EVALUATION OF IDENTIFICATION AND MOLECULAR CHARACTERIZATION OF FOODBORNE PATHOGENS BY WHOLE GENOME SEQUENCING THROUGH ILLUMINA AND NANOPORE

By

NICOLAS JAVIER LOPEZ GUERRA

Bachelor of Science in Biotechnology Engineering

Universidad de las Fuerzas Armadas – ESPE

Quito, Ecuador

2019

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
July, 2021

EVALUATION OF IDENTIFICATION AND MOLECULAR CHARACTERIZATION OF FOODBORNE PATHOGENS BY WHOLE GENOME SEQUENCING THROUGH ILLUMINA AND NANOPORE

Thesis Approved:

Li Maria Ma, Ph.D.

Thesis Adviser

Peter Muriana, Ph.D.

Andres Espindola, Ph.D.

ACKNOWLEDGEMENTS

I would like to express my immense gratitude my advisor Dr. Li Maria Ma, for giving me the opportunity to join Oklahoma State University and for her continuous support, motivation and valuable time during these two years of my master's program.

I would like to extent my deepest thanks to the other members of my advisory committee: Dr. Peter Muriana, for his guidance and insightful comments. And Dr. Andres Espindola who was always there to support and giving me advice that would help me to move forward with my project.

I want to thank the entire Institute of Biosecurity and Microbial Forensics for keeping an excellent environment for the students and professors.

The last, but the most important thanks to my family, especially my parents and sister, Javier, Evangelina, and Evita, for your love, support, patience, and encouragement, despite the distance, you were always there for me at any time. I would not have made it this far without you.

Name: NICOLAS JAVIER LOPEZ GUERRA

Date of Degree: JULY, 2021

Title of Study: EVALUATION OF IDENTIFICATION AND MOLECULAR CHARACTERIZATION OF FOODBORNE PATHOGENS BY WHOLE GENOME SEQUENCING THROUGH ILLUMINA AND NANOPORE

Major Field: FOOD SCIENCE

Abstract:

As next-generation sequencing (NGS) costs has dropped in recent years, whole genome sequencing (WGS) has been adopted as the primary typing method for pathogens of interest in outbreak surveillance in the U.S. In this study, we aimed to evaluate the performance of Nanopore sequencing for rapid identification and molecular characterization of *E. coli* and *Salmonella*, two major foodborne bacterial pathogens. Eleven *E. coli* and ten *Salmonella* isolates obtained from pecan orchards were sequenced using MinION and Illumina NextSeq 500. As MinION allows real-time reads analysis, the reads were time-based subsampled to determine the earliest identification turnaround time for each isolate. Species level identification was achieved at 15 mins of sequencing run. Complete antigenic profile and variants of major virulence genes were detected in 16 and 25 hours using assemblies obtained from the subsampled-reads of *E. coli* and *Salmonella*, respectively. Additionally, comparisons of the Nanopore-based assemblies against hybrid assemblies from the combined reads of MinION and Illumina showed that the best values of continuity were obtained at 4 and 8 hours; whereas, the best value of annotated features were obtained in 16 and 25 hours for *E. coli* and *Salmonella*, respectively ($p < 0.05$). By using these assemblies as input in a stringent BLASTn search (percentage of identity of 95 % and query coverage of 85 %) against the Comprehensive Antibiotic Resistance Database (CARD), we could find significantly similar results to those obtained from the hybrid assemblies of *Salmonella* but not for *E. coli* isolates. However, the hits obtained from the search against the Virulence Factor Database (VFDB) were not sufficient to generate results significantly similar for both species. Finally, the results of phylogeny analysis obtained from assemblies created with reads produced in 3 hours of sequencing process from both species, were significantly similar to those of the results with hybrid genomes ($p < 0.05$). These results demonstrated that Nanopore can offer an effective sequencing platform for the rapid identification of *E. coli* and Nontyphoidal *Salmonella* isolates, with certain capabilities for their molecular characterization.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

LIST OF APPENDICES

Appendix Page

CHAPTER I

INTRODUCTION

Annually, contaminated food results in 600 million cases of foodborne diseases and 420,000 deaths worldwide (Lee & Yoon, 2021; WHO, 2015). Meanwhile, in the U.S., the Center for Disease Control and Prevention (CDC) estimates that each year 48 million Americans get sick from a foodborne illness resulting in 128,000 hospitalizations and 3,000 deaths (CDC, 2019; Scallan et al., 2011). Rapid identification and molecular subtyping of foodborne bacterial pathogens is essential for timely response to outbreaks, therebyCDC created PulseNet, a molecular subtyping network of federal, state, and local public health laboratories designed to facilitate such process (CDC, 2019). The mission of this network is the early identification and recall of contaminated foods at the national level in order to reduce the total number of people affected by the consumption of these contaminated products (Scharff et al., 2016). It is estimated that more than 270,000 foodborne illnesses have been prevented annually in the U.S. since its implementation (Ribot & Hise, 2016). Additional economic evaluations of the PulseNet impacts by Scharff and Hedberg (2018) revealed that $5.4 billion are saved owing to improved recalls of *Escherichia coli* and *Salmonella.*

Pulsed Field Gel Electrophoresis (PFGE) was the gold standard subtyping technique used by the PulseNet, assisting in the detection and investigation of outbreaks caused by foodborne bacterial pathogens across the U.S. for more than 20 years (CDC, 2019; Gerner-Smidt et al., 2006). PFGE is a genotyping technique used for the separation of large DNA molecules after treating them with unique restriction enzymes and letting the reaction product migrate in a gel matrix under the electric field that periodically changes direction, thus generating band patterns that can be compared among different strains (Sharma-Kuinkel et al., 2016; Tang et al., 2019). Despite the efficiency with which it was possible to identify the origin of different foodborne pathogens with the use of PFGE (Gerner-Smidt et al., 2006; Ribot et al., 2019), it still has limitations related to its chemistry. The bands that are generated do not always represent homologous genetic material, therefore decreasing the discrimination power of PFGE ted such as in the studies carried out in strains of *E. coli* O157: H7 (Davis Margaret et al., 2003), *Salmonella enterica* (Hedberg et al., 2001) and *Yersinia enterocolotica* (Gilpin et al., 2014).

The emergence of affordable next-generation sequencing (NGS) technologies in recent years has opened the door to whole genome sequencing (WGS) as a viable and cost-effective subtyping approach for foodborne bacterial pathogens surveillance (Ribot et al., 2019). Unlike PFGE, WGS encompasses more information from an organism's genetic material, and its utility for outbreak investigations has already been demonstrated for several gastrointestinal pathogens (Jenkins et al., 2015; McDonnell et al., 2013; Moura et al., 2016; Quick et al., 2015). Incentivized by affordable costs and potential increase in discrimination, PulseNet is transitioning to WGS, wherein this technique has

already been standardized as the primary typing method for *Listeria, Salmonella, E. coli, Shigella,* and *Campylobacter* (Tolar et al., 2019). Additionally, the Food and Drug Administration (FDA) funds a network called GenomeTrakr which also stores data obtained from WGS of foodborne bacterial pathogens (Timme et al., 2019). The information from both platforms, PulseNet and GenomeTrakr, is yielded through MiSeq, i.e., an Illumina-based sequencing platform, and stored in the National Center of Biotechnology Information (NCBI) virtual repository (Tolar et al., 2019).

Illumina offers high quality short reads with an average error rate of ~1% (Stoler & Nekrutenko, 2021); however, the length of the reads can hamper sequencing complex or highly repetitive regions of the genome, which constitutes a major challenge for *de novo* sequencing of bacterial genomes, because bacterial chromosomes contain up to several dozens of intragenic and intergenic tandem repeats (Adewale, 2020; Alkan et al., 2011) These regions can offer valuable clues about co-regulated gene clusters or the presence of a gene of interest within a transmissible mobile genetic element (Kuśmirek & Nowak, 2018; Zhou et al., 2014). On the other hand, Nanopore sequencing platform overcomes this limitation by producing reads that can span thousands of nucleotides, allowing complex regions to be sequenced more easily (Logsdon et al., 2020). Additionally, this sequencing platform offers the ability to work with sequencing reads in real time, thus allowing the turnaround time in outbreak response to be shortened (Logsdon et al., 2020; Taylor et al., 2019). Nonetheless, the tradeoff of this practically new technology is its high sequencing error rate of between 5 and 15% (Rang et al., 2018). However, constant developments in Nanopore chemistry and the development of new bioinformatics tools have provided an improved landscape for the analysis of reads

from this platform (Adewale, 2020; Goldstein et al., 2019; Taylor et al., 2019). Therefore, it is imperative to explore the scope of this technology in foodborne bacterial surveillance.

In this study, we aimed to evaluate the performance of Nanopore sequencing for rapid identification and molecular characterization of *E. coli* and *Salmonella,* i.e., two major foodborne bacterial pathogens by PulseNet. Noteworthy, the bioinformatics work in this study was performed using the sequencing reads generated from the doctoral dissertation of Diaz-Proano (2019). Whereby, the information generated from the *E. coli* and *Salmonella* isolates were analyzed in chapter 3 and chapter 4, respectively. Wherein the Nanopore reads were subsampled with respect to time, and the subsequent analyzes were compared to the data generated from hybrid genomes that were created using Illumina in conjunction with the Nanopore reads. Overall, the results from this study should contribute as an example of the advantages and limitations of using Nanopore technology for foodborne bacterial pathogens surveillance

REFERENCES

Adewale, B. A. (2020). Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years? *African journal of laboratory medicine, 9*(1), 1340-1340. https://doi.org/10.4102/ajlm.v9i1.1340

Alkan, C., Sajjadian, S., & Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nature Methods*, *8*(1), 61-65. https://doi.org/10.1038/nmeth.1527

CDC. (2019). *About PulseNet*. Retrieved 04/02/2021 from www.cdc.gov/pulsenet/about/index.html

Davis Margaret, A., Hancock Dale, D., Besser Thomas, E., & Call Douglas, R. (2003). Evaluation of Pulsed-Field Gel Electrophoresis as a Tool for Determining the Degree of Genetic Relatedness between Strains of Escherichia coli O157:H7. *Journal of Clinical Microbiology*, *41*(5), 1843-1849. https://doi.org/10.1128/JCM.41.5.1843-1849.2003

Diaz-Proano, C. (2019). *Prevalence, molecular characterization and inactivation of foodborne pathogens on native pecans* Oklahoma State University].

Gerner-Smidt, P., Hise, K., Kincaid, J., Hunter, S., Rolando, S., Hyytiä-Trees, E., Ribot, E. M., & Swaminathan, B. (2006). PulseNet USA: a five-year update. *Foodborne Pathog Dis, 3*(1), 9-19. https://doi.org/10.1089/fpd.2006.3.9

Gilpin, B. J., Robson, B., Lin, S., Hudson, J. A., Weaver, L., Dufour, M., & Strydom, H. (2014). The limitations of pulsed-field gel electrophoresis for analysis of Yersinia enterocolitica isolates. *Zoonoses Public Health, 61*(6), 405-410. https://doi.org/10.1111/zph.12085

Goldstein, S., Beka, L., Graf, J., & Klassen, J. L. (2019). 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

Hedberg, C. W., Smith, K. E., Besser, J. M., Boxrud, D. J., Hennessy, T. W., Bender, J. B., Anderson, F. A., & Osterholm, M. T. (2001). Limitations of Pulsed-Field Gel Electrophoresis for the Routine Surveillance of Campylobacter Infections. *The Journal of Infectious Diseases, 184*(2), 242-243. https://doi.org/10.1086/322005

Jenkins, C., Dallman Timothy, J., Launders, N., Willis, C., Byrne, L., Jorgensen, F., Eppinger, M., Adak Goutam, K., Aird, H., Elviss, N., Grant Kathie, A., Morgan, D., McLauchlin, J., & Elkins, C. A. (2015). Public Health Investigation of Two Outbreaks of Shiga Toxin-Producing Escherichia coli O157 Associated with Consumption of Watercress. *Applied and Environmental Microbiology, 81*(12), 3946-3952. https://doi.org/10.1128/AEM.04188-14

Kuśmirek, W., & Nowak, R. (2018). De novo assembly of bacterial genomes with repetitive DNA regions by dnaasm application. *BMC Bioinformatics, 19*(1), 273. https://doi.org/10.1186/s12859-018-2281-4

Lee, H., & Yoon, Y. (2021). Etiological Agents Implicated in Foodborne Illness World Wide. *Food science of animal resources, 41*(1), 1-7. https://doi.org/10.5851/kosfa.2020.e75

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics, 21*(10), 597-614. https://doi.org/10.1038/s41576-020-0236-x

McDonnell, J., Dallman, T., Atkin, S., Turbitt, D. A., Connor, T. R., Grant, K. A., Thomson, N. R., & Jenkins, C. (2013). Retrospective analysis of whole genome sequencing compared to prospective typing data in further informing the epidemiological investigation of an outbreak of Shigella sonnei in the UK. *Epidemiology and Infection, 141*(12), 2568-2575. https://doi.org/10.1017/S0950268813000137

Moura, A., Criscuolo, A., Pouseele, H., Maury, M. M., Leclercq, A., Tarr, C., Björkman, J. T., Dallman, T., Reimer, A., Enouf, V., Larsonneur, E., Carleton, H., Bracq-Dieye, H., Katz, L. S., Jones, L., Touchon, M., Tourdjman, M., Walker, M., Stroika, S., Cantinelli, T., Chenal-Francisque, V., Kucerova, Z., Rocha, E. P., Nadon, C., Grant, K., Nielsen, E. M., Pot, B., Gerner-Smidt, P., Lecuit, M., & Brisse, S. (2016). Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes. *Nat Microbiol, 2*, 16185. https://doi.org/10.1038/nmicrobiol.2016.185

Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., Nair, S., Neal, K., Nye, K., Peters, T., De Pinna, E., Robinson, E., Struthers, K., Webber, M., Catto, A., Dallman, T. J., Hawkey, P., & Loman, N. J. (2015). Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome biology, 16*(1), 114. https://doi.org/10.1186/s13059-015-0677-2

Rang, F. J., Kloosterman, W. P., & de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome biology, 19*(1), 90. https://doi.org/10.1186/s13059-018-1462-9

Ribot, E. M., Freeman, M., Hise, K. B., & Gerner-Smidt, P. (2019). PulseNet: Entering the Age of Next-Generation Sequencing. *Foodborne pathogens and disease, 16*(7), 451-456. https://doi.org/10.1089/fpd.2019.2634

Ribot, E. M., & Hise, K. B. (2016). Future challenges for tracking foodborne diseases [https://doi.org/10.15252/embr.201643128]. *EMBO reports, 17*(11), 1499-1505. https://doi.org/https://doi.org/10.15252/embr.201643128

Scallan, E., Griffin, P. M., Angulo, F. J., Tauxe, R. V., & Hoekstra, R. M. (2011). Foodborne illness acquired in the United States--unspecified agents. *Emerging infectious diseases, 17*(1), 16-22. https://doi.org/10.3201/eid1701.091101p2

Scharff, R. L., Besser, J., Sharp, D. J., Jones, T. F., Peter, G.-S., & Hedberg, C. W. (2016). An Economic Evaluation of PulseNet: A Network for Foodborne Disease Surveillance. *American Journal of Preventive Medicine, 50*(5, Supplement 1), S66-S73. https://doi.org/https://doi.org/10.1016/j.amepre.2015.09.018

