UNIVERSITY OF CALIFORNIA, MERCED

Building Accurate Evolutionary Estimates for Large Genomic Data Sets and Applying These Inferences to Pathogen Surveillance

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Quantitative and Systems Biology

by

Jasper Toscani Field

Committee in charge:

Professor Chris Amemiya, Chair of Advisory Committee

Professor Gordon Bennett

Professor Siavash Mirarab

Professor Emily Jane McTavish, Supervisor

The Dissertation of Jasper Toscani Field is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Gordon Bennett, Ph.D

Siavash Mirarab, Ph.D

Emily Jane McTavish, Ph.D. Supervisor

Chris Amemiya, Ph.D. Chair

University of California, Merced
2022

Table of Contents

List of Figures

List of Tables

Acknowledgments

Funding

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. Jasper was supported by the NSF NRT Grant DGE-1633722.

Personal Acknowledgements

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 dissertation are those of the authors and do not necessarily represent the official position of the CDC.

Vita for Jasper S. Toscani Field

Education:

PhD	University of California, Merced, Quantitative and Systems Biology	2022
MS	San Francisco State University, Biology	2017
BS	San Francisco State University, Environmental Studies	2014

List of Publications:

Field, J. T., Abrams, A. J., Cartee, J. C., & McTavish, E. J. (2022). Rapid alignment updating with Extensiphy. *Methods in Ecology and Evolution*. *13*(3), 682-693.

Weinberg, J., **Field, J. T.**, Ilgūnas, M., Bukauskaitė, D., Iezhova, T., Valkiūnas, G., & Sehgal, R. N. (2019). De novo transcriptome assembly and preliminary analyses of two avian malaria parasites, Plasmodium delichoni and Plasmodium homocircumflexum. *Genomics*, *111*(6), 1815-1823.

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.

Awards and Fellowships:

NSF Research Training Program in Intelligent Adaptive Systems 2021
UC Merced Graduate Division

NSF Research Training Program in Intelligent Adaptive Systems 2020
UC Merced Graduate Division

JGI Distinguished Graduate Internship 2020
Joint Genome Institute

Best Talk 2019
Northern California Computational Biology Symposium

ASSURE Graduate Student Research Fellowship 2019
Michigan State University College of Natural Science

Graduate Distinguished Achievement Award 2017
San Francisco State University Department of Biology

IRA Research Award Grant 2015
San Francisco State University Division of Graduate Studies

Abstract

Tracing the evolutionary history of pathogen outbreaks allows researchers to develop appropriate public health interventions. For example, phylogenetic inferences have been key data informing the response to the on-going Covid-19 pandemic. I worked with researchers at the CDC to develop and test tools to rapidly infer phylogenies for large genomic data sets. I applied these new tools to understand the evolution of gonorrhea (*Neisseria gonorrhoeae*), a pathogen of major public health importance, which is increasing both in prevalence, and in rate of anti-microbial resistance. I found that our tools reduced program runtime and data set fragmentation while producing reliable phylogenetic estimates. I also investigated the underlying approach used by our methods to assemble genomic sequences. I found that reference choice is an important consideration when assembling sequences, as greater evolutionary distance to reference genome leads to an increase in errors. However, I found that while errors increase with evolutionary distance to reference genome, overall phylogenetic topology is largely unaffected. Finally, having shown that my original tools are reliable, I extended the methods and applied them to analyzing the evolutionary relationships of over 1,000 *N. gonorrhoeae* isolates in order to map gain and loss of anti-microbial resistant alleles data. Together these results demonstrate that the tools I have developed can be used to rapidly and accurately analyze genome scale data for thousands of lineages, and link those evolutionary inferences with important metadata to better inform public health interventions.

Chapter 1

Rapid Alignment Updating with Extensiphy

Published; Field, J. T., Abrams, A. J., Cartee, J. C., & McTavish, E. J. (2022). Rapid alignment updating with Extensiphy. *Methods in Ecology and Evolution*, *13*(3), 682-693.

1.1 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 analyzing 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.

Key words: Sequence Alignment, Genomes, Phylogenetics, Evolutionary Biology, Software, Monitoring

1.2 Background

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; Dunn et al., 2015). Phylogenetic estimates are essential frameworks for comparative genetics and genomics (Soltis & Soltis, 2003; Hardison, 2003; Dunn et al., 2018; Smith et al., 2020). 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., 2017a, 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 pair-wise awareness of sequence homology (Swofford et al., 1996; Chenna et al., 2003). Alignment as a procedure is the process of finding homology between two or more DNA sequences (Kim et al., 2015; Misra et al., 2019). The procedure of multiple sequence alignment is computationally challenging, which must be repeated when new data are added to existing alignments (Wang & Jiang, 1994; Chenna et al., 2003; Liu et al., 2012; Treangen et al., 2014; Field et al., 2018). 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; Liu et al., 2012; Nguyen et al., 2015; Grad et al., 2016; Hadfield et al., 2018; Leebens-Mack et al., 2019; NCBI, 2021). 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, 2021). 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 artifact such as low sequencing coverage or read quality in some genomic regions. It has been demonstrated that biological reality and bioinformatic artifacts 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 (Wilkinson, 1995; Driskell et al., 2004; Lemmon et al., 2009; Huang & Knowles, 2016; Xi et al., 2016; Molloy & Warnow, 2018). As such, the effect of missing data on evolutionary analyses has been hotly debated (Castresana, 2000; Talavera & Castresana, 2007; Capella-Gutiérrez et al., 2009; Lemmon et al., 2009; Treangen et al., 2014; Huang & Knowles, 2016; Xi et al., 2016; Molloy & Warnow, 2018). Some studies laud the effects of removing alignment regions with high proportions of missing data as improving phylogenetic estimations (Castresana, 2000; Talavera & Castresana, 2007; Capella-Gutiérrez et al., 2009; Criscuolo & Gribaldo, 2010; 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 (Castresana, 2000; Capella-Gutiérrez et al., 2009; 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 tends 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 (Lemmon et al., 2009; Huang & Knowles, 2016; 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; Lewis, 2001; Leaché et al., 2015). 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 (Rocha et al., 2006; Briggs et al., 2009; Huerta-Cepas et al., 2016). 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 (Williamson et al., 2014; Murolo & Romanazzi, 2015; Hodgins et al., 2016). 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 is 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 (Seemann, 2021; Bankevich et al., 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.

1.3 Methods and Materials

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 Fig 1.1. Alternative options for Extensiphy parameters and functionality are described in the following sections.

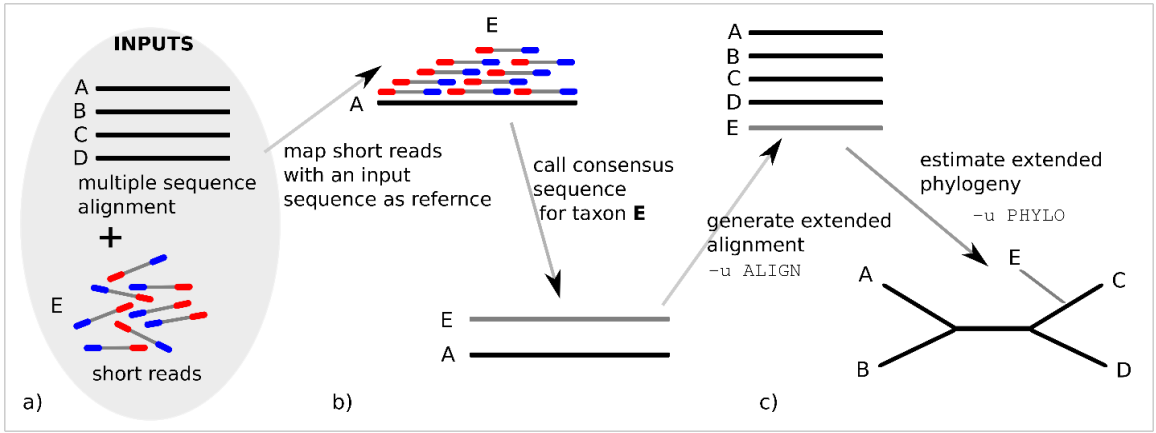

Figure 1.1: Default workflow of Extensiphy. a) Input an alignment file and new raw reads. b) Align reads to reference and call the consensus sequence. c) Output updated alignment and tree files.

Description of Extensiphy

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 (Misra 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.

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 converstion 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 vcfutils.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.

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, e.g. 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 phylogentic 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.

Program Comparison

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 single-nucleotide polymorphism (SNP) calling (Seemann, 2021; Castresana, 2000; Treangen et al., 2014; Bush et al., 2020). 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. There are two versions of the code, one for

analyzing each simulated and empirical sequence data. The empirical data comparison software requires whole-genome short-read sequences. The software for analyzing 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.

Datasets

To construct our simulated high-throughput dataset with a known phylogenetic topology, we used TreeToReads (McTavish et al., 2017b). 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, U.S.A) 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 51,924bp 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 and Grassly 1997). The nucleotide rate matrix of was: 1.039821, 5.116539, 0.339204, 0.910812, 5.291090, 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 data set 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, 2021). 14 isolates were identified as *N. meningitidis* and were removed from subsequent analyses. The final empirical dataset consisted of 1,223 samples.

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.

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 were 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.

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.

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.

Program Output Comparisons

Program Output Comparisons 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 programs 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.

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.

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 analyzed, 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 analyzed, 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.

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.

1.4 Results

Simulated Dataset Results

Runtime

Using Extensiphy, individual sequences were assembled at a mean rate of four seconds per sequence and the overall program runtime was completed in 6 minutes and 45 seconds (Table 1.1). *De novo* pipeline runtimes were a mean of eight seconds per individual sequence and a complete program runtime of 21 minutes. Snippy's mean individual sequence assembly time was three seconds per sequence and a complete program runtime of 10 minutes and 28 seconds.

Table 1.1: 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	6m 45s	21m	10m 28s
Individual sequence runtime	4s	8s	4s
Total miscalled bases	15	21	359
Total bases per taxon	51157	50245	51191
Total bases analyzed	10691913	10500766	10698919
RF distance to true tree	56	55	98

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 1.1). 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.

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 1.1). 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.

Missing data

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

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 1.1).

Empirical Dataset Results

Runtime

When processing and analyzing data from the empirical dataset, Extensiphy produced consensus sequences in a mean time of slightly over 6 minutes and produced a complete alignment in 38 hours (Fig 1.2, Table 1.2). The *de novo* pipeline assembled sequences in a mean time of 41 minutes and produced a complete alignment in 236 hours. Snippy produced individual sequences in a mean time of 41 seconds and produced a complete alignment in 18 hours.

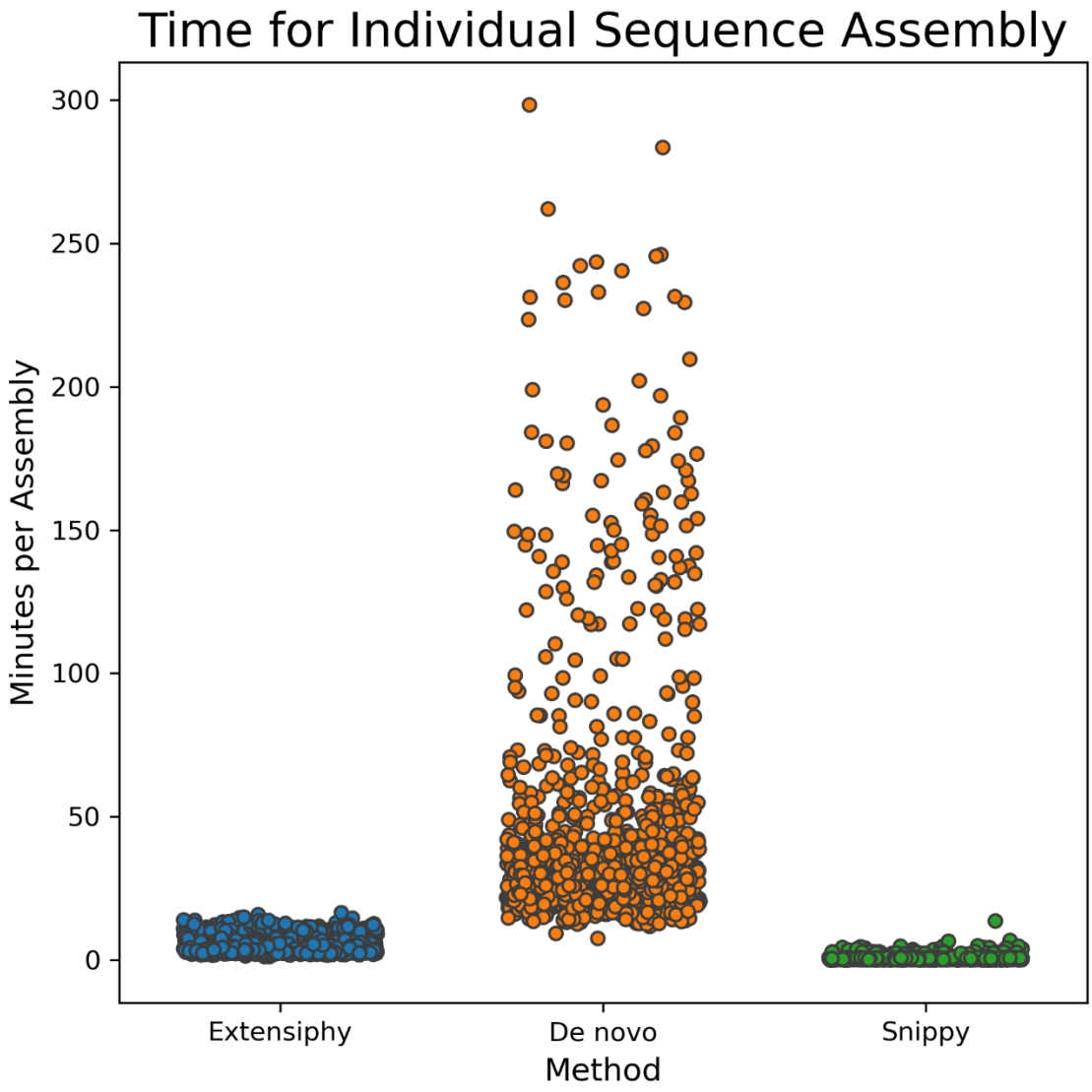

Figure 1.2: The time required by each method to assemble all sequences associated with each taxon in the empirical dataset.

Table 1.2: 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	38h	236h	18h
Average individual sequence runtime	6m 21s	41m	41s

Alignment length

Individual sequences produced by Extensiphy were all of 1,859,910 nucleotides in length for a total of $2.293x10^9$ nucleotides in the final alignment (Table 1.3). 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 1.3). The *de novo* pipeline returned individual sequences of 751,033 nucleotides and a total of $9.215x10^8$ 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 $2.732x10^9$ nucleotides in the final alignment. Locus values were not reported for Snippy as Snippy operates using whole-genome inputs and outputs.

Table 1.3: 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	2293269030	921517491	2732911282
Total gaps or degenerate bases	4891739	3469861	224835516
Average locus length	5868	1465	-
Loci output per program	317	522	-

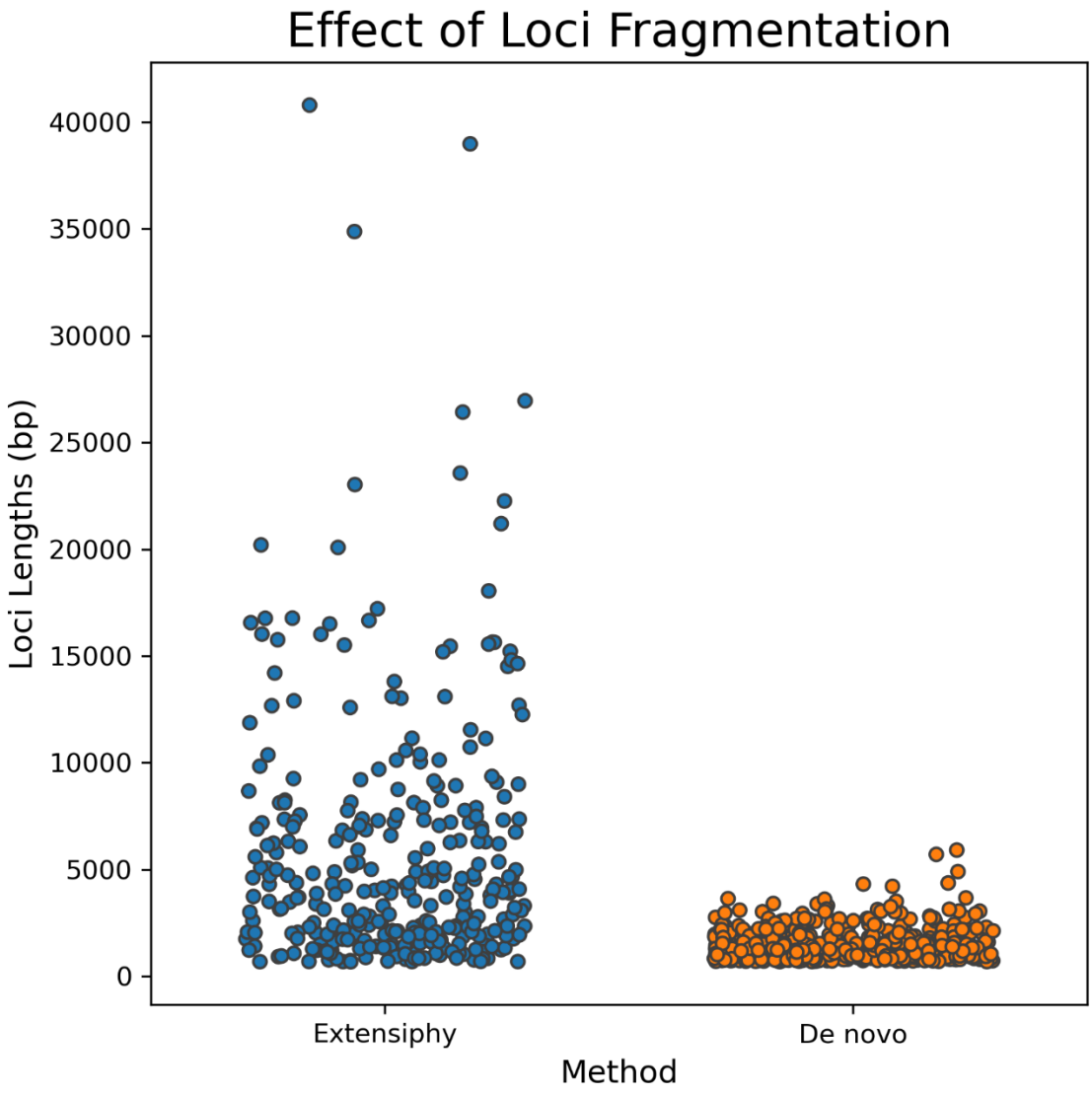

Figure 1.3: Empirical dataset locus lengths returned by Extensiphy and the *de novo* assembly pipeline.

