No part gets left behind: Tiled nanopore sequencing of whole ASFV genomes stitched together using Lilo

Running Title: Tiled amplicon sequencing with improved assembly of African Swine Fever Virus

Amanda Warr[a], Caitlin Newman[a], Nicky Craig[a], Ingrida Vendelė[a,♦], Rizalee Pilare[b], Lilet Cariazo Cruz[b], Twinkle Galase Barangan[b], Reildrin G Morales[d], Tanja Opriessnig[a], Virginia Mauro Venturina[b], Milagros R Mananggit[c], Samantha Lycett[a], Clarissa YJ Domingo[b], Christine Tait-Burkard[a#]

[a], The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian, UK

[b], College of Veterinary Science and Medicine, Central Luzon State University, Science City of Muñoz, Nueva Ecija, The Philippines

[c], Regional Animal Disease Diagnostic Laboratory, Department of Agriculture Regional Field Office III, City of San Fernando, Pampanga, The Philippines

[d], Bureau of Animal Industry, Department of Agriculture, Visayas Avenue, Diliman, Quezon City, The Philippines

♦Current address: Thermo Fisher Scientific, Vilnius, Lithuania

Address correspondence to Christine Tait-Burkard, christine.burkard@roslin.ed.ac.uk .

Abstract

African Swine Fever virus (ASFV) is the causative agent of a deadly, panzootic disease, infecting wild and domesticated suid populations. Contained for a long time to the African continent, an outbreak of a particularly infectious variant in Georgia in 2007 initiated the spread of the virus around the globe, severely impacting pork production and local economies. The virus is highly contagious and has a mortality of up to 100% in domestic pigs. It is critical to track the spread of the virus, detect variants associated with pathology, and implement biosecurity measures in the most effective way to limit its spread. Due to its size and other limitations, the 170-190kbp large DNA virus has not been well sequenced with fewer than 200 genome sequences available in public repositories. Here we present an efficient, low-cost method of sequencing ASFV at scale. The method uses tiled PCR amplification of the virus to achieve greater coverage, multiplexability and accuracy on a portable sequencer than achievable using shotgun sequencing. We also present Lilo, a pipeline for assembling tiled amplicon data from viral or microbial genomes without relying on polishing against a reference, allowing for structural variation and hypervariable region assembly other methods fail on. The resulting ASFV genomes are near complete, lacking only parts of the highly repetitive 3'- and 5'telomeric regions, and have a high level of accuracy. Our results will allow sequencing of ASFV at optimal efficiency and high throughput to monitor and act on the spread of the virus.

Main

African Swine Fever (ASF) is a viral hemorrhagic disease leading to an extremely high mortality of up to 100% within 7-10 days. Clinical signs of infection are non-specific, including fever, ataxia, anorexia, cyanosis, respiratory symptoms, gastrointestinal symptoms and death [1]. The causative agent of ASF is African Swine Fever Virus (ASFV), the only member of the *Asfivirus* genus and the *Asfarviridae* family. The virion is large and complex with a diameter of around 175-215nm containing a large, double stranded DNA genome of around 170-190kb encoding over 150 open reading frames [1,2].

ASF is endemic in Africa where a sylvatic cycle between *Ornithodoros spp.* soft ticks and the natural reservoir, African wild suids, maintains its presence [3]. Other routes of infection and spread include physical contact, fluids, excretions, contaminated feed and fomites. The virus is extremely resilient and can survive for prolonged periods in a range of environmental conditions in carcasses and pork product. The disease is highly contagious and can be transmitted through relatively low infectious dose in feed and water [4]. Wild suid species are susceptible to disease but only domestic and feral pigs, as well as Eurasian wild boar show symptoms. ASFV is therefore very difficult to contain.

ASF was first discovered in East Africa, with symptoms reported in Kenya in 1914 and the disease described in 1921 [5]. Outbreaks in other parts of Africa, Europe, Brazil and the Caribbean islands occurred in the 20th century, with African countries being the worst affected. The virus was almost completely eradicated from non-African countries by the end of the century, but an outbreak in the Republic of Georgia in 2007 has since lead to widespread outbreaks in other countries. Since 2018 major outbreaks have been occurring in China, the world's largest pork producer, and it has been reported that up to half of the pigs in the country, representing roughly a quarter of the world's population, died or were culled to contain the outbreak in 2019 [6]. The spread of the virus in East and Southeast Asia however could not be halted and since further countries including Mongolia, Vietnam, Cambodia, North Korea, Laos and island nations including The Philippines, Indonesia,

Timor-Leste and Papua New Guinea have reported outbreaks [7,8]. After the initial outbreaks in Eastern Europe and continuing spread, the disease reached the European Union in 2014 and continues to spread, not least through the wild boar population [7,9]. In late 2020 the virus reached the largest producer of pork in Europe, Germany [10]. The virus is also edging closer to the USA, one of the world's main exporters and importers of pork, having been recently detected in Haiti and the Dominican Republic, only 381km by air from the US territory of Puerto Rico. It is clear, that the disease represents a serious panzootic threat impacting the pork industry and threatening economies already shaken by the SARS-CoV-2 pandemic.

To understand the genetic and genomic variation for ASFV, sequencing is primarily focused on a roughly 400bp fragment of the B646L gene, encoding for the major capsid protein p72. This fragment, representing <0.25% of the total genome, is the basis of the current genotyping system, which has identified 24 genotypes so far [11-14]. To further discriminate, additional fragments of the E183L gene (p54, ~630bp), CP205L (p30, ~510bp), and B602L (gp83 ~800bp) are used, adding up to less than 1.4% of the genome characterized. Whilst being a DNA virus, antigenic diversity, the ability to acquire large deletions or insertions, and the presence of highly mutagenic hypervariable regions urge the need for whole genome sequencing for virus characterization and epidemiological studies [15,16]. To do this at the scale required, there is a need for a cheap and efficient method to sequence the large ASFV genome, whilst high abundance of homopolymers and hypervariable region require highest accuracy [17].

The availability of a portable sequencing technology opens new doors to travel to outbreak locations, sequence, and analyze samples without needing to transport them. The MinION sequencer from Oxford Nanopore Technologies (ONT) can be carried easily in a pocket or carryon bag. This avoids complications of challenging transportation of biological samples, highly contagious agents or the requirement of a cold chain. There is the potential for a fast turnaround from sample collection to analysis, allowing for near live-monitoring of outbreak situations, as observed in during the Western African Ebola virus epidemic 2013-2016, or of

course the COVID-19 pandemic. Furthermore, multiplexing and the washing and reuse of the most expensive component of sequencing, the flow cells, allowing for cheaper sequencing than other methods. Finally, the sequencer can produce very long reads which improves assembly potential, particularly of highly repetitive genomes.

Whilst it is possible to obtain whole genome sequences of ASFV directly from blood- and tissue extract DNA, the high prevalence of pig DNA and the need for baits or other methods to enrich ASFV DNA render that method inapplicable for high-throughput, fast, sequencing.

Here, we present a method to sequence the near complete genomes, excluding only the highly repetitive, variable length telomeric 3' and 5' regions, of ASFV using ONT's MinION sequencing device using a tiled amplicon approach. The genome is amplified in 32 large fragments 7kb in length, amplified simultaneously in two PCR pools. We propose this method as an efficient, highly adaptable, more accurate, fast, and cost-effective option for sequencing of continuing ASFV outbreaks as well as historic samples. We present 10 complete ASFV genome assemblies from samples from the early stages of the ASFV outbreak in the Philippines in 2019 assembled either with the tiled sequencing approach or a whole genome sequencing shotgun approach. The portability of Nanopore sequencing makes it ideal for exploring the dynamics of ASFV infections as outbreaks emerge. As ASFV continues to spread around the world, efficient methods of sequencing the genome are essential to improve our understanding of the virus and the ongoing global spread. Our primer sets have been optimized for relatively even coverage and have been designed to bind outside of hypervariable regions. They only anneal to roughly 0.8% of the genome and are designed to be well suited to the current outbreak, able to at least partially sequence other genotypes and be easily modifiable should the virus mutate.

Finally, we present the Lilo pipeline. While pipelines exist to assemble genomes from tiled amplicons, they rely on aligning reads to a reference and using polishing tools to generate a consensus from the reads. This method works well for producing a genome sequence with SNPs representative of the sequenced genome, however large indels, structural variants,

and hypervariable regions that may be difficult to align to a reference are not accurately represented. For ASFV, whole genes can be inserted or deleted and due to homologous recombination it can carry large structural variations, with indels likely being more important than SNPs in creating viral diversity [18]. Therefore, we designed Lilo, which aligns reads to a reference in order to assign them to an amplicon, selects the read with the highest base quality and of the expected length for each amplicon, polishes the read with the remaining reads, removes primers and stitches them together at overlaps ordered and oriented by a reference. This approach makes the pipeline more adaptable to large structural variation and hypervariable regions in genomes than currently available methods.

Shotgun sequencing of ASFV directly from blood

In field sequencing, particularly in developing countries, limits the availability of tools and reagents. During the first outbreaks in the Philippines whole DNA was isolated from the highly hemolysed blood collected from ASFV positive pigs. Samples were digested overnight with proteinase K at 55°C prior to phenol/chloroform/isoamyl alcohol extraction and precipitation with isopropanol before washing with 70% ethanol. Whole DNA samples were prepared for sequencing using the ligation sequencing kit (LSK) LSK109 before sequencing samples on a R9.4 flow cell using a MinION mk1b. The data were basecalled and demultiplexed using Guppy (ONT) and the reads assembled with Flye and polished with medaka. (Figure 1A)

The time between the beginning of sequencing and detection of the first ASFV read from whole blood ranged from 19 seconds to 3 minutes. As seen in the example of PHL-1969 (Figure 1B) the percentage of reads that came from ASFV ranged from 0.006% to 0.24%, likely dependent on the viral titers of the animals culled. ASFV samples show a similar size distribution to other DNA found in the samples, if anything a second small peak of larger fragments can be observed (Figure 1C). All four sequenced blood samples assembled into a

149 whole genome, however, due to variable coverage, the number of mismatches and indels
150 found in some of the samples were high (Figure 3B).

151

152 **Tiled amplicon sequencing of ASFV**

153 Given the low yield of ASFV sequences from shotgun sequencing, as demonstrated by us
154 and others [19-21], and the high expense per sample, this sequencing approach was not fit for
155 purpose for high-throughput screening of an ongoing virus outbreak. Therefore, we
156 developed a method to amplify, sequence, and assemble ASFV genomes from pigs.

157 In order to enrich ASFV from the sample easily, a PCR amplification approach was chosen,
158 due to its ease of use and usually readily available tools in many countries and labs. Tiling
159 primers were designed targeting 7kb amplicon length and 1kb amplicon overlap using primal
160 scheme using a set of 26 ASFV reference sequences (Figure 2A). The primers are well
161 suited to genotype II, from the current outbreak, but also cover the majority of the genome
162 for at least genotypes I and IV (Figure 2B). This relatively long amplicon size was chosen to
163 reduce the number of primer pairs but also to span potential hypervariable regions. After
164 initial individual performance tests, several primers were redesigned from the original set of
165 primers produced by primal scheme, however the majority of them worked well from the
166 beginning. Fragments were amplified using the PCRBio VeriFi Hot Start high fidelity
167 polymerase according to the manufacture's instruction. Following redesign, all primers
168 amplified their targets, however, they did so at different efficiencies leading to uneven
169 coverage over the genome. To test this, evenly concentrated pools of primers (pool 1 and
170 pool 2, Figure 2A and Figure 2C) were used to amplify blood DNA extract samples from
171 ASFV-infected pigs. Following initial amplifications, pools were split into three pools with
172 primer pair 1, producing a shorter 4kb fragment continuously outperforming the others in a
173 mixed reaction on its own, and primer concentrations in pool 1(Pair 1) and pool 2 were
174 gradually adjusted according to their performance. PCR products per sample were

combined, libraries prepared using the LSK109 kit in an R9.4 flow cell. Figure 2D demonstrates the improvement that can be gained by tweaking primer concentrations from evenly represented primer pairs (purple) to optimized primer concentrations (green). These optimizations improve performance for multiplexing of multiple samples on one flow cell.

