A comparison of short- and long-read whole genome sequencing for microbial pathogen epidemiology

Andrea M. Schiffer, Arafat Rahman, Wendy Sutton, Melodie L. Putnam, and Alexandra J. Weisberg

Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, USA.
Correspondence: Alexandra.Weisberg@oregonstate.edu

Abstract

Whole genome sequencing provides the highest resolution for characterizing pathogen evolution, epidemiology, and diagnostics. Genome assemblies contain information on the identity and potential phenotypes of a pathogen. Likewise, variant calling can inform on transmission patterns and evolutionary relationships. Recent improvements in Oxford Nanopore long-read sequencing have made its use attractive for genomic epidemiology. However, the accuracy and optimal strategy for analysis of Nanopore reads remains to be determined. We compared the use of Illumina short reads and Oxford Nanopore long reads for genome assembly and variant calling of phytopathogenic bacteria. We generated short- and long-read datasets for diverse phytopathogenic *Agrobacterium* strains. We then analyzed these data using multiple pipelines designed for either short or long reads and compared the results. We found that assemblies made from long reads were more complete than those made from short-read data and contained few sequence errors. Variant calling pipelines differed in their ability to accurately call variants and infer genotypes from long reads. Results suggest that computationally fragmenting long reads can improve the accuracy of variant calling in population-level studies. Using fragmented long reads, pipelines designed for short reads were more accurate at recovering genotypes than pipelines designed for long reads. Further, short- and long-read datasets can be analyzed together with the same pipelines. These findings show that Oxford Nanopore sequencing is accurate and can be sufficient for microbial pathogen genomics and epidemiology. Ultimately, this enhances the ability of researchers and clinicians to understand and mitigate the spread of pathogens.

Introduction

Whole genome sequencing (WGS) has driven major advances in understanding plant and clinical pathogen evolution, epidemiology, and diagnostics (Everhart et al. 2021; McLeish et al. 2021; Parkhill and Wren 2011; Straub et al. 2021). Whole genome sequences can be used to infer host range, antimicrobial resistance, and other characteristics of microbial plant and clinical pathogens. WGS data are advantageous because they simultaneously provide the highest resolution for epidemiological analysis and the characterization of pathogen transmission patterns. Recent technological developments have made it possible for plant and clinical diagnostic clinics to incorporate these tools into their arsenal. These data have already informed on emerging pathogens, such as the recent outbreak of multidrug-resistant *Pseudomonas aeruginosa* in contaminated eye drops, and an introduction of *Xanthomonas hortorum* pv. *pelargonii* into the U.S. on geranium (Grossman et al. 2024; Iruegas-Bocardo et al. 2023).

Traditionally, short-read Illumina sequencing has been considered the gold standard for whole genome sequencing. Illumina sequencing produces reads with high accuracy, which can be used to assemble genomes with high consensus sequence confidence or for whole genome variant calling (Goodwin et al. 2016). However, Illumina reads are at most 300 bp in length, which limits their use in assembling repetitive regions of genomes. Therefore, most genome assemblies made from short reads are fragmented into multiple pieces, called contigs. Chromosome structure or the presence of mobile genetic elements such as plasmids can be difficult to infer from short read assemblies.

The introduction of long-read sequencing aimed to resolve many of the issues inherent to short-read sequencing. Oxford Nanopore sequencing has the advantage of producing very long reads, up to a megabase in length, and on sequencers that are affordable to most labs. However, previous studies found that the quality of Nanopore reads was not high enough to be used on their own (Luan et al. 2024; Sereika et al. 2022). Nanopore reads also contained homopolymer repeat errors that can result in indels in assemblies

(Huang et al. 2021; Luan et al. 2024). Therefore, short-read sequencing data was often necessary for polishing long-read assemblies (Wick and Holt 2022). However, recent advances in Oxford Nanopore chemistry and basecalling have changed the landscape of long-read sequencing. Assemblies produced from Nanopore long reads are now nearly perfect for many organisms (Wick et al. 2023).

Epidemiology and pathogen transmission can be characterized to the highest resolution using whole genome variant calling. Variant calling pipelines typically use the alignment of reads to a common reference genome to identify single nucleotide polymorphisms (SNPs) for each strain (Olson et al. 2015). Strains differing by less than a determined threshold, often 10-15 SNPs, are then grouped into genotypes (Stimson et al. 2019). Mapping host or location information onto genotypes can then be used to characterize pathogen transmission. Short-read data has long been used for variant calling, and mature tools are available for this analysis (Danecek et al. 2021; Eggertsson et al. 2017; McKenna et al. 2010). However, the historically poor quality of long reads has meant that their value and use in variant calling is not well characterized. Tools for variant calling with Nanopore reads are now available (Zheng et al. 2022).

In this study, we assessed the quality of long reads in genome assembly and variant calling for population-level genotyping analysis. We compared genome assemblies made from either short or long reads. We also compared different strategies for variant calling using tools designed for either Illumina or Nanopore reads. Finally, we highlight the potential benefits that plant and medical clinics can derive from adding whole genome sequencing to their pathogen diagnostic testing.

Results

To test the viability of long-read sequencing for bacterial epidemiology, we analyzed *Agrobacterium* strains isolated from plant samples submitted to the OSU Plant Clinic. We also included strains from the Larry Moore and Thomas Burr culture collections, and from the California Department of Food and Agriculture (CDFA) collection. These 116 strains

were selected to maximize diversity in plant host, location, and year of isolation (**Table S1**). We generated Illumina short-read data for all strains, and Oxford Nanopore long-reads for a subset of strains and analyzed them independently using multiple established pipelines. The 35 paired datasets allowed us to compare standard pipelines using short reads with assembly and variant calling pipelines using long reads. The accuracy and completeness of each approach were assessed using multiple metrics.

We first assessed the quality of short- and long-read datasets. The average read length of Nanopore reads was 6,835 bp (mean N50 read length 13,385 bp; **Table S2**). Average coverage for Illumina read datasets varied from 22X to 106X (mean 52X coverage) and average coverage for Nanopore read datasets varied from 17X to 173X (mean 61X coverage; **Table S3**). During this study, multiple Nanopore basecalling models were released, so reads were rebasecalled and retested in the pipeline. We compared the "SUP" versions of production model v4.2.0, a beta-release bacterial methylation-aware model (hereafter "bacmethyl"), and the latest production model v5.0.0. The quality of reads was consistent or improved with each new model, particularly with v5.0.0 (**Table S2**). Nanopore simplex reads basecalled with the v5.0.0 SUP model had an average quality score of Q19.12, or a per-read accuracy of ~98.7%.

Assemblies made with Nanopore data reveal complete chromosome structure
Genomes were assembled from either short- or long-read data, or a hybrid approach combining both, and then compared to assess completeness and accuracy (Kolmogorov et al. 2019; Prjibelski et al. 2020; Wick et al. 2017). As expected, genomes assembled from long-read data or hybrid data were more complete and had higher N50 values than those assembled from short reads (**Figure 1A and B**; **Table S3**). The average N50 was 483,620 bp for short-read SPAdes assemblies and 3,784,004 bp for long-read Flye assemblies. Short-read datasets assembled into an average of 55 contigs (**Table S3**). In contrast, nearly all long-read assemblies contained structurally complete chromosomes and plasmids, if present (**Table S3**). Total assembly size was comparable between all assemblies, but Flye and hybrid assemblies were generally slightly larger (**Table S3**).

Improved basecalling reduces assembly errors in long read-only assemblies Genomes assembled from long reads have historically contained many single nucleotide polymorphism (SNP) and small insertion/deletion (indel) errors (Delahaye and Nicolas 2021). However, recent improvements in Nanopore basecalling models have reduced this error rate. We assessed the number of assembly errors in each long-read assembly by polishing with high quality Illumina short reads and counting the number of changes made (Wick and Holt 2022).

