



## RESEARCH ARTICLE

## Rapid alignment updating with Extensiphy

Jasper Toscani Field<sup>1</sup>  | A. Jeanine Abrams<sup>2</sup> | John C. Cartee<sup>2</sup> | Emily Jane McTavish<sup>3</sup> <sup>1</sup>Quantitative and Systems Biology Program, School of Natural Sciences, University of California, Merced, CA, USA<sup>2</sup>Division of STD Prevention, National Centers for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, Atlanta, GA, USA<sup>3</sup>Life and Environmental Sciences Department, School of Natural Sciences, University of California, Merced, CA, USA

## Correspondence

Jasper Toscani Field  
Email: jtoscanifield@ucmerced.edu

## Funding information

United States National Science Foundation, Grant/Award Number: 1759846

Handling Editor: Pablo Duchon

## Abstract

1. High-throughput sequencing has become commonplace in evolutionary studies. Large, rapidly collected genomic datasets are used to capture biodiversity and for monitoring global and national scale disease transmission patterns, among many other applications. Updating homologous sequence datasets with new samples is cumbersome, requiring excessive program runtimes and data processing. We describe Extensiphy, a bioinformatics tool to efficiently update multiple sequence alignments with whole-genome short-read data. Extensiphy performs reference based sequence assembly and alignment in one process while maintaining the alignment length of the original alignment. Input data-types for Extensiphy are any multiple sequence alignment in fasta format and whole-genome, short-read fastq sequences.
2. To validate Extensiphy, we compared its results to those produced by two other methods that construct whole-genome scale multiple sequence alignments. We measured our comparisons by analysing program runtimes, base-call accuracy, dataset retention in the presence of missing data and phylogenetic accuracy.
3. We found that Extensiphy rapidly produces high-quality updated sequence alignments while preventing alignment shrinkage due to missing data. Phylogenies estimated from alignments produced by Extensiphy show similar accuracy to other commonly used alignment construction methods.
4. Extensiphy is suitable for updating large sequence alignments and is ideal for studies of biodiversity, ecology and epidemiological monitoring efforts.

## KEYWORDS

evolutionary biology, genomes, monitoring, phylogenetics, sequence alignment, software

## 1 | INTRODUCTION

The development of genomic methods has revolutionized virtually all fields of biology and lead to an abundance of DNA sequence data available to researchers (Goodwin et al., 2016; Mardis, 2017). This genomic data can be used to estimate phylogenies, which

describe the evolutionary relationships of multiple lineages (Chan & Ragan, 2013). Phylogenies have a wide range of applications across ecology and evolutionary biology. Recent developments in genome scale phylogenetics have upended long held beliefs about deep evolutionary history (Dunn et al., 2008, 2015). Phylogenetic estimates are essential frameworks for comparative genetics and

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society

genomics (Dunn et al., 2018; Hardison, 2003; Smith et al., 2020; Soltis & Soltis, 2003). Large-scale phylogenies have long been recognized as a key tool when addressing gaps in knowledge of biodiversity (Drew et al., 2013; Hortal et al., 2015; McTavish et al., 2017; Sánchez-Reyes et al., 2021). Accurate trees provide context for ecologists seeking to understand community assembly and stability, trophic interactions and ecosystem function (Cavender-Bares et al., 2012). From a human health perspective, rapidly updated phylogenies are pivotal to tracing and understanding pathogen outbreaks (Hadfield et al., 2018). With sequencing rates producing more genomic data than ever before, the barrier for studies of ecology, evolution and biodiversity is now the process of organizing and manipulating data prior to estimating phylogenies (Hodcroft et al., 2021).

Adding new data to a phylogeny first requires that the new data to be incorporated into a key underlying data structure, the homologous sequence alignment. Homologous sequence alignments, also known as multiple sequence alignments, capture the shared evolutionary origin of any number of sequences arranged with pairwise awareness of sequence homology (Chenna et al., 2003; Swofford et al., 1996). Alignment as a procedure is the process of finding homology between two or more DNA sequences (Kim et al., 2015; Vasimuddin et al., 2019). The procedure of multiple sequence alignment is computationally challenging, which must be repeated when new data are added to existing alignments (Chenna et al., 2003; Field et al., 2018; Liu et al., 2012; Treangen et al., 2014; Wang & Jiang, 1994). While recent methods have improved the efficiency of aligning datasets of many taxa and long sequences, the continuing expansion of empirical genomic datasets make the necessary data processing cumbersome (Eddy, 2009; Grad et al., 2016; Hadfield et al., 2018; Leebens-Mack et al., 2019; Liu et al., 2012; NCBI, 2020; Nguyen et al., 2015). The National Center for Biotechnology Information (NCBI) pathogen database contains 14,915 *Neisseria gonorrhoeae* samples along with other pathogens with more than 340,000 samples (NCBI, 2020). The task of assembling these genomes, extracting loci-of-interest and aligning the updated datasets, while not intractable, will be formidable and highlights why novel methods for updating genomic datasets are necessary.

An additional problem when updating an existing MSA with large, rapidly growing genomic databases is the probability of introducing missing data or incomplete data. 'Missing data' may be due to biological reality, such as the evolutionary process of insertions and deletions, or can be a bioinformatic artefact such as low sequencing coverage or read quality in some genomic regions. It has been demonstrated that biological reality and bioinformatic artefacts can interact in driving patterns of missing data across the genome, as rapidly evolving regions are more likely to have reads fail to map, resulting in the appearance of missing data (Huang & Knowles, 2016). Researchers have studied the effect of missing data in evolutionary analyses for decades (Driskell et al., 2004; Huang & Knowles, 2016; Lemmon et al., 2009; Molloy & Warnow, 2018; Wilkinson, 1995; Xi et al., 2016). As such, the effect of missing data on evolutionary analyses has been hotly debated (Capella-Gutiérrez et al., 2009; Castresana, 2000; Huang & Knowles, 2016; Lemmon et al., 2009; Molloy & Warnow, 2018;

Talavera & Castresana, 2007; Treangen et al., 2014; Xi et al., 2016). Some studies laud the effects of removing alignment regions with high proportions of missing data as improving phylogenetic estimations (Capella-Gutiérrez et al., 2009; Castresana, 2000; Criscuolo & Gribaldo, 2010; Talavera & Castresana, 2007; Treangen et al., 2014). Methods of alignment trimming are based on cutoffs of the number of taxa which are missing a particular locus, removing the locus for all taxa (Capella-Gutiérrez et al., 2009; Castresana, 2000; Criscuolo & Gribaldo, 2010; Treangen et al., 2014). Alignment trimming programs often include strict default settings but allow for user specified inputs in order to tailor datasets for the question at hand (Castresana, 2000; Treangen et al., 2014). In general, missing data tend to be less problematic for phylogenetic estimation when it is randomly distributed across the phylogeny, and more problematic when there is a correlation between phylogeny and missingness (Huang & Knowles, 2016; Lemmon et al., 2009; Streicher et al., 2016). Wholesale removal of these regions from analyses can therefore bias estimates of evolutionary rate, affecting branch lengths, topology and bootstrap support (Huang & Knowles, 2016; Streicher et al., 2016). This bias can shorten branch lengths if predominantly variable regions are removed (Huang & Knowles, 2016), or lengthen branch lengths if invariant characters are dropped from the analysis (Felsenstein, 1992; Leaché et al., 2015; Lewis, 2001). Moreover, trimming alignment regions with high proportions of missing data can preclude potentially informative downstream analyses. Analyses of sequence selection and adaptation, often assessed using ratios of synonymous and non-synonymous mutations between taxa, also rely on multiple sequence alignments as statements of orthology (Briggs et al., 2009; Huerta-Cepas et al., 2016; Rocha et al., 2006). Studies in various biological fields describe removing missing data from selection analyses, either by the removal of any missing data or by cutoff values for the number of taxa with missing data at a site (Hodgins et al., 2016; Murolo & Romanazzi, 2015; Williamson et al., 2014). While these methods may be appropriate for within-locus missing data, the automated removal of sequences flanking missing data sites could bias investigations of adaptation. Simply put, if a locus has been removed from an alignment, no further analyses may be performed using it once new data are added to the alignment.

