APPLYING THIRD-GENERATION SEQUENCING TO PATHOGEN SURVEILLANCE AND MIXED INFECTION DETECTION

By

Jeremy B. Buttler, B.S.

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in

Biological Sciences

University of Alaska Fairbanks

August 2022

APPROVED:

Devin M. Drown, Committee Chair
Eric Bortz, Committee Co-chair
Naoki Takebayashi, Committee Member
Molly Murphy, Committee Member
Diane Wagner, Chair
Department of Biology and Wildlife
Karsten Hueffer, Interim Dean
College of Natural Science and Mathematics
Richard Collins, Director of the Graduate School

Abstract

One Health is the concept of interconnected health between plants, animals, humans, microorganisms and the environments they live in. One Health issues surround many important viral pathogens, including influenza, SARS-CoV-2, and Ebola, that have likely come from zoonotic spillovers. Genomic epidemiology combines pathogen genomes with metadata to forecast, track, and prepare for future pathogens and pathogen variants that may cause epidemics. Genomic epidemiology has been used to detect and track viral variants that have the potential to escape vaccines for viruses like porcine circovirus type 2 (PCV2). PCV2 causes porcine circovirus associated diseases (PCVAD), which results in weight loss and death in pigs around the world. The correlation between PCVAD and mixed infections shows that disease severity is linked to the microbial community in a host. Metagenomics allows researchers to sequence samples and sort out the individual community member genomes by bioinformatic analyses, allowing the study of the host microbiome. In this thesis, I tested if long read nanopore sequencing can uncover PCV2 diversity and reliably detect co-infections. I also assessed the accuracy and efficiency of long read metagenomic assemblers as a potential method for detecting mixed infections. In my first chapter, I found that nanopore sequencing can be used to understand PCV2 diversity and detect co-infections. This evidence shows that nanopore sequencing is a viable alternative to Sanger sequencing for PCV2 surveillance. In my second chapter, I found Flye built the most complete and accurate genomes for bacterial community members and their plasmids. Throughout my thesis I have shown that nanopore sequencing is a viable solution for modern surveillance. The lower cost of nanopore sequencing may allow more specific pathogen and metagenomic surveillance in regions with high risk of zoonotic spillovers, which may allow early detection of epidemic causing pathogens.

Table of Contents

List of Figures

Page

List of Tables

Page

List of Acronyms

Acronym	Definition
ASFV	African swine fever virus
BB	Boot strap
mb	Mega base pairs
bp	basepair
Cap	Capsid
HWM	High molecular weight
kb	Kilo base pairs
moDC	Monocyte derived dendritic cell
No.	Number
ORF	Open reading frame (encodes a gene)
ONT	Oxford Nanopore Technologies
PCV2	Porcine circovirus type 2
PCV3	Porcine circovirus type 3
PCVAD	Porcine circovirus associated disease
Q-score	Quality score
RCR	Rolling circle replication

Acknowledgements

I acknowledge the generous support that made this work possible from the Institute of Arctic Biology, Alaska IDeA Network of Biomedical Research Excellence (INBRE), and the University of Alaska Fairbanks (UAF) Biomedical Learning and Student Training (BLaST) program. Research reported here was supported by BLaST through the National Institute of General Medical Sciences of the National Institutes of Health under awards UL1GM118991, TL4GM118992, and RL5GM118990. Research reported here was also supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant 2P20GM103395. I thank the staff at Metabiota, Black & Veatch, and Battelle in Ukraine for administrative support. Research reported herein was in part supported by the US Defense Threat Reduction Agency (DTRA) Biological Threat Reduction Program in Ukraine (BTRP), through a subaward to the University of Alaska. I also thank the University of Alaska Fairbanks Biology and Wildlife Department and University of Alaska Anchorage Department of Biological sciences for providing TAships that supported my education and provided me valuable experience in teaching.

I also acknowledge my mentors; Devin Drown and Eric Bortz, who provided feedback and help throughout my research. My committee members, Molly Murphy and Naoki Takebayashi, who were always willing to help. Scientists in Ukraine including Nataliia Rudova, Ganna Kovalenko, Mykola Sushko, Vitaliy Bolotin, Larysa Muzykina, Oleksandr Zinenko, Borys Stegniy, Yurii Dunaiev, Mykola Sytiuk, Anton Gerilovych, and Oleksii Solodiankin, who contributed to the study design, sequencing, and sample collection needed for chapter one. I am thankful for Tracie Haan who provided feedback on my thesis and did sequencing for chapter 2 and Taylor Seitz, who also did the sequencing for chapter two. I am grateful for the support and

feedback from Bevyn Cover. Finally, I am grateful Jack Chen who supported my entry into UAF, for my family, the members of the Drown lab, and members of the Bortz lab, who gave feedback, support, and helpful ideas that improved my research.

General Introduction:

Many human pathogens of concern, such as influenza, SARS-CoV-2, and Ebola, have likely come from zoonotic spillovers and from animal reservoirs (Cunningham et al., 2017; Giovanetti et al., 2021). These spillovers and animal reservoirs show that our health is affected by the animals we come into contact with and the surrounding environment (Cunningham et al., 2017). One Health is the concept of interconnected health between plants, animals, humans, microorganisms, and the environments they live in (Cunningham et al., 2017). A modern problem in animal disease that requires a One Health perspective to solve is the transmission of African swine fever virus (ASFV) throughout the pig industry (Gaudreault et al., 2020).

ASFV is a DNA virus that cause hemorrhagic fevers in domestic pigs in Europe and Asia (Cisek et al., 2016; Gaudreault et al., 2020). The mortality rate of AFSV infected pigs is between 30% to 100% (Cisek et al., 2016). Transmission of ASFV to domestic pigs can occur through contaminated pork products, such as feeding food waste to pigs, soft ticks from the *Ornithodoros* genus, or through wild boar contact (Gaudreault et al., 2020). Through trade and wild boar migrations, the highly pathogenic p72 genotype II of ASFV has spread from Africa to Europe, China, and even across the ocean to the Dominican Republic (Gaudreault et al., 2020; USDA 2021).

Halting trade in live pigs and pork products with countries with ASFV is an example of a biosecurity intervention that may reduce pathogen transmission (Kedkovid et al., 2020). However, this intervention will not prevent pathogen transmission through other means, such as wild boar migrations between countries. Implementation of effective biosecurity interventions requires knowledge of how a pathogen is transmitted.

When pathogens replicate, the progeny receive copies of the parental genetic material (DNA or RNA), which may include mutations. These mutations in the progeny can distinguish a single progeny's lineage from other lineages that came from the parental genome (Gardy 2018). These mutations often fix within populations at regular rates and thus, act as a molecular clock that allows researchers to estimate when the progeny lineage diverged from the parent lineage (Gardy 2018). Mutations and molecular clocks allow for tracking of viral lineages and detecting changes in the viral lineages through time (Gardy 2018). Genomic epidemiology is the concept of using a pathogen's genome and metadata, such as time, location, and host type, to track the spread of pathogen lineages through a population (Gardy 2018; Hill et al., 2021). Another use for genomic epidemiology is identifying locations and events that have increased risk of pathogen transmission and alternative hosts that may transmit pathogens (Gardy 2018). Once the method and mode of transmission of a pathogen is known, effective biosecurity interventions can be applied to reduce further pathogen transmission.

Genomic epidemiology has been used to forecast future pathogen variants that may cause epidemics or pandemics, like in the SARS-CoV-2 pandemic (Hill et al., 2021). In the SARS-CoV-2 pandemic, genomic epidemiology has been used to identify SARS-CoV-2 variants that had increased pathogenicity before they became dominant variants in other countries (Giovanetti et al., 2021). This information may have given other countries advanced warning of potential SARS-CoV-2 waves. A country may use this information to prepare for a SARS-CoV-2 wave by stockpiling hospital supplies, limiting gatherings, and closing borders to countries with the variant of concern.

Genomic epidemiology has also been used to prepare for or prevent future epidemics of influenza (Hay and McCauley 2018). For influenza, genomic epidemiology is used to identify

variants that are common and likely to cause the next seasonal epidemic. These variants are then used to aid in vaccine design, several months before the typical Fall flu season in the Northern Hemisphere (Hay and McCauley 2018). This can allow for vaccines to be distributed early to at-risk populations, such as the elderly, before flu season, which reduces hospital cases and mortality in the elderly population (Hay and McCauley 2018). Both SARS-CoV-2 and influenza demonstrate how genomic epidemiology can help predict, prepare, and reduce the impact of future pathogens of concern at the global level.

Genomic epidemiology also aids in the detection of pathogens in otherwise isolated countries. An example being the spread of porcine circovirus type 2 (PCV2) into the island of Sardinia, Italy, which has no pig exports and little pig imports (Franzo et al., 2020). PCV2 causes a group of diseases known as porcine circovirus associated diseases (PCVAD) that can cause reduced weight gain, respiratory distress, abortions, and death in pigs (Karuppannan and Opriessnig 2017; Opriessnig and Langohr 2013). The globally distributed PCV2 can have a large economic impact when uncontrolled (Opriessnig et al., 2020). An example is PCV2 infections in the English pig industry, which was estimated to cost between 52 to 88 million euros yearly before heavy vaccine use in 2008 (Alarcon et al., 2013). Using genomic epidemiology, Franzo et al. (2020) tracked new lineages (genotypes) of PCV2 throughout the wild boar and domestic pig populations of Sardinia, and found that PCV2 introduction into Sardinia was likely by the small number of pig imports to backyard farms, which then infected wild boars (Franzo et al., 2020).

Genomic epidemiology requires genomic information, such as whole genome sequences, to analyze. However, whole genome sequencing may require the pathogen to be isolated, which is time consuming, or the use of metagenomics, which is expensive (Morris et al., 2019; Papaiakovou et al., 2022). An alternative to genome sequencing is genotyping, which uses one or

more partial or complete genes to determine a pathogens lineage (Janezic and Rupnik 2019). Using only a few genes allows more samples to be sequenced together, which reduces the sequencing cost (Papaiakovou et al., 2022). However, for viruses, genotyping often uses primers that are specific to a single pathogen or single family of pathogens, which prevents genotyping from identifying novel or unexpected pathogens (Janezic and Rupnik 2019).

After an epidemic or pandemic, genotyping has been used to detect viral variants that have the potential to escape known vaccines, one example being PCV2 (Franzo and Segalés 2020). Though PCVADs have been heavily reduced by vaccines, PCV2 transmission and PCVADs have not been completely prevented in vaccinated animals (Franzo and Segalés 2020; Segalés and Sibila 2022). As a result, there is strong selection pressure on PCV2 variants to escape the vaccine in vaccinated animals (Kekarainen et al., 2014). This selective pressure has resulted in surveillance of PCV2 being maintained in many countries (Franzo and Segalés 2020; Xiao et al., 2016; Song et al., 2020; Franzo et al., 2020).

PCV2 has a circular, single-stranded DNA genome that is 1766 bases (nucelotides) long (Breitbart et al., 2017). The genome of PCV2 contains two major open reading frames (ORFs) and at least four other functional ORFs (Breitbart et al., 2017; He et al., 2013; Liu, Chen, and Kwang 2005; Lv et al., 2015; Li et al., 2018). ORF2 encodes for the capsid protein, which is targeted by the immune system, as evidenced by a higher substitution rate than the rest of the genome (Franzo and Segalés 2020; Kekarainen et al., 2014). The high substitution rate in ORF2 also makes this gene an appropriate target for measuring diversity. To date, eight PCV2 genotypes (a-h) have been described (Franzo and Segalés 2018).

Though PCV2 is needed for a PCVAD to develop, PCV2 infections are often subclinical and do not develop into a PCVAD (Ouyang et al., 2019). To develop an overt PCVAD, it has been hypothesized that an animal needs to be infected with PCV2 and a bacterial pathogen (mixed infection) or other viruses (co-infection) (Ouyang et al., 2019). Sometimes co-infections can include other subclinical viruses, like porcine parvovirus, torque teno virus and other variants of PCV2 (Ouyang et al., 2019).

The correlation between PCVADs and mixed infections or co-infections shows that disease severity is linked to the bacterial and viral community in a host. Bacterial community members in a host can be found by sequencing the bacterial 16S rRNA gene or culturing (Bharti and Grimm 2021; Garmendía et al., 2012). Viral communities can be found by sequencing conserved genes or using assays that target multiple viral genes (Chiu and Miller 2019; Bharti and Grimm 2021). However, culturing takes multiple days to complete, and many microbes are difficult to culture, thus some pathogenic microbes might be missed (Garmendía et al., 2012). Amplification of 16S rRNA or other conserved genes may be limited to broad levels of classification (e.g., genus) and may not identify novel bacteria or viruses, respectively (Bharti and Grimm 2021). For viruses, sequence discovery may be limited to the virus families the assay is designed for and have limited ability to detect novel or unexpected viruses (Chiu and Miller 2019).

One solution is metagenomics, in which all genetic material in a sample is sequenced without targeting specific pathogens, and the individual community members' genomes are assembled through bioinformatic analyses (Bai et al., 2022). Metagenomics has been used to detect new viruses and show possible correlations between disease and viruses, such as PCVAD and co-infections involving PCV2 and torque teno virus (Bai et al., 2022; Qin et al., 2018).

Metagenomics has also been used to sequence and detect novel viruses that may be of concern, such as PCV3, which commonly co-infects with PCV2 and may contribute to PCVAD development (Chen et al., 2021; Palinski et al., 2017).In this thesis, I examined the reliability of long read sequencing for PCV2 genomic surveillance. I also extended my analysis to test how metagenomics could be used to sequence complex bacterial and viral communities. In Chapter 1, I tested if nanopore sequencing is a reliable option to detect PCV2 co-infections and measure diversity. In Chapter 2, I assessed the accuracy and efficiency of long read metagenomic assemblers as a potential method for detecting mixed infections.

References:

Alarcon, P., J. Rushton, and B. Wieland. 2013. "Cost of Post-Weaning Multi-Systemic Wasting Syndrome and Porcine Circovirus Type-2 Subclinical Infection in England - an Economic Disease Model." Preventive Veterinary Medicine 110 (2): 88–102. https://doi.org/10.1016/j.prevetmed.2013.02.010.

Bai, G., S. Lin, Y. Hsu, and S. Chen. 2022. "The Human Virome: Viral Metagenomics, Relations with Human Diseases, and Therapeutic Applications." Viruses 14 (2): 278.

Bharti, R., and D. G. Grimm. 2021. "Current Challenges and Best-Practice Protocols for Microbiome Analysis." Briefings in Bioinformatics 22 (1): 178–93. https://doi.org/10.1093/bib/bbz155.

Breitbart, M., E. Delwart, K. Rosario, J. Segalés, and A. Varsani. 2017. "ICTV Virus Taxonomy Profile: Circoviridae." The Journal of General Virology 98 (8): 1997–98.

Chen, S., L. Zhang, X. Li, G. Niu, and L. Ren. 2021. "Recent Progress on Epidemiology and Pathobiology of Porcine Circovirus 3." Viruses 13 (10): 1944. https://doi.org/10.3390/v13101944.

Chiu, C. Y., and S. A. Miller. 2019. "Clinical Metagenomics." Nature Reviews. Genetics 20 (6): 341–55. https://doi.org/10.1038/s41576-019-0113-7.

Cisek, A. A., I. Dąbrowska, K. P. Gregorczyk, and Z. Wyżewski. 2016. "African Swine Fever Virus: A New Old Enemy of Europe." Annals of Parasitology 62 (3): 161–67. https://doi.org/10.17420/ap6203.49.

Cunningham, A. A., P. Daszak, and J. L. N. Wood. 2017. "One Health, Emerging Infectious Diseases and Wildlife: Two Decades of Progress?" Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 372 (1725): 20160167. https://doi.org/10.1098/rstb.2016.0167.

Franzo, G., and J. Segalés. 2020. "Porcine Circovirus 2 Genotypes, Immunity and Vaccines: Multiple Genotypes but One Single Serotype." Pathogens (Basel, Switzerland) 9 (12): 1049. https://doi.org/10.3390/pathogens9121049.

Franzo, G., and J. Segalés. 2018. "Porcine Circovirus 2 (PCV-2) Genotype Update and Proposal of a New Genotyping Methodology." PloS One 13 (12): e0208585.

Franzo, G., S. Tinello, L. Grassi, C. M. Tucciarone, M. Legnardi, M. Cecchinato, G. Dotto, et al., 2020. "Free to Circulate: An Update on the Epidemiological Dynamics of Porcine Circovirus 2 (PCV-2) in Italy Reveals the Role of Local Spreading, Wild Populations, and Foreign Countries." Pathogens 9 (3): 221. https://doi.org/10.3390/pathogens9030221.

Gardy, J. L., and N. J. Loman. 2018. "Towards a Genomics-Informed, Real-Time, Global Pathogen Surveillance System." Nature Reviews. Genetics 19 (1): 9–20. https://doi.org/10.1038/nrg.2017.88.

Garmendía, L., A. Hernández, M. B. Sánchez, and J. L. Martínez. 2012. "Metagenomics and Antibiotics." Clinical Microbiology and Infection: European Society of Clinical Microbiology and Infectious Diseases 18 Suppl 4: 27–31. https://doi.org/10.1111/j.1469-0691.2012.03868.x.

Gaudreault, N. N., D. W. Madden, W. C. Wilson, J. D. Trujillo, and J. A. Richt. 2020. "African Swine Fever Virus: An Emerging DNA Arbovirus." Frontiers in Veterinary Science 7 (May): 215. https://doi.org/10.3389/fvets.2020.00215.

Giovanetti, M., F. Benedetti, G. Campisi, A. Ciccozzi, S. Fabris, G. Ceccarelli, V. Tambone, et al., 2021. "Evolution Patterns of SARS-CoV-2: Snapshot on Its Genome Variants." Biochemical and Biophysical Research Communications 538 (January): 88–91. https://doi.org/10.1016/j.bbrc.2020.10.102.

Hay, A. J., and J. W. McCauley. 2018. "The WHO Global Influenza Surveillance and Response System (GISRS)-a Future Perspective." Influenza and Other Respiratory Viruses 12 (5): 551–57. https://doi.org/10.1111/irv.12565.

He, J., J. Cao, N. Zhou, Y. Jin, J. Wu, and J. Zhou. 2013. "Identification and Functional Analysis of the Novel Orf4 Protein Encoded by Porcine Circovirus Type 2." J Virol. 87 (3): 1420 - 1429. http://dx.doi.org/10.1128/JVI.01443-12.

Hill, V., C. Ruis, S. Bajaj, O. G. Pybus, and M. U. G. Kraemer. 2021. "Progress and Challenges in Virus Genomic Epidemiology." Trends in Parasitology 37 (12): 1038–49. https://doi.org/10.1016/j.pt.2021.08.007.

Janezic, S., and M. Rupnik. 2019. "Development and Implementation of Whole Genome Sequencing-Based Typing Schemes for Clostridioides Difficile." Frontiers in Public Health 7: 309. https://doi.org/10.3389/fpubh.2019.00309.

Karuppannan, A. K, and T. Opriessnig. 2017. "Porcine Circovirus Type 2 (Pcv2) Vaccines in the Context of Current Molecular Epidemiology." Viruses 9 (5): 99. https://doi.org/10.3390/v9050099.

Kedkovid, R., C. Sirisereewan, and R. Thanawongnuwech. 2020. "Major Swine Viral Diseases: An Asian Perspective After the African Swine Fever." Porcine Health Management 6 (1): 20. https://doi.org/10.1186/s40813-020-00159-x.

Kekarainen, T., A. Gonzalez, A. Llorens, and J. Segalés. 2014. "Genetic Variability of Porcine Circovirus 2 in Vaccinating and Non-Vaccinating Commercial Farms." The Journal of General Virology 95 (PT 8): 1734–42. http://dx.doi.org/10.1099/vir.0.065318-0.

Li, D., J. Wang, S. Xu, S. Cai, C. Ao, L. Fang, S. Xiao, H. Chen, and Y. Jiang. 2018. "Identification and Functional Analysis of the Novel Orf6 Protein of Porcine Circovirus Type 2 in Vitro." Veterinary Research Communications 42 (1): 1–10. https://doi.org/10.1007/s11259-017-9702-0.

Liu, J., I. Chen, and J. Kwang. 2005. "Characterization of a Previously Unidentified Viral Protein in Porcine Circovirus Type 2-Infected Cells and Its Role in Virus-Induced Apoptosis." J Virol. 79 (13): 8262 - 8274. http://dx.doi.org/10.1128/JVI.79.13.8262-8274.2005.

Lv, Q., K. Guo, H. Xu, T. Wang, and Y. Zhang. 2015. "Identification of Putative Orf5 Protein of Porcine Circovirus Type 2 and Functional Analysis of GFP-Fused Orf5 Protein." PloS One 10 (6): e0127859. https://doi.org/10.1371/journal.pone.0127859.

Morris, A., G. Robinson, M. T. Swain, and R. M. Chalmers. 2019. "Direct Sequencing of Cryptosporidium in Stool Samples for Public Health." Frontiers in Public Health 7 (December): 360. https://doi.org/10.3389/fpubh.2019.00360.

Opriessnig, T., A. K. Karuppannan, A. M. M. G. Castro, and C. Xiao. 2020. "Porcine Circoviruses: Current Status, Knowledge Gaps and Challenges." Virus Research 286 (September): 198044. https://doi.org/10.1016/j.virusres.2020.198044.

Opriessnig, T., and I. Langohr. 2013. "Current State of Knowledge on Porcine Circovirus Type 2-Associated Lesions." Veterinary Pathology 50 (1): 23–38. https://doi.org/10.1177/0300985812450726.

Ouyang, T., X. Zhang, X. Liu, and L. Ren. 2019. "Co-Infection of Swine with Porcine Circovirus Type 2 and Other Swine Viruses." Viruses 11 (2): 185. https://doi.org/10.3390/v11020185.