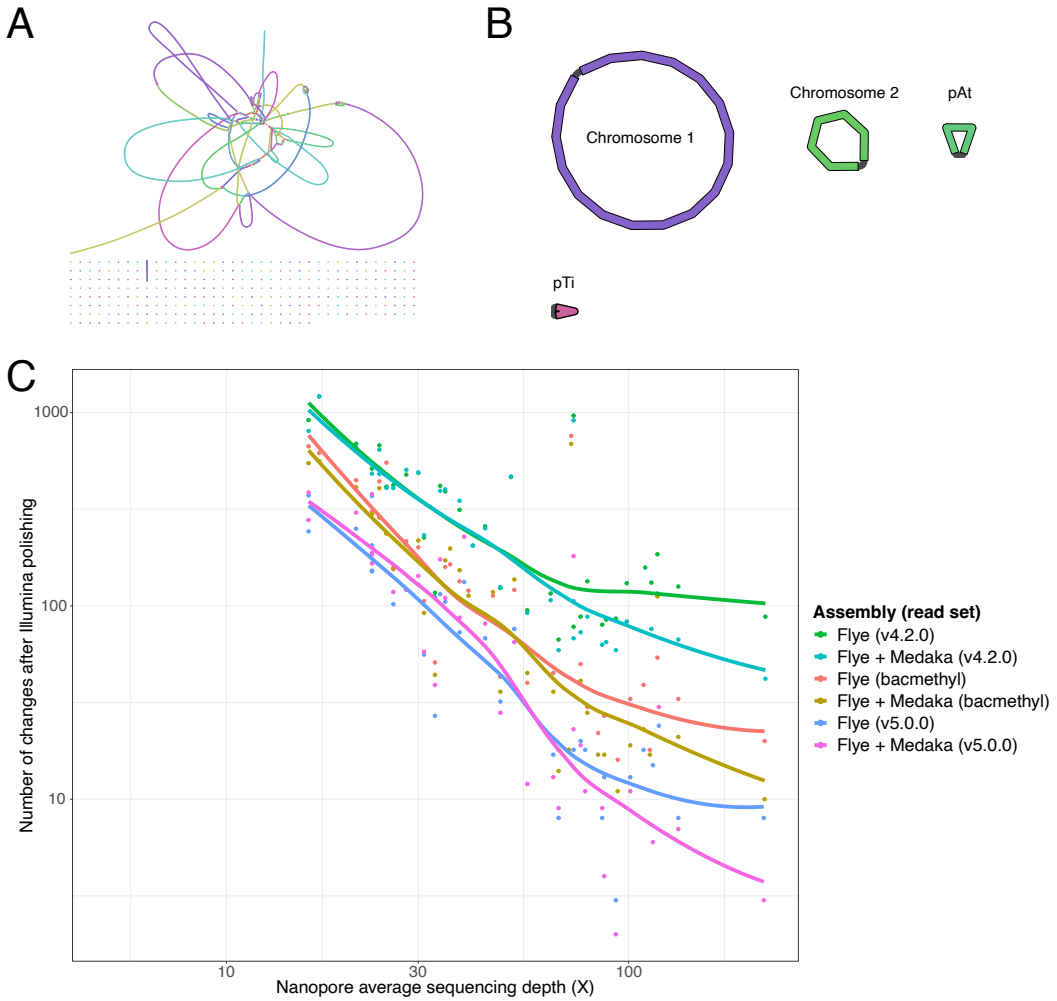

Figure 1: Nanopore-only assemblies with sufficient depth of sequencing are complete and have few errors. Assembly graphs of strain CG154 assembled from **A.** Illumina short reads or **B.** Nanopore long reads. **C.** Assemblies of sequenced strains plotted as points on a graph of long read sequencing depth (X) by the number of changes made to that assembly following Polypolish polishing with Illumina short reads. The long-read assembly method (Flye or Flye with Medaka polishing) and basecalling model version (v4.2.0, beta methylation-aware bacteria model [bacmethyl], or v5.0.0) is indicated by color. A best-fit line was plotted for each assembly method and model combination using locally estimated scatterplot smoothing.

Assemblies made from reads basecalled with the most recent model (v5.0.0) had the fewest assembly errors (**Table S4**). Long-read assemblies with at least 50x coverage had on average 23 SNP errors, with as few as 3 errors. The number of errors in long-read assemblies declined with increasing sequencing depth of long reads (**Figure 1C**). Strains with less than 30X long-read coverage had many errors, while strains with >30X sequencing depth had fewer errors (mean 43 errors). Above 80X long-read coverage, the number of errors did not dramatically decrease as sequencing depth increased. We also tested the effectiveness of polishing long-read assemblies with long reads using Medaka (Oxford Nanopore Technologies 2023). Medaka polishing slightly reduced the number of assembly errors when long-read datasets were above 80X coverage (**Figure 1C; Table S4**).

We noticed an unusually high number of errors in the assembly of strain CG160, despite this sample having >100X long-read coverage. We identified low levels of contamination in the short and long reads of this sample. This contamination primarily affected the assembly of one plasmid, which contained nearly all assembly errors. Manual examination of read alignments revealed that errors were located at conserved plasmid replication and conjugation loci. After filtering out reads that mapped to the contaminants, a new assembly made from the filtered reads reduced the errors from 469 to 24 (**Table S4**).

We next tried to identify motifs in long-read assemblies associated with errors. Previous studies identified that errors in Nanopore reads occur more frequently near methylated nucleotides (Chiou et al. 2023). Bacterial DNA is often methylated at specific sequence motifs (Sánchez-Romero and Casadesús 2020). The motif “GANTC” or “GATC” is methylated as part of cell cycle regulation in many Alpha- and Gammaproteobacteria, respectively (Gonzalez et al. 2014; Kahng and Shapiro 2001; Palmer and Marinus 1994). Sequence motifs targeted by restriction-modification (R-M) defense systems are also

methylated (Vasu and Nagaraja 2013; Williams 2003). We characterized the sequence context around errors in long-read assemblies and identified conserved motifs (**Table S5**). Most assembly SNP errors are near sites likely to be targeted for methylation. Nearly all identified motifs are either known to be methylated in agrobacteria, such as “GANTC”, or are short palindromes that resemble R-M system methylation targets, such as “CTGCAG”. Overall, assembly errors were reduced in datasets made from basecalling models that account for methylation (bacmethyl, v5.0.0) relative to models that do not (v4.2.0; **Figure 1C; Table S4**). Most improvements in assembly error rates are in sites with predicted methylated motifs (**Table S5**). Other assembly errors were identified as small indels or homopolymer errors, though these also were reduced with more recent basecalling models (**Table S4**).

We compared gene annotations of SPAdes short-read assemblies, short-read polished and unpolished Flye long-read assemblies, and Unicycler hybrid assemblies to characterize the quality of gene calls. Flye assemblies, both polished and unpolished, consistently had slightly greater numbers of annotated gene features than SPAdes assemblies (**Figure S1**). There were few or no differences in the number of genes called between polished and unpolished Flye assemblies (**Table S3; Figure S1**). BUSCO scores were comparable across all read-type and assembly methods (**Table S3; Figure S2**).

SNP calling pipelines vary in accuracy when used with long-read data

We next compared short and long reads for whole genome variant calling using a variety of pipelines. We tested four SNP calling pipelines with short reads and v5.0.0 long reads. Three of the pipelines, GATK, Graphtyper, and Bcftools, were originally designed for Illumina short-read data (Danecek et al. 2021; Eggertsson et al. 2017; McKenna et al. 2010). The other tool, Clair3, is the only one specifically designed for variant calling of Nanopore reads (Zheng et al. 2022). We previously found Graphtyper to be the most consistently accurate for inferring clonal genotypes of bacterial pathogens from short-read data (Weisberg et al. 2020, 2021). GATK was also accurate if datasets were first

partitioned into below species-level groups. While neither tool was designed for long reads, we tested these pipelines with our short- and long-read datasets. We then assessed the consistency of variant calls across short and long read datasets.