Fresher samples amplify more cleanly, but older, degraded samples will still amplify sufficiently. To show this, we highlight two samples; sample PHL-126, which has been heavily used and degraded, and sample PHL-261, which has been used less frequently (aliquot stored in freezer without frequent use) and is of better quality. As can be seen in the automated electrophoresis result of a tapestation (Figure 2E), PHL-126 shows poor amplification and relatively many amplicons <7kb. Good amplification can be seen for the shorter amplicon pair 1 still. PHL-261 on the other hand shows continued good amplification of the desired 7kb and 4kb products of pool 1 (odd), pool 2 (even) and pair 1, respectively. These samples were prepared with the LSK109 kit and multiplexed using native barcoding and run on a R9.4 flow cell with 3 other ASFV genomes having been pooled in representative quantities, the poorer amplification of PHL-126 had lower sequencing throughput than the better quality PHL-261, but was still assembled into a near-complete genome. Figure 2F shows sequencing coverage of the same two samples and the proportion of total reads for each that was assigned to each of the 32 amplicons.

The post-amplification DNA Integrity Number (DIN) can be used to help predict multiplexability, as different quality samples will impact the needed throughput. Figure 2G shows the relationship between the post-amplification DIN and throughput <3kb, and Figure 2H demonstrates the number of gaps for different throughputs of reads >3kb. Samples PHL-126 (orange) and PHL-261 (blue) have been highlighted in pale (super accuracy base calling (SAC)) and strong (high accuracy base calling (HAC)) colors, respectively. Figure 2H demonstrates the required throughput per sample and the gaps that can be expected in genomes produced with lower throughputs. Assuming a theoretical MinION throughput of 30GB, it should be possible to multiplex over 24 samples, and potentially up to 48 samples

(including PCR control), however we would recommend starting with fewer and assessing achievable throughput for each sequencing location as there is variability in expected throughput between users, flow cells, and geographical location. The integrity of the sample will also impact the throughput with degraded samples leading to sequencing capacity being taken up by shorter fragments instead of the required full length amplicons (Figure 2G). For very poor samples, more stringent size selection with AMPure XP beads prior to sequencing may be necessary if samples are to be multiplexed. While it is impractical to run an automated, high resolution electrophoresis, such as a tapestation, after every amplification, users can test a typical sample type (e.g. from decomposing wild boar, from farm culled pigs, with/without cold chain) to predict the likely multiplexability and clean-up steps of similar samples.

Lilo assembly of genomes from tiled amplicons

Comparing ASFV genomes we found major variation of the genome often originating from indels. Available assembly pipelines were struggling with such variation when it did not correspond to the reference sequence. Therefore, we developed the Lilo pipeline (Figure 3A) to assemble the tiled amplicons (https://github.com/amandawarr/Lilo). Whilst Lilo uses a reference alignment to sort the amplicons, it polishes against the highest quality reads rather than a reference sequence. Using this pipeline, highly accurate genomes were obtained with mismatch accuracy approaching Q50 when using SAC (Figure 2I) and indel accuracy up to Q40 when compared to a closely related publicly available ASFV genome assembly (MN715134.1) [21], which may still be quite divergent from these samples in truth.

QUAST [22] (v5.0.2; quality assessment tool for genome assemblies) results demonstrate that the increased coverage of the tiled amplicons produced a more accurate assembly than shotgun sequencing of the virus using a whole flow cell sequencing directly from extracted DNA. Shotgun sequencing however, was able to highlight some samples with longer

telomeric regions, such as PHL-237, which is a clear advantage of long-read sequencing technology and something that should be explored for more in-detail investigations into the role of the ASFV telomeric regions. Overall, SAC produced fewer mismatches and indels than HAC and should be the preferred method, however, the time for base calling is a trade-off. Samples with high percentages of unassigned bases (N's) clearly correspond to DIN numbers (Figure 3B).

The assembled genomes had excellent agreement on genome structure with the same samples assembled from shotgun sequencing (Figures 3B & 3C). The repetitive content shown at the edges of Figures 3B and 3C are sequences from the telomeres, showing that despite the sequences being from tiled amplicons, they do cover the majority of the genome and part of the telomeres.

Accuracy of Lilo and ARTIC assembled genomes

We assessed the quality of Lilo assemblies against those produced with the ARTIC pipeline (v1.2.1). A selection of the ASFV sequencing data were assembled using the ARTIC pipeline, as well as using Lilo, both using the assembled shotgun sequence PHL-1969 as a reference.

QUAST analysis shows lower numbers of mismatches against the closest reference (MN715134.1) but higher indels. The percentage of unassigned bases is much higher for ARTIC at around 2.4% whereas Lilo is at 0 or nearly 0%.(Figure 3B)

Comparing Lilo-assembled genomes and ARTIC-assembled genomes to a reference (MN715134.1) a number of indels can be observed. Figure 4A shows a likely real indel in the PHL ASFV samples which all assemblies agree on and which is well supported by the reads. In contrast, Figure 4B shows the only indel unique to almost all of the assemblies produced by the Lilo pipeline while being absent from all artic assemblies and occurs in a homopolymer, Most reads appear to support the deletion assembled by Lilo, whether this is

a real sequence or a result of poor accuracy of Nanopore sequencing of homopolymeric regions is a more difficult question. Figure 4C shows an extreme example of a very long homopolymeric region, ASFV has several of these and typically neither assembly method agrees on the length of the homopolymer, with the reads lending no strong support to either assembly. While errors from the Lilo pipeline tended to be randomly dispersed among homopolymers, ARTIC errors tended to be more systematic, appearing consistently across the assembled genomes. Frequently, homopolymers lead to the ARTIC pipeline replacing the base immediately before the homopolymer and the first base of the homopolymer with a pair of N's, as can be seen in Figure 4E and 4F. There were also occasions when reads did not support an indel, and there wasn't a clear cause of an indel in the reads or reference, as can be seen in Figures 4D and 4E. There are several of these types of indels throughout the assembly, where Lilo assemblies better agree with both the reads and the reference, likely contributing to the lower QUAST scores of ARTIC on homopolymers and the percentage of undefined bases.

Phylogenomics

A maximum likelihood phylogenomic tree was constructed incorporating the newly assembled genomes from tiled data of R9.4 flow cells with HAC with publicly available whole genome ASFV sequences using iqtree (v2.0.5; Figure 5A), and for B646L gene, encoding for the major capsid protein p72, genotype II sequences specifically (Figure 5B).

As observed in Figure 5A, p72 genotypes do not correspond to the clustering. For example, the E75 strain Spain 1975 isolate, an early genotype II, is grouping with genotype I's. Unfortunately, the phylogeny contains many gaps and lacks both timely and geographic resolution, showing that much more sampling is required. PHL samples clearly cluster within the highly virulent, novel p72 genotype II cluster.

Resolving the tree further, selecting only those clustering with the novel p72 genotype II genomes two distinct clusters of PHL sequences can be observed. Whilst, due to the similarity of the genomes, the orders of lower branches are of lower confidence than those of higher branches, reanalysis still suggest two different introductions into the Philippines (internal branch lengths may be found in supplementary documents S1 and S2). Whilst indications are that one cluster is closer related to Asian isolates, the other one showing a more likely Eastern European origin, the lack of sample numbers to fill branch gaps and common ancestors makes conclusive interpretation impossible.

Discussion

ASFV is a serious threat to the global pork industry and consequences of depopulation, reduced availability of pork, and increased prices of other animal protein affect local economies, especially in low- and middle income countries with high reliance on pig protein. Tracing the spread of the virus, understanding more about genome-pathology links, and consequently implementing targeted biosafety measures are paramount to combat the disease. As demonstrated in Figure 5, current genotyping methods based on partial sequencing of the B646L gene, encoding for the major capsid protein p72, do not correspond to whole-genome sequencing and are therefore inadequate to trace virus evolution.

Here, we have demonstrated an efficient method for sequencing the ASFV genome. Despite the use of tiled amplicons, part of the 3'- and 5' telomeric regions of the virus are included in the tiled amplicon assemblies, meaning the majority of the genome is included.

As demonstrated by us and others, ASFV sequences can be obtained by direct sequencing from blood or other tissue samples of infected pigs [19-21]. The resulting sequence includes interesting information on the lengths and repeats found in the telomeric regions, which may be helpful for more in-depth investigation into the virus pathology and spread. However, without enrichment for ASFV [17] or depletion of host-methylated DNA [23] sample percentage for ASFV is low relative to host DNA in the samples, meaning that obtaining sufficient ASFV reads to assemble the genome from shotgun sequencing usually requires an entire MinION flow cell, or more, depending on viral titer and original sample type. Bone marrow or blood will likely yield the best virus:host ratio with spleen or muscle, whilst good sources of viral DNA [24], also contain a large number of nucleated host cells. Even if sufficient data is obtained to assemble the genome, the coverage is likely too poor to sufficiently polish the genome. In contrast, the tiled amplicon method can be used on samples with lower viral titers or degraded DNA, selectively sequences the virus, and can be multiplexed on a flow cell to simultaneously sequence multiple samples at high enough coverage for good

polishing. Especially in countries where ASFV is circulating in wild boar or feral pigs, samples may be collected from infected animals that have been dead for a prolonged period of time. It is important that the method is capable of amplifying virus from both high- and low quality samples. Figure 2E demonstrates the variability of DNA integrity post-amplification and that even poor samples that have been degraded amplify and produce near complete genome assemblies.

Overall, the PCR amplification method increases coverage, is less prone to exhaust flow cells quickly, allows for multiplexing, and consequently reduces costs, improves genome accuracy, and removes the need for specialized enrichment or depletion methods.

Whilst ~7kb amplicons are very large compared to other comparable methods for other viruses, the size of the ASFV genome, the stability of DNA, the relatively low numbers of primer pairs, and the advantages of long reads detecting recombinants more easily make this the best approach. Especially with the small, medium, and large indels that can occur in ASFV [18], it is important to get good resolution across these regions, which can be achieved easily with large amplicons. It is important though to choose the right, high accuracy polymerase capable of amplifying such long amplicons. We found PCRBio VeriFi to be highly capable of this with the hot start version producing very few non-specific products, whilst the non-hot start version can produce more non-specific product, which may be an advantage for variant testing. As demonstrated in Figures 2E and F and 4B show that even low quality samples can produce whole genome assemblies with few gaps. However, a limitation of the large tiled amplicon method is that should a variant occur at the site of a primer, the amplification of a relatively large section of genome will fail. While this is an inconvenience, it will be simple to redesign a primer to replace the failed one or to act as an alternate primer. It is also possible to amplify across a larger region using the existing primers either side of the failed one, generating a 14kb product, to sequence a larger region and design a primer from the sequenced amplicon. This was found to be possible using the VeriFi HS polymerase and allows for the method to adapt as the virus changes.

Whilst Nanopore sequencing methods provide a lot of advantages, such as sequencing on site, portability, and accessibility to less specialist communities, there are, as for any sequencing method, drawbacks.