To address the problem of rapidly updating sequence alignments with unprocessed whole-genome sequence data while maintaining input alignment length, we introduce Extensiphy. Extensiphy uses efficient reference based sequence assembly to add homologous loci to existing multiple sequence alignments. Extensiphy performs sequence assembly, locus extraction and alignment of new data to the original dataset in a single process. The intended utility of Extensiphy is to incorporate new un-assembled sequence (e.g. raw reads) data into existing alignments for phylogenetic analyses. Here we describe the Extensiphy method and compare its speed and accuracy to a standard de novo assembly workflow and a commonly used reference alignment method for calling single nucleotide polymorphisms (SNPs); Snippy (Bankevich et al., 2012; Seemann, 2021; Treangen et al., 2014). We investigate Extensiphy's performance compared to these other methods by running each workflow on an empirical *N. gonorrhoeae* dataset as well as a simulated sequence dataset. Each method was

assessed using metrics of program runtime, dataset retention, base-call comparison and phylogenetic distances.

## 2 | MATERIALS AND METHODS

### 2.1 | Overview of Extensiphy

A standard run of Extensiphy accepts a multiple sequence alignment (MSA) and any number of high-throughput read files for newly sequenced samples. The MSA may contain any number of concatenated loci, here referring to genes or lengths of DNA sequences appended together. Extensiphy can accept both paired-end and single-end high-throughput short-read files. An arbitrary reference sequence is chosen from the taxa in the alignment for read alignment. After a reference is selected, all reads are aligned to the concatenated reference sequence. Following read alignment, nucleotides are called to create a consensus sequence that is homologous to all the sequences in the original MSA. All new consensus sequences are added to the multiple sequence alignment, completing assembly and sequence alignment as part of the same process. Finally, if the user opts to automate phylogeny estimation, a phylogeny based on the newly created and extended sequence alignment is estimated using a maximum-likelihood framework. A default run of Extensiphy is visually described in Figure 1. Alternative options for Extensiphy parameters and functionality are described in the following sections.

### 2.2 | Description of Extensiphy

#### 2.2.1 | File inputs, reference selection and read alignment

Extensiphy takes as input a single, concatenated MSA file or any number of unconcatenated single-locus MSA files with identical

taxon labels. If multiple single-locus files are chosen, sequences corresponding to each taxon are concatenated into a single sequence and all sequences are combined into a single multiple sequence alignment containing all sequences for all taxa. Reference selection by default selects the first taxon in the alignment to use as the reference. The user may also specify the selection of a specific reference. Read alignment is performed by BWA-MEM2 (Vasimuddin et al., 2019). A reference index is constructed for the chosen reference sequence and paired-end or single-end reads are aligned. The output of read alignment to the reference sequence is in the sequence alignment mapping (SAM) file format and no un-aligned sequences are output. The number of threads specified for each parallel run of Extensiphy are allocated to BWA-MEM2. All other settings are left as default.

#### 2.2.2 | Variant calling and consensus sequence construction

Following read alignment, SAM files are passed to programs for variant calling. Reference sequence indexing is performed by Samtools Faidx (Li et al., 2009). SAM files are converted to binary alignment mapping (BAM) files by Samtools View (Li et al., 2009). Once SAM to BAM conversion is complete, BAM file organizing is performed by Samtools Index (Li et al., 2009). Variant nucleotide calling is performed by Mpileup from the Bcftools suite (Li et al., 2009). Mpileup produces a Variant Call File (VCF; Danecek et al., 2011). Following VCF production, insertions and deletions are removed as these events usually prevent shared synteny between aligned sequences. The cleaned VCF is then converted to a fastq format file by vcftools.pl and then to a fasta format file by seqtk (Danecek et al., 2011; Gordon & Hannon, 2021; Heng, 2021). Finally, gaps in the original reference sequence are added to the new consensus sequence to preserve synteny. The fully constructed consensus sequence is then appended to the updated alignment file.

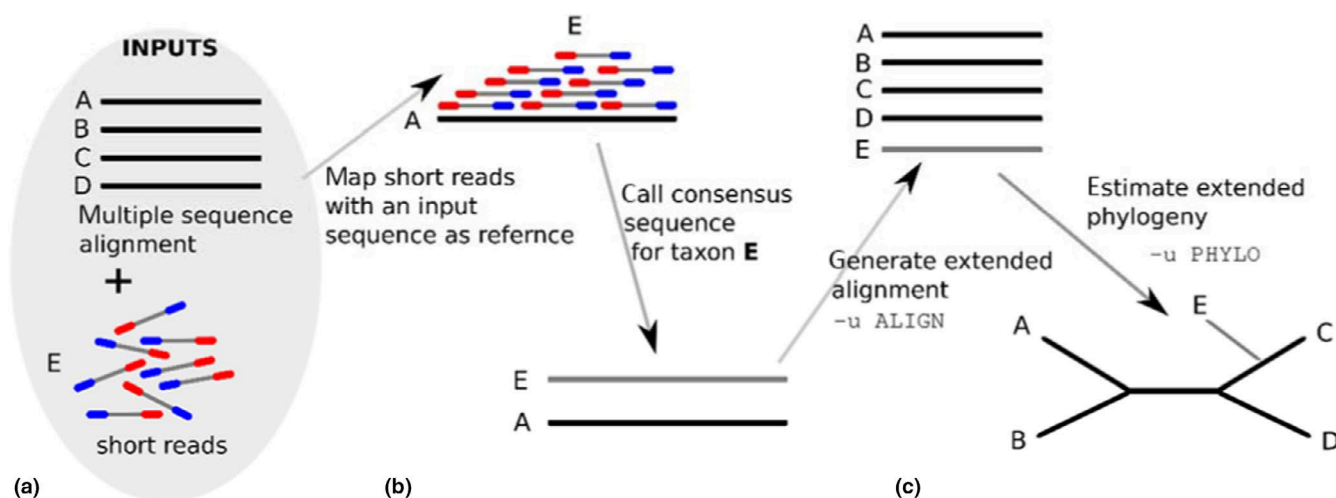


FIGURE 1 Default workflow of Extensiphy. (a) Input an alignment file and new raw reads. (b) Align reads to reference and call the consensus sequence. (c) Output updated alignment and tree files

## 2.2.3 | Phylogenetic estimation and output settings

If selected, phylogenetic estimations are performed using RAXML with the GTRGAMMA model of nucleotide substitution (Stamatakis, 2014). Extensiphy can perform a de novo phylogenetic estimation or, when updating a extant phylogeny, Extensiphy may use a tree produced by the original MSA as a starting tree to improve the search of tree space. The purpose of the starting tree is to build on the evolutionary estimations of the original phylogeny. If the input was multiple single-locus alignment files, the user may also choose to split the final, updated alignment back into single-locus multiple sequence alignment files, for example, for the estimation of gene trees or a species tree by way of summary methods (Yin et al., 2019). RAXML using the GTRGAMMA model is the only option for phylogenetic estimation currently implemented within Extensiphy. However, as a default execution of Extensiphy outputs an updated alignment, users are free to apply any available method of phylogenetic estimation, by using the output alignment as the input for an alternative method. For example, when updating multiple single-locus alignment files a more appropriate method of estimation may be available for inferring a species tree from single-locus alignments. While Extensiphy does not automate running a placement algorithm, the updated alignment and original phylogeny can be easily used as inputs software to place the new sequences without updating the input relationships (Matsen et al., 2010). Due to Extensiphy's focus on adding large amounts of new sequence data to existing alignments, users may specify removing intermediate output files used during consensus sequence production to reduce unnecessary on-disk storage. Phylogenetic inference may be skipped altogether if only an updated sequence alignment is desired.