Palinski, R., P. Piñeyro, P. Shang, F. Yuan, R. Guo, Y. Fang, E. Byers, and B. M. Hause. 2017. "A Novel Porcine Circovirus Distantly Related to Known Circoviruses Is Associated with Porcine Dermatitis and Nephropathy Syndrome and Reproductive Failure." Journal of Virology 91 (1): e01879-16. https://doi.org/10.1128/JVI.01879-16.

Papaiakovou, M., D. T. J. Littlewood, S. R. Doyle, R. B. Gasser, and C. Cantacessi. 2022. "Worms and Bugs of the Gut: The Search for Diagnostic Signatures Using Barcoding, and Metagenomics-Metabolomics." Parasites & Vectors 15 (1): 118. https://doi.org/10.1186/s13071-022-05225-7.

Qin, S., W. Ruan, H. Yue, C. Tang, K. Zhou, and B. Zhang. 2018. "Viral Communities Associated with Porcine Respiratory Disease Complex in Intensive Commercial Farms in Sichuan Province, China." Scientific Reports 8 (1): 13341. https://doi.org/10.1038/s41598-018-31554-8.

Segalés, J., and M. Sibila. 2022. "Revisiting Porcine Circovirus Disease Diagnostic Criteria in the Current Porcine Circovirus 2 Epidemiological Context." Veterinary Sciences 9 (3): 110. https://doi.org/10.3390/vetsci9030110.

Song, S., G. Park, S. Choe, R. M. Cha, S. Kim, B. Hyun, B. Park, and D. An. 2020. "Genetic Diversity of Porcine Circovirus Isolated from Korean Wild Boars." Pathogens 9 (6): 457. https://doi.org/10.3390/pathogens9060457.

USDA. 2021. "USDA Statement on Confirmation of African Swine Fever in the Dominican Republic," July. https://www.aphis.usda.gov/aphis/newsroom/news/sa_by_date/sa-2021/asf-confirm.

Xiao, C., K. M. Harmon, P. G. Halbur, and T. Opriessnig. 2016. "PCV2d-2 Is the Predominant Type of Pcv2 DNA in Pig Samples Collected in the u.s. During 2014-2016." Veterinary Microbiology 197 (December): 72–77. https://doi.org/10.1016/j.vetmic.2016.11.009.

Chapter 1: Genetic Diversity of Porcine Circovirus 2 in Wild Boars and Domestic Pigs in Ukraine.[1]

Abstract:

Between 2017 and 2021, the world produced 96 million to 112 million tons of pork per year. The pork industry uses vaccines to avoid the increased cost from animal death and weight loss caused by pathogens, such as porcine circovirus type 2 (PCV2). However, the PCV2 vaccine only reduces PCV2 infections, allowing for the possibility of escape variants to come from vaccinated pigs. New variants of PCV2 can spread within a country or to other countries through wild boar migration and the domestic pig trade. The last PCV2 surveillance study from Ukraine found genotypes a, b, d, and g in 2015.

To determine if other genotypes of PCV2 have circulated in Ukraine, we sequenced open reading frame 2, which encodes the capsid gene from 11 wild boar samples collected in 2012 and 6 domestic pig samples collected in 2019 on a third-generation sequencer, a MinION (Oxford Nanopore Technologies). The most common PCV2 genotype within our samples was genotype b, with the next most common being genotype a in the 2013 wild boar samples and genotype d in the 2019 domestic pig samples.

---

[1] Article published as: Rudova, N.*, J. Buttler*, G. Kovalenko, M. Sushko, V. Bolotin, L. Muzykina, O. Zinenko, et al., 2022. "Genetic Diversity of Porcine Circovirus 2 in Wild Boar and Domestic Pigs in Ukraine." Viruses 14 (5). https://doi.org/10.3390/v14050924.

* These authors contributed equally to this work

Introduction:

Between 2017 and 2021, the world produced 96 million to 112 million tons of pork per year (USDA 2021). The pork industry uses vaccines to avoid the increased cost from animal death and weight loss caused by pathogens, such as porcine circovirus type 2 (PCV2) (Gebhardt et al., 2020). PCV2 is often an asymptomatic infection that causes reduced weight gain. However, when PCV2 is combined with other pathogens or even different PCV2 variants, PCV2 can cause a group of diseases known as porcine circovirus associated diseases (PCAD) (Ouyang et al., 2019; Karuppannan and Opriessnig 2017). PCAD symptoms are often most severe in younger pigs and can include wasting, abortion, and death (Karuppannan and Opriessnig 2017). Vaccines have heavily reduced PCAD and subclinical PCV2 symptoms and have even reduced PCV2 infections (Franzo and Segalés 2020). However, vaccines have not prevented PCV2 infections, which has led to an increased selection pressure for escape variants in vaccinated pigs (Franzo and Segalés 2020).

PCV2 is a non-enveloped, icosahedral (T=1) capsid (Cap) virus that contains a single-stranded circular DNA genome that is 1766 bases (nucleotides) long (Breitbart et al., 2017). The genome of PCV2 contains two major open reading frames (ORFs), and at least four other functional ORFs (Breitbart et al., 2017; He et al., 2013; Liu et al., 2005; Lv et al., 2015; D. Li et al., 2018). ORF2 encodes the Cap gene and is used to determine which of the eight genotypes (a-h) a PCV2 isolate belongs to (Franzo and Segalés 2018).

New PCV2 variants and genotypes can spread to other countries through wild boar migrations and by mixing of pigs from different farms (Franzo et al., 2020; Correa-Fiz et al., 2020). The spread of PCV2 by the transfer of subclinically infected domestic pigs between farms

and wild boar migrations means that no country is isolated, instead all countries are a potential source, or at least at risk, of harboring new PCV2 variants. In order to detect new PCV2 variants that may replicate better in vaccinated pigs early, we need good surveillance across more than just a few countries. Despite modern surveillance of PCV2, there are still countries, such as Ukraine, that have not had surveillance on samples collected past 2015.

In 2015, Ukraine's major PCV2 genotype was b, with genotypes a, d, and g also in circulation (Kleymann et al., 2020). However, genotypes e, f, h, and i were not detected in Ukraine and genotype d was found to be rare in Ukraine (Dudar et al., 2018). Genotype d is currently a common, global genotype, which may replicate better in vaccinated pigs, and was rare before 2014 (Karuppannan and Opriessnig 2017; Xiao et al., 2016). Therefore, it is possible that genotype d would be rare in the samples on GenBank and the wild boar samples from Dudar et al., (2018), which were collected before 2014 and more common in recent samples, which are non-existent.

Our goal for this study was to determine if nanopore sequencing is a reliable option to genotype PCV2, detect co-infections, and contribute to understanding of viral diversity. We also wanted to discover if genotypes e, f, h, and i were in Ukraine. To determine if genotypes e, f, h, and i were in Ukraine, we estimated a phylogenetic tree using ORF2 sequences from GenBank, ORF2 sequences from archived Ukrainian wild boar samples, and ORF2 sequences from domestic pig samples collected in 2019 from Ukraine.

Methods:

Study sites and Sequencing:

We sequenced PCV2 from viremic blood samples from archived wild boar samples collected in 2012 from Ukraine and viremic domestic pig liver samples collected from a local market in Karkhiv in 2019, using an Oxford Nanopore Technologies (ONT) MinION. The wild boar samples we sequenced came from the Chernihiv, Chernivtsi, Luhansk, Poltava, Volyn, and Zaporizhia oblasts (administrative districts) of Ukraine (Figure 1-1). All wild boar and domestic pig samples were collected opportunistically and thus, the health of the domestic pig or wild boar was not known during collection. For further details about sample collection or sample processing, see Rudova et al., (2019). We enriched for ORF2 and whole PCV2 genomes by PCR, using primers from Rudova et al., (2019) and Yang et al., (2018). All samples were sequenced using the LSK-109 library kit (Oxford Nanopore Technologies, Oxford, UK). We basecalled and demultiplexed the reads with Guppy version 3.5 (ONT) using the r941_min_hight_g351 model.

Database:

Our database used for co-infection detection and tree building contained 5862 ORF2 sequences downloaded from GenBank. ORF2 sequences in our database were aligned with Mafft V7.407 (Katoh and Standley 2013) and manually inspected to remove sequences with early stop codons or incomplete reading frames with Geneious v2020.2.1 (Kearse et al., 2012). We removed natural and artificial recombinant sequences from the inspected ORF2 sequences with RDP4 v4.101 (Martin et al., 2015) using settings similar to (Franzo and Segalés 2018). We then removed all gaps and stop codons in our remaining, aligned ORF2 sequences using Geneious

v2020.2.1. Our final database contained 429 ORF2 sequences (Table 0-2: Sequences after recombinant removal).

Figure 1-1: Ukraine sample collection map. Oblasts that contributed PCV2 ORF2 sequences to our study are labeled. A red * indicates an oblast that we collected wild boar sequences from. An orange + indicates an oblast we collected domestic pig samples from. The outline for our map was bought from https://freevectormaps.com/ukraine/UA-EPS-01-1002?ref=atr. The map of Europe was downloaded from Google Maps in 2021.

Table 1-1: References used to identify genotypes. "-" indicates that the reference was detected as a recombinant genome and removed by RDP4. References are from Franzo and Segalés (2018).

Reference	Genotype	In final database
HM038034	a	yes
HQ202949	a	-
KX828215	a	yes
EU450638	b	-
KP768478	b	-
KY806003	b	yes
KJ094599	c	-
EU148503	c	-
EU148504	c	-
MF314285	d	-
KX960929	d	-
KC515014	d	yes
KT867799	e	-
KT870147	e	-
KT795280	e	-
LC004750	f	yes
HQ202949	f	-
LC008135	f	yes
KP420197	g	yes
FJ998185	g	-
JX099786	g	-

We labeled the seven sequences listed in Table 1-1 that were in our final database with their genotype. Sequences in our database from Ukraine were also labeled with their country of origin and oblast they were sampled from.

Co-infection Detection:

We detected co-infections between different PCV2 variants by mapping reads to the ORF2 sequences in our database (See Figure 1-2 for a flowchart). To reduce the mis-binning of reads from a single variant to a group of very similar ORF2 sequences, we took a sub-sample of ORF2 sequences, with a maximum similarity of 98% between any two sequences from our database with cd-hits using parameter -c 0.98 (W. Li and Godzik 2006; Fu et al., 2012). The sub-sampled database had 66 ORF2 sequences.

We mapped our reads to the ORF2 sequences in our sub-sampled database with minimap2 v2.22 (H. Li 2018) using parameters -ax map-ont. Low-quality reads were removed with samtools v1.14 (H. Li et al., 2009; Danecek et al., 2021), using parameters view -q 30 (mapping quality), –min-qlen 2000 (min length), -e "avg(qual)>=30" (average Q-score), and -F 0x04 (remove unmapped reads). The bam file from samtools was then split into separate bins by reference with bamtools v2.5.1 using parameter -reference (https://github.com/pezmaster31/bamtools). Each bin contained all the reads that were mapped to a single ORF2 sequence. We removed bins with fewer than 50 reads or fewer than 0.8% of reads that mapped to any PCV2 ORF2 sequence. Sequencing statistics for the remaining bins and un-binned reads were found using NanoStat --fastq (Coster et al., 2018).

We built consensus genomes from the remaining bins by polishing the best read from each bin with the reads in its bin (See Figure 1-2 for a flowchart). The 300 reads with the highest final score from the table output by filtlong v0.2 using parameter --verbose (https://github.com/rrwick/Filtlong) were extracted from their bin with grep using parameters --no-group-separator -A 4. We then polished the read with the highest score using the remaining 299 reads with one round of Racon v1.4.21 using parameters -m 8 -x 6 -g 8 -w 500 (https://github.com/isovic/racon) and one round of Medaka v1.4.3 with the r941_min_high model (https://github.com/nanoporetech/medaka).

We detected and removed consensuses built from miss-binned reads by removing consensuses that were very similar. The number of mismatches and aligned length were found using a blastn (NCBI Resource Coordinators 2016) query of a consensus genome against all consensus genomes in a sample. When two or more consensus genomes had fewer than 1.5% mismatches (100 * number mismatches / aligned length), we kept the consensus genome with the

most reads. We automated our co-infection detection pipeline using bash scripts (https://github.com/jeremyButtler/find--Co-infections).

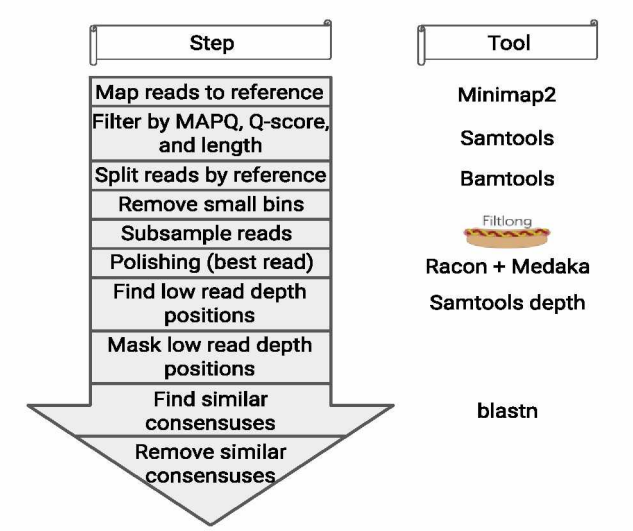

Figure 1-2: Co-infection pipeline steps.

Manual Curation:

We found references for manual curation by blasting each consensus against the PCV2 database on GenBank (taxid: 85708). The top hit for each consensus was then used to detect and remove indels in the consensuses. We also checked and corrected reading frames using transeq from the emboss package v6.6 (Rice et al., 2000).

Maximum Likelihood Tree:

We made a maximum likelihood tree using IQtree (Hoang et al., 2018; Minh et al., 2020; Kalyaanamoorthy et al., 2017; Chernomor et al., 2016), our consensuses, and a sub-sample of our ORF2 sequence database. Sub-sampling was done with cd-hit using parameter -c 0.985 to reduce our database to a manageable size. After sub-sampling, the maximum difference for any two ORF2 sequences was 98.5%. We added ORF2 sequences from Ukraine that were removed

by sub-sampling back into our sub-sampled database. We estimated a consensus tree with IQtree v1.6.12 (Hoang et al., 2018; Minh et al., 2020; Kalyaanamoorthy et al., 2017; Chernomor et al., 2016) using parameters -st CODON -m MFP+MERGE -bb 1000, the sub-sampled ORF2 database, and the ORF2 genes from our consensuses. The consensus tree was edited in R v4.1.1 using the treeio (L. Wang et al., 2020), ape (Paradis and Schliep 2019), ggplot2 (Wickham 2016), and ggtree (Yu 2020; Yu et al., 2018, 2017) packages.

Results:

Sequencing:

We enriched and sequenced the PCV2 ORF2 gene from six domestic pig samples and 12 infected wild boar blood samples from the Chernihiv, Chernivtsi, Luhansk, Poltava, Volyn, and Zaporizhia oblasts of Ukraine on a Minion sequencer. During quality control, we removed the Luhansk 1 sample because it had no reads that mapped to PCV2 genome. The remaining samples had read depths between 14941 to 635035 and mean Q-scores between 12.9 to 14.6 for the major strain after binning (Table 1-2).

We also enriched and sequenced whole genomes from 10 of our wild boar samples. Six samples (Chernihiv 1 to 3; Chernivtsi 1, and 2; and Poltava 1) had a read depth between 97 to 3194 reads and a mean Q-score between 12.3 to 12.9 after binning (Table 1-3). Five of our six samples (Chernihiv 1 to 3 and Chernivtsi 1 and 2) had at least one read that was between 1745 to 1787 bases long, which almost covers the entire PCV2 genome (Table 1-3). However, only the Chernihiv 1 sample had enough near full genome length reads to build a consensus genome (Table 1-3). The only disagreements in ORF2 gene from our whole genome samples and their ORF2 enriched replicate samples was an indel and an anonymous base (Table 1-4).

Table 1-2: ORF2 PCR sequencing results. QC indicates after the quality control step. No.reads is the number of reads

Sample	QC	No.reads	Yield (bases)	Mean length (bases)	Mean quality (Q-score)
Chernihiv 1	No	184466	145456780	788.5	13
Chernihiv 1	Yes	170304	134933726	792.3	13
Chernihiv 2	No	123054	97389870	791.4	12.9
Chernihiv 2	Yes	70235	55802830	794.5	13.2
Chernihiv 3	No	131087	121272848	925.4	12.9
Chernihiv 3	Yes	71070	56809157	799.3	13.2
Chernihiv 4	No	155582	172464376	1108.5	12.6
Chernihiv 4	Yes	47930	38305973	799.2	13.1
Chernivtsi 1	No	240443	181401635	754	13
Chernivtsi 1	Yes	144338	114925229	796.2	13.1
Chernivtsi 2	No	226523	179205173	791.1	13
Chernivtsi 2	Yes	146489	116358358	794.3	13.2
Luhansk 1	No	38948	88934212	2283.4	13.3
Luhansk 1	Yes	NA	NA	NA	NA
Luhansk 2	No	113234	141050360	1245.7	13
Luhansk 2	Yes	44527	36505679	819.9	13.2
Poltava 1	No	170553	136186715	798.5	13
Poltava 1	Yes	87848	69690980	793.3	13
Poltava 2	No	101124	108341638	1071.4	13
Poltava 2	Yes	76231	60518027	793.9	12.9
Volyn	No	89270	113902664	1275.9	12.9
Volyn	Yes	39156	31997967	817.2	13.1
Zaporizhzhia	No	282747	253514624	896.6	13
Zaporizhzhia	Yes	14941	12107298	810.3	13.1
Kharkiv 1	No	564644	441090027	781.2	14.2
Kharkiv 1	Yes	453425	359957451	793.9	14.5
Kharkiv 2	No	307150	239663591	780.3	14.3
Kharkiv 2	Yes	156283	124091978	794	14.6
Kharkiv 3	No	280582	221393187	789	14.3
Kharkiv 3	Yes	229888	182845561	795.4	14.6
Kharkiv 4	No	284647	219919196	772.6	14.3
Kharkiv 4	Yes	70490	55998055	794.4	14.6
Kharkiv 5	No	890841	700464609	786.3	14.2
Kharkiv 5	Yes	635035	503643792	793.1	14.5
Kharkiv 6	No	135297	87801240	649	14.2
Kharkiv 6	Yes	26903	21389410	795.1	14.6

Table 1-3: Whole genome and ORF2 enriched replicate differences. No.reads is number of reads.

Sample	QC	No.reads	Yield (bases)	Mean length (bases)	Mean quality (Q-score)	Longest read	No.Reads > 1700 bases
Chernihiv 1	No	263934	159310070	603.6	12.7	NA	NA
Chernihiv 1	Yes	3194	4023081	1259.6	12.5	1787	32
Chernihiv 2	No	295011	147685127	500.6	12.9	NA	NA
Chernihiv 2	Yes	401	505994	1261.8	12.5	1756	1
Chernihiv 3	No	179477	104622440	582.9	12.7	NA	NA
Chernihiv 3	Yes	546	686376	1257.1	12.6	1745	1
Chernivtsi 1	No	353680	149116019	421.6	12.3	NA	NA
Chernivtsi 1	Yes	390	493934	1266.5	12.5	1746	1
Chernivtsi 2	No	296593	140612668	474.1	12.4	NA	NA
Chernivtsi 2	Yes	97	123328	1271.4	12.8	1770	4
Luhansk 1	No	NA	NA	NA	NA	NA	NA
Luhansk 2	No	NA	NA	NA	NA	NA	NA
Poltava 1	No	322507	168769281	523.3	12.8	NA	NA
Poltava 1	Yes	108	134893	1249	12.5	1250	0
Volyn	No	NA	NA	NA	NA	NA	NA
Zaporizhzhia	No	NA	NA	NA	NA	NA	NA

Table 1-4: Whole genome and ORF2 enriched replicate differences. No.indels is number of indels. No.mismatches is number of mismatches. No.anonymous bases is number of anonymous bases.

Sample	No.indels	No.mismatches	No.anonymous bases
Chernihiv 1	0	0	0
Chernihiv 2	0	0	0
Chernihiv 3	0	0	0
Chernivtsi 1	1	0	0
Chernivtsi 2	0	0	1
Poltava 1	0	0	1

Co-infections:

We tested the accuracy of our co-infection pipeline mentioned in methods using simulated reads (Appendix: Chapter 1: Supplementary methods). We found that our co-infection pipeline could accurately detect co-infections with less conservative settings than we used, when each consensus was polished with at least 100 reads (Figure A-2 and Figure A-3). Most

consensuses had at least one indel, but very few consensuses had mismatches. Most of the mismatches were in consensuses with low read depths or consensus that missed a co-infection, which resulted in a hybrid consensus (Figure A-4 and Figure A-5).

We tested co-infections in our samples with our co-infection pipeline. Both of our wild boar samples from Chernivtsi had at 3% of their reads from a co-infection (Table 1-5). While half of our domestic pig samples had 5% to 28% of their reads from a co-infection (Table 1-5). All detected co-infections had at least a 1000 reads (Table 1-5) and were also detected by medaka_variant (Table A-1).

Table 1-5: Percent of minor variant reads. No.reads is the number of reads.