As demonstrated in Figure 1I & 1J the choice of flow cell and basecalling method has an impact on the accuracy of SNPs and indels. Generally, throughput is lower and required input DNA is higher for R10.3 flow cells, however HAC on R10.3 is of comparable accuracy to SAC on R9.4, with super accuracy on R10.3 being of even higher quality. SAC is very slow and resource hungry, and when speed is important such as in an active outbreak, R10.3 with HAC may be a good compromise for maximising efficiency with minimal sacrifice of accuracy. However where the highest accuracy is required we recommend using R10.3 flow cells and basecalling with a super accuracy model.

All of our assemblies have indels compared to the reference partially stemming from systematic errors in Nanopore sequencing in the abundant homopolymers and repeats in the ASFV genome (e.g. PHL-1969 contains 6.8% homopolymers of 4bp or more), however with the latest kits and flow cells from Nanopore, and without additional costs, we would expect future developments to continually reduce the indel errors without major alterations to the wet or dry lab methodologies described here. Additionally, some indels are likely true variations, and in many cases these indels are clearly present across all of our genomes (Figure 4A), and are not in homopolymeric regions, suggesting true variation between samples and the reference used. Additional sequencing of the same amplicons with a higher accuracy technology, such as Illumina could be used to polish the assemblies, however this would add time and cost. Reducing the number of false indels, where possible, is important, as the ASFV genome is known to have functionally relevant indels [25]. Polishing with an accurate reference can produce assemblies that are very accurate, however, these methods do not handle structural variants and hypervariable regions well. While the genomes sequenced here do not have any major indels compared to the reference used, diversity in ASFV is partially driven by small, medium and large indels [18] and increased sequencing of

368 samples is likely to reveal more of them. While errors from the Lilo pipeline tended to be
369 randomly dispersed among homopolymers, ARTIC errors tended to be more systematic,
370 appearing consistently across the assembled genomes. Errors occurring in the same
371 position between genomes may be more likely to impact phylogenomic analysis than
372 relatively random errors. The only consistent indel error found across the majority of the Lilo
373 assembled genomes that was always absent in the artic genomes is shown in Figure 4B.
374 This region contains a homopolymers, which is typically difficult to correct from Nanopore
375 sequencing data, however while the ARTIC assembly more closely agrees with the
376 reference, the reads are well-supporting of the deletion found in the Lilo assemblies. It is not
377 unusual when carrying out multi-sequence alignments between whole ASFV genome
378 sequences, even those constructed from reads from a higher accuracy sequencing
379 technology, to find large homopolymers of variable length and it is unclear to what degree
380 these are limitations of sequencing technologies as opposed to real variation.

381 The Lilo pipeline also has some limitations, it currently assumes that any structural variants
382 will not change the length of any given amplicon by more than 5%, it assumes that structural
383 variants will not be dramatic enough to prevent alignment to the reference for the purposes
384 of assigning reads to amplicons and ordering and orienting the polished amplicons. Lilo also
385 assumes the reads will be the full length of the amplicon, making it incompatible with ONT
386 rapid kits that utilize transposases. However, the strength of not relying on polishing reads
387 aligned to a reference is beneficial for genomes where structural variation is expected to be
388 important, and for species with hypervariable regions which may not align and polish well
389 with a reference. The pipeline has been tested on tiled sequences from ASFV, Porcine
390 Reproductive and Respiratory Syndrome-1 & -2, and SARS-CoV-2 (data not shown here)
391 and can handle custom schemes for other viruses.

392 ASFV is very under-sequenced with only a small number of whole genome sequences
393 available and there is a need for an affordable way to sequence the virus at scale, as
394 previously discussed [17]. While the majority of sequences produced from sequencing

individual genes in the virus have been frustratingly similar [26] reducing their usefulness in epidemiological studies, variants and large deletions have been observed across the genome, and these have been found to affect phenotypes [25]. As outbreaks continue to spread around the world and the amount of virus in circulation increases, these variations will likely increase in frequency. Additionally, there are few of the ancestral viruses from Africa sequenced, and these should be sequenced to understand the evolution of ASFV, particularly the loss of its dependence on the sylvatic cycle. Given the slow mutational rate of the virus, sequencing individual genes is unlikely to be informative and so to have a chance of seeing variants in the virus the whole genome must be sequenced. The ability to amplify the genotypes with our current scheme decreases with distance from genotype II, and additional primers will need designing in the future to improve coverage over other genotypes, however current coverage using this primer scheme is still likely to be of more use than the p72 gene alone. Coverage gaps can be resolved relatively easily as larger amplicons can be generated with flanking primers. Should primers on older or emerging samples fail, the altered region can be amplified using primers from either side of the failed amplicon, spanning the region, and the sequenced amplicon can be used to design new primers for the region.

Phylogenetic trees of currently available whole-genome ASFV sequences highlight the inadequacy of the p72 genotyping in reflecting similarities of ASFV on a genomic level. The sparsity of whole-genome sequences hampers the ability to trace virus movement and results in high levels of uncertainty in phylogenetic analyses. Running the maximum likelihood tree analysis including the whole-genome sequences obtained from tiled amplification in this manuscript reliably grouped samples from the early Philippines outbreak of ASFV in November 2019 into two clusters. This indicates at least two potential introductions of ASFV into the Philippines. Due to the lack of samples and resolution in the phylogenetic tree, no conclusion about countries or region of origins is possible.

We have presented an efficient, low cost method for sequencing and assembling ASFV which can be carried out in the lab or in the field during outbreaks. The Lilo pipeline is a lightweight pipeline that can be run on a standard laptop with 16GB RAM and no internet connection, making it ideal for in field bioinformatic analysis of ASFV and other viruses.

Methods

Samples

Blood samples from outbreaks in central Luzon (Philippines) were collected following depopulation of pigs within a defined containment radius. Blood samples were tested for ASFV by PCR. Blood samples from ASFV-positive pigs were pooled at equal amounts by farm before further processing.

DNA extraction

Blood samples were spun for 20min at 3,000rcf before decanting the supernatant. 5xTEN buffer(0.05M EDTA, 0.5M NaCl, 20mg/ml Proteinase K, 20% SDS, in 0.05M Trix-HCl, pH8.0) were added to a 1x final concentration before incubation overnight at 55°C in a shaking water bath. Equal volumes of phenol were added and gently mixed. Following 20min centrifugation at 3,000rcf the aqueous phase was transferred to a fresh tube. If the phase was very viscous, the phenol phase was re-extracted to improve yields. An equal volume of phenol/chloroform/isoamyl alcohol (25:24:1) was added to the aqueous phase before mixing and separation by centrifugation, 10min, 3,000rcf. The aqueous phase was transferred to a fresh tube before addition of 1:10 3M sodium acetate and an equal amount of isopropanol. Following 1h incubation at -20°C, samples were spun for 10min at 16,000rcf before washing the pellet with 70% Ethanol. The pellet was dried and resuspended in nuclease-free water.

Nanopore sequencing directly from DNA extracted from blood in the Philippines

Samples were sequenced following Nanopore's SQK-LSK109 protocol on R9.4 flow cells on a MinION mk1b. The protocol was started with 1ug of DNA as measured on an Implen NanoPhotometer P330. The protocol was carried out as recommended by ONT with the following modifications: The 20°C and 60°C incubations after the addition of NEB's FFPE repair and End-prep reagents were done for 30 minutes at each temperature instead of 5 minutes, and the room temperature incubation for the ligation reaction was done for 20 minutes instead of 10. During sequencing, two USB desk fans were pointed at the MinION to

assist with maintaining appropriate temperature for the run in the above average "room temperature" in the lab in the Philippines.

Designing primers for tiled amplification

Tiled primers were designed using Primal Scheme (v1.3.2) [27]. A set of 28 complete African Swine Fever genomes (listed in Supplementary Document S3) were downloaded from NCBI for primer design, which at the time were all that were available. Additionally three Filipino whole ASFV genomes we had assembled from shotgun sequencing data were included. A multi sequence alignment was carried out with Clustal Omega (in MEGA v7.0.2) [28]. Primal Scheme was run to produce 7kb amplicons with 1kb overlap resulting in 32 overlapping primer pairs in two non-overlapping pools. Primers were tested on samples and while the majority worked first time, several had to be redesigned due to failed amplification or preferential amplification of off-target regions. Redesigns were done using Primer-BLAST [29], targeting a similar region of the genome to the failed amplicon. Some primers amplified more efficiently than others and in order to make the coverage of these as even as possible, some primers were tweaked to have a different concentration. One primer pair (amplicon 1) was shorter than the others in order to avoid highly repetitive sequence in the telomeres and it is recommended to amplify it in a separate reaction to pools 1 and 2 to avoid overrepresentation.

Amplification, library prep and sequencing of tiled amplicons

Tiled primers were initially tested individually at 200nM concentration using approximately 90ng ASF DNA and Phusion High-Fidelity PCR Master Mix with HF Buffer with 1mM added $MgCl_2$ (both New England Biolabs, Ipswich, MA, USA). Individual PCRs, in a 25µl reaction volume, underwent initial denaturation of 2 minutes at 98°C, followed by 33 cycles of 10 seconds at 98°C, 30 seconds annealing at 63°C, and 4 minutes and 40 seconds extension at 72°C, followed by a final extension for 10 minutes at 72°C. PCR products were then examined using Tapestation Genomic DNA analysis (Agilent, Santa Clara, CA, USA). While

a small amount of off-target amplification was tolerated, primers which produced strong off-target bands or weak bands of the correct 7kb size were redesigned.

Once the complete set of primers had been successfully designed to cover the complete genome, the primers were pooled in equal amounts into two pools of non-overlapping primers. These pools were tested using the same conditions as the individual PCRs, but in a 50µl reaction volume and using 1µM of the primer pool. The resulting PCR products were cleaned using 0.4× volume AMPure XP beads (Beckman Coulter, Indianapolis, IN, USA) to remove products smaller than approximately 2kb in length, then pooled equally prior to sequencing. The cleaned PCR products were quantified using a Qubit ds DNA BR assay (Invitrogen, Waltham, MA, USA) and combined in equimolar amounts to a total of 700ng for library preparation according to the Native barcoding genomic DNA (with EXP-NBD104, EXP-NBD114, and SQK-LSK109)-Nanopore protocol.

Following bioinformatic analysis of sequencing data, primers which were found to be over- or under-performing were either redesigned or their contribution to the pool was adjusted accordingly, and the new primer pool tested as above in an iterative fashion. Ultimately 2 non-overlapping pools and a separate reaction for primer pair 1 were used to obtain the most even coverage and were processed as above, and pooled proportionally to the number of amplicons in each pool prior to sequencing. Additionally the polymerase was swapped from Phusion to VeriFi (PCRBIO) in a 25ul reaction using 2ul DNA per reaction, which has markedly better performance on the amplicons with far less off-target amplification. The PCR conditions for this polymerase were an initial denaturation of 1 minute at 98°C, followed by 40 cycles of 15 seconds at 98°C, 15 seconds annealing at 60°C, and 4 minutes and 40 seconds extension at 72°C, followed by a final extension for 5 minutes at 72°C. AMPure XP bead cleanup after PCR is optional, but recommended in samples with low DIN. Primer sequences, recommended primer concentrations and recommended pooling quantities are described in supplementary table S1, and any updates to these will be released on Lilo's github page.

