Insertions and Deletions: Computational Methods, Evolutionary Dynamics, and Biological Applications

Benjamin D. Redelings [1] Ian Holmes [2,3] Gerton Lunter [4] Tal Pupko [5]
Maria Anisimova [6,7,*]

[1] Department of Mathematics, Tulane University, New Orleans, LA 70118, USA
[2] Department of Bioengineering, University of California, Berkeley, CA 94720, USA
[3] Calico Life Sciences LLC, South San Francisco, CA 94080, USA
[4] Department of Epidemiology, University Medical Center Groningen, University of Groningen, Groningen 9713 GZ, The Netherlands
[5] The Shmunis School of Biomedicine and Cancer Research, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel
[6] Institute of Computational Life Sciences, Zurich University of Applied Sciences, Wädenswil, Switzerland
[7] Swiss Institute of Bioinformatics, Lausanne, Switzerland

***Corresponding author:** E-mail: maria.anisimova@zhaw.ch.

Associate editor: Brandon Gaut

Abstract

Insertions and deletions constitute the second most important source of natural genomic variation. Insertions and deletions make up to 25% of genomic variants in humans and are involved in complex evolutionary processes including genomic rearrangements, adaptation, and speciation. Recent advances in long-read sequencing technologies allow detailed inference of insertions and deletion variation in species and populations. Yet, despite their importance, evolutionary studies have traditionally ignored or mishandled insertions and deletions due to a lack of comprehensive methodologies and statistical models of insertions and deletion dynamics. Here, we discuss methods for describing insertions and deletion variation and modeling insertions and deletions over evolutionary time. We provide practical advice for tackling insertions and deletions in genomic sequences and illustrate our discussion with examples of insertions and deletion-induced effects in human and other natural populations and their contribution to evolutionary processes. We outline promising directions for future developments in statistical methodologies that would allow researchers to analyze insertions and deletion variation and their effects in large genomic data sets and to incorporate insertions and deletions in evolutionary inference.

Key words: **indel, insertion, deletion, evolution, phylogeny, alignment.**

Introduction

After point mutations, insertions and deletions (a.k.a. indels) constitute the second most important source of genomic variation in populations and species. Indel events appear to dominate the early stages of species divergence (Britten et al. 2003), and in human populations, between 16% and 25% of all genomic variations are indels, many of which are functionally important (Mills et al. 2006; Mullaney et al. 2010). Consequently, indels could be used to infer patterns informative of human diseases, thus potentially discovering indel biomarkers (Sehn 2015). Indel diversity is often understudied in viral sequences, due to their presumed highly deleterious effects, as well as technical difficulties. However, recent reports show that indel patterns can shed light on viral dynamics and evolution (Elena 2023), with indels having important phenotypic consequences. Several studies have evaluated and reported natural selection on indels in a variety of species (Haerty and Golding 2010; Mularoni et al. 2010; Barton and Zeng 2019). At deeper evolutionary divergences, indels lead to structural variation in protein superfamilies affecting binding, catalysis, and protein–protein interaction (Copley 2010). In some cases, multiple indels seem to have little effect on the overall folded protein structure, yet they noticeably affect the interaction between different proteins (Sandhya et al. 2009; Studer et al. 2013).

The treatment of indels in key bioinformatic tasks has direct influence on the quality of all downstream inferences. In particular, one such core task is multiple sequence alignment (MSA), since it is used extensively in genomics, evolutionary

Received: April 10, 2024. **Revised:** July 02, 2024. **Accepted:** July 09, 2024

biology, biochemistry, microbiology, and medicine. A few examples for cases in which MSAs are widely used are as follows: (i) phylogenetic tree inference (Kapli et al. 2020); (ii) detecting the selection regime of specific regions within proteins, genes, and genomes, e.g. patterns of positive selection (Ashkenazy et al. 2010; Kosiol and Anisimova 2019); (iii) the inference of remote homology (Eddy 1998); (iv) molecular dating (dos Reis et al. 2016); (v) structural predictions (Jumper et al. 2021); (vi) inference of lateral gene transfer (Dagan 2011); (vii) ancestral sequence reconstruction (Selberg et al. 2021); (viii) the detecting of coevolving sites (de Juan et al. 2013); (ix) genome assembly (Rice and Green 2019); and (x) whole-genome alignment and genome annotation (Angiuoli et al. 2011). The inference of MSA and the study of indel dynamics are deeply intertwined because an MSA implicitly reflects inference regarding the assignment of indel events along the evolutionary history that led to the analyzed sequences. Thus, better inference of indel dynamics should, in theory, lead to more accurate inferred MSAs.

In MSAs, indels create gap patterns that can be highly informative for phylogenetic inference, providing a strong signal that saturates at substantially deeper divergences compared to point mutations, and that can help to resolve disputed species relationships (Rokas and Holland 2000; Simmons and Ochoterena 2000; Belinky et al. 2010). However, the use of indel information poses a challenge because indels are not directly observed and because incorrectly placed indels can mislead inference (Westesson et al. 2012). A large-scale study (Dessimoz and Gil 2010) used real data from eukaryotes, fungi, and bacteria to evaluate the effect of gap placement in MSAs on the accuracy of phylogeny inference. They discovered that by excluding gap-rich and variable regions, valuable information from substitutions and gap patterns is often discarded, which can be detrimental. Other studies have shown that indels provide a significant phylogenetic signal (Birth et al. 2022) and should be better utilized by methods for MSA and tree inference. Phylogenetic information from indels can be used not only to study deep evolutionary history (Rivera and Lake 1992; Rokas and Holland 2000) but also to distinguish closely related species (Gaya et al. 2011). Indels may also prove useful for studies within species, for example for medical applications in human (Mullaney et al. 2010). More generally, indel markers are increasingly used to study population diversity and structure (Yang et al. 2014; Lü et al. 2015; Zhou et al. 2015; Vishwakarma et al. 2017; Jain et al. 2019), with a variety of applications, such as in germplasm management for purposes of conservation and crop improvement (Wang, Zhou, et al. 2023). In addition, indel calling is often done by mapping short reads to a reference genome, for example in bacteria where the inferred indel variation can shed light on processes such as the acquisition of antibiotic resistance and reveal epistatic interactions that involve both indels and substitutions (Godfroid et al. 2020).

There is also increased awareness of the importance of indels in ancestral sequence reconstruction, prompting a push for new methods. Inferred indel histories can be highly informative for studies of gene and protein evolution (Savino et al. 2022), providing insights into protein engineering strategies (Boersma and Plückthun 2011). Finally, indels are a major contributor to functional protein changes (Lin et al. 2017), incident disease and disease susceptibility within a population (Roos 2010; Ferlaino et al. 2017; Dai et al. 2020; Kundu et al. 2022).

Overall, genome research across a variety of species demonstrates that indels provide substantial insight into evolutionary relationships, trait discovery, and drug resistance and have multifaceted applications, such as in protein bioengineering, forensics, and breeding. Nevertheless, dealing with indels is challenging, and therefore, most studies tend to focus on point mutations (or substitutions) and remove or heavily trim gap-rich regions. There is no commonly accepted gold standard for treating indels. The simple reason for this is the inherent difficulty in properly handling complex gap patterns that arise over time due to indels, overlapping in space and altering sequence length (Redelings and Suchard 2009). Disentangling individual indel histories based on the observed gap distributions in a sequence alignment requires stochastic models of sequence evolution that explicitly include indel events over time. The first stochastic evolutionary models with indels were proposed over 30 years ago (Thorne et al. 1991, 1992), demonstrating the computational complexity of the problem, and subsequent development has focused on further improving model accuracy while keeping computations tractable. In this review, we summarize recent and notable models and methods that enable researchers to properly exploit patterns generated by indels in phylogenetic inferences and downstream applications. We consider methodological assumptions and potential issues with different approaches and discuss practical considerations in phylogenetic inference with indels. Finally, we outline future directions and call for closer collaborations between method developers and experimental and evolutionary biology experts.

Indels: Definition, Mutagenesis, and Representation

Mutations in genome sequences are often classified, from small to large, into single-nucleotide variants (SNVs), indels, and structural variants (SVs). Here, SNVs are substitutions of one nucleotide for another, while indels are typically defined as local indels of “short” DNA segments, and SVs involve longer sequence segments, or modify a genome more globally, e.g. through inversions or translocations. The meaning of “short” here is arbitrary, and sizes of up to 10 kb have been used (Mills et al. 2011; Sehn 2015) although a threshold of 50 bp is commonly accepted (Montgomery et al. 2013; Mahmoud et al. 2019; Ebert et al. 2021). In particular, this puts transposable element (TE) insertions into the class of SVs. First discovered by Barbara McClintock in maize (McClintock 1950), TEs are selfish genetic elements of 100 to over 10 kb in size (Wells and Feschotte 2020) that constitute a substantial fraction of many species’ genomes (Schnable et al. 2009) and are important drivers of evolution (Kazazian 2004) as well as reliable phylogenetic markers (Lunter 2007; Nystedt et al. 2013; Jarvis et al. 2014).

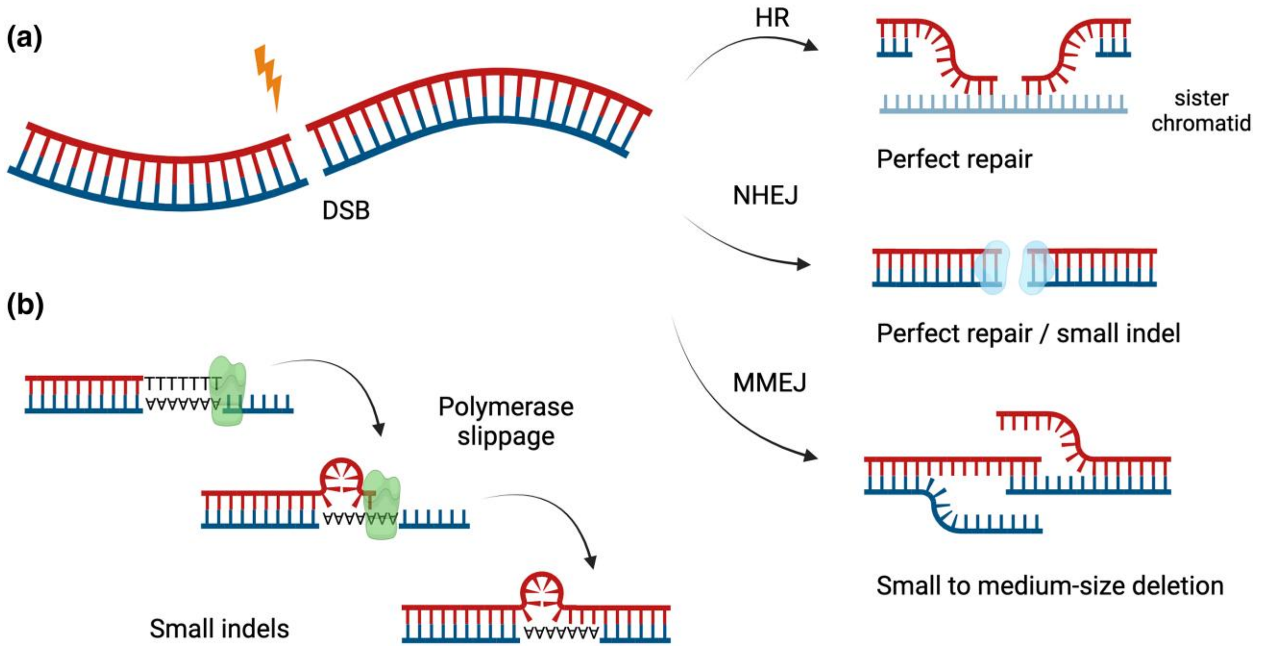

Fig. 1. Main mechanisms for indel formation. a) Double-stranded breaks (DSBs) can result from various mutagenic mechanisms, as well as normal cell processes, and are usually resolved without mutation by HR. Complementary repair pathways include NHEJ often leading to small indels and MMEJ, which typically results in larger indels. b) Indels also result from errors during replication. Polymerase slippage is a major cause for indels in repetitive DNA.

However, in this review, we focus on short indels, and we will not discuss TEs or other SVs further.

While we have different terms for single base pair mutations and substitutions (single base pair mutations that have become fixed in a population), we do not have similar terms for indels, and the term indel is used to refer to both indel mutations and indels that became fixed in a population. We also often use the term "gap" to indicate the representation of indel events in an alignment. We often lump indels together, since the observed alignment gaps can be caused by either. However, when we discuss mutational processes, we often distinguish indels.

Which mutation processes are known to lead to indels? A common source of indels is DNA slippage during replication, resulting from denaturation followed by mispairing of the nascent DNA strand (Levinson and Gutman 1987). Since mispairing is more likely in repetitive regions, slippage occurs frequently in *short tandem repeats* (STRs), i.e. multiple adjacent copies of short sequence motifs, also known as *microsatellites*. Conversely, the process of DNA replication slippage is thought to be responsible for the abundance of STRs in many genomes (Levinson and Gutman 1987). A similar process, termed template switching, is responsible for more complex mutations, including length-preserving micro-inversions (Chaisson et al. 2006; Cooke et al. 2021) and mutation clusters (Löytynoja and Goldman 2017).

A second cause of indels is DNA damage followed by imperfect repair. Damage, such as double-stranded breaks, occurs naturally due to metabolic processes, transcription, and replication stresses requiring untangling of DNA by topoisomerase, as well as through exogenous sources such as radiation and toxins (Mehta and Haber 2014). In most cases, the resulting lesions are repaired by one of several homologous repair (HR) pathways, which use a sister chromatid as template (Mehta and Haber 2014), generally leading to complete repair. In nondividing cells, this pathway is not accessible, and an alternative repair pathway, nonhomologous end joining (NHEJ) is used instead. NHEJ can handle most lesions and often leads to perfect repair (Bétermier et al. 2014; Bhargava et al. 2018), and if not results in small (few base pairs) indels often flanked by small (1 to 2 bp) microhomologies (Bennett et al. 2020). A third mechanism, microhomology-mediated end joining (MMEJ) serves as a fallback mechanism that is more error prone and requires longer (1 to 16 bp) microhomologies for repair (Sfeir and Symington 2015). An overview of these mechanisms is shown in Fig. 1. A range of other mutagenic mechanisms, including nonallelic homologous recombination and template switching during replication (FoSTeS) typically result in large structural variation but also contribute indels of all sizes, including small ones (Burssed et al. 2022).

The overall result of these mutagenic and repair mechanisms is that short indels are mostly found in repetitive regions (such as STRs), are most frequently very short (1 to 2 bp), and are biased toward deletions, while insertions are typically copies of a nearby (often adjacent) sequence (Verbiest et al. 2022).

A single nucleotide variant with respect to a reference sequence is naturally represented as a base change (e.g. T → A) at a certain position (e.g. hg18.chr1:10001), and this is essentially how SNVs are represented in the variant call format (VCF) (Danecek et al. 2011). However, it is less straightforward to represent indels as occurring at a specific base position in a reference genome. Is CAT → CAAT an insertion of an A next to the C or next to the existing A? For alignments, the equivalent question is whether to left- or right-align gaps. What if the haplotype carrying the indel also carries a C → A (or T → A) SNV? The VCF format has disambiguated the representation in many cases, but issues remain for complex haplotypes, which affect for instance the benchmarking of indel calling pipelines (Krusche et al. 2019; Cooke et al. 2021).

One way forward is to use more flexible graph-based representations, such as partial order graphs (POGs) introduced in the context of structure alignment (Ye and Godzik 2005). Instead of aligning linear sequences, graph-based methods allow taking alternative paths to indicate homology, which can be helpful for divergent sequences, e.g. for constructing MSAs (Löytynoja et al. 2012; Szalkowski and Anisimova 2013; Hickey et al. 2020). A sequence alignment is a hypothesis regarding the homology of individual characters in a given set of homologous sequences. Even if correct (that is, aligned characters are direct descendants of one another, i.e. "homologous"), it leaves much uncertainty about the actual historical mutation events, since many indel and substitution events can give rise to the same homology relationships. For substitution models that make the assumption that sites evolve independently of one another, it is possible to analytically integrate out this uncertainty, but for indels (as for substitution models that allow interactions between sites), things are more complicated, as we discuss later. However, before diving into statistical models for indels, we first discuss several widely used parsimony-based approaches, with a focus on phylogenetic inference.

Indel-Based Phylogeny via Indel Coding

Indels harbor information for phylogenetic reconstruction (Dessimoz and Gil 2010; Saurabh et al. 2012). Consider, say, 40 diverged sequences, in which only two sequences harbor an identical inserted sequence between the exact same positions. If these two sequences are not placed together as a clade in the phylogenetic tree, then one has to assume that these putative insertion events reflect a case of convergent evolution, in which exactly the same insertion occurred independently along two lineages in a tree, which is clearly unlikely. This logic has motivated researchers to search for phylogenetically informative indels to strengthen the phylogenetic signal inferred from substitution events. For example, Vogler and DeSalle (1994) examined the internal transcribed spacer region 1 in 50 clones of tiger beetles. They tested various ways to account for gap characters in the phylogenetic tree: treating each gap position of an indel as a different character, coding each gap length as a different character, and treating gap characters ("–") as unknown characters ("N"). Their parsimony-based analysis showed that the tree was less resolved when gaps were ignored (treated as unknown characters) compared to the two other indel-coding methods. This led them to conclude that accounting for gap characters within phylogenetic inference increases the amount of phylogenetic information.

Such studies motivated the development of automated indel-coding methods. These methods take MSAs as input, and output a data matrix, in which each row is a sequence and each column corresponds to a specific gap. In its simplest form, 0 and 1 represent the absence and presence, respectively, of a specific DNA or protein segment. This indel information matrix is then added to the substitution matrix as input to phylogenetic tree reconstruction algorithms. A variety of different indel-coding methods have been developed, which differ in the underlying assumptions used when coding indels (Simmons et al. 2007) compared several such methods in a simulation study.