Scharff, R. L., & Hedberg, C. (2018). The Role of Surveillance in Promoting Food Safety. In T. Roberts (Ed.), *Food Safety Economics: Incentives for a Safer Food Supply* (pp. 251-265). Springer International Publishing. https://doi.org/10.1007/978-3-319-92138-9_13

Sharma-Kuinkel, B. K., Rude, T. H., & Fowler, V. G., Jr. (2016). Pulse Field Gel Electrophoresis. *Methods in molecular biology (Clifton, N.J.), 1373*, 117-130. https://doi.org/10.1007/7651_2014_191

Stoler, N., & Nekrutenko, A. (2021). Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics, 3*(1). https://doi.org/10.1093/nargab/lqab019

Tang, S., Orsi, R. H., Luo, H., Ge, C., Zhang, G., Baker, R. C., Stevenson, A., & Wiedmann, M. (2019). Assessment and Comparison of Molecular Subtyping and Characterization Methods for Salmonella. *Frontiers in Microbiology, 10*, 1591-1591. https://doi.org/10.3389/fmicb.2019.01591

Taylor, T. L., Volkening, J. D., DeJesus, E., Simmons, M., Dimitrov, K. M., Tillman, G. E., Suarez, D. L., & Afonso, C. L. (2019). Rapid, multiplexed, whole genome and plasmid sequencing of foodborne pathogens using long-read nanopore technology. *Scientific Reports, 9*(1), 16350. https://doi.org/10.1038/s41598-019-52424-x

Timme, R. E., Sanchez Leon, M., & Allard, M. W. (2019). Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback. *Methods Mol Biol, 1918*, 201-212. https://doi.org/10.1007/978-1-4939-9000-9_17

Tolar, B., Joseph, L. A., Schroeder, M. N., Stroika, S., Ribot, E. M., Hise, K. B., & Gerner-Smidt, P. (2019). An Overview of PulseNet USA Databases. *Foodborne pathogens and disease, 16*(7), 457-462. https://doi.org/10.1089/fpd.2019.2637

WHO. (2015). WHO estimates of the global burden of foodborne diseases. https://apps.who.int/iris/bitstream/handle/10665/199350/9789241565165_eng.pdf?sequence=1

Zhou, K., Aertsen, A., & Michiels, C. W. (2014). The role of variable DNA tandem repeats in bacterial adaptation. *FEMS Microbiol Rev, 38*(1), 119-141. https://doi.org/10.1111/1574-6976.12036

CHAPTER II

LITERATURE REVIEW

OVERVIEW OF FOODBORNE PATHOGENS

Apart from the intake of food contaminated with chemicals or other agents detrimental to human health, foodborne illnesses can be caused by consuming food contaminated by pathogenic microorganisms or their toxins. Those pathogenic microorganisms, including viruses, bacteria, parasites, and fungi, are responsible for millions of reported cases of foodborne illnesses and/or chronic complications in many countries (Heredia & García, 2018; Schirone et al., 2019). However, these are not enemies that have arisen recently, throughout the history of mankind, foodborne pathogens have accompanied us, being even essential factors that shaped our past. As one of the most emblematic examples stands out is the death of Alexander the Great in 323 B.C. Regardless of his unprecedented military campaign through Western Asia and Northeastern Africa, he is believed to have been a victim of typhoid fever, a disease caused by *Salmonella* Typhi, one of the well-characterized foodborne bacterial pathogens nowadays (University Of Maryland Medical Center, 1998). Despite the ongoing risk posed by these microorganisms, it is of utmost relevance to understand that pathogens

are the minority among harmless microorganisms that cohabit the biosphere with humans, therefore, it is imperative to differentiate those that are truly a risk to human health from those that are not (National Academies Press, 2010; Pepper & Gentry, 2015).

At present, the biological agents most frequently involved in foodborne diseases have been identified and characterized (Table 1.), of which those that are reported more frequently worldwide are bacteria, such as *Salmonella*, *Campylobacter*, and Enterohemorrhagic *Escherichia coli*; and viruses, specially Noroviruses (WHO, 2020). The largest number of outbreaks caused by foodborne pathogens in China, the Republic of Korea, and the European Union were due to bacteria (European Food Safety Authority, 2018; Kim & Kim, 2021; W. Li et al., 2020); whereas viruses accompany bacteria as the protagonists of the largest number of outbreaks caused by foodborne pathogens in the US. However, in the US, foodborne bacterial pathogens were responsible for 90% of the total hospitalizations related to foodborne illnesses, with *Salmonella* being the most predominant among them (CDC, 2019b).

Table 1. Estimated annual percentage of the total number of domestically acquired foodborne illnesses, hospitalizations, and deaths caused in the United States (Elaine Scallan et al., 2011). STEC: Shiga toxin–producing *Escherichia coli*; ETEC: enterotoxigenic *E. coli*.

Pathogen	**Illnesses (%)**	**Hospitalizations (%)**	**Deaths (%)**
Viruses			
Norovirus	58.18	26.20	11.04
Astrovirus	0.16	0.16	0.01>
Rotavirus	0.16	0.62	0.01>
Sapovirus	0.16	0.16	0.01>
Hepatitis A virus	0.02	0.18	0.52
Bacteria			
Salmonella spp. nontyphoidal	10.95	34.55	28.00
Clostridium perfringens	10.29	0.78	1.93
Campylobacter spp.	9.00	15.12	5.63
Staphylococcus aureus	2.57	1.90	0.44
Shigella spp.	1.40	2.60	0.74

STEC non-O157	1.20	0.48	0.01>
Yersinia enterocolitica	1.04	0.95	2.15
Bacillus cereus	0.68	0.04	0.01>
STEC O157	0.67	3.82	1.48
V. parahaemolyticus	0.37	0.18	0.30
ETEC	0.19	0.02	0.01>
Vibrio spp. other	0.19	0.15	0.59
Diarrheagenic *E. coli* other than STEC and ETEC	0.13	0.01	0.01>
Streptococcus spp. group A	0.12	0.01>	0.01>
S. enterica serotype Typhi	0.02	0.35	0.01>
Listeria monocytogenes	0.02	2.60	18.89
Brucella spp.	0.01	0.10	0.07
V. vulnificus	0.01>	0.17	2.67
Vibrio cholerae toxigenic	0.01>	0.01>	0.01>
Mycobacterium bovis	0.01>	0.06	0.22
Clostridium botulinum	0.01>	0.08	0.67

Parasites			
Toxoplasma gondii	0.92	7.91	24.22
Giardia intestinalis	0.82	0.40	0.15
Cryptosporidium spp.	0.61	0.38	0.30
Cyclospora cayetanensis	0.12	0.02	0.01>
Trichinella spp.	0.01>	0.01	0.01>

Virulence factors

Virulence is the relative capacity of a microorganism to cause damage in a host (Casadevall & Pirofski, 2003). Hence, bacterial virulence factors (VFs) are components that increase the chances of survival of pathogenic bacteria that express them while infecting the host so that these bacteria replicate and disseminate within it successfully (Cross, 2008; Peterson, 1996; Sharma et al., 2017). The function they provide are various but each one can play a fundamental role in the efficiency of the infection through different mechanisms among which we commonly find: Uptake of essential nutrients from the infected host, colonization of target tissues, invasion through the host, or protection against host defenses (Peterson, 1996; Sharma et al., 2017).

Despite the correlation among VFs and bacterial pathogens, the presence of several VFs alone in a microorganism does not indicate that the microorganism is pathogenic to a host, because the relationship and interaction between a microorganism and its host are what determine whether a strain is pathogenic or not (Ho Sui et al., 2009;

Pallen & Wren, 2007; Zhang & Zhang, 2006). Some VFs may only have evolved as host-interaction factors in commensal strains that diverged into pathogenic strains, such as those involved in adhesion or different metabolic pathways, whereas others evolved to have more "offensive" functions with the host, such as those involved in active invasion or directly causing damage. (Ho Sui et al., 2009). For example, *Neisseria meningitidis* harboring the factors associated with virulence (e.g. capsule and type IV pili) can remain non-invasive in human carriers (Laver et al., 2015). Another special case is that of non-pathogenic *E. coli* that have evolved together with humans, gaining several VFs to thrive as a commensal in our gut, among them we can find adhesins or siderophores, VFs that are also part of the pathogenetic machinery of other microorganisms, such as *Salmonella* spp. or *Klebsiella* spp. (Sarowska et al., 2019). Nevertheless, pathogenic *E. coli* strains have acquired additional genetic material through horizontal gene transfer (HGT) that over time were not rejected but rather incorporated as useful tools when surviving more effectively in their niche or infecting their hosts (Diard & Hardt, 2017).

VFs detection is unlikely to be a routine requirement for most industrial laboratories but can be essential for overall risk assessment, whether for an individual strain or a population. For instance, VFs detection can be performed when it is necessary to take precautionary measures to reduce or eradicate microorganisms that carry particular VFs, such as for the development of immunization programs for animal production (Crasta et al., 2008; Luo et al., 2015; Rabinovitz et al., 2012), or for the evaluate virulence in environmental samples (Kimani et al., 2014; Menezes et al., 2014; Prieto et al., 2016). VFs can be detected by methods based on their exerted activity, such as the Vero cell assay or the reversed passive latex agglutination (RPLA) test (Liptáková

et al., 2002; To & Bhunia, 2019), or through molecular techniques, such as the use of multiplex PCR or the use of proteomes or genomes for detection to higher resolution compared to traditional techniques (Allen et al., 2020; Scheutz et al., 2012; Zheng et al., 2012). Currently, there are national surveillance systems to monitor the incidence of foodborne diseases related to microorganisms that harbor particular VFs, such as the National Surveillance of Shiga Toxin-producing Escherichia coli (STEC); or diseases caused by the presence of their toxins in food products, such as the National Botulism Surveillance, both led by the Center of Disease Control & Prevention (CDC) in order to provide a national snapshot of the occurrence of infections transmitted primarily through food (CDC).

Antibiotic resistance

Antibiotics are drugs that treat bacterial infections in humans and animals by either killing the bacteria or hindering their growth and multiplication, thereby they have been widely used either as therapeutic or as prophylactic agents, especially in agricultural animal production as growth promoters (van Hoek et al., 2011). Six major cell functions have been targeted by antibiotics: inhibitors of DNA replication (DNA synthesis and DNA gyrase), RNA synthesis, protein synthesis (50S or 30S ribosomal subunit inhibitors), cell wall biosynthesis, cell membrane biosynthesis, and fatty acid synthesis (O'Rourke et al., 2020).

Antibiotic resistance (AR) phenotypes can arise in microorganisms from chromosomal DNA mutations, which alter existing bacterial proteins or non-coding regions involved in transcriptional or translational regulation of the targets, as well as a

result of the acquisition of new genetic material between bacteria of the same or different species or genera (A. Ghosh et al., 2020; Maiden, 1998; van Hoek et al., 2011). The main mechanisms of resistance are: permeability changes in the bacterial cell wall achieved by new porin variants due to mutations in the genes that encode them, as well as by the downregulation of porins produced by mutations in genes or regions that regulate their expression which hinders drug access to target sites (Lavigne et al., 2013; Novais et al., 2012; Tamber & Hancock, 2003), active efflux of the antibiotic from inside of the bacteria resulting from acquired novel efflux pumps genes through HGT or due to their overexpression caused by mutations in regulatory genes (Abouzeed Yousef et al., 2008; Hung et al., 2013; Pomposiello Pablo et al., 2001), acquired genes that encode enzymes capable of altering or degrading the antibiotic (Johnson & Woodford, 2013; Poirel et al., 2012; Wright, 2005), single point mutations in genes encoding antibiotic targets or their regulatory mechanisms leading to an overexpression of the target, and the acquisition of genes homologous to the original target that can counteract the effect of the antibiotic (Gao et al., 2010; Katayama et al., 2000; Shore Anna et al., 2011).

Resistance to antibiotics in bacteria is currently a global crisis (Martens & Demain, 2017; Podolsky, 2018). Worldwide, at least 700,000 people die annually due to bacterial infections unsuccessfully treated related to AR, whereby these values are estimated to reach 10 million per year by 2050 if the necessary measures are not taken to stop this crisis (Strathdee et al., 2020). Although resistant bacteria can occur in nature (Allen et al., 2010), the main driver for the accelerated appearance of strains resistant to one or more antibiotics has been contributed to anthropogenic activities, such as the excessive use of antibiotics since their discovery (Roberts & Zembower, 2021; Van

Boeckel et al., 2014), inappropriate antibiotics prescription (Milani et al., 2019), extensive agricultural usage (Spellberg & Gilbert, 2014; C. Lee Ventola, 2015), improper handling of waste fluid containing active pharmaceutical ingredients from antibiotic manufacturing plants (Ahmad et al., 2017), or acquisition of resistance induced by exposure to disinfectants that are used ubiquitously (Amsalu et al., 2020; Jin et al., 2020; Kim et al., 2018).

The techniques that are currently used for the detection of antibiotic resistance (broth dilution and disk diffusion techniques) have a slow turnaround time, as well as they can be nonconclusive and not broad enough (Anjum et al., 2018; Hashempour-Baltork et al., 2019); consequently, such culture-dependent phenotypic methods can thus delay decision-making in the medical or agricultural field, so molecular analyzes such as PCR or DNA chips have been implemented, as they can be used to investigate the presence of a resistance gene or point mutation, providing direct support to ensure that an optimal treatment or control strategy is executed in a timely manner (Hashempour-Baltork et al., 2019; Woolhouse et al., 2015). Moreover, molecular characterization is regularly adopted as an indirect method to assist in epidemiological investigations after an outbreak complementary to phenotypic tests, since *in vitro* phenotypic methods can be sometimes not sufficiently conclusive to rule out that the bacterium analyzed is resistant *in vivo* or not, thus requiring additional information about its genotype (Anjum et al., 2018; Petersen et al., 2011; Robinson et al., 1999). The recent advancements in rapid and affordable DNA sequencing technologies, known as Next Generation Sequencing (NGS), have offered a better resolution at the entire genome level, hence Whole Genome Sequencing (WGS) is being used for the characterization of AR strains in local, national,

or even global surveillance of pathogens (Ribot et al., 2019). However, the molecular characterization of AR is limited based on the information already characterized regarding the genotype of resistant strains, as well as the lack of ensuring that the expression of a resistance gene will be favored by bacterial regulatory systems, therefore molecular techniques will continue to be a complement and not a total replacement for phenotypic methods. (Lepuschitz et al., 2019; Palmer & Kishony, 2013; Urmi et al., 2020).

Currently, in the U.S., the Center for Disease Control and Prevention (CDC) is involved in monitoring and coordinating surveillance of AR in important zoonotic bacteria isolated from animals intended for human consumption and human clinic samples. With several tracking systems in place, such as the Antibiotic Resistance Laboratory Network (AR Lab Network) and the National Antimicrobial Resistance Monitoring System for Enteric Bacteria (NARMS), data regarding the national prevalence of AR is being generated which would streamline decision-making and strengthen the surveillance system (CDC).

