral genetic transfer? *PLoS One* 2009;**4**:e4524.
18. Lawrence JG, Ochman H. Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* 1997;**44**:383–97.
19. Gogarten JP, Doolittle WF, Lawrence JG. Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* 2002;**19**:2226–38.
20. Skippington E, Ragan MA. Within-species lateral genetic transfer and the evolution of transcriptional regulation in *Escherichia coli* and *Shigella*. *BMC Genomics* 2011;**12**:532.
21. Hide W, Burke J, Davison DB. Biological evaluation of d^2, an algorithm for high-performance sequence comparison. *J Comput Biol* 1994;**1**:199–215.
22. Myers EW. A sublinear algorithm for approximate keyword searching. *Algorithmica* 1994;**12**:345–74.
23. Miller RT, Christoffels AG, Gopalakrishnan C, *et al*. A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res* 1999;**9**:1143–55.
24. Sameith K, Roscito JG, Hiller M. Iterative error correction of long sequencing reads maximizes accuracy and improves contig assembly. *Brief Bioinform* 2017;**18**:1–8.
25. Compeau PE, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol* 2011;**29**:987–91.
26. Lee D, Gorkin DU, Baker M, *et al*. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet* 2015;**47**:955–61.

27. Lee D, Karchin R, Beer MA. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res* 2011;**21**:2167–80.
28. Haubold B, Krause L, Horn T, *et al*. An alignment-free test for recombination. *Bioinformatics* 2013;**29**:3121–7.
29. Haubold B. Alignment-free phylogenetics and population genetics. *Brief Bioinform* 2014;**15**:407–18.
30. Vinga S, Almeida J. Alignment-free sequence comparison—a review. *Bioinformatics* 2003;**19**:513–23.
31. Greenfield P, Roehm U. Answering biological questions by querying *k*-mer databases. *Concurr Comput Pract Exper* 2013;**25**:497–509.
32. Höhl M, Rigoutsos I, Ragan MA. Pattern-based phylogenetic distance estimation and tree reconstruction. *Evol Bioinform Online* 2006;**2**:359–75.
33. Cong Y, Chan YB, Ragan MA. A novel alignment-free method for detection of lateral genetic transfer based on TF–IDF. *Sci Rep* 2016;**6**:30308.
34. Cong Y, Chan YB, Ragan MA. Exploring lateral genetic transfer among microbial genomes using TF–IDF. *Sci Rep* 2016;**6**:29319.
35. Rigoutsos I, Huynh T, Floratos A, *et al*. Dictionary-driven protein annotation. *Nucleic Acids Res* 2002;**30**:3901–16.
36. Maetschke SR, Kassahn KS, Dunn JA, *et al*. A visual framework for sequence analysis using *n*-grams and spectral rearrangement. *Bioinformatics* 2010;**26**:737–44.
37. Giegerich R, Kurtz S. From Ukkonen to McCreight and Weiner: a unifying view of linear-time suffix tree construction. *Algorithmica* 1997;**19**:331–53.
38. Manber U, Myers G. Suffix arrays: a new method for on-line string searches. *SIAM J Comput* 1993;**22**:935–48.
39. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* 2011;**27**:764–70.
40. Greenfield P, Duesing K, Papanicolaou A, *et al*. Blue: correcting sequencing errors using consensus and context. *Bioinformatics* 2014;**30**:2723–32.
41. Melsted P, Pritchard JK. Efficient counting of *k*-mers in DNA sequences using a bloom filter. *BMC Bioinformatics* 2011;**12**:333.
42. Bloom BH. Space/time trade-offs in hash coding with allowable errors. *Commun ACM* 1970;**13**:422–6.
43. Holley G, Wittler R, Stoye J. Bloom Filter Trie: an alignment-free and reference-free data structure for pan-genome storage. *Algorithms Mol Biol* 2016;**11**:3.
44. Delcher AL, Kasif S, Fleischmann RD, *et al*. Alignment of whole genomes. *Nucleic Acids Res* 1999;**27**:2369–76.
45. Miller JR, Delcher AL, Koren S, *et al*. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* 2008;**24**: 2818–24.
46. Kurtz S, Narechania A, Stein JC, *et al*. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics* 2008;**9**:517.
47. Jun SR, Sims GE, Wu GA, *et al*. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: an alignment-free method with optimal feature resolution. *Proc Natl Acad Sci USA* 2010;**107**:133–8.
48. Reinert G, Chew D, Sun F, *et al*. Alignment-free sequence comparison (I): statistics and power. *J Comput Biol* 2009;**16**:1615–34.
49. Russell DJ, Way SF, Benson AK, *et al*. A grammar-based distance metric enables fast and accurate clustering of large sets of 16S sequences. *BMC Bioinformatics* 2010;**11**:601.
50. Wang H, Xu Z, Gao L, *et al*. A fungal phylogeny based on 82 complete genomes using the composition vector method. *BMC Evol Biol* 2009;**9**:195.
51. Bromberg R, Grishin NV, Otwinowski Z. Phylogeny reconstruction with alignment-free method that corrects for horizontal gene transfer. *PLoS Comput Biol* 2016;**12**:e1004985.