Samples were sequenced following Nanopore's SQK-LSK109 or SQK-LSK110 protocol on R9.4 or R10.3 flow cells (which combination is specified alongside relevant results) on a MinION mk1b or mk1c. The protocol was started with 1ug of pooled amplicons as measured on a qubit using broad range reagents. For samples using multiplexing, the native barcoding expansion kit from Nanopore was used following Nanopore's instructions when using SQK-LSK109. For using the barcodes with SQK-LSK110, the instructions for SQK-109 were followed until after the barcodes had been ligated on, at which stage the end prep was repeated and we follow the standard protocol for library prep with SQK-LSK110 from after the end prep step.

Bioinformatic processing of ASFV genomes sequenced with shotgun sequencing

The data were basecalled and demultiplexed using MinKNOW (v19.06.8; ONT) using "fast" basecalling. Following basecalling the reads were aligned to an ASFV genome using minimap2 to identify ASFV reads, the fast5s for these reads were extracted using fast5_subset from the ont_fast5_api (https://github.com/nanoporetech/ont_fast5_api) and these were basecalled again using high accuracy basecalling. This was done to reduce basecalling time, as this work was done locally in the field on a laptop without a GPU. The reads were assembled with Flye (v2.6) [30] and polished 3 times with Medaka (v0.7.1; ONT). Comparisons of quantity of data produced and the proportion of which were ASFV reads were done using NanoComp (v1.28.1) [31].

Bioinformatic processing of ASFV genomes from tiled amplicons with Lilo

The data were basecalled and demultiplexed using Guppy (v5.0.14; ONT) using high or super accuracy model on a GPU. The snakemake pipeline, Lilo (https://github.com/amandawarr/Lilo), was developed and as summarised in Figure 3A, takes the following steps:

1. Use Porechop (v0.2.3) to remove any sequencing adapters or barcodes that have made it through demultiplexing.

2. Align to a reference with minimap2 (v2.22) [32] and samtools (v1.12) [33] and separate reads into amplicons by alignment position with bedtools (v2.30.0).[34]
3. Select reads of the expected amplicon length (+/-5%) and subset to 300X
4. Select the read with highest average base quality within +/-1% of the median length of reads for the amplicon to be the "reference" (with bioawk v1); https://github.com/lh3/bioawk), remove any amplicons with fewer than 40 reads. Targeting the median length allows for flexibility for large insertions or deletions.
5. Pool amplicon reads and references back into their original non-overlapping pools.
6. Polish the pools 3x with medaka (v1.4.4; ONT) and combine resulting polished amplicons.
7. Align to the reference with minimap2 and remove soft clipped bases (these likely represent missed barcodes or adapters)
8. Run porechop (specific fork: https://github.com/sclamons/Porechop-1) to remove primers from the amplicons.
9. Merge the amplicons with scaffold_builder (v2.3) [35].

The required input to Lilo are demultiplexed reads in fastq format in a directory named "raw/", a reference fasta, a bed file of primer alignments (as output by primal scheme), and a csv of primer sequences (if there are ambiguous bases it is advised to expand them first) and a config file, described on the github page. It is adaptable to any species (with a single genome fragment/chromosome) with any tiled primer scheme. The pipeline outputs a fasta file containing the assembled genome.

ARTIC assemblies

A subset of genomes were also assembled using the Artic pipeline (https://artic.network/ncov-2019; v1.2.1) following the bioinformatics SOP using the medaka method.

Quality control of assembled genomes

Quast (v5.0.2) was used to compare the assembled genomes to the most closely related publicly available ASFV assembly according to BLAST alignment (MN715134.1) [21]. Samples where both WGS and tiled sequencing were used were compared for overall structure using nucmer (v4.0.0beta2) [36].

Phylogeny

The phylogeny analysis was limited to the tiled genomes, as these were the most accurate assemblies, and publicly available genomes. These were aligned using Mafft (v7.467)[37] and maximum likelihood trees constructed using iqtree (v2.0.5)[38].

Acknowledgments

We would like to thank the Bureau of Animal Industry and Milagros Mananggit for providing us access to the valuable ASFV blood samples. We would also like to thank Central Luzon State University and Clarissa Yvonne Domingo and Virginia Venturina and their families for their amazing hospitality during all our visits to the Philippines.

We acknowledge financial support from the BBSRC Institute Strategic Programme grant funding to The Roslin Institute (BBS/E/D/20241866, BBS/E/D/20002172, and BBS/E/D/20002174) and BBSRC / Newton Fund Swine and Poultry research initiative grant (BB/R013187/1).

Author Contributions

Outlined the study AW, NC, IV, CYJD, CTB; collected and provided samples RGM, MRM, CYJD; performed experiments AW, NC, CN, IV, RP, LCC, TGB, and CTB; analyzed data AW, NC, SL, and CTB; interpreted data AW, NC, CN, SL, VMV, CYJD and CTB; wrote manuscript, AW and CTB, with contribution of other authors.

Competing Interests Statement

The authors declare no competing interests.

References

1 Blome, S., Franzke, K. & Beer, M. African swine fever – A review of current knowledge. *Virus Research* **287**, 198099, doi:10.1016/j.virusres.2020.198099 (2020).

2 Sánchez-Vizcaíno, J. M., Mur, L., Gomez-Villamandos, J. C. & Carrasco, L. An Update on the Epidemiology and Pathology of African Swine Fever. *Journal of Comparative Pathology* **152**, 9-21, doi:10.1016/j.jcpa.2014.09.003 (2015).

3 Jori, F. *et al.* Review of the sylvatic cycle of African swine fever in sub-Saharan Africa and the Indian ocean. *Virus Res* **173**, 212-227, doi:10.1016/j.virusres.2012.10.005 (2013).

4 Niederwerder, M. C. *et al.* Infectious Dose of African Swine Fever Virus When Consumed Naturally in Liquid or Feed. *Emerg Infect Dis* **25**, 891-897, doi:10.3201/eid2505.181495 (2019).

5 Eustace Montgomery, R. On A Form of Swine Fever Occurring in British East Africa (Kenya Colony). *Journal of Comparative Pathology and Therapeutics* **34**, 159-191, doi:https://doi.org/10.1016/S0368-1742(21)80031-4 (1921).

6 Tian, X. & von Cramon-Taubadel, S. Economic consequences of African swine fever. *Nature Food* **1**, 196-197, doi:10.1038/s43016-020-0061-6 (2020).

7 Penrith, M. L. Current status of African swine fever. *CABI Agriculture and Bioscience* **1**, 11, doi:10.1186/s43170-020-00011-w (2020).

8 Mighell, E. & Ward, M. P. African Swine Fever spread across Asia, 2018-2019. *Transbound Emerg Dis* **68**, 2722-2732, doi:10.1111/tbed.14039 (2021).

9 Gaudreault, N. N., Madden, D. W., Wilson, W. C., Trujillo, J. D. & Richt, J. A. African Swine Fever Virus: An Emerging DNA Arbovirus. *Front Vet Sci* **7**, 215, doi:10.3389/fvets.2020.00215 (2020).

10 Sauter-Louis, C. *et al.* Joining the club: First detection of African swine fever in wild boar in Germany. *Transbound Emerg Dis* **68**, 1744-1752, doi:10.1111/tbed.13890 (2021).

11 Bastos, A. D. *et al.* Genotyping field strains of African swine fever virus by partial p72 gene characterisation. *Arch Virol* **148**, 693-706, doi:10.1007/s00705-002-0946-8 (2003).

12 Quembo, C. J., Jori, F., Vosloo, W. & Heath, L. Genetic characterization of African swine fever virus isolates from soft ticks at the wildlife/domestic interface in Mozambique and identification of a novel genotype. *Transbound Emerg Dis* **65**, 420-431, doi:10.1111/tbed.12700 (2018).

13 Boshoff, C. I., Bastos, A. D., Gerber, L. J. & Vosloo, W. Genetic characterisation of African swine fever viruses from outbreaks in southern Africa (1973-1999). *Vet Microbiol* **121**, 45-55, doi:10.1016/j.vetmic.2006.11.007 (2007).

14 Achenbach, J. E. *et al.* Identification of a New Genotype of African Swine Fever Virus in Domestic Pigs from Ethiopia. *Transbound Emerg Dis* **64**, 1393-1404, doi:10.1111/tbed.12511 (2017).

15 Malogolovkin, A. & Kolbasov, D. Genetic and antigenic diversity of African swine fever virus. *Virus Res* **271**, 197673, doi:10.1016/j.virusres.2019.197673 (2019).

16 Netherton, C. L., Connell, S., Benfield, C. T. O. & Dixon, L. K. The Genetics of Life and Death: Virus-Host Interactions Underpinning Resistance to African Swine Fever, a Viral Hemorrhagic Disease. *Front Genet* **10**, 402, doi:10.3389/fgene.2019.00402 (2019).

17 Forth, J. H., Forth, L. F., Blome, S., Höper, D. & Beer, M. African swine fever whole-genome sequencing—Quantity wanted but quality needed. *PLOS Pathogens* **16**, e1008779, doi:10.1371/journal.ppat.1008779 (2020).

18 Zhu, Z. *et al.* Homologous recombination shapes the genetic diversity of African swine fever viruses. *Veterinary Microbiology* **236**, 108380, doi:https://doi.org/10.1016/j.vetmic.2019.08.003 (2019).

19 Jia, L. *et al.* Nanopore sequencing of African swine fever virus. *Sci China Life Sci* **63**, 160-164, doi:10.1007/s11427-019-9828-1 (2020).
20 O'Donnell, V. K. *et al.* Rapid Sequence-Based Characterization of African Swine Fever Virus by Use of the Oxford Nanopore MinION Sequence Sensing Device and a Companion Analysis Software Tool. *J Clin Microbiol* **58**, doi:10.1128/JCM.01104-19 (2019).
21 Olasz, F. *et al.* Short and Long-Read Sequencing Survey of the Dynamic Transcriptomes of African Swine Fever Virus and the Host Cells. *Front Genet* **11**, 758, doi:10.3389/fgene.2020.00758 (2020).
22 Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-1075, doi:10.1093/bioinformatics/btt086 (2013).
23 O'Donnell Vivian, K. *et al.* Rapid Sequence-Based Characterization of African Swine Fever Virus by Use of the Oxford Nanopore MinION Sequence Sensing Device and a Companion Analysis Software Tool. *Journal of Clinical Microbiology* **58**, e01104-01119, doi:10.1128/JCM.01104-19 (2021).
24 Fischer, M., Huhr, J., Blome, S., Conraths, F. J. & Probst, C. Stability of African Swine Fever Virus in Carcasses of Domestic Pigs and Wild Boar Experimentally Infected with the ASFV "Estonia 2014" Isolate. *Viruses* **12**, doi:10.3390/v12101118 (2020).
25 Gallardo, C. *et al.* Attenuated and non-haemadsorbing (non-HAD) genotype II African swine fever virus (ASFV) isolated in Europe, Latvia 2017. *Transbound Emerg Dis* **66**, 1399-1404, doi:10.1111/tbed.13132 (2019).
26 Gallardo, C. *et al.* Genetic variation among African swine fever genotype II viruses, eastern and central Europe. *Emerging infectious diseases* **20**, 1544-1547, doi:10.3201/eid2009.140554 (2014).
27 Quick, J. *et al.* Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nature Protocols* **12**, 1261-1276, doi:10.1038/nprot.2017.066 (2017).
28 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539, doi:10.1038/msb.2011.75 (2011).
29 Ye, J. *et al.* Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**, 134, doi:10.1186/1471-2105-13-134 (2012).
30 Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* **37**, 540-546, doi:10.1038/s41587-019-0072-8 (2019).
31 De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666-2669, doi:10.1093/bioinformatics/bty149 (2018).
32 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).
33 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
34 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
35 Silva, G. G. *et al.* Combining de novo and reference-guided assembly with scaffold_builder. *Source Code Biol Med* **8**, 23-23, doi:10.1186/1751-0473-8-23 (2013).
36 Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology* **14**, e1005944, doi:10.1371/journal.pcbi.1005944 (2018).
37 Katoh, K., Misawa, K., Kuma, K. i. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059-3066, doi:10.1093/nar/gkf436 (2002).