While it is obvious that indel characters are highly informative for phylogenetic reconstruction, the above indel-coding-based approaches are fraught with potential biases, statistical caveats, and methodological limitations. For example, indel-coding approaches generally code indels in a single fixed alignment estimate instead of accounting for alignment uncertainty. The placement of indels in this single alignment was shown to strongly depend on the algorithm used to align the sequences, and indeed methods were developed to quantify the reliability of each "indel character" (Ashkenazy et al. 2014).

Indel-coding methods are mainly used in parsimony-based tree searches, and thus such analyses inherit the statistical limitations of parsimony (Felsenstein 2004). In probabilistic-based approaches, handling indels is challenging, as indels violate the assumption of independence among positions, and thus, indel positions are often excluded from the data matrix before the tree inference procedure or, alternatively, are treated as unknown characters. Ideally, the recent improvements in probabilistic modeling of indel dynamics as well as in integrating tree search and alignment algorithms should make the inclusion of indel information within phylogenetic tree search a standard.

Shared indels have a stronger effect on phylogeny than shared substitutions, so it is essential for gaps to be placed without error. Most alignment software places gaps too inaccurately to use them for phylogenetic information (Ashkenazy et al. 2014). One reason for this is that most alignment algorithms employed today use progressive alignment, which can be biased toward the guide tree (Lake 1991). When gaps are treated as phylogenetically informative, this bias is even stronger, since the profile alignment step of progressive alignment introduces joint gaps into entire subtrees of the guide tree. Another reason is that most alignment algorithms are not phylogeny aware in the sense that they do not use the tree to determine the number of substitutions and indels implied by a specific alignment. As a result, most aligners tend to infer alignments where characters are often independently deleted many times when considered on a tree; see for example fig. 1 in Löytynoja and Goldman (2008). By taking into account the underlying phylogeny, phylogeny-aware programs such as PRANK can infer more realistic alignments. However, these alignments depend even more strongly on the phylogeny used as input. Thus, when inferring the phylogeny from the alignment, it is essential to score each phylogeny on an alignment inferred assuming that phylogeny. It is therefore important to consider possible errors and biases that creep in during the construction of these alignments. We discuss such alignment errors in the next section.

Alignment Biases, and Errors, and Their Consequences

Many bioinformatic inference pipelines take an MSA as input, e.g. for the inference of phylogenies, divergence times, ancestral states, and positive selection. It is important to remember that MSAs are estimated, not observed, and can therefore contain errors. MSA errors can propagate through the inference pipeline, undermining the final result.

As mentioned above, an MSA is a hypothesis about the homology of individual characters in a collection of homologous sequences. The homology information in an MSA is usually expressed by writing the sequences as a matrix so that each cell contains one character, each row contains one sequence, and each column contains characters that derive from one character in the common ancestor. When no character from a sequence is homologous to other characters in a column, we write a gap character ("–"). When a single column contains multiple letters, we can infer the presence of a substitution event in the history of the characters in that column. Likewise, when a column contains a gap, we can infer the presence of an indel event affecting the history of characters in that column.

Effects on Downstream Inferences

A wide range of bioinformatic inferences can be affected by errors in MSA inputs. For example, alignment error can negatively affect ancestral sequence reconstruction (Vialle et al. 2018; Aadland and Kolaczkowski 2020; Spence et al. 2021), the inference of positive selection (Wong et al. 2008; Fletcher and Yang 2010; Jordan and Goldman 2012; Privman et al. 2012; Redelings 2014), topology inference (Lake 1991; Morrison and Ellis 1997; Mugridge et al. 2000; Wong et al. 2008), and statistical tests to compare trees (Levy Karin et al. 2014). Alignment errors have a large effect on analyses that explicitly take indels into account. This includes the inference of indel rates (Westesson et al. 2012; Holmes 2017a) as well as phylogeny inference when indel information is used to place taxa (Ashkenazy et al. 2014).

Alignment Error

Alignment errors come from a variety of sources. First, many alignment inference algorithms seek to optimize a score function. Ideally, alignments with higher scores would represent a more plausible evolutionary scenario. However, most alignment inference software uses heuristics that do not reflect the evolutionary process very well, leading to a flawed optimality criterion.

We currently lack sophisticated stochastic models for indel dynamics. Ideally, we would like to evaluate an alignment using models that account for both substitutions and indels, allow for spatially varying indel rates, change in the insertion/deletion rate over time, context dependent indel rates, etc. We currently lack such models. Furthermore, the more accurate models that we do have tend to be too computationally expensive for practical use. Even getting a biologically realistic penalty for indel length can dramatically slow down computations (Cartwright 2007).

Additionally, the score for an MSA depends on the tree. Thus, to infer an MSA, we need to account for the phylogeny underlying the evolutionary relationship among the sequences at hand. The mutual dependence of the alignment and the phylogeny creates a "chicken-and-egg" problem. In order to account for the lack of knowledge of the tree, most alignment inference software packages use some ad hoc scores instead, such as the sum-of-pairs score, which is the sum of pairwise alignment scores over all pairs in an MSA.

Second, most MSA inference software programs fail to fully optimize the score function. The number of alignments grows much faster than exponentially in the sequence length and in the number of taxa, so that it is impossible to examine the entire set of MSAs and report the one with the highest score. Instead, heuristics are used to search a small fraction of the set of alternative alignments. For example, the progressive heuristic reduces the MSA problem to a sequence of pairwise alignment problems moving along the "guide tree" from the tips to the root. Despite various attempts to introduce "refinement" of MSAs following progressive alignment, the inability to thoroughly search the MSA space is a serious problem.

Alignment Uncertainty

MSA estimates can be incorrect when we attempt to choose a single MSA in the presence of alignment ambiguity. Alignment ambiguity occurs when multiple alternative MSAs represent plausible evolutionary scenarios (Redelings and Suchard 2009). Even with a perfect score function and the ability to search all possible alignments, it is impossible to eliminate alignment ambiguity.

There are at least two sources of alignment ambiguity. First, the optimality criterion may be sensitive to a large collection of parameters, such as the guide tree, gap-opening penalties, gap-extension penalties, mismatch penalties, and so on. For example, Lake (1991) showed that inferring a phylogeny from an MSA tends to recover whichever tree was used as the guide tree. Uncertainty about these parameter values then leads to uncertainty about which alignment is correct. Interestingly, most of these tunable parameters serve to describe the evolutionary process under which the sequences have evolved. Thus, the guide tree is a proxy for the unknown phylogeny, gap-opening penalties are a proxy for the unknown indel rate, and gap-extension penalties are a proxy for the unknown indel length distribution. Since these are often the parameters that we seek to estimate from the alignment, it is difficult to know what values to use when inferring the alignment.

Most bioinformatic workflows output a single MSA that they submit as input to downstream inference steps, thus the myriad of alignments that are near-optimal or equally optimal. Near-optimal alignments are also plausible evolutionary scenarios but place gaps differently. Ignoring near-optimal alignments makes downstream inferences sensitive to unreliably placed gaps even when the scoring parameters are reliable.

Mitigating Alignment Uncertainty

Because of the importance of MSA error on downstream analyses, biologists initially attempted to identify unreliable alignment regions "by eye." However, visual inspection and manual editing were subjective and nonrepeatable, leading different researchers to exclude different regions (Gatesy et al. 1993). Researchers therefore attempted to design objective and repeatable algorithms for identifying error-prone regions (Castresana 2000). Evaluation of these algorithms initially focused on chronicling the sensitivity of downstream analyses to ambiguous alignment regions. However, after the development of metrics that evaluated the algorithms based on curated alignment databases and alignment simulation tools, researchers attempted to (i) more accurately identify error-prone alignment regions and (ii) mitigate the effects of alignment error by removing error-prone regions. These tasks have both proven more difficult than expected.

Programs such as GBLOCKS attempt to identify and remove ambiguously aligned regions by identifying columns with multiple residues or regions with gaps, to some degree of success (Talavera and Castresana 2007). More sophisticated methods identify badly aligned regions either based on sensitivity to parameters (Talavera and Castresana 2007; Penn et al. 2010), based on the effect of incorporating alternative near-optimal pairwise alignments during MSA inference (Landan and Graur 2007), or based on a combination of both parameter sensitivity and near-optimal alignment uncertainty (Sela et al. 2015). When identifying ambiguously aligned regions, it is important to take into account all sources of alignment error and alignment ambiguity.

Researchers have also attempted to improve downstream estimates of phylogenies and positive selection by automatically identifying and "filtering" alignment columns or residues suspected of being incorrectly aligned (Jordan and Goldman 2012; Privman et al. 2012). Unfortunately, filtering does not seem to substantially improve accuracy of phylogeny estimation or power to detect positive selection when high-quality alignment estimation programs are used (Spielman et al. 2014; Tan et al. 2015). Indeed, filtering may even decrease inference quality by removing regions with many phylogenetically informative characters. Thus, it does not appear that filtering approaches are succeeding at screening out error while still preserving high power to detect biological phenomena.

The failure of filtering/masking has led to an interest in alternative remedies. One idea is that instead of removing uncertain regions, the uncertainty should be accounted for by considering a sample of alternative alignments (Wheeler et al. 1995). This approach has been embraced by MUSCLE version 5, which enables users to create alignment "ensembles" that contain several alternative MSAs for the same sequences (Edgar 2022). This is close in spirit to a Bayesian approach, but here, ad hoc methodologies are used to generate the alternative MSAs. It has the benefit that information in ambiguous regions is not completely discarded. However, the features of the alignment that vary between alternative MSAs are still downweighed. Averaging alignments using GUIDANCE2 was shown to yield more accurate trees compared to using a single best MSA (Ashkenazy et al. 2019).

Statistical approaches to sequence alignment address many of these problems. Statistical approaches are based on the probability of the alignment and of the data. Here, the log of the probability corresponds to traditional score functions that add up penalties for observed sequence changes. Statistical approaches allow estimating the relative rates of substitutions and indels from the data instead of specifying arbitrary penalties a priori. Statistical approaches also improve the accuracy of the score function by taking the phylogeny into account when scoring indels and substitutions. They are able to take advantage of information in shared indels to group taxa on the phylogeny without explicit indel coding and to infer parameters such as the phylogeny and indel rates from the data.

Bayesian statistical alignment involves jointly estimating the alignment and phylogeny via Markov chain Monte Carlo (MCMC). Joint inference solves the "chicken-and-egg" problem by coestimating the alignment and tree. MCMC allows a more thorough exploration of alternative alignments than progressive alignment with refinement. Bayesian approaches naturally account for near-optimal alignments by integrating over alternative alignments (weighted by their posterior probability) when inferring the tree and evolutionary parameters. They also naturally incorporate uncertainty in indel and substitution model parameters, uncertainty in the phylogeny, and the stochasticity of the evolutionary process.

Bayesian statistical alignment is able to accomplish what filtering-based methods cannot, including a low false positive rate and a high true positive rates in estimating positive selection (Redelings 2014). Ancestral sequences based on MSAs from BAli-Phy are nearly indistinguishable from inferences based on the true simulated alignment (Aadland and Kolaczkowski 2020). However, these benefits come at the cost of increasing computation time.

Alignment and Underparameterized Substitution Models

One issue that has not been explored is the effect of underparameterized substitution models on alignment inference. Huelsenbeck and Rannala (2004) and Lemmon and Moriarty (2004) have examined the effect of underparameterized substitution models on phylogeny inference. These papers found that ignoring biological phenomena such as across-site rate variation (ASRV) can lead both to increased bias and to increased confidence in the biased answer. In contrast, including parameters for phenomena that are not present has a much smaller effect.

However, the effect of underparameterized substitution models on MSA inference has yet to be explored. Until recently, MSA inference programs did not use statistical models of substitution to score alignment columns. The advent of programs using such models means that it is now possible to perform sequence alignment under ASRV models. Such models allow more substitutions to occur in certain columns that appear to be less conserved.

ASRV substitution models have the limitation that each column has a *single* rate. Thus, the sequences would still fail to fit this model in cases where an amino acid residue differs in its degree of conservation across the tree. Susko et al. (2002) attempted to remove model violations by inferring site rates for different segments of the data set and removing columns where the inferred rate differs between subsets. We suggest that, when the alignment is inferred assuming rate variation, the alignment may split such columns automatically to create subcolumns that have a single rate. While this may produce a less accurate MSA, it could produce a more accurate tree. Such problems with the alignment could then be addressed by adding extra parameters to describe the rate of rate switching across the tree.

Evolutionary Models of Indels

The TKF91 Model

As discussed above, many phylogenetic reconstruction approaches optimize a score function, a heuristic measure of the plausibility of the alignment and phylogeny as an evolutionary explanation of the observed sequence data. Ideally, however, phylogenetic inferences should rely on explicit evolutionary models of indels on par with Markov substitution models. However, compared to substitutions, indels are much more challenging to model, as multiple-character indels violate the site independence assumption, which is used to factorize substitution likelihoods. Over time, indels may merge and overlap, making it impossible to reconstruct indel history fully without considering the entire sequence itself as an evolving state in a Markov process.

With statistical models of point substitution in DNA, RNA, amino acid, and codon sequences (Jukes and Cantor 1969; Kimura 1980; Felsenstein 1981; Goldman and Yang 1994) successfully deployed in phylogenetics, it was natural to attempt to generalize these models beyond point substitutions. A key topic of consideration here is the conditional distribution $P(S(t)|S(0), \Theta)$ where $\{S(t)\}_{t\geq 0}$ is a sequence-valued random process and Θ is the model parameters. The special case where the sequence $S(t)$ contains exactly one character represents the family of point substitution models, which can be analyzed by the standard continuous-time Markov chain techniques including finite matrix exponentiation (Moler and Van Loan 2006; Pupko and Mayrose 2020). When the sequence is allowed to grow and shrink by indel events, then the analysis becomes more complicated, since the state space is no longer finite. In addition, the likelihood function cannot be expressed as a product of terms for independently evolving sites (except in special cases), making the likelihood computation much more challenging compared to point substitution models.

Ideally, we would like a statistical treatment that yields posterior distributions for model parameters $P(\Theta|S(0), S(t))$, a prior distribution for $S(0)$, a posterior distribution overalignments between $S(0)$ and $S(t)$, and a way of extending all these calculations to multiple sequences related by a phylogenetic tree. Bishop and Thompson (1986) used dynamic programming (DP) to compute the likelihood and posterior probability distribution overalignments for two sequences separated by a fixed evolutionary time interval. Their approach can be considered to be related to inference under a type of hidden Markov model (HMM) that emits two paired sequences, known as a pair HMM (Durbin 1998). However, neither the general pair HMM nor this specific model of Bishop and Thompson includes a time parameter, so their model is not a continuous time stochastic process.

The first treatment of this kind to include a time parameter was the TKF91 model (Thorne et al. 1991). In this model, the indel rates are assumed independent of sequence context, and indel events only involve single residues. Under such restrictive assumptions, the fate of each ancestral residue in $S(0)$ can be handled as an independently evolving zone (or "link"). The number of offspring $n(t)$ of each link at time t is a linear birth-and-death process whose finite-time transition probability $P(n(t) = k \mid n(0) = 1)$ is a geometric distribution, with a parameter that is a rational function of exponentials of the indel rates (Metzler 2003).

The emergence of the geometric distribution in the TKF91 model means that the joint distribution $P(S(t)|S(0))$ can be modeled by a pair HMM. The TKF91 paper motivated attempts to develop a statistical phylogenetic basis to sequence alignment, dubbed "statistical alignment" (Hein et al. 2000). TKF91 was extended to align multiple sequences on a star-shaped phylogeny (Steel and Hein 2001) and a binary tree (Hein 2001). Although the time and memory complexity of these algorithms (including TKF91) is $O(L^N)$ for N sequences of length L , which is prohibitively expensive, subsequent works by various authors succeeded in developing computationally cheaper inference processes, by developing tools to factorize, marginalize, and otherwise manipulate the phylogenetic likelihood. Specifically, Holmes and colleagues (Holmes and Bruno 2001; Holmes 2003; Westesson et al. 2012) showed the equivalence of TKF91 to a pair HMM and introduced the first practical MCMC samplers for statistical MSA, constructing phylogenetic likelihoods via combinations of pair HMMs (or more precisely input-conditioned pair HMMs, also called "transducers"); Redelings and Suchard (Redelings and Suchard 2005, 2007; Suchard and Redelings 2006) showed how to use these pair HMM combinatorics to perform MCMC sampling over tree topologies in time $O(L^2N)$, and Lunter et al. (2003) showed how to calculate the likelihood of a given MSA under TKF91 in time $O(LN)$.

Statistical alignment methods can be applied not only to MSA inference (Fleissner et al. 2005; Lunter et al. 2005; Novák et al. 2008) but also to perform robust statistical inferences of selection (Lunter et al. 2006; de Groot et al. 2008; Satija et al. 2008) and mutation rates (Metzler et al. 2001; Redelings 2014; Seo et al. 2022) in the presence of MSA uncertainty. While these methods—particularly those involving MCMC—are computationally expensive, they offer the most principled and apparently the most accurate way to deal with the chicken-and-egg problem that is the statistical entanglement of alignments, trees, and evolutionary parameters, as has been demonstrated (Nute et al. 2019; Gupta et al. 2021). Indeed, even without rigorous Bayesian inference via MCMC, various greedy heuristic MSA algorithms that are "tree-aware"—that is, algorithms whose scoring scheme is structured according to a phylogenetic tree—seem to outperform "tree-unaware" MSA algorithms when it comes to problems such as the reconstruction of ancestral histories and the estimation of insertion and deletion rates (Löytynoja and Goldman 2005).