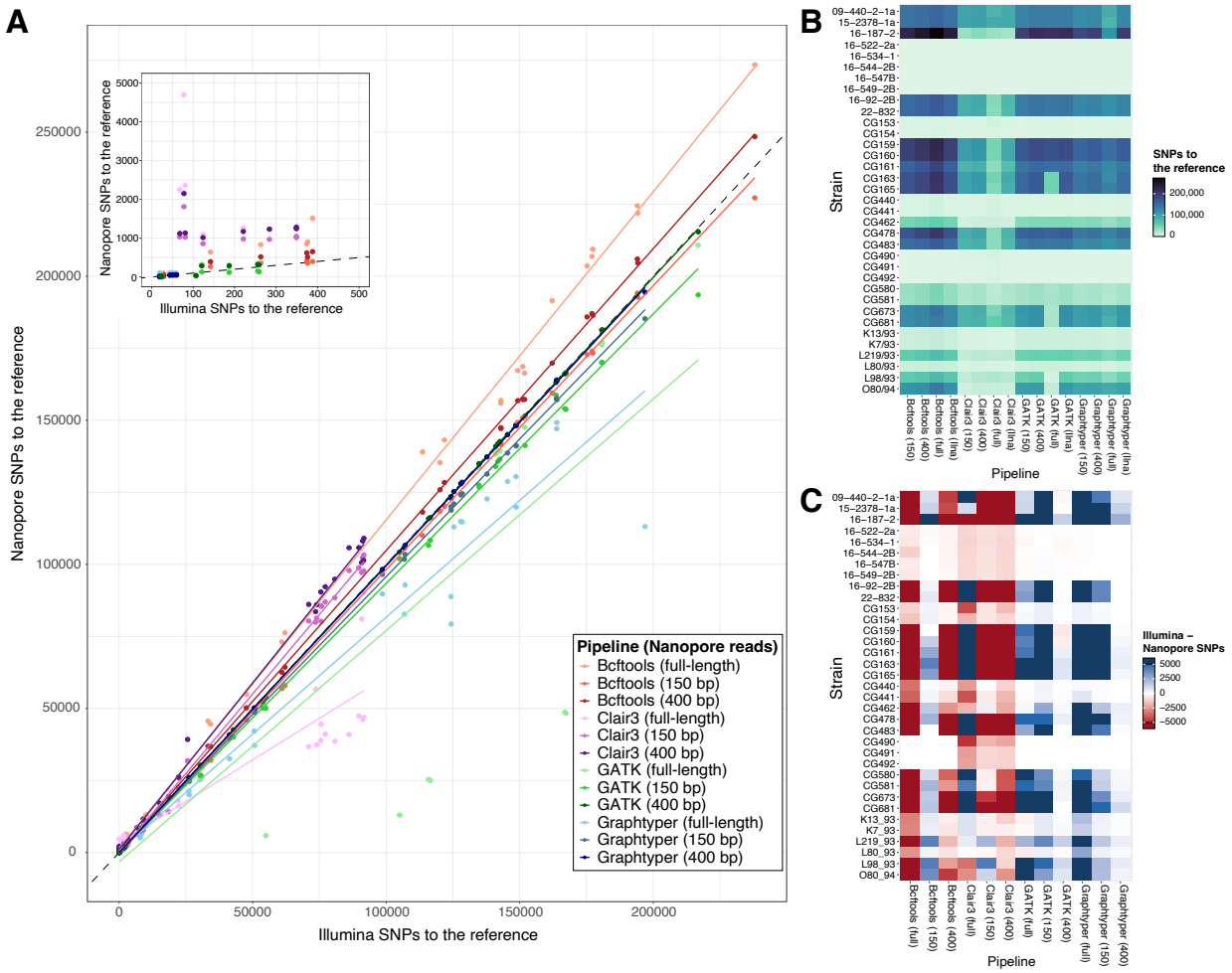

Figure 2: **Comparison of variant calling pipelines. A.** Plot of SNP calls to the reference for Illumina vs Nanopore reads for each strain and pipeline. Each point represents an individual strain, colors indicate variant calling pipeline and long-read dataset type. Long reads were either full length, or fragmented into 150 bp (150) or 400 bp (400) pieces. The black line represents a predicted 1 to 1 ratio between Illumina SNPs to the reference and Nanopore SNPs to the reference. The embedded graph is zoomed to show the 0 to 500 SNPs region. **B.** A heatmap comparing the number of SNPs called to a reference using different variant calling pipelines and read sets. Each row is a strain, each column is the number of SNPs called for that strain using each pipeline and read type. **C.** A heatmap representing the difference in SNP calls to the reference made with Illumina and Nanopore reads for each strain. Rows are strains, columns are variant calling pipelines and long-read type. Color indicates whether more SNPs were called with Illumina reads (blue) or Nanopore reads (red). Values above 5000 SNP differences are colored dark red and dark blue.

The analyzed strains represent genetically diverse agrobacteria (**Figure S3**). Variant calling was initially performed within species-level groups (ANI >95%). For each strain,

short and long reads were analyzed separately in each pipeline, along with strains from NCBI. We then compared the number of SNPs to the reference inferred for each strain with either short- or long-read data (**Figure 2A; Table S6**). Overall, results were largely comparable across short and long reads. However, each pipeline differed in how consistently calls were made with each read type. Bcftools consistently called more SNPs with full-length long reads than with short reads for each strain (**Figure 2A and C**). In contrast, Graphtyper, GATK, and Clair3 consistently called fewer variants with full-length long reads than with short reads. Results from Clair3 were unusual in that overall fewer variants to the reference were called with short or long reads relative to all other pipelines (**Figure 2B**). This was particularly the case for strains distantly related to their reference genome, such as those with >50,000 SNPs to the reference in Graphtyper or GATK analyses (**Table S6**). When considering only comparisons where strains are closely related to the reference genome (<500 SNPs), both Graphtyper and GATK produced very consistent SNP calls using either short-read or full-length long-read data (**Figure 2**). Clair3 and Bcftools consistently called a greater number of SNPs to the reference using full-length long reads than with short reads.

Fragmented long reads are comparable to short reads in variant calling pipelines

We hypothesized that poor read mapping of long reads to distantly related reference genomes was responsible for the under-calling of variants. We tested this by fragmenting full-length long reads into either 150 bp or 400 bp sequences and called variants using each pipeline. In all pipelines, fragmentation of long reads led to more consistent SNP calls across short- and long-read datasets (**Figure 2; Tables S11-14**). Except for Bcftools, variant calling with each pipeline using 400 bp fragments was more consistent with calls for short-read data than variant calling with 150 bp fragments. Variant calling with Graphtyper or GATK pipelines using 400 bp fragments most closely matched short-read SNP calls (**Figure 2A, S4**).