38 Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530-1534, doi:10.1093/molbev/msaa015 (2020).

Figure Legends

Figure 1

A) DNA was extracted from blood and sequenced with Nanopore's LSK109 on an m1kb before analysis and assembly using Flye and polishing with medaka. B) Read length histograms for the dataset demonstrating total throughput (blue) and ASFV reads throughput (orange) C) Normalized counts by dataset of reads for total throughput (blue) and ASFV reads throughput (orange). D) Throughput over time for total read count (blue) and ASFV read count (orange).

Figure 2

A) Design of the tiled primer scheme for ASFV with ~7kb amplicons and ~1kb overlaps. B) Predicted primer binding in correct region for representative ASFV genotypes. C) Workflow with extraction from blood, PCR amplification of primer pools, pooling, sequencing and bioinformatic analysis D) Coverage of one sample (PHL-3142) amplified with either evenly represented primer pairs (purple) or optimized proportions of primer pairs (green). E) Tapestation capillary electrophoresis of amplified pools of two different samples, the low quality PHL-126 and the high quality PHL-261, and statistics of the resulting assemblies of each of these. F) Coverage of amplicons using optimized primer concentrations for the two samples from figure 1E. G) Impact of post-amplification DIN on proportion of reads <3kb in length, essentially wasted sequencing capacity, using R9.4 flow cells and LSK-109 (magenta) or R10.4 and LSK-110 (blue). Semilog fit analysis shows an R squared correlation of 0.6420 or 0.7244, for R9.4_109 and R10.4_110, respectively. H) Association between overall throughput >3kb and number of gaps in final assembly using HAC (black/bold) and SAC (grey/faint). A log-log analysis shows an R squared correlation of 0.9680 and 0.9358 for HAC and SAC, respectively. PHL-126 is highlighted in orange and PHL-261 in blue. I) Assembly accuracy based on proportion of mismatches against reference (MN715134.1), with lines showing Q40 and Q50 PHRED scores. J) Assembly

718 accuracy based on proportion of indels against reference (MN715134.1), with lines showing
719 Q30 and Q40 PHRED scores. A&C Created with BioRender.com

720 **Figure 3**

721 A) Directed acyclic graph showing the steps the Lilo pipeline takes during assembly. The
722 graph has been simplified to show assembly of a genome containing 2 amplicons
723 (amplicon_01 and amplicon_n) for a single sample. B) Quast results for genomes
724 sequenced with tiled amplicons or from shotgun sequencing on R9.4 flow cells using SQK-
725 LSK109. Note PHL-10 and PHL-30 were sequenced with an earlier version of the primer
726 scheme with one primer different and are expected to have a 25bp gap C) Nucmer
727 alignment of PHL-261 genomes assembled from WGS or Lilo tiled assembly. D) Nucmer
728 alignment of PHL-237 genomes assembled from WGS or Lilo tiled assembly.Figure 4

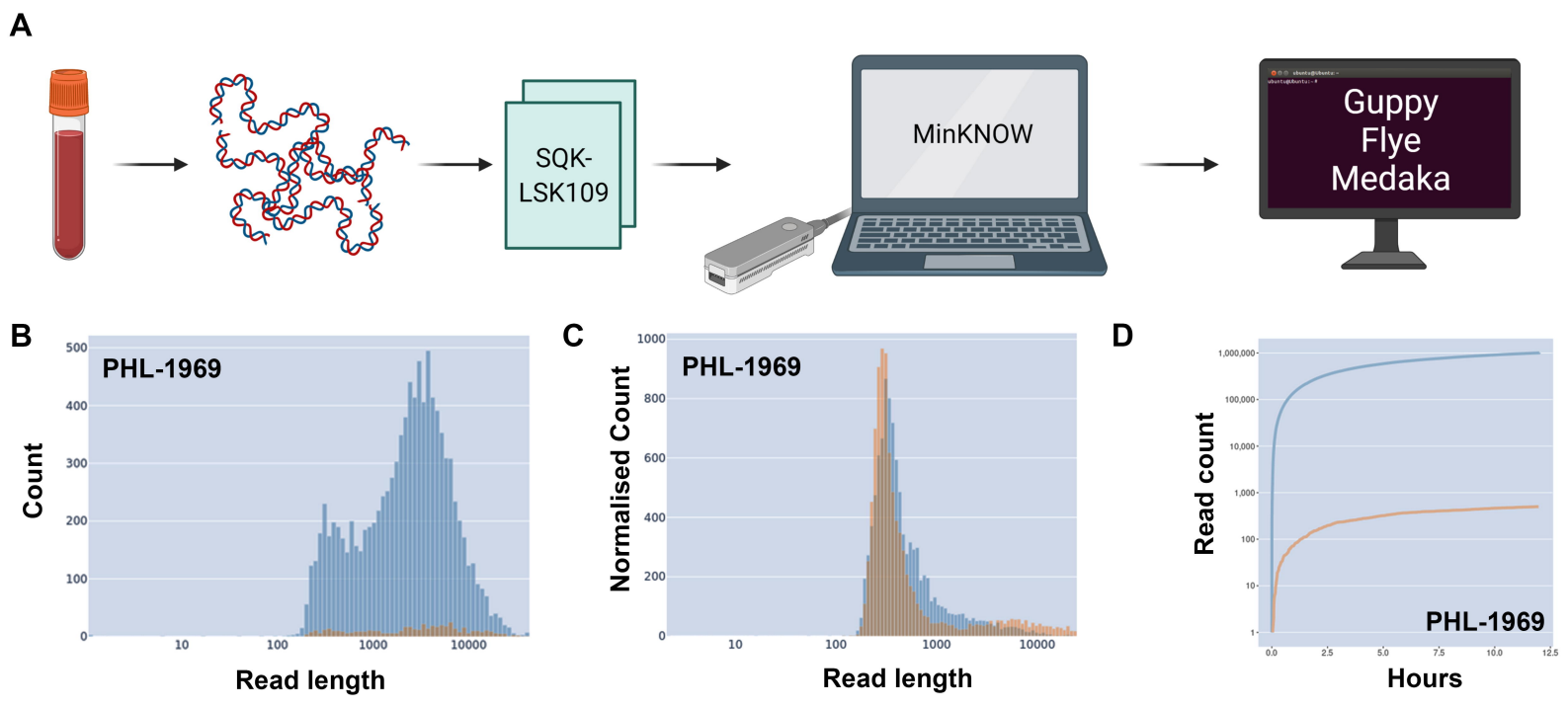
729 B) IGV image showing alignment of genomes assembled with Lilo or Artic (top) and
730 assemblies and reads for a single sample (PHL-10) aligned to a reference (MN715134.1).
731 This image shows a likely real indel present in all assemblies and supported by the reads. C)
732 as in B, but showing an indel common in Lilo assemblies and missing in Artic assemblies
733 and the reference. D) as in B, but showing a long homopolymer with poor consensus from
734 the reads and inconsistent results in assemblies. E, F & G) Examples of indels specific to the
735 Artic pipeline assemblies which do not agree with the reference.

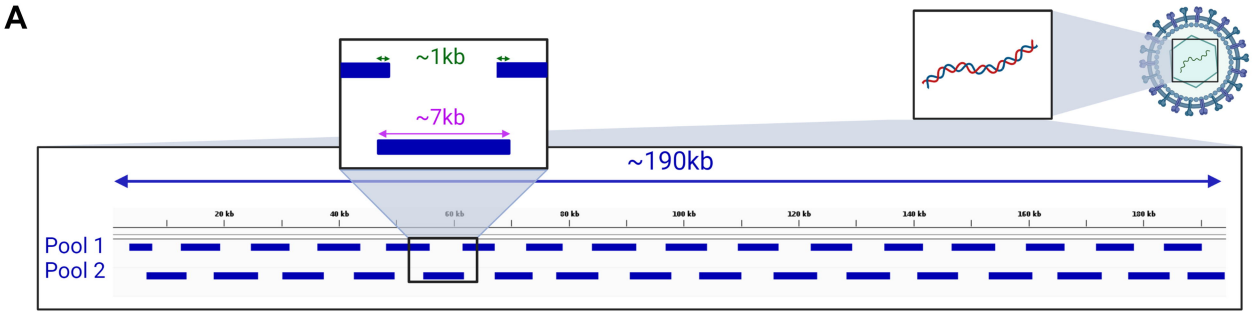
736 **Figure 5**

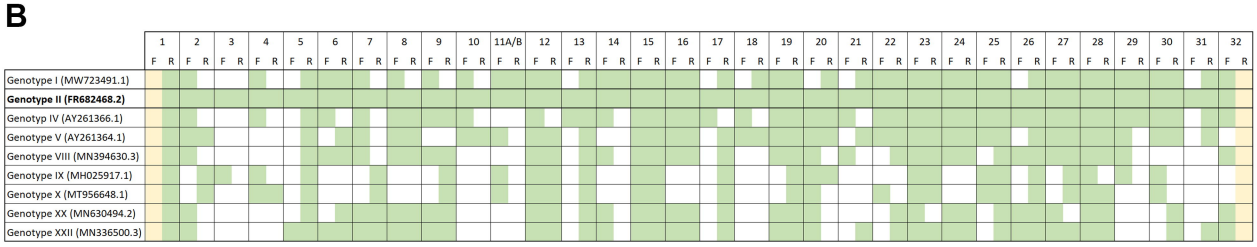
737 Maximum likelihood trees for our R9.4/SQK-LSK109 genomes and A) all available ASFV
738 genomes downloaded from NCBI (09/11/2021) or B) those specifically clustering with
739 genotype II.