Beyond the TKF91 Model: the GGI Model

While studies of statistical alignment drew much inspiration from Thorne et al.'s (1991) work, very few of them used the TKF91 model directly without modification. This is because TKF91 tends to produce poor alignments (Holmes and Bruno 2001), since it allows only independent single-character indels, but not instantaneous multiple-character indels. TKF91's log-likelihood, when used as an objective function for MSA optimization, implies a linear gap penalty, whereas better alignments can be inferred using an affine gap penalty (Vingron and Waterman 1994), since this avoids overpenalization of long gaps through distinguishing gap opening and extension penalties.

While an affine gap penalty is consistent with a geometric distribution for lengths of indel events, this too is an approximation: empirical indel length distributions (Benner et al. 1993; Qian and Goldstein 2001) tend to have fatter tails than simple geometric distributions and may be better described by the power law. Recent work has shown that some data sets are better described by a power law, while others are better described by a geometric distribution; i.e. no single length distribution fits all empirical data sets (Wygoda et al. 2024). A more sophisticated model should also take into account variation in sequence divergence, along with asymmetries between the insertion and deletion rates, sizes, dependence on sequence context, and potentially structural information. Certainly, empirical data exist to inform such models; however, tractability of statistical analysis becomes an issue.

To make the TKF91 model more realistic, as a first step toward more sophisticated empirical models, a natural starting place is to allow the lengths of instantaneous indel events to be geometrically distributed (instead of just being single characters). The geometric distribution is moderately tractable and well motivated: it is the maximum entropy distribution over integers with a given mean. In fact, since the assumption that indel sites are uniformly distributed in the sequence is also a maximum entropy assumption, a generalization of TKF91 that allows geometrically distributed indel lengths may be regarded as the maximum entropy indel model with given indel rate parameters (λ, μ) and mean indel lengths ($\bar{X}$, $\bar{Y}$). The resulting general geometric indel (GGI) model (Miklós et al. 2004; Holmes 2020; De Maio 2021) with parameters $\theta = \{\lambda, \mu, \bar{X}, \bar{Y}\}$ is time reversible if, and only if, $\lambda(\bar{X} - 1) = \mu(\bar{Y} - 1)$. The TKF91 model is a special (reversible) case of this model when $\bar{X} = \bar{Y} = 1$, in which case the probability of extending a gap is 0.

An exact solution to the GGI model has proven elusive, despite ongoing study (Rivas 2005). The underlying mathematical problem is that when deletions are allowed to remove more than one consecutive residue in a single event, the fates of adjacent residues can no longer be considered independent. Furthermore, even though the length distribution of instantaneous indels is geometric, the finite-time probability distribution over gap lengths no longer has a simple geometric form, so that no finite-state pair HMM can fit the finite-time gap length distributions exactly (Rivas 2005; Rivas and Eddy 2008, 2015). This is because an alignment gap potentially represents the accumulation of multiple overlapping indels, so the size of such a gap is a convolution over some number of instantaneous indel lengths (and this convolution no longer yields a geometric distribution).

Some approximations have been introduced to deal with this intractability. As an immediate follow-up of TKF91, the TKF92 model approximated multiple-character indels by imagining that the sequence consists of a number of indivisible multiple-character fragments (Thorne et al. 1992). Others attempted to guess forms for a pair HMM inspired by TKF91 or TKF92 (Knudsen and Miyamoto 2003; Löytynoja and Goldman 2005, 2008; Redelings and Suchard 2005, 2007; Suchard and Redelings 2006; Holmes 2017b). Most recently, De Maio (2021) used a moment-matching approach to derive differential equations for a best-fit pair HMM. As noted, a finite-state pair HMM with geometric waiting times cannot fit the gap length distributions exactly, but the earlier work of Miklós et al (2004) had shown that the alignment likelihood is still factorizable using a "generalized" pair HMM (i.e. one with nongeometrically distributed waiting times). Holmes (2020) further used De Maio's technique to develop refined ordinary differential equations and a pair HMM which, as current evidence suggests (Holmes 2020), is the best approximation to the GGI model so far.

It is clear that GGI is still a considerable simplification of true biological sequence evolution. To begin with, analyses of homologous sequences suggest that in some empirical data sets, indel lengths follow a power law distribution (Benner et al. 1993; Gu and Li 1995; Chang and Benner 2004). Furthermore, as discussed above, many inserted DNA sequences are local duplications (e.g. Messer and Arndt 2007; Vaughn and Bennetzen 2014), and microsatellite expansions/contractions account for a lot of genomic indels. This suggests that—in a realistic model—flanking context should influence indel rates and inserted sequence content. Finally, when modeling evolution at the protein level, it is crucial to consider the effect of flanking sequence context and indeed of all intramolecular and intermolecular contacts in suppressing indel mutations by selection.

Several algorithms have been developed that integrate probabilistic models of local duplication and sequence alignment, including some based on stochastic grammars (Hickey and Blanchette 2011) or other probabilistic models (Nánási et al. 2014), and others based on sequence-to-sequence neural networks (Lim and Blanchette 2020). An obvious but ambitious goal is to integrate such approaches with advances in understanding epistatic interactions in proteins, such as Potts models (Levy et al. 2017). High-quality data sets of aligned and phylogenetically resolved protein homologs, such as TreeFam (Schreiber et al. 2014) or OPTIC (Heger and Ponting 2008), could be useful to inform such efforts. An alternative to modeling the full GGI process was introduced in the Poisson Indel Process, dubbed the PIP model (Bouchard-Côté and Jordan 2013), which sacrifices locality (the indel rate per site varies inversely with sequence length) in favor of computational tractability.

At the time of writing, tools available for statistical alignment (listed in Table 1) have not yet caught up to the advances in approximating the GGI model and instead use lower-quality approximations to the GGI such as RS07 (Redelings and Suchard 2007).

Generalizing Felsenstein's Pruning Algorithm from Single Sites to Entire Sequences

The Felsenstein (1981) pruning algorithm for computing the likelihood of biological sequences operates on a single site. The ancestral state at each node of the tree may take on a small, finite number of states such as A, C, G, or T for DNA. The pruning algorithm then constructs a conditional likelihood vector $F_i^n = Pr(\text{Data}|X = i)$ for the state i at an internal node n in terms of a likelihood profile for the left and right child.

But what happens if we consider the internal node state to represent a complete DNA sequence of unknown length? First, the number of states becomes infinite, because the length of sequences is unbounded. Second, state changes on the branch to the left or right child now include not just substitutions such as A → G but also length changes such as GV → GIV. Third, we must consider the alignment of the sequence at a node to the sequences at its left and right child.

Westesson et al. (2012) show that for many statistical models, we can generalize Felsenstein's algorithm from single sites to entire sequences. The fundamental insight is that we can generalize conditional likelihood *vectors* to conditional likelihood *HMMs*. Such HMMs differ from traditional HMMs in that they do not *emit* an observed character with a certain probability but *absorb* an input character with a certain likelihood. In this paradigm, the transition probability matrix $exp(Rt_b)_{ij}$ along branch b in the pruning algorithm is generalized to a pair HMM. The pair HMM specifies the probability that a sequence i will evolve to a different sequence j along branch b by emitting two aligned sequences instead of just a single sequence (Thorne et al 1991; Holmes and Bruno 2001).

The conditional likelihood HMM F^n for a leaf node n has a relatively simple structure: for an observed sequence s of length L, it will have L HMM states in linear order, where

Table 1 Statistical phylogenetic software for analyzing indel evolution

Inference goals	Software	Indel model/representation	Approach	Special features
MSA	PRANK (+F)	HMM	Progressive MSA Tree-aware	Defaults to ancestral residues being absent when the presence/absence is ambiguous (with +F).
MSA Gap penalties	ProGRAPH (+TR)	HMM, POG	Progressive MSA Tree-aware	Similar to PRANK; Handles tandem repeats, alternative splicing
MSA, EP	ProPIP	PIP	ML Progressive MSA Tree-aware	Estimates indel rates, allows Gamma rate heterogeneity
MSA, Tree, ASR	StatAlign	TKF92	Bayesian MCMC	GUI; allows to incorporate structural information
MSA, Tree, ASR, EP	BAli-Phy	RS07	Bayesian MCMC	Multiple partitions; can specify priors on all parameters
MSA, Tree, ASR, EP	Historian	RS07	Progressive MSA followed by MCMC	Fast initial ASR, mixture models for substitutions
ASR	ARPIP	PIP	ML	Provides uncertainty profiles for inferred sequences
EP, simulations	SPARTA-ABC	Continuous time Markov process of indel evolution	ABC	Allows simulations with indel dynamics inferred from empirical data sets
ASR	FastML	Indel coding, 2-state models	ML	Web server, allows reinsertion of deleted characters
ASR	GRASP	POG	ML	Allows reinsertion of deleted characters
Pairwise alignment	SimBa-SAl	GGI	ML	...

Disclaimer: inferring ancestral states is not ML because ancestral states are random variables, not parameters.
Disclaimer: inferring a tree by integrating over internal node states is not ML unless you integrate out the alignment of leaf sequences.
ABC, approximate Bayesian computation; EP, evolutionary parameters; GGI, general geometric indel; ML, maximum likelihood; MP, maximum parsimony; POG, partial order graph.

the ith state has a nonzero likelihood only for the ith observed letter (Fig. 2a). We denote such an HMM as U_s to indicate that it assigns a likelihood of 1 to the sequence s and 0 to all other sequences. The conditional likelihood HMM for an internal node combines the HMMs for the left and right child nodes by taking the Cartesian product of their state spaces. If the left and right children are tip nodes with states X_i and Y_j, respectively, then the product HMM will contain states of the form (X_i, Y_j, H), where H represents an evolutionary history. Thus, we end up with a 2D DP matrix (Fig. 2b). The final likelihood profile F^n at the root node of a tree with N leaves will have $O(L^N)$ states. Felsenstein's (1981) algorithm thus generalizes to the construction of an HMM with an N-dimensional hypercube structure, which turns out to be very closely related to Sankoff's DP matrix (Westesson et al 2012). The connection between these algorithms is illustrated visually for the two- and three-sequence cases in Fig. 2.

Much of the subsequent MSA literature can be viewed as a cumulative effort to find good approximate solutions in sub-$O(L^N)$ time to the general NP-complete problem of finding the highest-scoring path through an N-dimensional hypercube of the same general structure as in Sankoff (1975). Notably, some of the employed heuristics include greedy optimization along a postorder tree traversal, often called "progressive alignment" (Thompson et al. 1994), further rounds of iterative refinement (Edgar 2004), approximate posterior decoding (Do et al. 2005), Gibbs sampling and other MCMC kernels (Holmes and Bruno 2001; Redelings and Suchard 2005), and transformations of the data including Fourier transforms (Katoh et al. 2002; Maiolo et al. 2020).

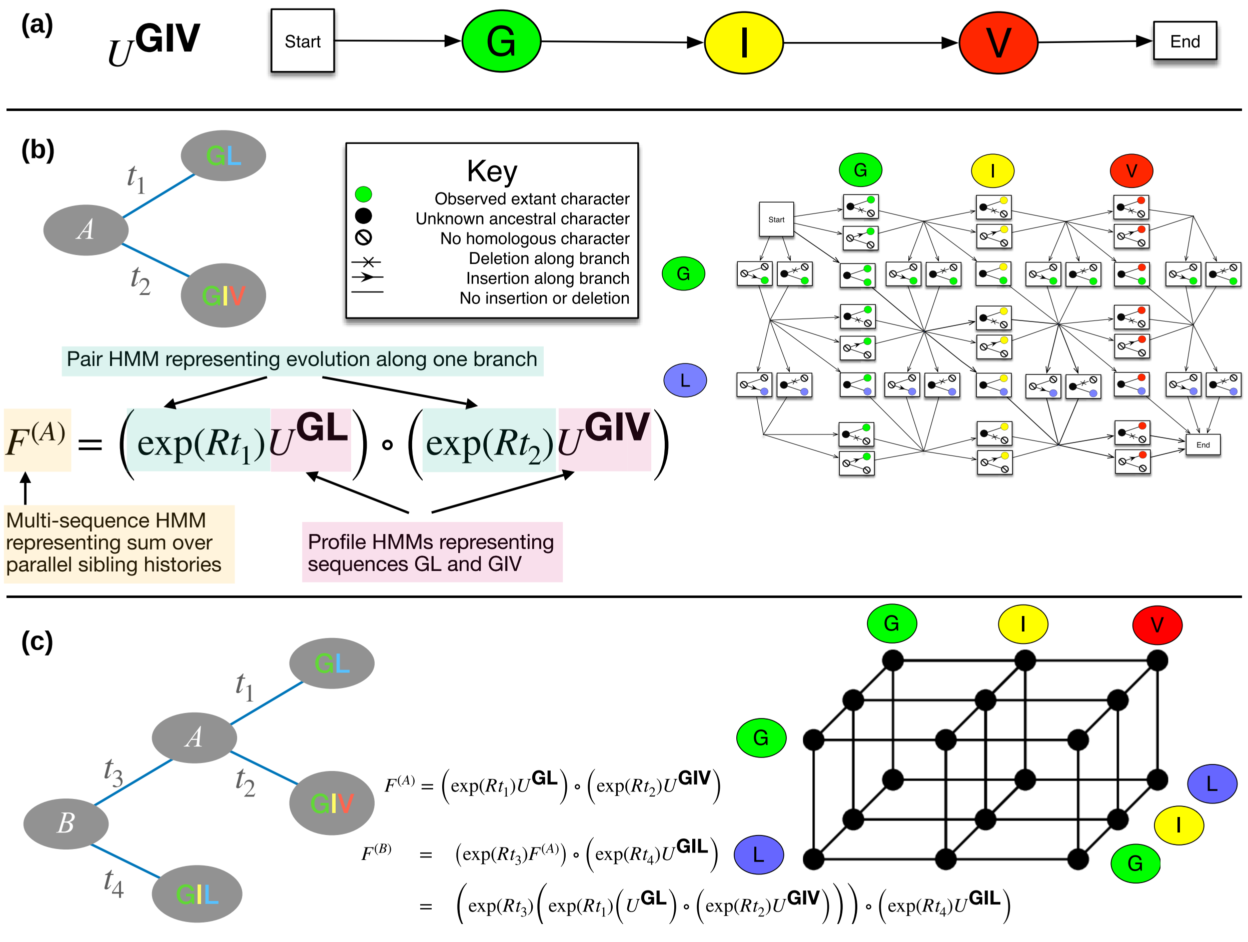
(a)
U^GIV
Start
G
I
V
End
(b)
GL
t1
A
t2
GIV
Key
● Observed extant character
● Unknown ancestral character
⊘ No homologous character
Deletion along branch
Insertion along branch
No insertion or deletion
Pair HMM representing evolution along one branch
F^(A) = (exp(Rt1)U^GL) ∘ (exp(Rt2)U^GIV)
Multi-sequence HMM representing sum over parallel sibling histories
Profile HMMs representing sequences GL and GIV
G
I
V
G
L
Start
End
(c)
GL
t1
A
t3
t2
GIV
B
t4
GIL
F^(A) = (exp(Rt1)U^GL) ∘ (exp(Rt2)U^GIV)
F^(B) = (exp(Rt3)F^(A)) ∘ (exp(Rt4)U^GIL)
= (exp(Rt3)(exp(Rt1)(U^GL) ∘ (exp(Rt2)U^GIV))) ∘ (exp(Rt4)U^GIL)
G
I
V
G
L
L
I
G

Fig. 2. Likelihood calculations on sequence characters. a) The likelihood profile HMM for a single observed sequence. b) The likelihood profile HMM for the ancestor A of two homologous sequences GL and GIV has the structure of a 2D DP matrix. A path through the HMM specifies the alignment of sequence A to the child sequences, as well as the likelihood of letters at each position of sequence A. Different states at the same matrix coordinates specify whether a character that is deleted in one child is present or absent in the ancestor. c) Aligning three sequences GL, GIV, and GIL yields two ancestors a) and b) This requires a 3D DP matrix. Here, multiple states at the same coordinates have been collapsed into single nodes to reduce clutter. The algebraic equations represent Felsenstein's algorithm interpreted through the lens of likelihood profile HMMs. F^n traditionally represents a conditional probability vector for a single letter; here, it is an input–output machine (transducer) that accepts the sequence at node n as an input. U traditionally represents a unit vector corresponding to the single observed letter; here, it is a deterministic state machine that accepts only the observed sequence at that leaf node. The exponentials in Felsenstein's recursion are finite matrix exponentials; here, they are HMMs (or more precisely, stochastic finite-state transducers) whose transition weights may be derived exactly (as in Thorne et al. 1991) or approximated systematically (as in Holmes 2020).

Simulating Sequence Using Indel Models

While inferring the alignment under an evolutionary model is very hard, simulating related sequences even under very complex models is much more tractable (Table 2). Simulations are used for various reasons, including, for example evaluating the accuracy of alignment algorithms (e.g. Maiolo et al. 2018; Gupta et al. 2021), comparing tree reconstruction algorithms (e.g. Takahashi and Nei 2000; Huelsenbeck and Rannala 2004; Azouri et al. 2021), evaluating the performance of algorithms aimed at ancestral sequence reconstruction (e.g. Moshe and Pupko 2019; Foley et al. 2022; Jowkar et al. 2023) and inference of selective forces (e.g. Anisimova et al. 2001; Spielman et al. 2016), testing evolutionary hypotheses (e.g. Goldman 1993), testing the effect of model misspecification (Lemmon and Moriarty 2004; Magee et al. 2021), and for the inference of model parameters (e.g. Arenas 2015; Levy Karin et al. 2015), to name a few. The very first sequence simulators totally ignored indels (Rambaut and Grassly 1997; Yang 1997). Other early sequence simulators such as MySPP (Rambaut and Grassly 1997; Rosenberg 2007) did not explicitly describe the details of the assumed indel dynamics process. Yet, other programs, such as Rose (Stoye et al. 1998) and EvolveAGene3 (Hall 2008), did not implement a continuous time Markov process for indel dynamics but rather introduced ad hoc and often unrealistic assumptions regarding the indel process. For example, in ROSE, insertion dynamics is controlled by a parameter that dictates the probability that a sequence will experience an insertion event. This probability is assumed to be fixed for the entire course of evolution. However, it is clear that the probability of insertion should depend on the sequence length, which varies along the tree. This is ignored in ROSE. Similarly, in each branch, a single insertion (and a single deletion) event is allowed, but in fact, a continuous Markov process should allow multiple events to occur along each branch. Assumptions regarding the indel length distribution also failed to reflect reality. For example, in EvolveAGene3, it is assumed that deletions of length between 2 and 23 bases have equal probabilities, which contradicts empirical observations that shorter indels are substantially more common than longer ones (Pascarella and Argos 1992; Benner et al. 1993; Loewenthal et al. 2021).