Sample	% minor variant	No.reads	Source
Chernivtsi 1	3.47	5191	Boar
Chernivtsi 2	3.93	6000	Boar
Karhkiv 4	28.41	28356	Pig
Karhkiv 5	5.81	39560	Pig
Karhkiv 6	25.82	9398	Pig

Maximum Likelihood Tree:

We used a maximum likelihood tree to determine the diversity of PCV2 genotypes in our samples and to detect any genotypes of PCV2 that are new to Ukraine (Figure 1-3). We found that our samples and Ukraine ORF2 sequences from GenBank grouped into the clades containing genotypes a, b, d, f, and g; which are all the genotypes in our tree (Figure 1-3). Our samples were grouped into the clades containing references from genotypes a (HM038034, Bootstrap (BB) = 100), b (KY806003, BB == 90), d (KC515014, BB = 91), and f (LC004750, BB = 100, Figure 1-3). With, the clade containing genotype b (KP420197) having most of our wild boar and domestic pig samples (BB = 90, Figure 1-3).

The clade with the genotype d reference (KC515014), had one non-co-infected domestic pig sample (Kharkiv 2), a Ukraine ORF2 sequence from GenBank (KP420187), and the major or minor variant from all three co-infected domestic pig samples in our tree (Kharkiv 4, 5, and 6; BB = 91, Figure 1-3). All the co-infected samples also had their major (2) or minor (1) variant group into the clade with the genotype b reference (KP420197, BB = 90, Figure 1-3).

Only wild boar samples were grouped into clades holding the genotypes a, f, and g references (Figure 1-3). For genotype a (HM038034) we had two minor variants from a co-infection (Chernivitsi 2 and 4) and the Ukraine ORF2 sequences from GenBank (KP20202, KP20203, KP420186, KP420194, and KP420199, Figure 1-3). The clade with the genotype f (LC004750) reference had only our Poltava (2) sample (BB = 93, Figure 1-3).

In clade containing the genotype b reference (KY806003), our second sample from Luhansk and three of our samples from Chernihiv (2-4) had a branch length of zero (were identical). Other ORF2 sequences that were identical included Luhansk (KP420189) and Cherkasy (KP420191) (BB = 98) in the clade with the genotype b reference (KY806003) and Lviv (KP420203) and Dnipropetrovsk (KP420199) (BB = 100) in the clade with the genotype a reference (HM038034, Figure 1-3).

Discussion

Using genomics, we sampled the diversity of PCV2 circulating in Ukrainian domestic pigs and wild boars. Previous studies identified the most common genotype in Ukraine as genotype b, with genotypes a, d, and g also being present in Ukraine (Dudar et al., 2018; Kleymann et al., 2020). Among our wild boar samples, we found that genotype b was the most common genotype detected. Other genotypes detected in our wild boar samples were genotypes a

and f. In our domestic pig samples, we found only genotype b and d which were at near equal frequencies.

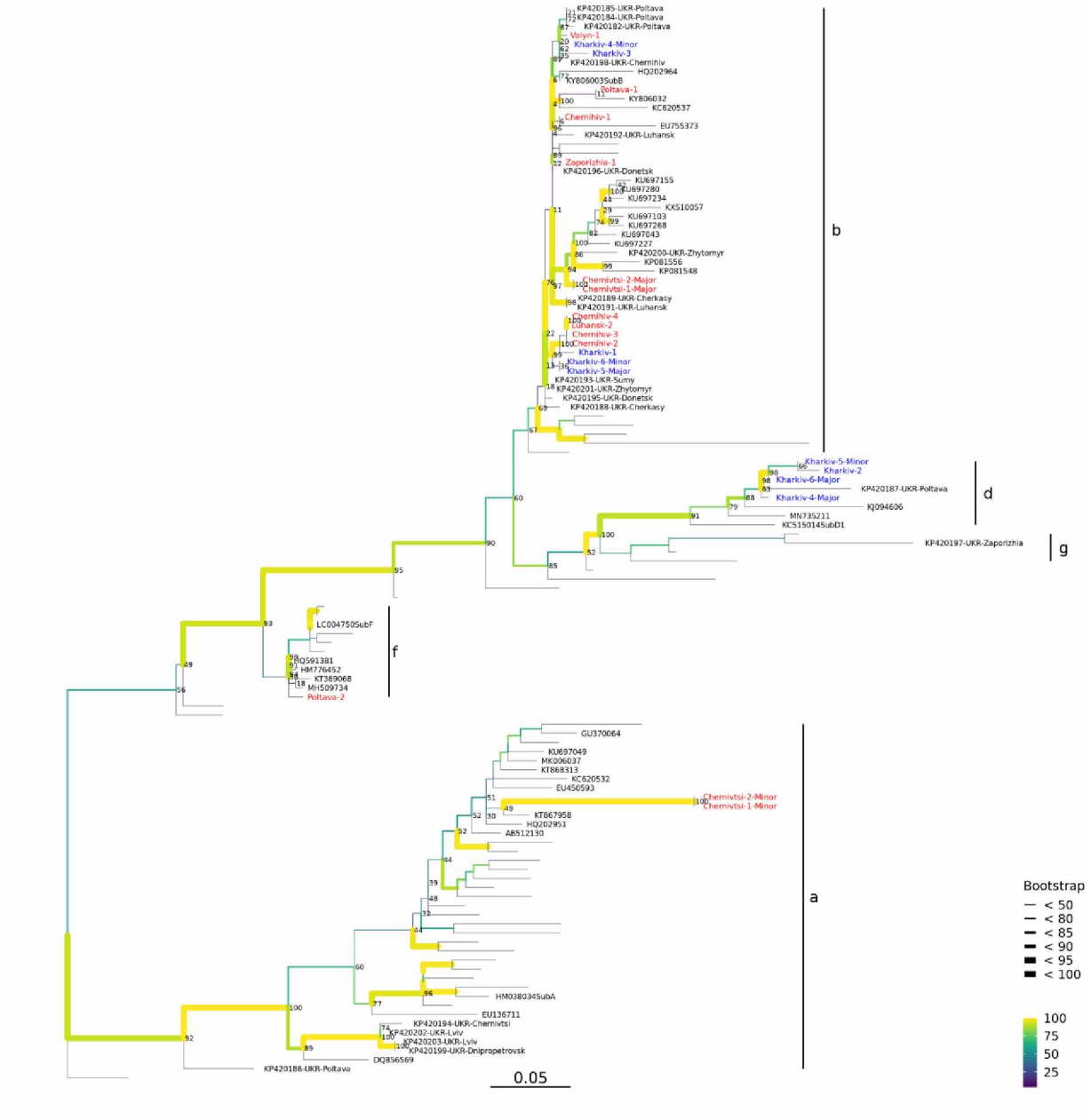

Figure 1-3: PCV2 ORF 2 maximum likelihood tree. Wild boar sequences from our study are highlighted in red. Domestic pig sequences from our study are highlighted in blue.

Genotypes:

We detected PCV2 genotype d in half of our domestic pig samples from 2019 and one pre-2013 Ukrainian wild boar sequence downloaded from GenBank. Before 2013, genotype d was rare and genotype b was common in domestic pigs and wild boars from many countries (Franzo and Segalés 2018; Xiao et al., 2016; Song et al., 2020; Franzo et al., 2020). However, after 2015, genotype d became a common global genotype, while genotype b decreased in prevalence (Franzo and Segalés 2018). So, it is likely that we observed a similar shift from genotype b to genotype d in our samples. However, our 2013 samples are from wild boars and our 2019 samples are from six domestic pigs. So, we would need more domestic pig samples, archived domestic pig samples from 2012, and wild boar samples from 2019 to confirm there was a shift from genotype b to d in Ukraine.

Currently, all Ukrainian wild boar sequences on GenBank and in our study are from samples collected before 2013. So, we cannot determine the current genotypes of PCV2 circulating in the Ukrainian wild boar population. However, between 2010 and 2013, genotype d became a more common genotype globally (Franzo and Segalés 2018). There was also an increase in genotype d in Italian and Korean wild boar populations after 2013 (Song et al., 2020; Franzo et al., 2020). This suggests that genotype d may be a common genotype in modern day Ukrainian wild boars. However, surveillance is needed to confirm if genotype d is a common genotype in the current day Ukraine wild boar population.

We found that genotype f, a genotype not previously detected in Ukraine, was in Ukraine during 2012. Genotype f was first detected in China in 2008, but archived samples show that genotype f could be found in China in 1996 (Zhao et al., 2010; Bao et al., 2018). Genotype f has

also been detected in India, Indonesia, Croatia, and the USA (Bhattacharjee et al., 2021; Nugroho et al., 2016; Y. Wang et al., 2020). However, none of these countries are neighbors to Ukraine. This suggests that either genotype f is undetected in many countries or is spreading through human activity. More surveillance is needed in countries around Ukraine to detect if genotype f may have been transmitted from neighboring countries or through trade from distant countries.

Before our study, we knew genotype d was present in wild boars and domestic pigs in Ukraine, while genotype g was present in wild boars (Dudar et al., 2018; Kleymann et al., 2020). However, we only detected genotypes d and g through the sequences we download from GenBank. This shows that genotypes d and g were in the wild boar populations in Ukraine before 2013, but that our sample size was too small to detect rare genotypes reliably. The lack of other studies to detect genotype f in Ukraine and the inability of our study to detect genotypes d and g in wild boar samples suggests that PCV2 diversity in Ukraine is under-sampled.

Porcine Circovirus Type 2 Transmission:

We detected identical PCV2 sequences in wild boars in Luhansk and Chernihiv, Lviv and Dnipropetrovsk, and Cherkasy and Luhansk. A previous study found identical PCV2 sequences in Zaporizhzhia and Chernigiv; and Cherasky and Kharkiv (Dudar et al., 2018). None of these oblasts are neighboring oblasts and Lviv and Dnipropetrovsk are on opposite sides of Ukraine, suggesting that PCV2 may be transmitted between oblasts in Ukraine. Alternatively, PCV2 transmission to different oblasts may be from similar sources, like neighboring countries, outside of Ukraine. More genomic surveillance in Ukraine and in neighboring countries is needed to understand PCV2 circulation in Ukraine.

PCV2 can be transmitted by mixing of farm populations, trade, and by wild boar movements (Franzo et al., 2020; Correa-Fiz et al., 2018). Our sample size prevents us from determining if the spread of PCV2 in Ukraine is through wild boar movements or trade. However, highly similar PCV2 sequences shared between domestic pigs and wild boars have been found in multiple countries, including Ukraine (Franzo et al., 2020; Cságola et al., 2006; Dudar et al., 2018). Showing that transmission between wild boars and domestic pigs is possible and that some PCV2 transmission between wild boars and domestic pigs may have happened in Ukraine. However, it is possible, but less likely, that both wild boars and domestic pigs may be infected with PCV2 from domestic pigs in neighboring countries.

Co-infections:

We detected co-infections between subtypes b and d in almost half of our domestic pig samples. Co-infections between genotypes b and d were a common co-infection seen in the US during 2015 (Xiao et al., 2016). However, other studies have also detected co-infections between genotypes a and d, a and b, and e and d (Xiao et al., 2016; Park and Chae 2021; Correa-Fiz et al., 2018). Showing that our detected domestic pig co-infections are not unexpected and that we likely missed co-infections involving rare genotypes.

There are multiple PCV2 studies that did not use methods that could detect co-infections between PCV2 variants (Zheng et al., 2020; Raev, Yuzhakov, and Aliper 2021; Song et al., 2020; Franzo et al., 2020). However, our only sequences with genotype a were from wild boar co-infections or were downloaded from GenBank. The detection of genotype a in our samples suggests that detecting co-infections increases the chances of detecting less common genotypes.

This also shows the utility of using nanopore sequencing, since our methods used to detect co-infections required few additional resources, mainly bioinformatic analyses.

The lower read accuracy of nanopore sequencers, relative to short-read and Sanger sequencing technologies, may seem to limit our ability to detect co-infections. However, we detected the same co-infections when we used Medaka_variant, which is a diploid variant caller (Appendix Table 1). Also, we found our pipeline could detect co-infections and ignore noise when we tested it with simulated co-infections (Appendix methods and Figures A-4 to A-5). Finally, Leigh et al., (2020) showed blastn could assign raw reads from nanopore sequencers to the correct genotype with fewer than 1% of reads being miss-assigned. All our minor variants in our co-infections had over 2% of reads, which is above the expected 1% of miss-assigned reads found by Leigh et al., (2020). Between Leigh et al., (2020), Medaka_variant, and testing with simulated co-infections, we have strong support that the co-infections we detected are not due to the error rate in nanopore sequencing. However, this does not eliminate our detected co-infections being from sample contamination or crosstalk between barcodes. Though, crosstalk is not likely in wild boar co-infections, due to only detecting genotype a in co-infections.

Full Genomes:

Using nanopore sequencing allowed us to sequence almost the entire PCV2 genome for five of our samples. However, low amplification prevented us from building consensus genomes for all but one sample (Cherihiv 1). Our PCV2 whole genome was from genotype b, had 3 single nucleotide variations, and no amino acid changes to its closest match on GenBank (Accession JX406426.1). Once the amplification step is improved, nanopore sequencing could be a reliable method to sequence whole PCV2 genomes.

One additional future improvement could be the use of rolling circle replication (RCR) as an enrichment method for PCV2. RCR has been used to enrich PCV2 samples in the past and would allow for reads with multiple copies of a single PCV2 genome (Navidad et al., 2008; Dezen et al., 2010). These reads can be split up and converted to a single, more accurate consensus (Gallardo et al., 2021). This approach may effectively reduce the high read error rate in nanopore sequencing.

Conclusions:

We have found that PCV2 genotype b was common in all our sequences from Ukraine and that genotype d was also common in our domestic pig samples. Also, that genotypes a, d, g, and f were circulating in the pre-2013 wild boar population. However, our low sample size likely limited our ability to detect rare genotypes in the domestic pig population. So, there may be more genotypes than b and d in the Ukrainian domestic pig population.

For circulation, we found that identical PCV2 variants have been transmitted multiple times to different oblasts. However, low sample sizes within oblasts prevented us from detecting transmission chains. So, we do not know if the transmission is between oblasts or from countries outside of Ukraine. Showing that more PCV2 surveillance is needed in Ukraine and the countries neighboring Ukraine to understand PCV2 transmission in Ukraine.

Our small sample size was the biggest limitation in our study, preventing us from detecting rare genotypes in the domestic pig populations. By detecting co-infections, we increased the number of PCV2 sequences in our study. This allowed us to detect genotype a in our wild boar samples. This evidence shows that detecting and sequencing co-infections can reduce, but not prevent, the effects of low sample sizes on estimating PCV2 diversity. We would

recommend that future PCV2 surveillance studies detect and sequence co-infections. One solution to detecting co-infections is nanopore sequencing, which gives full genome length reads and does not need the cloning step that Sanger sequencing requires to detect co-infections. The read error of nanopore sequencing can be improved with RCR, which has been used to enrich for PCV2 genomes in the past. For future studies looking at Ukraine or its neighboring countries, we would recommend larger sample sizes, detecting co-infections, and using nanopore sequencing instead of Sanger sequencing.

References:

Bao, F., S. Mi, Q. Luo, H. Guo, C. Tu, G. Zhu, and W. Gong. 2018. "Retrospective Study of Porcine Circovirus Type 2 Infection Reveals a Novel Genotype PCV2f." Transboundary and Emerging Diseases 65 (2): 432–40. https://doi.org/10.1111/tbed.12721.

Bhattacharjee, U., A. Sen, and I. Sharma. 2021. "A Retrospective Study Reveals the Porcine Circovirus-2f Genotype Predominant in the Indigenous Pig Population of North-Eastern India." Infection, Genetics and Evolution 96: 105100. https://doi.org/https://doi.org/10.1016/j.meegid.2021.105100.

Breitbart, M., E. Delwart, K. Rosario, J. Segalés, and A. Varsani. 2017. "ICTV Virus Taxonomy Profile: Circoviridae." The Journal of General Virology 98 (8): 1997–98.

Chernomor, O., A. v. Haeseler, and B. Q. Minh. 2016. "Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices." Systematic Biology 65 (6): 997–1008.

Correa-Fiz, F., G. Franzo, A. Llorens, E. Huerta, M. Sibila, T. Kekarainen, and J. Segalés. 2020. "Porcine Circovirus 2 (Pcv2) Population Study in Experimentally Infected Pigs Developing Pcv2-Systemic Disease or a Subclinical Infection." Scientific Reports 10 (1): 17747. https://doi.org/10.1038/s41598-020-74627-3.

Correa-Fiz, F., G. Franzo, A. Llorens, J. Segalés, and T. Kekarainen. 2018. "Porcine Circovirus 2 (PCV-2) Genetic Variability Under Natural Infection Scenario Reveals a Complex Network of Viral Quasispecies." Scientific Reports 8 (1): 15469. https://doi.org/10.1038/s41598-018-33849-2.

Coster, W. D. , S. D'Hert, D. T. Schultz, M. Cruts, C. V. Broeckhoven. 2018. "NanoPack: visualizing and processing long-read sequencing data." Bioinformatics 34 (15): 2666–69. https://doi.org/10.1093/bioinformatics/bty149.

Cságola, A., S Kecskeméti, G. Kardos, and T. Tuboly. 2006. "Genetic Characterization of Type 2 Porcine Circoviruses Detected in Hungarian Wild Boars." Archives of Virology 151 (3): 495–507.

Danecek, P., J. K Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, et al., 2021. "Twelve Years of SAMtools and BCFtools." GigaScience 10 (2): giab008. https://doi.org/10.1093/gigascience/giab008.

Dezen, D., F. A. M. Rijsewijk, T. F. Teixeira, C. L. Holz, S. P. Cibulski, O. A. F. Cláudia, O. A. Dellagostin, and P. M. Roehe. 2010. "Multiply-Primed Rolling-Circle Amplification (MPRCA) of Pcv2 Genomes: Applications on Detection, Sequencing and Virus Isolation." Research in Veterinary Science 88 (3): 436–40. https://doi.org/10.1016/j.rvsc.2009.10.006.

Dudar, L. V., I. Budzanivska, and V. P. Polishchuk. 2018. "Genetic Characterization of Porcine Circovirus Type 2 (Pcv2) from Wild Boars Detected in Different Regions of Ukraine." Biopolymers and Cell 34 (1): 41–48. http://dx.doi.org/10.7124/bc.00096F.

Franzo, G., and J. Segalés. 2020. "Porcine Circovirus 2 Genotypes, Immunity and Vaccines: Multiple Genotypes but One Single Serotype." Pathogens (Basel, Switzerland) 9 (12): 1049. https://doi.org/10.3390/pathogens9121049.

Franzo, G., and J. Segalés. 2018. "Porcine Circovirus 2 (PCV-2) Genotype Update and Proposal of a New Genotyping Methodology." PloS One 13 (12): e0208585.

Franzo, G., S. Tinello, L. Grassi, C. M. Tucciarone, M. Legnardi, M. Cecchinato, G. Dotto, et al., 2020. "Free to Circulate: An Update on the Epidemiological Dynamics of Porcine Circovirus 2 (PCV-2) in Italy Reveals the Role of Local Spreading, Wild Populations, and Foreign Countries." Pathogens 9 (3): 221. https://doi.org/10.3390/pathogens9030221.

Fu, L., B. Niu, Z. Zhu, S. Wu, and W. Li. 2012. "CD-HIT: Accelerated for Clustering the Next-Generation Sequencing Data." Bioinformatics (Oxford, England) 28 (23): 3150–52.

Gallardo, C. M, S. Wang, D. J. Montiel-Garcia, and B. E. Torbett. 2021. "MrHAMER Yields Highly Accurate Single Molecule Viral Sequences Enabling Analysis of Intra-Host Evolution." Nucleic Acids Research 49 (12): e70. https://dio.org/10.1093/nar/gkab231.

Gebhardt, J. T., M. D. Tokach, S. S. Dritz, J. M. DeRouchey, J. C. Woodworth, R. D. Goodband, and S. C. Henry. 2020. "Postweaning Mortality in Commercial Swine Production II: Review of Infectious Contributing Factors." Translational Animal Science 4 (2): txaa052. https://doi.org/10.1093/tas/txaa052.

He, J., J. Cao, N. Zhou, Y. Jin, J. Wu, and J. Zhou. 2013. "Identification and Functional Analysis of the Novel Orf4 Protein Encoded by Porcine Circovirus Type 2." J Virol. 87 (3): 1420 - 1429. http://dx.doi.org/10.1128/JVI.01443-12.

Hoang, D. T., O. Chernomor, A. Haeseler, B. Q. Minh, and L. S. Vinh. 2018. "UFBoot2: Improving the Ultrafast Bootstrap Approximation." Molecular Biology and Evolution 35 (2): 518–22.

Kalyaanamoorthy, S., B. Q. Minh, T. K. F. Wong, A. Haeseler, and L. S. Jermiin. 2017. "ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates." Nature Methods 14 (6): 587–89.

Karuppannan, A. K., and T. Opriessnig 2017. "Porcine Circovirus Type 2 (Pcv2) Vaccines in the Context of Current Molecular Epidemiology." Viruses 9 (5): 99. https://doi.org/10.3390/v9050099.

Katoh, K., and D. M. Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." Molecular Biology and Evolution 30 (4): 772–80.

Kearse, M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, et al., 2012. "Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data." Bioinformatics (Oxford, England) 28 (12): 1647–49.

Kleymann, A., E. Soto, O. Illanes, Y. S. Malik, C. Fuentealba, and S. Ghosh. 2020. "High Rates of Detection and Complete Genomic Analysis of Porcine Circovirus 2 (Pcv2) in the Lesser Antilles Island of St. Kitts: Identification of PCV2b-PCV2d Recombinants." Transboundary and Emerging Diseases 67 (6): 2282–89. https://doi.org/10.1111/tbed.13583.

Leigh, D. M., C. Schefer, and C. Cornejo. 2020. "Determining the Suitability of MinION's Direct RNA and DNA Amplicon Sequencing for Viral Subtype Identification." Viruses 12 (8): 801. https://doi.org/10.3390/v12080801.