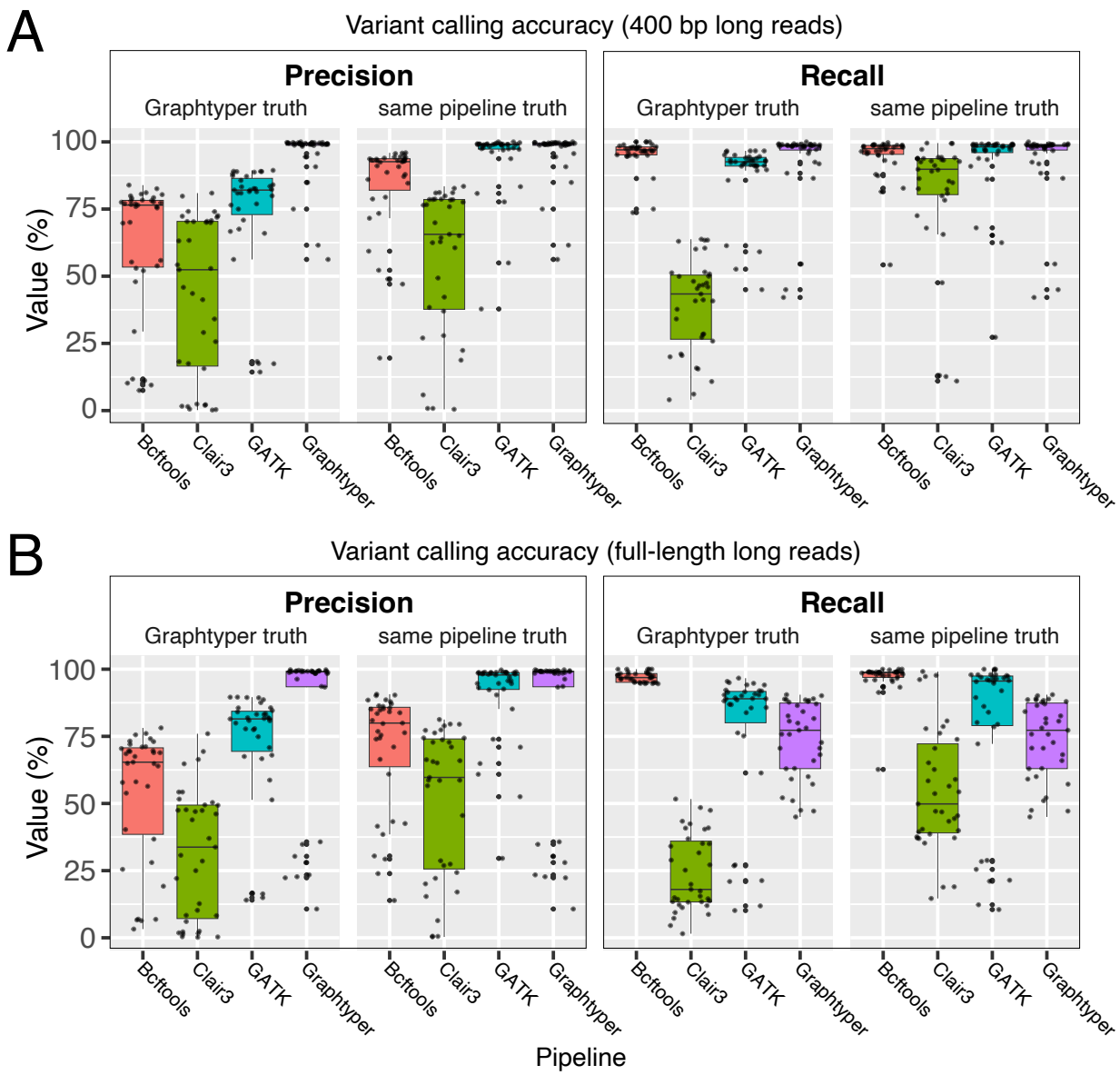

Figure 3: Variant calling accuracy across pipelines. Precision and recall scores for per-strain variant calls produced by each pipeline using either **A.** 400 bp fragmented long reads or **B.** full-length long reads. Truth sets were either variant calls produced by Graphtyper from Illumina short reads or variant calls produced by the tested pipeline using Illumina short reads. Precision and recall scores are represented as percentages. Scores for each strain are represented as points and summarized as boxplots.

To infer the accuracy of variant calls with long reads, precision and recall were calculated for each pipeline. First, variant calls produced from short reads with each pipeline were used as the truth set and variants produced from either full-length or 400 bp long reads were used as the test set. Precision and recall were highest for all pipelines when using 400 bp fragments relative to full-length long reads (**Figure 3**). Graphtyper with 400 bp

long reads had the greatest precision (95%) and recall (93%) of all pipelines (**Table S7, Figure 3B**). GATK with 400 bp long reads was the next most accurate, with an average precision of 94% and recall of 93%. Bcftools had a precision of 84% and recall of 95%, while Clair3 had a precision of 57% and recall of 81%. Precision vs. recall curves of analyses of 400 bp read fragments were consistent with these results, with Graphtyper and GATK having the highest accuracy (**Figure S5**). Precision vs. recall curves made using Graphtyper with short reads as the truth set were largely consistent with the analyses using within-pipeline truth sets (**Figure 3, S6**). In this comparison, GATK with 400 bp long reads had a slightly lower precision of 74% and recall of 89%. Bcftools had a precision of 62% and recall of 96%, while Clair3 had a precision of 43% and recall of 39%.

Long reads are sufficient for accurately inferring genotypes

We next compared pipelines for inferring genotypes from SNP data. Genotyping, identifying sets of genetically identical strains, is important for characterizing outbreaks and understanding pathogen transmission (Kamvar et al. 2014; Stimson et al. 2019). For each pipeline or approach, we characterized whether short- and long-read datasets for each strain were correctly inferred to be the same genotype. We used a threshold of <15 relative pairwise SNP differences to infer two strains as belonging to the same genotype. The Graphtyper and GATK pipelines with 400 bp long-read fragments correctly paired the most pairs of short and long read datasets into genotypes (**Table S8**). Graphtyper correctly inferred short-read and 400 bp long-read data from the same strain into genotypes for 16 of 35 strains, and GATK with the same data correctly inferred the same genotype for 9 of 35 strains. Bcftools correctly inferred genotypes for just 2 of 35 strains. Clair3 did not correctly pair short-read and long-read datasets into genotypes for any of the 35 strains. All pipelines struggled to correctly genotype full-length long-read datasets with their short-read counterparts. GATK, Bcftools, Graphtyper, and Clair3 correctly inferred just 8, 2, 1, and 0 of 35 genotypes, respectively. Long-read datasets fragmented to 150 bp had intermediate results between the full-length and 400 bp fragments in inferring genotypes in each pipeline (**Table S8**).

While Graphtyper with 400 bp long-read fragments correctly paired short and long reads into genotypes for most strains, it did not group all. We hypothesized that this may be due to comparing strains to distantly related references. Using a 95% ANI threshold, strains with 5-7 Mb genomes can be grouped with those that differ by ~100,000 SNPs. Therefore, we tested partitioning the strains into groups using a 99% ANI threshold. We tested Graphtyper with these groupings using 400 bp long-read fragments. While some strains could not be analyzed because the 99% ANI threshold grouped them as singletons, Graphtyper correctly grouped short-read and 400 bp long-read datasets for 25 of 28 strains (**Table S8**).

We also characterized how consistently each pipeline grouped different strains into genotypes. We previously found Graphtyper to be the most accurate at bacterial genotyping with short read data (Weisberg et al. 2020, 2021). Therefore, we selected genotypes inferred by Graphtyper from short-read data in the 95% ANI groupings to be the truth dataset. Graphtyper with short reads identified 17 pairs of strains that belong to the same genotype (**Table S9**). We characterized how many of these genotypes were inferred by the other pipelines. Graphtyper with 400 bp fragmented long reads was the only pipeline to correctly infer all 17 truth-set genotypes (**Table S7**). GATK with 400 bp fragments correctly genotyped 13. Graphtyper and GATK with 150 bp long-read fragments correctly genotyped 16 and 14 truth-set genotypes, respectively. Graphtyper and GATK with full-length long reads correctly genotyped 11 and 9 genotypes, respectively. Bcftools with each of full-length, 150 bp fragment, or 400 bp fragment long reads only correctly genotyped 1 truth set genotype. Clair3 failed to correctly genotype any trust set pairs with any of the long-read datasets. GATK, Bcftools, and Clair3 with short reads correctly inferred 13, 13, and 10 truth set genotypes, respectively.