More recent simulators assume a continuous time Markov process of indel dynamics (Fletcher and Yang 2009; Dalquen et al. 2012; Ly-Trong et al. 2022). In general, such simulators allow generating data with parameter-rich indel models, in which each branch or clade of the tree and each data partition are associated with their own set of indel model parameters. Although not yet implemented in available sequence simulators, simulations can also introduce context-dependent indel models, in which the probability of indels depends on the specific flanking characters. Moreover, structural-aware simulators are plausible, in which the probability of indels in protein sequences depends on whether, for example, they reside in exposed versus buried regions of the proteins. Thus, there is a gap between our ability to simulate sequence data along a phylogenetic tree using various indel models and to account for such rich models within bioinformatic tasks such as tree reconstruction, alignment, and the inference of selective forces.

Simulated sequences depend critically on the parameters that are used to simulate them. Data sets simulated under different parameter values can have substantially different properties. In some cases, it is important to simulate sequences that have the same parameters as a particular biological data set. This requires first inferring the parameters from the biological data set so that they can later be used during simulation. As always, when conditioning on a particular alignment, errors in the alignment estimate

Table 2 Sequence simulators with evolutionary indel models

Simulator tool	Indel model	Special features	Citation; availability
Indelible	multiple-character insertions and deletions with separate rates	…	(Fletcher and Yang 2009) http://abacus.gene.ucl.ac.uk/software/indelible/
AliSim	Similar to Indelible	Lower RAM requirement, different indel rates and length distributions	(Ly-Trong et al. 2022) http://www.iqtree.org/doc/AliSim
ALF	Similar to Indelible	Whole-genome simulator with heterogeneous rates and composition biases	(Dalquen et al. 2012) http://alfsim.org/#index
PhastSIM	Similar to Indelible	Fast for very large data sets (>10,000 tips) of low divergences, typical in epidemiology	https://github.com/NicolaDM/phastSim
JavaPIP	PIP	Only single-residue insertions	(Bouchard-Côté and Jordan 2013); Available from the authors on request
EvoLSTM	Recurrent neural networks	Context-dependent 1- to 2-bp indels	(Lim and Blanchette 2020) https://github.com/DongjoonLim/EvoLSTM
SpartaABC	Continuous time Markov model	Allows inference of model parameters from empirical data sets and simulations using these parameters	(Ashkenazy et al. 2017) https://github.com/gilloe/SpartaABC
SLiM	User-defined through a script	Support for complex demographic models and selection	(Haller and Messer 2023) https://messerlab.org/slim/
Part of SimBa_SAl_sim	No explicit model	Uses simulations to shorten running times for pairwise statistical alignment	(Levy Karin et al. 2019) https://github.com/elileka

can affect the inference of parameters (Westesson et al. 2012; Holmes 2017b).

Traditionally, model parameters have been estimated via inference methods such as MCMC or maximum likelihood (ML) that explicitly compute the probability of the data. For example, ProPIP infers indel rate parameters under the PIP (Bouchard-Côté and Jordan 2013), while Historian and BAli-Phy infer rates and indel lengths under the RS07 model (Redelings and Suchard 2007). Recently inference methods that are themselves based on simulation have been introduced, such as approximate Bayesian computation (ABC) methods (Ashkenazy et al. 2017; Loewenthal et al. 2021). Such methods allow parameter inference under models where it is very difficult to compute the likelihood or prior probability. However, the downside of such methods is that they condition on summary statistics of the data instead of the full data, and these summary statistics may not capture all of the relevant information in the data.

Indels and Bioinformatic Inference Pipelines

Alignment inference methods are an essential part of the modern bioinformatic toolkit. Most bioinformatic inference pipelines follow a sequential estimation approach where an MSA is first constructed without knowledge of the phylogeny, and this single estimate is fed into downstream analyses and treated as certain. Often, the second state of the pipeline is to infer a phylogeny from the single MSA estimate, after which an estimate of the phylogeny may be fed into further inference stages. Yet, the alignment and the tree are interconnected and should be considered jointly. Moreover, most phylogenetic software only model single-letter substitutions and do not account for indels. The failure to model indels and to jointly infer the tree and the alignment causes a number of problems.

First, the failure to model indels means that information in shared indels cannot be used to group taxa on the tree. Indels show up as gap characters in the alignment matrix, but such gaps are treated as letters that are present but known. This is unfortunate, as shared indels yield stronger evidence to group taxa on the tree than shared substitutions because of their lower rate. Shared indels can be especially useful in grouping highly divergent sequences (Rokas and Holland 2000; Simmons and Ochoterena 2000; Belinky et al. 2010). Second, this approach fails to account for alignment uncertainty. While there may be myriads of near-optimal alignments, only one of these is given as input to the next stage of the pipeline. The homology information expressed by the MSA is itself estimated, and it is important to take this uncertainty into account. Third, with the exception of a few programs such as PRANK (Löytynoja and Goldman 2008), alignment inference programs do not place indel events on branches of the tree and instead simply place gap characters in a matrix. This leads to alignment matrices that imply an excessive number of indel events. As mentioned above (see Section "Indel-Based Phylogeny via Indel Coding"), the effect of low-quality alignment estimates is magnified when shared gaps are used to group taxa through indel coding.

The first joint estimation approaches used parsimony (Wheeler 2005a, 2005b). This was soon followed up via Bayesian MCMC sampling (Lunter et al. 2005; Suchard and Redelings 2006). A statistical approach to alignment requires a phylogenetic model of the unaligned sequences:

$$\Pr(\text{Seqs} \mid \text{Tree}, \Theta).$$

If we do not have a rate matrix on the (infinite) space of sequences, then such a model will include an alignment:

$$\Pr(\text{Seqs}, \text{Align} \mid \text{Tree}, \Theta) = \Pr(\text{Seqs} \mid \text{Tree}, \text{Align}, \Theta) * \Pr(\text{Align} \mid \text{Tree}, \Theta).$$

Here, the term $\Pr(\text{Seqs} \mid \text{Tree}, \text{Align}, \Theta)$ indicates the traditional likelihood function that only accounts for substitutions and usually assumes that substitutions at different sites are independent. The term $\Pr(\text{Align} \mid \text{Tree}, \Theta)$ models indel events on a tree; it is the probabilistic analog to traditional gap penalties. This leads to a more accurate score function than the ad hoc score functions usually optimized by traditional alignment software. It can therefore be used both for alignment and for tree inference.

Bayesian Approaches for Coestimating Alignment and Phylogeny

Bayesian approaches can perform inference under probabilistic models of indel evolution using MCMC. Bayesian approaches naturally account for alignment uncertainty. Uncertainty over parameters such as the tree, indel rates, and substitution rates is also automatically accounted for.

However, MCMC accounting for alignment uncertainty is slower than traditional fixed-alignment MCMC for a variety of reasons: (i) It requires MCMC proposals for new alignments as well as trees and parameters. (ii) MCMC proposals for the tree must now integrate out the alignment on the part of the tree that changes. It is too hard to integrate out the alignment on the whole tree unless all indels are assumed to be only one character. (iii) Moves that make substantial changes to the indel rates or branch lengths will have lower acceptance rates if they do not propose new alignments as well, but this is tricky and slow (Redelings and Suchard 2009).

Bayesian approaches typically augment the MCMC state space with homology information for internal node sequences, following Holmes and Bruno (2001). With this approach, we have a pairwise alignment along each branch of the tree. This allows proposing new pairwise alignments along each branch relatively easily. However, proposing new tree topologies becomes more difficult, as we must integrate out pairwise alignments on the part of the tree that changes. Since we cannot integrate out the entire alignment, proposed topologies are sometimes rejected because they do not fit well with the alignment from the source topology.

Computational Alternatives to Bayesian Approaches

Maximim Likelihood Inference

A joint maximum likelihood tree and MSA search could provide a faster method since it avoids MCMC sampling. Further, in Bayesian MCMC methods, more effort is needed to monitor the convergence of multiple parameters including structural parameters such as MSAs and trees, compared to ML which is nevertheless also affected by multiple optima. For a joint ML approach to be computationally feasible, one needs to use an indel and substitution model with a reasonable time complexity for summing out both substitution and indel events for a fixed alignment. It is possible to compute such an approximate marginal likelihood in linear time under the TKF91 model (Lunter et al. 2005) as well as the PIP (Bouchard-Côté and Jordan 2013). By using progressive alignment to construct an MSA for a given tree, this property has been successfully exploited for inferring MSAs (Maiolo et al. 2018; Maiolo et al. 2020, 2021), phylogeny (Zhai and Alexandre 2017), ancestral sequence reconstruction (Jowkar et al. 2023), and joint MSA-tree inference (Pečerska et al. 2021) using an ML approach. For example, the PIP-based tree-aware progressive aligner ProPIP uses an ML optimization function. This method appears to produce alignments with gap patterns consistent with the underlying phylogeny and does not suffer from overalignment, producing MSAs of similar lengths to the tree-aware aligner PRANK. Assuming a single-residue indel model means that ProPIP tends to infer on average shorter gaps as well as fewer substitutions. However, since the method operates in the ML framework, indel rates can be estimated from data. In contrast, most other aligners including PRANK rely on fixed default gap penalties, ignoring data set-specific indel features. This means that the user remains uninformed about indel rates or gap penalties specific to their data, reducing the possibilities of utilizing evolutionary information from indel patterns.

Simulation-Based Inference

A parallel line of research attempts to bypass the need to compute likelihood functions (Cranmer et al. 2020) so that we can examine models where the likelihood functions are unknown or slow to compute. Assume a continuous-time Markov process of substitution and indels. To fully describe the evolutionary dynamics along a given phylogenetic tree, such a model must specify the length of the sequence at the root, parameters describing substitution events, and parameters describing indel events. Indel parameters include the insertion and deletion rates at each position in the sequence, as well as separate distributions for the number of characters to be inserted or deleted. Two distributions are commonly assumed: a geometric or a Zipf (power law). Each of these distributions is governed by a single parameter. Thus, assuming for example a Zipf distribution, the indel model parameters include a parameter dictating the shape of Zipf distribution for insertions and similarly for deletions.

Levy Karin et al. (2015) first used such an approach to infer indel parameters assuming a Zipf distribution, in which it was assumed that the insertion rate equals the deletion rate and that the exact same distribution dictates the insertion and deletion sizes. Next, Levy Karin et al. (2017) implemented an ABC approach, which also relies on repeated simulations, for inferring indel model parameters. The advantages of the ABC approach are that it is a robust inference methodology and it provides an estimate of the posterior distribution of the model parameters. This ABC approach was generalized (Loewenthal et al. 2021) by allowing different indel dynamics for insertions and for deletions. It was shown that, as previously reported using ad hoc methodologies, for a large number of empirical data sets, the deletion rate is higher than the insertion rate. Finally, Wygoda et al. (2024) applied an ABC model-selection approach to determine which indel length distribution best fits empirical data sets. It was shown that for most, but not all, the Zipf distribution provides better fit than the geometric distribution. Unfortunately, currently, there are no efficient alignment programs that assume a Zipf distribution of indel lengths. A web server implementing these algorithms, called SpartaABC, allows users to upload their empirical data set and obtain estimates of the underlying indel model parameters (Ashkenazy et al. 2017).

Graph-Based Methods

Representing MSAs through partial order graphs offers another promising direction for phylogenetic methods. Edges connecting nonadjacent residues in such a POG correspond to indels or rather to "gaps" resulting from potentially multiple overlapping indel events. Pangenome representation projects are now steering toward graph representations that facilitate the handling of indels and alterations of repeat numbers (Hickey et al. 2020). The idea of representing alignments as graphs dates to as early as over three decades ago (Hein 1989), and the theme was picked up again in the 2000s (Lee et al. 2002; Grasso and Lee 2004). Combining POG representation with a phylogeny-aware algorithm appears to be particularly successful. For example, the PAGAN method relies on a POG representation of sequences to extend existing MSAs with short fragmented or noisy NGS sequences, guided by a phylogeny (Löytynoja et al. 2012). Another phylogeny-aware graph-based aligner ProgdraphMSA (Szalkowski and Anisimova 2013) naturally accommodates alternative splicing and repeat unit gain–loss events, even when unit boundaries are distorted. This allows adapting indel penalties related to different mechanisms and does not require modeling of the indel process over time.

wMost recently, the method GRASP has capitalized on the advantages of a POG-based representation to infer ancestral sequences with indels (Foley et al. 2022). Given an input MSA, GRASP reconstructs ancestral characters in a probabilistic framework, independently for each site. Indel events in GRASP are not modeled explicitly. Rather, the gaps are represented via the POGs, which allows inferring indel histories from the extant and ancestral POGs. In turn, the ancestral POGs can be inferred either by parsimony or ML using a so-called bidirectional edge encoding, similar to position-specific encoding in FastML (Ashkenazy et al. 2012), but using POGs over the branches of the tree. For each site, GRASP determines the optimal neighbors, designating gap states as absent or present in the ancestor sequence. Speed is therefore an important advantage of this approach. Nevertheless, this method cannot handle individual indel events and therefore cannot accommodate nested indels nor individual or short overlapping indels.

The Historian method (Holmes 2017a) uses a graph data structure to retain multiple alternative possible alignments for subtrees in a progressive alignment approach. Alternative paths do not represent a subset of sequences to align to, but alternative alignments of all sequences in the subtree, weighted by their probabilities. This allows Historian to delay committing to a specific alignment for sequences in a subtree before data outside the subtree is accounted for.

Parsimony

As the need for fast scalable methods increases, we also see a surprising revival of novel methods based on the parsimony principle. For example, the IndelMaP method (Iglhaut et al. 2024) for MSA and ancestral sequence reconstruction implements a scoring criterion with multiresidue indels as separate events, using the Dollo principle; i.e. once a character is deleted, it cannot reappear. IndelMap disentangles the overlapping indel events by inferring their location on the tree using the affine gap penalties. On large densely sampled data sets, this method outperforms its competitors, including FastML, ARPIP, and GRASP.

Fast parsimony algorithms for simultaneous alignment and tree inference have been further optimized in POY5 (Wheeler et al. 2015). The new method PhyG extends POY5 by extending trees to phylogenetic networks that allow representing hybridization. Thus, PhyG enables simultaneous alignment and phylogenetic network inference (Wheeler et al. 2024).

Indels in Bioinformatics, Health, and Evolution

Indels and Effect on the Phenotype

While the focus of this review is on interspecies evolutionary dynamics, indels play a substantial role in population genetics too, both as genetic markers and in how they shape phenotypes. Indels can affect any functional sequence and can have a more substantial functional impact than SNVs do, in both coding and noncoding regions. When indels occur in protein-coding sequence, they can add or remove amino acids, potentially changing the conformation and function of proteins. It was previously shown that indels occur more in disordered regions of a protein and in regions that are exposed to solvent (Benner et al. 1993; Kim and Guo 2010; Light et al. 2013). Comparative genomics of protein domains suggests that indels have a major role in diversifying the function of domains across the tree of life (Wolf et al. 2007). In recent years, methods that can predict the phenotypic outcome of both nonsynonymous SNVs and short indels in protein-coding sequence have emerged (Choi et al. 2012; Hu and Ng 2012; Li et al. 2022). In addition, indels often change the reading frame, which can result in a complete abrogation of transcription due to nonsense-mediated decay of the mRNA transcript. This high functional potency of indel variants is reflected in the relatively high contribution of somatic indel variants to the development of cancer (Yang et al. 2010). Similarly, indels are strongly purified from most human genes, except from genes that experienced substantial relaxation of selective constraints, such as most human olfactory genes (Lin et al. 2017). Within exons, trinucleotide or homopolymer repetitive regions are particularly prone to indels, often leading to disease (Gall-Duncan et al. 2022; Elena-Real et al. 2023).

Indels in noncoding sequences also contribute to disease, and methods have been proposed to evaluate the pathogenicity of noncoding indel variants in the human genome (Ferlaino et al. 2017). In addition, since indels contribute to phenotype, they are among the causal mutations targeted in genome-wide association studies (GWAS). Relatively few indels are directly typed by the probe-based genotyping chips used in GWAS, which typically include mostly SNPs. To some degree, the status of indels can be imputed from nearby SNPs (Lu et al. 2012), and it has been shown that this indeed improves the power of GWAS (Song et al. 2018; Dai et al. 2020; Kundu et al. 2022; Boatwright et al. 2023). However, indel mutation rates vary by several orders of magnitude (Montgomery et al. 2013), and particularly STRs are highly mutable and show reduced linkage disequilibrium with nearby SNP markers, making imputation difficult (Gymrek et al. 2016). Nevertheless, polymorphic STRs are associated with gene expression differences (Bilgin Sonay et al. 2015; Gymrek et al. 2016) and modulate complex disease risk (Sonay et al. 2015; Jakubosky et al. 2020; Horton et al. 2023; Verbiest et al. 2024), and their impact on monogenic disorders is increasingly recognized (Trost et al. 2020; Depienne and Mandel 2021; Marwaha et al. 2022; Elena-Real et al. 2023). STRs cannot be typed using standard probe-based genotyping techniques, and whole-genome sequencing (WGS) is required to genotype these loci. So far, the phenotypic impact of STRs has been investigated mainly using relatively expensive WGS technologies in large consortia (Halldorsson et al. 2022). We expect that novel technologies such as long-read sequencing (Hon et al. 2020; Sereika et al. 2022) will eventually enable cost-effective genotyping of STRs.