Mobile genetic elementsMobile Genetic Elements (MGEs) are segments of DNA with varying lengths (1 to several hundred kb) that encode their machinery to thus move within genomes (intracellular mobility) or between bacterial cells (intercellular mobility) (Frost et al., 2005; Miller & Capy, 2004). Intercellular mobility can be achieved through transformation, conjugation or transduction (Frost et al., 2005). Transformation involves the uptake of extracellular DNA from closely related bacteria and is mediated by chromosomally encoded proteins from some naturally transformable bacteria (Frost et al.,

2005; Snyder & Champness, 2007). Whereas, conjugation depends independently on replicating genetic elements called conjugative plasmids, or chromosomally integrated conjugative elements (ICEs), which harbor genes that facilitate their transfer and sporadically the transfer of other cellular DNA from a plasmid-carrying cell, also called donor, to a recipient cell that lacks the plasmid or ICE (Botelho & Schulenburg, 2021; Frost et al., 2005; Snyder & Champness, 2007). On the other hand, transduction is the acquisition of DNA mediated by bacterial viruses (bacteriophages or phages). At low frequency, these obligate intracellular parasites can encapsidate segments of host DNA, which would subsequently be transferred to other bacteria when infected by this new phage particle that contains part of the genome of the previous host. The DNA can then recombine into the chromosome or replicate as a plasmid in the new host cell (Miller & Capy, 2004; Snyder & Champness, 2007). Finally, intracellular movement of DNA is a property of loci with high recombination rates commonly known as transposons, which randomly recombine or 'jump' between replicons, DNA fragments that have at least have one origin of replication and one binding site wherein DNA-binding proteins called initiators will recruit the additional factors necessary to open the double-stranded chromosomal DNA and begin synthesis (Maga, 2017). Since these promiscuous elements can come into contact with phages or plasmids, they can also be transferred into other cells (Frost et al., 2005; Sabbagh et al., 2021).

The acquisition of preexisting genetic determinants, such as genes encoding AR or VFs, through MGEs in conjunction with mutations and selective pressure are the key elements in bacterial evolution (Leplae et al., 2004). For instance, the laboratory *E. coli* strain K12 surprisingly differs genomically by about 20-30% compared to pathogenic *E.*

coli O157:H7, of which the main differences come from prophages integrated into the chromosome of the pathogenic strain (Hayashi et al., 2001; Kudva et al., 2002). Similarly, *Listeria monocytogenes* and *Listeria innocua* differ mainly to prophages that are only present in the latter (den Bakker et al., 2010; Glaser et al., 2001). Extra chromosomic genetic material can also play a key role in species differentiation, such as the case of pathogenic *Bacillus anthracis* and its close relative *Bacillus thuringiensis*, in which despite the similarity between their chromosomes, the type of plasmids that they harbor is different, thus providing essential dissimilitude that has a strong impact on the virulence of the former (Luna et al., 2006; Read et al., 2003).

Plasmids and other conjugative elements

A plasmid is a DNA molecule composed of functional genetic modules that give rise to a stable and self-replicating entity complex, which is smaller than bacterial chromosomes and usually lacks essential genes for basal cell functions (Frost et al., 2005). Most often, plasmids are covalently closed into circular double-stranded DNA molecules, but linear double-stranded DNA plasmids can also be found in some species (Frost et al., 2005; Partridge et al., 2018; Shintani et al., 2015). Genes involved in the replication of plasmids represent a core of plasmid housekeeping functions, also known as the "backbone". Additionally, a plasmid can harbor a wide variety of accessory genes, including those involved in niche adaptive functions that might benefit the host cell (Frost et al., 2005). Replication of plasmids starts at a specific region called "origin" (*ori*), caused by the

coupling of an initiation protein (Rep) to a proximal iterated DNA repeat sequences called iterons which hijacks the replication machinery of the cell for its benefit (Snyder & Champness, 2007). The presence of the gene encoding Rep in the plasmid broadens the host range of the plasmid, yet dependence on host-encoded DNA replication proteins is a fundamental factor limiting the host range of plasmids (Frost et al., 2005; Partridge et al., 2018; Snyder & Champness, 2007). Nonetheless, plasmids with multiple replication regions are quite common in both Gram-negative and Gram-positive bacteria, suggesting that fusions/cointegrations between plasmids are a common phenomenon (Johnson et al., 2007; Partridge et al., 2018; Villa et al., 2010). Three modes of plasmid replication have been described for circular plasmids: 1) rolling circle (RC) replication which is frequently used by small plasmids in Gram-positive and, less commonly in Gram-negative bacteria; 2) theta-mode replication, which resembles circular chromosome replication and is widely used by small to very large plasmids, and 3) strand displacement which is commonly found in small plasmids (del Solar et al., 1998; Partridge et al., 2018). In order to balance the competing demands of effective plasmid inheritance and metabolic burden impost on the host, plasmids control their copy number by using a wide variety of strategies, although two basic strategies have been elucidated: in the first, an antisense RNA binds to the transcribed Rep mRNA thus inhibiting its translation and indirectly restricting replication, and in the second mechanism, the Rep proteins bind together and seize plasmids through their iterons, restricting directly their replication (del Solar & Espinosa, 2000; Frost et al., 2005; Partridge et al., 2018). Once replicated, plasmids need a mechanism that ensures their maintenance in daughter cells during cell division. Small plasmids are commonly maintained at a high copy number, wherein random segregation

in both daughter cells is sufficient to achieve an efficient inheritance (Million-Weaver & Camps, 2014; Münch et al., 2019; Wang, 2017). On the other hand, large plasmids are present in a low copy number, so they require certain functional modules that contribute to their maintenance. These include multimer resolution systems (res) that recombinantly separate multimeric plasmids into monomers that are segregated independently to daughter cells, partitioning (par) systems that actively distribute plasmid copies to daughter cells, and postsegregational killing systems that hinder the fitness of the progeny cells that fail to inherit a copy of the plasmid (Million-Weaver & Camps, 2014; Shintani et al., 2015). Plasmid propagation is facilitated not only through vertical transmission via cell division but also via horizontal transmission to other bacterial cells, where although conjugation is the main responsible for the transfer of plasmids, there is evidence that bacteria can also uptake plasmids through natural transformation (Hasegawa et al., 2018; Nolan et al., 2020; Partridge et al., 2018). Conjugative plasmids are self-transmissible for which they possess additional components, thus increasing significantly the size of their conserved backbone (Partridge et al., 2018). Among the most essential plasmid parts for conjugation, we can find the transfer regions (tra) which encode proteins for the formation of mating pairs (MPF), a protein complex that functions as a specialized pore of the type IV secretion system (T4SS), by which a conjugative pilus is assembled, forming a filamentous surface appendage that mediates interactions with recipient cells (Alvarez-Martinez & Christie, 2009; Komano et al., 2000; Kurenbach et al., 2002); as well as DNA transfer replication proteins (DTR) that process plasmid DNA, such as a relaxase that specifically cuts the origin of transfer (oriT) present in the DNA strand to be exported to the recipient cell (Cabezón et al.,

2014; Giusti Mde et al., 2012; Partridge et al., 2018). Some non-conjugative plasmids can still be transferred horizontally by exploiting the MPF apparatus contributed by a conjugative plasmid present in the same cell. These plasmids carry only a subset of the DTR functions (usually termed Mobility or MOB), including an oriT and a gene for a corresponding relaxase (Frost et al., 2005; Partridge et al., 2018; Snyder & Champness, 2007).

Plasmids with the same replication and partition systems can't be propagated stably in the same host cell line, a phenomenon termed 'incompatibility' (Inc) (Shintani et al., 2015). Hence a classification system based on Inc has been widely adopted with great success, however, the laborious nature of incompatibility testing resulted in it being superseded by hybridization, then PCR-based replicon typing (PBRT), and ultimately sequencing-based approaches. Nowadays, because only the amino acid sequence of the Rep protein is taken into account to group different plasmids into Inc types, it is not necessarily confirmed by conventional methods if the plasmid shows incompatibility with the same plasmid of the Inc group in the same host cell line (Shintani et al., 2015). The drawbacks of using a replicon typing to classify plasmids rely on the inability to distinguish plasmids that carry more than one replicons, as well as there is not enough information about Rep types present among several microbial taxonomies (Million-Weaver & Camps, 2014; Rawlings & Tietze, 2001; Shintani et al., 2015), thus hindering the identification of other types of plasmids. Therefore, classification of plasmids based on MOB typing, which encompass conjugative and mobilization relaxase genes, and Mpf classes have emerged, yet these methods are not appropriate for non-transmissible plasmids (Orlek, Phan, et al., 2017; Orlek, Stoesser, et al., 2017; Shintani et al., 2015).

Bacteriophages and transposons

As discussed before, phages can play a key role in the intercellular transfer of genetic elements through transduction. Bacteriophage genomes can be composed of either single- or double-stranded DNA or RNA, with a genomic size ranging from a few to several 100 kb (Snyder & Champness, 2007). Commonly, bacteriophages need to harbor core genes involved in the expression of machinery used to hijack the host cell replicative, specific replicase genes, and the components that are part of their capsid (Canchaya et al., 2003; Chiang et al., 2019). Depending on their life cycle, they can be classified as virulent bacteriophages due to their vigorous replication and lysis of the host once inside them, or temperate bacteriophages which can be found in a quiescent, non-lytic growth mode called lysogeny (Chiang et al., 2019). The lysogenic bacteriophages are responsible for HGT, as they integrate their genome into the bacterial chromosome and replicate with it as a prophage, where eventually stress conditions, such as DNA damage, can induce the reassembly of viral particles into new phages in which portions of the DNA from the host cell can be accidentally packaged and later injected into a new host (Partridge et al., 2018). The ability to transduce host DNA seems to be limited to relatively large (50–100 kb) double-stranded DNA phages, and the transduced chunks of DNA must be able to recombine with the genome of the recipient host to prevail, hence transduction is limited to members of the same bacterial species (Canchaya et al., 2003; Frost et al., 2005; Hendrix, 2003). The mosaic structure of phages is the result of recombination between prophages and other mobile elements that reside in the same bacterial host (Belcaid et al., 2010; Casjens & Thuman-Commike, 2011). None of the phage genes is sufficiently specific or highly conserved to be employed as a single

marker for prophages detection (Canchaya et al., 2003), thereby, multiple strategies have been used to create algorithms that can identify prophages in data obtained from Next Generation Sequencing (NGS), such as the use of random forest machine learning to predict phage sequences (Amgarten et al., 2018), identification of assembled genomic fragments of phage origin by comparison with whole genome bacteriophage sequences (Jurtz et al., 2016), or the assessment of characteristics of prophages that exhibit no similarity to sequence genome and complies certain scores to be classified as prophages (Akhter et al., 2012).

Intracellular transfer of genetic elements is mainly carried out by transposons and integrons, which give rise to more complex mobile genetic elements such as genomic islands (GI) or integrative conjugative elements (ICE) (Frost et al., 2005; Partridge et al., 2018; Snyder & Champness, 2007). Transposons are mobile elements that harbor a site-specific tyrosine and serine recombinases transposase (tnp) gene, which produce either a tyrosine or a serine recombinase that is site-specific to the flanking direct (DR) or indirect repeat (IR) sequences and is accompanied by other genes that are not involved in the transposition process. When transposons only possess the tnp gene and not accessory genes, they are called Inserted Sequences (IS), contrarily, they can also be formed by the integration of one or various transposons into another which is known as a composite transposon (Frost et al., 2005; Partridge et al., 2018; Snyder & Champness, 2007). These MGEs can be divided into groups based on two different active site amino acidic motifs in Tnp, most commonly DDE (Asp, Asp, and Glu) but also DEDD and HUH (His, U: Large hydrophobic amino acid, and His); or based on whether transposition is a conservative, cut-and-paste mechanism, or replicative process (Babakhani & Oloomi,

2018; Rice & Baker, 2001). Miniature inverted-repeat transposable elements (MITEs) are non-autonomous derivatives of bacterial IS or transposons that retain the IR but which have lost central parts, including the transposase gene (Delihas, 2008). On the other hand, integrons are mobile DNA elements with the ability to capture genes by site-specific recombination mediated by a site-specific tyrosine recombinase encoded by an integrase gene (int) which, unlike transposases, does not recognize IR or DR, but multiple recombination sites (attI from the integrin and attC from a cassette) (Collis et al., 1998; Hall & Collis, 1995; Rice & Baker, 2001). A GI is a chromosomal region that has been acquired via horizontal transfer; in many cases, GIs are flanked by DRs (Malachowa & DeLeo, 2010; Partridge et al., 2018). GIs has a variable size due to their variable genetic contents and can be classified based on their encoded phenotype(s). For instance, resistance islands are GIs that harbor multiple resistance determinants, meanwhile, those that contain virulence factors are often called pathogenicity islands (Frost et al., 2005). Finally, ICEs are integrative mobilizable elements found in both Gram-negative and Gram-positive bacteria that are self-transmissible by conjugation due to the presence of encoded factors essential for their transmission, including transcription activators to induce MGI excision, a specific relaxase to initiate the transfer at the ICE integrated oriT and the conjugation machinery to transport them to recipient cells, wherein they are integrated into the new host bacterial chromosome and replicated as part of it (Cameron et al., 2019; Johnson & Grossman, 2015; Partridge et al., 2018). ICEs are commonly found at the 3′ ends of tRNA genes, and integration creates DR at the ends of the ICE, called *attL* and *attR* (Cameron et al., 2019; Johnson & Grossman, 2015). For the detection of transposons and integrons, there are databases available that allow the

detection of these elements in WGS, but for larger MGEs the commonly used technique is to identify the presence of MGE signatures, and evidence of phylogenetic incongruence such as highly identical copies of specific elements present in multiple species (Jiang et al., 2019); additionally, databases containing predicted and detected genomic islands or ICEs are also available (Hur et al., 2019; M. Liu et al., 2019; Yoon et al., 2015).

ESCHERICHIA COLI

E. coli is a non-spore-forming facultative anaerobic gram-negative rod-shaped bacterium, a member of the family Enterobacteriaceae. In nature, *E. coli* is principally a constituent of the mammalian gut microbiome, but it is also found, although less frequently, in the gut microbiomes of birds, reptiles, and fish, as well as in soil, water, plants, and food (Blount, 2015; Hartl & Dykhuizen, 1984; Heredia & García, 2018; Leimbach et al., 2013; Mageiros et al., 2021). Owing to their relationship with the human and animal digestive tracts, they are commonly used as an indicator of the sanitary quality of foods and water. Although generic *E. coli* is not normally the cause of serious illness, the ease with which they are grown makes their presence used as an indicator that other pathogenic organisms of fecal origin may be present (Li & Liu, 2019).

The presence of different combinations of accessory genes has shaped some *E. coli* into a variety of pathogenic strains, which are influenced by the selective pressure of their niches and the occurrence of HGT among bacterial communities with whom they cohabit (Blount, 2015; Mageiros et al., 2021). Notably, it seems that most pathogenic *E. coli* strains do not share a single evolutionary origin, rather they are the result of different DNA transfer events, and that even strains capable of causing the same disease do not constitute a monophyletic group (Shannon D. Manning et al., 2008). Hence, several

subtyping methods have helped to distinguish the strains that are involved in outbreaks and pose an increased risk, which has led to the emergence of different classification systems.