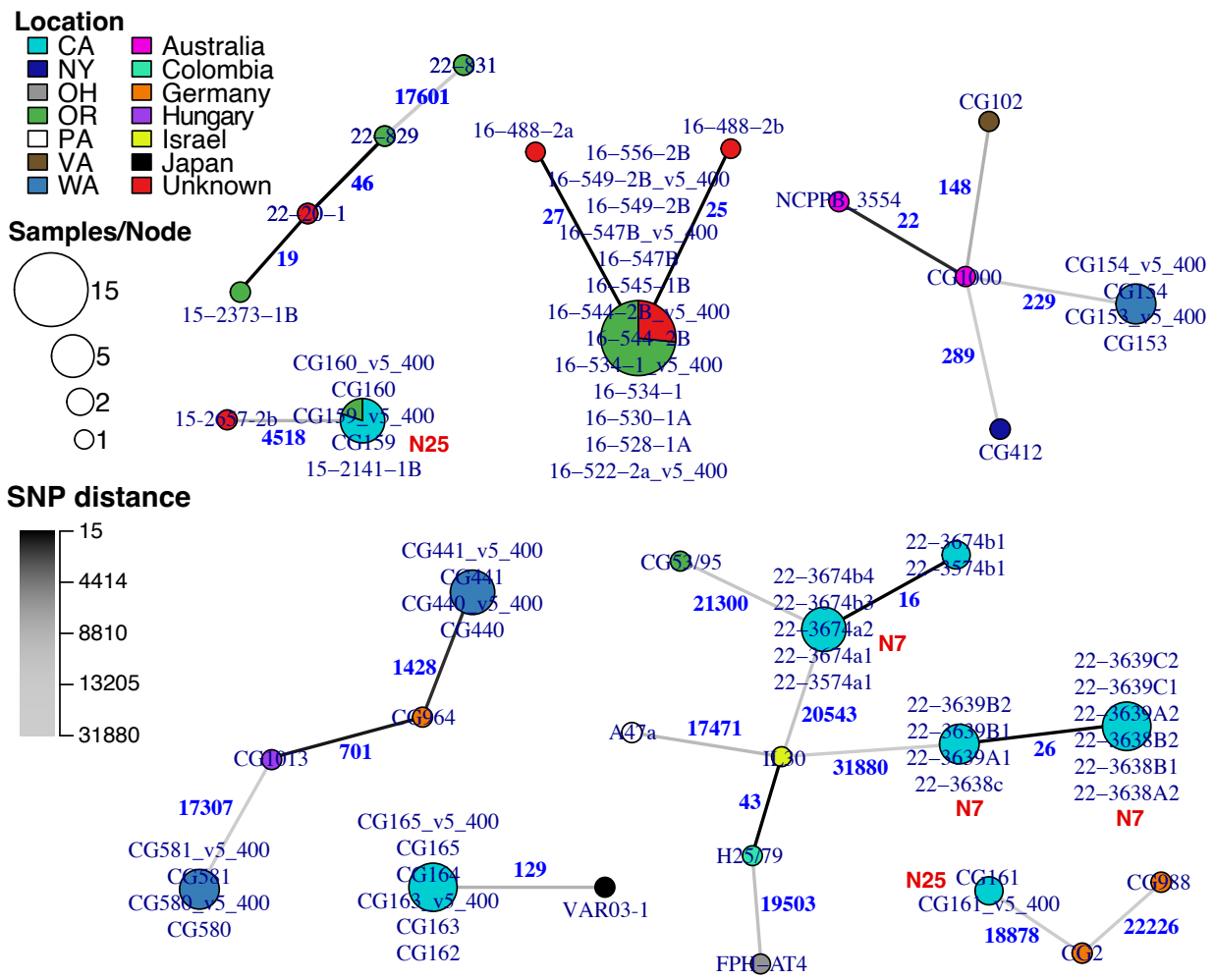

Figure 4: Whole genome sequencing data reveals epidemiological patterns. Minimum spanning networks representing relationships between select strains. Nodes represent whole-genome genotypes inferred from strains with ≤15 SNP differences. Edge labels and color represent the number of SNP differences between genotypes. Node size indicates the number of strains of each genotype, the proportion of each color indicates the location of isolation for strains of that genotype. Genotypes from select nurseries are labelled in red text. Strains were analyzed in groups with >99% ANI similarity. Genotypes were inferred from variant calls produced by Graphtyper with either short or 400 bp long read data (labeled _v5_400).

Epidemiology of Plant Clinic strains

Once we had identified an optimal pipeline and strategy for variant calling, we characterized the identity and inferred transmission patterns for the newly sequenced strains. The 116 sequenced strains are genetically diverse and represent each of the major lineages of agrobacteria (**Figure S3**). Many strains are closely related to previously sequenced agrobacteria from other hosts or locations. Most strains carry oncogenic

plasmids, suggesting that they are likely to be pathogenic. Genotypes inferred from the Graphtyper analysis of strains partitioned using a 99% ANI threshold were visualized as minimum spanning networks (**Figure 4; Table S15**). Both short reads and 400 bp long-read fragments were consistent in identifying epidemiological links between or within locations. Strains CG163 and CG165 belong to the same genotype but were collected from two different vineyards. These data suggest a potential transmission event or acquisition from a common source. Likewise, two non-pathogenic strains (CG159, CG160) sampled from grapevine in California in 2000 are genetically identical to a non-pathogen strain also sampled from grapevine in Oregon in 2015 (15-2141-1B). Three strains (CG159, CG160, and CG161) collected from nursery N25 in 2000 represent multiple genotypes (**Figure 4**). Likewise, several genotypes of strains isolated from nursery N7 in 2022 differ by more than 30,000 SNPs (**Figure 4**). Both examples represent co-occurring infections at a single location.

Discussion

These results and other recent studies suggest that Nanopore long reads are now of high enough quality to be used for genome assembly and variant calling on their own (Bloomfield et al. 2024; Bogaerts et al. 2024; Hall et al. 2023, 2024; Luan et al. 2024). With the latest basecalling models and sufficient depth of sequencing, long reads can be used to assemble complete genomes with very few errors. The most serious errors are indel errors, as these can cause apparent frameshift mutations that disrupt gene annotation. However, these errors are also reduced in reads produced by the latest basecalling models (**Table S4**). While long-read assemblies are likely of acceptable quality on their own, shallow Illumina sequencing can also be added to ensure perfect assemblies (Wick et al. 2023). However, for most use cases this may not be necessary. Long-read assemblies are sufficient for identifying an organism, inferring phylogenetic relationships, and characterizing the presence of virulence genes or genes involved in antimicrobial resistance (AMR).

Results from this study show that fragmenting Nanopore reads into shorter sequences is important for ensuring accurate variant calling. This is especially important when performing variant calling with more distantly related references, as would be done in population-level analyses. Once long reads were fragmented, pipelines designed for short reads were more accurate than tools designed for long reads. The improvements in variant calling are likely due to improved read mapping. As evolutionary distance increases, synteny between strains is likely to decrease as rearrangements and insertions or deletions accumulate. Longer reads are more likely to span breaks in synteny and therefore not map or map poorly. Long reads may also map poorly in regions with a high density of variants. By fragmenting long reads, read fragments are less likely to span breaks in contiguity or genome rearrangements. Similar strategies have been used for pangenome construction, where tools that build pangenomes at a per-gene level can be more accurate with distantly related strains (Colquhoun et al. 2020). While fragmenting reads loses information about synteny and may result in reads mapping to multiple regions of the reference, it may be necessary for ensuring adequate mapping. Alternative strategies could include using less-stringent read mapping algorithms.

It is unclear why Clair3 performed poorly with long reads relative to the other pipelines. These results contrast with previous studies, which found Clair3 to be the most accurate variant caller with long reads (Hall et al. 2024). In our datasets, Clair3 called far fewer variants than other pipelines with the same strains (**Figure 2B; Table S6**). It is possible that the reference genomes chosen for this analysis were too different for the models in Clair3 to call SNPs. Previous studies comparing variant calling pipelines for Nanopore reads used closely related reference genomes (Bloomfield et al. 2024; Bogaerts et al. 2024; Hall et al. 2024). In this study, we tested datasets representing population-level studies rather than a direct comparison of strains known to be nearly identical. However, the three other pipelines tested in this study performed better than Clair3 even when strains were closely related to the reference.

Most variant calling tools are only designed for either short or long reads. Few tools explicitly have options for mixed datasets. This results in compromises in choosing parameters or settings for each run. For example, using long reads with Graphtyper requires setting an option that ignores filtering by read pair mapping ("--no_filter_on_proper_pairs"). If a Graphtyper analysis has strains with paired short reads and other strains with long reads, read pairing will be ignored for the short-read datasets. This parameter did not dramatically impact variant calling with short reads in our analyses; however, it may be a problem for other datasets. Pipelines that include per-sample options for either short or long reads could improve results for mixed datasets. Alternatively, users should use only one read type or the other.