52. Göke J, Schulz MH, Lasserre J, *et al*. Estimation of pairwise sequence similarity of mammalian enhancers with word neighbourhood counts. *Bioinformatics* 2012;**28**:656–63.
53. Leimeister CA, Sohrabi-Jahromi S, Morgenstern B. Fast and accurate phylogeny reconstruction using filtered spaced-word matches. *Bioinformatics* 2017;**33**:971–9.
54. Yi H, Jin L. Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic Acids Res* 2013;**41**:e75.
55. Ulitsky I, Burstein D, Tuller T, *et al*. The average common substring approach to phylogenomic reconstruction. *J Comput Biol* 2006;**13**:336–50.
56. Haubold B, Pfaffelhuber P, Domazet-Lošo M, *et al*. Estimating mutation distances from unaligned genomes. *J Comput Biol* 2009;**16**:1487–500.
57. Torney DC, Burks C, Davison D, *et al*. Computation of D2—a measure of sequence dissimilarity. In: G Bell, R Marr (eds). *Computers and DNA: the Proceedings of the Interface between Computation Science and Nucleic Acid Sequencing Workshop*. Santa Fe: Addison-Wesley, 1990, 109–25.
58. Forêt S, Wilson SR, Burden CJ. Characterizing the D2 statistic: word matches in biological sequences. *Stat Appl Genet Mol Biol* 2009;**8**:43.
59. Lippert RA, Huang H, Waterman MS. Distributional regimes for the number of *k*-word matches between two random sequences. *Proc Natl Acad Sci USA* 2002;**99**:13980–9.
60. Song K, Ren J, Reinert G, *et al*. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Brief Bioinform* 2014;**15**:343–53.
61. Blaisdell BE. A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc Natl Acad Sci USA* 1986;**83**:5155–9.
62. Wan L, Reinert G, Sun F, *et al*. Alignment-free sequence comparison (II): theoretical power of comparison statistics. *J Comput Biol* 2010;**17**:1467–90.
63. Chan CX, Bernard G, Poirion O, *et al*. Inferring phylogenies of evolving sequences without multiple sequence alignment. *Sci Rep* 2014;**4**:6504.
64. Burden CJ, Jing J, Wilson SR. Alignment-free sequence comparison for biologically realistic sequences of moderate length. *Stat Appl Genet Mol Biol* 2012;**11**:3.
65. Burden CJ, Leopardi P, Forêt S. The distribution of word matches between Markovian sequences with periodic boundary conditions. *J Comput Biol* 2014;**21**:41–63.
66. Chor B, Horn D, Goldman N, *et al*. Genomic DNA *k*-mer spectra: models and modalities. *Genome Biol* 2009;**10**:R108.
67. Liu B, Shi Y, Yuan J, *et al*. Estimation of genomic characteristics by analyzing *k*-mer frequency in de novo genome projects. *arXiv* 2013;1308.2012.
68. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;**4**:406–25.
69. Studier JA, Keppler KJ. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol Biol Evol* 1988;**5**:729–31.
70. Felsenstein J. *Inferring Phylogenies*. Sunderland MA: Sinauer, 2004.
71. Höhl M, Ragan MA. Is multiple-sequence alignment required for accurate inference of phylogeny? *Syst Biol* 2007;**56**:206–21.
72. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 2007;**24**:1586–91.

73. Bernard G, Chan CX, Ragan MA. Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer. *Sci Rep* 2016;**6**:28970.
74. Horwege S, Lindner S, Boden M, *et al*. *Spaced words* and *kmacs*: fast alignment-free sequence comparison based on inexact word matches. *Nucleic Acids Res* 2014;**42**:W7–11.
75. Sims GE, Kim SH. Whole-genome phylogeny of *Escherichia coli/Shigella* group by feature frequency profiles (FFPs). *Proc Natl Acad Sci USA* 2011;**108**:8329–34.
76. Bernard G, Ragan MA, Chan CX. Recapitulating phylogenies using *k*-mers: from trees to networks [version 2; referees: 2 approved]. *F1000Res* 2016;**5**:2789.
77. Beiko RG, Ragan MA. Detecting lateral genetic transfer: a phylogenetic approach. *Methods Mol Biol* 2008;**452**:457–69.
78. Chan CX, Beiko RG, Ragan MA. Scaling up the phylogenetic detection of lateral gene transfer events. *Methods Mol Biol* 2017;**1525**:421–32.
79. Chan CX, Beiko RG, Ragan MA. Detecting recombination in evolving nucleotide sequences. *BMC Bioinformatics* 2006;**7**:412.
80. Clarke GD, Beiko RG, Ragan MA, *et al*. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *J Bacteriol* 2002;**184**:2072–80.
81. Ragan MA. Detection of lateral gene transfer among microbial genomes. *Curr Opin Genet Dev* 2001;**11**:620–6.
82. Ragan MA. On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol Lett* 2001;**201**:187–91.
83. Dalquen DA, Anisimova M, Gonnet GH, *et al*. ALF—a simulation framework for genome evolution. *Mol Biol Evol* 2012;**29**:1115–23.