One of the most common systems used for classification is serotyping, a method that groups *E. coli* based on its antigenic variation in the surface O- (LPS) and H- (flagella) antigen caused by differences in the O-antigen gene cluster (O-AGC) and the flagellin-associated genes (Fratamico et al., 2016), although HGT contribute to the high plasticity of strains with even the same antigenic profile that may differ in the pathogenic profile due to the gain or loss of MGEs that carry genes that increase virulence (dos Santos et al., 2007; Wu et al., 2008; Yang et al., 2020). Particular sets of VFs and characteristic diseases typical of certain strains can be used to group *E. coli* into different pathotypes (Table 2); but the identification of some subclones or clades, in particular, can be difficult despite having the serotypes and VFs identified, in these cases, additional methods are usually required to increase the resolution (S. D. Manning et al., 2008). Furthermore, other phenotypic techniques including phage typing, multilocus enzyme electrophoresis, biochemical-based testing, or culture methods can provide alternatives to discern the relationship between different strains, nonetheless, they are time and labor intensive and may not be discriminatory enough (Fratamico et al., 2016). On the other hand, genetic typing methods make use of DNA fingerprinting to correlate strains in a more efficient process thus overcoming the limitations from the aforementioned techniques (CDC, 2012). For this reason, the CDC has adopted and standardized protocols of the process called Pulse Field Gel Electrophoresis (PFGE) as a gold standard in the detection and investigation of foodborne disease outbreaks caused by *E. coli* O157

and non-O157 STEC under the PulseNet network (CDC, 2012, 2016). Currently, PulseNet is in transition to WGS, where the new data obtained is stored in a combined database called *Escherichia*, which stores information on *E. coli* O157, non-O157 STEC, *Shigella* non-*flexneri* species, and *Shigella flexneri* (Tolar et al., 2019). Additionally, the Food and Drug Administration (FDA) funds a network called GenomeTrakr which also stores data obtained from WGS of *E. coli* strains (Timme et al., 2019). The information from PulseNet and GenomeTrakr is stored in the National Center of Biotechnology Information (NCBI) virtual repository along with other genomes sequenced by laboratories not certified by the CDC or FDA (Tolar et al., 2019).

Table 2. Pathogenic *E. coli* Pathotypes in Humans. * Diarrheagenic *E. coli* (Puente & Finlay, 2001).

Type of *E. coli*	Disease	Virulence factors
***Enterotoxigenic (ETEC)**	Watery to cholera-like diarrhea	Heat-labile toxin (LT), heat-stable toxin (ST), colonization factors (CFs)
***Enteroinvasive (EIEC)**	Watery diarrhea to dysentery	Ipas, type III secretion (Mxi and Spa), VirG/IcsA
***Enteropathogenic (EPEC)**	Watery diarrhea	Esps, type III secretion (Sep and Esc), intimin, Tir, and BFP
***Enterohemorrhagic (EHEC)**	Hemorrhagic colitis, hemolytic uremic syndrome (HUS)	Above EPEC factors and Shiga toxin, hemolysin
***Enteroaggregative (EAEC)**	Watery to mucoid diarrhea	AAF adhesins, EAST-1, Pet, Pic, hemolysin
Diffusely adhering (DAEC)	Watery diarrhea	F1845 and AIDA-I fimbriae

Uropathogenic (UPEC)	Urinary tract infections	Type I pili, P pili, Afimbrial adhesins (Afa), hemolysin, CNF-1
Septic (SEC)	Neonatal sepsis, meningitis	Capsule, type I pili, S-fimbrial adhesin, IbeA and IbeB (invasion proteins)

The presence of the gene encoding Shiga toxins (*stx 1* and/or *stx 2*), commonly incorporated via a prophage (lambdoid bacteriophage), causes an *E. coli* strain to be named as Shiga toxin-producing *E. coli* (STEC) or verotoxin-producing *E. coli* (Nguyen & Sperandio, 2012). This group is the most important *E. coli* at the surveillance level since it is responsible for an estimated 265,000 illnesses each year in the United States, with more than 3,600 hospitalizations and 30 deaths (CDC, 2012). Apart from bloody diarrhea, in around 5-10% of diagnosed patients, STEC can lead to hemolytic uremic syndrome (HUS), a serious complication characterized by renal failure, hemolytic anemia, and thrombocytopenia that can be fatal (Nguyen & Sperandio, 2012). The most common STEC serogroup implicated in severe illness in humans is O157, although other 400 STEC serotypes have been found, of which the most common non-O157 STEC serogroups are O26, O45, O103, O111, O121, and O145, also known as the Big 6 (CDC, 2012). Big 6 STECs and other non-O157 STECs have surpassed the number of annual infections caused by O157 STEC strains according to surveillance data collected in the U.S. (CDC, 2018). The modes by which STEC infection is transmitted in human populations include foodborne transmission, environmental transmission from contaminated animals or water, and transmission through person-to-person contact

(DuPont, 2007). Because O157 presents resistance mechanisms to acidic environments thanks to adaptations acquired when growing in the rumen of cattle, it makes it prevalent in several products with high acidity and even this benefits this group when colonizing the human digestive tract (Jones et al., 2020; Leyer et al., 1995; Price et al., 2004).

SALMONELLA ENTERICA

S. enterica is a non-spore-forming rod-shaped facultative anaerobic gram-negative bacterial species. Based on biochemical and genomic relatedness, it is divided into six subspecies: *S. enterica* subsp. *enterica*, *S. enterica* subsp. *salamae*, *S. enterica* subsp. *arizonae*, *S. enterica* subsp. *diarizonae*, *S. enterica* subsp. *houtenae*, and *S. enterica* subsp. *Indica* (Brenner et al., 2000). In total, there are more than 2600 serotypes based on the O-antigen and the H-antigen in which depending on the strain, an additional H-antigen may be present as a result of flagellar phase variation (Andino & Hanning, 2015; Brenner et al., 2000; Crump & Wain, 2017). Almost 60% of *Salmonella* serotypes belong to *S. enterica* subsp. e*nterica*, which is the only subspecies that has named serovars depending on whether they have certain antigenic profiles that meet the full antigenic definition for a serovar, whereas, other serovars from the *S. enterica* subsp. *enterica* without a defined antigenic profile as well as the remaining subspecies are named specifying the O-, and H-antigens separated by colons (Brenner et al., 2000). Human infection with most *S. enterica* produces an exudative intestinal inflammation that causes gastroenteritis also known as non-typhoidal Salmonellosis (Crump & Wain, 2017). As farm and wild animals can be reservoirs for the serovars that cause non-typhoideal Salmonellosis, these pathogens are commonly found in contaminated foods of animal origin, mainly eggs, meat, poultry and milk. However, plants such as fresh

produce can also become contaminated by exposure to manure, whereas exposure to infected pets can also be a cause of the disease (WHO, 2018). While all serovars can cause disease in humans, a few are host-specific and can reside in only one or a few animal species: for instance, *Salmonella* Dublin in cattle (Nielsen et al., 2004), *Salmonella Gallinarum* in poultry (Chaudhari et al., 2012), and *Salmonella* Choleraesuis in pigs (Leekitcharoenphon et al., 2019). When these particular serovars cause disease in humans, it is often invasive and can be life-threatening (Bäumler & Fang, 2013; Huang et al., 2019; Tanner & Kingsley, 2018).

Pathogenic islands form part of the genome of *Salmonella*, these horizontally acquired loci encode genes facilitating several virulence mechanisms, including the expression of secretion systems, fimbriae, flagella, and capsules; serotype conversion; and host colonization and subsequent survival within the host (Cheng et al., 2019; van Asten & van Dijk, 2005). Among 24 identified *Salmonella* pathogenicity islands (SPIs), only SPI-1 and SPI-2 are ubiquitously found in *S. enterica*, while SPI-22 only corresponds to *S. bongori;* whereas the remaining SPIs can be variably present among *S. enterica* or they can be only found in specific serovars (Cheng et al., 2019). SPI-1 encodes a type three secretion system (T3SS), which is essential for the export of effector proteins required for invasion of host cells (Amavisit et al., 2003; Cheng et al., 2019; Lou et al., 2019). SPI-2 encodes an additional T3SS, harboring genes that are essential for intracellular survival and for preventing acidification of the *Salmonella* containing vacuole (SCV) (Cheng et al., 2019; Marcus et al., 2000). Further virulence traits, such as the pSLT virulence plasmid, adhesins, flagella, and biofilm-related proteins, also contribute to success within the host (Cheng et al., 2019; Fàbrega & Vila, 2013; van

Asten & van Dijk, 2005). This huge armamentarium of virulence factors is under the control of an extremely complicated regulatory network, which coordinates and synchronizes all the elements involved (Cheng et al., 2019; Lou et al., 2019).

While typhoid fever and paratyphoid fever are most common in parts of the world that lack strict sanitation programs for food and water, infections with notyphoidal Salmonellosis account for just over one-fifth of all bacterial foodborne illnesses worldwide, causing an estimated 78.7 million cases per year (Havelaar et al., 2015; Jong, 2012). In the U.S., *Salmonella* is the leading cause of bacterial foodborne illness with the largest number of deaths and the largest economic losses with an annual estimate of $2.71 billion for 1.4 million cases (Andino & Hanning, 2015). The highest numbers of *Salmonella* outbreaks from the past decade are related to land animals, with poultry as the main reservoir (Andino & Hanning, 2015; Crump & Wain, 2017; WHO, 2018). More than 70% of human salmonellosis in the US has been attributed to the consumption of contaminated chicken, turkey, or eggs (Whiley & Ross, 2015). The predominant subspecies associated with severe disease is *S. enterica* subsp. *Enterica*, in which there are discrepancies between different levels of severe outcomes produced by different serovars. For example, *S. enterica* serovar Heidelberg contributes to about 7% of the *Salmonella*-related deaths in the U.S. and 11% of reported invasive infections, which are relatively high percentages considering that they generally cause less than 5% of infections (Aljahdali et al., 2020). The methods for discrimination within serovars of clinical and epidemiological importance include established tests such as phage typing and PulseNet standardized PFGE protocols, nonetheless, DNA sequencing is replacing them, either as sample sequencing such as multilocus sequence typing or increasingly

with WGS (Banerji et al., 2020; Rabsch, 2007; Tolar et al., 2019). As in *E. coli*, PulseNet has elaborated a WGS database specific for *Salmonella*, which together with the FDA-funded network, GenomeTrakr, can be accessed through NCBI along with other genomes that were obtained by laboratories not certified by either agency (Banerji et al., 2020; Tolar et al., 2019).

FOODBORNE DISEASE SURVEILLANCE

Federal and state agencies, including the U.S. Department of Health and Human Services' agencies, CDC, and the FDA, and the United States Department of Agriculture's Food Safety and Inspection Service (USDA/FSIS) cooperate to ensure safety measures are followed to protect the U.S. population (Institute of Medicine (US) Forum on Microbial Threats, 2006). Surveillance of the food supply is the integral process of searching for the pathogens that cause foodborne disease, when the surveillance occurs before consumer consumption it is called food monitoring which implies the direct detection of microbial pathogens along the food chain (Bishop & Tritscher, 2012; Institute of Medicine (US) Forum on Microbial Threats, 2006). On the other hand, if the surveillance process takes place after the people consume a contaminated product, the process is referred to as foodborne disease surveillance, which is the collection of human or animal disease data, followed by analyses of case clusters and disease trends (Bishop & Tritscher, 2012; Institute of Medicine (US) Forum on Microbial Threats, 2006). Despite the theoretically aim to provide primary prevention against foodborne disease by food monitoring, many technical challenges hamper the detection of foodborne pathogens in food, such as the limited sample size for testing (Institute of Medicine (US) Forum on

Microbial Threats, 2006; Zwietering et al., 2016), the presence of viable but non-culturable pathogens (Fakruddin et al., 2013; Năşcuţiu, 2010), or the low cell numbers in food to produce severe disease when ingested (Cooke & Slack, 2017; Doyle, 2013; Hara-Kudo & Takatori, 2011). On the other hand, foodborne disease surveillance networks overcome the sensitivity or sampling limitations of food monitoring through a continuous screening of foodborne disease cases and a rapid decision-making program (Bishop & Tritscher, 2012; Institute of Medicine (US) Forum on Microbial Threats, 2006). Although these networks cannot prevent initial cases due to the time interval between contamination event and the surveillance signal issued, they facilitate the prevention of ongoing pathogen transmission, and the identification of unforeseen problems in the food system, as well as, trends in foodborne diseases that can direct public health policymaking (Bishop & Tritscher, 2012; Institute of Medicine (US) Forum on Microbial Threats, 2006). The three most common foodborne disease surveillance strategies are: complaint or notification systems based on reports from diarrheal illnesses possibly linked to foodborne exposure, pathogen-specific surveillance, and syndromic surveillance, which, unlike complaint or notification systems, uses non-specific health data (Institute of Medicine (US) Forum on Microbial Threats, 2006).

Among different networks and resources for food safety (see Table 3), the PulseNet network created by the CDC is a powerful molecular subtyping network consisting of state public health laboratories in all 50 states and food regulatory laboratories within the FDA and USDA designed to identify and facilitate investigation of foodborne disease outbreaks (CDC, 2019a). As evidence of its effectiveness, from the time it was implemented it has been possible to perceive a reduction of reported illnesses

due to improved information, enhanced industry accountability, and more rapid recalls; furthermore, economic impacts attributable to PulseNet include medical costs and productivity losses averted due to reduced illness (Boxrud et al., 2010; Tolar et al., 2019). Scharff et al. (2016) estimated that because of the prevention of foodborne diseases resulting from PulseNet surveillance, a reduction of medical and productivity costs by $507 million was achieved in the period between 1994 and 2009. The PulseNet system is currently used to track nine organisms by use of standardized PFGE protocols (*E. coli* O157, non-O157 STEC, non-flexneri *Shigella* species, *Shigella flexneri*, *Vibrio cholerae*, *Vibrio parahaemolyticus*, *Salmonella*, *Listeria*, or *Campylobacter*) (Institute of Medicine (US) Forum on Microbial Threats, 2006). Potentially, PulseNet can be used to track any infectious disease confirmed by detection of a specific microorganism (CDC, 2019a; Institute of Medicine (US) Forum on Microbial Threats, 2006).