Li, D., J. Wang, S. Xu, S. Cai, C. Ao, L. Fang, S. Xiao, H. Chen, and Y. Jiang. 2018. "Identification and Functional Analysis of the Novel Orf6 Protein of Porcine Circovirus Type 2 in Vitro." Veterinary Research Communications 42 (1): 1–10. https://doi.org/10.1007/s11259-017-9702-0.

Li, H.. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." Bioinformatics (Oxford, England) 34 (18): 3094–3100.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." Bioinformatics (Oxford, England) 25 (16): 2078–79.

Li, W., and A. Godzik. 2006. "Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences." Bioinformatics (Oxford, England) 22 (13): 1658–59.

Liu, J., I. Chen, and J. Kwang. 2005. "Characterization of a Previously Unidentified Viral Protein in Porcine Circovirus Type 2-Infected Cells and Its Role in Virus-Induced Apoptosis." J Virol. 79 (13): 8262-74. http://dx.doi.org/10.1128/JVI.79.13.8262-8274.2005.

Lv, Q., K. Guo, H. Xu, T. Wang, and Y. Zhang. 2015. "Identification of Putative Orf5 Protein of Porcine Circovirus Type 2 and Functional Analysis of GFP-Fused Orf5 Protein." PloS One 10 (6): e0127859. https://doi.org/10.1371/journal.pone.0127859.

Martin, D. P., B. Murrell, M. Golden, A. Khoosal, and B. Muhire. 2015. "Rdp4: Detection and Analysis of Recombination Patterns in Virus Genomes." Virus Evolution 1 (1): vev003. https://doi.org/10.1093/ve/vev003.

Minh, B. Q., A. S. Heiko, O. Chernomor, D. Schrempf, M. D. Woodhams, A. Haeseler, and R. Lanfear. 2020. "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era." Molecular Biology and Evolution 37 (5): 1530–34.

Navidad, P. D., H. Li, A. Mankertz, and B. Meehan. 2008. "Rolling-Circle Amplification for the Detection of Active Porcine Circovirus Type 2 DNA Replication in Vitro." Journal of Virological Methods 152 (1-2): 112–16. https://doi.org/10.1016/j.jviromet.2008.05.026.

NCBI Resource Coordinators. 2016. "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res 44(D1): D7-19. https://doi.org/10.1093/nar/gkv1290.

Nugroho, W., F. Hemmatzadeh, S. Artanto, and M. P. Reichel. 2016. "Complete Genome Characteristics of Porcine Circovirus Type 2 (Pcv2) Isolates from Papuan Pigs, Indonesia." International Journal of Advanced Veterinary Science and Technology 5 (1): 239–47.

Ouyang, T., X. Zhang, X. Liu, and L. Ren. 2019. "Co-Infection of Swine with Porcine Circovirus Type 2 and Other Swine Viruses." Viruses 11 (2): 185. https://doi.org/10.3390/v11020185.

Paradis, E., Schliep K. 2019. "ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R." Bioinformatics, 35 (3):526-528. https://doi.org/10.1093/bioinformatics/bty633.

Park, K. H., and C. Chae. 2021. "The Prevalence of Porcine Circovirus Type 2e (PCV2e) in Korean Slaughter Pig Lymph Nodes When Compared with Other Pcv2 Genotypes." Transboundary and Emerging Diseases 68 (6): 3043–47. *https://doi.org/10.1111/tbed.13975*.

Raev, S., A. Yuzhakov, and T. Aliper. 2021. "Whole-Genome Analysis of Porcine Circovirus Type 2 in Russia." Pathogens (Basel, Switzerland) 10 (12): 1631. https://doi.org/10.3390/pathogens10121631.

Rice, P., I. Logden, and A. Bleasby. 2000. "EMBOSS: The European Molecular Biology Open Software Suite." Trends in Genetics 16 (6): 276–77.

Rudova, N. G., V. I. Bolotin, O. S. Solodiankin, and A. P. Gerliovych. 2019. "Porcine Circovirus Type II Screening in Feral Swine Populations in Ukraine." Journal for Veterinary Medicine, Biotechnology and Biosafety 5: 10–12. https://doi.org/10.36016/JVMBBS-2019-5-3-2.

Song, S., G. Park, S. Choe, R. M. Cha, S. Kim, B. Hyun, B. Park, and D. An. 2020. "Genetic Diversity of Porcine Circovirus Isolated from Korean Wild Boars." Article. Pathogens 9 (6): 457. https://doi.org/10.3390/pathogens9060457.

USDA. 2021. "Livestock and Poultry: World Markets and Trade." Global Market Analysis. https://apps.fas.usda.gov/psdonline/circulars/livestock_poultry.pdf.

Wang, L. G., T. T. Y. Lam, S. Xu, Z. Dai, L. Zhou, T. Feng, P. Guo, et al., 2020. "Treeio: An r Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data." Molecular Biology and Evolution 37 (2): 599–603. https://doi.org/10.1093/molbev/msz240.

Wang, Y., L. Noll, N. Lu, E. Porter, C. Stoy, W. Zheng, X. Liu, M. Peddireddi, M. Niederwerder, and J. Bai. 2020. "Genetic Diversity and Prevalence of Porcine Circovirus Type 3 (Pcv3) and Type 2 (Pcv2) in the Midwest of the USA During 2016-2018." Transbound Emerg Dis. 67 (3): 1284–94. https://doi.org/10.1111/tbed.13467.

Wickham, H. 2016. "ggplot2: Elegant Graphics for Data Analysis." Springer-Verlag New York 17: 160-167.

Xiao, C., K. M. Harmon, P. G. Halbur, and T. Opriessnig. 2016. "PCV2d-2 Is the Predominant Type of Pcv2 DNA in Pig Samples Collected in the U.S. During 2014-2016." Veterinary Microbiology 197 (December): 72–77. https://doi.org/10.1016/j.vetmic.2016.11.009.

Yang, S., S. Yin, Y. Shang, B. Liu, L. Yuan, M. U. Z. Khan, X. Liu, and J. Cai. 2018. "Phylogenetic and Genetic Variation Analyses of Porcine Circovirus Type 2 Isolated from China." Transboundary and Emerging Diseases 65 (2): e383–92. https://doi.org/10.1111/tbed.12768.

Yu, G. 2020. "Using Ggtree to Visualize Data on Tree-Like Structures. Current Protocols in Bioinformatics." Current Protocols in Bioinformatics, 69 (1): e96. https://doi.org/10.1002/cpbi.96.

Yu, G., D. K. Smith, Z. Huachen, Y. Guan, and T. T. Larn. 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 (1): 28–36. https://doi.org/10.1111/2041-210X.12628.

Yu, G., T. T. Larn, Z. Huachen, and Y. Guan. 2018. "Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree." Molecular Biology and Evolution 35 (12): 3041–43. https://doi.org/10.1093/molbev/msy194.

Zhao, G. H., W. Cheng, P. J. Zhang, Y. S. Han, and D. K. Chen. 2010. "Novel Genotypes of Type 2 Porcine Circovirus (Pcv2) in PMWS Pigs in China Between 2008 and 2009." Journal of Animal and Veterinary Advances 9 (24): 3083–91. https://doi.org/10.3923/javaa.2010.3083.3091.

Zheng, G., Q. Lu, F. Wang, G. Xing, H. Feng, Q. Jin, Z. Guo, et al., 2020. "Phylogenetic Analysis of Porcine Circovirus Type 2 (Pcv2) Between 2015 and 2018 in Henan Province, China." BMC Veterinary Research 16 (1): 6. https://doi.org/10.1186/s12917-019-2193-1.

Chapter 2: Accuracy and Completeness of Long Read Metagenomic Assemblies.[2]

Abstract:

By looking at interactions between microbes, we can learn how microbes influence the surrounding environment, contribute to human health, and understand which pathogen interactions result in differences in disease severity. Metagenomics has been utilized as a tool to explore such interactions. Metagenomic assemblies built using long read nanopore data depend on the read level accuracy. The read level accuracy of nanopore data has made dramatic improvements. However, we do not know if the increased read level accuracy allows for faster assemblers to make as accurate metagenomic assemblies as slower assemblers. Here, we present the results of a benchmarking study comparing three commonly used long read assemblers, Flye, Raven, and Redbean. We used a prepared DNA standard of seven bacteria as our input community. We prepared a sequencing library on the VolTRAX V2 sequence using a MinION mk1b. We basecalled using the latest version of Guppy with the super-accuracy model. We found that increasing read depth benefited each of the assemblers, and nearly complete community member chromosomes were assembled with as little as 10x read depth. Polishing assemblies using Medaka had a predictable improvement. Among the bacterial community, some assemblers struggled with particular members, but we found Flye to be the most robust across taxa. We found Flye was the most effective assembler for recovering plasmids. Based on Flye's consistency for chromosomes and increased effectiveness at assembling plasmids, we would recommend using Flye in future metagenomic studies.

---

[2]Article to be published by Buttler, J. and D. M. Drown

Introduction:

Current methods for sequencing microbes involve isolating and sequencing individual community members, sequencing 16S rRNA genes, or metagenomics (Garmendía et al., (2012); Petersen et al., 2019). Isolating individual microbes requires culturing, which is often difficult or practically impossible (Garmendía et al., 2012). Sequencing 16S rRNA genes cannot provide information on the entire genomes, such as genes that might increase virulence or provide antibiotic resistance (Petersen et al., 2019). Metagenomics is a method where an entire sample is sequenced and the individual community members are sorted out later with bioinformatic analyses (Bai et al., 2022). Metagenomic sequencing can detect most DNA community members, including unculturable microbes and novel community members (Garmendía et al., 2012). The individual community member sequences can be studied to identify pathogens in difficult to diagnose diseases, genes that may increase virulence, and look for correlations between co-infecting pathogens that increase disease severity (Garmendía et al., 2012; Kumar et al., 2018; Petersen et al., 2019; Qin et al., 2018). Currently, most metagenomic studies use Illumina based technology, which produces highly accurate, short reads (Petersen et al., 2019). Over the past several years, Oxford Nanopore has increased sequencing throughput and yield to be reasonable for metagenomic studies, but these reads are error prone, but are also orders of magnitude longer than short read platforms (Petersen et al., 2019).

The short reads (150-300 base pairs) from Illumina sequencing make genome assembly difficult for complex communities. Short read lengths do not facilitate merging the multiple contigs built for a genome into a single contig, resulting in highly fragmented assemblies (Goldstein et al., 2018). Also, short reads cannot span long repeat regions, causing repeat regions to shrink, providing less complete assemblies (Sevim et al., 2019). More complete genomes can

be made using long read sequencing technologies, such as Oxford Nanopore or PacBio (Goldstein et al., 2018). Oxford Nanopore sequencers platforms (e.g., MinION) have produced reads greater than 2 mb long in length and can easily produce libraries with mean read lengths of greater than 16 kb, which makes it possible to assemble long repeat regions (Amarasinghe et al., 2020; Payne et al., 2019; Jain et al., 2018). However, the high error rates of nanopore sequencing prevent short read assemblers from producing quality assemblies with long read data (Jain et al., 2018; Latorre-Pérez et al., 2020).

Three commonly used long read specific assemblers include Flye, Raven, and Redbean (Yang et al., 2021; Latorre-Pérez et al., 2020; Breckell and Silander 2021; Chen et al., 2020b). Flye is a long read metagenomic assembler that constructs a repeat graph to assemble and polish contigs. These contigs are then used to build an assembly graph with A-Bruijn (Kolmogorov et al., 2019; Kolmogorov et al., 2020). Previous studies found that while Flye can build more accurate metagenomic assemblies than Raven or Redbean, it also takes more time and more memory (Wick and Holt 2019; Latorre-Pérez et al., 2020). Raven is a fast assembler that uses an Overlap-Layout-Consensus (OLC) approach to build an assembly graph from raw reads (Vaser and Sikić 2021). For some individual assemblies Raven can have comparable accuracy to Flye after the assemblies are polished, but has less accuracy for metagenomic assemblies (Wick and Holt 2019; Breckell and Silander 2021). Redbean is another fast assembler that follows the OLC concept by using a fuzzy de Bruijn graph to build assemblies from raw reads (Ruan and Li 2020; Rizzi et al., 2019). Previous studies have found that Redbean uses more memory and builds less accurate assemblies than Raven (Wick and Holt 2019; Latorre-Pérez et al., 2020).

Benchmarking is used to compare bioinformatics tools and to determine which tool is best suited for a particular task (Yang et al., 2021; Aniba et al., 2010). Benchmarking studies for

metagenomic assemblers often include well characterized communities or mock communities, like one of the many ZymoBIOMICS Microbial Community Standards (Latorre-Pérez et al., 2020; Sereika et al., 2021; Goldstein et al., 2018; Kolmogorov et al., 2020). Mock communities are synthetic communities composed of multiple known microbes, with known genome sequences and abundances (Bokulich et al., 2016). This information allows for accurate assessment and comparison of assemblers for metagenomic data.

A past benchmarking study using a ZymoBIOMICS Microbial Community Standard found that Raven and Redbean could not build complete assemblies for the *E. coli* and *Salmonella enterica* community members (Latorre-Pérez et al., 2020). Raven did well for the other community members in the Zymo mock community (Latorre-Pérez et al., 2020). Raven also performs well for individual assemblies of E. coli (Breckell and Silander 2021; Chen et al., 2020b). These differences in performance suggest that the high read error rate may cause Raven to confuse genome fragments from other community members with *E. coli* fragments (Latorre-Pérez et al., 2020; Breckell and Silander 2021; Chen et al., 2020b). If so, a higher read accuracy, as produced by the current versions of Guppy, may allow Raven to assemble all community members from the mock community with similar accuracy to Flye. Another weakness of Raven and Redbean is that they often fail to build assemblies for plasmids (Wick and Holt 2019). These weaknesses may limit the performance of Raven and Redbean for complex metagenomic assemblies, where plasmids may be common and particular community members may be present.

Improvements in converting the electrical signal from nanopore sequencing to nucleotides (basecalling) have led to increased read level nanopore sequence accuracy. The release of the super-accuracy model to Guppy has pushed modal accuracy to 98%

(https://nanoporetech.com/accuracy). As the individual reads improve in quality, faster assemblers, like Raven, may be able to build assemblies of problematic community members, like E. coli, with comparable accuracy to slower, but more accurate assemblers, like Flye.

Here, we compare the completeness and accuracy of metagenomic assemblies built with Flye, Raven, and Redbean. We used data basecalled with the super-accuracy model. From this comparison, we will contrast the areas of strength and weakness of long read metagenomic assemblers

Methods:

Sequencing:

We sequenced a mock community standard (ZymoBIOMICS HMW DNA Standard, catalog #D6322) using long read sequencing to compare metagenomic assembly methods. The HMW DNA standard is a synthetic microbial community comprising three gram-negative bacteria, four gram-positive bacteria, and one yeast (Table 2-1). Bacterial community members have a genome size between 2.73 mb to 6.792 mb, a GC content between 32.9% to 66.2% (Table 2-1). Each bacteria community also contributed 14% of nucleotides in the mock community (Table 2-1). The template DNA in the community has a mean length of 24 kb. Sequences can be found at https://s3.amazonaws.com/zymo-files/BioPool/D6322.refseq.zip,

We used 1 ug of the HMW DNA standard as input for the VolTRAX V2 (Oxford Nanopore Technologies [ONT]) to prepare a sequencing library using the (VSK-VSK002 workflow). The VolTRAX library is analogous to the Rapid Sequencing library and results in additional DNA template fragmentation as the library is prepared. We sequenced the prepared library using the MinION mk1b (ONT) on a r9.4.1 flow cell (FLO-MIN106) for 48 hours

(VSK002 script). We basecalled the reads using Guppy version 5.0.7 with the super-accuracy model (-c dna_r9.4.1_450bps_sup.cfg). We set a minimum quality filter of ≥ 10 (-min_qscore 10).

Table 2-1: Community members in the HMW DNA standard. No. genomes refers the number of genomes.

NRRL Accession No.	Organism	Number plasmids	GC Content (%)	Genome size (mb)	Gram stain	% nucleotides	No. genomes
B-354	*Bacillus subtilis*	0	43.9	04.045	+	14	13.20
B-537	*Enterococcus faecalis*	0	37.5	02.845	+	14	18.80
B-1109	*Escherichia coli*	1	46.7	04.875	-	14	10.90
B-33116	*Listeria monocytogenes*	0	38.0	02.992	+	14	17.80
B-3509	*Pseudomonas aeruginosa*	0	66.2	06.792	-	14	07.80
B-4212	*Salmonella enterica*	0	52.2	04.760	-	14	11.20
B-41012	*Staphylococcus aureus*	3	32.9	02.730	+	14	19.60
Y-567	*Saccharomyces cerevisiae*	NA	38.3	12.100	NA	02	00.63

To generate a subsample of reads, we used trycycler (Wick et al., 2021). We used a genome size of 42 mb and the –min read depth parameter to generate sub-samples of 420 mb, 840 mb, 1260 mb, 2100 mb, 4200 mb, and 8400 mb. These total yields should theoretically represent 10x, 20x, 30x, 50x, 100x, and 200x read depths. At each read depth, we produced 12 subsamples for a total of 72 datasets. The mean number of bases, mean longest read length, and mean N50 for each read depth was found using NanoStat --fastq (Coster et al., 2018).

Assembly and Polishing:

For this comparison, we used three commonly used long read assemblers to construct metagenomic assemblies of our datasets, Flye, Raven, and Redbean. We used Flye version

v2.8.3 (Kolmogorov et al., 2020) with default parameters specifying nanopore reads (-nano-raw) and the following options in recover plasmids (-plasmids) and metagenomes (-meta). We used Raven v1.5.1 (Vaser and Sikić 2021) with default parameters. We used Redbean v2.5 (Ruan and Li 2020) with default parameters specifying nanopore reads (-x ont), and a genome size of 42 mb (-g 42m).

We polished all assemblies using one round of Racon v1.4.22 (Vaser et al., 2017) followed by one round of Medaka v1.4.3 (https://github.com/nanoporetech/medaka), specifying the super-accuracy model (-m r941_min_sup_g507). For Racon we used the ONT suggested parameters: score for matching bases (-m 8), score for mismatching bases (-x -6), gap penalty (-g -8), window size (-w 500), and mean quality threshold for each window (-q -1).

Quality Assessment:

We measured assembly quality and completeness with the genome fraction output by MetaQuast v5.1.0 (Mikheenko et al., 2016). For MetaQuast we used the references in Table 2-1 to measure the completeness of both the polished and unpolished metagenomic assemblies.

We measured assembly accuracy with the median Q-score output by Pomoxis (https://nanoporetech.github.io/pomoxis) assess_assembly. Pomoxis was used with the references in Table 2-1 to find the quality scores (Q-scores) of the assemblies. For each assembly, we calculated Q-scores for chromosomes and plasmids separately.

We completed all analysis, including assembly, polishing, and assembly quality assessment on a server with an Intel Core i9 9900K 3.6GHz Eight Core (16 thread) CPU, a Nvidia Quadro GV100 GPU, and 128 GB of ram. We measured the time required and the maximum memory used to build each assembly using GNU time with parameter -f %ee. The

time, assembly, polishing, MetaQuast, and Pomoxis steps were automated using custom bash scripts (https://github.com/jeremyButtler/assembler-scripts).

We used R v4.1.1 (R Core Team 2022) with ggplot2 (Wickham 2016) to build graphs for the metagenome fraction, genome fraction, median Q-scores, number of misassemblies, time, and maximum memory usage. The metagenome fraction was found by dividing the number of bases that were aligned to a community member in a replicate by the total bases in the community.

Results:

Subsampling Statistics:

We sequence the ZymoBIOMICS HMW DNA Standard on a nanopore sequencer and subsampled reads into subsamples of 420 mb (~10x read depth), 840 mb (~20x read depth), 1260 mb (~30x read depth), 2100 mb (~50x read depth), 4200 mb (~100x read depth), and 8400 mb (~200x read depth). For each targeted read depth, our mean number of bases was very close to are target number of bases pairs (Table 2-2). The mean N50 between our read depths only differed by 18 base pairs (5012 to 15030 bp) (Table 2-2). The mean number of reads for each read depth ranged between with 45673 reads at 10x depth and 913452 at 200x read depth (Table 2-2). Each time the read depth was doubled, we saw a two-fold increase in the mean number of reads (Table 2-2).

Table 2-2: Subsample statistics for each read depth. Each read depth had 12 subsamples. No. bases is number of bases

Read depth	Mean No. reads	Mean N50	Mean No. base pairs (mb)
10x	45673	15012.25	419.28
20x	91345	15020.75	838.87
30x	137018	15020.08	1258.72
50x	228363	15023.42	2099.35
100x	456726	15027.83	4198.77
200x	913452	15030.75	8398.78

Chromosome:

Genome Fraction:

Across all read depths, we found Flye produced assemblies with near 100% metagenome fractions (Figure 2-1 a). Even at our smallest read depth of 10x, Flye recovered nearly 100% of the metagenomic fraction (Figure 2-1 a). With increasing read depth, Raven and Redbean produced assemblies with improved metagenome fractions (Figure 2-1 a). Raven and Redbean reached a maximum metagenome fraction of 95% at 200x read depth (Figure 2-1 a). At the individual community member level, Raven and Redbean had the most difficulty in the assembly of *Escherichia coli* and *Salmonella enterica*, recovering less than 90% of the genome even at 200x read depth (Figure 2-1 b).

Accuracy (Q-score):

Across all read depths, we found Flye produced the most accurate metagenomic assemblies, followed by Raven, and then Redbean (Figure 2-2 a). Increased read depth and polishing, predicably improved the median quality scores (Q-scores) of assemblies from all assemblers (Figure 2-2 a). All assemblers had a large improvement in Q-scores between 10x and

50x read depth (Figure 2-2 a). At 200x read depth Flye reached a maximum Q-score of 50, while Raven and Redbean reached a maximum Q-score of 46 and 45 respectively (Figure 2-2 a).