## 2.3 | Program comparison

### 2.3.1 | Program comparison overview

Extensiphy produces an alignment of homologous sequence data. In order to assess Extensiphy's ability to produce useful data, we compared Extensiphy's alignment to similar alignments produced by contemporary programs and methodologies. In addition to comparing the alignments, we also compared phylogenies produced from alignments, and overall program runtimes. Based on previous literature, we identified two dominant approaches for constructing alignments with a focus on outputs used for evolutionary analyses: de novo sequence assembly followed by core genome alignment and read alignment to reference genome followed by SNP calling (Bush et al., 2020; Castresana, 2000; Seemann, 2021; Treangen et al., 2014). We chose the pipeline Snippy to represent read alignment and variant calling methodologies due to its results in program runtime and SNP calling accuracy (Bush et al., 2020). Following light quality trimming with BBDUK (Bushnell, 2021), we chose to perform de novo sequence assembly with SPAdes and homologous locus selection with ParSNP (Bankevich et al., 2021; Treangen et al., 2014). SPAdes has been used to assemble genomic sequences in

numerous studies for a variety of subject organisms. ParSNP is routinely cited in studies involving evolutionary analyses with topics on the microbial tree of life, the evolution of antibiotic resistance in *Staphylococcus aureus* and genomic analysis of antibiotic susceptibility in *N. gonorrhoeae* (Chen et al., 2020; Gernert et al., 2020; Shakya et al., 2020).

We ran each of these approaches on a simulated dataset and an empirical dataset and assessed the outputs. The simulated dataset was used to test all aspects of interest; program runtime, base-call accuracy, dataset retention and phylogenetic accuracy. The empirical dataset was used to test program runtime and the resulting alignments and phylogenies produced by each method were compared to each other to note discrepancies. The comparison software was primarily written in Bash shell scripts and Python, and these scripts as well as the configuration files for Tree to Reads are shared on GitHub at [https://github.com/jtfield/phylo\\_comparison](https://github.com/jtfield/phylo_comparison). There are two versions of the code, one for analysing each simulated and empirical sequence data. The empirical data comparison software requires whole-genome short-read sequences. The software for analysing simulated data required the same input parameters with the addition of the phylogeny and genomes that were used to simulate the raw read sequences. Details on configuring the comparison software are available in the manual packaged with the software.

### 2.3.2 | Datasets

To construct our simulated high-throughput dataset with a known phylogenetic topology, we used TreeToReads (McTavish, Pettengill, et al., 2017). TreeToReads takes as input a phylogeny, evolutionary model parameters and a reference sequence that serves as the template for simulating all additional sequences. In order to generate an input phylogeny for simulation, we obtained 209 *N. gonorrhoeae* raw read files in fastq format from the CDC (Centers for Disease Control and Prevention, USA) used in a 2016 study of the evolutionary relationships of antibiotic resistant *N. gonorrhoeae* (Grad et al., 2016). We replaced all isolate names with random identifiers before phylogenetic estimation. The resulting phylogeny was used as the input phylogeny for TreeToReads. We used a 51,924 bp segment of a complete *N. gonorrhoeae* genome (GenBank: NC\_002946.2) as the reference sequence. The NC\_002946.2 sample was also used as the reference in all instances of reference-based read alignment when processing the empirical dataset. To introduce sequence variation, 3,000 variant nucleotides were uniformly distributed throughout the reference genome and reads of 100 nucleotides were generated at an average of 20 reads per site. To simulate sequences and reads, we used the evolutionary rate model estimated by RAXML from the 2016 study isolates (Rambaut & Grassly, 1997). The nucleotide rate matrix of was: 1.039821, 5.116539, 0.339204, 0.910812, 5.291090 and 1.000000 with the default rate variation of 0.0200. Mutation cluster grouping was enabled with 25% variable site clustering. Sequence fragment size was set to 320 nucleotides and given a standard deviation of 50 nucleotides. We used the default Illumina sequencing error model packaged with ART (Huang et al., 2012).

The outputs of TreeToReads include simulated genome sequences in fasta format and raw read sequences for each simulated taxon. Our empirical dataset was comprised of 1,237 *N. gonorrhoeae* SRA files in fastq format collected from GenBank. Samples were chosen semi-randomly as the first 1,237 SRA numbers found on NCBI Pathogen Detection database under *Neisseria* (NCBI, 2020). Fourteen isolates were identified as *N. meningitidis* and were removed from subsequent analyses. The final empirical dataset consisted of 1,223 samples.

### 2.3.3 | De novo sequence assembly and selection of loci

During the de novo assembly and automated locus selection pipeline, for both the empirical and simulated datasets, bases were trimmed from the raw reads with a quality score of 10 or below. We also removed any sequencing adapters included in the BBDUK default adapters file (Bushnell, 2021). De novo sequence assembly was performed on the trimmed read files to construct contigs for all taxa in the dataset. De novo sequence assembly was performed by SPAdes using default parameters with the exception of additional computing cores (Bankevich et al., 2021). Following assembly, the core genome for all assembled sequences was selected using ParSNP (Treangen et al., 2014). Core genomes are defined as sets of orthologous sequences that are conserved in all included taxa (Hodgins et al., 2016). ParSNP identifies core genomes using a used maximal unique matches between sequences to capture conserved blocks of sequences in highly similar sets of genomes. Regions with missing data are not included in the final core genome, resulting in separate locus alignments. The selected loci were concatenated into a single alignment while the separate locus alignments were retained for downstream base-call analyses. While ParSNP includes options to alter the sequence distance between acceptable matches used for identifying core genome sequences, all options were left as defaults for our analyses.

### 2.3.4 | Read alignment and SNP calling with Snippy

For both the empirical and simulated datasets, Snippy was run using the chosen reference sequence and the raw reads as inputs. Snippy aligned reads to the reference and replaced reference nucleotides with taxon-specific variants where appropriate. The output of the Snippy runs was alignments with sequence lengths matching the reference sequence. The empirical dataset used a contiguous *N. gonorrhoeae* genome sequence as a reference while the simulated dataset used the sequence input into TreeToReads for sequence simulation.

### 2.3.5 | Read alignment and SNP calling with Extensiphy

In order to create an input alignment for use with Extensiphy, we took the assembled genomes for four random taxa and assembled them in

the same manner as the de novo assembly stage described above. We created a core genome alignment for these four taxa and the selected reference sequence using ParSNP (Treangen et al., 2014). This small set of taxa produced a set of loci that were influenced by the missing data found in the five included taxa. The homologous loci of this smaller dataset were concatenated and used as the input alignment for Extensiphy, along with raw read sequences corresponding to the rest of the taxa. Extensiphy processed the concatenated alignment, raw read input files and produced an updated multiple sequence alignment and phylogeny based on the alignment. Once phylogenetic estimation was complete, the concatenated sequence alignment was split into individual locus alignments in preparation for base-call comparisons.