Table 3. Networks and Resources for Food Safety in the U.S. (Institute of Medicine (US) Forum on Microbial Threats, 2006)

Acronym	Program Name
Biosense	Biosense (CDC)
CAHFSE	Collaboration in Animal Health and Food Safety Epidemiology (USDA; APHIS, ARS, FSIS)
CaliciNet	CaliciNet (CDC)
eFORS	Electronic Foodborne Outbreak Reporting System (CDC)
eLEXNET	Electronic Laboratory Exchange Network
eLRN	Environmental Laboratory Response Network
Epi-X	Epidemic Information Exchange

Essence	Electronic Surveillance System for Early Notification of Community-based Epidemics
FERN	Food Emergency Response Network
FoodNet	Foodborne Disease Active Surveillance Network
GEMS	Global Environmental Monitoring System
GenomeTrakr	GenomeTrakr (FDA)
Global Salm-Surv	Global Salmonella Survey (WHO)
GOARN	Global Outbreak Alert and Response Network
GPHIN	Global Public Health Intelligence Network
HAN	Health Alert Network
ICLN	Integrated Consortium of Laboratory Networks
IDSA-EIN	Infectious Disease Society of America Emerging Infections Network
INFOSAN	International Food Safety Authorities Network
LRN	Laboratory Response Network
NAHSS	National Animal Health Surveillance System
NARMS	National Antibiotic Resistance Monitoring System
NEDSS	National Electronic Disease Surveillance System
NETSS	National Electronic Telecommunications System for Surveillance
NPDN	National Plant Diagnostic Network
NRDM	National Retail Data Monitor
PulseNet	PulseNet (CDC)
RASFF	Rapid Alert System for Food and Feed
RODS	Real-time Outbreak and Disease Surveillance

TEPHINET	Training Programs in Epidemiology and Public Health Interventions Network
UNEX	Unexplained Death and Serious Illness

PFGE is a genotyping technique used for the separation of large DNA molecules including genomic DNA, after treating them with unique restriction enzymes and letting the reaction product migrate in a gel matrix under the electric field that periodically changes direction (Sharma-Kuinkel et al., 2016). In comparison to other genotyping methods, PFGE provides a good representation of the entire bacterial chromosome in a single gel with a highly reproducible restriction profile, providing distinct and well-resolved DNA fragments (Sharma-Kuinkel et al., 2016; Tang et al., 2019; WHO, 2009). The images of the PFGE patterns are electronically transferred to CDC, where they are analyzed, this enables rapid subtyping and comparison of PFGE patterns of bacteria isolated from ill persons, food, veterinary and environmental sources across the country, and the detection of clusters of cases with identical patterns to indicate that an outbreak might be occurring (Sharma-Kuinkel et al., 2016). Despite the introduction of PFGE to PulseNet revolutionized the detection, investigation, and control of outbreaks over the past two decades, it has inherent limitations for molecular characterization and subtyping of bacterial pathogens owing to its inability to infer phylogenetic relationships caused by the lack of power to resolve relationships between unrelated isolates with identical or nearly identical PFGE banding patterns (Bergholz et al., 2016; Oakeson et al., 2018).

Since the advent of NGS and the drastic decrease in its cost in recent years, numerous technical barriers related to Sanger sequencing have been overcome, such as sequencing speed, read length, throughput and especially cost, making it possible not

only at the detection of mutations in single base pairs but also at the whole genome scale, as well as the identification of key genes involved in the regulation of complex phenotypes (Li et al., 2019; Morozova et al., 2009; van Dijk et al., 2014). Consequently, WGS captured the interest from the food safety community due to the increased resolution in terms of foodborne pathogens subtyping and their molecular characterization, thus, the FDA's Center for Food Safety and Applied Nutrition (CFSAN) established in 2013 (Allard et al., 2016), the first integrated network of state and federal laboratories to use WGS to track foodborne pathogens to improve outbreak response activities related to FDA compliance and regulatory programs by providing more precise scientific traceback, and a publicly available global database containing the genetic makeup of thousands of foodborne disease-causing bacteria from food and environmental sources stored at the NCBI (Allard et al., 2016; Brown et al., 2019; Tolar et al., 2019). GenomeTrakr network is made up of 14 federal laboratories (including USDA-FSIS food laboratories), and state agriculture, food, environmental, and public health laboratories in 14 states, 1 U.S. hospital laboratory, and 9 international laboratories (including laboratories from Mexico, Argentina, and England) (Brown et al., 2019). In 2013, the PulseNet together with its partners at FDA, FSIS, and NCBI, and state laboratories participating in GenomeTrakr and PulseNet, launched a pilot project for sequencing and analyzing isolates of *L. monocytogenes* in real-time, in parallel with the current PFGE-based surveillance. This collaborative effort demonstrated that the application of WGS to laboratory surveillance contributes higher precision and resolution than PFGE, and as a result, more outbreaks could be detected, investigated, and controlled than ever before, and in addition, WGS could also identify false outbreak signals produced by PFGE

(Jackson et al., 2016). Thereby, PulseNet transitioned to WGS as the primary subtyping tool for surveillance of *Listeria*, *Salmonella*, *E. coli*, *Shigella,* and *Campylobacter* in 2019, thus merging PFGE and WGS databases for these organisms; although the use of WGS as the primary subtyping tool of *Vibrio*, *Yersinia*, and *Cronobacter* is not yet validated or established (Tolar et al., 2019).

WHOLE GENOME SEQUENCING (WGS)

With the implementation of modified nucleotides called dideoxy-nucleotides (dNTPs) to the amplification process of new DNA strands in which the ribose 3′-OH group is blocked, thus preventing elongation, the arrival of the first generation of sequencing by the hand of Sanger sequencing technology opened the door to the opportunity to characterize the genetic material of different species with even higher resolution than what was achieved with DNA banding pattern-based genotyping methods (Heather & Chain, 2016). Despite the Sanger sequencing was widely used for three decades and even today for single or low-throughput DNA sequencing, its limitations pose the difficulty of further improving the low throughput that does not allow the sequencing of complex genomes, in addition to which it can be significantly more expensive and slow compared to newer technologies (Canadian Agency for Drugs and Technologies in Health, 2014; Goodwin et al., 2016). The appearance of the second-generation sequencing facilitated the analysis of the entire genomic DNA sequence of a cell by reducing sequencing speed and costs while maintaining high accuracy, as well as, yielding high-throughput data, thus promoting the wide use of WGS for the study of different living organisms at a deeper level (Churko et al., 2013; Goodwin et al., 2016). In the food safety field, the high discriminatory power of WGS compared with traditional

molecular typing tools promotes its use as a tool for foodborne illness surveillance. Whereas, the microbiological testing of foods performed frequently in the food industry focuses more on the detection of well-characterized foodborne pathogens, thereby, a higher level of characterization is commonly omitted, leaving traditional techniques still preferred, unless it is necessary to assess the safety of probiotics or starter cultures (Chokesajjawatee et al., 2020; Lee et al., 2021; Surachat et al., 2021), as well as to track and trace the source of contamination (Jagadeesan et al., 2019).

Currently, two strategies have been developed in second generation sequencing: sequencing by ligation (SBL) and sequencing by synthesis (SBS). In SBL approaches (SOLiD and Complete Genomics), a probe sequence that is bound to a fluorophore hybridizes to a DNA fragment and ligates to an adjacent oligonucleotide for image capture. Then, the emission spectrum of the fluorophore indicates the identity of the complementary base(s) at specific positions within the probe (Goodwin et al., 2016). In SBS approaches (454 Roche pyrosequencing, Ion Torrent, Illumina and GeneReader), a polymerase is used and a signal, such as a fluorophore or a change in ionic concentration, identifies the incorporation of a nucleotide into an elongating strand (Goodwin et al., 2016). Over time, Illumina positioned itself as the predominant platform in short-read Next Generation Sequencing owing, in part, to its high level of cross-platform compatibility, good administrative decisions that allowed the constant improvement of technology, and a variety of equipment that meets different needs in the market (Dervan & Shendure, 2017; Goodwin et al., 2016; Sabino, 2020). Nonetheless, the advent of third generation sequencing has brought technologies that can offer longer reads, which in spite of the higher accuracy of short-read technologies, reads spanning thousands of

nucleotides can elucidate the complex structure of some regions in the genome thus eliminating ambiguity in the positions or size of genomic elements (Goodwin et al., 2016). This is how the Oxford Nanopore Technologies (ONT) and PacBio platforms have gained popularity in studies seeking to perform the highest resolution typing possible between closely related bacteria of the same species and those studying broad genomic arrangement (C. Li et al., 2020; Moss et al., 2020; Uelze et al., 2020), nonetheless, ONT offers fast and portable technologies that make it ideal for genomic sequencing applications in the surveillance of foodborne pathogens (Jagadeesan et al., 2019; Logsdon et al., 2020).

Table 4. Summary of commonly used Whole Genome Sequencing platforms (Jagadeesan et al., 2019**).**

Platform	**Sequencing technology**	**Read length**	**Output/run**	**Error rate**	**Example of use**	**Type of instrument and run time**
Illumina	Sequencing by synthesis	Short reads 1 × 36bp – 2 × 300bp	0.3–1000Gb	Low	Variant calling	Benchtop 2–29 h
PacBio	Single molecule	Long reads	0.5–10Gb	High	De novo assembly of	Large scale

	sequencing by synthesis	Up to 60kb			small bacterial genomes and large genome finishing	0.5–4 h
Oxford Nanopore	Single molecule	Long reads	0.1–20Gb	High	The complete genome of isolates and metagenomics	Portable
		Up to 100kb				1min-48 h

Illumina

Illumina sequencing technology is based on cyclic reversible termination (CRT), sequencing by synthesis approach that makes use of terminator molecules that are similar to those used in Sanger sequencing (Goodwin et al., 2016). For the preparation of the DNA to be sequenced, first, it has to be fragmented via mechanical or enzymatic shearing. Next, adaptors and barcodes sequences are ligated to the DNA fragments and then loaded to the flow-cell coated with primers complementary to the adaptor sequences, wherein exclusion amplification creates clonal clusters in each well from the individual library molecules (Illumina, 2013). The strands that were anchored to the wells are primed by a sequence that is hybridized to an adapter region, thus initiating polymerase binding to this double-stranded DNA (dsDNA) region. In each cycle, a mixture of all four individually labeled and 3′-blocked dNTPs are added, which will be incorporated one by one in the elongation process, so that later the nucleotides that did not bind to the template strand are removed, allowing subsequent excitation and imaging of the dNTPs incorporated at each cluster through total internal reflection fluorescence (TIRF)

microscopy using either two or four laser channels (Goodwin et al., 2016). Generally, each dNTP is bound to a single specific fluorophore for each nucleotide and requires a 4-channel system, whereas the NextSeq and Mini-Seq systems use a two-fluorophore system, thus only requiring a 2-channel system (Table 5) (Goodwin et al., 2016; Illumina, 2015). Implementation of the 2-channel system results in a faster sequencing process, although since mixing two light signals is commonly used to detect adenine, the phasing effect due to improper removal of blockers can more rapidly increase the error rate in comparison to 4-channel systems (ecSeq, 2017). Finally, the fluorophore and blocking group can be removed and a new cycle can begin. Base calling is obtained directly from the signal intensity measurements during each cycle, furthermore, Illumina has the option of generating single ended (SE), or paired-ended (PE) reads in which both ends of the anchored DNA fragment are sequenced (Illumina, 2013).

The suite of instruments offered by Illumina includes short-read sequencers with high precision (~ 99.9%), ideal for a varied range of yield requirements, from small low-throughput benchtop units to large ultra-high-throughput instruments dedicated to WGS at the population level (Table 5) (Wentz et al., 2019). Illumina has become widely used for the genotyping of strains suspected of being involved in an outbreak of foodborne diseases, in which detection of Single Nucleotide Polymorphisms (SNPs) and structural variations in their genomes can elucidate their relationship with already characterized pathogens (Brown et al., 2019), with MiSeq being the standardized platform for PulseNet and GenomeTrakr networks (Timme et al., 2020; Timme et al., 2018). Furthermore, gene expression and transcriptome analysis facilitate the characterization of all transcriptional activity (coding and non-coding) of microorganisms present in food, which can be useful

not only to broaden the understanding of the physiology of the diseases that pathogenic strains cause but also to decipher other complex biological processes, such as AR, food spoilage and biofilm formation (Puttamreddy et al., 2008; Sabino et al., 2019). As mentioned above, due to size limitations, short-read sequencing platforms have issues sequencing complex or highly repetitive regions of the genome, which constitutes a major challenge for *de novo* sequencing of bacterial genomes, because they contain up to several dozens of intragenic and intergenic tandem repeats (Adewale, 2020; Alkan et al., 2011). Regions that can be much longer than the maximum read length and the insert size of PE tags (Kuśmirek & Nowak, 2018). In addition, characterizing these highly repetitive regions and their instability can provide essential hints about the modulation of the function of specific genes that may be involved in bacterial adaptation to a new environment in a short term without complicated mutation (Kuśmirek & Nowak, 2018; Zhou et al., 2014).

Table 5. Illumina platforms specifications (Wentz et al., 2019).

Platform	Max read length (bp)	Type of Chemistry	Max reads produced	Max output	WGS applications
MiniSeq	2× 150 bp	2-channel SBS	25 million	7.5 Gb	Viruses, bacteria, small eukaryotes/targeted sequencing
NextSeq 500/550	2× 150 bp	2-channel SBS	800 million	100–120 Gb	Virus, bacteria, eukaryote

MiSeq	2× 300 bp	4-channel SBS	44–50 million	13.5–5 Gb	Virus, bacteria, small eukaryote/targeted sequencing
HiSeq 2500	2× 250 bp	4-channel SBS	4 billion	1000 Gb	Virus, bacteria, eukaryote

Oxford Nanopore Technology

Unlike PacBio, which is a long-read sequencing platform based on SBS, with the help of electrolytic solutions and the application of a constant electric field, ONT uses electrophoresis to mobilize native DNA molecules through a nanopore, which is connected to a motor protein that unzips the double stranded DNA at a Y adapter added during the library preparation step, directing just one strand at a time through the nanopore (Goodwin et al., 2016; Nygaard et al., 2020). The passage of the molecule to be sequenced through this nanopore blocks the flow of ions, which reduces the current for a length of time proportional to the size of the different nucleotides (Nygaard et al., 2020). Consequently, the change in the current pattern and magnitude is measured, providing a signal to be used for base calling (Deamer et al., 2016). Theoretically, sequencing continues until the end of the DNA fragment or until the pore becomes physically blocked, hence, during the library preparation step DNA molecules can be fragmented or kept intact in the case of seeking to obtain reads as long as possible (Canadian Agency for Drugs and Technologies in Health, 2014; Deamer et al., 2016; Nygaard et al., 2020). Furthermore, in order to improve accuracy, ONT uses a leader–hairpin structure formed during the library preparation process, which allows the forward DNA strand to pass

through the pore, followed by the reverse strand. This generates reads from both strands, also known as 1-dimensional (1D) reads with a ~20.19% error rate, from which a consensus sequence can be generated resulting in a 2-dimensional (2D) read with an error rate of ~13.40% (Cao et al., 2017; Weirather et al., 2017). ONT also offers a variety of equipment that adapts to the different needs of the market (Table 6), although MinION in particular is attracting interest for pathogen surveillance and diagnostics owing to the low investment cost required for its implementation and its portability (Goodwin et al., 2016). Despite the fact that ONT sequences still have notably higher error rates compared with second-generation sequencing platforms, it is expected that the precision will continue to increase due to constant development of improvements in its chemistry, as well as the active research of new base callers (Deamer et al., 2016; Nygaard et al., 2020). Apart from the aforementioned utility of long reads for *de novo* assembly, ONT has proven to be effective for the detection and differentiation of methylations resulting from epigenetic alterations in bacterial genomes (Dumschott et al., 2020), mechanisms that can be important to identify AR in bacteria (Fernández et al., 2011; D. Ghosh et al., 2020; Motta et al., 2015).

Table 6. Oxford Nanopore Technology (ONT) platforms specifications (ONT; van Dijk et al., 2014; Wentz et al., 2019).