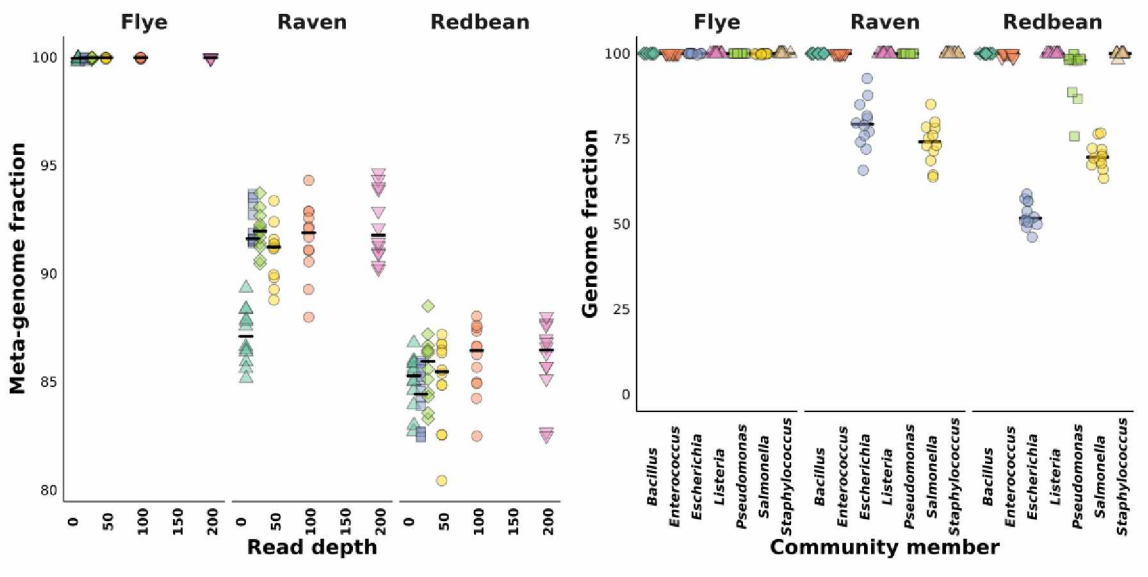

Figure 2-1: Chromosome completeness. a — The metagenome fractions for all isolates in a replicate. Horizontal bars indicate the medians across replicates. Color and shape indicate different read depths, b — The genome fraction for each isolate at 200x read depth. Horizontal bars indicate the median value across replicate samples. Color and shape indicate different community members.

At the individual community member level, Raven and Redbean had the most difficulty in the assembly of *E. coli* and *S. enterica* (Figure 2-2 b). *E. coli* assemblies produced with Flye were more accurate (median Q-score 42.81) than those from Raven (26.73) and Redbean (under 20). *S. enterica* assemblies produced by Flye were highly accurate (median Q-score 50) while Raven was slightly less accurate (42.54), but Redbean produced error prone assemblies (under

20) (Figure 2-2 b). We also found that Raven and Redbean, but not Flye, had over 10 miss-assemblies for *E. coli* and *S. enterica* (Figure A-6).

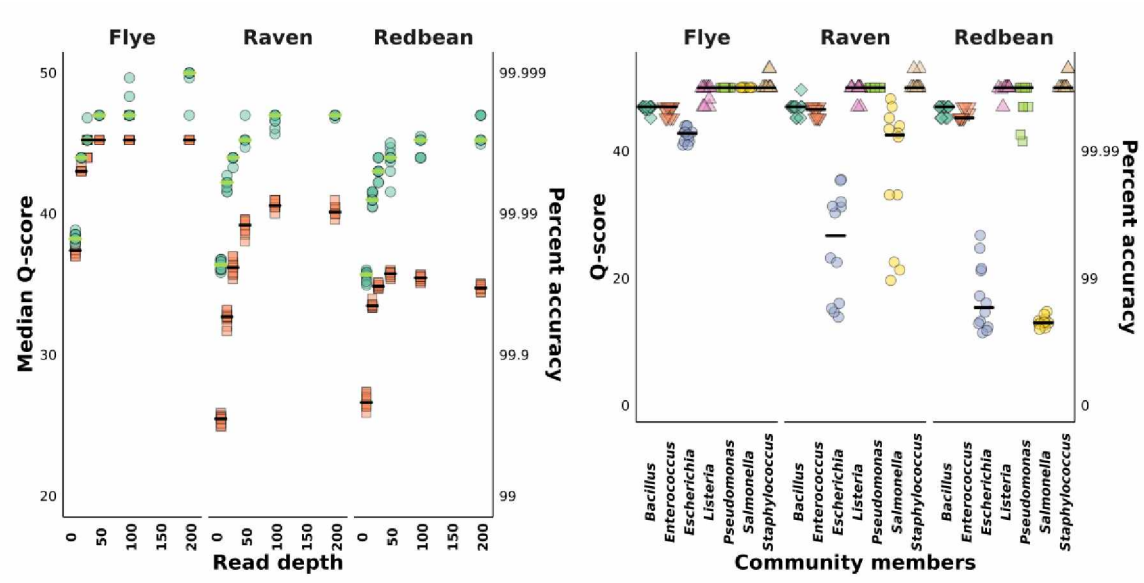

Figure 2-2: Chromosome accuracy. a — Q-score for all chromosomes in each replicate. Median quality score (Q-score) for chromosomes in each assembly. Horizontal bars indicate the median across replicates. Green circles are polished assemblies and green horizontal bars indicate polished assemblies., b — median Q-score for each isolate at 200x read depth. Horizontal bars indicate the median across replicates. Color and shape indicate different community members.

Plasmids:

Genome Fraction:

Across all read depths, we found Flye recovered over 94% of the plasmid genomes (Figure 2-3 a). At 50x read depth Flye recovered nearly 100% of the plasmid genomes (Figure 2-3 a). After 20x read depth, Raven and Redbean decreased the recovery of plasmid genomes (Figure 2-3 a). Raven and Redbean assembled a maximum of 95% of the plasmid genomes at 20x read depth (Figure 2-3 a).

At the individual plasmid level, Raven and Redbean both struggled with the plasmids smaller than 7 kb (Figure 2-3 b). Raven and Redbean assembled more of plasmids under 7 kb at 30x and 50x read depth than at 200x read depth (Figures A-7 b, A-7 a). Raven could assemble the 2995 bp plasmid for all replicates at 50x read depth, but not at 200x read depth (Figure A-7 b).

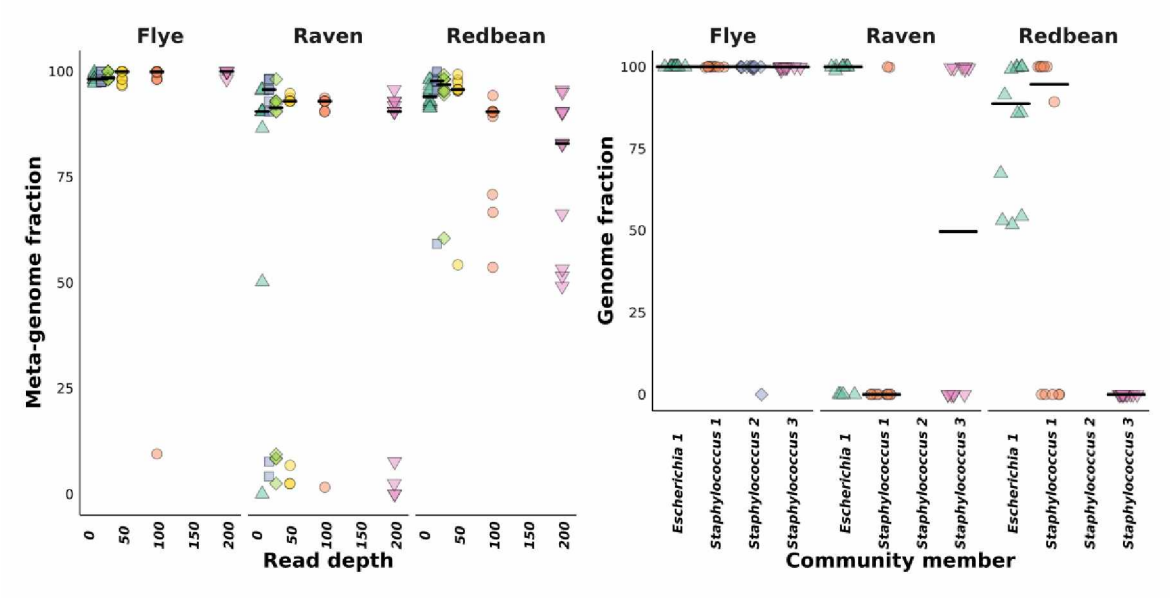

Figure 2-3: Plasmid completeness. a — Meta genome fraction for plasmids from each replicate. Horizontal bars indicate the median meta genome fraction across replicates., b — genome fraction for each plasmid at 200x read depth. Horizontal bars indicate the median genome fraction across replicates. *E. coli* is 110009 bases, *S. aureus* 1 is 6339 bases, *S. aureus* 2 is 2218 bases, and *S. aureus* 3 is 2995 bases long.

Accuracy (Q-score):

Across all read depths, we found that Flye assembled the most accurate plasmids (Figure 2-4 a). With increased read depth, Flye produced more accurate plasmid assemblies (Figure 2-4

a). However, polishing did not improve the accuracy of Flye plasmid assemblies (Figure 2-4 a). At 100x read depth, Flye plasmid assemblies had a median Q-score of 50 (Figure 2-4 a).

Across all read depths, polishing Raven and Redbean plasmid assemblies resulted in more accurate plasmid genomes (Figure 2-4 a). Increased read depth did not improve the accuracy of Raven produced plasmid assemblies (Figure 2-4 a). Beyond 50x read depth, Redbean produced more accurate plasmid assemblies than Raven (Figure 2-4 a). However, Raven built assemblies with higher accuracy than Redbean when the read depth was under 100x (Figure 2-4 a).

At the individual plasmid level, only the *E. coli* 110009 bp plasmid could be assembled by all assemblers (Figure 2-4 b). All assemblers had a similar accuracy for the *E. coli* plasmid, (Q-scores around 26, Figure 2-4 b). All assemblers were able to assemble the *E. coli* plasmid without any missassemblies, but Flye and Redbean did have misassemblies in assemblies of plasmids under 7 kb (Figure A-8). However, Flye assembled almost all replicates for each plasmid and had near perfect accuracy for plasmids under 7 kb (Figure 2-4 b).

Assembly Time and Memory Usage

Predictably, we found that assemblers needed more time and memory to build an assembly with greater initial read input (Figure 2-5 a and Figure 2-5 b). When the read depth was under 50x, all assemblers used less than 30 minutes to complete an assembly (Figure 2-5 a). At 200x read depth, Flye required over 400 minutes to complete an assembly. With that same input Raven required just 50 minutes and Redbean required only 25 minutes to complete an assembly (Figure 2-5 a).

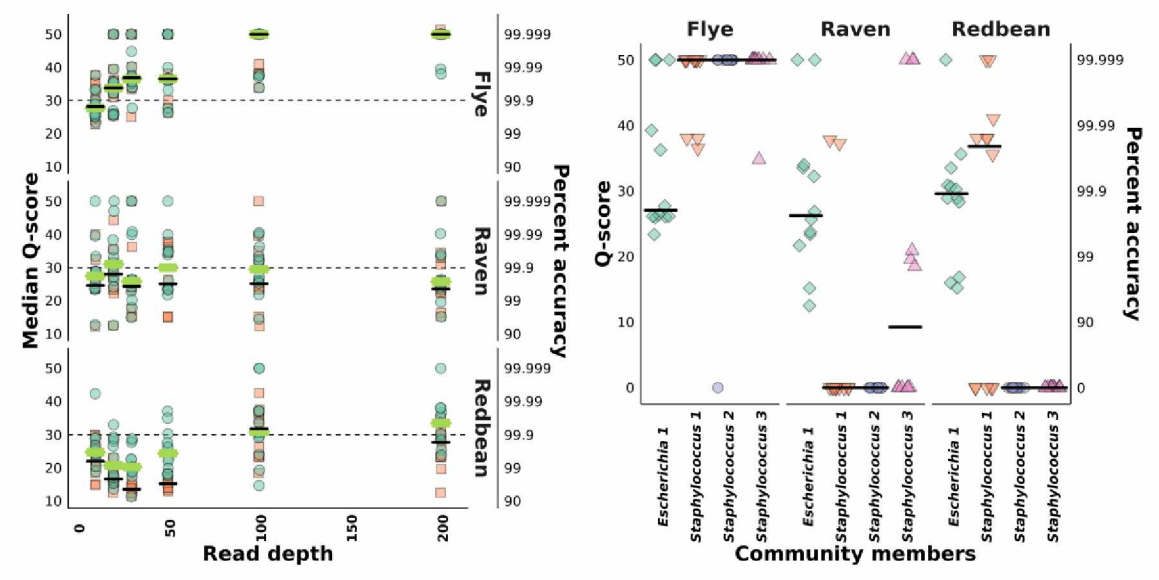

Figure 2-4: Plasmid accuracy. a — Median Q-score of all replicates at 10x, 20x, 30x, 50x, 100x, and 200x read depth. Horizontal bars indicate the median across replicates. Green circles and horizontal green bars indicated polished assemblies. The dashed line indicates the highest Q-score for Raven, b — median Q-score for each plasmid at 200x read depth. Q 50 is Q-score of infinity (100% accuracy). Horizontal bars indicate the median across replicates, *E. coli* is 110009 bases, *S. aureus* 1 is 6339 bases, *S. aureus* 2 is 2218 bases, and *S. aureus* 3 is 2995 bases long.

Across all read depths, Raven and Redbean needed less memory than Flye to build an assembly (Figure 2-5 b). At read depths under 100x, Raven needed less memory than Redbean to build an assembly (Figure 2-5 b). At 50x read depth, Raven used 5.5 Gb of memory to build an assembly, while Redbean used 7.7 Gb of memory to build an assembly (Figure 2-5 b). Beyond 50x read depth, Raven used more memory than Redbean to build an assembly (Figure 2-5 b). At 200x read depth, Raven used 15.6 Gb of memory to build an assembly, while Redbean used 10.5 Gb of memory to build an assembly (Figure 2-5 b). Flye used the most memory to build an assembly, requiring 10.6 Gb of memory at 10x read depth and 55.8 Gb of memory at 200x read depth (Figure 2-5 b).

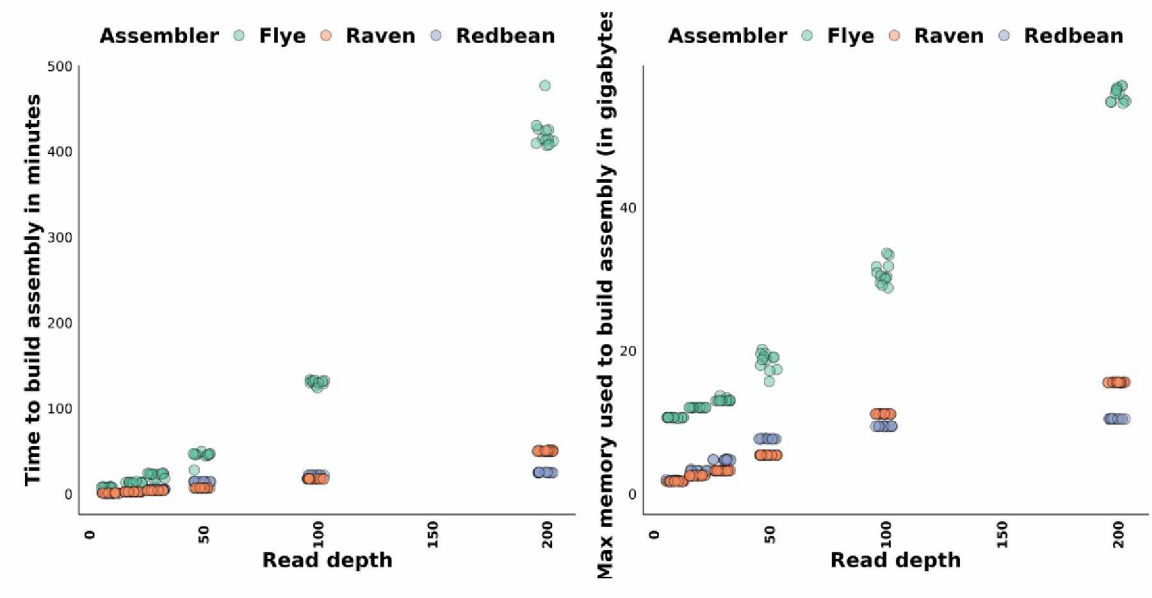

Figure 2-5: Time and memory usage of each assembler a — Amount of time to run each assembler in minutes., b — Amount of memory used by each assembler.

Discussion:

We compared the accuracy and completeness of assemblies built by three long read assemblers, Flye, Raven, and Redbean. For chromosomes, we found Flye was the only assembler that made near complete and accurate genomes for all community members. For plasmids, we found Flye was the only assembler that could assemble all plasmids reliably. However, Raven and Redbean were better than Flye in time and memory usage.

Effect of Read Depth:

For chromosomes, we found with increased read depth, all assemblers made more accurate and complete assemblies. We found that there was a sharp increase in accuracy between 10x and 50x read depth. At 10x read depth, Flye was the only assembler that had near complete metagenome fractions. Showing that Flye could be used for low read depth datasets. However,

for more accurate assemblies, future metagenomic studies should continue to aim for a read depth of at least 30x.

For plasmids, we found most plasmids under 7 kb were assembled best by Flye, with the most plasmids recovered at 200x read depth. However, Raven and Redbean had decreased small plasmid recovery at deeper read depths and performed best at read depths between 20x or 50x. The decrease in assembled plasmids under 7 kb at deeper read depths suggests that Raven and Redbean are discarding smaller reads and contigs at deeper read depths. This results in plasmids under 7 kb being missed at deeper read depths but being kept at more shallow read depths. These observations are consistent with Wick and Holt (2019) who also found that both Raven and Redbean struggled to complete assemblies of smaller plasmids. These results highlight the weakness in Raven and Redbean for recovering plasmids.

We found that the accuracy of the larger *E. coli* plasmid (Q-score under 30) was much lower than the chromosome assemblies (40 or 50). This suggests that the plasmids have more error prone regions, assemblers are more likely to make misassemblies for plasmids, or that the plasmid references have more errors than the chromosome references. For reference errors, Flye could often assemble plasmids under 3 kb with no indels or mismatches and only a few (2 to 3) misassemblies. So, errors in the references are a less likely, but still a potential explanation for why Flye, Raven, and Redbean had poor performance for the *E. coli* plasmid.

For misassembly errors, we found all assemblers had no misassemblies in *E. coli* plasmid assemblies at 200x read depth, showing that the problem is not from misassemblies in the *E. coli* plasmid. Other sources of errors in the *E. coli* plasmid could be from more error prone regions in the *E. coli* plasmid or errors inserted by the assemblers in the assembly of the *E. coli* plasmid.

During the process of generating these results a new version of Flye was released (v2.9), which included improvements for recovering plasmids and accounts for the improved accuracy from the super-accuracy model. However, more testing with a broader range of plasmid sizes is needed to determine if the errors are from error prone regions or from the assembler.

Metagenomics and Viruses:

Though our study looked at a mock microbial community mostly made of bacterial genomes, our results give insights in how reliable each assembler may be for viral metagenomic assemblies. The *E. coli* plasmid in our study is 110 kb long, which is close to or under the size of a large virus, such as the 170 to 190 kb long African swine fever virus genome (Gaudreault et al., 2020). While the smaller plasmids in our study are near the size of small viruses, such as porcine circovirus type 2, which is 1.76 kb long (Breitbart et al., 2017).

For the larger plasmids and likely larger viruses, we found that Raven or Redbean would likely work as well as Flye. However, only Flye could make reliable assemblies for the smaller plasmids and so, is the only reliable assembler for smaller viruses, such as porcine circovirus type 2. Even then Flye will often have a few misassemblies, so it might be best to use an assembler, like viralFlye that is designed for viruses (Antipov et al., 2022). However, viralFlye is specialized for virus detection and thus has limitations on the maximum genome size (Antipov et al., 2022). This may limit viralFlye's use for bacterial community members. Making Flye or assemblies made with both viralFlye and Flye the best option for sequencing mixed communities of viruses and bacteria.

Effect of Polishing:

We found that polishing improved the accuracy of all chromosome assemblies. However, for Flye and Redbean, polishing continued to improve the accuracy at 200x read depth, suggesting that even more data will improve the accuracy of polished Flye assemblies. To achieve highly accurate assemblies, we would recommend polishing and as much read depth as possible.

For Flye, polishing seemed to have little effect on the accuracy of plasmid assemblies. Instead, most plasmids smaller than 3 kb had no indels or mismatches at 200x read depth. Showing that polishing did not decrease the accuracy of the perfect assemblies. Likely, the high accuracy was due to the genome sizes of the plasmids being smaller than the error rate of nanopore consensuses assemblies (one error in 10000 bases for chromosomes). The idea of size is somewhat supported by the tenfold larger *E. coli* plasmid assemblies built by Flye having a much higher error rates (Median Q-score ~ 28) than the plasmids under 3 kb. Since polishing provides large improvements for chromosomes, while having no decrease in accuracy for plasmids, we would recommend polishing all metagenomic assemblies.