### 2.3.6 | Phylogenetic analysis

For all datasets, phylogenetic estimation was performed on the concatenated alignment using RAxML to produce a maximum-likelihood topology and a consensus topology based on 100 bootstrap replicates (Stamatakis, 2014). We used the GTRGAMMA model for all estimations as this model is the most flexible maximum-likelihood model, and the only one available in RAxML.

## 2.4 | Program output comparisons

### 2.4.1 | Program output comparison overview

We assessed each methodology using three metrics: program runtime, base-call accuracy and phylogenetic accuracy. The methods of measuring program runtime were identical regardless of the dataset. We assessed individual time to assemble each single sequence and the total time for a program to assemble a complete alignment. The time required for phylogenetic estimation was not included for any program. Base-call comparisons, when using the simulated dataset, benefit from comparing each program outputs to the original TreeToReads sequences used to simulate the input data for each program. By using the original TreeToReads sequences, we collected an accurate description of which nucleotides were correctly and incorrectly called. The true base-calls of any empirical sequence are unknown. With this limitation in mind, we compared the sequence outputs of each program to their counterparts from each other program when assessing sequences produced from the empirical dataset. We assessed base-calls pairwise from any locus present in the output of any two programs. This conservative comparison was necessary due to the variation in the length of the sequences output by each program. Consequently, each sequence comparison was limited to the length of the shortest sequence. Phylogenies produced from the simulated dataset were compared to the original topology used by TreeToReads for sequence simulation. For the empirical dataset, the phylogeny produced by each program was compared to each other program's phylogeny. We compared majority-rule consensus phylogenies on bootstrapped data for all comparisons to account for stochastic variation in inferences of very short branches.

## 2.4.2 | Program runtime comparisons

We defined program runtime as two values: the time taken to assemble and output the sequence associated with a single taxon and the total program runtime for assembling all taxon sequences and outputting a complete sequence alignment. All three programs reported the time required for individual sequence alignment and assembly. The total program runtimes to produce a complete alignment were recorded.

## 2.4.3 | Program base-call comparisons

For simulated dataset base-call comparisons, each taxon's sequences were aligned to the original genomes produced by TreeToReads. Extensiphy and de novo assembled sequences which were separate loci for each taxon. Snippy sequences, being duplicates of the reference sequence with variant nucleotides inserted, were the same length as the reference sequence. A base-call comparison was made once two sequences were aligned by noting which nucleotides in one sequence were identical to the paired sequence produced from the other program. Identical nucleotides, non-identical nucleotides, non-identical degenerate nucleotides and gaps within the sequences were counted and summed for each locus. The lengths of all loci were also recorded for Extensiphy and the de novo pipeline. Additional metrics collected from the simulated data analyses were the total number of bases analysed, the per-base miscall and missing data rate for each program and, when comparing Extensiphy and de novo assembled sequences, the discrepancy in the length between the sequences output each program and the sequences produced by TreeToReads. For empirical dataset base-call comparisons, each taxon's sequences were aligned to the sequences produced by both other programs. Additional metrics collected from the empirical data analyses were the total number of bases analysed, the per-base disagreement between each sequence and, when comparing Extensiphy and de novo assembled sequences, the discrepancy in the length of the compared loci.

## 2.4.4 | Phylogenetic comparisons

Phylogenies estimated from each program's alignment were compared using the Robinson–Foulds (RF) distance calculations, the symmetric

distance of partitions between two phylogenies, using the Dendropy Python library (Robinson & Foulds, 1981; Sukumaran & Holder, 2010). All RF distances were calculated as unweighted, expressing only the symmetric differences in branches between topologies.

# 3 | RESULTS

## 3.1 | Simulated dataset results

### 3.1.1 | Runtime

Using Extensiphy, individual sequences were assembled at a mean rate of 4 s per sequence and the overall program runtime was completed in 6 min and 45 s (Table 1). De novo pipeline runtimes were a mean of 8 s per individual sequence and a complete program runtime of 21 min. Snippy's mean individual sequence assembly time was 3 s per sequence and a complete program runtime of 10 min and 28 s.

### 3.1.2 | Alignment length

Extensiphy returned 209 sequences at 51,157 nucleotides each for a total of 10,691,913 nucleotides in the final alignment, including the reference sequence (Table 1). The de novo pipeline returned 209 sequences at 50,245 nucleotides for a total of 10,500,766 nucleotides. Snippy returned 209 full-length sequences at the same 51,191 nucleotide length as the simulated reference sequences as well as a 'core sites' alignment with 1,030 nucleotides per taxon. The full length alignment included 10,698,919 nucleotides excluding the reference sequence.

### 3.1.3 | Alignment accuracy

Extensiphy's sequences produced the lowest miscall rate at 15 nucleotides while the de novo pipeline's alignment contained 21 miscalled nucleotides (Table 1). Snippy produced an alignment with 359 miscalled nucleotides. Supplementary Table 1 contains more descriptive statistics from the simulated dataset base-call comparison of the three programs.

**TABLE 1** Simulated data comparison statistics. Results of comparison pipeline output after processing 209 taxa sequences. m, minutes; s, seconds

Comparison metrics	Extensiphy	De novo assembly	Snippy
Total program runtime	6 m 45 s	21 m	10 m 28 s
Individual sequence runtime	4 s	8 s	4 s
Total miscalled bases	15	21	359
Total bases per taxon	51,157	50,245	51,191
Total bases analysed	10,691,913	10,500,766	10,698,919
RF distance to true tree	56	55	98

### 3.1.4 | Missing data

Extensiphy returned 1,001 total gaps or degenerate nucleotides in the final alignment based on simulated data (Table S1). Snippy returned 163,545 gaps or degenerate nucleotides. The de novo pipeline's alignment contained no gaps or degenerate nucleotides.

### 3.1.5 | Phylogenetic accuracy

Extensiphy produced a phylogeny with an RF distance to the true topology of 56 while the de novo pipeline's phylogeny received an RF distance of 55 and Snippy produced a phylogeny with an RF distance of 98 (Table 1).

## 3.2 | Empirical dataset results

### 3.2.1 | Runtime

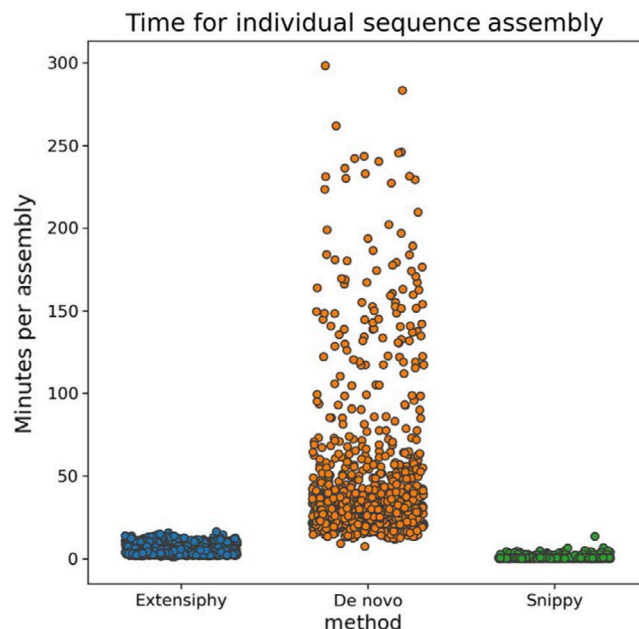
When processing and analysing data from the empirical dataset, Extensiphy produced consensus sequences in a mean time of slightly over 6 min and produced a complete alignment in 38 hr (Figure 2; Table 2). The de novo pipeline assembled sequences in a mean time of 41 min and produced a complete alignment in 236 hr. Snippy produced individual sequences in a mean time of 41 s and produced a complete alignment in 18 hr.