Alignment accuracy

We assessed empirical basecalls for the outputs of all three programs against each other as true basecalls cannot be described with certainty for empirical sequence data. The Extensiphy-*de novo* pipeline comparison contained 490 differing nucleotides from 31,909,017 analyzed sites between both alignments. The Extensiphy-Snippy comparison produced 27,778 differing nucleotides from 338,286,158 analyzed sites between both alignments. The comparison of Snippy and the *de novo* pipeline alignments contained 142 differing nucleotides from 31,974,892 sites analyzed between both alignments.

Missing data

We assessed empirical missing data in the same manner as empirical basecalls, 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 analyzed sites between both alignments. The Extensiphy-Snippy comparison produced 1,857,035 differing gaps or degenerate nucleotides from 338,286,158 analyzed 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 analyzed between both alignments. When analyzing the complete alignment for each program, the alignment produced by Extensiphy contained 4,891,739 gaps and degenerate nucleotides (Table 1.3). 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.

Phylogenetic accuracy

When analyzing 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 comparison produced an RF distance of 749 (Table 1.4). The *de novo* pipeline-Snippy comparison produced an RF distance of 676.

Table 1.4: 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	-

1.5 Discussion

Sequencing efforts are expanding for the collection of genomic data (Goodwin et al., 2016; Mardis, 2017; Hodcroft et al., 2021). 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 analyzed 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 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 analyzed 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 doesn't constrain the user to a specific method of phylogenetic estimation.

Extensiphy is designed to integrate new genomic data with existing data sets. 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. Extenisphy does not require a full reference genome and can be applied to integrating sequences form whole genome data into even single locus data sets. These few or single locus data sets 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 (Ros-Freixedes et al., 2018; Günther & Nettelblad, 2019). 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 researcher's goal.

1.6 Conclusion

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 basecall 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.

1.7 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., Kućan, Ž., & others. (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), giaa007. https://doi.org/10.1093/gigascience/giaa007

Bushnell, B. (Accessed 5/5/2021). *BBTools*. 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.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, *17*(4), 540–552.

Chan, C. X., & Ragan, M. A. (2013). Next-generation phylogenomics. *Biology Direct*, *8*(1), 3. https://doi.org/10.1186/1745-6150-8-3

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

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.

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

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.

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., & others. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158.

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., Hejnol, 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., & Hejnol, A. (2018). Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proceedings of the National Academy of Sciences*, *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

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. (2021), Data for: Rapid Alignment Updating with Extensiphy, Dryad, Dataset, https://doi.org/10.6071/M38T0T

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.

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

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. (Accessed 5/5/2021). *Fastq_toolkit*. 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.

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. (Accessed 5/5/2021). *Seqtk*. 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. In *Nature*. 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.

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 W., Li L, Myers J. R., Marth G. T. (2012). ART: a next-generation sequencing read simulator, *Bioinformatics* 28 (4): 593-594

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.

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.

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360.

Leaché, A. D., Banbury, B. L., Felsenstein, J., de Oca, A. nieto-Montes, & 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., & others. (2019). *One thousand plant transcriptomes and the phylogenomics of green plants*.

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.

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, Heng, 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

Li, Heng, Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

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. (2017a). How and Why to Build a Unified Tree of Life. *BioEssays : news and reviews in molecular, cellular and developmental biology*, *39*(11), 10.1002/bies.201700114. https://doi.org/10.1002/bies.201700114

McTavish, E. J., Pettengill, J., Davis, S., Rand, H., Strain, E., Allard, M., & Timme, R. E. (2017b). TreeToReads—A pipeline for simulating raw reads from phylogenies. *BMC Bioinformatics*, *18*(1). https://doi.org/10.1186/s12859-017-1592-1

Vasimuddin, M., Misra, S., Li, H., and Aluru, S. "*Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems*," *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2019, pp. 314-324, doi: 10.1109/IPDPS.2019.00041.

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.

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.

NCBI. (2020). *NCBI Pathogen Database*. 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., and 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

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

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.

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.

Sánchez-Reyes, L.L., Kandziora, M. & McTavish, E.J. Physcraper: a Python package for continually updated phylogenetic trees using the Open Tree of Life. *BMC Bioinformatics* 22, 355 (2021). https://doi.org/10.1186/s12859-021-04274-6

Seemann, T. (Accessed 5/5/2021). *Snippy*. 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.

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.

Sukumaran, J., & Holder, M. T. (2010). DendroPy: A Python library for phylogenetic computing. *Bioinformatics*, *26*(12), 1569–1571.

Swofford, D., Olsen, G., Waddell, P. & Hillis, D. M. (1996). Phylogenetic Inference. *Molecular Systematics, (ed. DM Hillis et. Al.), chap 5 pp. 407-514.* Sinauer Associates, Sunderland, Massachusettes.

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*. 15.

Wang, L., & Jiang, T. (1994). On the complexity of multiple sequence alignment. *Journal of Computational Biology*, *1*(4), 337–348.

Wilkinson, M. (1995). Coping with abundant missing entries in phylogenetic inference using parsimony. *Systematic Biology*, *44*(4), 501–514.

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 Genet*, *10*(9), e1004622.

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.

Yin, J., Zhang, C., & Mirarab, S. (2019). ASTRAL-MP: scaling ASTRAL to very large datasets using randomization and parallelization. *Bioinformatics*, *35*(20), 3961–3969.

Chapter 2

An Exploration of the Effects of Reference Choice on Phylogenomics

Field, J. T. & McTavish, E. J.

2.1 Abstract

Modern genomic studies often involve detailed analyses of single nucleotide polymorphisms (SNPs) or broader investigations of evolutionary history through phylogenetic estimation. The results of approaches can be influenced by dataset curation and sequence assembly. While *de novo* assembly is regarded as the gold standard of sequence assembly methods, the processing time and resources required to generate contiguous sequences can be prohibitive. Reference-guided sequence assembly offers an efficient alternative with the caveat that a reference taxon must be chosen before assembly can be performed. Previous work has shown that reference choice can introduce bias and that in certain contexts, the effects of reference choice are evident in both SNP and phylogenetic results. While recent studies have described many factors affecting reference bias, quantitative measures of sequence and phylogenetic variation due to reference bias have not been adequately pursued. We use an empirical dataset to investigate the effects of reference choice at varying levels of evolutionary distance on read coverage and error rates in multiple sequences alignments. Our study culminates with the assessment of topological and tree length changes based on reference choice. We find that evolutionary distance to the reference taxon has a significant effect on the number of errors in assembled sequences and that an overwhelming proportion of these errors are biased to the bases in the reference sequence. We demonstrate that read coverage is only weakly driven by distance to the reference taxon and that error rates quickly diminish at 20-25x coverage and above. Finally, we find that topological accuracy is largely resilient to reference choice but tree length variation does occur. Our findings suggest that reference choice is less important if the goal of the study is motivated by an assessment of evolutionary relationships but plays a larger role in sequence composition, potentially affecting SNP studies.

2.2 Introduction

All fields of biology are directly or indirectly influenced by the process of evolution. A critical data structure for studying evolutionary biology is the phylogeny, a tree-like graph describing the evolution of various selected organisms (O'hara. 1997). While phylogenies have been present in biology since the birth of evolutionary theory, the application of phylogenetic systematics to genetic information has broadened the field's influence dramatically and publications expanding these methods have become some of the most influential in the field of biology (Felsenstein. 1981; Felsenstein. 1985; Saitou & Nei. 1987; Swofford et al., 1996; Van Noorden et al., 2014). Accurate phylogenetic estimates are of great importance in varied fields such as vaccine development and understanding human evolution (Ovchinnikov et al., 2000; Rolland et al., 2011). However, choices made during data processing prior to phylogenetic estimation can affect valuable results.

Phylogenetic estimates can be influenced by many factors, including missing data, taxon sampling and the method used during estimation (Zwickl & Hillis. 2002; Leaché et al., 2015; Huang & Knowles. 2016). Previous research has shown that increased taxon sampling can improve phylogenetic accuracy (Zwickl & Hillis. 2002). Incomplete or non-random taxon sampling can affect the topology of your estimates, almost irrespective of sequence sizes (Zwickl & Hillis 2002; Heath 2008). While these statements about taxon sampling support broad inclusion of taxa, such expansive inclusion can lead to the second factor highlighted as affecting sequence dataset composition: missing data (Eaton et al., 2017). Succinctly, the amount of missing data is expected to increase as the evolutionary distance between taxa increases (Huang & Knowles. 2016; Eaton et al., 2017; Molloy & Warnow. 2018). Early approaches to missing data involved filtering loci containing missing data using varying degrees of strictness, resulting in truncated datasets that would frequently influence branch lengths (Huang & Knowles. 2016; Molloy & Warnow. 2018).

While addressing taxon sampling relies on researcher decisions prior to sample collection and well before computational data processing, concerns around missing data indicate that consideration of bioinformatic processing practices is warranted. ~~Ignoring the sample collection and wet lab practices that may affect data quality,~~ one aspect of missing data is rooted in a project upstream bioinformatics decisions on how to assemble sequences, specifically in cases when assembling with a reference sequence. Reference guided sequence assembly is a commonly performed assembly approach involving aligning reads to a trusted, high quality sequence to efficiently output relevant data. Greater evolutionary distance between reference and query can decrease the likelihood of successful read alignment (Bertels et al., 2014). Such mis-alignments can result in missing data or, perhaps more influentially, successfully aligned reads carrying the reference allele more often than alternatives, leading to biased basecalls (Gunter & Nettleblad. 2019). Therefore, the choice of reference sequence can greatly affect downstream analyses, including phylogenetic estimation.

Recent studies have investigated the specific effects of reference choice on a variety of datasets (Degner et al., 2009; Günther & Nettleblad. 2019; Prasad et al., 2021; Rick et al., 2022). Additionally, the effect on phylogenetic topology and branch length has also

been demonstrated. Work by Bertels et al., (2014) investigated the effects of reference choice on simulated and empirical bacterial sequence data. Bertels et al., (2014) found that by varying distance to the reference in four-taxon phylogenies caused attempts to align reads to more divergent genome regions to fail, leading to incorrectly estimated branch lengths and topologies (Bertels et al., 2014). In an investigation of the impact of reference choice on the historically convoluted topic of domestication, Gopalakrishnan et al., (2017) *de novo* assembled a wolf genome for use as a reference when assembling sequences from several published canine datasets. Gopalakrishnan et al., contrasted the topology based on these wolf-guided sequences with a topology based on sequences assembled using a boxer genome. This comparison of topologies highlighted the change of several monophyletic clades throughout the phylogeny (Gopalakrishnan et al., 2017). Supporting the established theory of reference bias as a broad effector of the results of read alignment, previous research has shown reference bias is not limited to one clade of the evolutionary tree of life. Many studies on the effects of reference choice use human datasets. Work by Degner et al., (2009), one of the earliest studies of short read alignment on reference sequences, used empirical and simulated samples from the Yorba peoples of Nigeria. Degner et al., (2009) found and subsequently verified that sequencing errors increased mapping preference for the reference allele. More recently, a 2021 study by Li et al (2021) noted discrepancies of variant calling when using previous and current versions of the human genome as a reference. Li et al., (2021) found that a small but significant number of variant basecalls had a noticeable effect on disease associated loci and genes, highlighting the effects of region-specific reference bias even at very low evolutionary distances. An investigation of multiple archaic human datasets by Gunther & Nettleblad (2019) found the effects of reference bias are particularly pronounced in conditions common to paleogenomic studies, including low sequencing coverage and shorter sequence fragment size. The results of Gunther & Nettleblad (2019) also supported the hypothesis that strict quality filters increase reference allele preference by reducing initial acceptable variant matches during read alignment. Taken together, this body of work describes the broad and continuing importance reference bias has on biological studies.

Biological studies would clearly benefit from quantitative information on how reference choice could potentially affect their dataset. Here we describe our investigation into the effects of how this choice may bias inferences at several stages of bioinformatics workflow. Using an empirical dataset of *Procellariiform* ultra conserved element (UCE) sequence data from Estandia et al., (2021), we investigated how reference choice at varying evolutionary distances alters sequence structure and to what extent bases in sequences assembled using a reference were biased towards the reference sequence. We also describe the downstream effect these bases have on phylogenetic topology by comparing our estimated phylogenies to the original phylogeny published by Estandia et al., (2021). Our goal with this work is to elucidate the potential effects a choice of reference could have on downstream analyses in order to improve inferences and methodological decisions.

2.3 Methods

Datasets

We used data published by Estandia et al., (2021) on *Procellariiform* seabirds. The focus of the Estandia dataset were ultra conserved elements from 54 species comprised of 51 species of *Procellariiform* and 3 outgroup species. We downloaded the data from the DataDryad repository which contained, among other study data, the homologous sequence alignments and associated phylogenies produced using those alignments (Estandia et al., 2021). We downloaded the original raw read files associated with each sample in the Estandia dataset from the National Center for Biotechnology Information Short Read Archive (NCBI SRA). To construct the sequence dataset for all subsequent analyses, we used each sequence in the Estandia alignment as a reference sequence to assemble all 54 sequences. The result of this reference guided sequence assembly process was 54 alignments, each containing with 54 sequences. Alignments were produced using the Extensiphy bioinformatics pipeline (Field et al., 2022).

Extensiphy builds or updates sequence alignments with new, whole-genome sequencing data. Extensiphy aligns a sample's raw reads to a reference sequence and calls a consensus sequence. The output consensus sequence is the same length and has the same gap positions as the reference sequence, ensuring synteny. Once all samples have been assembled in this manner, the sequences are combined with the original alignment or reference sequence to form a new, extended alignment. In this version of Extensiphy, read alignment is performed by BWA-MEM2 (Vasimuddin et al., 2019), the aligned read files are indexed by Faidx (Li et al., 2009) and sequence alignment files are converted to binary alignment files by Samtools View (Li et al., 2009). Variant nucleotide calling is performed by the Bcftools Mpileup (Li et al., 2009), which outputs a Variant Call File (VCF) (Danecek et al., 2011). Vcfutils.pl (Gordon & Hannon. 2021) then converts the VCF to a fastq format file and seqtk converts the fastq file to a fasta (Heng. 2021). A more in depth description of Extensiphy and the utility of the program are discussed in Field et al., (2022).

Phylogenetic Distance

Using the primary phylogeny published by Estandia, we calculated the phylogenetic distance of every taxon pair in the phylogeny using the Dendropy Python library (Sukumaran & Holder. 2016; Estandia et al., 2021). In summary, our dataset was comprised of 2,916 sequences in 54 separate alignments and the phylogenetic distance between the query taxon and the reference taxon that was used to assemble the query sequences.

Effect of Phylogenetic Distance on Basecall Errors

We compared each new Extensiphy-produced, reference-based sequence to the published sequence of the query taxon and counted the number of bases not matching the published, *de novo* assembled query taxon sequence, hereafter referred to as errors. We

also compared our basecalls to the reference sequence used for each assembly. We distinguished between two error types: errors of unambiguous bases and total errors. Unambiguous errors are a difference where both sequences have a basecall of one of the four DNA nucleotides: A, C, G or T that differ from the published sequence. Total errors include all of the unambiguous errors but also include differences where one or both sequences have a degenerate base or a gap, the most common of which is N.

Distance Correlation Analysis

First, we examined the correlation between the phylogenetic distance to the reference and the number of unambiguous errors for all sequence comparisons. Second, we examined the correlation between phylogenetic distance to the reference and the number of total errors for all sequence comparisons. The third and fourth analyses used the same predictor (distance to reference) and response (unambiguous errors) variables for testing unambiguous errors as the first analysis but were performed on data associated with each individual taxon rather than the entire dataset. These individual taxon datasets are inferred sequences or a single query that used each taxon as a reference. We performed simple linear regression (Freedman. 2009) in base R (R Core Team. 2022) to investigate the relationship between phylogenetic distance to the reference and the number of erroneous bases in each dataset with phylogenetic distance to the reference as the explanatory variable and the number of errors as the response variable.