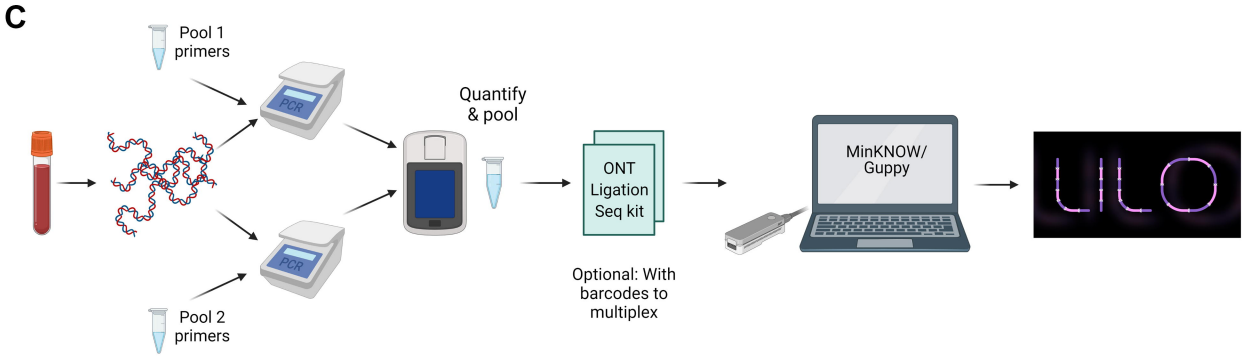
740

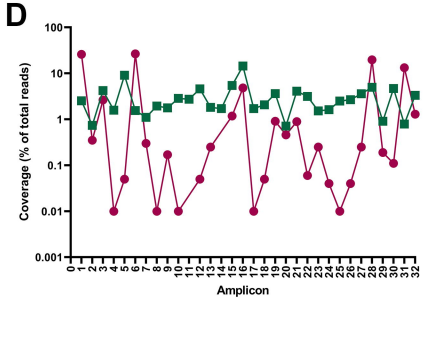
741

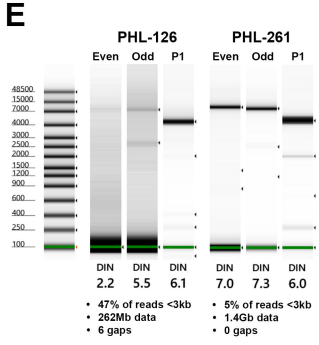

A
SQK-
LSK109
MinKNOW
Guppy
Flye
Medaka
B
PHL-1969
Count
Read length
C
PHL-1969
Normalised Count
Read length
D
Read count
PHL-1969
Hours

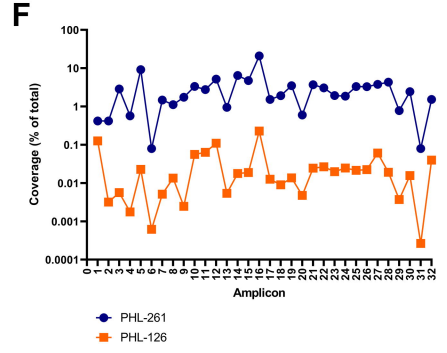

A
~1kb
~7kb
~190kb
Pool 1
Pool 2

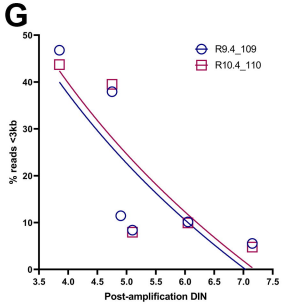

B
Genotype I (MW723491.1)
Genotype II (FR682468.2)
Genotype IV (AY261366.1)
Genotype V (AY261364.1)
Genotype VIII (MN394630.3)
Genotype IX (MH025917.1)
Genotype X (MT956648.1)
Genotype XX (MN630494.2)
Genotype XXII (MN336500.3)

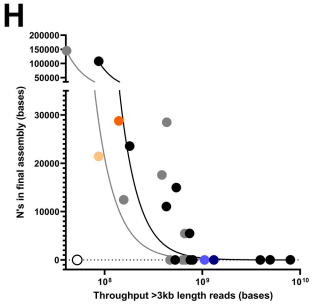

C
Pool 1 primers
Pool 2 primers
PCR
PCR
Quantify & pool
ONT Ligation Seq kit
Optional: With barcodes to multiplex
MinKNOW/ Guppy

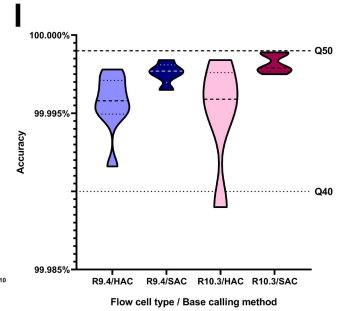

D
Coverage (% of total reads)
Amplicon

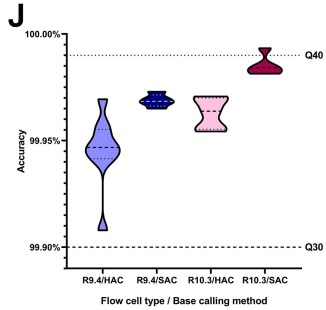

E
PHL-126
PHL-261
Even
Odd
P1
DIN 2.2
DIN 5.5
DIN 6.1
DIN 7.0
DIN 7.3
DIN 6.0
47% of reads <3kb
262Mb data
6 gaps
5% of reads <3kb
1.4Gb data
0 gaps


F
Coverage (% of total)
Amplicon
PHL-261
PHL-126

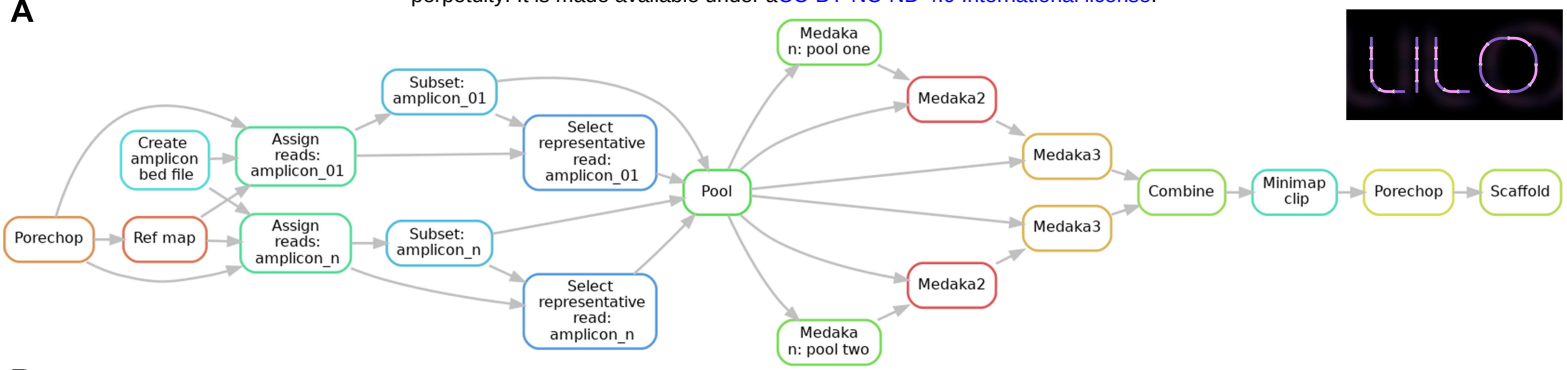

G
R9.4_109
R10.4_110
% reads <3kb
Post-amplification DIN


H
N's in final assembly (bases)
Throughput >3kb length reads (bases)


I
Accuracy
Q50
Q40
R9.4/HAC
R9.4/SAC
R10.3/HAC
R10.3/SAC
Flow cell type / Base calling method


J
Accuracy
Q40
Q30
R9.4/HAC
R9.4/SAC
R10.3/HAC
R10.3/SAC
Flow cell type / Base calling method