### 3.2.2 | Alignment length

Individual sequences produced by Extensiphy were all of 1,859,910 nucleotides in length for a total of  $2.293 \times 10^9$  nucleotides in the final alignment (Table 3). The Extensiphy alignment was composed of 317 loci with a mean length of 5,868 nucleotides and a range of lengths between 682 and 40,798 nucleotides (Figure 3). The de novo pipeline returned individual sequences of 751,033 nucleotides and a total of  $9.215 \times 10^8$  nucleotides in the final alignment. The de novo pipeline alignment was composed of 522 loci with a mean length of 1,465 and a range of lengths between 688 and 5,913 nucleotides. Individual sequences produced by Snippy were 2,180,847 nucleotides in length for a total of  $2.732 \times 10^9$  nucleotides in the final alignment. Locus values were not reported for Snippy as Snippy operates using whole-genome inputs and outputs.

### 3.2.3 | Alignment accuracy

We assessed empirical base-calls for the outputs of all three programs against each other as true base-calls cannot be described with certainty for empirical sequence data (Table S2). The Extensiphy-de novo pipeline comparison contained 490 differing nucleotides from 31,909,017 analysed sites between both alignments. The Extensiphy-Snippy comparison produced 27,778 differing nucleotides from



**FIGURE 2** The time required by each method to assemble all sequences associated with each taxon in the empirical dataset

338,286,158 analysed sites between both alignments. The comparison of Snippy and the de novo pipeline alignments contained 142 differing nucleotides from 31,974,892 sites analysed between both alignments.

### 3.2.4 | Missing data

We assessed empirical missing data in the same manner as empirical base-calls, that is, by comparing the outputs of each program against each other. The Extensiphy-de novo pipeline comparison contained 81,035 differing gaps or degenerate nucleotides from 31,909,017 analysed sites between both alignments (Table S2). The Extensiphy-Snippy comparison produced 1,857,035 differing gaps or degenerate nucleotides from 338,286,158 analysed sites between both alignments. The comparison of Snippy and the de novo pipeline alignments contained 105,875 differing gaps or degenerate nucleotides from 31,974,892 sites analysed between both alignments. When analysing the complete alignment for each program, the alignment produced by Extensiphy contained 4,891,739 gaps and degenerate nucleotides (Table 3). The de novo pipeline alignment contained 3,469,861 gaps and degenerate nucleotides and the Snippy alignment contained 224,835,516 gaps and degenerate nucleotides.

### 3.2.5 | Phylogenetic accuracy

When analysing the RF distances between the phylogenies produced by each program, the Extensiphy-de novo pipeline comparison produced an RF distance of 687 and the Extensiphy-Snippy

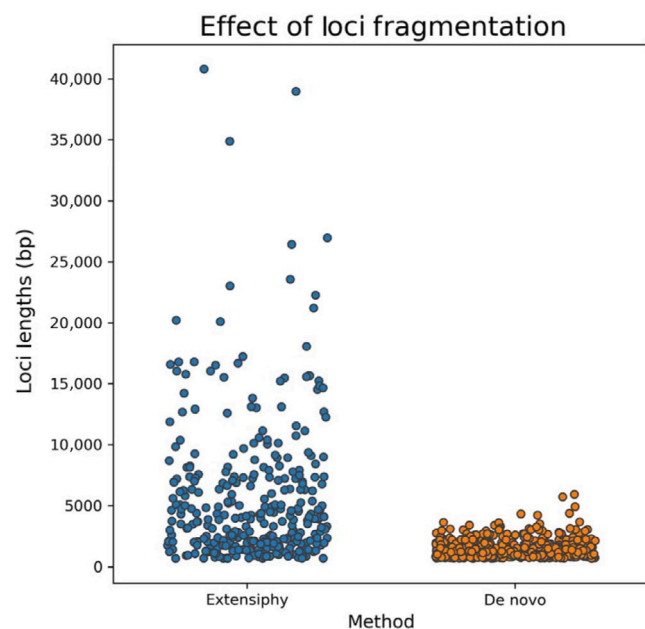


**TABLE 2** Empirical data runtime statistics. Results of program runtimes after processing 1,223 taxa sequences. h, hours; m, minutes; s, seconds

Comparison metrics	Extensiphy	De novo assembly	Snippy
Total program runtime	38 h	236 h	18 h
Average individual sequence runtime	6 m 21 s	41 m	41 s

**TABLE 3** Empirical data alignment statistics. Nucleotide and locus metrics for the alignments containing 1,223 sequences produced by each program. A '—' symbol indicates the value is not applicable

Comparison metrics	Extensiphy	De novo assembly	Snippy
Total bases per alignment	2,293,269,030	921,517,491	2,732,911,282
Total gaps or degenerate bases	4,891,739	3,469,861	224,835,516
Average locus length	5,868	1,465	—
Loci output per program	317	522	—



**FIGURE 3** Empirical dataset locus lengths returned by Extensiphy and the de novo assembly pipeline

comparison produced an RF distance of 749 (Table 4). The de novo pipeline-Snippy comparison produced an RF distance of 676.

## 4 | DISCUSSION

Sequencing efforts are expanding for the collection of genomic data (Goodwin et al., 2016; Hodcroft et al., 2021; Mardis, 2017). Current methods for incorporating new data into sequence alignments exist but are inadequate for whole-genome datasets with thousands of taxa (Eddy, 2009; Nguyen et al., 2015). While combining new and previously analysed data during de novo alignment construction is a routinely performed workflow, this process can result in alignment trimming that can remove potentially useful data from a dataset (Huang & Knowles, 2016). To address issues of expanding existing sequence alignments, we introduced the Extensiphy program and

**TABLE 4** Empirical data phylogeny RF distances. Unweighted Robinson–Foulds distances between phylogenies produced by each program. A '—' symbol indicates the value is not applicable

Comparison metrics	Extensiphy	De novo assembly	Snippy
Extensiphy	—	687	749
De novo assembly	687	—	676
Snippy	749	676	—

assessed its outputs to two workflows with comparable outputs. Our results show that Extensiphy balances between data retention, runtime efficiency and applicability to genomic datasets. Extensiphy returned alignments with sequence lengths matching those of the input alignment and containing a lower proportion of degenerate or gap sites than other methods. Extensiphy accommodated and returned an alignment with sequences of lengths comprising over 90% of the *N. gonorrhoeae* genome. All sequences were assembled in competitive times compared to other analysed methodologies. If the starting point of a study is an existing concatenated alignment or set of alignments for the same taxa and a set of whole-genome short-read data and the goal is to rapidly add the new data to the alignment, Extensiphy will produce the desired results. Additionally, we argue that the analyses of both the simulated and empirical datasets demonstrate that Extensiphy performs equally well when updating alignments with any number of loci and inputs of either separate alignments or a single, pre-concatenated alignment. While these two features are simple in terms of modern bioinformatics tools, their presence expands the scope of studies for which Extensiphy may be appropriate. By accommodating any number of loci, Extensiphy is applicable to any scale of project, from inquiries with a single or a few loci to full-scale epidemiological monitoring efforts (Grad et al., 2016; Hadfield et al., 2018; Hodcroft et al., 2021). By accepting either individual locus alignments or a concatenated alignment, Extensiphy does not constrain the user to a specific method of phylogenetic estimation.