Basecall Error Bias to the Reference

We used our sequence comparison data to investigate if unambiguous errors, sites that did not receive the same nucleotide as the Estandia nucleotide for that taxon, were identical to the reference nucleotide at an unexpected rate. During sequence comparison, we collected information on unambiguous errors that matched the reference sequence. We subtracted the number of unambiguous errors matching the reference from the total number of unambiguous errors to calculate the number of unambiguous errors not matching the reference sequence. We estimated that errors matching the reference sequence should occur at a rate of roughly 25% of the total number of errors if errors are random across all four possible bases. To compare the observed proportion of errors matching the reference to our expected proportion of errors matching the reference, we used a two-sided one proportion z-test (Sprinthall. 2003). We also performed a Pearson chi-squared test to examine the independence of individual comparison's errors matching and not matching the reference sequence (Balakrishnan et al., 2013).

The Effect of Coverage on Errors

We used data from our basecall analyses to investigate the relationship between read coverage and error rate. To assess the read coverage at each unambiguous error, we assessed the individual base coverage values in the variant call files (VCF) produced by the consensus calling portion of Extensiphy. We separated every unambiguous error into a

bin based on the integer level of read coverage of that base. Error rate was calculated by dividing the number of errors at each level of coverage by the total number of unambiguous identical (i.e. correct) bases at the same coverage level. To statistically investigate the relationship between coverage and error rate, we used coverage as the predictor variable and error rate as the response variable in a simple linear regression analysis.

Phylogenetic Analyses

To assess the effect of reference choice on phylognetic inference, we estimated phylogenies for every reference assembled alignment produced during our Extensiphy runs. Here reference assembled alignment indicates an alignment produced by Extensiphy using one of the sequences published by Estandia et al., (2021) as a reference sequence. We used the program RaxML for phylogenetic estimation using the GTRGAMMA model of nucleotide substitution (Stamatakis. 2014). We performed 100 bootstrap replicates and computed a majority rule consensus phylogeny. Our final phylogenetic estimation outputs were the best tree of each estimation run and the majority rule consensus tree produced by the bootstrap replicates. We used Dendropy (Sukumaran & Holder. 2010) to calculate the unweighted Robinson Foulds distance (Robinson & Foulds. 1981) for each phylogeny in the majority rule consensus dataset by comparing the phylogenies we produced to the original phylogeny published by Estandia et al., (2021). We also computed the weighted Robinson Foulds distance for phylogenies in the best tree dataset by comparing the phylogenies we estimated to the original phylogeny produced by Estandia et al., (2021). To explore the changes to overall tree lengths based on reference choice, we calculated the total tree length of each reference based phylogeny as the summed length of each branch in the phylogeny using Dendropy. We subtracted the total tree length of the Estandia phylogeny from the total tree length of each reference-based phylogeny, noting whether the value was positive or negative.

2.4 Results

Effect of Phylogenetic Distance to Reference on Basecall Errors

In our comparisons of reference-based sequences to sequences produced by Estandia et al., (2021), we separated errors into two groups: unambiguous errors and total errors. The range of unambiguous errors was between 4 and 4,037 bases in each comparison of 2,328,289 bases, resulting in a range of unambiguous error rates between 1.7e-06 and 0.0017 errors per base. The range of total errors recorded during sequence comparisons, including ambiguities, was between 878 and 1,101,714 bases per comparison, leading to a range of total error rates between 0.0003 and 0.473 errors per base. To investigate the effect of phylogenetic distance to the reference taxon on errors, we partitioned our error data using four schemes. The first two analyses used a dataset incorporating every sequence produced using every reference, leading to 2,916 comparisons to the published sequences in total. The results of simple linear regression on

phylogenetic distance to the reference as the predictor variable and unambiguous errors as the response variable produced a p-value of 2.2e-16, an r-squared value of 0.918 and an adjusted r-squared value of 0.918 (Figure 2.1). Simple linear regression using phylogenetic distance to the reference taxon as predictor variable and total errors as the response variable returned a p-value of 0.018, an r-squared of 0.001 and an adjusted r-squared of 0.001 (Figure 2.2).

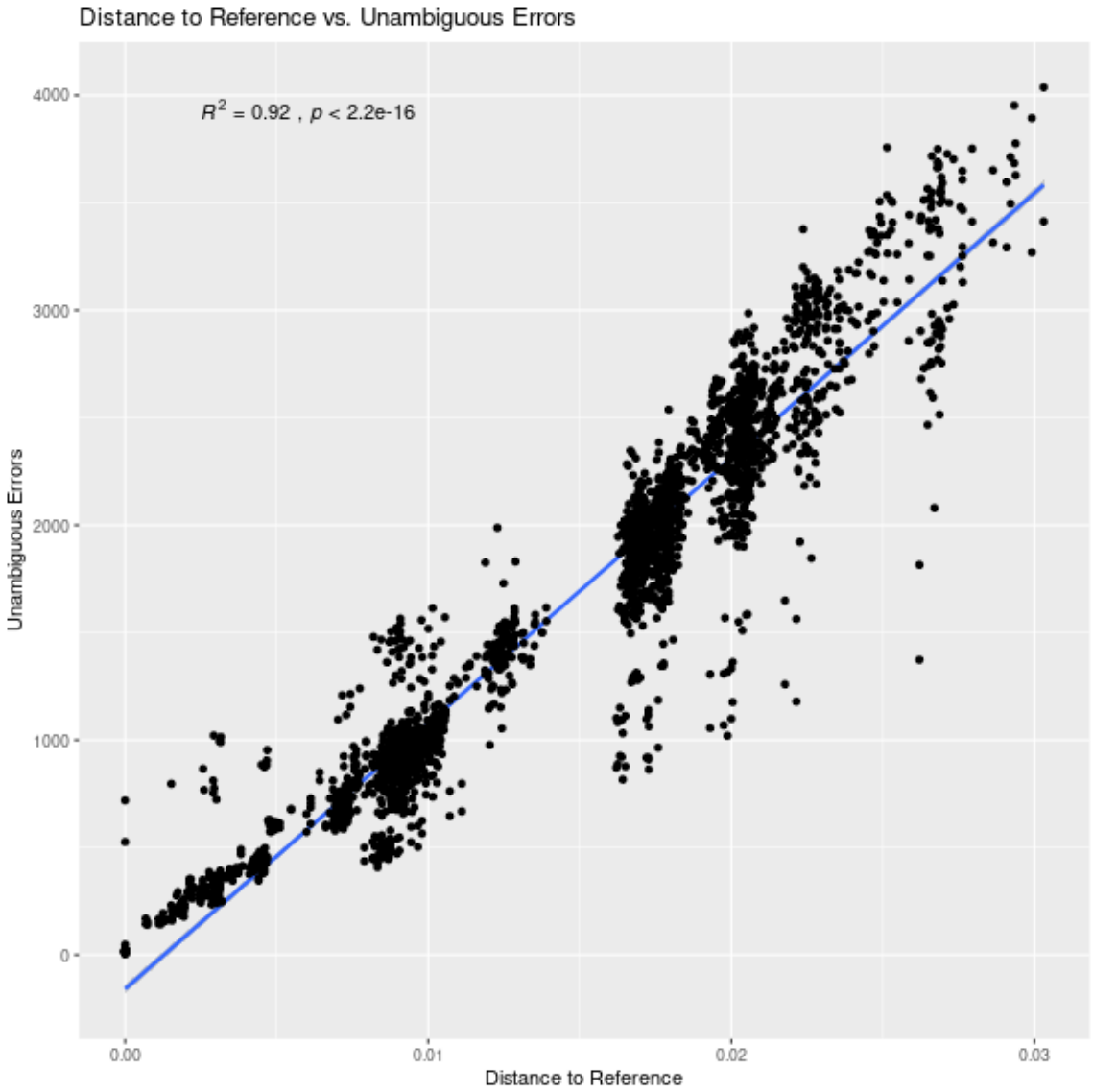

Figure 2.1: Distance to the reference has a significant relationship to the quantity of unambiguous errors. Phylogenetic distance to the reference is based on pairwise distance measurements collected from the Estandia et al (2021) phylogeny. Unambiguous errors only include the four primary nucleotides. The phylogenetic distance to the reference also explains a significant portion of the variation in error rates (R^2 = 0.92).

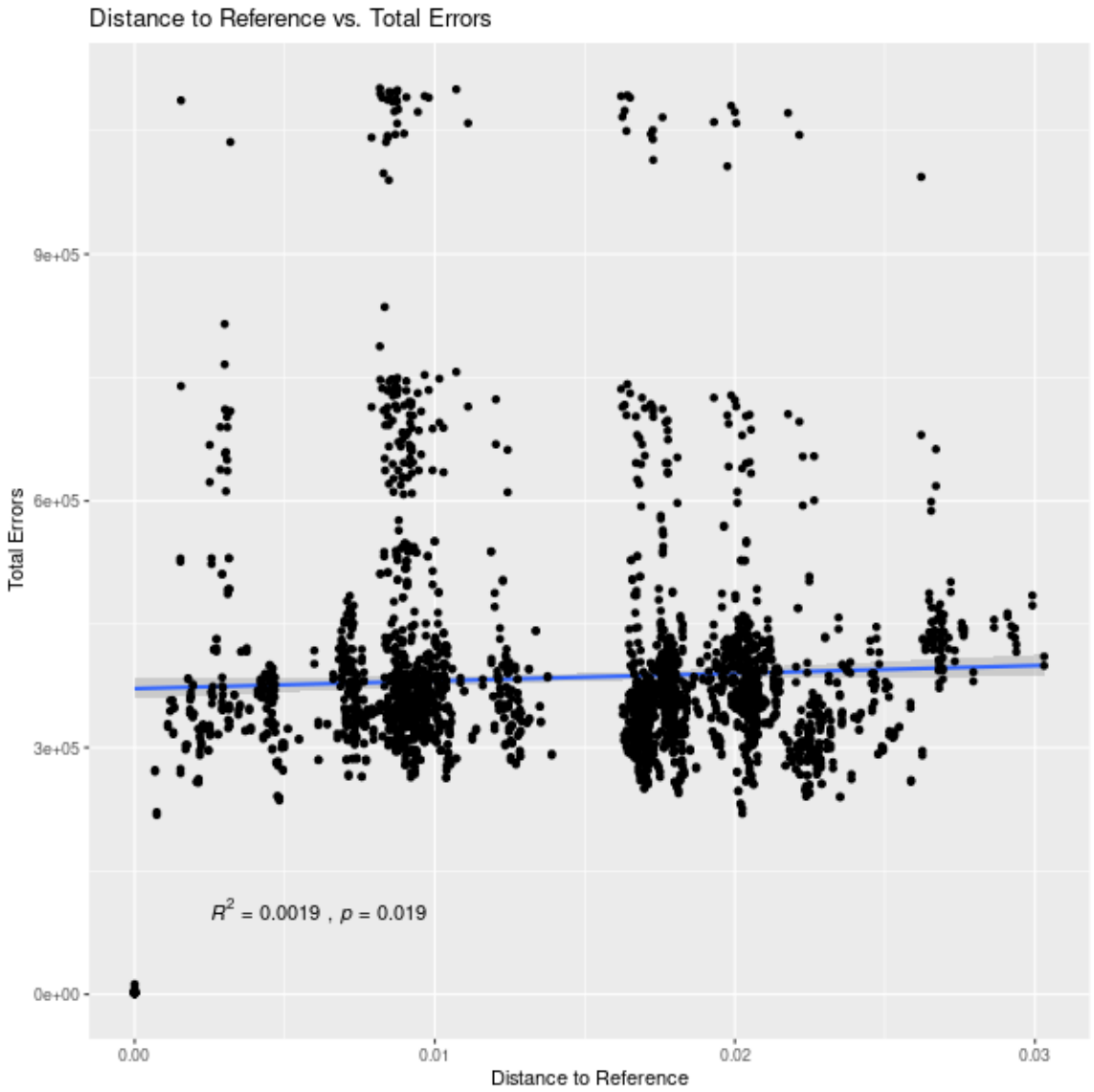

Figure 2.2: Distance to the reference shows no significant effect on ambiguous errors. Phylogenetic distance to the reference is based on pairwise distance measurements collected from the Estandia et al (2021) phylogeny. Ambiguous errors include gaps and ambiguous characters as well as the four nucleotides. Ambiguous errors are not correlated with distance (R^2 = 0.0019, p-value = 0.019)

The third data partition examined the sequence comparisons made by all query taxa assembled using one individual reference taxon. This partition lead to 54 sub-datasets of 54 sequence comparisons, each with a single reference taxon. The individual reference taxon regression analyses used the same predictor variable (phylogenetic distance to the reference taxon) and response variable (unambiguous and total errors) as the whole-dataset analyses. When examining unambiguous errors, the minimum p-value was 1.941e-42 and the maximum p-value was 6.714e-12, with an average of 1.832e-13. The r-squared values

of the reference taxon regression analyses describe that a majority of the variation in unambiguous errors can be explained by examining the phylogenetic distance to the reference. The minimum r-squared value was 0.599 and the maximum r-squared value was 0.971, with an average of 0.897. Of the 54 reference taxon regression analyses, all 54 displayed a positive correlation between distance to the reference genome and errors. The fourth data partition examined the individual taxon comparison values but incorporated sequences belonging to each query taxon regardless of reference choice. This partition lead to 54 sub-datasets of 54 sequence comparisons, each for a single query taxon. The minimum p-value was 1.081e-53 and the maximum p-value was 1.711e-18, with an average of 4.802e-20 The r-squared values for unambiguous error regressions again described that most of the variation in errors can be explained by examining the phylogenetic distance to the reference taxon. The r-squared values for query taxa regression analyses ranged from a minimum of 0.775 to a maximum of 0.990, with an average of 0.954. Of the 54 query taxon regression analyses, 54 showed a positive correlation between distance to the reference genome and unambiguous errors.

Bias to the Reference Base

Across all of our sequence comparisons, we found a total of 4,533,610 unambiguous errors. Of the total unambiguous errors, 4,402,487 matched the reference sequence and 131,123 did not match the reference sequence (Figure 2.3). The one proportion z-test examining the proportion of unambiguous errors matching the reference compared to the expected proportion of errors matching the reference produced a p-value of 2.2e-16. The observed proportion of errors matching the reference was 0.971 compared to our expected value of 0.25. Our test of the independence of individual comparisons separated into values of matching the reference and not matching the reference using a Pearsons chi-squared test returned an identical p-value of 2.2e-16. Both statistical tests reflect a high degree of preference for the reference base in unambiguous errors to the Estandia sequences.

The Effect of Coverage on Errors

When investigating the relationship between sequence coverage and unambiguous errors, we found a range of coverage values at individual unambiguous basecalls from 0x to 315x (Figure 2.4). Error rate was highest at the lowest levels of coverage and displayed an exponential decrease as coverage increased, shown in Figure 2.5. The results of the simple linear regression produced a p-value of 0.006 indicating a significant relationship between coverage and error rate. The R-squared value produced by this analysis was 0.023 and the adjusted R-squared values was 0.019, indicating that only a portion of the variance in error rate could be explained by coverage level. While the fitted linear regression was statistically significant, the distribution is clearly non-linear and bi-modal so we focus on the characteristics of the distribution instead.

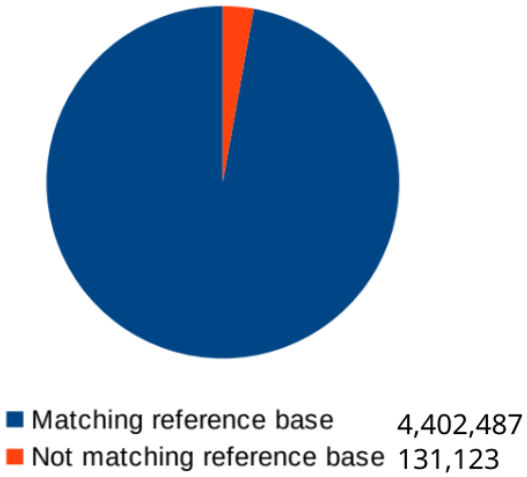

Figure 2.3: The number of unambiguous errors matching the reference genome. Errors match the reference genome more often than any other base.

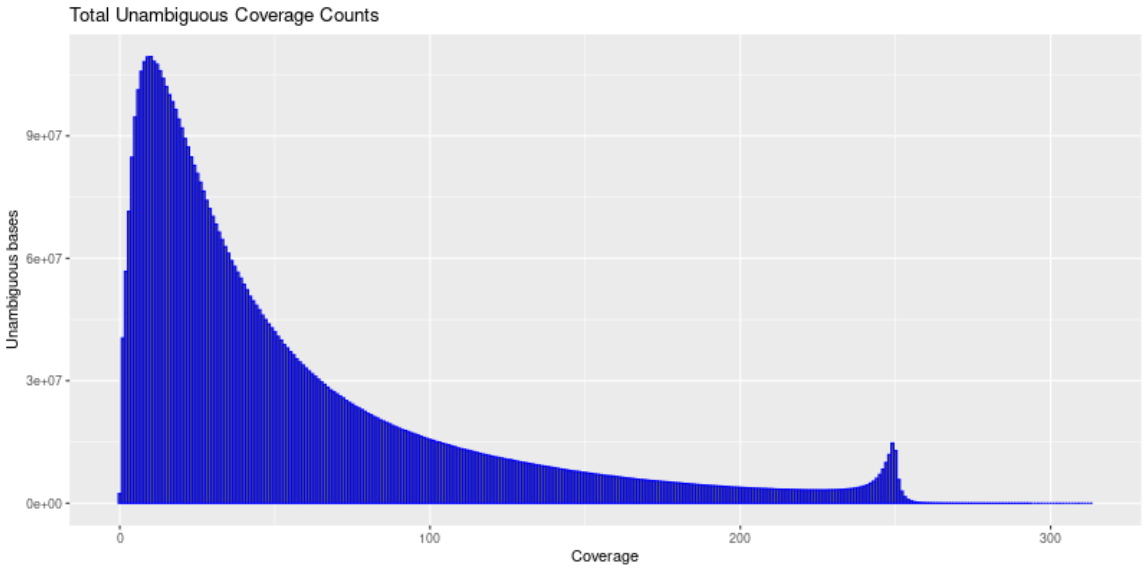

Figure 2.4: The total counts of unambiguous basecalls, both identical and nonidentical, to the sequences in Estandia et al (2021) at each level of coverage.