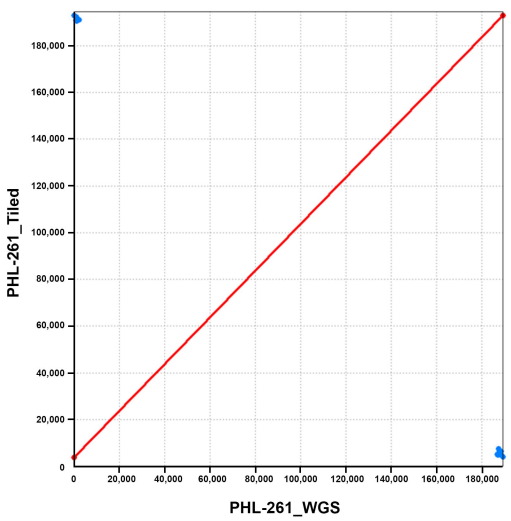
A



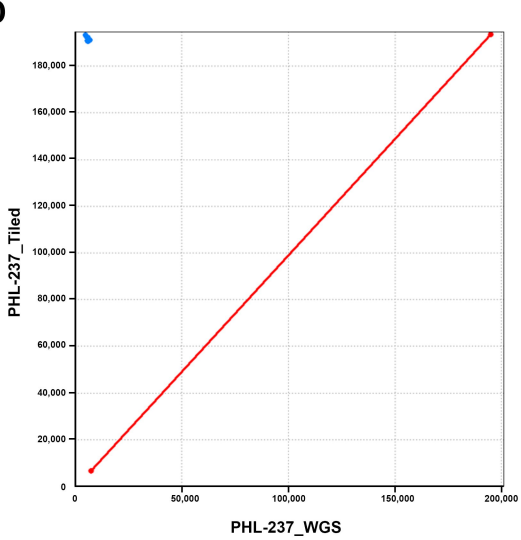
B

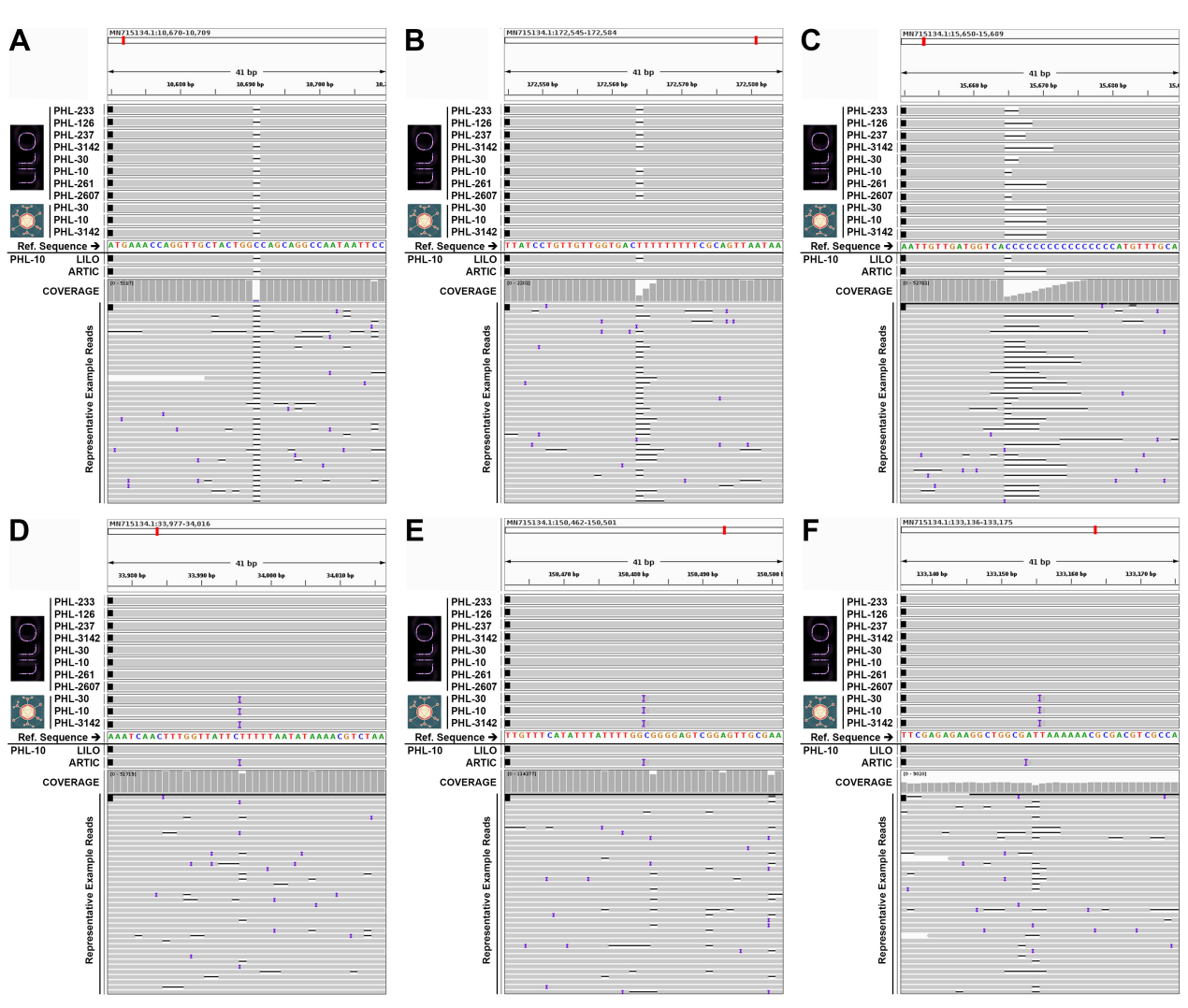
	Mismatches/ 100kb	Indels/ 100kb	% of genome N's	Post-amp DIN	% reads <3kb	Total Mismatches	Total Indels	Assembled Genome Length
Whole genome sequences								
PHL-100 109 R9.4 HAC LILO	63.54	578.23	0			120	1,092	191,202
PHL-237 109 R9.4 HAC LILO	17.90	102.12	0			34	194	201,056
PHL-261 109 R9.4 HAC LILO	10.04	199.26	0			19	377	189,291
PHL-2842 109 R9.4 HAC LILO	9.44	153.74	0			18	293	192,401
PHL-1969 109 R9.4 HAC LILO	18.89	81.33	0			36	155	194,709
Tiled - ARTIC versus LILO								
PHL-10 109 R9.4 HAC LILO	4.20	24.30	0.01		14.65	8	46	189,551
PHL-10 109 R9.4 HAC ARTIC	1.60	38.40	2.49		14.65	3	73	194,629
PHL-30 109 R9.4 HAC LILO	2.60	31.70	0.01		25.08	5	60	189,557
PHL-30 109 R9.4 HAC ARTIC	1.60	38.40	2.49		25.08	3	73	194,635
PHL-3142 109 R9.4 HAC LILO	5.30	37.40	0	6.70	16.71	10	71	189,861
PHL-3142 110 R9.4 HAC LILO	3.20	71.10	0	4.90	11.46	6	135	190,114
PHL-3142 110 R9.4 HAC ARTIC	1.10	38.40	2.40	6.70	16.71	2	73	194,631
Tiled - 109 / R9.4 vs 110 / R10.3								
PHL-126 109 R9.4 HAC LILO	3.70	42.60	0	2.40	24.37	7	81	189,902
PHL-126 109 R9.4 SAC LILO	2.60	25.20	0	2.40	23.86	5	48	189,909
PHL-126 110 R10.3 HAC LILO	4.30	24.20	17.84	3.85	46.80	7	39	161,294
PHL-126 110 R10.3 SAC LILO	1.30	9.40	11.28	3.85	43.72	2	14	190,035
PHL-233 109 R9.4 HAC LILO	4.80	49.90	12.42	1.80	39.71	8	83	189,872
PHL-233 109 R9.4 SAC LILO	2.40	31.90	6.57	1.80	39.50	4	53	189,520
PHL-233 110 R10.3 HAC LILO	11.00	25.70	56.96	4.75	37.94	9	21	190,074
PHL-233 110 R10.3 SAC LILO	2.20	6.70	76.42	4.75	39.52	1	3	190,087
PHL-237 109 R9.4 HAC LILO	8.10	47.20	2.98	5.15	18.92	15	87	184,406
PHL-237 109 R9.4 SAC LILO	2.20	22.80	2.88	5.15	18.62	4	42	189,818
PHL-237 110 R10.3 HAC LILO	1.60	28.40	0	6.05	10.16	3	54	190,090
PHL-237 110 R10.3 SAC LILO	1.00	10.50	0	6.05	9.94	2	20	189,962
PHL-2607 109 R9.4 HAC LILO	2.20	40.90	5.84	1.35	41.08	4	73	189,585
PHL-2607 109 R9.4 SAC LILO	3.50	23.80	9.28	1.35	40.69	6	41	189,503
PHL-2607 110 R10.3 HAC LILO	3.70	23.50	9.28	5.10	8.40	6	38	161,518
PHL-2607 110 R10.3 SAC LILO	2.50	12.40	15.00	5.10	7.95	4	20	189,995
PHL-261 109 R9.4 HAC LILO	4.70	43.70	0	6.15	11.37	9	83	189,878
PHL-261 109 R9.4 SAC LILO	1.60	22.60	0	6.15	10.95	3	43	189,881
PHL-261 110 R10.3 HAC LILO	3.20	21.10	0	7.15	5.55	6	40	189,886
PHL-261 110 R10.3 SAC LILO	2.10	8.40	0	7.15	4.78	4	16	189,933