Caution should be taken in interpreting variant calls and genotypes inferred from mixed datasets of short and long reads. In some cases, comparisons across data types did not correctly infer a genotype for pairs of strains, but comparisons within data types did. For example, with 95% ANI groups, analysis of short-read data for strains CG153 and CG154 inferred them to be the same genotype (**Table S9**). While the full-length long-read data of CG154 was included in this genotype, the full-length long read data of CG153 was not (**Table S9**). However, by fragmenting long reads to 400 bp and grouping strains with a 99% ANI threshold, we eliminated most issues with genotyping across read types. In this analysis, short and long read data correctly grouped CG153 with CG154 in all comparisons. Even at a 99% ANI threshold, strains can differ by >40,000 SNPs. We observed several cases where a 99% ANI threshold is still too distant for accurate variant calling. For example, using strain CA75/95 as reference for 99% ANI group 723, the short and long reads of strains L219/93 and L98/93 did not genotype together (**Table S15**). However, when variants were instead called using L219/93 as reference, the read sets and strains were genotyped together as expected (**Table S15**). If strains appear to be closely related but not the same genotype, we recommend re-analyzing with a reference genome more closely related to those strains.

Generating whole genome data is now easier and less expensive than before. Individual labs or plant clinics can prepare libraries and perform Illumina short read or Oxford Nanopore long read sequencing in the lab with relative ease. These results suggest that either technology is sufficient for genome assembly and variant calling in most cases. Both Illumina and Oxford Nanopore have released DNA sequencers that can be purchased by individual labs. An advantage of long read sequencing is that libraries are generally less time consuming and easier to prepare in the lab. If sequencing is performed in house, costs can be cheaper, and results acquired more quickly, than with short read sequencing. Rapid advances in sequencing technology provide a major opportunity for plant clinics and labs to generate data with the highest resolution for diagnostics and epidemiology.

Materials and Methods

Genome sequencing and analysis

Strains were selected from the Larry Moore culture collection or from those collected by the OSU Plant Clinic (**Table S1**). Additional strains were shared by the California Department of Food and Agriculture (CDFA). Strains were cultured in MGYs liquid media or plates at 28C (Moore et al. 1997). The Promega Wizard Genomic DNA Extraction kit (Promega; Madison, WI) was used to extract DNA from cell pellets of each strain. The protocol was modified to include an overnight lysis step at 37C and mixing by inverting the tube rather than pipetting. All strains were sequenced with both Illumina and Oxford Nanopore technologies. The SeqWell PurePlex kit was used to prepare libraries for Illumina sequencing. Illumina libraries were sequenced (2x150 bp) on an Illumina MiniSeq (Illumina; San Diego, California) in the OSU Plant Clinic. The Rapid Barcoding 96 V14 kit (SQK-RBK114.96) was used to prepare multiplexed libraries for Nanopore sequencing. Nanopore libraries were sequenced on a PromethION flow cell on a P2 solo sequencer (Oxford Nanopore; Oxford, UK). Dorado v.0.5.0 with the default parameters, SUP basecalling, and different basecalling models was used to basecall nanopore reads (Oxford Nanopore PLC. 2024).

FastQC was used to check Illumina reads for quality (Andrews 2024). NanoPlot v1.42.0 with the default parameters was used to check Nanopore reads for quality (De Coster and Rademakers 2023). Spades v3.15.3 with the parameters "--phred-offset 33 –careful -k 21,33,55,77,99" was used to assemble genomes from only Illumina reads (Prjibelski et al. 2020). For all assemblies, contigs with <500 bp in length or <5X coverage were removed. Flye v2.9.2-b1786 with the parameters "--scaffold --read-error 0.03 was used to assemble genomes from only Nanopore reads (Kolmogorov et al. 2019). Medaka v.2.0.0 with the model r1041_e82_400bps_sup_v4.2.0 or r1041_e82_400bps_sup_v5.0.0 was used to polish Flye assemblies (Oxford Nanopore Technologies 2023). Polypolish v0.5.0 with the default parameters was used to polish Flye assemblies with Illumina reads (Wick and Holt 2022). Unicycler v0.5.0 with the parameter "--mode normal" was used to create a hybrid assembly from both Illumina and Nanopore reads (Wick et al. 2017). Beav v0.2 with the parameter "--agrobacterium" and Bakta v.1.8.2 was used to annotate genomes (Jung et al. 2024; Schwengers et al. 2021). BUSCO with the parameters "-m genome -l rhizobium-agrobacterium_group_odb10", Quast v5.0.0 with the default parameters, and the Samtools v.1.19.2 subcommand depth were used to gather genome assembly statistics (Danecek et al. 2021; Gurevich et al. 2013; Simão et al. 2015). MEME v5.4.1 with the parameters "-dna -nmotifs 10" was used to identify motifs around putative Flye assembly errors identified by Polypolish polishing (Bailey and Elkan 1994). Motifs were filtered to those with an evalue < 1. Bandage v.0.8.1 with the default parameters was used to visualize assembly graphs (Wick et al. 2015). FastANI v1.1 with the default parameters was used to calculate pairwise average nucleotide identity (ANI) and cluster strains into groups based on 95% or 99% thresholds (Jain et al. 2018).

The assembly of strain CG160 was found to have an unusual number of errors. BBTools sendsketch.sh v39.06 with the default parameters was used to identify contaminants (Bushnell, Brian 2024). Sourmash v4.8.11 sketch and compare were used to generate and compare signatures for the contaminated plasmid and the plasmids of a closely

related strain (Brown and Irber 2016). BBTools seal.sh v39.06 with the parameters "pattern=out_%.fq ambig=first" and combined reference assemblies were used to filter out contaminated reads (Bushnell, Brian 2024). The assemblies of *Rhizobium rhizogenes* K84 (NCBI: GCF_000016265.1), *Rhizobium tumorigenes* B21/90 (NCBI: GCF_019355815.1), CG160 without the contaminated plasmid, and the equivalent plasmid from closely related strain CG159 were used as input to partition reads. Non-contaminant reads mapping to the CG160 assembly and CG159 plasmid were combined and used in the previously described analyses.

Phylogenetic analysis

NCBI Datasets was used to download all publicly available genomes of the *Agrobacterium*/*Rhizobium* complex from NCBI on July 19th, 2023 (O'Leary et al. 2024). Automlsa2 v0.8.1 with the parameters "--allow_missing 4 --missing_check" was used to generate a multi locus sequence analysis (MLSA) phylogeny including newly sequenced and NCBI strains (Davis II et al. 2016). The protein sequences AcnA, AroB, CgtA, CtaE, DnaK, GlgB1, GlyS, HemF, LeuS, LysC, MurC, PrfC, RecR, RplB, RpoB, RpoC, SecA, SMc00019, SMc00714, SMc01147, SMc01148, SMc02059, SMc02478, ThrA, and TruA from *Sinorhizobium meliloti* strain 1021 were used as references (Zhang et al. 2012). The R package ggtree was used to plot phylogenies (Yu et al. 2017).

Whole genome SNP calling and comparisons

Whole genome SNP analyses were performed using Illumina short reads, full-length Nanopore reads, or fragmented Nanopore reads. Seqkit v0.16.1 sliding with parameters "-W 150 -s 150" or "-W 400 -s 400" was used to split Nanopore reads into 150 bp and 400 bp fragments, respectively (Shen et al. 2016). Species groups were determined by ANI thresholds of 95% and 99% (Jain et al. 2018). Assemblies with the fewest contigs were selected as reference. Kingfisher v0.2.2 subcommand "get" and parameters "-m ena-ascp aws-http prefetch" was used to simulate Illumina reads for NCBI strains (Woodcroft et al. 2024). Bwa v0.7.17-r1188 mem with the parameter "-M" was used to map Illumina reads

to a reference within each group (Li 2013). Minimap2 v2.26-r1175 with parameter "-ax map-ont" was used to map Nanopore reads to a reference within each group (Li 2018).