Extensiphy is designed to integrate new genomic data with existing datasets. The approach targets computational effort to



regions which are homologous to existing data. This removes the computationally taxing requirement of a downstream multiple sequence alignment step, as the new reads are aligned to a sequence already included in the alignment. Extensiphy also packages a maximum-likelihood phylogenetic estimation method for streamlined results. While Extensiphy and Snippy share similar approaches to sequence construction, Extensiphy produces a homologous sequence alignment as opposed to genome-length sequences which require additional processing to identify and isolate loci-of-interest. Extensiphy assembles new loci directly aligned to existing loci, as opposed to a reference genome. Extensiphy does not require a full reference genome, and can be applied to integrating sequences from whole-genome data into even single-locus datasets. These few or single-locus datasets form the phylogenetic backbone of our understanding of many taxa.

As part of this framework, Extensiphy also allows for the selection of a reference sequence already found in an existing alignment. This provides an opportunity to assess the role of choice of reference sequence in consensus sequence inference. While reference-based read alignment is an excellent flexible method for many studies, the choice of reference sequence can inherently bias downstream analyses (Brandt et al., 2015; Günther & Nettelblad, 2019). Reference bias is a well-known potential influence on sequence structure during read alignment based on the structure of the reference (Günther & Nettelblad, 2019; Ros-Freixedes et al., 2018). The extent to which reference bias affects phylogenetic estimation is still ambiguous. Extensiphy paired with the methodologies of sequence and phylogenetic comparison we describe in this study offer an excellent opportunity to repeatedly measure the effects of constructing alignments based on diverse reference sequences. By running the same analyses using different references with known phylogenetic relationships to each other, it is straightforward to use Extensiphy to assess if this bias is playing a role in one's own dataset.

Acknowledging and addressing missing data are key issues in modern phylogenomics. Current research argues for a case-by-case strategy on including or excluding missing data (Huang & Knowles, 2016; Streicher et al., 2016). The distribution of missing data throughout an alignment influences such decisions (Lemmon et al., 2009). Assuming a relatively even distribution of missing data, alignment trimming may not be necessary and such trimming could remove valuable variant nucleotides from future analyses. In the presence of an uneven distribution of missing data, perhaps due to sequencing bias, a study could benefit from judicious locus removal (Streicher et al., 2016). Extensiphy finds an 'middle ground' in respect to retaining full loci-of-interest while introducing a minimum of missing data. Using Extensiphy, all input loci are maintained while updating an alignment, preventing loci from fragmenting into smaller sequence segments as seen when using ParSNP in the de novo pipeline. Moreover, a smaller percentage of missing data was found in the Extensiphy alignment compared to the alignment produced by Snippy. While the Snippy alignment did contain more sites, expressed as the full length of the reference sequence for each taxon, the difference in size between the Snippy alignment and the Extensiphy

alignment is modest compared to the amount of missing data found in the Snippy alignment. Such a percentage of missing data could affect inferred phylogenies by biasing branch lengths, potentially misleading conclusions based on those phylogenies. Extensiphy rapidly returns an updated alignment while minimizing missing data and enabling researchers to make decisions on the inclusion or excision of loci. Ultimately, all three methods tested here produced accurate estimates and useful alignments and the choice of application of any of the approaches described here depends on the researchers' goal.

## 5 | CONCLUSIONS

Updating a multiple sequence alignment previously required trade-offs of program runtime, reference sequence availability and dataset trimming and fragmentation. We have introduced Extensiphy, a program that updates alignments of loci with new data, and compared it to two popular alternative methods. Extensiphy is applicable to any project with a starting alignment and new whole-genome short-read data. Alignments may be concatenated or separate single-locus alignments. Extensiphy offers an efficient and flexible solution to any study producing high volumes of whole-genome data, particularly for disease monitoring purposes. Projects where maintaining locus length and preventing alignment trimming due to missing data are important will find Extensiphy particularly useful. Extensiphy produces updated alignments suitable for multiple methods of phylogenetic estimation and base-call accuracy comparable to standard methods in the field of bioinformatics. Updating sequence alignments with Extensiphy removes the burden of data processing from the researcher and enables them to focus on purpose and applications of their research.

## ACKNOWLEDGEMENTS

Research was supported by the grant 'Cultivating a sustainable Open Tree of Life', NSF ABI No. 1759846. Computer time was provided by the Multi-Environment Research Computer for Exploration and Discovery (MERCED) cluster from the University of California, Merced (UCM), supported by the NSF Grant No. ACI-1429783. J.T.F. was supported by the NSF NRT Grant DGE-1633722. We appreciate helpful feedback from Dr. Chris Amemiya, Dr Gordon Bennett, Dr Mark Sistrom, Dr Siavash Mirarab, Dr Jessica Blois and the members of the UC Merced Blois-McTavish Lab Group. The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of the CDC.

## CONFLICT OF INTEREST

None declared.

## AUTHORS' CONTRIBUTIONS

J.T.F., A.J.A., J.C.C. and E.J.M. designed Extensiphy; J.T.F. and E.J.M. programmed Extensiphy; J.T.F. performed all data collection and comparisons; J.T.F., A.J.A., J.C.C. and E.J.M. wrote and edited the manuscript.

## PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1111/2041-210X.13790>.

## DATA AVAILABILITY STATEMENT

Extensiphy is open source software utilizing software written by other developers. The Extensiphy pipeline itself is available on Github <https://github.com/McTavishLab/extensiphy> and on Zenodo <https://doi.org/10.5281/zenodo.5770686> (Field, 2021b). The comparison pipelines are also open source software pipelines and are available on Github [https://github.com/jtfield/phylo\\_comparison](https://github.com/jtfield/phylo_comparison) and on Zenodo <https://doi.org/10.5281/zenodo.5770698> (Field, 2021c). All accession numbers for samples and alignments, as well as the simulated data files used in this study are publicly available on Dryad Digital Repository <https://doi.org/10.6071/M38T0T> (Field, 2021a).