Problem Isolates:

We found that Raven and Redbean struggled to build assemblies of *E. coli* and *Salmonella enterica* Latorre-Pérez et al., (2020), also found that Raven and Redbean struggled with *E. coli* and *S. enterica* strains for the log and even mock communities from ZymoBIOMICS, both of which used the same *E. coli* and *S. enterica* strains as the HMW DNA Standard mock community. However, in a non-metagenomic study, Chen et al., (2020a) found that Raven could assemble complete genomes for a different strain of *E. coli* and a possibly a

different serovar of *Salmonella enterica* (*S.* Typhimurium). This suggests that either the strain of *E. coli* used in the mock community is a problematic strain or that assembling genomes of *E. coli* combined with *S. enterica* is difficult. Breckell and Silander (2021) found that strain specific characteristics of different *E. coli* strains made some *E. coli* strains harder for assemblers to assemble, so it is possible that the strain of *E. coli* in the mock community could be a more difficult strain to assemble. However, Breckell and Silander (2021) found that problematic strains of *E. coli* were problematic for all assemblers. Flye had very few misassemblies for *E. coli* at 200x read depth and had more accurate assemblies of *E. coli* than Raven or Redbean. This evidence is not consistent with a problematic strain of *E. coli*. However, we cannot fully eliminate the idea that the strain of *E. coli* in the mock community may be more difficult strain to assemble.

Other Studies:

To the best of our knowledge, our study is the first study to compare metagenomic assemblies made by Flye, Raven, and Redbean using super-accurate basecalled reads. We found Flye still made more accurate and complete genomes than Raven or Redbean when more accurate reads are used. This agrees with what Latorre-Pérez et al., (2020) found when comparing Flye, Raven, and Redbean assemblies made from the less accurate reads. Like Sereika et al., (2021) we found accurate genomes could be built from read depths as low as 30x using Flye (Q-score 45 at 30x). This is an improvement from the Q-score of 43.6 at 80x read depth seen by Broddrick et al., (2020). We also know from Sereika et al., (2021) that even higher accuracies can be achieved if a R10.4 flow cell is used instead of a R9.4 flow cell.

Like Breckell and Silander (2021) and Latorre-Pérez et al., (2020), we found Flye and Raven to be better than Redbean in assembling complete genomes. However, unlike Breckell and Silander (2021), but like Latorre-Pérez et al., (2020), we found Flye assembled more accurate assemblies than Raven. The difference may be that Breckell and Silander (2021) looked at assembling single isolates instead of metagenomes, like us and Latorre-Pérez et al., (2020). This suggests that Raven may be better suited for assembling single isolates than metagenomics.

Like Wick and Holt (2019), we found Flye needed more time and memory than Raven and Redbean to complete an assembly. The large time and memory demands of Flye may limit Flye to lab use or at least limit Flye to high end laptops. However, Flye was the only assembler able to assemble the entire mock community at Q-scores over 40. Also, the use of the super-accuracy super-accuracy basecalling model will likely require a higher end laptop with a good GPU. This makes the high time and memory usage of Flye less of an issue.

Summary:

We found Flye was more reliable than Raven or Redbean for building accurate and complete assemblies of both chromosomes and plasmids from metagenomic communities. We found that Raven and Redbean struggle to recover small plasmids. This suggests that Flye would be a better choice for assembling viral community members. For our study's community, Raven and Redbean only performed better than Flye in the amount of computational resources needed to build an assembly. However, for a metagenomic study using the super-accurate basecalling model, the extra time and memory usage needed to run Flye would likely be minimal. On the other hand, the cost in accuracy from problematic community members or missing small plasmid

and virus assemblies from Raven and Redbean could lead to misinterpretations. Thus, for future metagenomic studies that use the super-accurate basecalling model, we recommend using Flye.

References:

Amarasinghe, S. L., S. Su, X. Dong, L. Zappia, M. E. Ritchie, and Q. Gouil. 2020. "Opportunities and Challenges in Long-Read Sequencing Data Analysis." Genome Biology 21 (1): 30. https://doi.org/10.1186/s13059-020-1935-5.

Aniba, M. R., O. Poch, and J. D Thompson. 2010. "Issues in Bioinformatics Benchmarking: The Case Study of Multiple Sequence Alignment." Nucleic Acids Research 38 (21): 7353–63. https://doi.org/10.1093/nar/gkq625.

Antipov, D., M. Rayko, M. Kolmogorov, and P. A. Pevzner. 2022. "viralFlye: Assembling Viruses and Identifying Their Hosts from Long-Read Metagenomics Data." Genome Biology 23 (1): 57. https://doi.org/10.1186/s13059-021-02566-x.

Bai, G.-Hao, S. Lin, Y. Hsu, and S. Chen. 2022. "The Human Virome: Viral Metagenomics, Relations with Human Diseases, and Therapeutic Applications." Viruses 14 (2): 278.

Bokulich, N. A., J. R. Rideout, W. G. Mercurio, A. Shiffer, B. Wolfe, C. F. Maurice, R. J Dutton, P. J. Turnbaugh, R. Knight, and J. G. Caporaso. 2016. "Mockrobiota: A Public Resource for Microbiome Bioinformatics Benchmarking." mSystems 1 (5): e00062-16. https://doi.org/10.1128/mSystems.00062-16.

Breckell, G. L., and O. K. Silander. 2021. "Do You Want to Build a Genome? Benchmarking Hybrid Bacterial Genome Assembly Methods." bioRxiv. n. pag. https://doi.org/10.1101/2021.11.07.467652.

Breitbart, M., E. Delwart, K. Rosario, J. Segalés, and A. Varsani. 2017. "ICTV Virus Taxonomy Profile: Circoviridae." The Journal of General Virology 98 (8): 1997–98.

Broddrick, J. T., R. Szubin, C. J. Norsigian, J. M. Monk, B. O. Palsson, and M. N. Parenteau. 2020. "High-Quality Genome-Scale Models from Error-Prone, Long-Read Assemblies." Frontiers in Microbiology 11: 596626. https://doi.org/10.3389/fmicb.2020.596626.

Chen, Z., D. L. Erickson, and J. Meng. 2020a. "Benchmarking Long-Read Assemblers for Genomic Analyses of Bacterial Pathogens Using Oxford Nanopore Sequencing." International Journal of Molecular Sciences 21 (23): 9161. https://doi.org/10.3390/ijms21239161.

Chen, Z., L. D. Erickson, J. Meng. 2020b. "Benchmarking Hybrid Assembly Approaches for Genomic Analyses of Bacterial Pathogens Using Illumina and Oxford Nanopore Sequencing." BMC Genomics 21 (1): 631. https://doi.org/10.1186/s12864-020-07041-8.

Coster, W. D. , S. D'Hert, D. T. Schultz, M. Cruts, C. V. Broeckhoven. 2018. "NanoPack: visualizing and processing long-read sequencing data." Bioinformatics 34 (15): 2666–69. https://doi.org/10.1093/bioinformatics/bty149.

Garmendía, L., A. Hernández, M. B. Sánchez, and J. L. Martínez. 2012. "Metagenomics and Antibiotics." Clinical Microbiology and Infection: European Society of Clinical Microbiology and Infectious Diseases 18 Suppl 4: 27–31. https://doi.org/10.1111/j.1469-0691.2012.03868.x.

Gaudreault, N. N., D. W. Madden, W. C. Wilson, J. D. Trujillo, and J. A. Richt. 2020. "African Swine Fever Virus: An Emerging DNA Arbovirus." Frontiers in Veterinary Science 7 (May): 215. https://doi.org/10.3389/fvets.2020.00215.

Goldstein, S., L. Beka, J. Graf, and J. L. Klassen. 2018. "Evaluation of Strategies for the Assembly of Diverse Bacterial Genomes Using MinION Long-Read Sequencing." BMC Genomics 20 (1): 23. https://doi.org/10.1186/s12864-018-5381-7.

Jain, M., S. Koren, K. H. Miga, J. Quic, A. C. Rand, T. A. Sasani, J. R. Tyso, et al., 2018. "Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads." Nature Biotechnology 36 (4): 338–45. https://doi.org/10.1038/nbt.4060.

Kolmogorov, M., Jeffrey Y., Yu L., and P. Pevzner. A. 2019. "Assembly of Long Error-Prone Reads Using Repeat Graphs" Nature Biotechnology 37: 540-46. https://doi.org/10.1038/s41587-019-0072-8.

Kolmogorov, M., D. M. Bickhart, B. Behsaz, A. A. Gurevich, M. Rayko, S. B. Shin, K. L. Kuhn, et al., 2020. "metaFlye: Scalable Long-Read Metagenome Assembly Using Repeat Graphs." Nature Methods 17 (11): 1103–10. https://doi.org/10.1038/s41592-020-00971-x.

Kumar, N., S. Sharma, S. Barua, B. N. Tripathi, and B. T. Rouse. 2018. "Virological and Immunological Outcomes of Coinfections." Clinical Microbiology Reviews 31 (4): e00111–17. https://doi.org/10.1128/CMR.00111-17.

Latorre-Pérez, Adriel, P. Villalba-Bermell, J. Pascual, M. Porcar, and C. Vilanova. 2020. "Assembly Methods for Nanopore-Based Metagenomic Sequencing: A Comparative Study." Scientific Reports 10 (1): 13588. https://doi.org/10.1038/s41598-020-70491-3.

Mikheenko, A., V. Saveliev, and A. A. Gurevich. 2016. "MetaQUAST: Evaluation of Metagenome Assemblies." Bioinformatics 32 (7): 1088–90. https://doi.org/10.1093/bioinformatics/btv697.

Payne, A., N. Holmes, V. K. Rakyan, and M. W. Loose. 2019. "BulkVis: A Graphical Viewer for Oxford Nanopore Bulk Fast5 Files." Bioinformatics 35 (13): 2193–98. https://doi.org/10.1093/bioinformatics/bty841.

Petersen, L. M., I. W. Martin, W. E. Moschetti, C. M. Kershaw, and G. J. Tsongalis. 2019. "Third-Generation Sequencing in the Clinical Laboratory: Exploring the Advantages and Challenges of Nanopore Sequencing." Journal of Clinical Microbiology 58 (1): e01315-19. https://doi.org/10.1128/JCM.01315-19.

Qin, S., W. Ruan, H. Yue, C. Tang, K. Zhou, and B. Zhang. 2018. "Viral Communities Associated with Porcine Respiratory Disease Complex in Intensive Commercial Farms in Sichuan Province, China." Scientific Reports 8 (1): 13341. https://doi.org/10.1038/s41598-018-31554-8.

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

Rizzi, R., S. Beretta, M. Patterson, Y. Pirola, M. Previtali, G. D. Vedova, P. Bonizzoni. 2019. "Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era." Quantitative Biology 7(4): 278–92. https://doi.org/10.1007/s40484-019-0181-x

Ruan, Jue, and Heng Li. 2020. "Fast and Accurate Long-Read Assembly with Wtdbg2." Nature Methods 17 (2): 155–58. https://doi.org/10.1038/s41592-019-0669-3.

Ruan, J., and H. Li. 2020. "Fast and accurate long-read assembly with wtdbg2." Nature Methods 17:155-58. https://doi.org/10.1038/s41592-019-0669-3

Sereika, M., R. H. Kirkegaard, S. M. Karst, T. Y. Michaelsen, E. A. Sørensen, R. D. Wollenberg, and M. Albertsen. 2021. "Oxford Nanopore R10.4 Long-Read Sequencing Enables Near-Perfect Bacterial Genomes from Pure Cultures and Metagenomes Without Short-Read or Reference Polishing." bioRxiv n. pag. https://doi.org/10.1101/2021.10.27.466057.

Sevim, V., J. Lee, R. Egan, A. Clum, H. N. Hundley, J. Lee, R. C. Everroad, et al., 2019. "Shotgun Metagenome Data of a Defined Mock Community Using Oxford Nanopore, PacBio and Illumina Technologies." Scientific Data 6 (1): 285. https://doi.org/10.1038/s41597-019-0287-z.

Vaser, R., and M. Šikić. 2021. "Time and Memory-Efficient Genome Assembly with Raven." Nature Computational Science 1 (4): 332–36. https://doi.org/10.1038/s43588-021-00073-4.

Vaser, R., I. Sović, N. Nagarajan, and M. Šikić. 2017. "Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads." Genome Research 27 (5): 737–46. https://doi.org/10.1101/gr.214270.116.

Wick, R. R., and K. E. Holt. 2019. "Benchmarking of Long-Read Assemblers for Prokaryote Whole Genome Sequencing." F1000Research 8: 2138. https://doi.org/10.12688/f1000research.21782.4.

Wick, R. R., Judd L. M., Cerdeira L. T., Hawkey J., Méric G., Vezina B., Wyres K. L., Holt K.E. 2021. "Trycycler: consensus long-read assemblies for bacterial genomes." Genome Biology 22 (1): 266. https://doi.org/10.1186/s13059-021-02483-z.

Wickham, H. 2016. "ggplot2: Elegant Graphics for Data Analysis." Springer-Verlag New York 17: 160-67.

Yang, C., D. Chowdhury, Z. Zhang, W. K. Cheung, A. Lu, Z. Bian, and L. Zhang. 2021. "A Review of Computational Tools for Generating Metagenome-Assembled Genomes from Metagenomic Sequencing Data." Computational and Structural Biotechnology Journal 19: 6301–14. https://doi.org/10.1016/j.csbj.2021.11.028.

General Conclusions:

Genomic epidemiology offers a way to detect pathogens of concern before epidemics begin and to track pathogens during epidemics (Gardy 2018). By tracking pathogens and identifying transmission routes, countries can prepare for future epidemics by reducing points of pathogen transmission or by making vaccines (Giovanetti et al., 2021; Hay and McCauley 2018; Gardy 2018). However, emerging zoonotic pathogens that are transmitted from animals to humans (zoonotic spillover), such as Ebola, Zika, and SARS-CoV-2, have origins from regions of high biodiversity, like parts of Africa, Latin America, and Asia (Gardy 2018; Jones et al., 2008). Some of the countries in these regions do not have the resources to invest in large-scale use of costly sequencing technology (Gardy 2018; Jones et al., 2008). As a result, there is less surveillance in countries with higher risks of zoonotic spillover, than in countries, like the United States, that have a relatively lower zoonotic spill over risk, but more resources for surveillance (Jones et al., 2008).

This lack of surveillance may cause emerging zoonotic pathogens to not be detected until they become epidemics, like with Ebola and Zika, resulting in an increased cost of stopping viral transmission (Gardy 2018; Dobson et al., 2020). An example is the SARS-CoV-2 pandemic, which has cost the world over five trillion dollars (Dobson et al., 2020). This multi-trillion-dollar cost and the cost of other epidemics, such as Ebola, may have been prevented by a 26-billion-dollar yearly investments in surveillance and other preventative measures (Dobson et al., 2020).

Low investment, portable sequencers, like nanopore sequencing, offer a solution for low-resource countries to do their own surveillance (Sereika et al., 2021; Gardy 2018). The portability allows nanopore sequencers to be used on site, allowing for rapid testing and

sequencing in regions with minimal laboratory resources (Gardy 2018). The low investment may allow low-resource countries and labs to afford nanopore sequencers, which may increase surveillance (Gardy 2018). Increasing surveillance in low-resource countries may increase the genomic surveillance of locations that are high risk for zoonotic spillovers. This may allow for genomic detection of emerging pathogens before they become an epidemic or pandemic. Early detection may allow countries with more resources to respond to pathogens in the country of origin, before they cause an epidemic or pandemic and are transmitted to other countries (Dobson et al., 2020; Gardy 2018). This will reduce the cost of emerging pathogens to the original country, the responding country, and the global society.

In chapter 1, I showed that nanopore sequencing can be used to understand the diversity and transmission of porcine circovirus type 2 (PCV2) in Ukraine. The discovery of a PCV2 genotype (f) not previously found in Ukraine shows that PCV2 diversity in Ukraine was and still is likely underestimated. I also found that nanopore sequencers could detect potential co-infections involving multiple PCV2 genotypes. Finally, I found identical sequences in non-neighboring oblasts. This suggests that there may be PCV2 transmission between oblasts or that PCV2 may be transmitted into multiple oblasts from the same out of country source. However, none of our identical sequences were obvious transmission chains, suggesting that PCV2 transmission is poorly understood in Ukraine and in the countries neighboring Ukraine. The underestimation in diversity and PCV2 transmission shows that more PCV2 surveillance is needed in Ukraine and neighboring countries.

Rare genotypes, like PCV2 genotype f, can easily be missed in surveillance studies. However, when rare genotypes are missed, we miss a potential source of diversity that can give rise to new variants. These rare variants can recombine in co-infections with more common

genotypes or other rare variants to produce recombinant variants more fit than the parent variants (Franzo and Segalés 2020; Amoutzias et al., 2022). In the SARS-CoV-2 pandemic, the delta-omicron variant is an example of a recombinant variant that is a more fit variant than its parent variants (Amoutzias et al., 2022; Colson et al., 2022). Also, it is possible that rare variants may become common variants under the right conditions. An example of this is PCV2 genotype d, which became a common genotype after global PCV2 vaccination (Franzo and Segalés 2018; Franzo and Segalés 2020). Future PCV2 surveillance studies should use larger samples sizes to detect rare and recombinant variants.

In chapter 1, I also showed that nanopore sequencing can detect co-infections between PCV2 genotypes and possibly even within genotypes. These methods can be extended beyond PCV2 to other viruses, so long as there is at least a two to five percent difference between variants. My results show that nanopore sequencers are reliable in detecting co-infections between viral variants.

My pipeline is limited by the number of references input and the ability of minimap2 to accurately map reads to a set of references. Leigh et al., 2020 has shown that 5 kb reads could be correctly mapped to a set of references that were 98% similar. However, when the references were 99% similar, blastn could not reliably map 3 kb reads to the correct reference. This evidence shows that the read mapping step is likely a limitation in my pipeline. Especially since my PCV2 reference database could be easily expanded with many more sequences from GenBank. This limitation in read accuracy has been improved since my first thesis chapter.

My first thesis chapter did not take advantage of the recent advances in nanopore sequencing that have improved read accuracy (https://nanoporetech.com/accuracy). Also, my

second chapter did not take advantage of new updates to chemistry that have further increased read accuracy (https://nanoporetech.com/accuracy). It is likely that nanopore sequencers, with the recent updates, could detect more similar co-infections than I detected in my study.

Metagenomics offers a surveillance method that enables the detection of novel pathogens or pathogens that are often ignored in surveillance studies (Bai et al., 2022; Chiu and Miller 2019). Long read metagenomics build more complete genomes than Illumina sequencing, while having a lower cost than other sequencing methods, such as Illumina sequencing and PacBio sequencing (Sereika et al., 2021). However, the high error rate of nanopore sequencers is a limitation (Petersen et al., 2019).

In chapter 2, I looked at the current completeness and accuracy of long read metagenomic assemblies for a synthetic bacterial community sequenced using nanopore sequencing. I found that chromosome metagenomic assemblies from Flye were complete and had Q-scores between 40 (99.99% accurate) and 50 (99.999% accurate). I also found that Flye was the only assembler that could assemble small plasmids reliably. Plasmids are about the size of viruses, thus Flye is the only reliable assembler for viruses. This evidence shows that Flye is the best assembler for bacterial communities and the only reliable assembler for viral communities.

The completeness of long read metagenomic assemblies combined with the low cost suggests that long read metagenomics is a good method for detecting correlations between mixed infections or co-infections. Allowing nanopore sequencers to be used in finding correlations between porcine circovirus associated diseases (PCVAD) and other pathogens. One potential virus that has co-infected with PCV2 in the past is African swine fever virus (ASFV) (Dundon et al., 2022).

ASFV, like Ebola, causes hemorrhagic fever in domestic pigs and has a high lethality (Cisek et al., 2016; Goeijenbier et al., 2014). Both ASFV and PCV2 infect macrophage cells (Franzoni et al., 2019). However, PCV2 cannot replicate in the monocyte-derived dendritic cells (moDC) cells ASFV infects, so direct competition between PCV2 and ASFV is unlikely (Franzoni et al., 2019). Though, PCV2 infections do prevent moDC cell maturation and cause some immunosuppression, suggesting that there is some indirect interaction between PCV2 and ASFV. (Franzoni et al., 2019). Also, PCV2 infections have been shown to reduce the vaccine efficacy for the bone marrow derived-blood dendritic cells infecting classical swine fever virus (Franzoni et al., 2019; Ouyang et al., 2019). Thus, it is possible PCV2 infections may reduce the efficacy of future ASFV vaccines.

One additional area long read metagenomics may be useful for is detecting novel, not yet sequenced pathogens. The genomes of novel pathogens are unknown, thus sequencing requires metagenomics, culturing, or PCR targeting conserved regions, such as the bacterial 16S rRNA in pathogens (Bharti and Grimm 2021; Garmendía et al., 2012; Chiu and Miller 2019). However, not all microbes can be cultured and PCR targeting of conserved regions does not give full genomes (Garmendía et al., 2012; Bharti and Grimm 2021). Metagenomics gives full genomes and allows the scientist to detect pathogens that cannot be cultured (Bai et al., 2022). One example of using metagenomics to discover disease causing pathogens is SARS-CoV-2, which was identified as a corona virus with PCR and sequenced using metagenomics (Zhou et al., 2020). This shows that metagenomics is a powerful for viral discovery and surveillance. My thesis shows that using nanopore sequencers can improve surveillance for specific pathogens and through metagenomics pathogens that would be hard to capture.

Nanopore sequencers offer a portable, low investment tool for surveillance and detecting pathogen combinations that increase disease severity, allowing for countries to do both their own surveillance with genotyping and with long read metagenomics. This warning may allow us to detect zoonotic spillovers early, which may allow us to prevent or reduce the severity of future epidemics. Also, as surveillance increases, we may get a better understanding of how pathogens are transmitting through human and animal populations. This may allow us to develop and apply one health preventive measures that may reduce zoonotic spillovers in high-risk regions. Long read metagenomics also provides a way to detect correlations between mixed infections or co-infections that increase disease severity of pathogens. In this thesis, I have shown that long read metagenomic assemblies built using Flye could detect viral and bacterial pathogens. This can then be extended by others to find correlations between mixed infecting or co-infecting pathogens in samples. With the methods that I have described, a co-infection that would be interesting and important for future exploration would be ASFV and PCV2.