Indels as Phylogenetic Markers

Due to the improved sequencing methods and the ease of obtaining large amounts of genomic data, indels are now increasingly recognized as an independent source of information and potential molecular markers in medical and evolutionary studies. In human populations, indel variation is not only frequent but also affects many functional genomic regions (Mills et al. 2006). Selected functionally important indels may serve as biomarkers in personalized medicine (Chuzhanova et al. 2003; Mullaney et al. 2010; Sehn 2015). Indels have been linked to at least 22% of hereditary and complex somatic diseases (Stenson et al. 2009). For example, in human tumor samples, in addition to the traditionally used tumor mutational burden, the analysis of indel burden further informs cancer prognostics and stratification (Wu et al. 2019). As mentioned above, indels resulting from STR length variations may be involved in regulation of gene expression. Verbiest et al. (2024) derived a set of such eSTR loci, whose lengths have a linear relationship with expression levels in colorectal cancer tumors. These could be considered as candidate biomarkers, subject to further classification and clinical validation. The same study showed that indel patterns are predictive of the microsatellite instability status, which is used by clinicians for prognostics and therapy choices. Microsatellite unstable tumors appear to be dominated by short deletions.

The evolutionary dynamics of indels may be even more prominent in the genomes of pathogens. Viral sequences have also shown that indel patterns could be linked to the origin of new variants and changes in pathogenicity. The analysis of the SARS-CoV-2 spike protein showed an increase of indels over time, suggesting a selective advantage (Rao et al. 2021). Multiple reports detailing indels in the spike protein suggest that indel patterns can be used for the identification of coronavirus strains and their infectious properties (Andersen et al. 2020; Liu et al. 2020; Som et al. 2022). In the malaria parasite *Plasmodium falciparum*, indels are the most common polymorphism within the core genome, contributing to high genomic diversity and drug resistance (Miles et al. 2016). Indel diversity in HIV-1, much of which is presumably driven by selection, allows to identify patterns specific to different subtypes (e.g. Palmer and Poon 2019).

Deletion Bias

Already in 1973, based on the analysis of a few globin sequences, it was found that out of 13 indels, only two were insertions, suggesting that deletions are more common than insertions (Fitch 1973). This observation of an excess of deletions was subsequently confirmed by a study in which a much larger number of sequences were analyzed, which also proposed a mutational model that favors deletions as a result of a DNA repair mechanism (de Jong and Rydén 1981). In general, the indel ratio in proteins reflects the outcome of both mutation and selection processes. By demonstrating a strong deletion bias in processed pseudogenes, it was shown that deletion bias mainly reflects bias in the mutation process toward deletions (Graur et al. 1989). Research that followed generally confirmed the existence of a mutation bias toward deletions (Zhang and Gerstein 2003; Kuo and Ochman 2009) and proposed model-selection methods to test for the presence of such a bias in specific empirical data sets (Loewenthal et al. 2021).

Short indels (up to 50 bp) are only one of the factors that determine the lengths of genomic entities such as introns, exons, and intergenic noncoding regions. Other factors are, for example, microsatellite expansion, gene and domain duplications, insertions of mobile genetic elements, large deletions, and whole-genome duplication. While variation in the rate of short indel events clearly contributed to the observed differences in genome architectures and sizes among species, their relative contribution relative to other factors is unknown and likely varies among different phylogenetic clades. In mammals and aves, for example, it was estimated that microdeletions account for <10% of the total DNA lost over the past few dozen million years (Kapusta et al. 2017).

Variation in Indel Rates

Variation in substitution rates and patterns among different clades within the tree of life was extensively studied (Drake et al. 1998). In contrast, substantially less is known about variation in indel dynamics among different taxonomic clades. Gu and Li (1995) analyzed small data sets and used simple methods to infer indel rates; their estimates of the Zipfian length distribution parameter varied from 1.93 in noncoding mitochondrial DNA to 1.70 in primate globin noncoding regions. Fan et al. (2007) analyzed coding and noncoding indels in 18 mammalian genomes and found that the Zipfian length parameter ranged from 1.059 for deletions in chimpanzee to 1.883 for insertions in rabbit. Petrov et al. (1996) reported high intrinsic rate of DNA loss in *Drosophila*, an observation was later confirmed by Loewenthal et al. (2021), who studied differences in indel rates among 15 taxonomic groups and found that the insertion rates, the deletion rates, and the average size of an indel event substantially vary among these groups. Indel rates were also found to vary among different lineages of HIV-1 (Palmer and Poon 2019). Finally, indel rates are strongly correlated with substitution rates and were found to be negatively correlated with the effective population size (Lynch et al. 2023). Clearly, recent advances in methods to infer indel rates together with the wealth of genomic data should enable deeper understanding of the variation of indel rates across clades and the factors that drive such variation.

Conclusions

Overall, indels can often provide valuable evolutionary signals, aiding taxonomic classification and improving statistical support for phylogenetic branches that are weakly supported by substitutions. Increasingly, indel mutations are considered in biomedical applications. However, the analysis of indel variation is often hampered by the lack of appropriate bioinformatic pipelines; and, despite recent developments, statistical indel models and methods for practical applications are still lagging behind the more mature tools for analyzing SNVs. Therefore, it is important that researchers analyzing indel evolutionary dynamics are aware of current limitations and should follow “good-enough practices.”

Good-Enough Practices

First, indel information is only as good as the alignment that determines the placement of indels. As a result, indel information should only be used when the alignment was inferred with a phylogeny-aware aligner (see Table 1). However, some phylogeny-aware alignment software, such as PRANK, relies heavily on a guide tree to place indels. If this guide tree is wrong, then indels may be incorrectly placed. Ideally, each phylogeny should be scored using an MSA that is consistent with that particular phylogeny. Thus, for the best quality, the MSA and the tree should be inferred jointly (see Table 1). More specifically, when the goal is to infer the tree, uncertainty overalignments should be accounted for, and vice versa, when the goal is to infer alignments, uncertainty in the tree should be accounted for. At any rate, instead of point estimates of either the tree, the MSA, or both, it is best to consider the distribution of possible MSAs and trees, weighted by their probabilities in downstream analyses.

There is a high amount of phylogenetic information in shared indels. However, for this information to be accounted for, it is especially important to account for alternative placement of indels, i.e. to explicitly consider alignment uncertainty. This is true even when indels are inferred using a good model and state-of-the-art alignment programs. It is especially important to account for alignment uncertainty when analyzing homologous sequences with low sequence similarity. This includes sequences that diverged in the deep past and also sequences such as noncoding sequences that have diverged more recently but are not under strong selection. Perhaps the best method of accounting for alignment uncertainty is to average over alternative alignments using Bayesian methods. However, alternative methods that sacrifice some accuracy may be substantially faster, such as performing the same analysis on a collection of alternative alignments to see how much the result varies.

When using information from shared indels to group taxa on the phylogeny, it may be important to correctly assess the weight of evidence in shared indels. Models such as TKF1 or the model that treats "-" as a fifth nucleotide can exaggerate the evidence in longer indels, since they either explicitly or implicitly assume that gaps spanning multiple MSA columns evolved from multiple indel events of a single character. Since indels are rare events, this can substantially inflate the weight of longer indels. Thus, analyses that use information from shared indels to reconstruct phylogenies would ideally use models that distinguish the occurrence of an indel from the length of gaps within an MSA. Additionally, caution should be used when analyzing data sets that contain very long indels, such as the deletion of an entire domain. This is because current models of indel lengths often assume a geometric distribution on indel length, under which very long indels are not expected to occur.

Since genomic sequences are known to be highly heterogeneous, indel rates should be inferred from the data being analyzed rather than fixed to default values. These rates determine the relative weight of shared indels and shared substitutions in grouping taxa on the tree. However, indel rates that are inferred under single-character indel models do not have the same interpretation, as they attempt to fit the indel rate and the indel length with a single parameter. This single parameter then cannot be interpreted as the biological rate of indels. Further, as indels were shown to have different rates, models that assume the same rate for both should be avoided. Such considerations should also be accounted for when simulating MSAs, as otherwise, simulations fail to reflect the evolutionary patterns observed in empirical data sets.

Future Directions

Despite the remaining challenges, a good selection of statistical methods for analyzing indels is already available and should be used. Researchers can take the advantage of indel information by wider incorporation of suitable indel methods into all steps of bioinformatic pipelines, in accordance with "good-enough" practices and paying attention to inference uncertainty as discussed above.

In the medical field, the importance of indels is increasingly recognized, particularly when analyzing STRs, which due to their low information content were previously considered "junk DNA." The high mutation rate and complex error profiles of STRs make analysis of indels in these loci challenging. Better genotyping technologies and bioinformatic pipelines, including for imputation of multiallelic STRs, will open up new possibilities for understanding complex disease and monogenic disorders.

In evolutionary biology, statistical alignment approaches are limited by the accuracy of their indel models. Improved alignment accuracy depends on improved indel models. Such improvements can range from better approximations to the indel process at large branch lengths (Holmes 2020; De Maio 2021), to more realistic distributions of indel lengths (Cartwright 2007; Wygoda et al. 2024), to models of processes that we are currently ignoring, such changes in microsatellite length. These models are especially important in applications that use indels as informative characters, for example in phylogeny reconstruction.

In part because of the formidable technical challenges, existing models of indels have remained comparatively simple compared to single-nucleotide substitution models. These difficulties are exacerbated by the need to analyze data sets of growing sizes. In search for scalable solutions, fast approximations for phylogenetic likelihoods of substitution models were proposed (De Maio et al. 2023; Prillo et al. 2023), opening new avenues of research to include indels. Efficient representation of sequences can bring further advantages (Karasikov et al. 2020).

Big data support the emergence of novel data-driven machine learning approaches. This presents an opportunity to develop more sophisticated "black box" indel models, which could complement the more traditional partly mechanistic mutation models that have been used so far.

Machine Learning

In theory, machine learning could be used to predict insertion and deletion rates that depend on surrounding sequence context, as well as structural and functional features and signals of selection. While such factors could be explicitly modeled, deep networks may be able to extract relevant patterns directly from the analyzed sequences.

Currently machine learning models in evolutionary biology tend to be trained on simulated data (Suvorov et al. 2020; Azouri et al. 2021; Wang, Sun, et al. 2023). This may limit the degree to which unknown phenomena can be inferred, since we still cannot simulate sufficiently realistic MSAs (Trost et al. 2024). Further, recent work suggests that deep learning methods infer phylogenies similar to those inferred with ML, right or wrong (Thompson et al. 2024). However, for phylogeny and alignment reconstruction methods, we also lack reliable empirical benchmarks. We anticipate that machine learning algorithms will be increasingly introduced to molecular evolution inference algorithms and will enable more accurate and often faster inference of model parameters. Approach for representing MSAs in a form that can be processed by neural networks is just beginning. Current techniques involve converting input sequences to a string that is "translated" to a string representing the aligned sequences via "transformer" neural networks that are used in language translation (Dotan et al. 2024).

Neural networks do not provide a mechanistic model that explains why its predictions are correct. Thus, they might be able to predict that indels occur in particular places without an explanation of why they occur there.

Toward More Realistic Indel Models

One obvious step toward more realistic models would be relaxing the assumption of constant indel rate. In empirical protein data sets, indel rates are often higher in some regions (e.g. regions exposed to solvent in the tertiary structure) and lower in others (such as buried regions). In regions where the indel rate is higher, we expect to find more indels, but probably less shared indels. Future models should allow spatial variation of the indel rate along the sequence, possibly considering hotspot/coldspot regions. Such models should be fit to the empirical data, simultaneously with computing the alignment. This is a challenging problem because it involves *a priori* dividing the sequence into spatial regions while at the same time indels alter the space. Alignment methods that consider indel hotspots/coldspots are a step in this direction (Satija et al. 2008, 2009). Similarly, models like PIP, which rely on single-site indel events, can account for such factors by assigning site-specific indel evolutionary rate to each column in the MSA, similar to the common practice for substitutions. However, the practical value of such an approach has not yet been evaluated.

Previous reports clearly indicate the heterogeneity of indel rates also among taxa. For example, Petrov et al. (1996) reported an exceptionally high deletion rate in *Drosophila* and Loewenthal et al. (2021) showed that both the insertion rate and the deletion rate can have more than 2-fold differences among different clades. Together, such observations suggest that future models should account for shifts in indel rates within branches in the phylogenetic tree, possibly using similar approaches applied within covarion-like substitution models (Petrov et al. 1996; Galtier 2001; Loewenthal et al. 2021). Moreover, like substitutions, indels were observed to be context dependent (Chang and Benner 2004; de la Chaux et al. 2007; Messer and Arndt 2007; Kvikstad et al. 2009); therefore, including context dependency features could lead to better fit to data making a model more realistic.

More realistic models are expected to be more computationally demanding and applying them to large data sets may be computationally challenging. However, approximations that sacrifice some accuracy but nevertheless capture the main biological phenomena may be developed. Moreover, such realistic models could be very useful to study more intricate effects in targeted analyses of specific regions or proteins on smaller data sets.

In general, new models should aim to provide a better mechanistic understanding of indel evolutionary dynamics. Just like substitutions, indels are affected by a number of contacts in protein structure; therefore, one can anticipate the potential fitness effects of indels. Given this, is it possible to combine both substitutions and indels together in the mutation-selection framework (Teufel et al. 2018)? Good statistical models even without a mechanistic basis can be used for simulations that provide a better test of our current inference method.

Further biological realism can be achieved by bringing structural aspects into an indel model. Ideally, analyses of protein sequences would reflect the effects of selection due to changes in 3D structure and fitness of proteins when predicting the observed rates of substitution and indel mutations. One approach is to introduce the atomic coordinates as latent variables in a 3D diffusion process, initially focusing only on alpha carbons and assuming that each alpha carbon drifts independently (Challis and Schmidler 2012). Later models incorporate proximity constraints (Larson et al. 2020) and bond angles (Golden et al. 2017) between adjacent amino acids. There remain further technical challenges regarding consistency of such models, such as maintaining continuity of flanking atomic coordinates immediately prior and subsequent to indel events. One path forward is to continue development of these models, addressing these technical challenges; another is to develop alternative latent variable models based (for example) on pairwise interaction terms between amino acids (such as Potts models) or on residue contact graphs. Arguably, all latent variable approaches suffer from limitations inherent to the fact that the protein structure is, in reality, an emergent property of the biological sequence, not a jointly inherited one.

Correct handling of indels in comparative sequence analysis continues to present many open problems. Recent advances in machine learning have not yet solved this problem; indeed, many of the neural network methods used in genomic side step the handling of indels completely, assuming their input sequences are of fixed length. An improved treatment of this issue is an important challenge if we are to understand the observed range of natural genetic variation.

Acknowledgments

We would like to thank Jeffrey Thorne and three anonymous reviewers for constructive feedback that helped to improve this review.

Funding

B.D.R. acknowledges funding from the Louisiana Board of Regents RCS subprogram, NSF DEB1754142, and NIH R01AI153044. T.P. acknowledges the funding from Israel Science Foundation grant 2818/21. M.A. acknowledges the funding from the SNSF grant 315230_215379.

Data availability

There are no new data associated with this review article.

References

Aadland K, Kolaczkowski B. Alignment-integrated reconstruction of ancestral sequences improves accuracy. Genome Biol Evol. 2020:**12**(9):1549–1565. https://doi.org/10.1093/gbe/evaa164.

Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med. 2020:**26**(4):450–452. https://doi.org/10.1038/s41591-020-0820-9.

Angiuoli SV, Dunning Hotopp JC, Salzberg SL, Tettelin H. Improving pan-genome annotation using whole genome multiple alignment. BMC Bioinformatics. 2011:**12**(1):272. https://doi.org/10.1186/1471-2105-12-272.

Anisimova M, Bielawski JP, Yang Z. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol. 2001:**18**(8):1585–1592. https://doi.org/10.1093/oxfordjournals.molbev.a003945.

Arenas M. Advances in computer simulation of genome evolution: toward more realistic evolutionary genomics analysis by approximate Bayesian computation. J Mol Evol. 2015:**80**(3-4):189–192. https://doi.org/10.1007/s00239-015-9673-0.

Ashkenazy H, Cohen O, Pupko T, Huchon D. Indel reliability in indel-based phylogenetic inference. Genome Biol Evol. 2014:**6**(12): 3199–3209. https://doi.org/10.1093/gbe/evu252.

Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res. 2010:**38**(Web Server issue):W529–W533. https://doi.org/10.1093/nar/gkq399.

Ashkenazy H, Levy Karin E, Mertens Z, Cartwright RA, Pupko T. SpartaABC: a web server to simulate sequences with indel parameters inferred using an approximate Bayesian computation algorithm. Nucleic Acids Res. 2017:**45**(W1):W453–W457. https://doi.org/10.1093/nar/gkx322.

Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, Pupko T. FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res. 2012:**40**(W1):W580–W584. https://doi.org/10.1093/nar/gks498.

Ashkenazy H, Sela I, Levy Karin E, Landan G, Pupko T. Multiple sequence alignment averaging improves phylogeny reconstruction. Syst Biol. 2019:**68**(1):117–130. https://doi.org/10.1093/sysbio/syy036.

Azouri D, Abadi S, Mansour Y, Mayrose I, Pupko T. Harnessing machine learning to guide phylogenetic-tree search algorithms. Nat Commun. 2021:**12**(1):1983. https://doi.org/10.1038/s41467-021-22073-8.