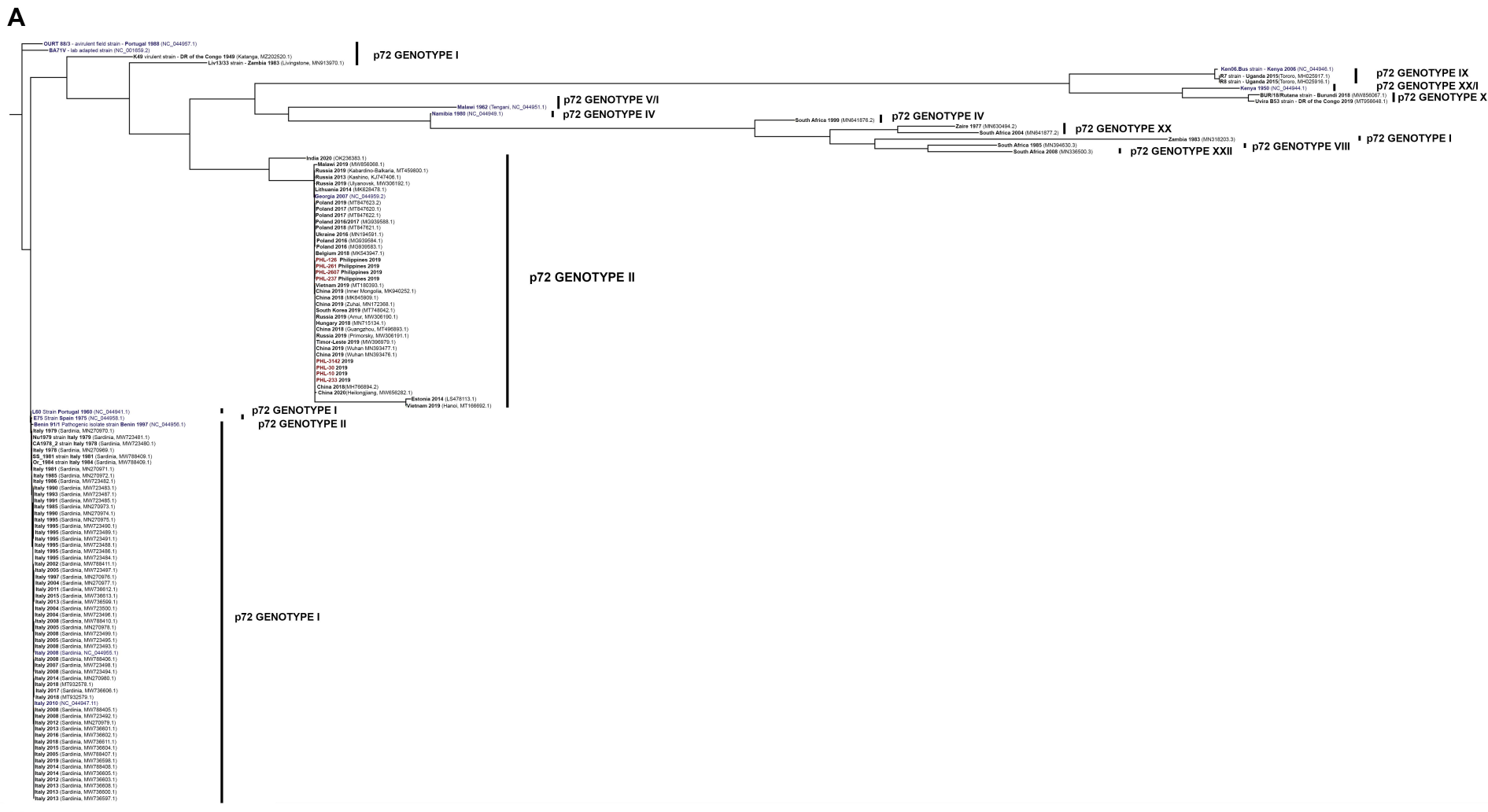
C

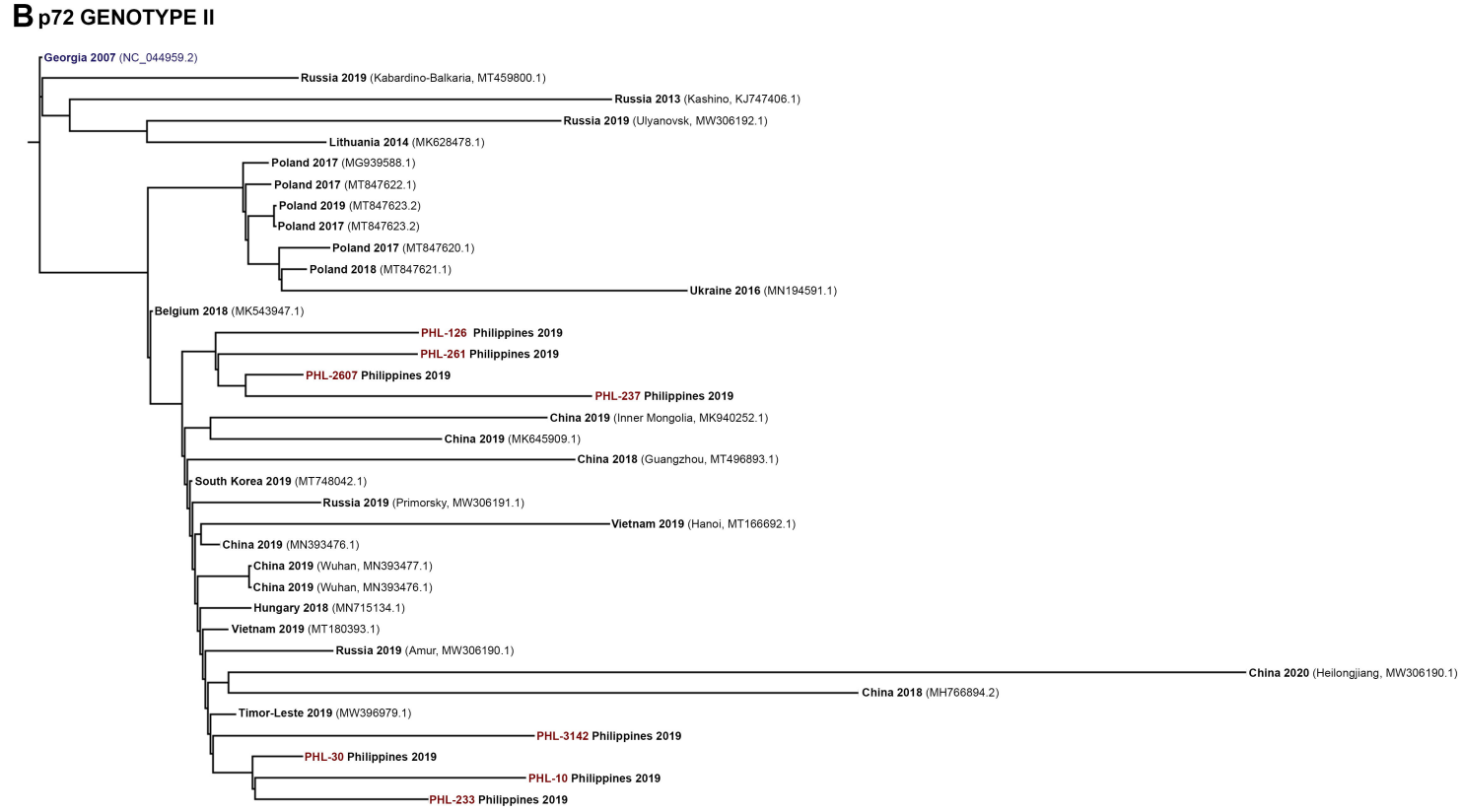


D




A
MN715134.1:10,670-10,709
41 bp
10,680 bp
10,690 bp
10,700 bp
PHL-233
PHL-126
PHL-237
PHL-3142
PHL-30
PHL-10
PHL-261
PHL-2607
PHL-30
PHL-10
PHL-3142
Ref. Sequence
ATGAAACCAGGTTGCTACTGGCCAGCAGGCCAATAATTCC
PHL-10
LILO
ARTIC
COVERAGE
Representative Example Reads
B
MN715134.1:172,545-172,584
41 bp
172,550 bp
172,560 bp
172,570 bp
172,580 bp
PHL-233
PHL-126
PHL-237
PHL-3142
PHL-30
PHL-10
PHL-261
PHL-2607
PHL-30
PHL-10
PHL-3142
Ref. Sequence
TTATCCTGTTGTTGGTGACTTTTTTTTTCGCAGTTAATAA
PHL-10
LILO
ARTIC
COVERAGE
Representative Example Reads
C
MN715134.1:15,650-15,689
41 bp
15,660 bp
15,670 bp
15,680 bp
PHL-233
PHL-126
PHL-237
PHL-3142
PHL-30
PHL-10
PHL-261
PHL-2607
PHL-30
PHL-10
PHL-3142
Ref. Sequence
AATTGTTGATGGTCACCCCCCCCCCCCCCCCATGTTTGCA
PHL-10
LILO
ARTIC
COVERAGE
Representative Example Reads
D
MN715134.1:33,977-34,016
41 bp
33,980 bp
33,990 bp
34,000 bp
34,010 bp
PHL-233
PHL-126
PHL-237
PHL-3142
PHL-30
PHL-10
PHL-261
PHL-2607
PHL-30
PHL-10
PHL-3142
Ref. Sequence
AAATCAACTTTGGTTATTCTTTTTAATATAAAACGTCTAA
PHL-10
LILO
ARTIC
COVERAGE
Representative Example Reads
E
MN715134.1:150,462-150,501
41 bp
150,470 bp
150,480 bp
150,490 bp
PHL-233
PHL-126
PHL-237
PHL-3142
PHL-30
PHL-10
PHL-261
PHL-2607
PHL-30
PHL-10
PHL-3142
Ref. Sequence
TTGTTTCATATTTATTTTGGCGGGGAGTCGGAGTTGCGAA
PHL-10
LILO
ARTIC
COVERAGE
Representative Example Reads
F
MN715134.1:133,136-133,175
41 bp
133,140 bp
133,150 bp
133,160 bp
133,170 bp
PHL-233
PHL-126
PHL-237
PHL-3142
PHL-30
PHL-10
PHL-261
PHL-2607
PHL-30
PHL-10
PHL-3142
Ref. Sequence
TTCGAGAGAAGGCTGGCGATTAAAAAACGCGACGTCGCCA
PHL-10
LILO
ARTIC
COVERAGE
Representative Example Reads


A
OURT 88/3 - avirulent field strain - Portugal 1988 (NC_044957.1)
BA71V - lab adapted strain (NC_001659.2)
K49 virulent strain - DR of the Congo 1949 (Katanga, MZ202520.1)
Liv1/33/3 strain - Zambia 1983 (Livingstone, MW913970.1)
p72 GENOTYPE I
Ken06.Bus strain - Kenya 2006 (NC_044946.1)
R7 strain - Uganda 2015(Tororo, MH025917.1)
R8 strain - Uganda 2015(Tororo, MH025916.1)
Kenya 1950 (NC_044944.1)
BUR/16/Rutana strain - Burundi 2018 (MW856067.1)
Uvira B53 strain - DR of the Congo 2019 (MT956648.1)
p72 GENOTYPE IX
p72 GENOTYPE XX/I
p72 GENOTYPE X
Malawi 1962 (Tengani, NC_044951.1)
Namibia 1980 (NC_044949.1)
p72 GENOTYPE V/I
p72 GENOTYPE IV
South Africa 1999 (MN641876.2)
Zaire 1977 (MN630494.2)
South Africa 2004 (MN641877.2)
Zambia 1983 (MN318203.3)
South Africa 1985 (MN394630.3)
South Africa 2008 (MN336500.3)
p72 GENOTYPE IV
p72 GENOTYPE XX
p72 GENOTYPE XXII
p72 GENOTYPE VIII
p72 GENOTYPE I
India 2020 (OK236383.1)
Malawi 2019 (MW856068.1)
Russia 2019 (Kabardino-Balkaria, MT459800.1)
Russia 2013 (Kashino, KJ747406.1)
Russia 2019 (Ulyanovsk, MW306192.1)
Lithuania 2014 (MK628478.1)
Georgia 2007 (NC_044959.2)
Poland 2019 (MT847623.2)
Poland 2017 (MT847620.1)
Poland 2017 (MT847622.1)
Poland 2016/2017 (MG939588.1)
Poland 2018 (MT847621.1)
Ukraine 2016 (MN194591.1)
Poland 2016 (MG939584.1)
Poland 2016 (MG939583.1)
Belgium 2018 (MK543947.1)
PHL-126 Philippines 2019
PHL-261 Philippines 2019
PHL-2607 Philippines 2019
PHL-237 Philippines 2019
Vietnam 2019 (MT180393.1)
China 2019 (Inner Mongolia, MK940252.1)
China 2018 (MK645909.1)
China 2019 (Zuhai, MN172368.1)
South Korea 2019 (MT748042.1)
Russia 2019 (Amur, MW306190.1)
Hungary 2018 (MN715134.1)
China 2018 (Guangzhou, MT496893.1)
Russia 2019 (Primorsky, MW306191.1)
Timor-Leste 2019 (MW396979.1)
China 2019 (Wuhan MN393477.1)
China 2019 (Wuhan MN393476.1)
PHL-3142 2019
PHL-30 2019
PHL-10 2019
PHL-233 2019
China 2018(MH766894.2)
China 2020(Heilongjiang, MW656282.1)
Estonia 2014 (LS478113.1)
Vietnam 2019 (Hanoi, MT166692.1)
p72 GENOTYPE II
L60 Strain Portugal 1960 (NC_044941.1)
E75 Strain Spain 1975 (NC_044958.1)
Benin 97/1 Pathogenic isolate strain Benin 1997 (NC_044956.1)
p72 GENOTYPE I
p72 GENOTYPE II
Italy 1979 (Sardinia, MN270970.1)
Nur1979 strain Italy 1979 (Sardinia, MW723481.1)
CA1978_2 strain Italy 1978 (Sardinia, MW723480.1)
Italy 1978 (Sardinia, MN270969.1)
SS_1981 strain Italy 1981 (Sardinia, MW788409.1)
Or_1984 strain Italy 1984 (Sardinia, MW788409.1)
Italy 1981 (Sardinia, MN270971.1)
Italy 1985 (Sardinia, MN270972.1)
Italy 1986 (Sardinia, MW723482.1)
Italy 1990 (Sardinia, MW723483.1)
Italy 1993 (Sardinia, MW723487.1)
Italy 1991 (Sardinia, MW723485.1)
Italy 1985 (Sardinia, MN270973.1)
Italy 1990 (Sardinia, MN270974.1)
Italy 1995 (Sardinia, MN270975.1)
Italy 1995 (Sardinia, MW723490.1)
Italy 1995 (Sardinia, MW723489.1)
Italy 1995 (Sardinia, MW723491.1)
Italy 1995 (Sardinia, MW723488.1)
Italy 1995 (Sardinia, MW723486.1)
Italy 1995 (Sardinia, MW723484.1)
Italy 2002 (Sardinia, MW788411.1)
Italy 2005 (Sardinia, MW723497.1)
Italy 1997 (Sardinia, MN270976.1)
Italy 2004 (Sardinia, MN270977.1)
Italy 2011 (Sardinia, MW736612.1)
Italy 2015 (Sardinia, MW736613.1)
Italy 2013 (Sardinia, MW736599.1)
Italy 2004 (Sardinia, MW723500.1)
Italy 2004 (Sardinia, MW723496.1)
Italy 2008 (Sardinia, MW788410.1)
Italy 2005 (Sardinia, MN270978.1)
Italy 2008 (Sardinia, MW723499.1)
Italy 2005 (Sardinia, MW723495.1)
Italy 2008 (Sardinia, MW723493.1)
Italy 2008 (Sardinia, NC_044955.1)
Italy 2008 (Sardinia, MW788406.1)
Italy 2007 (Sardinia, MW723498.1)
Italy 2008 (Sardinia, MW723494.1)
Italy 2014 (Sardinia, MN270980.1)
Italy 2018 (MT932578.1)
Italy 2017 (Sardinia, MW736606.1)
Italy 2018 (MT932579.1)
Italy 2010 (NC_044947.1)
Italy 2008 (Sardinia, MW788405.1)
Italy 2008 (Sardinia, MW723492.1)
Italy 2012 (Sardinia, MN270979.1)
Italy 2013 (Sardinia, MW736601.1)
Italy 2016 (Sardinia, MW736602.1)
Italy 2018 (Sardinia, MW736611.1)
Italy 2015 (Sardinia, MW736604.1)
Italy 2005 (Sardinia, MW788407.1)
Italy 2019 (Sardinia, MW736598.1)
Italy 2014 (Sardinia, MW788408.1)
Italy 2014 (Sardinia, MW736605.1)
Italy 2012 (Sardinia, MW736603.1)
Italy 2013 (Sardinia, MW736608.1)
Italy 2013 (Sardinia, MW736600.1)
Italy 2013 (Sardinia, MW736597.1)


B p72 GENOTYPE II
Georgia 2007 (NC_044959.2)
Russia 2019 (Kabardino-Balkaria, MT459800.1)
Russia 2013 (Kashino, KJ747406.1)
Russia 2019 (Ulyanovsk, MW306192.1)
Lithuania 2014 (MK628478.1)
Poland 2017 (MG939588.1)
Poland 2017 (MT847622.1)
Poland 2019 (MT847623.2)
Poland 2017 (MT847623.2)
Poland 2017 (MT847620.1)
Poland 2018 (MT847621.1)
Ukraine 2016 (MN194591.1)
Belgium 2018 (MK543947.1)
PHL-126 Philippines 2019
PHL-261 Philippines 2019
PHL-2607 Philippines 2019
PHL-237 Philippines 2019
China 2019 (Inner Mongolia, MK940252.1)
China 2019 (MK645909.1)
China 2018 (Guangzhou, MT496893.1)
South Korea 2019 (MT748042.1)
Russia 2019 (Primorsky, MW306191.1)
Vietnam 2019 (Hanoi, MT166692.1)
China 2019 (MN393476.1)
China 2019 (Wuhan, MN393477.1)
China 2019 (Wuhan, MN393476.1)
Hungary 2018 (MN715134.1)
Vietnam 2019 (MT180393.1)
Russia 2019 (Amur, MW306190.1)
China 2020 (Heilongjiang, MW306190.1)
China 2018 (MH766894.2)
Timor-Leste 2019 (MW396979.1)
PHL-3142 Philippines 2019
PHL-30 Philippines 2019
PHL-10 Philippines 2019
PHL-233 Philippines 2019