References:

Amoutzias, G. D., M. Nikolaidis, E. Tryfonopoulou, K. Chlichlia, P. Markoulatos, and S. G. Oliver. 2022. "The Remarkable Evolutionary Plasticity of Coronaviruses by Mutation and Recombination: Insights for the COVID-19 Pandemic and the Future Evolutionary Paths of SARS-CoV-2." Viruses 14 (1): 78. https://doi.org/10.3390/v14010078.

Bai, G., S. Lin, Y. Hsu, and S. Chen. 2022. "The Human Virome: Viral Metagenomics, Relations with Human Diseases, and Therapeutic Applications." Viruses 14 (2): 278.

Bharti, R., and D. G. Grimm. 2021. "Current Challenges and Best-Practice Protocols for Microbiome Analysis." Briefings in Bioinformatics 22 (1): 178–93. https://doi.org/10.1093/bib/bbz155.

Chiu, C. Y., and S. A. Miller. 2019. "Clinical Metagenomics." Nature Reviews. Genetics 20 (6): 341–55. https://doi.org/10.1038/s41576-019-0113-7.

Cisek, A. A., I. Dąbrowska, K. P. Gregorczyk, and Z. Wyżewski. 2016. "African Swine Fever Virus: A New Old Enemy of Europe." Annals of Parasitology 62 (3): 161–67. https://doi.org/10.17420/ap6203.49.

Colson, P., P. Fournier, J. Delerce, M. Million, M. Bedotto, L. Houhamdi, N. Yahi, et al., 2022. "Culture and Identification of a "Deltamicron" SARS-CoV-2 in a Three Cases Cluster in Southern France." Journal of Medical Virology 94 (8): 3739- 49. https://doi.org/10.1002/jmv.27789.

Dobson, A. P., S. L. Pimm, L. Hannah, L. Kaufman, J. A. Ahumada, A. W. Ando, A. Bernstein, et al., 2020. "Ecology and Economics for Pandemic Prevention." Science (New York, N.Y.) 369 (6502): 379–81. https://doi.org/10.1126/science.abc3189.

Dundon, W. G., G. Franzo, T. B. K. Settypalli, N. L. P. I. Dharmayanti, U. Ankhanbaatar, I. Sendow, A. Ratnawati, et al., 2022. "Evidence of Coinfection of Pigs with African Swine Fever Virus and Porcine Circovirus 2." Archives of Virology 167 (1): 207–11. https://doi.org/10.1007/s00705-021-05312-7.

Franzo, G., and J. Segalés. 2020. "Porcine Circovirus 2 Genotypes, Immunity and Vaccines: Multiple Genotypes but One Single Serotype." Pathogens (Basel, Switzerland) 9 (12): 1049. https://doi.org/10.3390/pathogens9121049.

Franzo, G., and J. Segalés. 2018. "Porcine Circovirus 2 (PCV-2) Genotype Update and Proposal of a New Genotyping Methodology." PloS One 13 (12): e0208585. https://doi.org/10.1371/journal.pone.0208585

Franzoni, G., S. P. Graham, S. D. Giudici, and A. Oggiano. 2019. "Porcine Dendritic Cells and Viruses: An Update." Viruses 11 (5): 445. https://doi.org/10.3390/v11050445.

Gardy, J. L., and N. J. Loman. 2018. "Towards a Genomics-Informed, Real-Time, Global Pathogen Surveillance System." Nature Reviews. Genetics 19 (1): 9–20. https://doi.org/10.1038/nrg.2017.88.

Garmendía, L., A. Hernández, M. Blanca Sánchez, and J. L. Martínez. 2012. "Metagenomics and Antibiotics." Clinical Microbiology and Infection: European Society of Clinical Microbiology and Infectious Diseases 18 Suppl 4: 27–31. https://doi.org/10.1111/j.1469-0691.2012.03868.x.

Giovanetti, M., F. Benedetti, G. Campisi, A. Ciccozzi, S. Fabris, G. Ceccarelli, V. Tambone, et al., 2021. "Evolution Patterns of SARS-CoV-2: Snapshot on Its Genome Variants." Biochemical and Biophysical Research Communications 538 (January): 88–91. https://doi.org/10.1016/j.bbrc.2020.10.102.

Goeijenbier, M, J. J. Kampen, C. B. Reusken, M. P. Koopmans, and E. C. Gorp. 2014. "Ebola Virus Disease: A Review on Epidemiology, Symptoms, Treatment and Pathogenesis." The Netherlands Journal of Medicine 72 (9): 442–48.

Hay, A. J., and J. W. McCauley. 2018. "The WHO Global Influenza Surveillance and Response System (GISRS)-a Future Perspective." Influenza and Other Respiratory Viruses 12 (5): 551–57. https://doi.org/10.1111/irv.12565.

Jones, K. E., N. G. Patel, M. A. Levy, A. Storeygard, D. Balk, J. L. Gittleman, and P. Daszak. 2008. "Global Trends in Emerging Infectious Diseases." Nature 451 (7181): 990–93. https://doi.org/10.1038/nature06536.

Leigh D. M., C. Schefer, C. Cornejo. 2020. "Determining the Suitability of MinION's Direct RNA and DNA Amplicon Sequencing for Viral Subtype Identification." Viruses 12(8):801. https://doi:org/10.3390/v12080801.

Ouyang, T., X. Zhang, X. Liu, and L. Ren. 2019. "Co-Infection of Swine with Porcine Circovirus Type 2 and Other Swine Viruses." Viruses 11 (2): 185. https://doi.org/10.3390/v11020185.

Petersen, L. M., I. W. Martin, W. E. Moschetti, C. M. Kershaw, and G. J. Tsongalis. 2019. "Third-Generation Sequencing in the Clinical Laboratory: Exploring the Advantages and Challenges of Nanopore Sequencing." Journal of Clinical Microbiology 58 (1): e01315-19. https://doi.org/10.1128/JCM.01315-19.

Sereika, M., R. H. Kirkegaard, S. M. Karst, T. Y. Michaelsen, E. A. Sørensen, R. D. Wollenberg, and M. Albertsen. 2021. "Oxford Nanopore R10.4 Long-Read Sequencing Enables Near-Perfect Bacterial Genomes from Pure Cultures and Metagenomes Without Short-Read or Reference Polishing." bioRxiv n. pag. https://doi.org/10.1101/2021.10.27.466057.

Zhou, P., X. Yang, X. Wang, B. Hu, L. Zhang, W. Zhang, H. Si, et al., 2020. "A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin." Nature 579 (7798): 270–73. https://doi.org/10.1038/s41586-020-2012-7.

Appendix

Chapter 1: Supplementary methods:

Read simulation:

We tested the accuracy of my co-infection pipeline mentioned in my first chapter using simulated reads, made from pairs of references in our database. Potential reference pairs were found by blasting our original database against itself with blastn (NCBI Resource Coordinators 2016) (See Figure A-1 for a flowchart). Reference pairs (hits with a query and subject) that had a percent identity score under 98% or greater than 98.5% were removed. For each query, I only kept the hit closest to 98 percent identity (n = 237).

I simulated three sets of reads for each reference pair using badread v0.2 (Wick 2019) with parameters –identity "90,97.5,5" and –quantity 20000x. The first set of simulated reads had 50% of reads coming from the minor variant (query). The second set of simulated reads had 5% of reads coming from the minor variant. Finally, the third simulated set of reads had 1% of reads coming from the minor variant.

I detected Co-infections using the steps mentioned in chapter 1's methods, except we used mapping qualities of 20 and 30, required at least 1% difference between consensuses, did not check for differences in mismatches, and required each bin to have at least 0.4% of all mapped reads. For each consensus, I found the number of mismatches and indels by blasting the consensus genome against the reference pair used to simulate its reads. I made graphs showing the accuracy of the built consensus and ability of my pipeline to detect co-infections with ggplot2 (Wickham 2016)

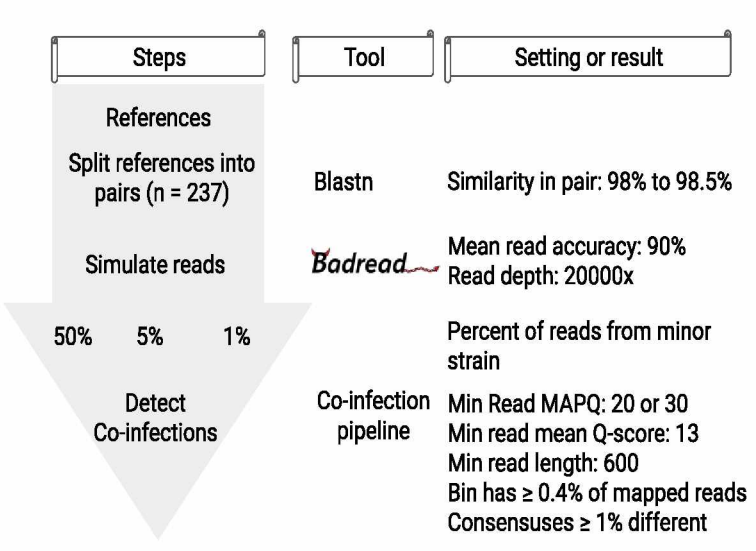

Figure 0-1: Co-infection pipeline testing methods.

Chapter 1: Supplementary figures and tables:

Table 0-1: Co-infections detected by Medaka_variant. Variant calling with medaka_variant was done on reads that mapped to a columns reference genome. Columns show the reference genotype used with Medaka_variant. Rows show the genotypes detected. Co-infections are shwon by genotype-1/genotype-2. Genotypes d1 and d2 distinguish between the two-gentoype d's in Franzo and Segalés (2018). Genotype d2 may actually be genotype e. Genotypes were found by building a maximum likelihood tree using RAxML with 1000 bootstraps and references from Franzo and Segalés (2018). Samples not shown had only one genotype detected.

Sample	genotype a	genotype b	genotype c	genotype d1	genotype d2	genotype f	genotype g	genotype h
Kharkiv 4	d1	b/d1	d1	d1	b/d1	d1	d1	d1
Kharkiv 5	b/d1	b	b/d1	d1	b/d1	d1	d1	d1
Kharkiv 6	d1	b	d1	b/d1	b/d1	d1	b/d1	d1
Chernivtsi 1	a/b	b	a/b	a/b	b	b	a/b	a/b
Chernivtsi 2	a/b	b	a/b	a/b	b	b	a/b	a/b

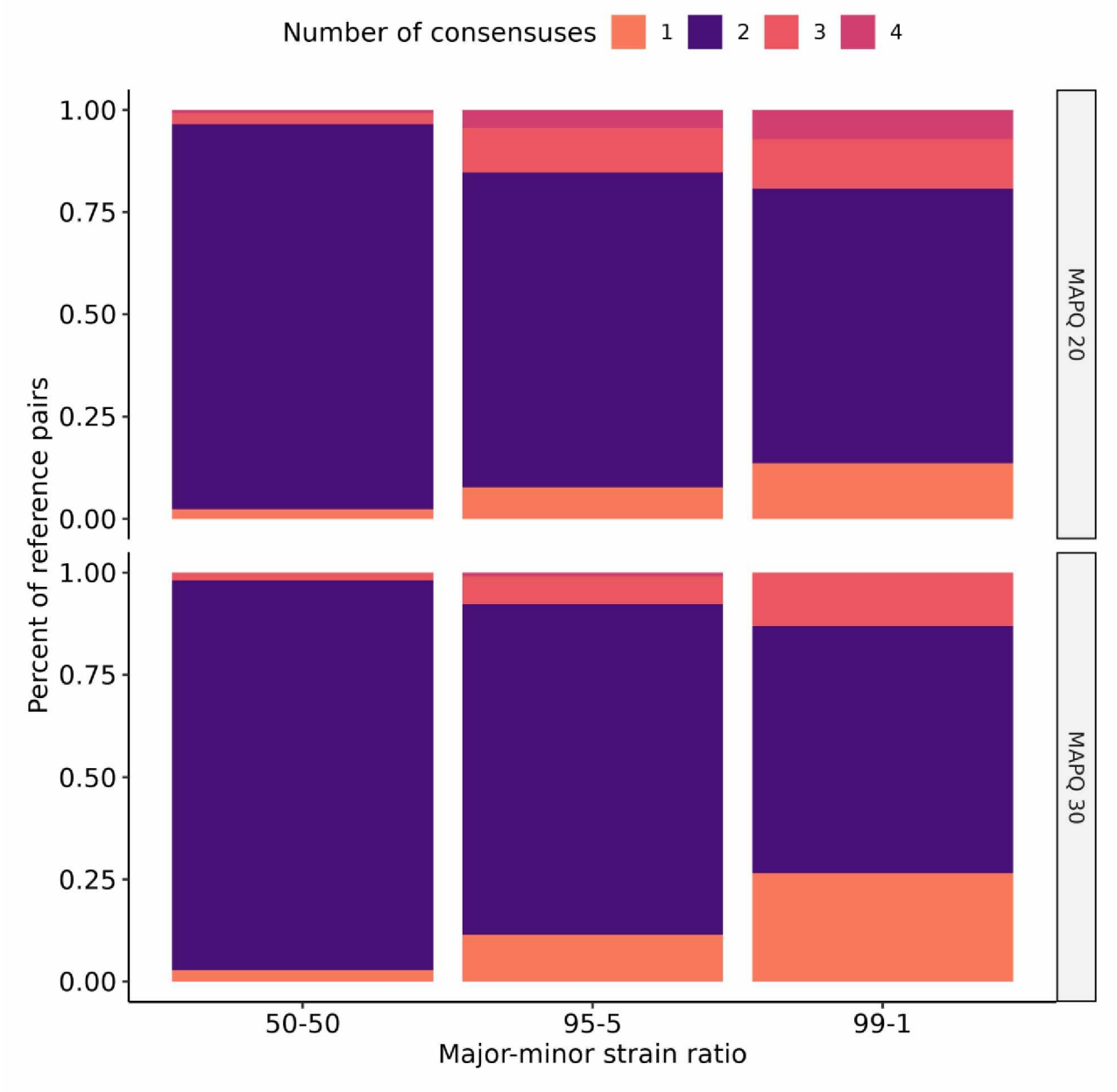

Figure 0-2: Percent of correctly detected co-infections. One consensus indicates a missed co-infection. More than two consensuses indicate extra co-infections (noise). Each major-minor strain ratio has sample size of 237 reference pairs.

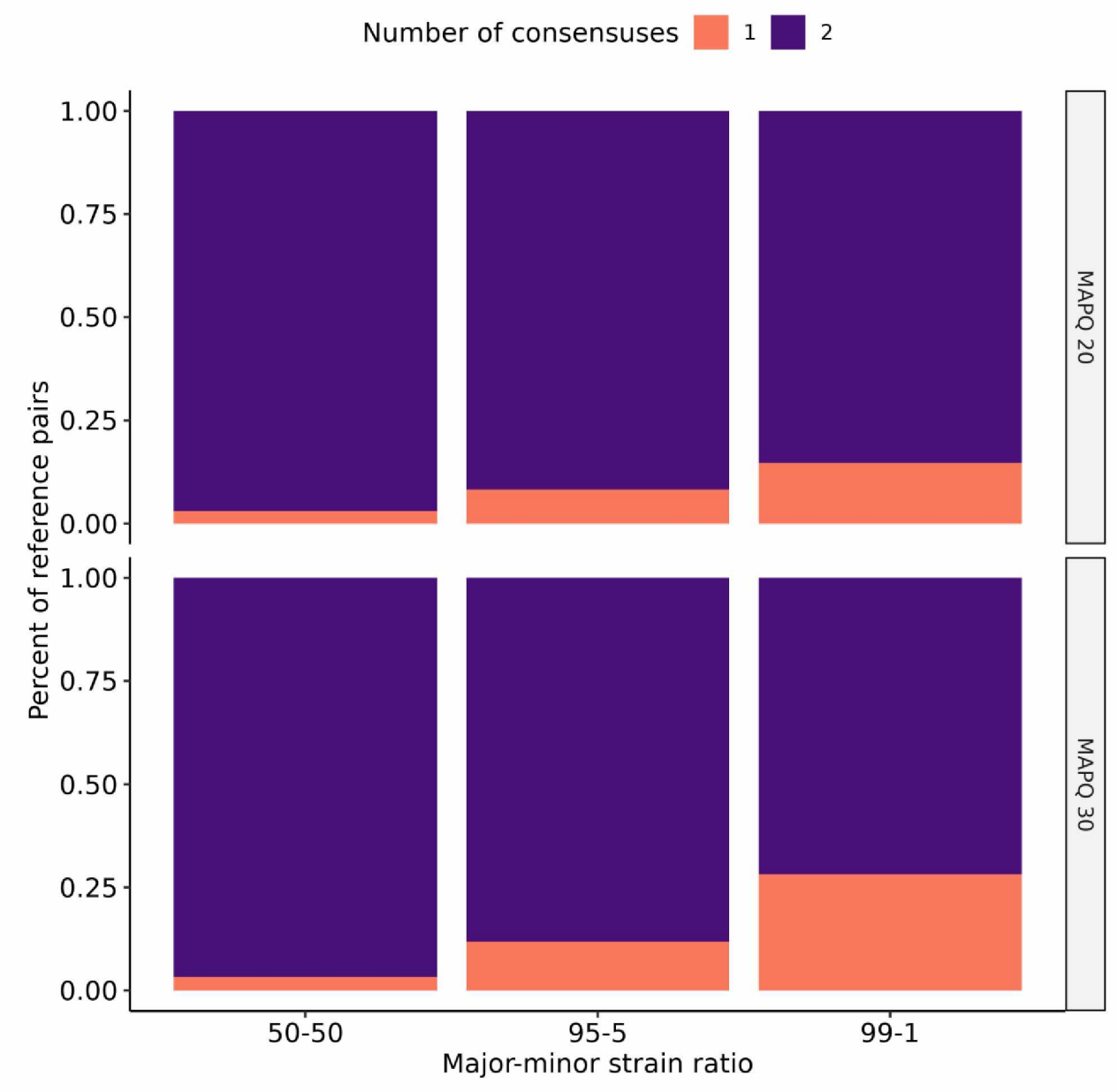

Figure 0-3: Extra filters removes co-infection noise. Extra filters: consensus genomes have at least 100 reads and 0.3% mismatches. One consensus indicates a missed co-infection. Each major-minor strain ratio has sample size of 237 reference pairs.

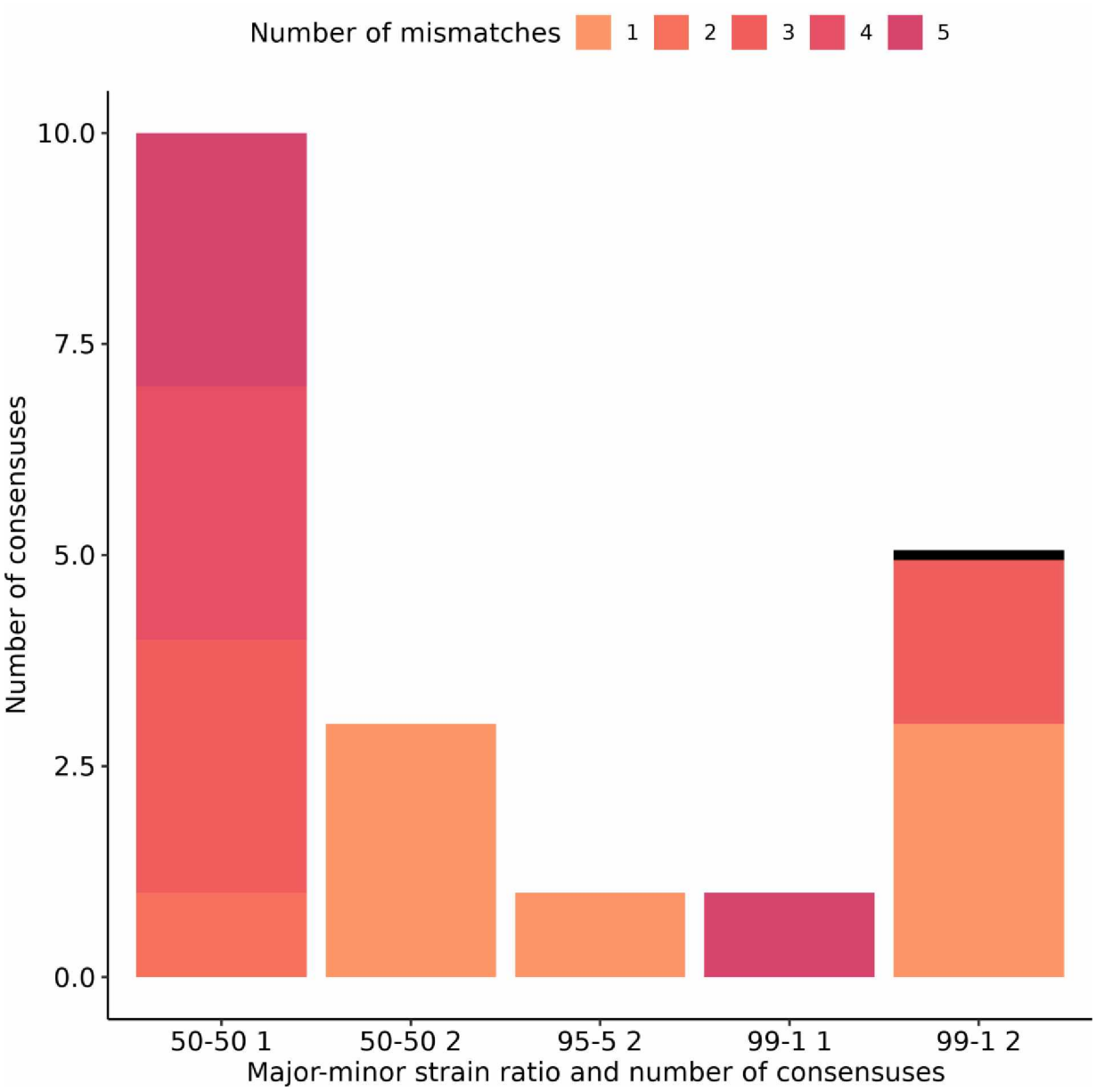

Figure 0-4: Number of mismatches in co-infection consensuses. Black bars indicate the number of minor strains with mismatches for the 95-5 and 99-1 major-minor rations. Consensus with 0 mismatches were removed. Sample size was 237 reference pairs (roughly 474 consensus) for each major-minor strain before filtering.