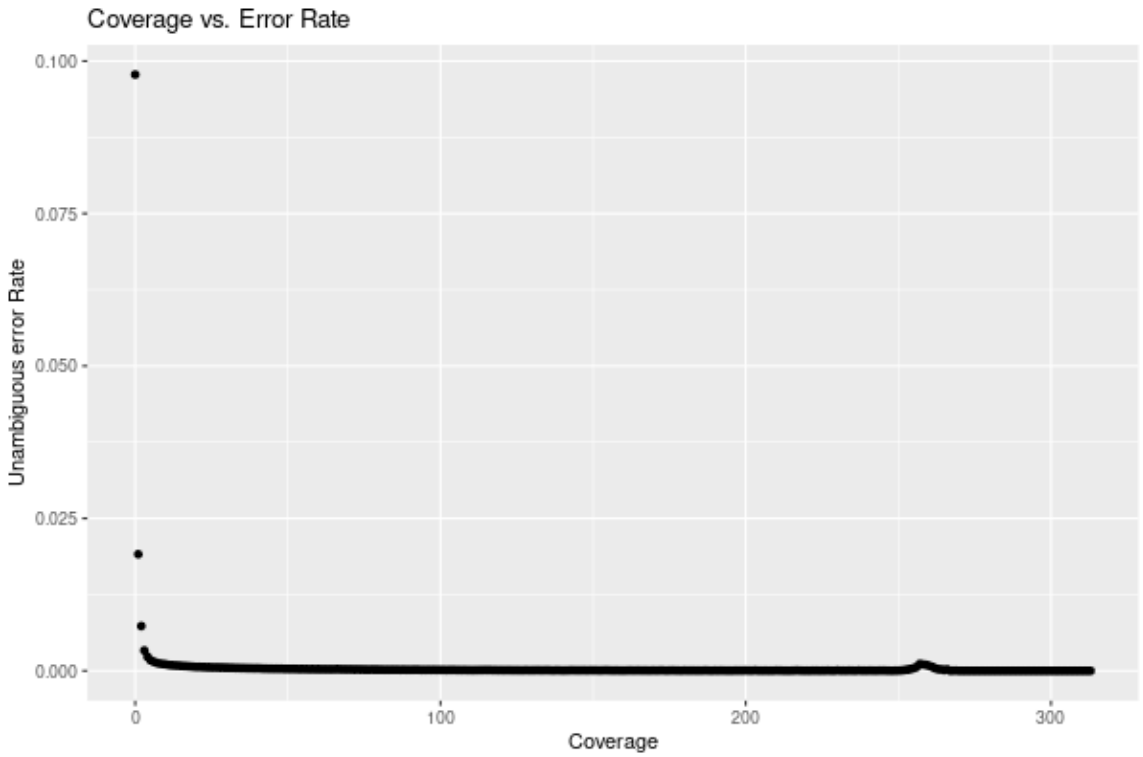

Figure 2.5: The relationship between read coverage at every unambiguous error and the rate of unambiguous errors in all unambiguous sequence comparisons at specified coverage levels. The plateau of error rates at ~25x coverage is in line with previous findings on error rate and coverage. While read coverage and unambiguous error rate are correlated (p-value = 1.3e-06), the coverage level explains relatively little variation in unambiguous error rate (R^2 = 0.073).

Phylogenetic Analyses

We estimated phylogenies using 54 alignments, one for each reference, separated into two datasets: For each data set we estimated the best maximum likelihood tree and the majority rule consensus across 100 bootstrap replicates. We compared these results to the phylogeny published by Estandia et al., (2021) and calculated the Robinson Foulds (RF) distance. The RF distances between the majority rule consensus phylogenies and the Estandia phylogeny were all 0 except for the phylogeny that used one of the outgroups, *Spheniscus demersus*, as a reference sequence, which received an RF distance of 2. The topological difference between the *S. demersus* phylogeny and the Estandia phylogeny involved the rearrangement of the branch leading to *S. demersus* itself to a different position relative to the other two outgroup taxa, specifically sister to *Sula leucaogaster*. The weighted RF distances between the best trees and the Estandia phylogeny were slightly more diverse. We obtained a range of RF distances with a minimum of 0.011, a maximum of 0.048 and an average of 0.013. These values indicate that there are branch length

differences between the reference assembled phylogenies but the topologies remained identical to the Estandia topology. Figure 2.6 displays the original phylogeny produced by Estandia et al., and the most disparate reference-based best tree phylogeny with lines connecting tip positions between the two phylogenies. Our exploration of branch length changes by subtracting the total tree length of the published Estandia et al., phylogeny from each of our reference-based phylogenies indicated a general reduction of branch lengths. However, the tree lengths of trees based on the out-group reference taxa (*S. leucaogaster*, *S. demersus* and *Ciconia Maguari*) were slightly increased over the tree length of the Estandia phylogeny. Other than the three outgroup-based reference phylogenies producing slightly longer phylogenies than the Estandia phylogeny, no meaningful taxonomic pattern is identifiable in the changes to the remaining tree lengths. Figure 2.7 describes these changes to tree length based on reference choice. Notably, the reference-based phylogeny with the largest tree length change was not based on *S. demersus*, the reference taxon that produced the only tree with topological variation but on *Pachyptila desolata*. The *P. desolata* sequence and sequences assembled using taxon as a reference contained the most missing data of any alignment.

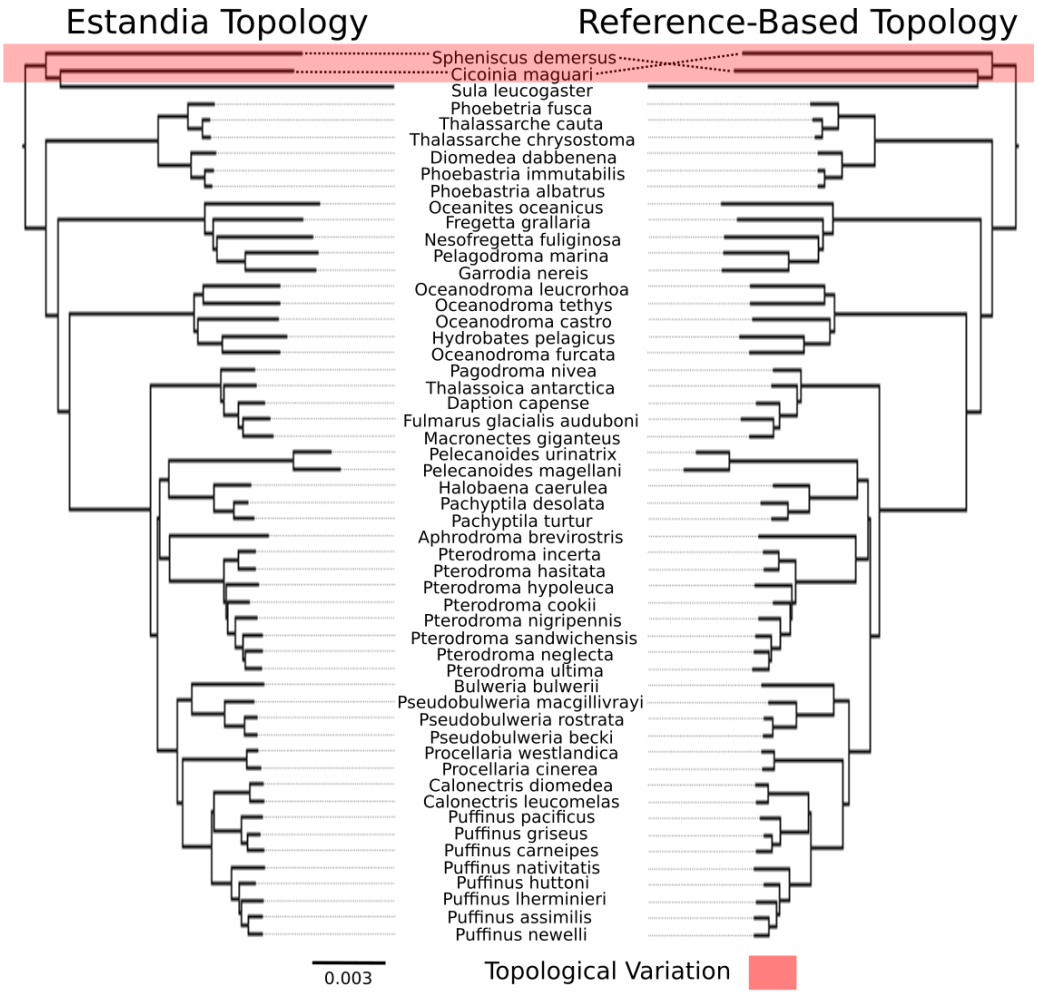

Figure 2.6: Comparison of the topologies of the Estandia published phylogeny and the only reference-based phylogeny with topological differences (based on reference taxon *S. demersus*). The Robinson Foulds distance between these two topologies is 2, reflected by the rearrangement of the branches leading to the outgroup taxa *S. demersus* and *C. Maguari*. The phylogenies estimated from the alignment using *S. demersus* were the only phylogenies to display topological differences to the published Estandia phylogeny.

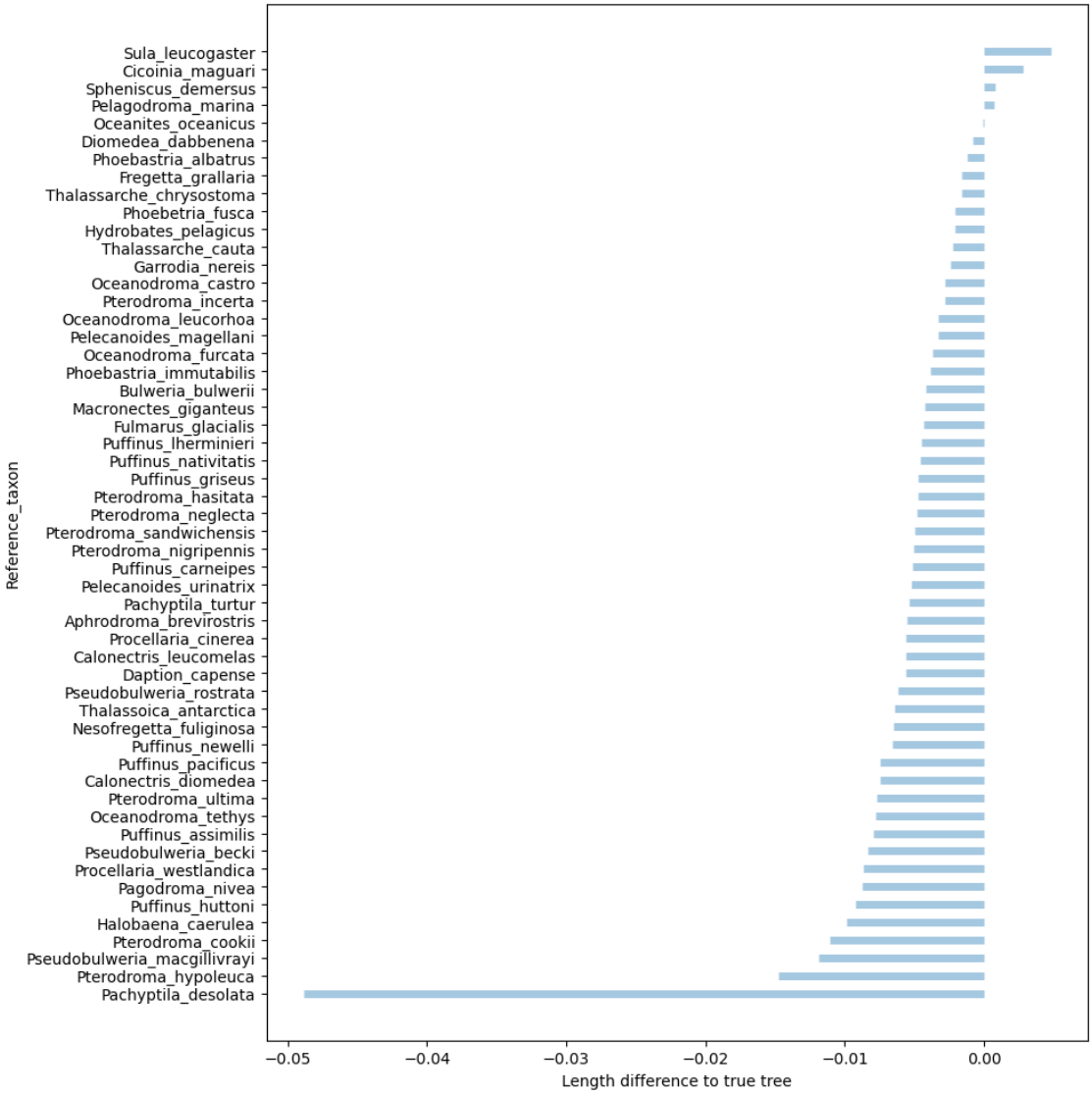

Figure 2.7: Variation in tree lengths of the reference-based phylogenies compared to the Estandia published phylogeny. All three phylogenies estimated using outgroup species as references resulted in a positive change in tree length, along with the tree estimated using Pelagodroma marina as the reference. No taxonomic pattern is visible in the negative change in tree lengths.

2.5 Discussion

Genome sequencing and particularly high throughput sequencing have greatly expanded our ability to investigate many topics in biology (Degner et al., 2009; Bertels et al., 2014). The variety of bioinformatics methods involved in processing high-throughput

sequencing data is extensive. Acknowledging and addressing the caveats of these bioinformatics methods can ensure the reliability of continuing research in biology. Our results quantify the effects of reference bias by focusing on metrics that are directly applicable to any genomic study in the form of basecall errors in relation to several motivating factors. We found that aligning reads to a more phylogenetically distant reference sequence will increase the number of erroneously called bases, with distance to reference as the strongest influence on basecalls. We also found that erroneous bases were highly biased towards the reference bases, regardless of distance to the reference. We demonstrated the relationship between coverage and error rates and while this relationship is less a direct predictor than distance to reference, the result agrees with previously established guidelines. Perhaps soothingly to researchers focusing on phylogenetic estimation as the final stage of evolutionary analyses, we found that despite increased errors at greater distances to the reference, phylogenetic topologies are largely resilient within the distances we investigated. Reference choice had more significant effect on sequence composition and branch lengths rather than affecting the topology itself. Taken together, these results support the previous findings and expectations of previous investigations into the influence of reference choice.

Genomic studies often rely on nuanced analyses of SNP data to investigate a variety of questions, whether focusing on taxon evolution or the effects of individual alleles (Brandt et al., 2015; Rick et al., 2022). Regardless of the application, the accuracy of SNP calls is an important aspect that can be affected by data processing choices prior to final basecalls. Our findings of a linear relationship between errors and distance to the reference support previous research on this subject, indicating that caution must be exercised when choosing reference taxa for sequence alignment. Moreover, as ultra conserved elements are expected to exhibit relatively little variation compared to other regions of an organism's genome, It's probable that a dataset with increased variety of genetic elements would likewise produce more errors when aligned to more distant references. In other studies, basecall errors have been discussed in terms of fluctuations in heterozygosity, with the expectation that as distance to the reference increases, heterozygosity will decrease (Ros-Freixedes et al., 2018; Gunther & Nettleblad. 2019; Prasad et al., 2021; Rick et al., 2022). This skewing of heterozygous sites has been described as less deleterious in cases of large population studies but potentially misleading when the study examines subtle variation, such as investigations of fewer loci (Gunther & Nettleblad. 2019; Prasad et al., 2021; Rick et al., 2022). Interestingly, Prasad et al., (2021) found that heterozygosity can increase as distance to the reference increases, although the proposed explanation is that misalignments are leading to erroneous bases not matching the reference, producing false estimates of diversity at these sites. The results of our investigation of error bias to the reference does not support this supposition as we found a significant majority of erroneous bases matching the reference rather than another possible base. Our results of errors biased to the reference base far exceed expected proportions of 50% found in other studies (Degner et al., 2009; Ros-Freixedes et al., 2018; Gunther & Nettleblad. 2019; Prasad et al., 2021) As we did not find a strong relationship between distance to the reference and coverage, the bias towards the reference bases may explain much of the strong relationship between unambiguous error rate and distance to the reference. When using a closely related

reference sequence, there are fewer differences between the query and the reference. Therefore the reference base is more likely to be the correct base, and that basecall is less likely to appear as an error.

The results of our investigation of the relationship between read coverage and error rates display a significant relationship between coverage and error rate. This finding is in line with expectations set by other studies, as sequence coverage is associated with error rate but is unlikely to be the primary cause of erroneously called bases (Brandt et al., 2015). Our coverage analyses state that while error rates are relatively low regardless of coverage, error rates begin to plateau around coverage values of 20x-25x, which is in line with assessments of the coverage-error relationship in earlier studies (Dohm et al., 2008; Luo et al., 2012; Bertels et al., 2014). Rigorous filtering based on coverage is advisable for studies that require high confidence in individual basecalls. While our results indicate that this issue is a minor influence on error rates in general, the cumulative effect could help explain bias to the reference base. The curious second peak in both unambiguous errors and unambiguous bases in both coverage figures (Figure 2.4, Figure 2.5) are most likely the result of read aggregation from paralogs. Reads from multiple regions in the genome align to the only available locus (Wu et al., 2017).