84. Beiko RG, Charlebois RL. A simulation test bed for hypotheses of genome evolution. *Bioinformatics* 2007;**23**:825–31.
85. Maetschke SR, McIntyre L, Chan CX, *et al*. LGTNet: fast inference of lateral genetic transfer networks. http://bioinformatics.org.au/tools/LGTNet/ (26 April 2017, date last accessed).
86. Becq J, Churlaud C, Deschavanne P. A benchmark of parametric methods for horizontal transfers detection. *PLoS One* 2010;**5**:e9989.
87. Dufraigne C, Fertil B, Lespinats S, *et al*. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res* 2005;**33**:e6.
88. Garcia-Vallvé S, Romeu A, Palau J. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res* 2000;**10**:1719–25.
89. Lawrence JG, Ochman H. Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci USA* 1998;**95**:9413–7.
90. Médigue C, Rouxel T, Vigier P, *et al*. Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J Mol Biol* 1991;**222**:851–6.
91. Ragan MA, Harlow TJ, Beiko RG. Do different surrogate methods detect lateral genetic transfer events of different relative ages? *Trends Microbiol* 2006;**14**:4–8.
92. Elhai J, Liu H, Taton A. Detection of horizontal transfer of individual genes by anomalous oligomer frequencies. *BMC Genomics* 2012;**13**:245.
93. Robinson PMW, O'Hara RJ. Cladistic analysis of an Old Norse manuscript tradition. *Res Human Comput* 1996;**4**:115–37.
94. Howe CJ, Barbrook AC, Spencer M, *et al*. Manuscript evolution. *Trends Genet* 2001;**17**:147–52.
95. Marmerola GD, Oikawa MA, Dias Z, *et al*. On the reconstruction of text phylogeny trees: evaluation and analysis of textual relationships. *PLoS One* 2016;**11**:e0167822.
96. Sims GE, Jun SR, Wu GA, *et al*. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc Natl Acad Sci USA* 2009;**106**:2677–82.
97. Lee AR. Numerical taxonomy revisited: John Griffith, cladistic analysis and St. Augustine's *Quaestiones in Heptateuchum*. *Studia Patristica* 1989;**20**:24–32.
98. Ragan MA, Lee AR. Making phylogenetic sense of biochemical and morphological diversity among the protists. In: TR Dudley (ed). *The Unity of Evolutionary Biology, Proceedings of Fourth International Congress of Systematic and Evolutionary Biology*. College Park MD: Dioscorides Press, 1991, 432–441.
99. Cong Y, Chan YB, Phillips CA, *et al*. Robust inference of genetic exchange communities from microbial genomes using TF–IDF. *Front Microbiol* 2017;**8**:21.
100. Lu Y, Phillips CA, Langston MA. GrAPPA. Graph Algorithms Pipeline for Pathway Analysis. https://grappa.eecs.utk.edu/ (26 April 2017, date last accessed).
101. Skippington E, Ragan MA. Lateral genetic transfer and the construction of genetic exchange communities. *FEMS Microbiol Rev* 2011;**35**:707–35.
102. Halary S, Leigh JW, Cheaib B, *et al*. Network analyses structure genetic diversity in independent genetic worlds. *Proc Natl Acad Sci USA* 2010;**107**:127–32.
103. Liu W, Pellegrini M, Wang X. Detecting communities based on network topology. *Sci Rep* 2014;**4**:5739.
104. Dagan T, Martin W. Getting a better picture of microbial evolution en route to a network of genomes. *Philos Trans R Soc Lond B Biol Sci* 2009;**364**:2187–96.
105. Fondi M, Fani R. The horizontal flow of the plasmid resistome: clues from inter-generic similarity networks. *Environ Microbiol* 2010;**12**:3228–42.
106. Koonin EV. The turbulent network dynamics of microbial evolution and the statistical Tree of Life. *J Mol Evol* 2015;**80**:244–50.
107. Puigbò P, Wolf YI, Koonin EV. The tree and net components of prokaryote evolution. *Genome Biol Evol* 2010;**2**:745–56.
108. Wong S, Ragan MA. MACHOS: Markov clusters of homologous subsequences. *Bioinformatics* 2008;**24**:i77–85.
109. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 2002;**30**:1575–84.
110. Shin CJ, Davis MJ, Ragan MA. Towards the mammalian interactome: inference of a core mammalian interaction set in mouse. *Proteomics* 2009;**9**:5256–66.
111. Doolittle WF. The practice of classification and the theory of evolution, and what the demise of Charles Darwin's tree of life hypothesis means for both of them. *Philos Trans R Soc Lond B Biol Sci* 2009;**364**:2221–8.
112. Doolittle WF, Bapteste E. Pattern pluralism and the Tree of Life hypothesis. *Proc Natl Acad Sci USA* 2007;**104**:2043–9.
113. Doolittle WF. Lateral gene transfer, genome surveys, and the phylogeny of prokaryotes. Response from Doolittle. *Science* 1999;**286**:1443a.