## ORCID

Jasper Toscani Field  <https://orcid.org/0000-0002-6457-4359>

Emily Jane McTavish  <https://orcid.org/0000-0001-9766-5727>

## REFERENCES

- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5), 455–477. <https://doi.org/10.1089/cmb.2012.0021>
- Brandt, D. Y., Aguiar, V. R., Bitarello, B. D., Nunes, K., Goudet, J., & Meyer, D. (2015). Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data. *G3: Genes, Genomes, Genetics*, 5(5), 931–941.
- Briggs, A. W., Good, J. M., Green, R. E., Krause, J., Maricic, T., Stenzel, U., Lalueza-Fox, C., Rudan, P., Brajković, D., Kučan, Ž., Gusic, I., Schmitz, R., Doronichev, V. B., Golovanova, L. V., De La Rasilla, M., Fortea, J., Rosas, A., & Pääbo, S. (2009). Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science*, 325(5938), 318–321.
- Bush, S. J., Foster, D., Eyre, D. W., Clark, E. L., De Maio, N., Shaw, L. P., Stoesser, N., Peto, T. E. A., Crook, D. W., & Walker, A. S. (2020). Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines. *GigaScience*, 9(2). <https://doi.org/10.1093/gigascience/giaa007>
- Bushnell, B. (2021). *BBTools*. Retrieved from <https://sourceforge.net/projects/bbmap/>
- Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15), 1972–1973. <https://doi.org/10.1093/bioinformatics/btp348>
- Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, 17(4), 540–552. <https://doi.org/10.1093/oxfordjournals.molbev.a026334>
- Cavender-Bares, J., Ackerly, D. D., & Kozak, K. H. (2012). Special Issue: Integrating ecology and phylogenetics: The footprint of history in modern-day communities. *Ecology*, 93(8), S1–S3. <http://www.jstor.org/stable/23229892>
- Chan, C. X., & Ragan, M. A. (2013). Next-generation phylogenomics. *Biology Direct*, 8(1), 3. <https://doi.org/10.1186/1745-6150-8-3>
- Chen, C.-J., Huang, Y.-C., & Shie, S.-S. (2020). Evolution of multi-resistance to vancomycin, daptomycin, and linezolid in methicillin-resistant staphylococcus aureus causing persistent bacteremia. *Frontiers in Microbiology*, 11, 1414. <https://doi.org/10.3389/fmicb.2020.01414>
- Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., & Thompson, J. D. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*, 31(13), 3497–3500. <https://doi.org/10.1093/nar/gkg500>
- Criscuolo, A., & Gribaldo, S. (2010). BMGE (block mapping and gathering with entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology*, 10(1), 1–21. <https://doi.org/10.1186/1471-2148-10-210>
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>
- Drew, B. T., Gazis, R., Cabezas, P., Swithers, K. S., Deng, J., Rodriguez, R., Katz, L. A., Crandall, K. A., Hibbett, D. S., & Soltis, D. E. (2013). Lost branches on the tree of life. *PLoS Biology*, 11(9), e1001636. <https://doi.org/10.1371/journal.pbio.1001636>
- Driskell, A. C., Ané, C., Burleigh, J. G., McMahon, M. M., O'meara, B. C., & Sanderson, M. J. (2004). Prospects for building the tree of life from large sequence databases. *Science*, 306(5699), 1172–1174.
- Dunn, C. W., Hejnal, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., Sørensen, M. V., Haddock, S. H. D., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R. M., Wheeler, W. C., Martindale, M. Q., & Giribet, G. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452(7188), 745–749. <https://doi.org/10.1038/nature06614>
- Dunn, C. W., Leys, S. P., & Haddock, S. H. D. (2015). The hidden biology of sponges and ctenophores. *Trends in Ecology & Evolution*, 30(5), 282–291. <https://doi.org/10.1016/j.tree.2015.03.003>
- Dunn, C. W., Zapata, F., Munro, C., Siebert, S., & Hejnal, A. (2018). Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proceedings of the National Academy of Sciences of the United States of America*, 115(3), E409–E417. <https://doi.org/10.1073/pnas.1707515115>
- Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Informatics*, 2009, 205–211. [https://doi.org/10.1142/9781848165632\\_0019](https://doi.org/10.1142/9781848165632_0019)
- Felsenstein, J. (1992). Phylogenies from restriction sites: A maximum-likelihood approach. *Evolution*, 46(1), 159–173. <https://doi.org/10.1111/j.1558-5646.1992.tb01991.x>
- Field, J. T. (2021a). Data from: Rapid alignment updating with extensiphy. *Dryad Digital Repository*, <https://doi.org/10.6071/M38T0T>
- Field, J. T. (2021b). McTavishLab/extensiphy: MEE submission (bioinformatics). *Zenodo*, <https://doi.org/10.5281/zenodo.5770686>
- Field, J. T. (2021c). jtfield/phylo\_comparison: MEE submission (bioinformatics). *Zenodo*, <https://doi.org/10.5281/zenodo.5770698>
- Field, J. T., Weinberg, J., Bensch, S., Matta, N. E., Valkiūnas, G., & Sehgal, R. N. (2018). Delineation of the genera Haemoproteus and Plasmodium using RNA-Seq and multi-gene phylogenetics. *Journal of Molecular Evolution*, 86(9), 646–654. <https://doi.org/10.1007/s00239-018-9875-3>
- Gernert, K. M., Seby, S., Schmerer, M. W., Thomas, J. C., Pham, C. D., St Cyr, S., Schlanger, K., Weinstock, H., Shafer, W. M., Raphael, B. H., Kersh, E. N., Hun, S., Hua, C., Ruiz, R., Soge, O. O., Dominguez, C., Patel, A., Loomis, J., Leavitt, J., ... Harvey, A. (2020). Azithromycin susceptibility of *Neisseria gonorrhoeae* in the USA in 2017: A genomic analysis of surveillance data. *The Lancet Microbe*, 1(4), e154–e164. [https://doi.org/10.1016/S2666-5247\(20\)30059-8](https://doi.org/10.1016/S2666-5247(20)30059-8)
- Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies.