Ultimately, for all of our findings quantifying the relationships between distance to the reference taxon and coverage to basecall errors, these influences have little effect on the overall topology of our estimated phylogenies. The presence of only one reference based phylogeny with one topological change is encouraging in the case of studies on non-model organisms that do not readily have a reference genome for their taxon. The stability of our estimations could be due in part to the inclusion of non-polymorphic sites as supported by previous findings by Bertels et al., (2014) and ultra conserved elements as the chosen dataset. These results do not indicate, however, that phylogenetic estimation is unaffected by reference choice. The weighted RF distances described minor but informative changes to branch lengths depending on reference choice, connecting the effect of our previous analyses on topological accuracy. Notably, while we cannot definitively state that missing data drives branch length changes, the effect of a reference taxon with a high proportion of missing data provided a strong enough outlier influence to skew the regression analysis of missing data to reference-based tree length change (Figure 2.8). While these changes to branch lengths may be a secondary consideration in an evolutionary study focusing on topological accuracy, the underlying cause of such branch length changes could be critical to a study involving targeted SNP analyses. Anecdotally, we combined sequences from the same taxon but assembled using every sequence in the dataset as a reference and inferred a phylogeny from this “one taxon” dataset. We found that the sequences of the single taxon reconstructed the published topology of the references used to assemble them. This unorthodox analysis underscores the importance of the results of our investigation on errors biased to the reference base and the effect of read coverage on error rates. Taken together, our results show the importance of reference choice on read alignment and sequence composition while also emphasizing the resilience of the correct topology.

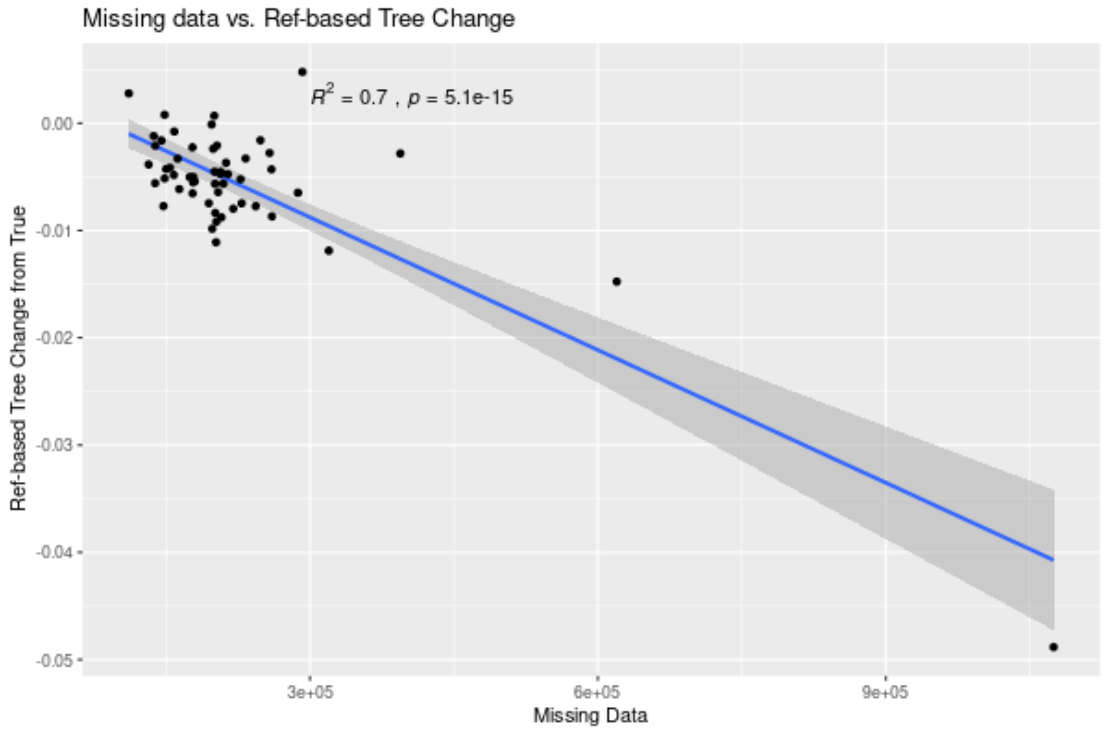

Figure 2.8: The amount of missing data is loosely correlated with total tree length change. The significant correlation is primarily driven by the tree based on the reference taxon with the most negative tree length change as seen in Figure 2.6. The tree with the most negative length change is based on the reference sequence with the most missing data.

While we have confidence in value of our findings, additional work is needed to explore the consistency of our findings in an expanded context. Our dataset consists of only ultra conserved elements from 54 bird species, representing a relatively invariant section of possible genomic regions. As reference-based sequence assembly sees broad use, we expect that future work on the effects of reference choice will need to expand to examine datasets exemplifying the diversity found in the tree of life. Additionally, the use of simulated sequences could add valuable context to our more unexpected results, such as the overwhelming bias towards the reference base in unambiguous errors. Moreover, an incorporation of multiple programs and methods at each stage of data processing workflows coupld help elucidate bias introduced by a single program. Ultimately, while the work we present here is a valuable addition to our understanding of genomic methods, more work would expand context of these results and increase our understanding of downstream results of reference choice.

2.6 Conclusion

Our results offer several quantitative measurements of the effect of reference choice on assembled sequences and phylogenetic estimations. While the choice of a UCE dataset is more conservative than compared to the variation we expect in a whole genome analysis, these findings offer a starting guide of what to expect when aligning short read, high throughput data to a reference sequence. In the context of evolutionary analyses, our more significant finding is that while distance to reference drives errors and ultimately error bias to the reference, phylogenetic estimation is generally resilient to the influence of distant but reasonable reference choices. However, reference choices made during SNP analyses should be considered more carefully as the small but highly biased errors could lead to incorrect assumptions. To summarize, a wider diversity of references is acceptable for phylogenetic estimation while SNP analyses should choose less distant references. Rigorous coverage cutoffs will decrease errors but will also decrease available data.

2.7 References

Balakrishnan, N., Voinov, V., & Nikulin, M. S. (2013). *Chi-squared goodness of fit tests with applications*. Academic Press.

Brandt, D. Y. C., Aguiar, V. R. C., 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*. 11.

Bertels, F., Silander, O. K., Pachkov, M., Rainey, P. B., & van Nimwegen, E. (2014). Automated Reconstruction of Whole-Genome Phylogenies from Short-Sequence Reads. *Molecular Biology and Evolution*, *31*(5), 1077–1088. https://doi.org/10.1093/molbev/msu088

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., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., & Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, *25*(24), 3207–3212. https://doi.org/10.1093/bioinformatics/btp579

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, *36*(16), e105. https://doi.org/10.1093/nar/gkn425

Eaton, D. A. R., Spriggs, E. L., Park, B., & Donoghue, M. J. (2017). Misconceptions on Missing Data in RAD-seq Phylogenetics with a Deep-scale Example from Flowering Plants. *Systematic Biology*, *66*(3), 399–412. https://doi.org/10.1093/sysbio/syw092

Estandía, A., Chesser, R. T., James, H. F., Levy, M. A., Obiol, J. F., Bretagnolle, V., González-Solís, J., & Welch, A. J. (2021). *Substitution Rate Variation in a Robust Procellariiform Seabird Phylogeny is not Solely Explained by Body Mass, Flight Efficiency, Population Size or Life History Traits* (p. 2021.07.27.453752). https://doi.org/10.1101/2021.07.27.453752

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, *17*(6), 368–376. https://doi.org/10.1007/BF01734359

Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, *39*(4), 783–791. https://doi.org/10.2307/2408678

Field, J. T., Abrams, A. J., Cartee, J. C., & McTavish, E. J. (2022). Rapid alignment updating with Extensiphy. *Methods in Ecology and Evolution*, *13*(3), 682–693. https://doi.org/10.1111/2041-210X.13790

Freedman, D. A. (2009). *Statistical models: theory and practice*. cambridge university press.

Gopalakrishnan, S., Samaniego Castruita, J. A., Sinding, M.-H. S., Kuderna, L. F. K., Räikkönen, J., Petersen, B., Sicheritz-Ponten, T., Larson, G., Orlando, L., Marques-Bonet, T., Hansen, A. J., Dalén, L., & Gilbert, M. T. P. (2017). The wolf reference genome sequence (Canis lupus lupus) and its implications for Canis spp. Population genomics. *BMC Genomics*, *18*(1), 495. https://doi.org/10.1186/s12864-017-3883-3

Gordon, A., & Hannon, G. J. (2021). Fastq_toolkit. Retrieved from http://hannonlab.cshl.edu/fastx_toolkit/download.html

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

Heath, T. A., Hedtke, S. M., & Hillis, D. M. (2008). *Taxon sampling and the accuracy of phylogenetic analyses*. Journal of Systematics and Evolution 46.

Heng, L. (2021). Seqtk. Retrieved from https://github.com/lh3/seqtk

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

Leaché, A. D., Banbury, B. L., Felsenstein, J., de Oca, A. nieto-Montes, & 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

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

Luo, C., Tsementzi, D., Kyrpides, N., Read, T., & Konstantinidis, K. T. (2012). Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. *PLOS ONE, 7*(2), e30087. https://doi.org/10.1371/journal.pone.0030087

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). IEEE.

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

O'hara, R. J. (1997). Population thinking and tree thinking in systematics. *Zoologica Scripta*, *26*(4), 323–329. https://doi.org/10.1111/j.1463-6409.1997.tb00422.x

Ovchinnikov, I. V., Götherström, A., Romanova, G. P., Kharitonov, V. M., Lidén, K., & Goodwin, W. (2000). Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature, 404*(6777), 490–493. https://doi.org/10.1038/35006625

Prasad, A., Lorenzen, E. D., & Westbury, M. V. (2021). Evaluating the role of reference-genome phylogenetic distance on evolutionary inference. *Molecular Ecology Resources*, *22*(1), 45–55. https://doi.org/10.1111/1755-0998.13457

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rick, J. A., Brock, C. D., Lewanski, A. L., Golcher-Benavides, J., & Wagner, C. E. (2022). *Reference genome choice and filtering thresholds jointly influence phylogenomic analyses* (p. 2022.03.10.483737). bioRxiv. https://doi.org/10.1101/2022.03.10.483737

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

Rolland, M., Tovanabutra, S., deCamp, A. C., Frahm, N., Gilbert, P. B., Sanders-Buell, E., Heath, L., Magaret, C. A., Bose, M., Bradfield, A., O'Sullivan, A., Crossler, J., Jones, T., Nau, M., Wong, K., Zhao, H., Raugi, D. N., Sorensen, S., Stoddard, J. N., … Mullins, J. I. (2011). Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. *Nature Medicine*, *17*(3), 366–371. https://doi.org/10.1038/nm.2316

Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*(4), 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454

Sprinthall, R. C. (2003). *Basic statistical analysis*. Allyn & Bacon.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313.

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. (1996). *Phylogenetic inference Molecular systematics, (ed. DM Hillis et. Al.), chap 5 pp. 407-514*. Sinauer Associates, Sunderland, Massachusettes.

Van Noorden, R., Maher, B., & Nuzzo, R. (2014). The top 100 papers. *Nature News*, *514*(7524), 550. https://doi.org/10.1038/514550a

Vasimuddin, Md., Misra, S., Li, H., & Aluru, S. (2019). Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 314–324. https://doi.org/10.1109/IPDPS.2019.00041

Wu, S. H., Schwartz, R. S., Winter, D. J., Conrad, D. F., & Cartwright, R. A. (2017). Estimating error models for whole genome sequencing using mixtures of Dirichlet-multinomial distributions. *Bioinformatics*, *33*(15), 2322–2329. https://doi.org/10.1093/bioinformatics/btx133

Wu, S. H., Schwartz, R. S., Winter, D. J., Conrad, D. F., & Cartwright, R. A. (2017). Estimating error models for whole genome sequencing using mixtures of Dirichlet-multinomial distributions. *Bioinformatics*, *33*(15), 2322–2329. https://doi.org/10.1093/bioinformatics/btx133

Zwickl, D. J., & Hillis, D. M. (2002). Increased Taxon Sampling Greatly Reduces Phylogenetic Error. *Systematic Biology*, *51*(4), 588–598. https://doi.org/10.1080/10635150290102339

Chapter 3

Elucidating the Evolutionary Relationships of Anti-Microbial Resistant *Neisseria gonorrhoeae* with Phylogeny-Aware Database Filtering

Field, J. T. & McTavish, E. J.

3.1 Abstract

Monitoring efforts focusing on anti-microbial resistant *Neisseria gonorrhoeae* are crucial for making informed policy decisions and alleviating community burden. Phylogenetic methods are a key aspect of exploring the evolutionary relationships of anti-microbial resistant pathogens. Methods of phylogenetic estimation must contend with large amounts of sequencing data produced by contemporary monitoring efforts. One such monitoring program is the NCBI? Pathogen Detection database. The Pathogen Detection database utilizes a data processing pipeline that applies kmer matching, sequence-sequence similarity approaches and maximum compatibility methods of phylogenetic estimation. While these methods are established and generally reliable, alternative methods may offer some benefits. Here, we describe an alternative approach that employs efficient and flexible reference guided sequence assembly and full maximum likelihood phylogenetic estimation methods to update a clade-of-interest from a previous study with new data. We further explore our dataset by examining the topologies of lineage clusters found throughout our chosen clade. We complete our investigation by inferring dated phylogenies for sub-clades of our dataset and overlaying anti-microbial resistance profiles that would be standard inclusions in any effort monitoring pathogen evolution.

3.2 Introduction

Anti-microbial resistant bacteria are an increasing threat to contemporary human society. Anti-microbial resistant gonorrhea, caused by the bacterium *Neisseria gonorrhoeae*, is a pathogen of highest concern to monitoring organizations (CDC. 2019). *N. gonorrhoeae* infections have been connected to increased infertility, ectopic pregnancy and play a facilitative roll in the transmission of human immunodeficiency virus (St. Cyr et al., 2020). Regarding treatment, the recommended dose of some antibiotics by the United States Centers for Disease Control and Prevention (CDC) has doubled since 2010 (St. Cyr et al., 2020). The number of anti-microbial resistant lineage transmissions were estimated to be 550,000 each year in 2019 (CDC. 2019). Efforts to monitor anti-microbial resistant *N. gonorrhoeae* lineages have been successful in guiding internationally collaborative policy, as evidenced by a number of international health organizations adopting similar policies (Unemo et al., 2019). Despite successful efforts to trace infections and set policies, pathogen evolutionary dynamics combined with growing populations ensure transmissions will continue. Monitoring efforts have increasingly turned to genomic data to investigate the life histories of collected *N. gonorrhoeae* and contemporary efforts produce prodigious quantities of raw data. Taken together, these data indicate a continued need for extensive pathogen surveillance efforts aimed at *N. gonorrhoeae*.

Phylogenetic analyses have played an increasingly important role in monitoring *N. gonorrhoeae*, primarily due to the field's utility in assessing patterns of antibiotic resistance through a shared life history (Grad et al., 2014; Grad et al., 2016; Harris et al., 2018). phylogenetic analyses are sensitive to emerging lineage's methods of anti-microbial resistance mechanisms such as horizontal gene transfer or clonal expansion through examinations of presence and absence of associated genes (Grad et al., 2016). More broadly, the identification of anti-microbial isolates in phylogenomic studies can elucidate transmission and distribution information, indicating potential driving forces in at-risk groups (Harris et al., 2018). Platforms such as Nextstrain have highlighted the importance of large data driven efforts in tracking the evolution of pathogens with its coverage of multiple viral pathogens including Covid-19 (Hadfield et al., 2018; Hodcroft et al., 2021). However, Nextstrain does not currently monitor *N. gonorrhoeae* evolution, a task generally performed by individual research groups in conjunction with government-backed monitoring programs (Grad et al., 2014; Grad et al., 2016; Harris et al., 2017; Papp et al., 2017). Recent studies have investigated the evolution of the penA gene, a gene associated with resistance to cephalosporin anti-microbials, over a specified time period and often in a specific geographic region (Whiley et al., 2018; Yahara et al., 2021). While studies almost always include penA as a gene-of-interest, other anti-microbial resistance genes are often investigated, displaying an awareness of historical treatment methods when examining the evolutionary history and origins of modern lineages (Unemo & Shafer. 2011; Yahara et al., 2021). Extensive historical research on anti-microbial resistance lineages of *N. gonorrhoeae* were ultimately successful in suggesting plausible entrance paths into the United States (Unemo & Shafer. 2011).

Efforts to investigate *N. gonorrhoeae* evolution in the context of anti-microbial resistance emergence and exchange are a continuing effort. Our recent work on the

software package Extensiphy produced a genome scale multi-locus phylogeny for 1,237 anti-microbial lineages (Field et al., 2022). Our data from this project was a semi-random selection of samples from the *N. gonorrhoeae* repository of the NCBI Pathogen Detection database (NCBI. 2022a). While no anti-microbial genes were specifically incorporated into our original dataset, the detailed metadata on the Pathogen Detection website describes a variety of pertinent information on the presence of particular anti-microbial genes. Beyond the metadata for the 1,237 samples we have already analyzed, the Pathogen Detection database contains sample information for more than 24,000 additional *Neisseria* isolates. Such large datasets present a challenge for contemporary phylogenetic analysis methods, particularly if the studies goal is to incorporate all samples in the dataset into a single, resolved phylogenetic estimation. The Pathogen Detection project addresses the issue of efficient data processing with a four stage pipeline of sequence assembly, sequence clustering, phylogenetic estimation and sequence annotation (NCBI. 2022). While the Pathogen Detection pipeline is ideal for adding new samples of high similarity to the extant database and identifying anti-microbial resistance alleles, phylogenetic inference is made using maximum compatibility (Cherry. 2017; NCBI. 2022). The maximum compatibility algorithm used by Pathogen Detection supports the rapid calculation of relationships between clustered sequences but intentionally avoids analyzing relevant biological data such as sites with more than two nucleotide variants and time reversibility (Cherry. 2017). With these limitations in mind, we decided to evaluate the similarity between the maximum compatibility trees produced by Pathogen Detection and trees estimated using full maximum likelihood methods. To make this comparison, we investigated the evolutionary relationships of a single clade selected from our Extensiphy phylogeny and expanded with all applicable samples from Pathogen Detection using full maximum likelihood methodologies. Samples were automatically downloaded, assembled and stored in a database over several weeks using a new software package, Intensiphy. Following clade updating, we investigated the relationships of the clade in a temporal context by building a dated phylogeny. Finally, we examined the distribution of anti-microbial resistance genes throughout the clade.