Barton HJ, Zeng K. The impact of natural selection on short insertion and deletion variation in the great tit genome. Genome Biol Evol. 2019:**11**(6):1514–1524. https://doi.org/10.1093/gbe/evz068.

Belinky F, Cohen O, Huchon D. Large-scale parsimony analysis of metazoan indels in protein-coding genes. Mol Biol Evol. 2010:**27**(2):441–451. https://doi.org/10.1093/molbev/msp263.

Benner SA, Cohen MA, Gonnet GH. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. J Mol Biol. 1993:**229**(4):1065–1082. https://doi.org/10.1006/jmbi.1993.1105.

Bennett EP, Petersen BL, Johansen IE, Niu Y, Yang Z, Chamberlain CA, Met Ö, Wandall HH, Frödin M. INDEL detection, the "Achilles heel" of precise genome editing: a survey of methods for accurate profiling of gene editing induced indels. Nucleic Acids Res. 2020:**48**(21):11958–11981. https://doi.org/10.1093/nar/gkaa975.

Bétermier M, Bertrand P, Lopez BS. Is non-homologous end-joining really an inherently error-prone process? PLoS Genet. 2014:**10**(1): e1004086. https://doi.org/10.1371/journal.pgen.1004086.

Bhargava R, Sandhu M, Muk S, Lee G, Vaidehi N, Stark JM. C-NHEJ without indels is robust and requires synergistic function of distinct XLF domains. Nat Commun. 2018:**9**(1):2484. https://doi.org/10.1038/s41467-018-04867-5.

Bilgin Sonay T, Carvalho T, Robinson MD, Greminger MP, Krützen M, Comas D, Highnam G, Mittelman D, Sharp A, Marques-Bonet T, et al. Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. Genome Res. 2015:**25**(11):1591–1599. https://doi.org/10.1101/gr.190868.115.

Birth N, Dencker T, Morgenstern B. Insertions and deletions as phylogenetic signal in an alignment-free context. PLoS Comput Biol. 2022:**18**(8):e1010303. https://doi.org/10.1371/journal.pcbi.1010303.

Bishop MJ, Thompson EA. Maximum likelihood alignment of DNA sequences. J Mol Biol. 1986:**190**(2):159–165. https://doi.org/10.1016/0022-2836(86)90289-5.

Boatwright JL, Sapkota S, Kresovich S. Functional genomic effects of indels using Bayesian genome-phenome wide association studies in sorghum. Front Genet. 2023:**14**:1143395. https://doi.org/10.3389/fgene.2023.1143395.

Boersma YL, Plückthun A. DARPins and other repeat protein scaffolds: advances in engineering and applications. Curr Opin Biotechnol. 2011:**22**(6):849–857. https://doi.org/10.1016/j.copbio.2011.06.004.

Bouchard-Côté A, Jordan MI. Evolutionary inference via the Poisson Indel Process. Proc Natl Acad Sci U S A. 2013:**110**(4):1160–1166. https://doi.org/10.1073/pnas.1220450110.

Britten RJ, Rowen L, Williams J, Cameron RA. Majority of divergence between closely related DNA samples is due to indels. Proc Natl Acad Sci U S A. 2003:**100**(8):4661–4665. https://doi.org/10.1073/pnas.0330964100.

Burssed B, Zamariolli M, Bellucco FT, Melaragno MI. Mechanisms of structural chromosomal rearrangement formation. Mol Cytogenet. 2022:**15**(1):23. https://doi.org/10.1186/s13039-022-00600-6.

Cartwright RA. Ngila: global pairwise alignments with logarithmic and affine gap costs. Bioinformatics. 2007:**23**(11):1427–1428. https://doi.org/10.1093/bioinformatics/btm095.

Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000:**17**(4): 540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334.

Chaisson MJ, Raphael BJ, Pevzner PA. Microinversions in mammalian evolution. Proc Natl Acad Sci U S A. 2006:**103**(52):19824–19829. https://doi.org/10.1073/pnas.0603984103.

Challis CJ, Schmidler SC. A stochastic evolutionary model for protein structure alignment and phylogeny. Mol Biol Evol. 2012:**29**(11): 3575–3587. https://doi.org/10.1093/molbev/mss167.

Chang MSS, Benner SA. Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments. J Mol Biol. 2004:**341**(2): 617–631. https://doi.org/10.1016/j.jmb.2004.05.045.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012:**7**(10):e46688. https://doi.org/10.1371/journal.pone.0046688.

Chuzhanova NA, Anassis EJ, Ball EV, Krawczak M, Cooper DN. Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. Hum Mutat. 2003:**21**(1):28–44. https://doi.org/10.1002/humu.10146.

Cooke DP, Wedge DC, Lunter G. A unified haplotype-based method for accurate and comprehensive variant calling. Nat Biotechnol. 2021:**39**(7):885–892. https://doi.org/10.1038/s41587-021-00861-3.

Copley SD. Evolution and the enzyme. In: Liu H-W (ben), Mander L, editors. Comprehensive natural products II. Cambridge, Massachusetts, United States: Elsevier; 2010. p. 9–46.

Cranmer K, Brehmer J, Louppe G. The frontier of simulation-based inference. Proc Natl Acad Sci U S A. 2020:**117**(48):30055–30062. https://doi.org/10.1073/pnas.1912789117.

Dagan T. Phylogenomic networks. Trends Microbiol. 2011:**19**(10): 483–491. https://doi.org/10.1016/j.tim.2011.07.001.

Dai J, Huang M, Amos CI, Hung RJ, Tardon A, Andrew A, Chen C, Christiani DC, Albanes D, Rennert G, et al. Genome-wide association study of INDELs identified four novel susceptibility loci associated with lung cancer risk. Int J Cancer. 2020:**146**(10): 2855–2864. https://doi.org/10.1002/ijc.32698.

Dalquen DA, Anisimova M, Gonnet GH, Dessimoz C. ALF—a simulation framework for genome evolution. Mol Biol Evol. 2012:**29**(4):1115–1123. https://doi.org/10.1093/molbev/msr268.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. The variant call format and VCFtools. Bioinformatics. 2011:**27**(15):2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

de Groot S, Mailund T, Lunter G, Hein J. Investigating selection on viruses: a statistical alignment approach. BMC Bioinformatics. 2008:**9**(1):304. https://doi.org/10.1186/1471-2105-9-304.

de Jong WW, Rydén L. Causes of more frequent deletions than insertions in mutations and protein evolution. Nature. 1981:**290**(5802): 157–159. https://doi.org/10.1038/290157a0.

de Juan D, Pazos F, Valencia A. Emerging methods in protein coevolution. Nat Rev Genet. 2013:**14**(4):249–261. https://doi.org/10.1038/nrg3414.

de la Chaux N, Messer PW, Arndt PF. DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. BMC Evol Biol. 2007:**7**(1):191. https://doi.org/10.1186/1471-2148-7-191.

De Maio N. The cumulative indel model: fast and accurate statistical evolutionary alignment. Syst Biol. 2021:**70**(2):236–257. https://doi.org/10.1093/sysbio/syaa050.

De Maio N, Kalaghatgi P, Turakhia Y, Corbett-Detig R, Minh BQ, Goldman N. Maximum likelihood pandemic-scale phylogenetics. Nat Genet. 2023:**55**(5):746–752. https://doi.org/10.1038/s41588-023-01368-0.

Depienne C, Mandel J-L. 30 years of repeat expansion disorders: what have we learned and what are the remaining challenges? Am J Hum Genet. 2021:**108**(5):764–785. https://doi.org/10.1016/j.ajhg.2021.03.011.

Dessimoz C, Gil M. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol. 2010:**11**(4):R37. https://doi.org/10.1186/gb-2010-11-4-r37.

Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S. ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res. 2005:**15**(2):330–340. https://doi.org/10.1101/gr.2821705.

dos Reis M, Donoghue PCJ, Yang Z. Bayesian molecular clock dating of species divergences in the genomics era. Nat Rev Genet. 2016:**17**(2):71–80. https://doi.org/10.1038/nrg.2015.8.

wDotan E, Wygoda E, Ecker N, Alburquerque M, Avram O, Belinkov Y, Pupko T. 2024. BetaAlign: a deep learning approach for multiple sequence alignment. bioRxiv 586462. https://doi.org/10.1101/2024.03.24.586462, 3 April 2024, preprint: not peer reviewed.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. Rates of spontaneous mutation. Genetics. 1998:**148**(4):1667–1686. https://doi.org/10.1093/genetics/148.4.1667.

Durbin R. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge, United Kingdom: Cambridge University Press; 1998.

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science. 2021:**372**(6537):eabf7117. https://doi.org/10.1126/science.abf7117.

Eddy SR. Profile hidden Markov models. Bioinformatics. 1998:**14**(9): 755–763. https://doi.org/10.1093/bioinformatics/14.9.755.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004:**32**(5):1792–1797. https://doi.org/10.1093/nar/gkh340.

Edgar RC. Muscle5: high-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. Nat Commun. 2022:**13**(1):6968. https://doi.org/10.1038/s41467-022-34630-w.

Elena-Real CA, Mier P, Sibille N, Andrade-Navarro MA, Bernadó P. Structure-function relationships in protein homorepeats. Curr Opin Struct Biol. 2023:**83**:102726. https://doi.org/10.1016/j.sbi.2023.102726.

Elena SF. The role of indels in evolution and pathogenicity of RNA viruses. Proc Natl Acad Sci U S A. 2023:**120**(33):e2310785120. https://doi.org/10.1073/pnas.2310785120.

Fan Y, Wang W, Ma G, Liang L, Shi Q, Tao S. Patterns of insertion and deletion in mammalian genomes. Curr Genomics. 2007:**8**(6): 370–378. https://doi.org/10.2174/138920207783406479.

Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981:**17**(6):368–376. https://doi.org/10.1007/BF01734359.

Felsenstein J. Inferring phylogenies. Sunderland (MA)): Sinauer Associates; 2004.

Ferlaino M, Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. An integrative approach to predicting the functional effects of small indels in non-coding regions of the human genome. BMC Bioinformatics. 2017:**18**(1):442. https://doi.org/10.1186/s12859-017-1862-y.

Fitch WM. Aspects of molecular evolution. Annu Rev Genet. 1973:**7**(1): 343–380. https://doi.org/10.1146/annurev.ge.07.120173.002015.

Fleissner R, Metzler D, von Haeseler A. Simultaneous statistical multiple alignment and phylogeny reconstruction. Syst Biol. 2005:**54**(4): 548–561. https://doi.org/10.1080/10635150590950371.

Fletcher W, Yang Z. INDELible: a flexible simulator of biological sequence evolution. Mol Biol Evol. 2009:**26**(8):1879–1888. https://doi.org/10.1093/molbev/msp098.

Fletcher W, Yang Z. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol Biol Evol. 2010:**27**(10):2257–2267. https://doi.org/10.1093/molbev/msq115.

Foley G, Mora A, Ross CM, Bottoms S, Sützl L, Lamprecht ML, Zaugg J, Essebier A, Balderson B, Newell R, et al. Engineering indel and substitution variants of diverse and ancient enzymes using Graphical Representation of Ancestral Sequence Predictions (GRASP). PLoS Comput Biol. 2022:**18**(10):e1010633. https://doi.org/10.1371/journal.pcbi.1010633.

Gall-Duncan T, Sato N, Yuen RKC, Pearson CE. Advancing genomic technologies and clinical awareness accelerates discovery of disease-associated tandem repeat sequences. Genome Res. 2022:**32**(1):1–27. https://doi.org/10.1101/gr.269530.120.

Galtier N. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol. 2001:**18**(5):866–873. https://doi.org/10.1093/oxfordjournals.molbev.a003868.

Gatesy J, DeSalle R, Wheeler W. Alignment-ambiguous nucleotide sites and the exclusion of systematic data. Mol Phylogenet Evol. 1993:**2**(2):152–157. https://doi.org/10.1006/mpev.1993.1015.

Gaya E, Redelings BD, Navarro-Rosinés P, Llimona X, De Cáceres M, Lutzoni F. Align or not to align? Resolving species complexes within the *Caloplaca saxicola* group as a case study. Mycologia. 2011:**103**(2):361–378. https://doi.org/10.3852/10-120.

Godfroid M, Dagan T, Merker M, Kohl TA, Diel R, Maurer FP, Niemann S, Kupczok A. Insertion and deletion evolution reflects antibiotics selection pressure in a *Mycobacterium tuberculosis* outbreak. PLoS Pathog. 2020:**16**(9):e1008357. https://doi.org/10.1371/journal.ppat.1008357.

Golden M, García-Portugués E, Sørensen M, Mardia KV, Hamelryck T, Hein J. A generative angular model of protein structure evolution. Mol Biol Evol. 2017:**34**(8):2085–2100. https://doi.org/10.1093/molbev/msx137.

Goldman N. Statistical tests of models of DNA substitution. J Mol Evol. 1993:**36**(2):182–198. https://doi.org/10.1007/BF00166252.

Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 1994:**11**(5): 725–736. https://doi.org/10.1093/oxfordjournals.molbev.a040153.

Grasso C, Lee C. Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. Bioinformatics. 2004:**20**(10): 1546–1556. https://doi.org/10.1093/bioinformatics/bth126.

Graur D, Shuali Y, Li WH. Deletions in processed pseudogenes accumulate faster in rodents than in humans. J Mol Evol. 1989:**28**(4): 279–285. https://doi.org/10.1007/BF02103423.

Gu X, Li WH. The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. J Mol Evol. 1995:**40**(4):464–473. https://doi.org/10.1007/BF00164032.

Gupta M, Zaharias P, Warnow T. Accurate large-scale phylogeny-aware alignment using BAli-Phy. Bioinformatics. 2021:**37**(24):4677–4683. https://doi.org/10.1093/bioinformatics/btab555.

Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet. 2016:**48**(1):22–29. https://doi.org/10.1038/ng.3461.

Haerty W, Golding GB. Genome-wide evidence for selection acting on single amino acid repeats. Genome Res. 2010:**20**(6):755–760. https://doi.org/10.1101/gr.101246.109.

Hall BG. Simulating DNA coding sequence evolution with EvolveAGene 3. Mol Biol Evol. 2008:**25**(4):688–695. https://doi.org/10.1093/molbev/msn008.

Halldorsson BV, Eggertsson HP, Moore KHS, Hauswedell H, Eiriksson O, Ulfarsson MO, Palsson G, Hardarson MT, Oddsson A, Jensson BO, et al. The sequences of 150,119 genomes in the UK Biobank. Nature. 2022:**607**(7920):732–740. https://doi.org/10.1038/s41586-022-04965-x.

Haller BC, Messer PW. SLiM 4: multispecies eco-evolutionary modeling. Am Nat. 2023:**201**(5):E127–E139. https://doi.org/10.1086/723601.

Heger A, Ponting CP. OPTIC: orthologous and paralogous transcripts in clades. Nucleic Acids Res. 2008:**36**(Database):D267–D270. https://doi.org/10.1093/nar/gkm852.

Hein J. A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. Mol Biol Evol. 1989:**6**(6):649–668. https://doi.org/10.1093/oxfordjournals.molbev.a040577.

Hein J. An algorithm for statistical alignment of sequences related by a binary tree. Pac Symp Biocomput. 2001:179–190. https://doi.org/10.1142/9789814447362_0019.

Hein J, Wiuf C, Knudsen B, Møller MB, Wibling G. Statistical alignment: computational properties, homology testing and goodness-of-fit. J Mol Biol. 2000:**302**(1):265–279. https://doi.org/10.1006/jmbi.2000.4061.

Hickey G, Blanchette M. A probabilistic model for sequence alignment with context-sensitive indels. J Comput Biol. 2011:**18**(11): 1449–1464. https://doi.org/10.1089/cmb.2011.0157.

Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, Dawson ET, Garrison E, Novak AM, Paten B. Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biol. 2020:**21**(1):35. https://doi.org/10.1186/s13059-020-1941-7.

Holmes IH. Using guide trees to construct multiple-sequence evolutionary HMMs. Bioinformatics. 2003:**19**(suppl_1):i147–i157. https://doi.org/10.1093/bioinformatics/btg1019.

Holmes IH. Historian: accurate reconstruction of ancestral sequences and evolutionary rates. Bioinformatics. 2017a:**33**(8):1227–1229. https://doi.org/10.1093/bioinformatics/btw791.

Holmes IH. Solving the master equation for indels. BMC Bioinformatics. 2017b:**18**(1):255. https://doi.org/10.1186/s12859-017-1665-1.

Holmes IH. A model of indel evolution by finite-state, continuous-time machines. Genetics. 2020:**216**(4):1187–1204. https://doi.org/10.1534/genetics.120.303630.

Holmes IH, Bruno WJ. Evolutionary HMMs: a Bayesian approach to multiple alignment. Bioinformatics. 2001:**17**(9):803–820. https://doi.org/10.1093/bioinformatics/17.9.803.

Hon T, Mars K, Young G, Tsai Y-C, Karalius JW, Landolin JM, Maurer N, Kudrna D, Hardigan MA, Steiner CC, et al. Highly accurate long-read HiFi sequencing data for five complex genomes. Sci Data. 2020:**7**(1):399. https://doi.org/10.1038/s41597-020-00743-4.

Horton CA, Alexandari AM, Hayes MGB, Marklund E, Schaepe JM, Aditham AK, Shah N, Suzuki PH, Shrikumar A, Afek A, et al. Short tandem repeats bind transcription factors to tune eukaryotic gene expression. Science. 2023:**381**(6664):eadd1250. https://doi.org/10.1126/science.add1250.

Hu J, Ng PC. Predicting the effects of frameshifting indels. Genome Biol. 2012:**13**(2):R9. https://doi.org/10.1186/gb-2012-13-2-r9.

Huelsenbeck J, Rannala B. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst Biol. 2004:**53**(6):904–913. https://doi.org/10.1080/10635150490522629.