GATK v3.7 HaplotypeCaller with the parameters "-ERC GVCF -ploidy 1 -allowPotentiallyMisencodedQuals" was used to call variants for each dataset (McKenna et al. 2010). GATK GenotypeGVCFs with the default parameters was used to combine partial variant calls across strains. GATK SelectVariants with parameter "-selectType SNP" was used to select SNP variants only. GATK VariantFiltration with parameter "--filterExpression QD < 2.0 SOR > 3.0 QUAL < 30.0 FS > 60.0 MQ < 40.0 MQRankSum < -12.5 ReadPosRankSum -8.0" was used to hard filter SNPs. The GATK tool SelectVariants was used with parameter "--excludefiltered". Graphtyper v2.7.3 with the parameter "--no_filter_on_proper_pairs" was used to call SNPs (Eggertsson et al. 2017). Bcftools v1.17 mpileup with parameters "-x -I -Q 13 -h 100 -M 10000 -a 'INFO/SCR,FORMAT/SP,INFO/ADR,INFO/ADF", subcommand call with parameters "-m -ploidy 1 -V indels", and the apply_filters.py script with parameters "-q 27 -M 55 -V 0.00001 -x 0.2 -K 0.9 -s 1 -d 5" were used to call SNPs (Danecek et al. 2021). Clair3 v.1.0.4 with parameters "--no_phasing_for_fa --include_all_ctgs --haploid_precise –gvcf" were used to perform SNP calls between strains within species groups (Zheng et al. 2022). Clair3 gvcf files were merged using the GATK pipeline and filtering steps. The R package poppr v.2.9.5 was used to calculate pairwise SNP distances and plot minimum spanning networks (Kamvar et al. 2014). Bcftools view with the parameter "--samples" was used to extract individual strain SNP calls for both short reads and long reads fragmented into 400 bp. These files were filtered to only contain non-reference SNP calls.

VCFCompare v.1.0 was used to calculate overall precision and accuracy scores for each strain with Nanopore reads (Nkambule 2023). Vcfdist v2.5.2 with the parameters "--max-qual 60 --min-qual 0 --cluster gap 50" and Illumina SNPs as the truth dataset was used to calculate accuracy scores for Nanopore SNPs across quality thresholds (Dunn and Narayanasamy 2023). For precision vs. recall curves, true positives, true negatives, false

positives, and false negatives were each summed across strains at each quality score threshold for calculation of precision and recall.

Acknowledgements

We would like to thank the California Department for Food and Agriculture (CDFA) for sharing *Agrobacterium* strains. We thank the Department of Botany and Plant Pathology for supporting the CQLS high performance computing infrastructure. This work was supported by startup funds from the Department of Botany and Plant Pathology at Oregon State University. This project was funded in part by the USDA National Institute of Food and Agriculture through the Western Integrated Pest Management Center, award SA 18-4060-41.

Competing interests

AJW and AR have received support from Oxford Nanopore Technologies to present findings at scientific conferences.

Bibliography

Andrews, S. 2024. FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bailey, T. L., and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2:28–36.

Bloomfield, M., Bakker, S., Burton, M., Castro, M. L., Dyet, K., Eustace, A., Hutton, S., Macartney-Coxson, D., Taylor, W., and White, R. T. 2024. Resolving a neonatal intensive care unit outbreak of methicillin-resistant *Staphylococcus aureus* to the SNV level using Oxford Nanopore simplex reads and HERRO error correction. 2024.07.11.603154. https://doi.org/10.1101/2024.07.11.603154.

Bogaerts, B., Van den Bossche, A., Verhaegen, B., Delbrassinne, L., Mattheus, W., Nouws, S., Godfroid, M., Hoffman, S., Roosens, N. H. C., De Keersmaecker, S. C. J., and Vanneste, K. 2024. Closing the gap: Oxford Nanopore Technologies R10 sequencing allows comparable

results to Illumina sequencing for SNP-based outbreak investigation of bacterial pathogens. Journal of Clinical Microbiology 62:e01576-23. https://doi.org/10.1128/jcm.01576-23.
Brown, C. T., and Irber, L. 2016. sourmash: a library for MinHash sketching of DNA. Journal of Open Source Software 1:27. https://doi.org/10/ghdrk5.
Bushnell, Brian. 2024. BBMap. https://sourceforge.net/projects/bbmap/.
Chiou, C.-S., Chen, B.-H., Wang, Y.-W., Kuo, N.-T., Chang, C.-H., and Huang, Y.-T. 2023. Correcting modification-mediated errors in nanopore sequencing by nucleotide demodification and reference-based correction. Commun Biol 6:1–9. https://doi.org/10.1038/s42003-023-05605-4.
Colquhoun, R. M., Hall, M. B., Lima, L., Roberts, L. W., Malone, K. M., Hunt, M., Letcher, B., Hawkey, J., George, S., Pankhurst, L., and Iqbal, Z. 2020. Nucleotide-resolution bacterial pan-genomics with reference graphs. bioRxiv 2020.11.12.380378. https://doi.org/10.1101/2020.11.12.380378.
Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., and Li, H. 2021. Twelve years of SAMtools and BCFtools. GigaScience 10:giab008. https://doi.org/10.1093/gigascience/giab008.
Davis II, E. W., Weisberg, A. J., Tabima, J. F., Grünwald, N. J., and Chang, J. H. 2016. Gall-ID: tools for genotyping gall-causing phytopathogenic bacteria. PeerJ 4:e2222. https://doi.org/10.7717/peerj.2222.
De Coster, W., and Rademakers, R. 2023. NanoPack2: population-scale evaluation of long-read sequencing data. Bioinformatics 39:btad311. https://doi.org/10.1093/bioinformatics/btad311.
Delahaye, C., and Nicolas, J. 2021. Sequencing DNA with nanopores: Troubles and biases. PLoS One 16:e0257521. https://doi.org/10.1371/journal.pone.0257521.
Dunn, T., and Narayanasamy, S. 2023. vcfdist: accurately benchmarking phased small variant calls in human genomes. Nat Commun 14:8149. https://doi.org/10.1038/s41467-023-43876-x.
Eggertsson, H. P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., Masson, G., Zink, F., Hjorleifsson, K. E., Jonasdottir, A., Jonasdottir, A., Jonsdottir, I., Gudbjartsson, D. F., Melsted, P., Stefansson, K., and Halldorsson, B. V. 2017. Graphtyper enables population-scale genotyping using pangenome graphs. Nature Genetics 49:1654–1660. https://doi.org/10/gbx7v6.
Everhart, S., Gambhir, N., and Stam, R. 2021. Population genomics of filamentous plant pathogens—a brief overview of research questions, approaches, and pitfalls. Phytopathology 111:12–22. https://doi.org/10.1094/PHYTO-11-20-0527-FI.

Gonzalez, D., Kozdon, J. B., McAdams, H. H., Shapiro, L., and Collier, J. 2014. The functions of DNA methylation by CcrM in *Caulobacter crescentus*: a global approach. Nucleic Acids Res 42:3720–3735. https://doi.org/10.1093/nar/gkt1352.

Goodwin, S., McPherson, J. D., and McCombie, W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333–351. https://doi.org/10.1038/nrg.2016.49.

Grossman, M. K., Rankin, D. A., Maloney, M., Stanton, R. A., Gable, P., Stevens, V. A., Ewing, T., Saunders, K., Kogut, S., Nazarian, E., Bhaurla, S., Mephors, J., Mongillo, J., Stonehocker, S., Prignano, J., Valencia, N., Charles, A., McNamara, K., Fritsch, W. A., Ruelle, S., Plucinski, C. A., Sosa, L. E., Ostrowsky, B., Ham, D. C., Walters, M. S., and for the Multistate Pseudomonas Outbreak Investigation Group. 2024. Extensively drug-resistant *Pseudomonas aeruginosa* outbreak associated with artificial tears. Clinical Infectious Diseases 79:6–14. https://doi.org/10.1093/cid/ciae052.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086.

Hall, M. B., Rabodoarivelo, M. S., Koch, A., Dippenaar, A., George, S., Grobbelaar, M., Warren, R., Walker, T. M., Cox, H., Gagneux, S., Crook, D., Peto, T., Rakotosamimanana, N., Lapierre, S. G., and Iqbal, Z. 2023. Evaluation of Nanopore sequencing for *Mycobacterium tuberculosis* drug susceptibility testing and outbreak investigation: a genomic analysis. The Lancet Microbe 4:e84–e92. https://doi.org/10.1016/S2666-5247(22)00301-9.

Hall, M. B., Wick, R. R., Judd, L. M., Nguyen, A. N., Steinig, E. J., Xie, O., Davies, M., Seemann, T., Stinear, T. P., and Coin, L. 2024. Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data. eLife 13:RP98300. https://doi.org/10.7554/eLife.98300.