3.3 Methods

2.1 Software Overview

Intensiphy is a software package for automatically downloading datasets from NCBI and incorporating them into a database with a focus on continuous updating. Intensiphy accepts a CSV formatted file of NCBI short read archive (SRA) numbers, a concatenated starting alignment and a phylogeny based on the concatenated alignment. Intensiphy automatically detects if the current run is a continuation of a previous run or the start of a new run of the program. If the current mode is a new run, Intensiphy separates the input alignment into individual sequence files, constructing the beginning of the sequence database. A reference sample is chosen at random or by user selection. Intensiphy then reads the input CSV file and identifies any samples not already in the sample database.

The stages of downloading data and assembling the sequences are intertwined, with Intensiphy downloading data for a subset of the complete sample set and using Extensiphy to align the newly downloaded data to the reference sample. The resulting consensus sequences are added to the database, the intermediate files are removed and the system downloads the next batch of samples data. Once all samples listed in the CSV have been assembled and added to the sequence database, the new samples are placed into the starting phylogeny. Once placement tree inference is complete, the Intensiphy run is complete and the database is ready for the next run. The user may construct an alignment of all samples in the database using included script options at any time.

Clade Selection

The starting point for our analyses was the phylogeny produced in our manuscript describing the Extensiphy software package (Field et al., 2022). This phylogeny described the relationships of 1,237 *N. gonorrhoeae* isolates collected from the Pathogen Detection database. Using the python phylogenetic manipulation library Dendropy (Sukumaran & Holder. 2010), we selected a clade from the Extensiphy phylogeny where the number of tips is between 50 and 150, the highest average number of anti-microbial resistance genes between all tips in the clade, the fewest tips with missing location data and the fewest tips missing from the Pathogen Detection database. Once the clade best fitting these criteria was identified, we pruned this clade sub-tree from the Extensiphy manuscript tree and isolated the sequences associated with the clade from the alignment used to infer the Extensiphy manuscript tree.

Expanded Dataset

We obtained the complete metadata of all *Neisseria* samples in the Pathogen Detection database, which on the date of March 29th 2022 contained 24,571 isolates. All *Neisseria* samples included in the Pathogen Detection database were sequenced across the whole-genome, making every sample suitable for Extensiphy and Intensiphy use. Following metadata collection, we filtered the dataset by removing samples that lacked associated NCBI Short Read Archive (SRA) numbers and location metadata. Once filtered on these criteria, the dataset included 16,110 isolates. The format of Pathogen Detection metadata tables are acceptable by Intensiphy by default. The reference sequence used for read alignment and sequence assembly was the same reference sequence used for our work on Extensiphy. This sequence contains concatenated loci found in the NCBI reference *N. gonorrhoeae* sequence (NCBI Genome ID: 864). More details on the selection of this sequence are available in the Extensiphy manuscript (Field et al., 2022).

Read Collection and Sequence Assembly

Intensiphy uses fasterq-dump from the NCBI SRA toolkit collection of programs to download raw read files directly from the NCBI short read archive (SRA) (NCBI. 2022b). Intensiphy downloaded each sample in the dataset and passed the raw read files to

Extensiphy for read alignment and assembly based on the reference sequence (Field et al., 2022). Extensiphy builds or updates sequence alignments with new, whole-genome sequencing data. Extensiphy aligns a sample's raw reads to a reference sequence and calls a consensus sequence. The output consensus sequence is the same length and has the same gap positions as the reference sequence, ensuring synteny. Once all samples have been assembled in this manner, the sequences are combined with the original alignment or reference sequence to form a new, extended alignment. In this version of Extensiphy, read alignment is performed by BWA-MEM2 (Vasimuddin et al., 2019), the aligned read files are indexed by Faidx (Li et al., 2009) and sequence alignment files are converted to binary alignment files by Samtools View (Li et al., 2009). Variant nucleotide calling is performed by the Bcftools Mpileup (Li et al., 2009), which outputs a Variant Call File (VCF) (Danecek et al., 2011). Vcfutils.pl (Gordon & Hannon. 2021) then converts the VCF to a fastq format file and seqtk converts the fastq file to a fasta (Heng. 2021). A more in depth description of Extensiphy and the utility of the program are discussed in Field et al., (2022). Once assembled, the sequences were stored in the sequence database for this project.

Updated Clade Isolation and Phylogenetic Estimation

Once all sequences were assembled and added to the sequence database, we estimated a resolved phylogeny from the chosen clade-of-interest using RaxML (Stamatakis. 2014). We used the GTRGAMMA model and 100 bootstrap replicates to produce a majority rule consensus phylogeny. The consensus phylogeny was used as the starting tree for the phylogenetic placement of all new samples in the Intensiphy sequence database. Placement was performed using the RAxML Evolutionary Placement Algorithm (EPA) (Stamatakis. 2014), also using the GTRGAMMA model. As the goal of this analysis was only to identify which samples were placed within the clade-of-interest and not to assess overall evolutionary relationships, batches of 500 samples from the sequence database were combined with the original clade-of-interest. After samples were placed in the starting clade-of-interest phylogeny, we identified the new samples placed in the clade-of-interest of each batch and pulled their sequences from the sequence database, creating an alignment containing only sequences in the clade-of-interest. This updated alignment was used to estimate a resolved phylogeny using RaxML using the same settings described above.

Phylogenetic Comparisons

Once clade estimation was complete, we identified all Pathogen Detection single nucleotide polymorphism (SNP) clusters found in the updated clade-of-interest. For disambiguation purposes, we will refer to NCBI SNP clusters as lineage clusters, and we will specify if the lineage cluster in question was collected from NCBI or from our own phylogenetic estimations (COI). The NCBI lineage cluster phylogenies and associated metadata were collected from Pathogen Detection for comparison to our results. We used Dendropy to identify the most recent common ancestor (MRCA) of all samples in each NCBI lineage cluster. Once the MRCA was identified, we isolated all samples descended

from the MRCA, including samples not included in the original NCBI lineage cluster dataset. Sub-phylogenies for these datasets were pruned from the updated clade-of-interest phylogeny and set aside for analyses and comparisons, hereafter referred to as clade-of-interest (COI) lineage clusters. We then compared the NCBI lineage cluster tree to the COI lineage cluster sub-tree. If the COI lineage cluster was found to be monophyletic, we calculated the Robinson-Foulds (RF) distance between both trees for each lineage cluster using Dendropy (Sukumaran & Holder. 2010). We ignored NCBI lineage clusters that contained three or fewer isolates.

Dated Phylogeny Estimation

We estimated a dated phylogeny for every COI lineage cluster found in our updated clade-of-interest phylogeny using the TreeTime software package (Sagulenko, Puller & Neher. 2018). We provided TreeTime with sample identification and date metadata, along with a starting phylogeny and the original alignment used to estimate the phylogeny. Sample collection date and SRA run metadata were collected from the Pathogen Detection database for each sample in the updated clade-of-interest. Collection dates are an included category of Pathogen Detection metadata but not all samples have recorded collection date values. Dated phylogenies were also estimated for the starting clade-of-interest and the updated clade-of-interest.

Anti-Microbial Resistance Overlay

Anti-microbial resistance gene profiles were collected from the metadata included for each sample in the clade-of-interest. The anti-microbial resistance gene data was reformatted to a presence-absence heat map and overlaid on any applicable COI lineage cluster phylogenies using the Interactive Tree of Life tree manipulation platform (Letunic. & Bork. 2021). To examine overarching dynamics of anti-microbial resistance genes found throughout the starting and updated clade-of-interest phylogenies, the total counts of each anti-microbial resistance gene found in the starting and updated clade were summed and visualized.

3.4 Results

Updated Clade-of-Interest Dataset

The clade-of-interest started with 69 samples, excluding two outgroup lineages, prior to updating (Figure 3.1). Upon completion of updating with samples from NCBI Pathogen Detection, the clade-of-interest contained 1,092 samples and two outgroup lineages for a total of 1,094 samples (Figure 3.2). Sequences in the dataset all matched the reference length of 1,859,910 bases, comprising 317 loci. Pathogen Detection metadata dated these samples as being collected between 2016 and 2021. The starting clade-of-interest contained an average of 12 anti-microbial resistance genes per isolate prior to

updating and contained an average of 14 anti-microbial resistance genes after updating. There were samples corresponding to 6 NCBI lineage clusters in the starting clade-of-interest and samples from to 33 NCBI lineage clusters in the updated clade-of-interest. Of these 33 COI lineage clusters, we identified 15 COI lineage clusters containing three or more lineages (Figure 3.3). Of these 15 COI lineage clusters, two were recovered with all isolates listed in the corresponding Pathogen Detection NCBI lineage clusters and as monophyletic clades: NCBI lineage clusters PDS000104762 and PDS000104763, labeled in the clade-of-interest as COI cluster 4 and COI cluster 5. The remaining 13 COI lineage clusters were all found to be paraphyletic and a range of lineages included in the corresponding NCBI lineage clusters were not included in the updated clade-of-interest, having been placed outside of the clade-of-interest during the placement phase of the workflow. The number of isolates found in each COI lineage cluster was variable for both the starting and updated clade-of-interest. In the starting clade-of-interest, 5 COI lineage clusters had 6 or fewer associated samples and the remaining single COI lineage cluster contained 55 isolates. In the updated clade-of-interest, the smallest COI lineage clusters were represented by 1 sample, and the largest COI lineage cluster containing 401 isolates.

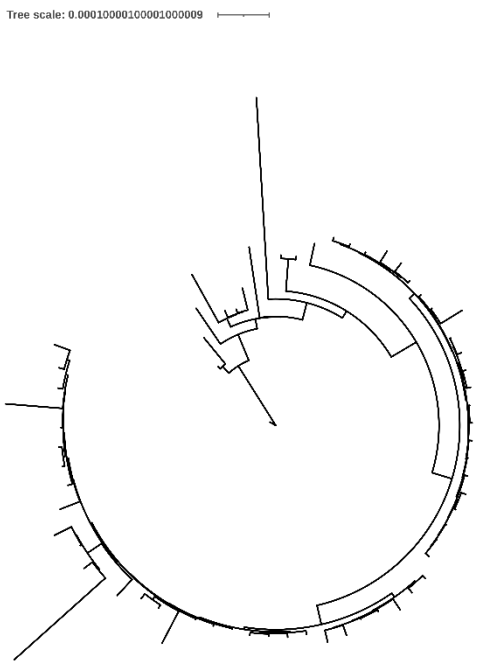

Figure 3.1: Starting clade-of-interest. This version includes the original 69 isolates and does not include the two outgroups.

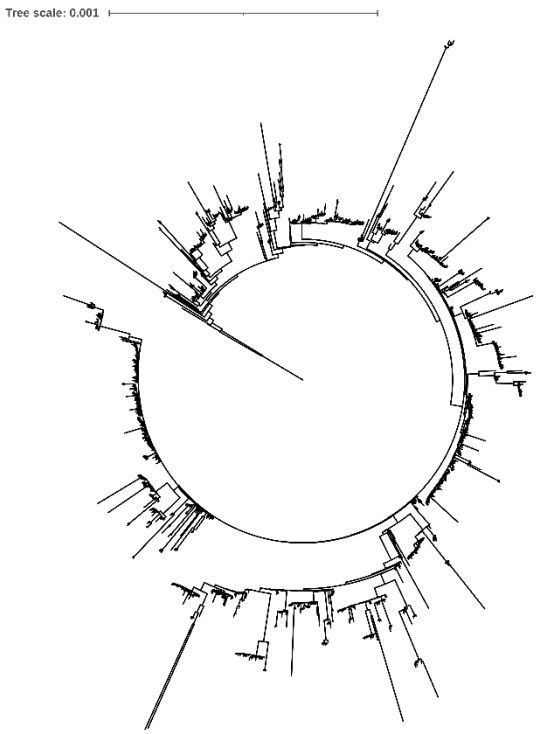

Figure 3.2: Updated clade-of-interest. Includes 1,092 isolates.

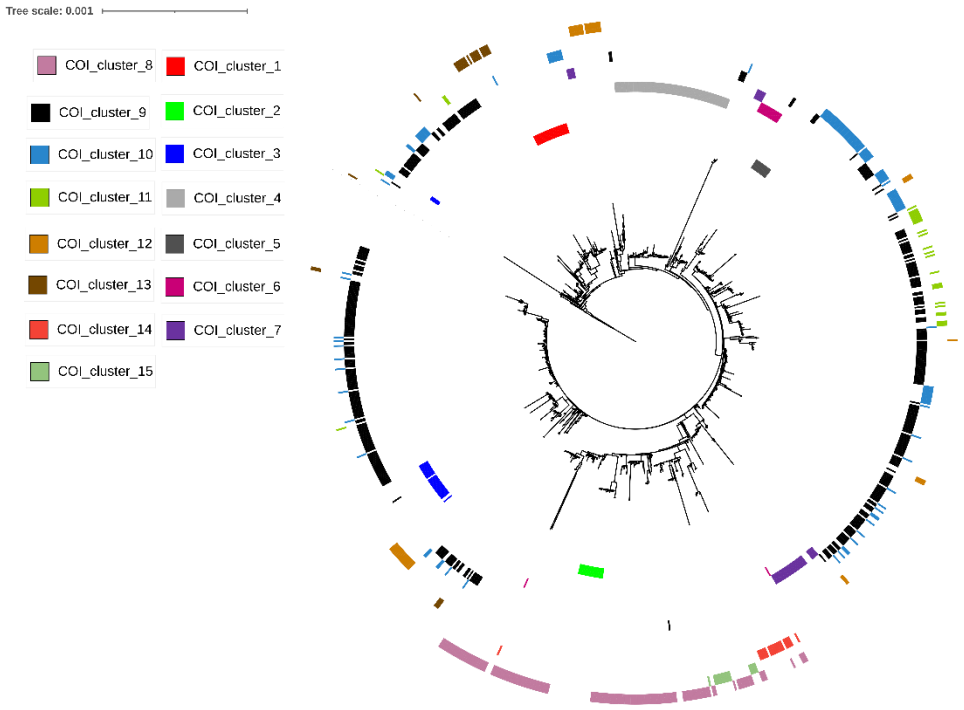

Figure 3.3: Updated clade-of-interest with Pathogen Detection SNP cluster locations displayed as separate colors. The two lineage clusters found with all lineages listed in the NCBI lineage cluster and as monophyletic clades are COI cluster 4 and COI cluster 5, individually displayed in Supplemental Figure 4 and Supplemental 5, respectively.

SNP Cluster Phylogenetic Comparisons

When making comparisons between the NCBI lineage clusters and our COI lineage clusters, we first examined the number of shared isolates. Of the 15 comparisons, we found that two lineage clusters had the same set of isolates between the NCBI version and the COI version (Table 3.1). These two lineage clusters contained 81 and 15 isolates and comparisons of the NCBI and COI versions returned RF distances of 16 and 5, respectively. Of the other 13, the range of isolates found in the NCBI lineage clusters was between 7 and 435. In 11 of the remaining 13 cluster comparisons, the differences between the number of isolates listed in the NCBI clusters and the number isolates found in the COI clusters ranged from 1 to 34 isolates. In the two remaining cluster comparisons with the most

extreme variation in found isolates in the COI cluster compared to the NCBI cluster, the NCBI clusters included 60 and 90 more isolates than the COI clusters (Table 3.1).

Table 3.1: Clade-of-interest lineage cluster descriptors. Two NCBI lineage clusters were recovered in our updated clade-of-interest with all listed samples.

Lineage Cluster NCBI ID	Count of lineages in NCBI cluster	Count of lineages in updated COI cluster	Lineages removed during COI placement step	Monophyletic in updated COI tree?	RF distance (if applicable)
PDS000104762	81	81	0	Yes	16
PDS000104763	15	15	0	Yes	5
PDS000062540	28	19	9	No	N/A
PDS000099711	33	32	1	No	N/A
PDS000104765	34	18	16	No	N/A
PDS000104770	7	6	1	No	N/A
PDS000105692	130	40	90	No	N/A
PDS000107814	435	401	34	No	N/A
PDS000107815	132	128	4	No	N/A
PDS000107816	30	29	1	No	N/A
PDS000107817	50	42	8	No	N/A
PDS000107819	38	25	13	No	N/A
PDS000109193	30	18	12	No	N/A
PDS000109194	16	14	2	No	N/A
PDS000109195	191	131	60	No	N/A

Dated Phylogeny Investigation

We estimated dated phylogenies for our updated clades-of-interest. The updated dated phylogeny included tips with estimated collection dates set in the future (Figure 3.4). Tip dates were considered misdated if the year date reported by TreeTime differed from dated reported in the Pathogen Detection metadata. We considered any dated phylogenies with any misdated collection dates as unreliable for additional analyses. We inferred dated phylogenies for the COI lineage clusters found in our updated clade-of-interest. Of the 15 lineage clusters for which we inferred dated phylogenies, two lineage clusters produced dated phylogenies without misdated tips or internal nodes. The NCBI IDs for these clusters were PDS000048684.14 and PDS000104763.1, which we have relabeled as lineage cluster 1 and lineage cluster 5 to clarify that these lineage clusters are found in our updated clade-of-interest and reflect the toplogies found there (Figures 3.5, 3.6). Lineage cluster 1 included samples with collection dates estimated between 2017 and 2021, with a clade root delineation set in 2011 (Figure 3.5). Lineage cluster 6 included samples with collection

dates estimated between 2018 and 2019, with a clade root delineation set in late 2017 (Figure 3.6).