- Nature Reviews Genetics*, 17(6), 333–351. <https://doi.org/10.1038/nrg.2016.49>
- Gordon, A., & Hannon, G. J. (2021). *Fastq\_toolkit*. Retrieved from [http://hannonlab.cshl.edu/fastx\\_toolkit/download.html](http://hannonlab.cshl.edu/fastx_toolkit/download.html)
- Grad, Y. H., Harris, S. R., Kirkcaldy, R. D., Green, A. G., Marks, D. S., Bentley, S. D., Trees, D., & Lipsitch, M. (2016). Genomic epidemiology of gonococcal resistance to extended-spectrum cephalosporins, macrolides, and fluoroquinolones in the United States, 2000–2013. *Journal of Infectious Diseases*, 214(10), 1579–1587. <https://doi.org/10.1093/infdis/jiw420>
- Günther, T., & Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genetics*, 15(7), e1008302. <https://doi.org/10.1371/journal.pgen.1008302>
- Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018). Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics*, 34(23), 4121–4123. <https://doi.org/10.1093/bioinformatics/bty407>
- Hardison, R. C. (2003). Comparative genomics. *PLOS Biology*, 1(2), e58. <https://doi.org/10.1371/journal.pbio.0000058>
- Heng, L. (2021). *Seqtk*. Retrieved from <https://github.com/lh3/seqtk>
- Hodcroft, E. B., De Maio, N., Lanfear, R., MacCannell, D. R., Minh, B. Q., Schmidt, H. A., Stamatakis, A., Goldman, N., & Dessimoz, C. (2021). Want to track pandemic variants faster? Fix the bioinformatics bottleneck. *Nature*, 591(7848), 30–33. Retrieved from <https://www.nature.com/articles/d41586-021-00525-x>
- Hodgins, K. A., Yeaman, S., Nurkowski, K. A., Rieseberg, L. H., & Aitken, S. N. (2016). Expression divergence is correlated with sequence evolution but not positive selection in conifers. *Molecular Biology and Evolution*, 33(6), 1502–1516. <https://doi.org/10.1093/molbev/msw032>
- Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M., & Ladle, R. J. (2015). Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, 46(1), 523–549. <https://doi.org/10.1146/annurev-ecolsys-112414-054400>
- Huang, H., & Knowles, L. L. (2016). Unforeseen consequences of excluding missing data from next-generation sequences: Simulation study of RAD sequences. *Systematic Biology*, 65(3), 357–365. <https://doi.org/10.1093/sysbio/syu046>
- Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: A next-generation sequencing read simulator. *Bioinformatics*, 28(4), 593–594. <https://doi.org/10.1093/bioinformatics/btr708>
- Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, 33(6), 1635–1638. <https://doi.org/10.1093/molbev/msw046>
- Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, 12(4), 357–360. <https://doi.org/10.1038/nmeth.3317>
- Leaché, A. D., Banbury, B. L., Felsenstein, J., de Oca, A.-N.-M., & Stamatakis, A. (2015). Short tree, long tree, right tree, wrong tree: New acquisition bias corrections for inferring SNP phylogenies. *Systematic Biology*, 64(6), 1032–1047. <https://doi.org/10.1093/sysbio/syv053>
- Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., Deyholos, M. K., Gitzendanner, M. A., Graham, S. W., Grosse, I., Li, Z., Melkonian, M., & Mirarab, S. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, 574, 679–685. <https://doi.org/10.1038/s41586-019-1693-2>
- Lemmon, A. R., Brown, J. M., Stanger-Hall, K., & Lemmon, E. M. (2009). The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Systematic Biology*, 58(1), 130–145. <https://doi.org/10.1093/sysbio/syp017>
- Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 50(6), 913–925. <https://doi.org/10.1080/106351501753462876>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Liu, K., Warnow, T. J., Holder, M. T., Nelesen, S. M., Yu, J., Stamatakis, A. P., & Linder, C. R. (2012). SATe-II: Very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic Biology*, 61(1), 90.
- Mardis, E. R. (2017). DNA sequencing technologies: 2006–2016. *Nature Protocols*, 12(2), 213–218. <https://doi.org/10.1038/nprot.2016.182>
- Matsen, F. A., Kodner, R. B., & Armbrust, E. V. (2010). pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11(1), 538. <https://doi.org/10.1186/1471-2105-11-538>
- McTavish, E. J., Drew, B. T., Redelings, B., & Cranston, K. A. (2017). How and why to build a unified tree of life. *BioEssays*, 39(11), 1700114–1700114. <https://doi.org/10.1002/bies.201700114>
- McTavish, E. J., Pettengill, J., Davis, S., Rand, H., Strain, E., Allard, M., & Timme, R. E. (2017). TreeToReads – A pipeline for simulating raw reads from phylogenies. *BMC Bioinformatics*, 18(1), <https://doi.org/10.1186/s12859-017-1592-1>
- Molloy, E. K., & Warnow, T. (2018). To include or not to include: The impact of gene filtering on species tree estimation methods. *Systematic Biology*, 67(2), 285–303. <https://doi.org/10.1093/sysbio/syx077>
- Murolo, S., & Romanazzi, G. (2015). In-vineyard population structure of ‘Candidatus Phytoplasma solani’ using multilocus sequence typing analysis. *Infection, Genetics and Evolution*, 31, 221–230. <https://doi.org/10.1016/j.meegid.2015.01.028>
- NCBI. (2020). *NCBI Pathogen database*. Retrieved from <https://www.ncbi.nlm.nih.gov/pathogens/organisms/>
- Nguyen, N. D., Mirarab, S., Kumar, K., & Warnow, T. (2015). Ultra-large alignments using phylogeny-aware profiles. *Genome Biology*, 16(1), 124. <https://doi.org/10.1186/s13059-015-0688-z>
- Rambaut, A., & Grassly, N. C. (1997). Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, 13, 235–238. <https://doi.org/10.1093/bioinformatics/13.3.235>
- Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1–2), 131–147. [https://doi.org/10.1016/0025-5564\(81\)90043-2](https://doi.org/10.1016/0025-5564(81)90043-2)
- Rocha, E. P., Smith, J. M., Hurst, L. D., Holden, M. T., Cooper, J. E., Smith, N. H., & Feil, E. J. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of Theoretical Biology*, 239(2), 226–235. <https://doi.org/10.1016/j.jtbi.2005.08.037>
- Ros-Freixedes, R., Battagin, M., Johnsson, M., Gorjanc, G., Mileham, A. J., Rounsley, S. D., & Hickey, J. M. (2018). Impact of index hopping and bias towards the reference allele on accuracy of genotype calls from low-coverage sequencing. *Genetics Selection Evolution*, 50(1), 1–14. <https://doi.org/10.1186/s12711-018-0436-4>
- Sánchez-Reyes, L. L., Kandziora, M., & McTavish, E. J. (2021). Physcraper: A Python package for continually updated phylogenetic trees using the Open Tree of Life. *BMC Bioinformatics*, 22, 355. <https://doi.org/10.1186/s12859-021-04274-6>
- Seemann, T. (2021). *Snippy*. Retrieved from <https://github.com/tseemann/snippy>
- Shakya, M., Ahmed, S. A., Davenport, K. W., Flynn, M. C., Lo, C.-C., & Chain, P. S. G. (2020). Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life. *Scientific Reports*, 10(1), 1723. <https://doi.org/10.1038/s41598-020-58356-1>
- Smith, S. D., Pennell, M. W., Dunn, C. W., & Edwards, S. V. (2020). Phylogenetics is the new genetics (for most of biodiversity). *Trends*

- in *Ecology & Evolution*, 35(5), 415–425. <https://doi.org/10.1016/j.tree.2020.01.005>
- Soltis, D. E., & Soltis, P. S. (2003). The role of phylogenetics in comparative genetics. *Plant Physiology*, 132(4), 1790–1800. <https://doi.org/10.1104/pp.103.022509>
- Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), 1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>
- Streicher, J. W., Schulte, J. A., & Wiens, J. J. (2016). How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. *Systematic Biology*, 65(1), 128–145. <https://doi.org/10.1093/sysbio/syv058>
- Sukumaran, J., & Holder, M. T. (2010). DendroPy: A Python library for phylogenetic computing. *Bioinformatics*, 26(12), 1569–1571. <https://doi.org/10.1093/bioinformatics/btq228>
- Swofford, D., Olsen, G., Waddell, P., & Hillis, D. M. (1996). Phylogenetic inference. In D. M. Hillis, C. Moritz, & B. K. Mable (Eds.), *Molecular systematics* (Chapter 5, pp. 407–514). Sinauer Associates.
- Talavera, G., & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, 56(4), 564–577. <https://doi.org/10.1080/10635150701472164>
- Treangen, T. J., Ondov, B. D., Koren, S., & Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology*, 15(11). <https://doi.org/10.1186/s13059-014-0524-x>
- Vasimuddin, M., Misra, S., Li, H., & Aluru, S. (2019). Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 314–324. <https://doi.org/10.1109/IPDPS.2019.00041>
- Wang, L., & Jiang, T. (1994). On the complexity of multiple sequence alignment. *Journal of Computational Biology*, 1(4), 337–348. <https://doi.org/10.1089/cmb.1994.1.337>
- Wilkinson, M. (1995). Coping with abundant missing entries in phylogenetic inference using parsimony. *Systematic Biology*, 44(4), 501–514. <https://doi.org/10.2307/2413657>
- Williamson, R. J., Josephs, E. B., Platts, A. E., Hazzouri, K. M., Haudry, A., Blanchette, M., & Wright, S. I. (2014). Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora*. *PLoS Genetics*, 10(9), e1004622. <https://doi.org/10.1371/journal.pgen.1004622>
- Xi, Z., Liu, L., & Davis, C. C. (2016). The impact of missing data on species tree estimation. *Molecular Biology and Evolution*, 33(3), 838–860. <https://doi.org/10.1093/molbev/msv266>
- Yin, J., Zhang, C., & Mirarab, S. (2019). ASTRAL-MP: Scaling ASTRAL to very large datasets using randomization and parallelization. *Bioinformatics*, 35(20), 3961–3969. <https://doi.org/10.1093/bioinformatics/btz211>

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Field, J. T., Abrams, A. J., Cartee, J. C., & McTavish, E. J. (2022). Rapid alignment updating with Extensiphy. *Methods in Ecology and Evolution*, 13, 682–693. <https://doi.org/10.1111/2041-210X.13790>