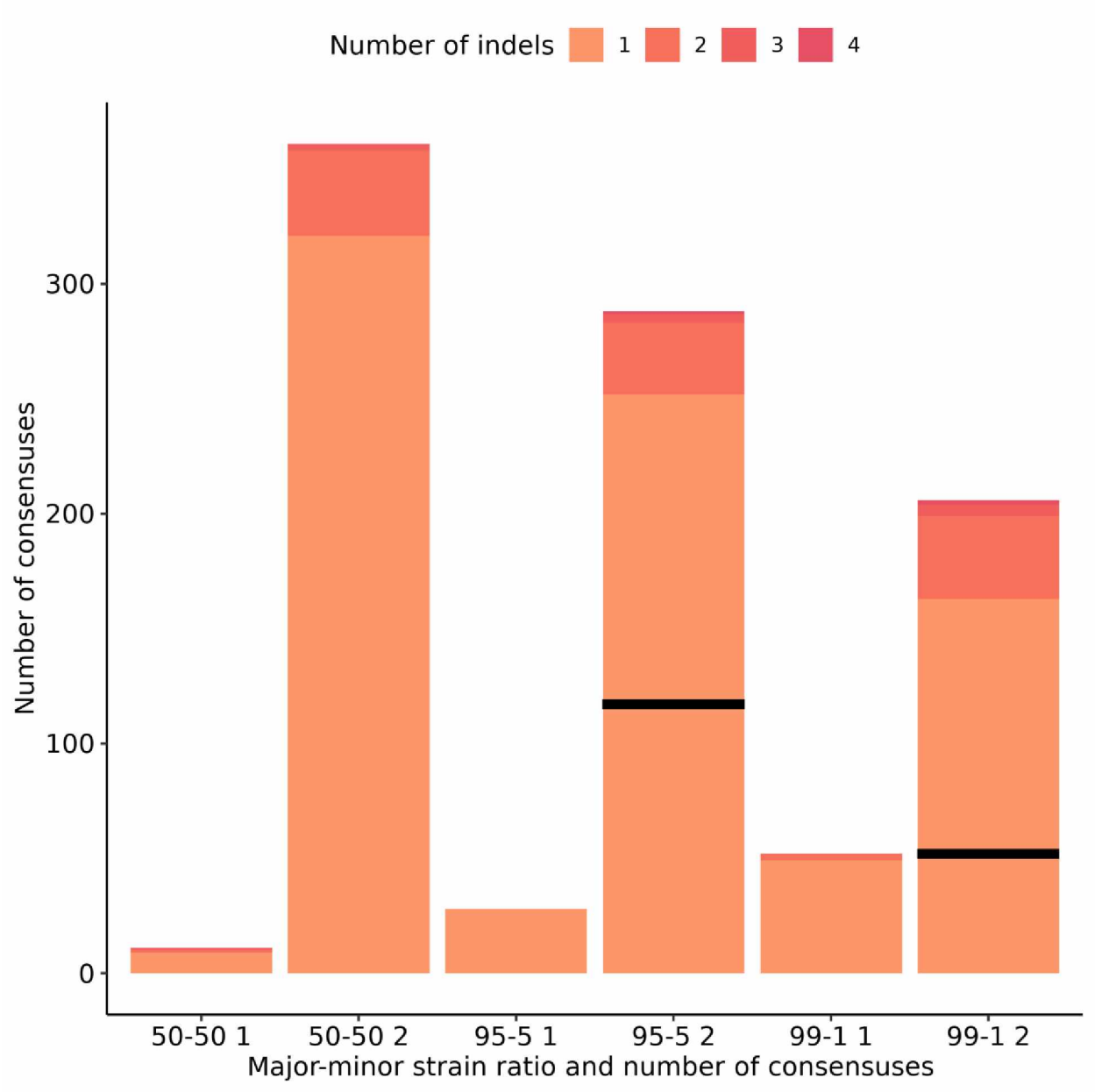

Figure 0-5: Number of indels in co-infection consensuses. Black bars indicate the number of minor strains with mismatches for the 95-5 Consensus with 0 indels were removed. Sample size was 237 reference pairs (roughly 474 consensus) for each major-minor strain before filtering.

Table 0-2: Sequences after recombinant removal

LC008137	KT868494	KX169308	KP081543	KP245920	HM003569	JF690916
LC008135	KC261601	MG182436	KJ139962	KC514991	HM003570	JF690918
LC004750	HQ591377	KU697267	KP081542	KC515014	GU370064	JF690919
KY806003	JN006457	KU697268	KP081556	KM190990	GU083582	FJ905459
KX828216	FJ804417	KU697227	KY947569	KJ679446	KP231108	FJ905463
KP768478	AB512130	KU317492	KM116513	KP231118	KY806028	FJ905468
KP420197	AB512139	KU697198	KM360053	JX945577	MF314200	AB361574
KC515014	LC310740	KU697201	KP231128	JX945576	KX161671	HM038027
HQ202949	MF737383	MH341484	KY940520	KP231116	KT867949	MH465433
HM038034	KY947570	KU317473	KM191084	JX274295	KJ679445	KX161686
EU450638	MG813259	MF169728	KM191108	KM042406	HQ231328	KT867856
MN482046	KY947565	KU697031	KP081541	KF850465	JF317565	KT867870
MN482054	KX814351	KU697049	KM191057	JN382159	JF317570	KT867898
MN196673	MG807481	KU697071	KM191071	JN382164	JF317573	KT867953
MK552324	MG813260	MT376375	KP081538	JN382175	HQ591379	KT867958
MN258752	KY656102	MF589532	KM191020	KP420191	GQ995584	KT867979
MN170517	KX510057	KY806064	MN935179	KP420193	GU325759	KT868003
MN482037	KX510064	KX928994	MN935183	JN382187	GU325766	KT868018
MT104513	KY656098	LC310732	MH465425	JN382189	KP231103	KT868056
MK305880	KX098776	KY656019	KY440166	KP420192	KP231110	KT868067
MK305883	KY655972	KY656009	KY388468	LC383443	HM038032	KT868095
MN735211	KX098762	KY656045	KU193766	KY806032	KP231102	KT868129
MK006039	KX098742	KP081551	KT867860	KY806033	KP231129	KT868142
MH920582	MK604508	KP081552	KT867867	KR559710	GQ404807	KT868178
MK504382	MK347352	KR058355	KT867888	KR559713	KF742546	KT868187
MK504383	MH465430	KP081550	KT867914	KR559714	HM038025	KT868200
MK604480	MH465431	KU311027	KT867931	KR559722	MH465402	KT868222
MK426838	MH351271	KP081549	KT867943	KC514972	MH465404	KT868234
MK347380	MG732823	KU697102	KT867952	KC514979	MH465420	KT868258
MH509735	MG893892	KP081546	KJ511872	JQ181595	KY806012	KT868270
MH509736	MG807610	KP081547	KF524259	JQ002672	KY806016	KT868271
MT104514	MH055402	KY655985	KC751546	KP231145	EU521707	KT868272
MT423827	MF631809	KP081548	KF035059	KX904946	HQ831526	KT868284
MK006037	KY810321	LC383446	KM191009	JF683406	JF683403	KT868296
MT376339	LC310734	MH465410	KM191004	HQ378162	GQ404799	KT868316
MK140461	KX855983	MH465438	KP768483	JQ866919	EU296794	KT868325
MH509732	KX981602	MF314288	KJ680360	KJ094600	AB361585	KT868441
MH509733	MF737379	KT819159	KP231164	KJ094603	FJ233908	KT868480
MG732802	KX169322	KT819160	KM924366	KJ094606	KM624030	KT868519
MF589528	KX169329	KT819161	KP420200	KC620511	KT868313	KC620544
KT868448	KX098689	KT819163	MF314229	JF317584	KT868321	KC620552
KT868482	KY940535	KX298474	KT867951	JN006464	KT868365	KT868055
KT868491	KX169298	KT868419	KT867972	GQ358997	KT868432	KT868058
KT868250	MK005836	EU450630	KT867997	KT868038	KT868454	KT868374
KT868254	MK005837	EF990646	KT868390	KT868079	KT868500	KC620541
KT868357	LC278328	EF452364	KT868401	KT868128	KT868509	FJ905464
KT868360	LC278333	EF524518	KT868440	KT868158	JF690911	FJ905466
KT868437	LC278346	EF524523	KT216672	KT868289	JF690912	FJ644556

Table 0-3 Continued: Sequences after recombinant removal

FJ644559	HQ202963	DQ629129
FJ644562	HQ202964	DQ648031
EU755372	HQ202966	DQ195679
EU755373	GU244506	DQ104420
EU755376	HQ395026	AY864814
EU755377	FN398026	AY484411
EU755381	EU747085	AY682993
KC620532	AB462385	AY556477
KC620537	AB462391	AY217743
KC620536	EU257515	AF465211
JF927976	EU057186	AF201311
JF927979	EU057188	
JF927978	AB426905	
JF683387	EU450585	
JF927977	EU450593	
JN133304	EU450595	
FJ644919	EU450597	
MF139077	EU450601	
JX512856	EU450613	
JX512855	EU450623	
MT769305	EU450627	
MK005838	EF565349	
MK005843	EF565351	
MK005848	EF565367	
MK005850	EU136711	
MK005854	EF619037	
KX641126	EF592575	
KX641138	EF560608	
LC278327	EF452360	
LC278348	EF524524	
MF589543	EF524528	
MF616428	EF524538	
MF142266	EF190926	
KR868575	EF190927	
HQ202944	EF064149	
HQ202947	EF067852	
HQ202948	DQ856564	
HQ202949	DQ856569	
HQ202950	DQ629114	
HQ202951	DQ629123	
HQ202957	DQ629127	

Chapter 1 Supplementary References:

Franzo, G., and J. Segalés. 2018. "Porcine Circovirus 2 (PCV-2) Genotype Update and Proposal of a New Genotyping Methodology." PloS One 13 (12): e0208585.

NCBI Resource Coordinators. 2016. "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res 44(D1): D7-19. https://doi.org/10.1093/nar/gkv1290.

Wick, R. R. 2019. "Badread: Simulation of Error-Prone Long Reads." Journal of Open Source Software 4 (36): 1316. https://doi.org/10.21105/joss.01316.

Wickham, H. 2016. "ggplot2: Elegant Graphics for Data Analysis." Springer-Verlag New York 17:160-167.

Chapter 2 Supplementary figures:

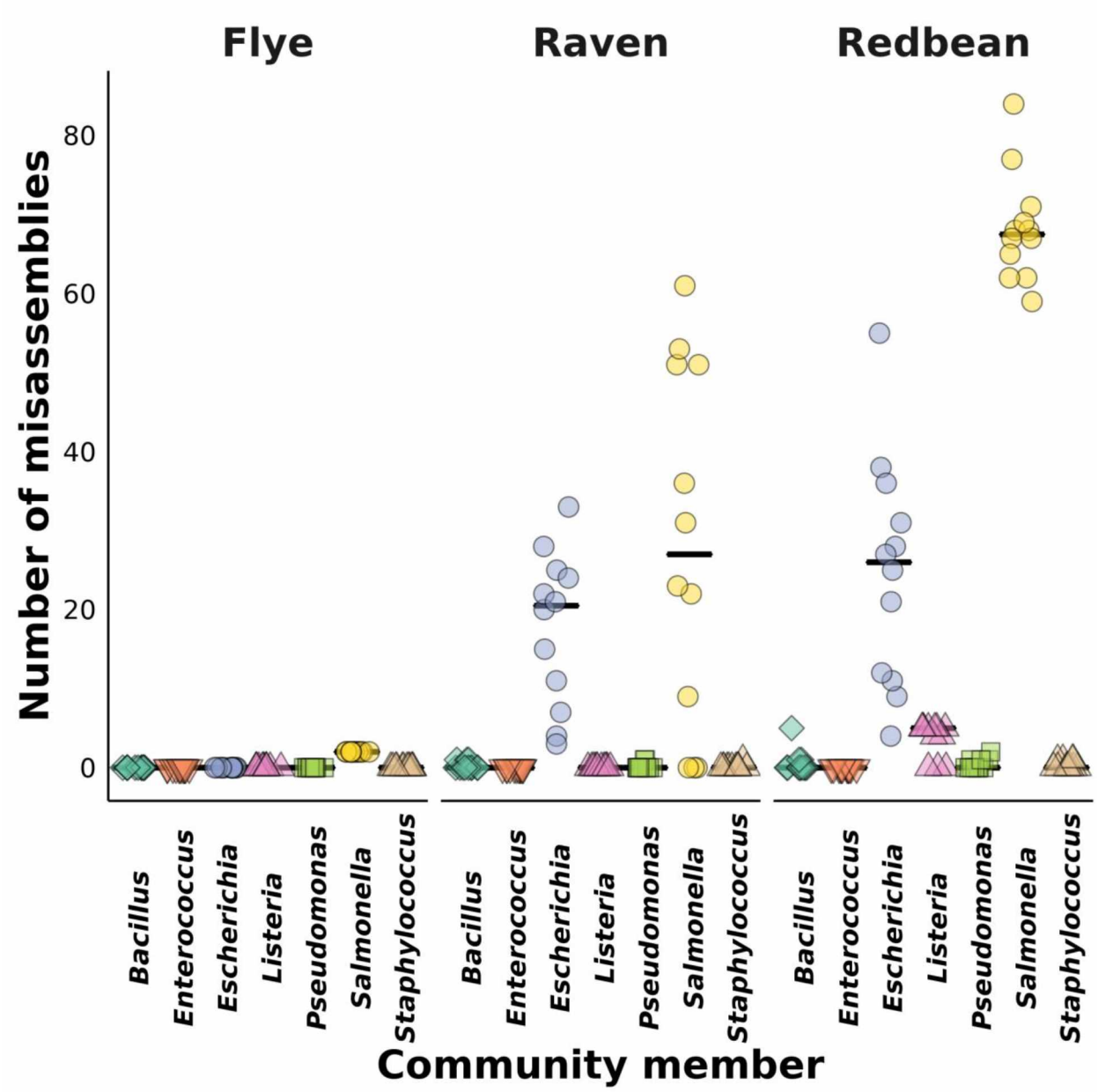

Figure 0-6: Chromosome mis-assemblies at 200x read depth. Horizontal bars indicate the median value across replicate samples.

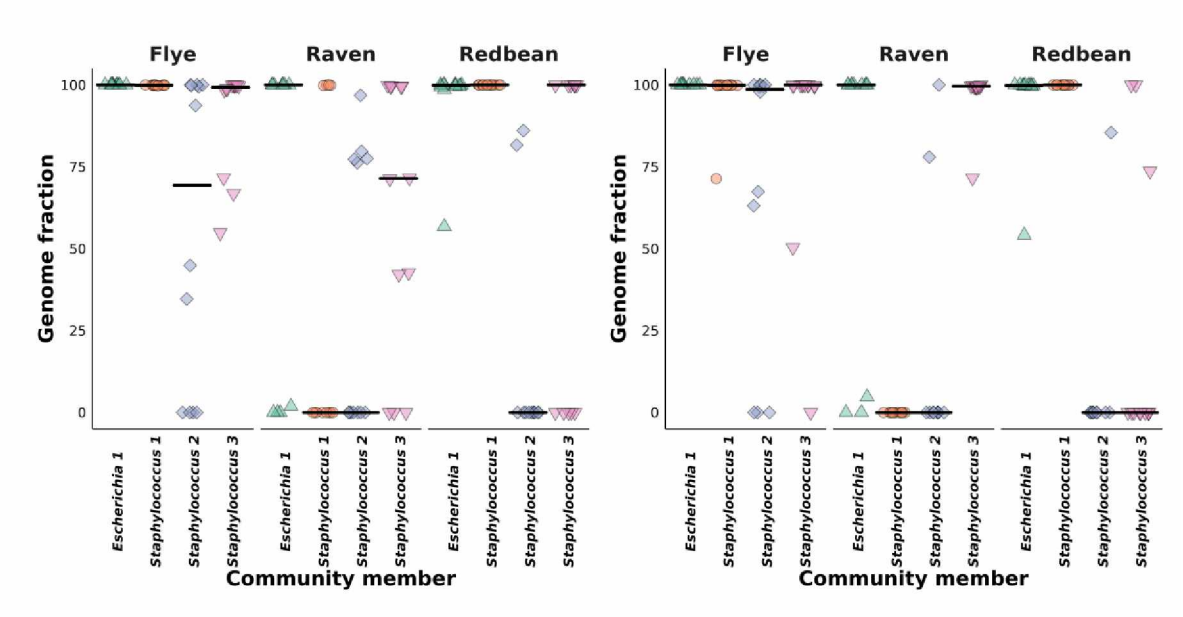

Figure 0-7: 30x and 50x plasmid completeness. Horizontal bars indicate the median across replicates. E. coli is 110009 bases, S. aureus 1 is 6339 bases, S. aureus 2 is 2218 bases, and S. aureus 3 is 2995 bases long. a — 30x read depth., b — 50x read depth.

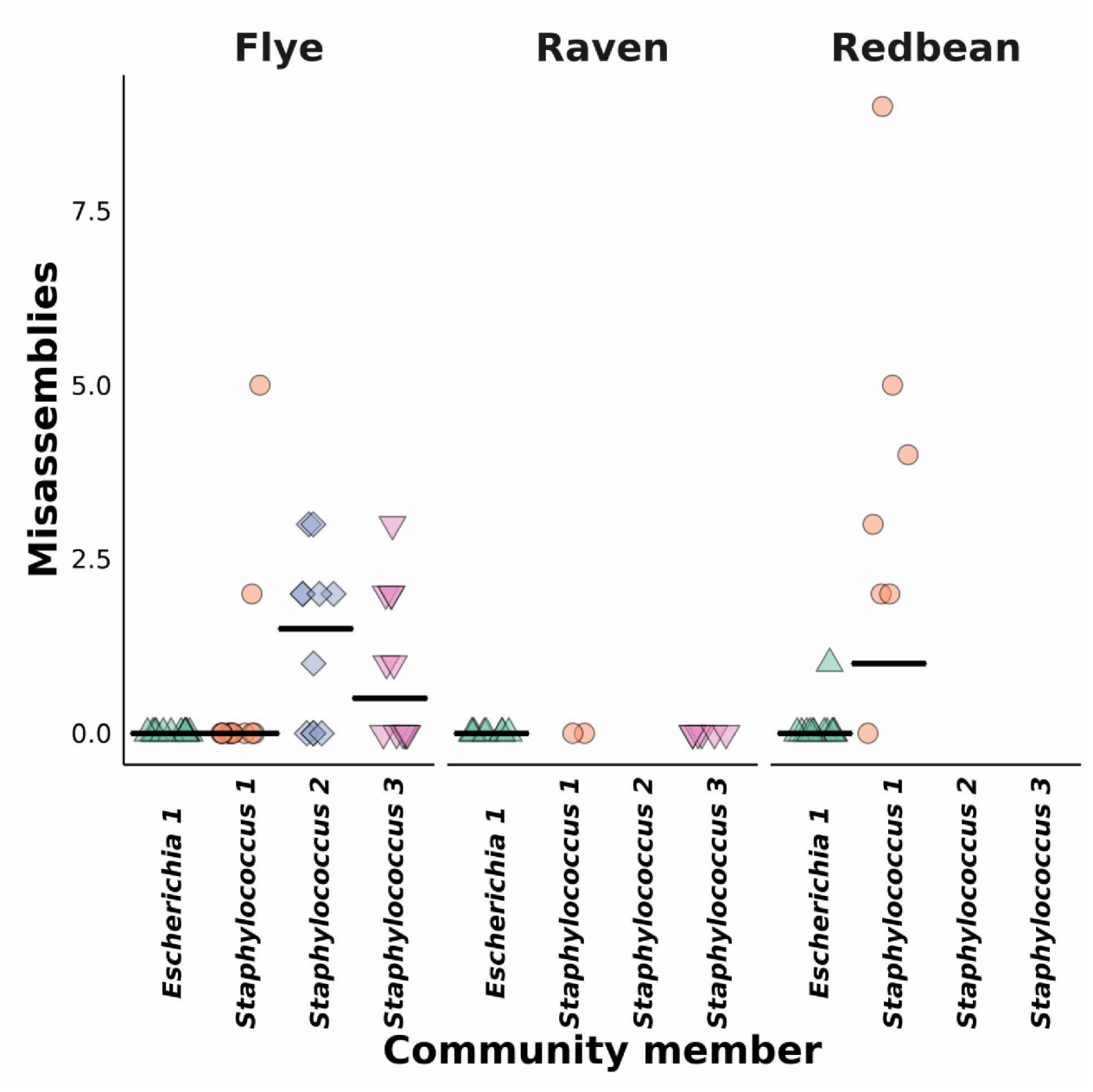

Figure 0-8: Plasmid mis-assemblies at 200x read depth. Horizontal bars indicate the median value across replicate samples.