Platform	Max read length	Max reads produced	Max Output	WGS applications
Flongle	Nanopores read the length	Read length dependent	2.8 Gb	Viruses, bacteria, targeted sequencing

	of DNA		Virus, bacteria, small
MinION	presented to	50 Gb	eukaryote/targeted
	them. Longest		sequencing
GridION	read so far: > 4	250 Gb	Virus, bacteria, eukaryote
	Mb.		
PromethION		14 Tb	Virus, bacteria, eukaryote

BIOINFORMATICS FOR WHOLE GENOME SEQUENCING ANALYSIS

The advances of WGS have resulted in a milestone in the resolution for surveillance and outbreak investigations, source attribution, genomic studies, as well as genomic information for phenotypic prediction (Uelze et al., 2020). The evolution of sequencing technologies was not the only reason for this unprecedented event, since an evolution alongside bioinformatics tools and computational resources has been fundamental to manage and analyze large amount of information that are commonly yielded in NGS platforms (Uelze et al., 2020). Among the most important steps to analyze sequencing data obtained from WGS are: quality control of reads, reads mapping, detection of allelic variants, genome assembly, genome annotation and WGS phylogenetic analyzes (Mohammed & Thapa, 2020; Wadapurkar & Vyas, 2018). User-friendly bioinformatics tools for high-throughput sequencing data analysis are available to be installed in Windows, such as the commercial systems BioNumerics, a comprehensive software package, which has many applications in different research fields of the biological sciences (Applied Maths, Sint-Martens-Latem, Belgium) and Ridom SeqSphereC a tool for automatic processing and analyzing of NGS sequence data which can be used for whole genome microbial typing or traditional Multiple Locus

Sequence Typing (MLST) projects (Ridom GmbH, Munster, Germany); or found as web services, such as Rapid Annotation using Subsystem Technology (RAST) a fully-automated service for annotating complete or nearly complete bacterial and archaeal genomes (Aziz et al., 2008), MG-RAST an automated server for phylogenetic and functional analysis of metagenomes (Meyer et al., 2008), Pathosystems Resource Integration Center (PATRIC) an information system designed to support biomedical work on bacterial infectious diseases through advanced searches based on specific pathogenic bacteria related information (Wattam et al., 2013), and EnteroBase an integrated software environment that supports the identification of global population structures within various bacterial genera including pathogens (Alikhan et al., 2018). However, these tools might not be flexible in terms of parameter modifications and require access accounts/licenses (Quijada et al., 2020). Furthermore, they are also limited to be used in conventional systems; or in the case of web based systems, the user depends on the availability of the computational resources of the external servers where the tools are allocated (Quijada et al., 2020). Therefore, most bioinformatic analyses are performed via command-line on UNIX operative systems, which provides more versatility and enable the optimal usage of the computational resources available (Quijada et al., 2020; Uelze et al., 2020). Nonetheless, command-line based tools have their limitations, since learning their use can be challenging for personnel who do not have experience in handling systems through command-line or in the use of different programming languages, such as R, Python, Perl, and Bash (Quijada et al., 2020). Consequently, pipelining is one of the main goals in bioinformatics, as it can create handy shortcuts to easily use multiple tools, as well as allowing automation and parallelization of the

analyses (Jagadeesan et al., 2019; Quainoo et al., 2017). Most of the bioinformatics software in development is available in open-source repositories such as Conda/Anaconda (https://anaconda.org/), GitHub (https://github.com) or SourceForge (https://sourceforge.net), so that the user can have free access to the tools generally developed by the bioinformatics community (Quijada et al., 2020).

Sequencing quality assessment and genome assembly

As the quality of the reads varies throughout the sequencing process, a quality score known as the Phred or Q score is used to estimate the probability that an error was made for each nucleotid**e.** Q scores are often represented as ASCII characters, so different reads will be made up of nucleotides with different score values (Illumina, 2011). These scores are used by tools to show graphically the quality and read length distributions, as well as counts of over-represented k-mers among the yielded reads for quality control of the sequencing process, among these tools we can find FastQC (https://github.com/s-andrews/FastQC) and PRINSEQ (Schmieder, 2013), but since these methods have not been fully optimized for characterizing long error-prone reads other tools have been created, such as LongQC (Fukasawa et al., 2020). Reads trimming, removal of adapters, and filtering based on quality are performed by platform specific tools, such as Trimmomatic (Bolger et al., 2014) and SolexaQA (Schmieder & Edwards, 2011) for Illumina; or Porechop (https://github.com/rrwick/Porechop), Filtlong (https://github.com/rrwick/Filtlong) and NanoFilt (De Coster et al., 2018) for ONT. Del Fabbro et al. (2013) confirmed the positive effect of trimming reads which increased the quality and reliability from SNP calling and genome assembly, although, it is worth

bearing in mind that for different platforms the trimming and filtering parameters are different due to the characteristics of the reads produced by contrasting technologies (Utturkar et al., 2017). Additional tools for long-reads accuracy correction have been developed, in which the main strategy is to find consensus corrections using read overlaps among the same long-reads dataset, such as the approach used by Canu (S. Koren et al., 2017) or with the aid of short-reads, such as proovread (Hackl et al., 2014).

Assembly algorithms are implemented to arrange reads into larger sequences (contigs), and this longer arrangement of reads can eventually be chained together in a process called scaffolding, wherein the contigs are joined typically employing a reference genome to obtain additional information on their relative position and orientation in the genome (PacBio, 2021; Quijada et al., 2020). However, scaffolding can lose critical information made up of missing gaps and can be misleading about the true gap size. Additionally, the gap-flanking scaffold sequences can be of low quality due to the presence of homopolymers that stopped the sequencing early or due to read length limitations (PacBio, 2021). For short reads, the preferred tools are de Bruijn graph-based (DBG) assemblers, because they break down original short reads into smaller sequences called *k*-mers, which are further reduced into *k*-1-mers. Subsequently, these fragments are joined via an Eulerian walk, which is the shortest possible path through these *k*-1-mers, thus decreasing the chance of an incorrect assembly of repeat regions (Quijada et al., 2020). Some of the assemblers that use this strategy and are tailored for bacterial chromosomes assembly are Velvet (Zerbino, 2010), Ray (Boisvert et al., 2010) and SPAdes (Bankevich et al., 2012), which is used by the Bionumerics distribution for bacterial genomic assembly. Due to the principle of this strategy, DBG assemblers are

dependent on high-quality reads thus limiting their use to short-reads platforms (Deamer et al., 2016). Therefore, due to the higher error rate of long-reads, a different strategy, known as overlap layout consensus (OLC), is implemented to assemble longer reads. The principle of OLC relies on finding overlaps between reads and producing contigs (Quijada et al., 2020). This algorithm is applied by most long-read assemblers, including Raven (Vaser & Šikić, 2020), Miniasm (Li, 2016), NextDenovo (https://github.com/Nextomics/NextDenovo), HINGE (Kamath et al., 2017), and Canu (Sergey Koren et al., 2017); with the exception of Flye which aims to produce repeating graphs in order to resolve not bridged contigs (Kolmogorov et al., 2019). Currently, Raven, Flye, Miniasm, and NextDenovo outperform other long-reads assemblers for prokaryotes, nonetheless, they are not yet free of specific limitations for each tool (Chen et al., 2020; Wick & Holt, 2021). Additionally, reads obtained from different platforms can be combined in order to obtain better quality assemblies, although resulting in higher sequencing costs (Goldstein et al., 2019; Quijada et al., 2020). Short-reads can be directly used to polish assemblies constructed with long-reads using Pilon (Goldstein et al., 2019), or some assemblers can use long-reads to close gaps between contigs generated using short-reads, such as hybridSPAdes (Antipov et al., 2016) or Unicycler (Wick et al., 2017).

Quality assessments for genome assemblies are based on metrics that evaluate the length of the assemblies and the annotation capacity that can be achieved using them (Manchanda et al., 2020). As for length metrics, N50 and L50 are commonly used to describe assembly contiguity, which represents the sequence length of the shortest contig at 50% of the total genome size and the smallest number of contigs whose length sum

makes up half of the genome size, respectively (Gurevich et al., 2013; Manchanda et al., 2020). Whereas, the accuracy of a genome can be assessed by annotation quality metrics which include a number of gene models, exons per gene model, and the average lengths of genes, exons, and transcripts (Yandell & Ence, 2012). However, genomic completeness is better estimated using a set of genes that are universally distributed as orthologs across particular clades of species, such as the strategy utilized by the Benchmarking Universal Single-Copy Orthologs (BUSCO) (Simão et al., 2015), a tool that provides a summary of complete single-copy, duplicated, fragmented, and missing housekeeping genes distributed in a specific clade annotated in an assembly.

Genome Characterization

Sequence alignment plays an essential role in genomics characterization. (Li & Homer, 2010). While finding how similar a sequence is to another and quantifying the degree of similarity, aligning two or multiple sequences can transfer already known annotations to newly generated sequences (Clausen et al., 2018; Li & Homer, 2010). Hence, WGS of foodborne pathogens can yield sufficient data to be aligned against databases containing genes relevant for their typing as well as others that are involved in virulence or AR (Kleinheinz et al., 2014). Currently, there are established pipelines for the annotation of bacterial genomes, a process that can be automated since the lack of post-transcriptional RNA modifications in bacteria allows the open reading frames to be easily assigned, unlike genomes from eukaryotic organisms where the process is usually more extensive (Ejigu & Jung, 2020). Two of the major tools used to annotate bacterial genomes are PROKKA (Seemann, 2014) and RAST (Aziz et al., 2008), which are used

with command-line and via online respectively. Both tools follow a similar process where genetic elements inthe genome are predicted and then compared with curated databases (Aziz et al., 2008; Seemann, 2014). Bacteria have dedicated curated databases that can contain reference genomes that meet standards for sequence quality, completeness, and freedom from contamination, such as RefSeq (Haft et al., 2018). Furthermore, there are databases dedicated to containing single or cluster of genes of interest that are annotated based on their function, including serotype specific genes (Banerji et al., 2020; Ingle et al., 2016), subtype determinants (Orlek, Phan, et al., 2017; Tang et al., 2019), virulence factors (Joensen et al., 2014; B. Liu et al., 2019) or AR genes (Alcock et al., 2019; Bortolaia et al., 2020), as well as databases harboring complete operons such as PADS arsenal, a database containing elements of the immune response of bacteria against viruses or heterologous DNA (Zhang et al., 2020), or entire mobile genetic elements such as the ACLAME database (Leplae et al., 2010). Indeed, there are several available databases for the same purpose that have unequal content, this could be detrimental for the consistency of similar studies performing the characterization of specific traits. For instance, in a comparative study of available resources for the detection of determinants involved in AR performed by Xavier et al. (2016), it was possible to verify that there are significant impacts on the results obtained from different databases thus demonstrating that the use of databases is limited to the information stored in them, which can be complemented in some cases with the addition of new sequences (Passarelli-Araujo et al., 2019).

Traditional methods for the characterization of genomes utilize the Basic Local Alignment Search Tool (BLAST) in assembled genomes, such as AR-ANNOT (Gupta et

al., 2014) or the former versions of ResFinder and VirulenceFinder (Kleinheinz et al., 2014). Although there are also tools that make use of BLAST but also apply a more elaborate process, such as the VFDB search tool that makes use of partitioned databases with multiple BLAST searches to elucidate not only well-characterized virulence factors but also predicted virulence genes (B. Liu et al., 2019). Nonetheless, the drawbacks of BLAST based tools rely on the necessity of high quality assemblies, as high-throughput sequencing produces large datasets that pose a challenge in terms of time and computational resources when using BLAST, in addition to the presence of gaps within the assembled chromosome due to a large number of repetitive regions that can lead to missing data (Clausen et al., 2016; Clausen et al., 2018; Inouye et al., 2014). On the other hand, read mapping overcomes those limitations, but previously read mappers were more focused on mapping in reference genomes and not against databases storing sequences of determinants for specific virulence factors or AR genes, which are mostly made up of groups of different sequences that can share more than 95% of similarity (Clausen et al., 2018; Scheutz et al., 2012). Whereby, most commonly used reads mappers, such as Bowtie2 (Langmead & Salzberg, 2012), BWA-MEM (https://github.com/lh3/bwa), or Minimap2 (Li, 2018), choose a random hit when there is a tie for the best match among similar determinants of a gene in those redundant databases. However, Clausen et al. (2018) developed KMA, which implements a novel sorting scheme to solve scenarios with hits that have similar scores of similarity, thus opening the possibility of using reads directly to detect different determinants. For the detection of other more extensive and complex genetic elements, it is necessary to execute strategies that combine sequence alignments with more complex analysis (Quijada et al., 2020). For instance , prediction of

prophages in the genome of a bacterium, prophage prediction tools, apart from carrying out searches for proteins homologous to the proteins housed in dedicated databases, usually include density calculations of elements related to phages in a determined region of the genome in order to confirm that it is a potential prophage (Akhter et al., 2012; Arndt et al., 2016; Lima-Mendez et al., 2008; Reis-Cunha et al., 2019), or they can also combine sequence similarity-based matching and genetic features-based machine learning classifications (Song et al., 2019).

Phylogenomics

In order to determine the relatedness among strains, there are two main strategies being used: SNP-based and gene by gene-based approaches (Uelze et al., 2020). The expected final results from those types of analyses can be either matrixes containing SNPs as well as alleles information, or phylogenetic trees (Quainoo et al., 2017; Uelze et al., 2020). In the SNP based approach, sequencing reads can be mapped directly to a reference genome that has to be as contiguous and complete as possible, and the genetically closelly related to the genomes being analyzed (Jagadeesan et al., 2019). SNPs calling relative to the reference genome are performed to each isolate, and thereafter, the identified variants are used to quantify the genetic relatedness between strains (Uelze et al., 2020). Some of the tools available for SNP calling are SAMtools (https://github.com/samtools/), GATK (Heldenbrand et al., 2019), and Freebayes (https://github.com/freebayes/freebayes). Moreover, there are specialized pipelines for SNP calling from bacterial genomes, such as Snippy (https://github.com/tseemann/snippy), CFSAN SNP pipeline (Davis et al., 2015), NASP (Sahl et al., 2016), and BactSNP (Yoshimura et al., 2019); nonetheless,

owing to tool-to-tool variability in SNPs calling of foodborne pathogens, it is recommended to use previously validated SNP-based tools, such as those developed by the FDA or the CDC (Timme et al., 2017). Additionally, the main limitation of SNP-based methods is the need for a highly related reference genome, which if it is too distant from the studied isolates, fewer reference positions will be covered and, subsequently, fewer SNPs will be discovered. Furthermore, if among the isolates analyzed there are one or more remotely linked isolates, the core SNPs that could be identified will be reduced (Quainoo et al., 2017; Uelze et al., 2020). However, kSNP3, a K-mer based tool, attempts to overcome those limitations by detecting core SNPs between strains without the need for a reference (Gardner et al., 2015). As evidence of its efficiency, kSNP3 has been successfully applied for retro-perspective outbreak detection (Carroll et al., 2019; Mercante et al., 2016).