Iglhaut C, Pečerska J, Gil M, Anisimova M. Please mind the gap: indel-aware parsimony for fast and accurate ancestral sequence reconstruction and multiple sequence alignment including long indels. Mol Biol Evol. 2024; **41**(7):msae109. https://doi.org/10.1093/molbev/msae109.

Jain A, Roorkiwal M, Kale S, Garg V, Yadala R, Varshney RK. Indel markers: an extended marker resource for molecular breeding in chickpea. PLoS One. 2019:**14**(3):e0213999. https://doi.org/10.1371/journal.pone.0213999.

Jakubosky D, D'Antonio M, Bonder MJ, Smail C, Donovan MKR, Young Greenwald WW, Matsui H, i2QTL Consortium, D'Antonio-Chronowska A, Stegle O, et al. Properties of structural variants and short tandem repeats associated with gene expression and complex traits. Nat Commun. 2020:**11**(1):2927. https://doi.org/10.1038/s41467-020-16482-4.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. 2014: **346**(6215):1320–1331. https://doi.org/10.1126/science.1253451.

Jordan G, Goldman N. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Mol Biol Evol. 2012:**29**(4):1125–1139. https://doi.org/10.1093/molbev/msr272.

Jowkar G, Pečerska J, Maiolo M, Gil M, Anisimova M. ARPIP: ancestral sequence reconstruction with insertions and deletions under the Poisson Indel Process. Syst Biol. 2023:**72**(2):307–318. https://doi.org/10.1093/sysbio/syac050.

Jukes TH, Cantor CR. Evolution of protein molecules. Mammalian protein metabolism. New York, United States: Academic Press; 1969. p. 21–132.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021:**596**(7873):583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kapli P, Yang Z, Telford MJ. Phylogenetic tree building in the genomic age. Nat Rev Genet. 2020:**21**(7):428–444. https://doi.org/10.1038/s41576-020-0233-0.

Kapusta A, Suh A, Feschotte C. Dynamics of genome size evolution in birds and mammals. Proc Natl Acad Sci U S A. 2017:**114**(8): E1460-E1469. https://doi.org/10.1073/pnas.1616702114.

Karasikov M, Mustafa H, Danciu D, Zimmermann M, Barber C, Rätsch G, Kahles A. 2020. Indexing all life's known biological sequences. bioRxiv 322164. https://doi.org/10.1101/2020.10.01.322164, 14 May 2024, preprint: not peer reviewed.

Katoh K, Misawa K, Kuma K-I, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002:**30**(14):3059–3066. https://doi.org/10.1093/nar/gkf436.

Kazazian HH Jr. Mobile elements: drivers of genome evolution. Science. 2004:**303**(5664):1626–1632. https://doi.org/10.1126/science.1089670.

Kim R, Guo J-T. Systematic analysis of short internal indels and their impact on protein folding. BMC Struct Biol. 2010:**10**(1):24. https://doi.org/10.1186/1472-6807-10-24.

Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980:**16**(2):111–120. https://doi.org/10.1007/BF01731581.

Knudsen B, Miyamoto MM. Sequence alignments and pair hidden Markov models using evolutionary history. J Mol Biol. 2003:**333**(2):453–460. https://doi.org/10.1016/j.jmb.2003.08.015.

Kosiol C, Anisimova M. Selection acting on genomes. Methods Mol. Biol. 2019:**1910**:373–397. https://doi.org/10.1007/978-1-4939-9074-0_12.

Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, Gonzalez-Porta M, Eberle MA, Tezak Z, Lababidi S, et al. Best practices for benchmarking germline small-variant calls in human genomes. Nat Biotechnol. 2019:**37**(5):555–560. https://doi.org/10.1038/s41587-019-0054-x.

Kundu K, Tardaguila M, Mann AL, Watt S, Ponstingl H, Vasquez L, Von Schiller D, Morrell NW, Stegle O, Pastinen T, et al. Genetic associations at regulatory phenotypes improve fine-mapping of causal variants for 12 immune-mediated diseases. Nat Genet. 2022:**54**(3):251–262. https://doi.org/10.1038/s41588-022-01025-y.

Kuo C-H, Ochman H. Deletional bias across the three domains of life. Genome Biol Evol. 2009:**1**:145–152. https://doi.org/10.1093/gbe/evp016.

Kvikstad EM, Chiaromonte F, Makova KD. Ride the wavelet: a multiscale analysis of genomic contexts flanking small insertions and deletions. Genome Res. 2009:**19**(7):1153–1164. https://doi.org/10.1101/gr.088922.108.

Lake JA. The order of sequence alignment can bias the selection of tree topology. Mol Biol Evol. 1991:**8**(3):378–385. https://doi.org/10.1093/oxfordjournals.molbev.a040654.

Landan G, Graur D. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol. 2007:**24**(6):1380–1383. https://doi.org/10.1093/molbev/msm060.

Larson G, Thorne JL, Schmidler S. Incorporating nearest-neighbor site dependence into protein evolution models. J Comput Biol. 2020:**27**(3):361–375. https://doi.org/10.1089/cmb.2019.0500.

Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. Bioinformatics. 2002:**18**(3):452–464. https://doi.org/10.1093/bioinformatics/18.3.452.

Lemmon AR, Moriarty EC. The importance of proper model assumption in Bayesian phylogenetics. Syst Biol. 2004:**53**(2):265–277. https://doi.org/10.1080/10635150490423520.

Levinson G, Gutman GA. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol. 1987:**4**(3):203–221. https://doi.org/10.1093/oxfordjournals.molbev.a040442.

Levy RM, Haldane A, Flynn WF. Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. Curr Opin Struct Biol. 2017:**43**:55–62. https://doi.org/10.1016/j.sbi.2016.11.004.

Levy Karin E, Ashkenazy H, Hein J, Pupko T. A simulation-based approach to statistical alignment. Syst Biol. 2019:**68**(2):252–266. https://doi.org/10.1093/sysbio/syy059.

Levy Karin E, Rabin A, Ashkenazy H, Shkedy D, Avram O, Cartwright RA, Pupko T. Inferring indel parameters using a simulation-based approach. Genome Biol Evol. 2015:**7**(12):3226–3238. https://doi.org/10.1093/gbe/evv212.

Levy Karin E, Shkedy D, Ashkenazy H, Cartwright RA, Pupko T. Inferring rates and length-distributions of indels using approximate Bayesian computation. Genome Biol Evol. 2017:**9**(5):1280–1294. https://doi.org/10.1093/gbe/evx084.

Levy Karin E, Susko E, Pupko T. Alignment errors strongly impact likelihood-based tests for comparing topologies. Mol Biol Evol. 2014:**31**(11):3057–3067. https://doi.org/10.1093/molbev/msu231.

Li C, Zhi D, Wang K, Liu X. MetaRNN: differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning. Genome Med. 2022:**14**(1):115. https://doi.org/10.1186/s13073-022-01120-z.

Light S, Sagit R, Sachenkova O, Ekman D, Elofsson A. Protein expansion is primarily due to indels in intrinsically disordered regions. Mol Biol Evol. 2013:**30**(12):2645–2653. https://doi.org/10.1093/molbev/mst157.

Lim D, Blanchette M. EvoLSTM: context-dependent models of sequence evolution using a sequence-to-sequence LSTM. Bioinformatics. 2020:**36**(Supplement_1):i353–i361. https://doi.org/10.1093/bioinformatics/btaa447.

Lin M, Whitmire S, Chen J, Farrel A, Shi X, Guo J-T. Effects of short indels on protein structure and function in human genomes. Sci Rep. 2017:**7**(1):9313. https://doi.org10.1038/s41598-017-09287-x.

Liu Z, Zheng H, Lin H, Li M, Yuan R, Peng J, Xiong Q, Sun J, Li B, Wu J, et al. Identification of common deletions in the spike protein of severe acute respiratory syndrome coronavirus 2. J Virol. 2020:**94**(17):e00790-20. https://doi.org/10.1128/JVI.00790-20.

Loewenthal G, Rapoport D, Avram O, Moshe A, Wygoda E, Itzkovitch A, Israeli O, Azouri D, Cartwright RA, Mayrose I, et al. A probabilistic model for indel evolution: differentiating insertions from deletions. Mol Biol Evol. 2021:**38**(12):5769–5781. https://doi.org/10.1093/molbev/msab266.

Löytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A. 2005:**102**(30):10557–10562. https://doi.org/10.1073/pnas.0409137102.

Löytynoja A, Goldman N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science. 2008:**320**(5883):1632–1635. https://doi.org/10.1126/science.1158395.

Löytynoja A, Goldman N. Short template switch events explain mutation clusters in the human genome. Genome Res. 2017:**27**(6):1039–1049. https://doi.org/10.1101/gr.214973.116.

Löytynoja A, Vilella AJ, Goldman N. Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. Bioinformatics. 2012:**28**(13):1684–1691. https://doi.org/10.1093/bioinformatics/bts198.

Lü Y, Cui X, Li R, Huang P, Zong J, Yao D, Li G, Zhang D, Yuan Z. Development of genome-wide insertion/deletion markers in rice based on graphic pipeline platform. J Integr Plant Biol. 2015:**57**(11):980–991. https://doi.org/10.1111/jipb.12354.

Lu JT, Wang Y, Gibbs RA, Yu F. Characterizing linkage disequilibrium and evaluating imputation power of human genomic insertion–deletion polymorphisms. Genome Biol. 2012:**13**(2):R15. https://doi.org/10.1186/gb-2012-13-2-r15.

Lunter G. Dog as an outgroup to human and mouse. PLoS Comput Biol. 2007:**3**(4):e74. https://doi.org/10.1371/journal.pcbi.0030074.

Lunter G, Miklós I, Drummond A, Jensen JL, Hein J. Bayesian coestimation of phylogeny and sequence alignment. BMC Bioinformatics. 2005:**6**(1):83. https://doi.org/10.1186/1471-2105-6-83.

Lunter GA, Miklós I, Song YS, Hein J. An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. J Comput Biol. 2003:**10**(6):869–889. https://doi.org/10.1089/106652703322756122.

Lunter G, Ponting CP, Hein J. Genome-wide identification of human functional DNA using a neutral indel model. PLoS Comput Biol. 2006:**2**(1):e5. https://doi.org/10.1371/journal.pcbi.0020005.

Ly-Trong N, Naser-Khdour S, Lanfear R, Minh BQ. AliSim: a fast and versatile phylogenetic sequence simulator for the genomic era. Mol Biol Evol. 2022:**39**(5):msac092. https://doi.org/10.1093/molbev/msac092.

Lynch M, Ali F, Lin T, Wang Y, Ni J, Long H. The divergence of mutation rates and spectra across the tree of life. EMBO Rep. 2023:**24**(10):e57561. https://doi.org/10.15252/embr.202357561.

Magee AF, Hilton SK, DeWitt WS. Robustness of phylogenetic inference to model misspecification caused by pairwise epistasis. Mol Biol Evol. 2021:**38**(10):4603–4615. https://doi.org/10.1093/molbev/msab163.

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019:**20**(1):246. https://doi.org/10.1186/s13059-019-1828-7.

Maiolo M, Gatti L, Frei D, Leidi T, Gil M, Anisimova M. ProPIP: a tool for progressive multiple sequence alignment with Poisson Indel Process. BMC Bioinformatics. 2021:**22**(1):518. https://doi.org/10.1186/s12859-021-04442-8.

Maiolo M, Ulzega S, Gil M, Anisimova M. Accelerating phylogeny-aware alignment with indel evolution using short time Fourier transform. NAR Genom Bioinform. 2020:**2**(4):lqaa092. https://doi.org/10.1093/nargab/lqaa092.

Maiolo M, Zhang X, Gil M, Anisimova M. Progressive multiple sequence alignment with indel evolution. BMC Bioinformatics. 2018:**19**(1):331. https://doi.org/10.1186/s12859-018-2357-1.

Marwaha S, Knowles JW, Ashley EA. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. Genome Med. 2022:**14**(1):23. https://doi.org/10.1186/s13073-022-01026-w.

McClintock B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A. 1950:**36**(6):344–355. https://doi.org/10.1073/pnas.36.6.344.

Mehta A, Haber JE. Sources of DNA double-strand breaks and models of recombinational DNA repair. Cold Spring Harb Perspect Biol. 2014:**6**(9):a016428. https://doi.org/10.1101/cshperspect.a016428.

Messer PW, Arndt PF. The majority of recent short DNA insertions in the human genome are tandem duplications. Mol Biol Evol. 2007:**24**(5):1190–1197. https://doi.org/10.1093/molbev/msm035.

Metzler D. Statistical alignment based on fragment insertion and deletion models. Bioinformatics. 2003:**19**(4):490–499. https://doi.org/10.1093/bioinformatics/btg026.

Metzler D, Fleißner R, Wakolbinger A, von Haeseler A. Assessing variability by joint sampling of alignments and mutation rates. J Mol Evol. 2001:**53**(6):660–669. https://doi.org/10.1007/s002390010253.

Miklós I, Lunter GA, Holmes I. A "long indel" model for evolutionary sequence alignment. Mol Biol Evol. 2004:**21**(3):529–540. https://doi.org/10.1093/molbev/msh043.

Miles A, Iqbal Z, Vauterin P, Pearson R, Campino S, Theron M, Gould K, Mead D, Drury E, O'Brien J, et al. Indels, structural variation, and recombination drive genomic diversity in *Plasmodium falciparum*. Genome Res. 2016:**26**(9):1288–1299. https://doi.org/10.1101/gr.203711.115.

Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. 2006:**16**(9):1182–1190. https://doi.org/10.1101/gr.4565806.

Mills RE, Pittard WS, Mullaney JM, Farooq U, Creasy TH, Mahurkar AA, Kemeza DM, Strassler DS, Ponting CP, Webber C, et al. Natural genetic variation caused by small insertions and deletions in the human genome. Genome Res. 2011:**21**(6):830–839. https://doi.org/10.1101/gr.115907.110.

Moler C, Van Loan C. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Rev. 2006:**45**:3–49. https://doi.org/10.1137/S00361445024180.

Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, et al. The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. Genome Res. 2013:**23**(5):749–761. https://doi.org/10.1101/gr.148718.112.

Morrison DA, Ellis JT. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. Mol Biol Evol. 1997:**14**(4):428–441. https://doi.org/10.1093/oxfordjournals.molbev.a025779.

Moshe A, Pupko T. Ancestral sequence reconstruction: accounting for structural information by averaging over replacement matrices. Bioinformatics. 2019:**35**(15):2562–2568. https://doi.org/10.1093/bioinformatics/bty1031.

Mugridge NB, Morrison DA, Jäkel T, Heckeroth AR, Tenter AM, Johnson AM. Effects of sequence alignment and structural domains of ribosomal DNA on phylogeny reconstruction for the protozoan family sarcocystidae. Mol Biol Evol. 2000:**17**(12):1842–1853. https://doi.org/10.1093/oxfordjournals.molbev.a026285.

Mularoni L, Ledda A, Toll-Riera M, Albà MM. Natural selection drives the accumulation of amino acid tandem repeats in human proteins. Genome Res. 2010:**20**(6):745–754. https://doi.org/10.1101/gr.101261.109.

Mullaney JM, Mills RE, Pittard WS, Devine SE. Small insertions and deletions (INDELs) in human genomes. Hum Mol Genet. 2010:**19**(R2):R131–R136. https://doi.org/10.1093/hmg/ddq400.

Nánási M, Vinař T, Brejová B. Probabilistic approaches to alignment with tandem repeats. Algorithms Mol Biol. 2014:**9**(1):3. https://doi.org/10.1186/1748-7188-9-3.

Novák A, Miklós I, Lyngsø R, Hein J. StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. Bioinformatics. 2008:**24**(20):2403–2404. https://doi.org/10.1093/bioinformatics/btn457.

Nute M, Saleh E, Warnow T. Evaluating statistical multiple sequence alignment in comparison to other alignment methods on protein data sets. Syst Biol. 2019:**68**(3):396–411. https://doi.org/10.1093/sysbio/syy068.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, et al. The Norway spruce genome sequence and conifer genome evolution. Nature. 2013:**497**(7451):579–584. https://doi.org/10.1038/nature12211.

Palmer J, Poon AFY. Phylogenetic measures of indel rate variation among the HIV-1 group M subtypes. Virus Evol. 2019:**5**(2):vez022. https://doi.org/10.1093/ve/vez022.

Pascarella S, Argos P. Analysis of insertions/deletions in protein structures. J Mol Biol. 1992:**224**(2):461–471. https://doi.org/10.1016/0022-2836(92)91008-D.

Pečerska J, Gil M, Anisimova M. 2021. Joint alignment and tree inference. bioRxiv 462230. https://doi.org/10.1101/2021.09.28.462230, 30 September 2021, preprint: not peer reviewed.

Penn O, Privman E, Landan G, Graur D, Pupko T. An alignment confidence score capturing robustness to guide tree uncertainty. Mol Biol Evol. 2010:**27**(8):1759–1767. https://doi.org/10.1093/molbev/msq066.

Petrov DA, Lozovskaya ER, Hartl DL. High intrinsic rate of DNA loss in *Drosophila*. Nature. 1996:**384**(6607):346–349. https://doi.org/10.1038/384346a0.

Prillo S, Deng Y, Boyeau P, Li X, Chen P-Y, Song YS. CherryML: scalable maximum likelihood estimation of phylogenetic models. Nat Methods. 2023:**20**(8):1232–1236. https://doi.org/10.1038/s41592-023-01917-9.

Privman E, Penn O, Pupko T. Improving the performance of positive selection inference by filtering unreliable alignment regions. Mol Biol Evol. 2012:**29**(1):1–5. https://doi.org/10.1093/molbev/msr177.

Pupko T, Mayrose I. A gentle introduction to probabilistic evolutionary models. In: Scornavacca C, Delsuc F, Galtier N, editors. Phylogenetics in the genomic era. HAL open science. Authors open access book: No commercial publisher; 2020. p. 1.1:1–1.1:21. Available from: https://hal.science/hal-02535102/.