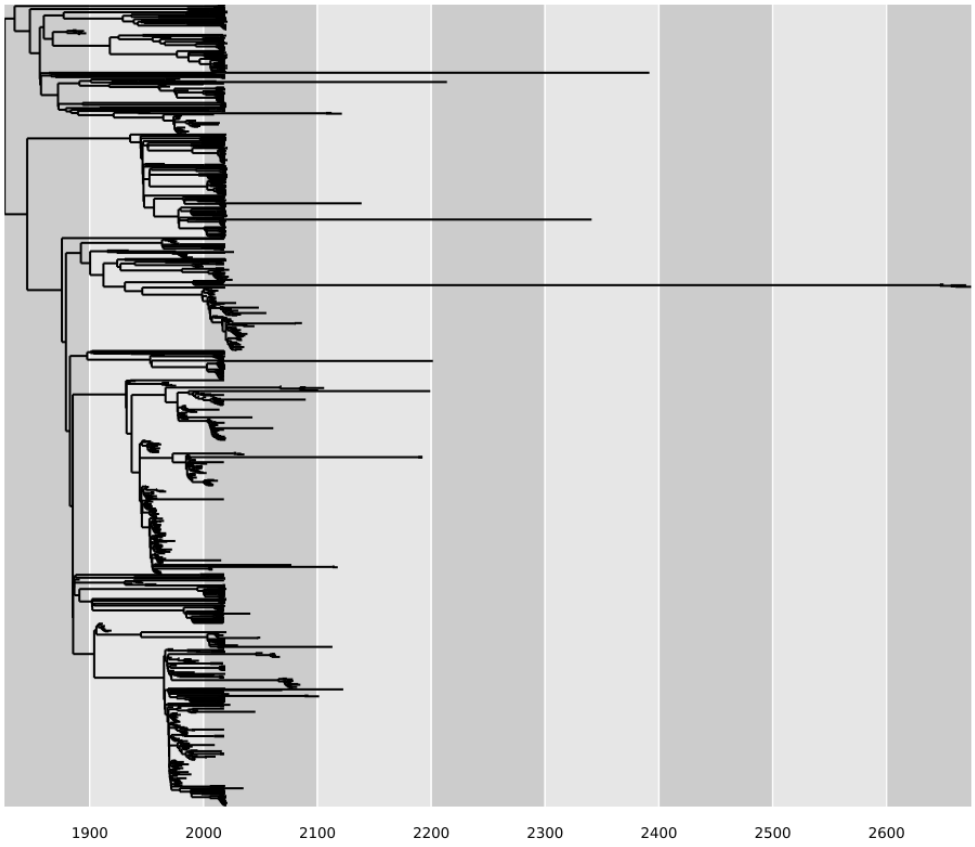

Figure 3.4: Dated phylogeny estimated for the updated clade-of-interest. Several branches can be seen as dated many years into the future, the longest of which has a collection date in the Pathogen Detection metadata of October 2018 but is dated in this figure almost 600 years in the future.

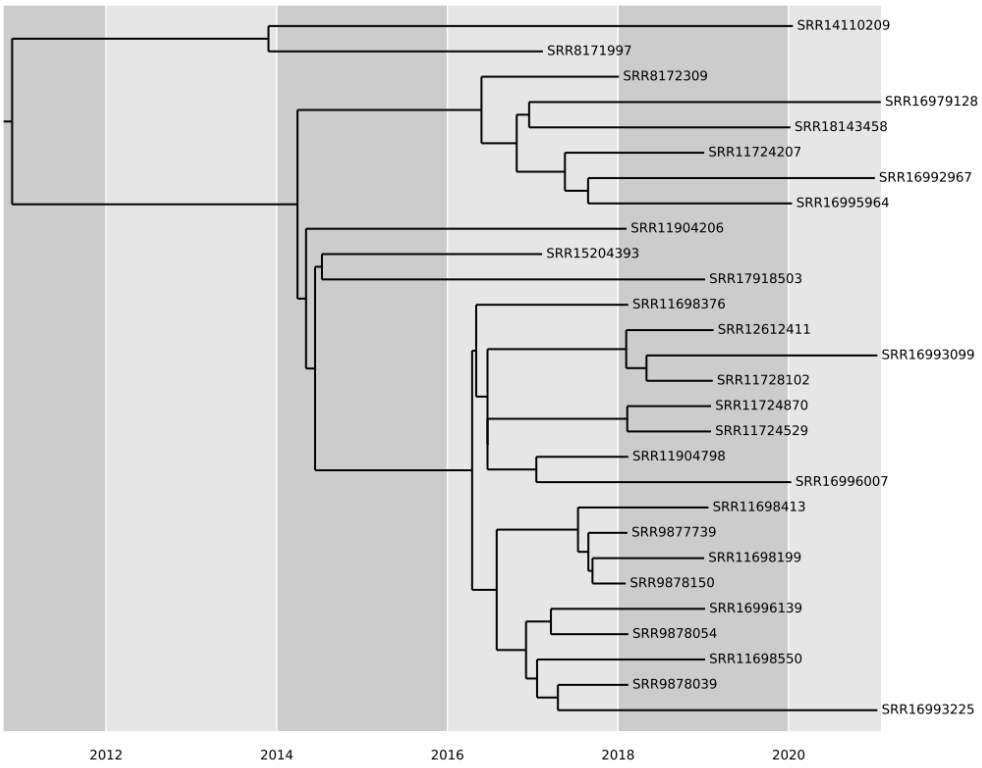

Figure 3.5: COI lineage cluster 1 dated phylogeny. All dates that were provided by the Pathogen Detection database metadata are accurate in this dated phylogeny.

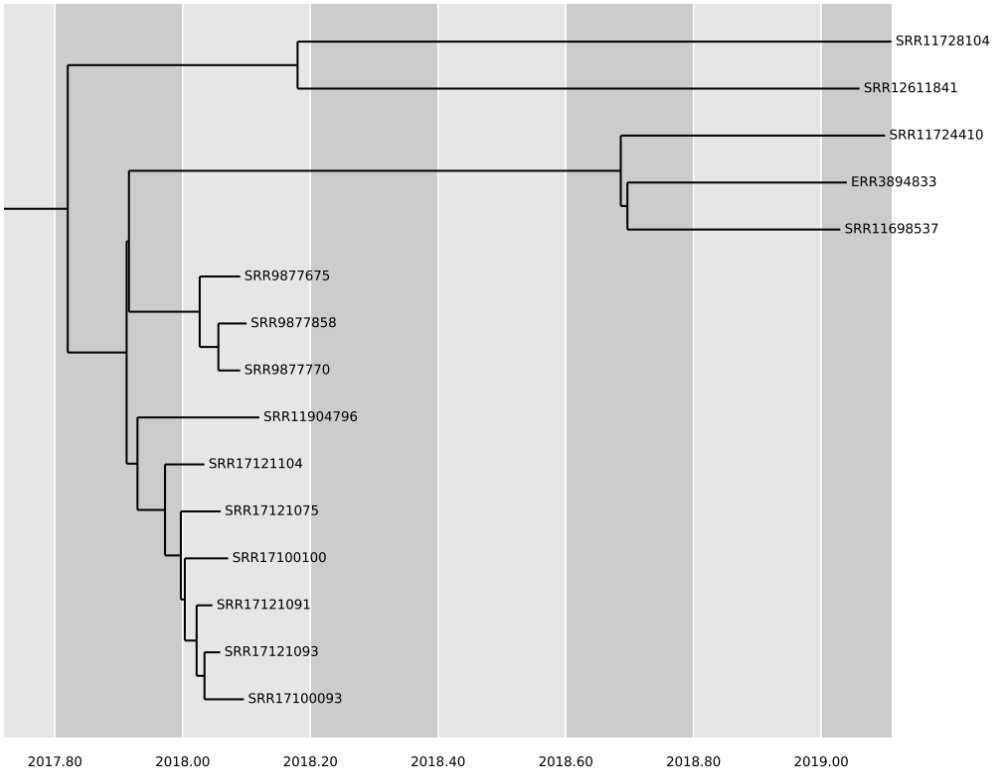

Figure 3.6: COI lineage cluster 5 dated phylogeny. All dates that were provided by the Pathogen Detection database metadata are accurate in this dated phylogeny. This lineage cluster was also one of two clusters in the updated clade-of-interest that contained all the samples listed in the NCBI lineage cluster.

Anti-Microbial Resistance Overlay

We collected anti-microbial resistance gene profiles from the Pathogen Detection metadata associated with each sample in the two COI lineage clusters for which we obtained reliable dated phylogenies. We constructed Interactive Tree of Life annotation files for the entire updated clade-of-interest and applied the annotation files to the three dated phylogenies. Both lineage clusters displayed high retention of anti-microbial genes throughout each cluster (Figures 3.7, 3.8). Isolates in lineage cluster 1 include 16 anti-microbial resistance genes, 14 of which are found in all isolates in the cluster. Of the two anti-microbial resistance genes that were not ubiquitous, one was not observed in one isolate and the other was missing from two isolates (Figure 3.7). The isolates were not sister to each other and the bifurcation leading to these samples occurred in 2011. Isolates

in lineage cluster 6 were found to possess 16 anti-microbial resistance genes. All 16 resistance genes were found in all the isolates of the lineage cluster (Figure 3.8).

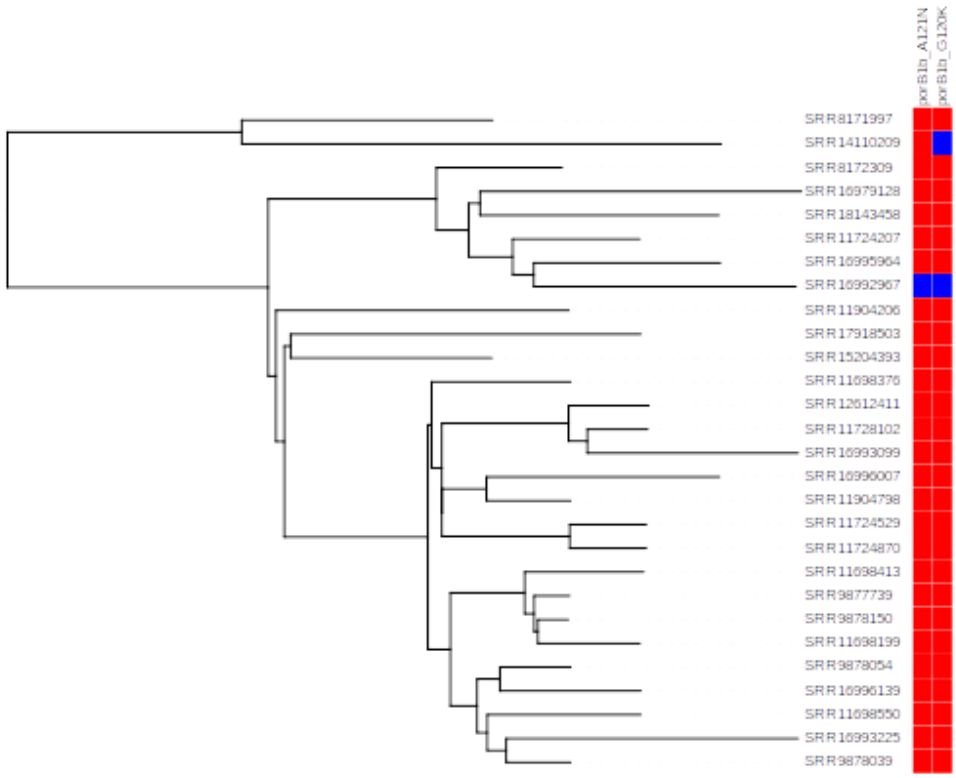

Figure 3.7: COI lineage cluster 1 dated phylogeny with anti-microbial resistance gene heat map. Red indicates the presence of an anti-microbial resistance gene and blue indicates a gene is absent.

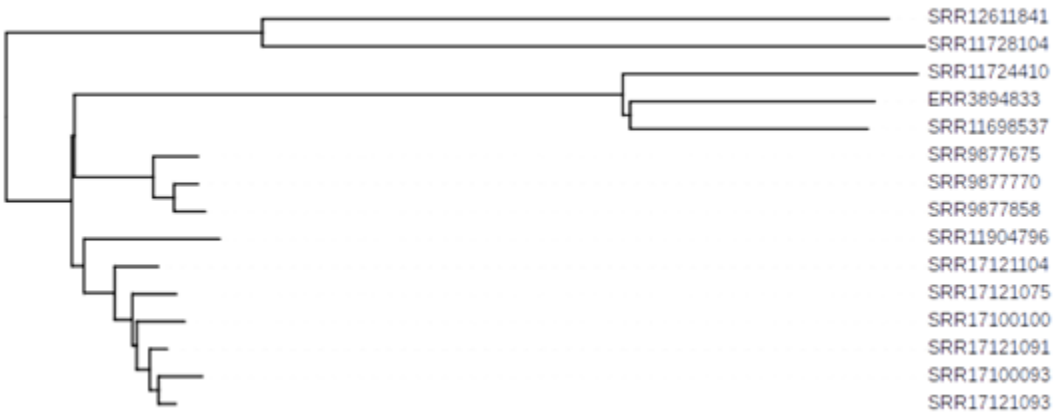

Figure 3.8: COI lineage cluster 5 dated phylogeny with anti-microbial resistance gene heat map. Isolates found in this lineage cluster are positive for all 16 anti-microbial resistance genes

When examining the overarching antimicrobial dynamics of the starting and updated clade-of-interest (Figures 3.9, 3.10), we found that the starting clade contained 16 anti-microbial resistance genes in total (Figure 3.9). Of the 16 anti-microbial resistance genes found in the starting clade, eleven were found in all isolates while another 5 were found at variable levels throughout the clade. The updated clade-of-interest hosts isolates with 26 anti-microbial resistance genes. Ten anti-microbial genes were found in all isolates while 6 genes were found in variable numbers of isolates and another ten isolates were found in few isolates (Figure 3.10).

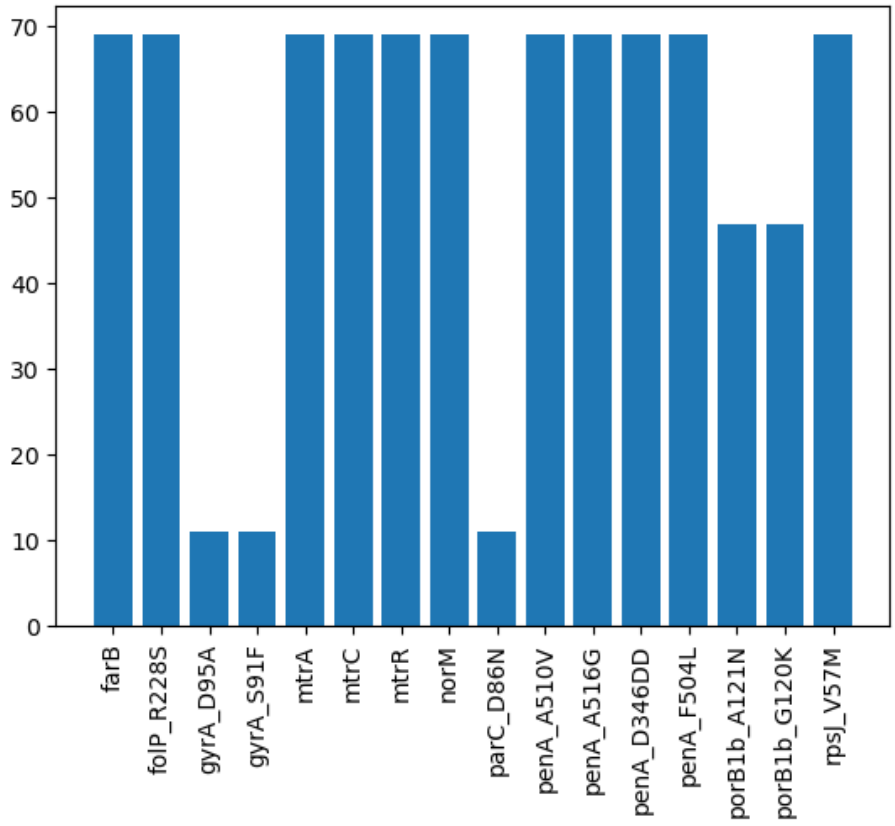

Figure 3.9: Summed anti-microbial resistance gene prevalence throughout the starting clade-of-interest.