On the other hand, gene by gene analysis assesses the variation in delimited genes of a draft bacterial genome (Maiden et al., 2013). With a similar model to the used in the traditional 7-loci multi-locus sequence typing (MLST), the genes in either a defined core genome (cgMLST) or the whole genome (wgMLST), which includes the accessory genes of the analyzed isolates, are compared against a reference database of all known alleles for a particular species. To assign an MLST type, the assembled reads are compared using BLAST to a reference allele database, also known as MLST scheme, which has all characterized allelic variants for each locus defined for a specific species like the enterobase cgMLST scheme (http://enterobase.warwick.ac.uk) (Quainoo et al., 2017; Uelze et al., 2020). Variations, including SNPs, indels and recombinations in the same gene are deemed as a single allele difference (Uelze et al., 2020). A number is assigned to

each gene or allele sequence, which allows the genomes to be compared based on the number of allele differences that exist, thereby the sum of differently assigned allele numbers between a pair of samples determines the allele difference providing the data to create allele distance matrixes of a set of samples (Quainoo et al., 2017; Uelze et al., 2020). A major advantage of cg/wgMLST over the SNP-based approach is that it can be standardized and harmonized by using unique MLST schemes (Uelze et al., 2020). Conversely, an allele difference between two strains may be explained by one or several mutations, thus indicating the intrinsically higher discriminatory power of SNP analyses (Jagadeesan et al., 2019).

Finally, several algorithms to perform phylogenetic analyses via either Bayesian methods or ML methods are available (Holder & Lewis, 2003). These phylogeny algorithms are capable to model the evolutionary signal more accurately than neighbor-joining and parsimony methods, although they can be computationally demanding (Williams & Moret, 2003). Various models of nucleotide substitution are applied in phylogenetic analysis, which attempts to simplify the actual evolutionary signal, such as the general time reversible model (GTR), a model often used for inferring phylogeny from nucleotide and SNP data (Quainoo et al., 2017). Phylogenetic inferences based on the core SNP alignment can provide more detailed evolutionary models; therefore, in practice, SNP analyses may be applied after defining a potential phylogenetic cluster after pre-clustering with cgMLST (Tolar et al., 2019; Uelze et al., 2020).

REFERENCES

Abouzeed Yousef, M., Baucheron, S., & Cloeckaert, A. (2008). ramR Mutations Involved in Efflux-Mediated Multidrug Resistance in Salmonella enterica Serovar Typhimurium. Antimicrobial Agents and Chemotherapy, 52(7), 2428-2434. https://doi.org/10.1128/AAC.00084-08

Adewale, B. A. (2020). Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years? African journal of laboratory medicine, 9(1), 1340-1340. https://doi.org/10.4102/ajlm.v9i1.1340

Ahmad, A., Patel, I., Khan, M. U., & Babar, Z. u.-d. (2017). Pharmaceutical waste and antimicrobial resistance. The Lancet Infectious Diseases, 17(6), 578-579. https://doi.org/10.1016/S1473-3099(17)30268-2

Akhter, S., Aziz, R. K., & Edwards, R. A. (2012). PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic acids research, 40(16), e126-e126. https://doi.org/10.1093/nar/gks406

Akhter, S., Aziz, R. K., & Edwards, R. A. (2012). PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Research, 40(16), e126-e126. https://doi.org/10.1093/nar/gks406

Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A.-L. V., Cheng, A. A., Liu, S., Min, S. Y., Miroshnichenko, A., Tran, H.-K., Werfalli, R. E., Nasir, J. A., Oloni, M., Speicher, D. J., Florescu, A., Singh, B., Faltyn, M., Hernandez-Koutoucheva, A., Sharma, A. N., Bordeleau, E., Pawlowski, A. C., Zubyk, H. L., Dooley, D., Griffiths, E., Maguire, F., Winsor, G. L., Beiko, R. G., Brinkman, F. S. L., Hsiao, W. W. L., Domselaar, G. V., & McArthur, A. G. (2019). CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic acids research, 48(D1), D517-D525. https://doi.org/10.1093/nar/gkz935

Alikhan, N.-F., Zhou, Z., Sergeant, M. J., & Achtman, M. (2018). A genomic overview of the population structure of Salmonella. PLOS Genetics, 14(4), e1007261. https://doi.org/10.1371/journal.pgen.1007261

Aljahdali, N. H., Sanad, Y. M., Han, J., & Foley, S. L. (2020, 2020/11/17). Current knowledge and perspectives of potential impacts of Salmonella enterica on the profile of the gut microbiota. BMC Microbiology, 20(1), 353. https://doi.org/10.1186/s12866-020-02008-x

Alkan, C., Sajjadian, S., & Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. Nature Methods, 8(1), 61-65. https://doi.org/10.1038/nmeth.1527

Allard, M. W., Strain, E., Melka, D., Bunning, K., Musser, S. M., Brown, E. W., & Timme, R. (2016, Aug). Practical Value of Food Pathogen Traceability through Building a Whole-Genome Sequencing Network and Database. J Clin Microbiol, 54(8), 1975-1983. https://doi.org/10.1128/jcm.00081-16

Allen, H. K., Donato, J., Wang, H. H., Cloud-Hansen, K. A., Davies, J., & Handelsman, J. (2010, 2010/04/01). Call of the wild: antibiotic resistance genes in natural environments. Nature Reviews Microbiology, 8(4), 251-259. https://doi.org/10.1038/nrmicro2312

Allen, J. P., Ozer, E. A., Minasov, G., Shuvalova, L., Kiryukhina, O., Anderson, W. F., Satchell, K. J. F., & Hauser, A. R. (2020). A comparative genomics approach identifies contact-dependent growth inhibition as a virulence determinant. Proceedings of the National Academy of Sciences, 117(12), 6811. https://doi.org/10.1073/pnas.1919198117

Alvarez-Martinez, C. E., & Christie, P. J. (2009). Biological Diversity of Prokaryotic Type IV Secretion Systems. Microbiology and Molecular Biology Reviews, 73(4), 775. https://doi.org/10.1128/MMBR.00023-09

Amavisit, P., Lightfoot, D., Browning, G. F., & Markham, P. F. (2003). Variation between Pathogenic Serovars within Salmonella Pathogenicity Islands. Journal of Bacteriology, 185(12), 3624. https://doi.org/10.1128/JB.185.12.3624-3635.2003

Amgarten, D., Braga, L. P. P., da Silva, A. M., & Setubal, J. C. (2018). MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins [Original Research]. Frontiers in Genetics, 9(304). https://doi.org/10.3389/fgene.2018.00304

Amsalu, A., Sapula, S. A., De Barros Lopes, M., Hart, B. J., Nguyen, A. H., Drigo, B., Turnidge, J., Leong, L. E., & Venter, H. (2020). Efflux Pump-Driven Antibiotic and Biocide Cross-Resistance in Pseudomonas aeruginosa Isolated from Different Ecological Niches: A Case Study in the Development of Multidrug Resistance in Environmental Hotspots. Microorganisms, 8(11), 1647. https://www.mdpi.com/2076-2607/8/11/1647

Andino, A., & Hanning, I. (2015). Salmonella enterica: survival, colonization, and virulence differences among serovars. TheScientificWorldJournal, 2015, 520179-520179. https://doi.org/10.1155/2015/520179

Anjum, M. F., Zankari, E., & Hasman, H. (2018). Molecular Methods for Detection of Antimicrobial Resistance. In Antimicrobial Resistance in Bacteria from Livestock and Companion Animals. American Society of Microbiology. https://doi.org/doi:https://doi.org/10.1128/microbiolspec.ARBA-0011-2017

Antipov, D., Korobeynikov, A., McLean, J. S., & Pevzner, P. A. (2016, Apr 1). hybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics, 32(7), 1009-1015. https://doi.org/10.1093/bioinformatics/btv688

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., & Wishart, D. S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. Nucleic acids research, 44(W1), W16-W21. https://doi.org/10.1093/nar/gkw387

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., & Zagnitko, O. (2008, 2008/02/08). The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics, 9(1), 75. https://doi.org/10.1186/1471-2164-9-75

Babakhani, S., & Oloomi, M. (2018, Nov). Transposons: the agents of antibiotic resistance in bacteria. J Basic Microbiol, 58(11), 905-917. https://doi.org/10.1002/jobm.201800204

Banerji, S., Simon, S., Tille, A., Fruth, A., & Flieger, A. (2020, 2020/03/09). Genome-based Salmonella serotyping as the new gold standard. Scientific Reports, 10(1), 4333. https://doi.org/10.1038/s41598-020-61254-1

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A.

(2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology : a journal of computational molecular cell biology, 19(5), 455-477. https://doi.org/10.1089/cmb.2012.0021

Bäumler, A., & Fang, F. C. (2013). Host specificity of bacterial pathogens. Cold Spring Harbor perspectives in medicine, 3(12), a010041-a010041. https://doi.org/10.1101/cshperspect.a010041

Belcaid, M., Bergeron, A., & Poisson, G. (2010). Mosaic graphs and comparative genomics in phage communities. Journal of computational biology : a journal of computational molecular cell biology, 17(9), 1315-1326. https://doi.org/10.1089/cmb.2010.0108

Bergholz, T. M., den Bakker, H. C., Katz, L. S., Silk, B. J., Jackson, K. A., Kucerova, Z., Joseph, L. A., Turnsek, M., Gladney, L. M., Halpin, J. L., Xavier, K., Gossack, J., Ward, T. J., Frace, M., & Tarr, C. L. (2016, Feb 1). Determination of Evolutionary Relationships of Outbreak-Associated Listeria monocytogenes Strains of Serotypes 1/2a and 1/2b by Whole-Genome Sequencing. Appl Environ Microbiol, 82(3), 928-938. https://doi.org/10.1128/aem.02440-15

Bintsis, T. (2017). Foodborne pathogens. AIMS microbiology, 3(3), 529-563. https://doi.org/10.3934/microbiol.2017.3.529

Bishop, J., & Tritscher, A. (2012). Food safety surveillance and response. Western Pacific surveillance and response journal : WPSAR, 3(2), 1-3. https://doi.org/10.5365/WPSAR.2012.3.2.012

Blount, Z. D. (2015). The unexhausted potential of E. coli. eLife, 4, e05826. https://doi.org/10.7554/eLife.05826

Boisvert, S., Laviolette, F., & Corbeil, J. (2010). Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. Journal of computational biology : a journal of computational molecular cell biology, 17(11), 1519-1533. https://doi.org/10.1089/cmb.2009.0238

Bolger, A. M., Lohse, M., & Usadel, B. (2014, Aug 1). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120. https://doi.org/10.1093/bioinformatics/btu170

Bortolaia, V., Kaas, R. S., Ruppe, E., Roberts, M. C., Schwarz, S., Cattoir, V., Philippon, A., Allesoe, R. L., Rebelo, A. R., Florensa, A. F., Fagelhauer, L., Chakraborty, T., Neumann, B., Werner, G., Bender, J. K., Stingl, K., Nguyen, M., Coppens, J., Xavier, B. B., Malhotra-Kumar, S., Westh, H., Pinholt, M., Anjum, M. F., Duggett, N. A., Kempf, I., Nykäsenoja, S., Olkkola, S., Wieczorek, K., Amaro,

A., Clemente, L., Mossong, J., Losch, S., Ragimbeau, C., Lund, O., & Aarestrup, F. M. (2020, Dec 1). ResFinder 4.0 for predictions of phenotypes from genotypes. J Antimicrob Chemother, 75(12), 3491-3500. https://doi.org/10.1093/jac/dkaa345

Botelho, J., & Schulenburg, H. (2021, 2021/01/01/). The Role of Integrative and Conjugative Elements in Antibiotic Resistance Evolution. Trends in Microbiology, 29(1), 8-18. https://doi.org/https://doi.org/10.1016/j.tim.2020.05.011

Boxrud, D., Monson, T., Stiles, T., & Besser, J. (2010, May-Jun). The role, challenges, and support of pulsenet laboratories in detecting foodborne disease outbreaks. Public health reports (Washington, D.C. : 1974), 125 Suppl 2(Suppl 2), 57-62. https://doi.org/10.1177/00333549101250S207

Brenner, F. W., Villar, R. G., Angulo, F. J., Tauxe, R., & Swaminathan, B. (2000). Salmonella nomenclature. Journal of Clinical Microbiology, 38(7), 2465-2467. https://doi.org/10.1128/JCM.38.7.2465-2467.2000

Brown, E., Dessai, U., McGarry, S., & Gerner-Smidt, P. (2019). Use of Whole-Genome Sequencing for Food Safety and Public Health in the United States. Foodborne Pathogens and Disease, 16(7), 441-450. https://doi.org/10.1089/fpd.2019.2662

Cabezón, E., Ripoll-Rozada, J., Peña, A., de la Cruz, F., & Arechaga, I. (2014). Towards an integrated model of bacterial conjugation. FEMS Microbiology Reviews, 39(1), 81-95. https://doi.org/10.1111/1574-6976.12085

Cameron, A., Zaheer, R., & McAllister, T. A. (2019). Emerging Variants of the Integrative and Conjugant Element ICEMh1 in Livestock Pathogens: Structural Insights, Potential Host Range, and Implications for Bacterial Fitness and Antimicrobial Therapy. Frontiers in Microbiology, 10, 2608-2608. https://doi.org/10.3389/fmicb.2019.02608

Canadian Agency for Drugs and Technologies in Health. (2014). Appendix 5, Summary of findings – cost-effectiveness of next generation sequencing. Retrieved 04/09/2021 from https://www.ncbi.nlm.nih.gov/books/NBK274079/

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., & Brüssow, H. (2003). Prophage genomics. Microbiology and molecular biology reviews : MMBR, 67(2), 238-276. https://doi.org/10.1128/mmbr.67.2.238-276.2003

Cao, Y., Fanning, S., Proos, S., Jordan, K., & Srikumar, S. (2017, 2017-September-21). A Review on the Applications of Next Generation Sequencing Technologies as Applied to Food-Related Microbiome Studies [Review]. Frontiers in Microbiology, 8(1829). https://doi.org/10.3389/fmicb.2017.01829

Carroll, L. M., Wiedmann, M., Mukherjee, M., Nicholas, D. C., Mingle, L. A., Dumas, N. B., Cole, J. A., & Kovac, J. (2019). Characterization of Emetic and Diarrheal Bacillus cereus Strains From a 2016 Foodborne Outbreak Using Whole-Genome Sequencing: Addressing the Microbiological, Epidemiological, and Bioinformatic Challenges. Front Microbiol, 10, 144. https://doi.org/10.3389/fmicb.2019.00144

Casadevall, A., & Pirofski, L.-a. (2003, 2003/10/01). The damage-response framework of microbial pathogenesis. Nature Reviews Microbiology, 1(1), 17-24. https://doi.org/10.1038/nrmicro732

Casalta, J.-P., Zaratzian, C., Hubert, S., Thuny, F., Gouriet, F., Habib, G., Grisoli, D., Deharo, J.-C., & Raoult, D. (2013, 2013/08/01/). Treatment of Staphylococcus aureus endocarditis with high doses of trimethoprim/sulfamethoxazole and clindamycin—Preliminary report. International Journal of Antimicrobial Agents, 42(2), 190-191. https://doi.org/https://doi.org/10.1016/j.ijantimicag.2013.05.002

Casjens, S. R., & Thuman-Commike, P. A. (2011, 2011/03/15/). Evolution of mosaically related tailed bacteriophage genomes seen through the lens of phage P22 virion assembly. Virology, 411(2), 393-415. https://doi.org/https://doi.org/10.1016/j.virol.2010.12.046

CDC. (2012). National Enteric Disease Surveillance: STEC Surveillance Overview. Atlanta, Georgia: US Department of Health and Human Services.