Qian B, Goldstein RA. Distribution of indel lengths. Proteins. 2001:**45**(1):102–104. https://doi.org/10.1002/prot.1129.

Rambaut A, Grassly NC. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci. 1997:**13**(3):235–238. https://doi.org/10.1093/bioinformatics/13.3.235.

Rao RSP, Ahsan N, Xu C, Su L, Verburgt J, Fornelli L, Kihara D, Xu D. Evolutionary dynamics of indels in SARS-CoV-2 spike glycoprotein. Evol Bioinform Online. 2021:**17**:11769343211064616. https://doi.org/10.1177/11769343211064616.

Redelings B. Erasing errors due to alignment ambiguity when estimating positive selection. Mol Biol Evol. 2014:**31**(8):1979–1993. https://doi.org/10.1093/molbev/msu174.

Redelings BD, Suchard MA. Joint Bayesian estimation of alignment and phylogeny. Syst Biol. 2005:**54**(3):401–418. https://doi.org/10.1080/10635150590947041.

Redelings BD, Suchard MA. Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. BMC

Evol Biol. 2007:**7**(1):40. https://doi.org/10.1186/1471-2148-7-40.

Redelings BD, Suchard MA. Robust inferences from ambiguous alignments. In: Rosenberg MS, editors. Sequence alignment: methods, concepts, and strategies. Berkeley: University of California Press; 2009. p. 209–270.

Rice ES, Green RE. New approaches for genome assembly and scaffolding. Annu Rev Anim Biosci. 2019:**7**(1):17–40. https://doi.org/10.1146/annurev-animal-020518-115344.

Rivas E. Evolutionary models for insertions and deletions in a probabilistic modeling framework. BMC Bioinformatics. 2005:**6**(1):63. https://doi.org/10.1186/1471-2105-6-63.

Rivas E, Eddy SR. Probabilistic phylogenetic inference with insertions and deletions. PLoS Comput Biol. 2008:**4**(9):e1000172. https://doi.org/10.1371/journal.pcbi.1000172.

Rivas E, Eddy SR. Parameterizing sequence alignment with an explicit evolutionary model. BMC Bioinformatics. 2015:**16**(1):406. https://doi.org/10.1186/s12859-015-0832-5.

Rivera MC, Lake JA. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science. 1992:**257**(5066):74–76. https://doi.org/10.1126/science.1621096.

Rokas A, Holland PW. Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol. 2000:**15**(11):454–459. https://doi.org/10.1016/S0169-5347(00)01967-4.

Roos RAC. Huntington's disease: a clinical review. Orphanet J Rare Dis. 2010:**5**(1):40. https://doi.org/10.1186/1750-1172-5-40.

Rosenberg MS. MySSP: non-stationary evolutionary sequence simulation, including indels. Evol Bioinform Online. 2007:**1**:81–83. https://doi.org/10.1177/117693430500100007.

Sandhya S, Rani SS, Pankaj B, Govind MK, Offmann B, Srinivasan N, Sowdhamini R. Length variations amongst protein domain superfamilies and consequences on structure and function. PLoS One. 2009:**4**(3):e4981. https://doi.org/10.1371/journal.pone.0004981.

Sankoff D. Minimal mutation trees of sequences. SIAM J Appl Math. 1975:**28**(1):35–42. https://doi.org/10.1137/0128004.

Satija R, Novák A, Miklós I, Lyngsø R, Hein J. BigFoot: Bayesian alignment and phylogenetic footprinting with MCMC. BMC Evol Biol. 2009:**9**(1):217. https://doi.org/10.1186/1471-2148-9-217.

Satija R, Pachter L, Hein J. Combining statistical alignment and phylogenetic footprinting to detect regulatory elements. Bioinformatics. 2008:**24**(10):1236–1242. https://doi.org/10.1093/bioinformatics/btn104.

Saurabh K, Holland BR, Gibb GC, Penny D. Gaps: an elusive source of phylogenetic information. Syst Biol. 2012:**61**(6):1075–1082. https://doi.org/10.1093/sysbio/sys043.

Savino S, Desmet T, Franceus J. Insertions and deletions in protein evolution and engineering. Biotechnol Adv. 2022:**60**:108010. https://doi.org/10.1016/j.biotechadv.2022.108010.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009:**326**(5956):1112–1115. https://doi.org/10.1126/science.1178534.

Schreiber F, Patricio M, Muffato M, Pignatelli M, Bateman A. TreeFam v9: a new website, more species and orthology-on-the-fly. Nucleic Acids Res. 2014:**42**(D1):D922–D925. https://doi.org/10.1093/nar/gkt1055.

Sehn JK. Insertions and deletions (indels). In: Kulkarni S, Pfeifer J, editors. Clinical genomics. Elsevier; 2015. p. 129–150.

Sela I, Ashkenazy H, Katoh K, Pupko T. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Res. 2015:**43**(W1):W7–W14. https://doi.org/10.1093/nar/gkv318.

Selberg AGA, Gaucher EA, Liberles DA. Ancestral sequence reconstruction: from chemical paleogenetics to maximum likelihood algorithms and beyond. J Mol Evol. 2021:**89**(3):157–164. https://doi.org/10.1007/s00239-021-09993-1.

Seo T-K, Redelings BD, Thorne JL. Correlations between alignment gaps and nucleotide substitution or amino acid replacement. Proc Natl Acad Sci U S A. 2022:**119**(34):e2204435119. https://doi.org/10.1073/pnas.2204435119.

Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, Albertsen M. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods. 2022:**19**(7):823–826. https://doi.org/10.1038/s41592-022-01539-7.

Sfeir A, Symington LS. Microhomology-mediated end joining: a back-up survival mechanism or dedicated pathway? Trends Biochem Sci. 2015:**40**(11):701–714. https://doi.org/10.1016/j.tibs.2015.08.006.

Simmons MP, Müller K, Norton AP. The relative performance of indel-coding methods in simulations. Mol Phylogenet Evol. 2007:**44**(2):724–740. https://doi.org/10.1016/j.ympev.2007.04.001.

Simmons MP, Ochoterena H. Gaps as characters in sequence-based phylogenetic analyses. Syst Biol. 2000:**49**(2):369–381. https://doi.org/10.1093/sysbio/49.2.369.

Som A, Sharma AK, Kumari P. Recombination in Sarbecovirus lineage and mutations/insertions in spike protein are linked to the emergence and adaptation of SARS-CoV-2. Bioinformation. 2022:**18**(10):951–961. https://doi.org/10.6026/97320630018951.

Sonay TB, Koletou M, Wagner A. A survey of tandem repeat instabilities and associated gene expression changes in 35 colorectal cancers. BMC Genomics. 2015:**16**(1):702. https://doi.org/10.1186/s12864-015-1902-9.

Song B, Mott R, Gan X. Recovery of novel association loci in *Arabidopsis thaliana* and *Drosophila melanogaster* through leveraging INDELs association and integrated burden test. PLoS Genet. 2018:**14**(10):e1007699. https://doi.org/10.1371/journal.pgen.1007699.

Spence MA, Kaczmarski JA, Saunders JW, Jackson CJ. Ancestral sequence reconstruction for protein engineers. Curr Opin Struct Biol. 2021:**69**:131–141. https://doi.org/10.1016/j.sbi.2021.04.001.

Spielman SJ, Dawson ET, Wilke CO. Limited utility of residue masking for positive-selection inference. Mol Biol Evol. 2014:**31**(9):2496–2500. https://doi.org/10.1093/molbev/msu183.

Spielman SJ, Wan S, Wilke CO. A comparison of one-rate and two-rate inference frameworks for site-specific dN/dS estimation. Genetics. 2016:**204**(2):499–511. https://doi.org/10.1534/genetics.115.185264.

Steel M, Hein J. Applying the Thorne–Kishino–Felsenstein model to sequence evolution on a star-shaped tree. Appl Math Lett. 2001:**14**(6):679–684. https://doi.org/10.1016/S0893-9659(01)80026-4.

Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN. The human gene mutation database: 2008 update. Genome Med. 2009:**1**(1):13. https://doi.org/10.1186/gm13.

Stoye J, Evers D, Meyer F. Rose: generating sequence families. Bioinformatics. 1998:**14**(2):157–163. https://doi.org/10.1093/bioinformatics/14.2.157.

Studer RA, Dessailly BH, Orengo CA. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. Biochem J. 2013:**449**(3):581–594. https://doi.org/10.1042/BJ20121221.

Suchard MA, Redelings BD. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics. 2006:**22**(16):2047–2048. https://doi.org/10.1093/bioinformatics/btl175.

Susko E, Inagaki Y, Field C, Holder ME, Roger AJ. Testing for differences in rates-across-sites distributions in phylogenetic subtrees. Mol Biol Evol. 2002:**19**(9):1514–1523. https://doi.org/10.1093/oxfordjournals.molbev.a004214.

Suvorov A, Hochuli J, Schrider DR. Accurate inference of tree topologies from multiple sequence alignments using deep learning. Syst Biol. 2020:**69**(2):221–233. https://doi.org/10.1093/sysbio/syz060.

Szalkowski AM, Anisimova M. Graph-based modeling of tandem repeats improves global multiple sequence alignment. Nucleic Acids Res. 2013:**41**(17):e162. https://doi.org/10.1093/nar/gkt628.

Takahashi K, Nei M. Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. Mol Biol Evol. 2000:**17**(8):1251–1258. https://doi.org/10.1093/oxfordjournals.molbev.a026408.

Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 2007:**56**(4):564–577. https://doi.org/10.1080/10635150701472164.

Tan G, Muffato M, Ledergerber C, Herrero J, Goldman N, Gil M, Dessimoz C. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. Syst Biol. 2015:**64**(5):778–791. https://doi.org/10.1093/sysbio/syv033.

Teufel AI, Ritchie AM, Wilke CO, Liberles DA. Using the mutation-selection framework to characterize selection on protein sequences. Genes. 2018:**9**(8):409. https://doi.org/10.3390/genes9080409.

Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994:**22**(22):4673–4680. https://doi.org/10.1093/nar/22.22.4673.

Thompson A, Liebeskind BJ, Scully EJ, Landis MJ. Deep learning and likelihood approaches for viral phylogeography converge on the same answers whether the inference model is right or wrong. Syst Biol. 2024:**73**(1):183–206. https://doi.org/10.1093/sysbio/syad074.

Thorne JL, Kishino H, Felsenstein J. An evolutionary model for maximum likelihood alignment of DNA sequences. J Mol Evol. 1991:**33**(2):114–124. https://doi.org/10.1007/BF02193625.

Thorne JL, Kishino H, Felsenstein J. Inching toward reality: an improved likelihood model of sequence evolution. J Mol Evol. 1992:**34**(1):3–16. https://doi.org/10.1007/BF00163848.

Trost B, Engchuan W, Nguyen CM, Thiruvahindrapuram B, Dolzhenko E, Backstrom I, Mirceta M, Mojarad BA, Yin Y, Dov A, et al. Genome-wide detection of tandem DNA repeats that are expanded in autism. Nature. 2020:**586**(7827):80–86. https://doi.org/10.1038/s41586-020-2579-z.

Trost J, Haag J, Höhler D, Jacob L, Stamatakis A, Boussau B. Simulations of sequence evolution: how (un)realistic they are and why. Mol Biol Evol. 2024:**41**(1):msad277. https://doi.org10.1093/molbev/msad277.

Vaughn JN, Bennetzen JL. Natural insertions in rice commonly form tandem duplications indicative of patch-mediated double-strand break induction and repair. Proc Natl Acad Sci U S A. 2014:**111**(18):6684–6689. https://doi.org/10.1073/pnas.1321854111.

Verbiest MA, Lundström O, Xia F, Baudis M, Sonay TB, Anisimova M. Short tandem repeat mutations regulate gene expression in colorectal cancer. Sci Rep. 2024:**14**(1):3331. https://doi.org/10.1038/s41598-024-53739-0.

Verbiest M, Maksimov M, Jin Y, Anisimova M, Gymrek M, Bilgin Sonay T. Mutation and selection processes regulating short tandem repeats give rise to genetic and phenotypic diversity across species. J Evol Biol. 2022:**36**(2):321–336. https://doi.org/10.1111/jeb.14106.

Vialle RA, Tamuri AU, Goldman N. Alignment modulates ancestral sequence reconstruction accuracy. Mol Biol Evol. 2018:**35**(7):1783–1797. https://doi.org/10.1093/molbev/msy055.

Vingron M, Waterman MS. Sequence alignment and penalty choice. Review of concepts, case studies and implications. J Mol Biol. 1994:**235**(1):1–12. https://doi.org/10.1016/S0022-2836(05)80006-3.

Vishwakarma MK, Kale SM, Sriswathi M, Naresh T, Shasidhar Y, Garg V, Pandey MK, Varshney RK. Genome-wide discovery and deployment of insertions and deletions markers provided greater insights on species, genomes, and sections relationships in the genus *Arachis*. Front Plant Sci. 2017:**8**:290580. https://doi.org/10.3389/fpls.2017.02064.

Vogler AP, DeSalle R. Evolution and phylogenetic information content of the ITS-1 region in the tiger beetle *Cicindela dorsalis*. Mol. Biol. Evol. 1994:**11**(3):393–405. https://doi.org/10.1093/oxfordjournals.molbev.a040121.

Wang Z, Sun J, Gao Y, Xue Y, Zhang Y, Li K, Zhang W, Zhang C, Zu J, Zhang L. Fusang: a framework for phylogenetic tree inference via deep learning. Nucleic Acids Res. 2023:**51**(20):10909–10923. https://doi.org/10.1093/nar/gkad805.

Wang D, Zhou Q, Le L, Fu F, Wang G, Cao F, Yang X. Molecular characterization and genetic diversity of *Ginkgo* (L.) based on insertions and deletions (indel) markers. Plants. 2023:**12**(13):2567. https://doi.org/10.3390/plants12132567.

Wells JN, Feschotte C. A field guide to eukaryotic transposable elements. Annu Rev Genet. 2020:**54**(1):539–561. https://doi.org/10.1146/annurev-genet-040620-022145.

Westesson O, Lunter G, Paten B, Holmes I. Accurate reconstruction of insertion–deletion histories by statistical phylogenetics. PLoS One. 2012:**7**(4):e34572. https://doi.org/10.1371/journal.pone.0034572.

Wheeler WC. Iterative pass optimization of sequence data. Cladistics. 2005a:**19**(3):254–260. https://doi.org/10.1111/j.1096-0031.2003.tb00368.x.

Wheeler WC. Implied alignment: a synapomorphy-based multiple-sequence alignment method and its use in cladogram search. Cladistics. 2005b:**19**:261–268. https://doi.org/10.1111/j.1096-0031.2003.tb00369.x.

Wheeler WC, Gatesy J, DeSalle R. Elision: a method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. Mol Phylogenet Evol. 1995:**4**(1):1–9. https://doi.org/10.1006/mpev.1995.1001.

Wheeler WC, Lucaroni N, Hong L, Crowley LM, Varón A. POY version 5: phylogenetic analysis using dynamic homologies under multiple optimality criteria. Cladistics. 2015:**31**(2):189–196. https://doi.org/10.1111/cla.12083.

Wheeler WC, Washburn A, Crowley LM. PhylogeneticGraph (PhyG) a new phylogenetic graph search and optimization program. Cladistics. 2024:**40**(1):97–105. https://doi.org/10.1111/cla.12560.

Wolf Y, Madej T, Babenko V, Shoemaker B, Panchenko AR. Long-term trends in evolution of indels in protein sequences. BMC Evol Biol. 2007:**7**(1):19. https://doi.org10.1186/1471-2148-7-19.

Wong KM, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. Science. 2008:**319**(5862):473–476. https://doi.org/10.1126/science.1151532.

Wu H-X, Wang Z-X, Zhao Q, Chen D-L, He M-M, Yang L-P, Wang Y-N, Jin Y, Ren C, Luo H-Y, et al. Tumor mutational and indel burden: a systematic pan-cancer evaluation as prognostic biomarkers. Ann Transl Med. 2019:**7**(22):640. https://doi.org/10.21037/atm.2019.10.116.

Wygoda E, Loewenthal G, Moshe A, Alburquerque M, Mayrose I, Pupko T. Statistical framework to determine indel-length distribution. Bioinformatics. 2024:**40**(2):btae043. https://doi.org/10.1093/bioinformatics/btae043.

Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 1997:**13**(5):555–556. https://doi.org/10.1093/bioinformatics/13.5.555.

Yang J, Wang Y, Shen H, Yang W. In silico identification and experimental validation of insertion–deletion polymorphisms in tomato genome. DNA Res. 2014:**21**(4):429–438. https://doi.org/10.1093/dnares/dsu008.

Yang H, Zhong Y, Peng C, Chen J-Q, Tian D. Important role of indels in somatic mutations of human cancer genes. BMC Med Genet. 2010:**11**(1):128. https://doi.org/10.1186/1471-2350-11-128.

Ye Y, Godzik A. Multiple flexible structure alignment using partial order graphs. Bioinformatics. 2005:**21**(10):2362–2369. https://doi.org/10.1093/bioinformatics/bti353.

Zhai Y, Alexandre B-C. A poissonian model of indel rate variation for phylogenetic tree inference. Syst Biol. 2017:**66**(5):698–714. https://doi.org/10.1093/sysbio/syx033.

Zhang Z, Gerstein M. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. Nucleic Acids Res. 2003:**31**(18):5338–5348. https://doi.org/10.1093/nar/gkg745.

Zhou G, Zhang Q, Tan C, Zhang X-Q, Li C. Development of genome-wide InDel markers and their integration with SSR, DArT and SNP markers in single barley map. BMC Genomics. 2015:**16**(1):1–8. https://doi.org/10.1186/s12864-015-2027-x.