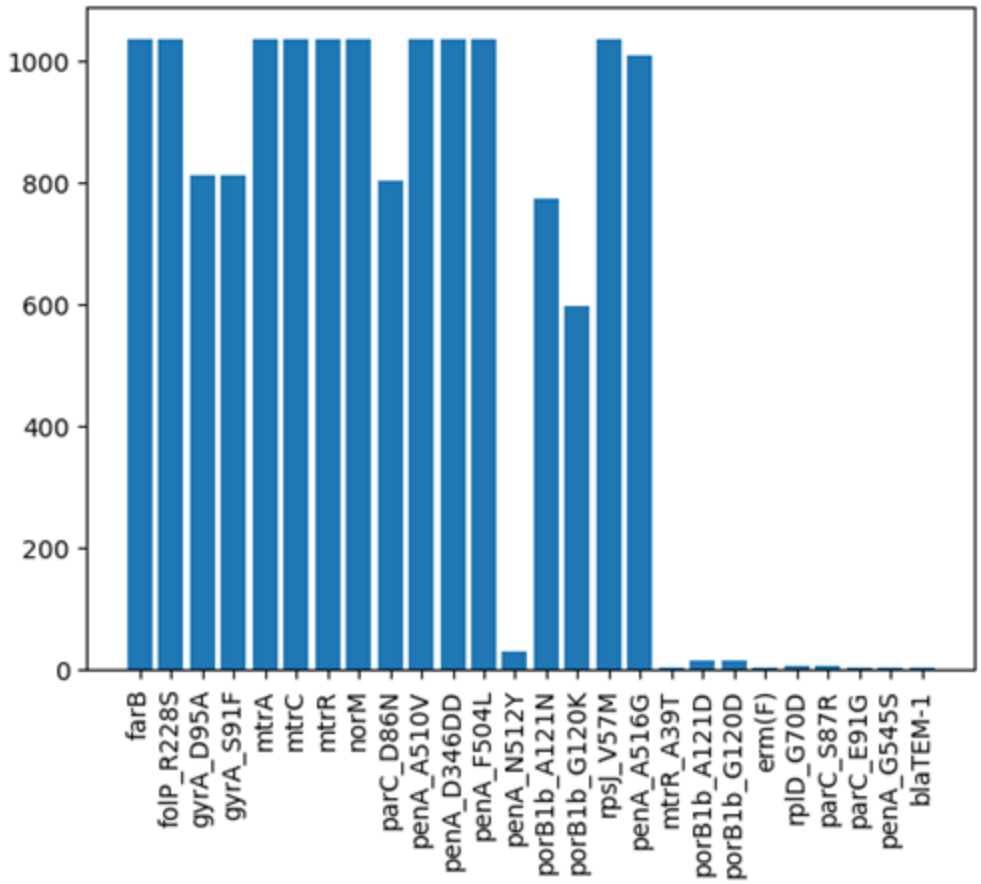

Figure 3.10: Summed anti-microbial resistance gene prevalence throughout the starting clade-of-interest. Ten anti-microbial loci are found ubiquitously, six are found at varying levels. Ten loci were found in only a few isolates.

3.5 Discussion

Investigating *N. gonorrhoeae* evolutionary relationships is a crucial task for researchers in organizations that inform policy makers. Existing methods of investigating gonorrhea have proven effective but are not perfect. The NCBI pathogen Detection program is suitable for rapidly connecting samples and housing significant metadata. The NCBI Pathogen Detection pipeline processes sequence data by assembling genomes *de novo* before using a series of kmer matching and SNP comparison stages to identify and refine lineage groupings (NCBIa. 2022). Finally, the Pathogen Detection pipeline estimates a maximum compatibility phylogeny for each lineage group (Cherry. 2017). We present an alternative method that improves on several aspects of the pipeline and offers researchers a flexible tool to build homologous sequence databases that can easily be used to

investigate evolutionary dynamics throughout a large set of lineages. Our method uses reference guided assembly to rapidly assemble loci-of-interest into a concatenated sequence for each lineage. The sequence of each lineage is added to a database which can be reconstituted into a multiple sequence alignment at any time. Following database construction, any set of loci can be used with a scaffold phylogeny to investigate the relationships of any set of lineages in the database by selecting a clade-of-interest in the scaffold phylogeny and using phylogenetic placement to partition lineages within and outside of a clade-of-interest. Once the clade-of-interest has been expanded with new lineages from in the database, a full maximum likelihood estimation is performed to provide high quality evolutionary relationships. If the clade-of-interest has grown too large for efficient estimation as a single clade, researchers may split the clade into sub-clades to improve processing.

By identifying the most-recent-common-ancestor of all lineages in NCBI lineage clusters in our own clade-of-interest, we were able to identify and collect the composition of the monophyletic group containing these lineages, including lineages that were identified by NCBI as belonging to a different cluster. We used this method to investigate several NCBI identified lineage clusters in our own clade-of-interest, comparing the composition of the NCBI identified lineage clusters to those found in our own clade-of-interest. We found that several lineage clusters from NCBI Pathogen Detection are paraphyletic in our clade-of-interest, although the majority of lineage clusters found in our clade-of-interest are closely related rather than distributed across the entire clade. We used the metadata associated with the samples in our clade-of-interest to build dated phylogenies for several lineage clusters and found that erratic tip dating based on branch length and cluster size. Based on our exploration of dated lineage clusters, we found that smaller sub-clades with relatively equivalent evolution rates lead to the most accurate dated phylogeny estimates. We also used associated anti-microbial resistance metadata to build resistance profiles for lineage clusters in our clade-of-interest, finding little variation in individual resistance genes in these sub-clades.

Our investigation of individual lineage clusters found in our clade-of-interest produced interesting results, regarding the topologies of the lineage clusters themselves and of how our method handles sample classification in and outside of the clade-of-interest. We limited our analyses to the 15 lineage clusters that returned with four or more lineages from the NCBI identified clusters. Only two lineage clusters as identified by NCBI were found in the clade-of-interest with all member lineages included and as monophyletic clades. The remaining 13 clusters had varying numbers of NCBI identified member lineages placed outside of the clade-of-interest (Table 3.1).

Partitioning of lineages included and excluded from the clade-of-interest is a key stage of our approach and we argue its beneficial to the efficiency of the method. Attempts to place lineages into the full 1,237 lineage phylogeny from our work in Field et al., (2022) were unsuccessful. We expect this failure of program performance was driven by dataset size, both in terms of number of included lineages and the sequence lengths associated with each lineage. To accommodate the extreme size of the sample set included in the Pathogen Detection database that we sourced samples from, we removed all samples from the starting phylogeny except the starting clade-of-interest and a single outgroup lineage.

Taxon sampling dictates that by doing this, we skewed the preference of placement locations (Heath et al., 2008). However, as the majority of lineages in this filtering tree were included in the clade-of-interest itself, the expectation of bias in placements is that lineages would place into the clade with more starting lineages. As the majority of the clade-of-interest is included in 15 lineage clusters and none of the 15 analyzed lineage clusters lost a majority of the lineages described by the Pathogen Detection pipeline, we are confident that our filtration approach introduced minimal sampling bias. We also found that the 13 lineage clusters that lost lineages during our placement stage were also paraphyletic compared to their versions produced by the Pathogen Detection pipeline. While these clades were found as paraphyletic, none of lineages from these clusters were widely distributed around the clade-of-interest except cluster 9. Lineage cluster 9, the largest lineage cluster, also includes many short branches throughout the phylogeny, possibly contributing to its intermingling with other lineage clusters. The paraphyletic nature of some lineage clusters could also be due to our dataset's expanded sequence lengths and maximum likelihood estimation method (Bertels et al., 2014). Regardless of the methodology used, questions involving more detailed examinations of lineage relationships are likely to value an examination of sub-datasets including fewer lineages.

Estimating dated phylogenies from the lineage clusters required selecting smaller lineage clusters. Initially, our estimations of full clade-of-interest dated phylogenies included tips dated in the future. Generally, these branches were longer than most other branches in the clade-of-interest, presumably caused by a rate increase in these lineages. While the dates attached to lineages on shorter branches appeared reasonable given the provided date metadata, we chose a cautious approach to handling these erroneously dated long branches by focusing date estimations on smaller, monophyletic lineage clusters. By choosing smaller clades and avoiding excessively long branches, we obtained reliable dated lineage cluster estimates that would be useful for any additional analyses. Additionally, future studies are not limited to lineage clusters as these were an artificial constraint on our analyses. Smaller clades or sub-divisions of clades could be used to estimate dates piecemeal throughout the clade-of-interest, avoiding issues involving long branches. The issue of increased rates might also be avoided by selecting fewer or shorter loci and including fewer invariant sites in the final sequence alignment. Biological mechanisms such a horizontal gene transfer could also play a role in obfuscating accurately estimated dates. Future analyses of horizontal gene transfer could emphasize the careful selection of loci for future analyses or an accommodation of such a mechanism's influence on the dataset. Regardless, the method we have described here is applicable for efficiently building dated phylogenies if used appropriately.

When analyzing the change in anti-microbial resistance profiles of isolates in the clade, we found that while the average number of anti-microbial resistance genes found throughout the clade had increased, the new anti-microbial resistance genes found in the updated clade-of-interest were generally found in only a few isolates. These low prevalence anti-microbial resistance genes are interesting due to their potential as fringe resistance genes that could be in the process of being eliminated from the clade as no longer useful or as the first stages of resistance to a new or rarely used anti-microbial being transmitted to a new clade. The difference in resistance gene prevalence throughout the updated clade

also displays the selection of one resistance allele over another as some variants of a gene are found in almost all isolates while other variants are found in almost none. When we examined the anti-microbial resistance profiles of the two dated phylogenies, we found that both sub-clades primarily reflected larger clade-of-interest anti-microbial resistance dynamics as both contained resistance genes that were found in all isolates across the clade. Lineage cluster 1 contained two resistance genes with high but not ubiquitous prevalence throughout the sub-clade. These two resistance genes were found to have variable prevalence throughout the clade-of-interest and could reflect the last two isolates to not receive the resistance genes through horizontal gene transfer.

The results we've provided here are valuable for gonorrhea surveillance moving forward and highlight some potential next steps for surveillance organizations to consider. The continued proliferation of this clade over time displays that rapid analyses of large bacterial datasets are feasible when using some of the tools we outlined. While our study focused on *N. gonorrhoeae* surveillance, other bacterial pathogens are also recognized as threats to nations around the world. We expect our research can be applied to any bacterial pathogen and prove particularly useful in examining anti-microbial resistant pathogens that are not currently tracked by the Pathogen Detection database. Intensiphy downloads sequences directly from the NCBI SRA, giving future surveillance projects the freedom to collect data for any pathogen. This simple but effective data collection method paired with Extensiphy's ability to accommodate an alignment with any incorporated loci points to easy assembly of anti-microbial loci if present in any query taxon as long as they're included in the reference sequence. The utility of analyzing any bacterial pathogen with no included metadata would improve a surveillance program's ability to monitor outbreaks and transmissions when circumstances are not ideal and metadata is not collected. Taken together, our results have quantitatively described the expansion of a clade of *N. gonorrhoeae*, with an analysis of sub-clade temporal analyses and anti-microbial resistance profiling. Our results also describe the utility that any pathogen surveillance program can gain by employing our methods.

3.6 Conclusions

We developed and described an approach for rapidly processing sequence data to investigate evolutionary relationships. We applied this method to a dataset of over 16,000 *N. gonorrhoeae* samples collected from the NCBI Pathogen Detection database and specifically applied our method in a manner suitable for pathogen monitoring efforts. The method of phylogenetic sample filtration we introduced can accommodate large databases with relative ease and is applicable beyond investigations of just *N. gonorrhoeae*. We compared the results of our clade updating process to the results of an established NCBI pipeline and found that while our estimates differed, they were of comparable utility. The results of our method shed light on an interesting challenge in large scale dated phylogeny estimation that should be acknowledged when selecting loci and lineages for investigations. While researchers will have to choose the method best suited for their questions, we feel we have provided a flexible and reliable framework for investigating

evolutionary relationships at multiple scales. Our approach makes it straightforward to link extensive metadata on sample provenance with core genome alignments and statistically rigorous phylogenetic inferences.

3.7 References

Bertels, F., Silander, O. K., Pachkov, M., Rainey, P. B., & van Nimwegen, E. (2014). Automated Reconstruction of Whole-Genome Phylogenies from Short-Sequence Reads. *Molecular Biology and Evolution*, *31*(5), 1077–1088. https://doi.org/10.1093/molbev/msu088

Budkaew, J., Chumworathayi, B., Pientong, C., & Ekalaksananan, T. (2019). Prevalence and factors associated with gonorrhea infection with respect to anatomic distributions among men who have sex with men. *PLOS ONE*, *14*(4), e0211682. https://doi.org/10.1371/journal.pone.0211682

Centers for Disease Control and Prevention. (2019). *Antibiotic resistance threats in the United States, 2019*. US Department of Health and Human Services, Centres for Disease Control and Prevention.

Cherry, J. L. (2017). A practical exact maximum compatibility algorithm for reconstruction of recent evolutionary history. *BMC Bioinformatics*, *18*(1), 127. https://doi.org/10.1186/s12859-017-1520-4

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., & others. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158.

Field, J. T., Abrams, A. J., Cartee, J. C., & McTavish, E. J. (2022). Rapid alignment updating with Extensiphy. *Methods in Ecology and Evolution*, *13*(3), 682–693. https://doi.org/10.1111/2041-210X.13790

Gordon, A., & Hannon, G. J. (n.d.). *Fastq_toolkit*. http://hannonlab.cshl.edu/fastx_toolkit/download.html

Grad, Y. H., Kirkcaldy, R. D., Trees, D., Dordel, J., Harris, S. R., Goldstein, E., Weinstock, H., Parkhill, J., Hanage, W. P., Bentley, S., & Lipsitch, M. (2014). Genomic epidemiology of Neisseria gonorrhoeae with reduced susceptibility to cefixime in the USA: A retrospective observational study. *The Lancet Infectious Diseases*, *14*(3), 220–226. https://doi.org/10.1016/S1473-3099(13)70693-5

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

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

Harris, S. R., Cole, M. J., Spiteri, G., Sánchez-Busó, L., Golparian, D., Jacobsson, S., Goater, R., Abudahab, K., Yeats, C. A., Bercot, B., Borrego, M. J., Crowley, B., Stefanelli, P., Tripodo, F., Abad, R., Aanensen, D. M., Unemo, M., Azevedo, J., Balla, E., … Verbrugge, R. (2018). Public health surveillance of multidrug-resistant clones of Neisseria gonorrhoeae in Europe: A genomic survey. *The Lancet Infectious Diseases*, *18*(7), 758–768. https://doi.org/10.1016/S1473-3099(18)30225-1

Heath, T. A., Hedtke, S. M., & Hillis, D. M. (2008). Taxon sampling and the accuracy of phylogenetic analyses. *Journal of systematics and evolution*, *46*(3), 239-257.

Heng, L. (2022). *Seqtk*. https://github.com/lh3/seqtk

Hodcroft, E. B., Maio, N. D., 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. https://doi.org/10.1038/d41586-021-00525-x

Leaché, A. D., Banbury, B. L., Felsenstein, J., de Oca, A. nieto-Montes, & 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

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293–W296. https://doi.org/10.1093/nar/gkab301

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

National Center for Biotechnology Information (2022a) *Pathogen Detection.* https://www.ncbi.nlm.nih.gov/pathogens/

National Center for Biotechnology Information (2022b) *NCBI SRA Toolkit*. https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

Papp, J. R., Abrams, A. J., Nash, E., Katz, A. R., Kirkcaldy, R. D., O'Connor, N. P., O'Brien, P. S., Harauchi, D. H., Maningas, E. V., Soge, O. O., Kersh, E. N., Komeya, A., Tomas, J. E., Wasserman, G. M., Kunimoto, G. Y., Trees, D. L., & Whelen, A. C. (2017). Azithromycin Resistance and Decreased Ceftriaxone Susceptibility in Neisseria gonorrhoeae, Hawaii, USA. *Emerging Infectious Diseases*, *23*(5), 830–832. https://doi.org/10.3201/eid2305.170088

Sagulenko, P., Puller, V., & Neher, R. A. (2018). TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution*, *4*(1), vex042. https://doi.org/10.1093/ve/vex042

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

St. Cyr, S., Barbee, L., Workowski, K. A., Bachmann, L. H., Pham, C., Schlanger, K., Torrone, E., Weinstock, H., Kersh, E. N., & Thorpe, P. (2020). Update to CDC's Treatment Guidelines for Gonococcal Infection, 2020. *Morbidity and Mortality Weekly Report*, *69*(50), 1911–1916. https://doi.org/10.15585/mmwr.mm6950a6

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

Unemo, M., Lahra, M. M., Cole, M., Galarza, P., Ndowa, F., Martin, I., Dillon, J.-A. R., Ramon-Pardo, P., Bolan, G., Wi, T., Unemo, M., Lahra, M. M., Cole, M., Galarza, P., Ndowa, F., Martin, I., Dillon, J.-A. R., Ramon-Pardo, P., Bolan, G., & Wi, T. (2019). World Health Organization Global Gonococcal Antimicrobial Surveillance Program (WHO GASP): Review of new data and evidence to inform international collaborative actions and research efforts. *Sexual Health*, *16*(5), 412–425. https://doi.org/10.1071/SH19023

Unemo, M., & Shafer, W. M. (2011). Antibiotic resistance in Neisseria gonorrhoeae: Origin, evolution, and lessons learned for the future. *Annals of the New York Academy of Sciences*, *1230*(1), E19–E28. https://doi.org/10.1111/j.1749-6632.2011.06215.x

Vasimuddin, Md., Misra, S., Li, H., & Aluru, S. (2019). Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 314–324. https://doi.org/10.1109/IPDPS.2019.00041

Whiley, D. M., Mhango, L., Jennison, A. V., Nimmo, G., & Lahra, M. M. (2018). *Direct Detection of penA Gene Associated with Ceftriaxone-Resistant Neisseria gonorrhoeae FC428 Strain by Using PCR - Volume 24, Number 8—August 2018—Emerging Infectious Diseases journal—CDC*. https://doi.org/10.3201/eid2408.180295