Huang, Y.-T., Liu, P.-Y., and Shih, P.-W. 2021. Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing. Genome Biol 22:95. https://doi.org/10.1186/s13059-021-02282-6.

Iruegas-Bocardo, F., Weisberg, A. J., Riutta, E. R., Kilday, K., Bonkowski, J. C., Creswell, T., Daughtrey, M. L., Rane, K., Grünwald, N. J., Chang, J. H., and Putnam, M. L. 2023. Whole genome sequencing-based tracing of a 2022 introduction and outbreak of *Xanthomonas hortorum* pv. *pelargonii*. Phytopathology 113:975–984. https://doi.org/10.1094/PHYTO-09-22-0321-R.

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Communications 9:5114. https://doi.org/10/gfknmg.

Jung, J. M., Rahman, A., Schiffer, A. M., and Weisberg, A. J. 2024. Beav: a bacterial genome and mobile element annotation pipeline. mSphere 9:e00209-24. https://doi.org/10.1128/msphere.00209-24.

Kahng, L. S., and Shapiro, L. 2001. The CcrM DNA methyltransferase of *Agrobacterium tumefaciens* is essential, and its activity is cell cycle regulated. J Bacteriol 183:3065–3075. https://doi.org/10.1128/JB.183.10.3065-3075.2001.

Kamvar, Z. N., Tabima, J. F., and Grünwald, N. J. 2014. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ 2:e281. https://doi.org/10/c57j.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. 2019. Assembly of long, error-prone reads using repeat graphs. Nature Biotechnology 37:540–546. https://doi.org/10/gfzbrd.

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio].

Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Luan, T., Commichaux, S., Hoffmann, M., Jayeola, V., Jang, J. H., Pop, M., Rand, H., and Luo, Y. 2024. Benchmarking short and long read polishing tools for nanopore assemblies: achieving near-perfect genomes for outbreak isolates. BMC Genomics 25:679. https://doi.org/10.1186/s12864-024-10582-x.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303. https://doi.org/10/bnzbn6.

McLeish, M. J., Fraile, A., and García-Arenal, F. 2021. Population genomics of plant viruses: the ecology and evolution of virus emergence. Phytopathology 111:32–39. https://doi.org/10.1094/PHYTO-08-20-0355-FI.

Moore, L. W., Chilton, W. S., and Canfield, M. L. 1997. Diversity of opines and opine-catabolizing bacteria isolated from naturally occurring crown gall tumors. Appl Environ Microbiol 63:201–207. https://doi.org/10.1128/aem.63.1.201-207.1997.

Nkambule, L. 2023. VCFCompare. https://github.com/LindoNkambule/VCFCompare.

O'Leary, N. A., Cox, E., Holmes, J. B., Anderson, W. R., Falk, R., Hem, V., Tsuchiya, M. T. N., Schuler, G. D., Zhang, X., Torcivia, J., Ketter, A., Breen, L., Cothran, J., Bajwa, H., Tinne, J., Meric, P. A., Hlavina, W., and Schneider, V. A. 2024. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. Sci Data 11:732. https://doi.org/10.1038/s41597-024-03571-y.

Olson, N. D., Lund, S. P., Colman, R. E., Foster, J. T., Sahl, J. W., Schupp, J. M., Keim, P., Morrow, J. B., Salit, M. L., and Zook, J. M. 2015. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Front Genet 6:235. https://doi.org/10.3389/fgene.2015.00235.

Oxford Nanopore PLC. 2024. Dorado. https://github.com/nanoporetech/dorado.

Oxford Nanopore Technologies. 2023. Medaka. https://github.com/nanoporetech/medaka.

Palmer, B. R., and Marinus, M. G. 1994. The dam and dcm strains of *Escherichia coli*--a review. Gene 143:1–12. https://doi.org/10.1016/0378-1119(94)90597-5.

Parkhill, J., and Wren, B. W. 2011. Bacterial epidemiology and biology - lessons from genome sequencing. Genome Biology 12:230. https://doi.org/10.1186/gb-2011-12-10-230.

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., and Korobeynikov, A. 2020. Using SPAdes *de novo* assembler. Current Protocols in Bioinformatics 70:e102. https://doi.org/10/gg8s3z.

Sánchez-Romero, M. A., and Casadesús, J. 2020. The bacterial epigenome. Nat Rev Microbiol 18:7–20. https://doi.org/10.1038/s41579-019-0286-2.

Schwengers, O., Jelonek, L., Dieckmann, M. A., Beyvers, S., Blom, J., and Goesmann, A. 2021. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microbial Genomics 7:000685. https://doi.org/10.1099/mgen.0.000685.

Sereika, M., Kirkegaard, R. H., Karst, S. M., Michaelsen, T. Y., Sørensen, E. A., Wollenberg, R. D., and Albertsen, M. 2022. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods 19:823–826. https://doi.org/10.1038/s41592-022-01539-7.

Shen, W., Le, S., Li, Y., and Hu, F. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One 11:e0163962. https://doi.org/10.1371/journal.pone.0163962.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

Stimson, J., Gardy, J., Mathema, B., Crudu, V., Cohen, T., and Colijn, C. 2019. Beyond the SNP threshold: identifying outbreak clusters using inferred transmissions. Mol Biol Evol 36:587–603. https://doi.org/10.1093/molbev/msy242.

Straub, C., Colombi, E., and McCann, H. C. 2021. Population genomics of bacterial plant pathogens. Phytopathology 111:23–31. https://doi.org/10.1094/PHYTO-09-20-0412-RVW.

Vasu, K., and Nagaraja, V. 2013. Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense. Microbiology and Molecular Biology Reviews 77:53–72. https://doi.org/10.1128/mmbr.00044-12.

Weisberg, A. J., Davis, E. W., Tabima, J., Belcher, M. S., Miller, M., Kuo, C.-H., Loper, J. E., Grünwald, N. J., Putnam, M. L., and Chang, J. H. 2020. Unexpected conservation and global transmission of agrobacterial virulence plasmids. Science 368:eaba5256. https://doi.org/10.1126/science.aba5256.

Weisberg, A. J., Grünwald, N. J., Savory, E. A., Putnam, M. L., and Chang, J. H. 2021. Genomic approaches to plant-pathogen epidemiology and diagnostics. Annual Review of Phytopathology 59:311–332. https://doi.org/10.1146/annurev-phyto-020620-121736.

Wick, R. R., and Holt, K. E. 2022. Polypolish: Short-read polishing of long-read bacterial genome assemblies. PLOS Computational Biology 18:e1009802. https://doi.org/10.1371/journal.pcbi.1009802.

Wick, R. R., Judd, L. M., Gorrie, C. L., and Holt, K. E. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLOS Computational Biology 13:e1005595. https://doi.org/10/gbj5jz.

Wick, R. R., Judd, L. M., and Holt, K. E. 2023. Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing. PLOS Computational Biology 19:e1010905. https://doi.org/10.1371/journal.pcbi.1010905.

Wick, R. R., Schultz, M. B., Zobel, J., and Holt, K. E. 2015. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics 31:3350–3352. https://doi.org/10.1093/bioinformatics/btv383.

Williams, R. J. 2003. Restriction endonuclease. Mol Biotechnol 23:225–243. https://doi.org/10.1385/MB:23:3:225.

Woodcroft, B. J., Cunningham, M., Gans, J. D., Bolduc, B. B., and Hodgkins, S. B. 2024. Kingfisher: A utility for procurement of public sequencing data. https://zenodo.org/doi/10.5281/zenodo.10525085 (accessed 18 Sep 2024).

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T.-Y. 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 8:28–36. https://doi.org/10/f9qv8x.

Zhang, Y. M., Tian, C. F., Sui, X. H., Chen, W. F., and Chen, W. X. 2012. Robust markers reflecting phylogeny and taxonomy of rhizobia. PLOS ONE 7:e44936. https://doi.org/10.1371/journal.pone.0044936.

Zheng, Z., Li, S., Su, J., Leung, A. W.-S., Lam, T.-W., and Luo, R. 2022. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. Nat Comput Sci 2:797–803. https://doi.org/10.1038/s43588-022-00387-x.