CDC. (2016). Standard Operating Procedure for PulseNet PFGE of Escherichia coli O157:H7, Escherichia coli non-O157 (STEC), Salmonella serotypes, Shigella sonnei and Shigella flexneri. Atlanta, Georgia: US Department of Health and Human Services.

CDC. (2018). National Enteric Disease Surveillance: Shiga Toxin-producing Escherichia coli (STEC) Annual Report, 2016. Atlanta, Georgia: U.S. Department of Health and Human Services.

CDC. (2019a). About PulseNet. Retrieved 04/02/2021 from www.cdc.gov/pulsenet/about/index.html

CDC. (2019b). Surveillance for Foodborne Disease Outbreaks, United States, 2017, Annual Report. Atlanta, Georgia: U.S. Department of Health and Human Services.

CDC. Antibiotic / Antimicrobial Resistance (AR / AMR). Retrieved 03/05/2021 from https://www.cdc.gov/drugresistance/biggest-threats/tracking.html

Chaudhari, A. A., Jawale, C. V., Kim, S. W., & Lee, J. H. (2012). Construction of a Salmonella Gallinarum ghost as a novel inactivated vaccine candidate and its protective efficacy against fowl typhoid in chickens. Veterinary Research, 43(1), 44. https://doi.org/10.1186/1297-9716-43-44

Chen, Z., Erickson, D. L., & Meng, J. (2020). 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

Cheng, R. A., Eade, C. R., & Wiedmann, M. (2019, 2019-June-26). Embracing Diversity: Differences in Virulence Mechanisms, Disease Severity, and Host Adaptations Contribute to the Success of Nontyphoidal Salmonella as a Foodborne Pathogen [Review]. Frontiers in Microbiology, 10(1368). https://doi.org/10.3389/fmicb.2019.01368

Chiang, Y. N., Penadés, J. R., & Chen, J. (2019). Genetic transduction by phages and chromosomal islands: The new and noncanonical. PLOS Pathogens, 15(8), e1007878. https://doi.org/10.1371/journal.ppat.1007878

Chokesajjawatee, N., Santiyanont, P., Chantarasakha, K., Kocharin, K., Thammarongtham, C., Lertampaiporn, S., Vorapreeda, T., Srisuk, T., Wongsurawat, T., Jenjaroenpun, P., Nookaew, I., & Visessanguan, W. (2020, 2020/06/24). Safety Assessment of a Nham Starter Culture Lactobacillus plantarum BCC9546 via Whole-genome Analysis. Scientific Reports, 10(1), 10241. https://doi.org/10.1038/s41598-020-66857-2

Churko, J. M., Mantalas, G. L., Snyder, M. P., & Wu, J. C. (2013). Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases. Circulation research, 112(12), 1613-1623. https://doi.org/10.1161/CIRCRESAHA.113.300939

Clausen, P. T. L. C., Aarestrup, F. M., & Lund, O. (2018, 2018/08/29). Rapid and precise alignment of raw reads against redundant databases with KMA. BMC Bioinformatics, 19(1), 307. https://doi.org/10.1186/s12859-018-2336-6

Clausen, P. T., Zankari, E., Aarestrup, F. M., & Lund, O. (2016, Sep). Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. J Antimicrob Chemother, 71(9), 2484-2488. https://doi.org/10.1093/jac/dkw184

Collis, C. M., Kim, M. J., Stokes, H. W., & Hall, R. M. (1998, Jul). Binding of the purified integron DNA integrase Intl1 to integron- and cassette-associated

recombination sites. Mol Microbiol, 29(2), 477-490. https://doi.org/10.1046/j.1365-2958.1998.00936.x

Cooke, F. J., & Slack, M. P. E. (2017). 183 - Gram-Negative Coccobacilli. In J. Cohen, W. G. Powderly, & S. M. Opal (Eds.), Infectious Diseases (Fourth Edition) (pp. 1611-1627.e1611). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-7020-6285-8.00183-0

Crasta, O. R., Folkerts, O., Fei, Z., Mane, S. P., Evans, C., Martino-Catt, S., Bricker, B., Yu, G., Du, L., & Sobral, B. W. (2008). Genome Sequence of Brucella abortus Vaccine Strain S19 Compared to Virulent Strains Yields Candidate Virulence Genes. PloS one, 3(5), e2193. https://doi.org/10.1371/journal.pone.0002193

Cross, A. S. (2008). What is a virulence factor? Critical care (London, England), 12(6), 196-196. https://doi.org/10.1186/cc7127

Crump, J. A., & Wain, J. (2017). Salmonella. In S. R. Quah (Ed.), International Encyclopedia of Public Health (Second Edition) (pp. 425-433). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-803678-5.00394-5

Davis, S., Pettengill, J. B., Luo, Y., Payne, J., Shpuntoff, A., Rand, H., & Strain, E. (2015, 2015/08/26). CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. PeerJ Computer Science, 1, e20. https://doi.org/10.7717/peerj-cs.20

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. Bioinformatics, 34(15), 2666-2669. https://doi.org/10.1093/bioinformatics/bty149

Deamer, D., Akeson, M., & Branton, D. (2016, May 6). Three decades of nanopore sequencing. Nat Biotechnol, 34(5), 518-524. https://doi.org/10.1038/nbt.3423

Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. PloS one, 8(12), e85024. https://doi.org/10.1371/journal.pone.0085024

del Solar, G., & Espinosa, M. (2000, Aug). Plasmid copy number control: an ever-growing story. Mol Microbiol, 37(3), 492-500. https://doi.org/10.1046/j.1365-2958.2000.02005.x

del Solar, G., Giraldo, R., Ruiz-Echevarría, M. J., Espinosa, M., & Díaz-Orejas, R. (1998). Replication and control of circular bacterial plasmids. Microbiology and molecular biology reviews : MMBR, 62(2), 434-464. https://pubmed.ncbi.nlm.nih.gov/9618448

Delihas, N. (2008). Small mobile sequences in bacteria display diverse structure/function motifs. Molecular Microbiology, 67(3), 475-481. https://doi.org/https://doi.org/10.1111/j.1365-2958.2007.06068.x

den Bakker, H. C., Cummings, C. A., Ferreira, V., Vatta, P., Orsi, R. H., Degoricija, L., Barker, M., Petrauskene, O., Furtado, M. R., & Wiedmann, M. (2010, 2010/12/02). Comparative genomics of the bacterial genus Listeria: Genome evolution is characterized by limited gene acquisition and limited gene loss. BMC Genomics, 11(1), 688. https://doi.org/10.1186/1471-2164-11-688

Dervan, A., & Shendure, J. (2017). Chapter 3 - The State of Whole-Genome Sequencing. In G. S. Ginsburg & H. F. Willard (Eds.), Genomic and Precision Medicine (Third Edition) (pp. 45-62). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-800681-8.00003-7

Diard, M., & Hardt, W.-D. (2017). Evolution of bacterial virulence. FEMS Microbiology Reviews, 41(5), 679-697. https://doi.org/10.1093/femsre/fux023

dos Santos, L. F., Gonçalves, E. M., Vaz, T. M. I., Irino, K., & Guth, B. E. C. (2007). Distinct Pathotypes of O113 Escherichia coli Strains Isolated from Humans and Animals in Brazil. Journal of Clinical Microbiology, 45(6), 2028. https://doi.org/10.1128/JCM.00340-07

Doyle, M. P. (2013). Food Safety: Bacterial Contamination. In B. Caballero (Ed.), Encyclopedia of Human Nutrition (Third Edition) (pp. 322-330). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-375083-9.00124-0

Dumitrescu, O., Boisset, S., Badiou, C., Bes, M., Benito, Y., Reverdy, M.-E., Vandenesch, F., Etienne, J., & Lina, G. (2007). Effect of Antibiotics on Staphylococcus aureus Producing Panton-Valentine Leukocidin. Antimicrobial agents and chemotherapy, 51(4), 1515. https://doi.org/10.1128/AAC.01201-06

Dumschott, K., Schmidt, M. H. W., Chawla, H. S., Snowdon, R., & Usadel, B. (2020). Oxford Nanopore sequencing: new opportunities for plant genomics? Journal of Experimental Botany, 71(18), 5313-5322. https://doi.org/10.1093/jxb/eraa263

DuPont, H. L. (2007, Nov 15). The growing threat of foodborne bacterial enteropathogens of animal origin. Clin Infect Dis, 45(10), 1353-1361. https://doi.org/10.1086/522662

Dutta, A., More, D., Tupaki-Sreepurna, A., Sinha, B., Goyal, N., & Rongsen-Chandola, T. (2020). Typhoid and paratyphoid fever co-infection in children from an urban

slum of Delhi. IDCases, 20, e00717-e00717. https://doi.org/10.1016/j.idcr.2020.e00717

ecSeq. (2017). Do you have two colors or four colors in Illumina? Retrieved 04/11/2021 from https://www.ecseq.com/support/ngs/do_you_have_two_colors_or_four_colors_in_Illumina

Ejigu, G. F., & Jung, J. (2020). Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. Biology, 9(9), 295. https://doi.org/10.3390/biology9090295

European Food Safety Authority. (2018). The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2017. EFSA Journal, 16(12), e05500. https://doi.org/https://doi.org/10.2903/j.efsa.2018.5500

Fàbrega, A., & Vila, J. (2013). Salmonella enterica serovar Typhimurium skills to succeed in the host: virulence and regulation. Clinical Microbiology Reviews, 26(2), 308-341. https://doi.org/10.1128/CMR.00066-12

Fakruddin, M., Mannan, K. S. B., & Andrews, S. (2013, 2013/09/26). Viable but Nonculturable Bacteria: Food Safety and Public Health Perspective. ISRN Microbiology, 2013, 703813. https://doi.org/10.1155/2013/703813

Fernández, L., Breidenstein, E. B. M., & Hancock, R. E. W. (2011, 2011/02/01/). Creeping baselines and adaptive resistance to antibiotics. Drug Resistance Updates, 14(1), 1-21. https://doi.org/https://doi.org/10.1016/j.drup.2011.01.001

Fournier, P.-E., Dubourg, G., & Raoult, D. (2014). Clinical detection and characterization of bacterial pathogens in the genomics era. Genome medicine, 6(11), 114-114. https://doi.org/10.1186/s13073-014-0114-2

Fratamico, P. M., DebRoy, C., Liu, Y., Needleman, D. S., Baranzoni, G. M., & Feng, P. (2016, 2016-May-03). Advances in Molecular Serotyping and Subtyping of Escherichia coli† [Mini Review]. Frontiers in Microbiology, 7(644). https://doi.org/10.3389/fmicb.2016.00644

Frost, L. S., Leplae, R., Summers, A. O., & Toussaint, A. (2005, 2005/09/01). Mobile genetic elements: the agents of open source evolution. Nature Reviews Microbiology, 3(9), 722-732. https://doi.org/10.1038/nrmicro1235

Fukasawa, Y., Ermini, L., Wang, H., Carty, K., & Cheung, M.-S. (2020). LongQC: A Quality Control Tool for Third Generation Sequencing Long Read Data. G3 (Bethesda, Md.), 10(4), 1193-1196. https://doi.org/10.1534/g3.119.400864

Gao, W., Chua, K., Davies, J. K., Newton, H. J., Seemann, T., Harrison, P. F., Holmes, N. E., Rhee, H.-W., Hong, J.-I., Hartland, E. L., Stinear, T. P., & Howden, B. P. (2010). Two Novel Point Mutations in Clinical Staphylococcus aureus Reduce Linezolid Susceptibility and Switch on the Stringent Response to Promote Persistent Infection. PLOS Pathogens, 6(6), e1000944. https://doi.org/10.1371/journal.ppat.1000944

Gardner, S. N., Slezak, T., & Hall, B. G. (2015). kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics, 31(17), 2877-2878. https://doi.org/10.1093/bioinformatics/btv271

Ghosh, A., N, S., & Saha, S. (2020, 2020/06/02). Survey of drug resistance associated gene mutations in Mycobacterium tuberculosis, ESKAPE and other bacterial species. Scientific Reports, 10(1), 8957. https://doi.org/10.1038/s41598-020-65766-8

Ghosh, D., Veeraraghavan, B., Elangovan, R., & Vivekanandan, P. (2020). Antibiotic Resistance and Epigenetics: More to It than Meets the Eye. Antimicrobial agents and chemotherapy, 64(2), e02225-02219. https://doi.org/10.1128/AAC.02225-19

Giusti Mde, L., Pistorio, M., Lozano, M. J., Tejerizo, G. A., Salas, M. E., Martini, M. C., López, J. L., Draghi, W. O., Del Papa, M. F., Pérez-Mendoza, D., Sanjuán, J., & Lagares, A. (2012, May). Genetic and functional characterization of a yet-unclassified rhizobial Dtr (DNA-transfer-and-replication) region from a ubiquitous plasmid conjugal system present in Sinorhizobium meliloti, in Sinorhizobium medicae, and in other nonrhizobial Gram-negative bacteria. Plasmid, 67(3), 199-210. https://doi.org/10.1016/j.plasmid.2011.12.010

Glaser, P., Frangeul, L., Buchrieser, C., Rusniok, C., Amend, A., Baquero, F., Berche, P., Bloecker, H., Brandt, P., Chakraborty, T., Charbit, A., Chetouani, F., Couvé, E., de Daruvar, A., Dehoux, P., Domann, E., Domínguez-Bernal, G., Duchaud, E., Durant, L., Dussurget, O., Entian, K. D., Fsihi, H., García-del Portillo, F., Garrido, P., Gautier, L., Goebel, W., Gómez-López, N., Hain, T., Hauf, J., Jackson, D., Jones, L. M., Kaerst, U., Kreft, J., Kuhn, M., Kunst, F., Kurapkat, G., Madueno, E., Maitournam, A., Vicente, J. M., Ng, E., Nedjari, H., Nordsiek, G., Novella, S., de Pablos, B., Pérez-Diaz, J. C., Purcell, R., Remmel, B., Rose, M., Schlueter, T., Simoes, N., Tierrez, A., Vázquez-Boland, J. A., Voss, H., Wehland,

J., & Cossart, P. (2001, Oct 26). Comparative genomics of Listeria species. Science, 294(5543), 849-852. https://doi.org/10.1126/science.1063447

Goldstein, S., Beka, L., Graf, J., & Klassen, J. L. (2019, 2019/01/09). 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

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016, 2016/06/01). Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics, 17(6), 333-351. https://doi.org/10.1038/nrg.2016.49

Gupta, S. K., Padmanabhan, B. R., Diene, S. M., Lopez-Rojas, R., Kempf, M., Landraud, L., & Rolain, J. M. (2014). ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrobial agents and chemotherapy, 58(1), 212-220. https://doi.org/10.1128/aac.01310-13

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072-1075. https://doi.org/10.1093/bioinformatics/btt086

Hackl, T., Hedrich, R., Schultz, J., & Förster, F. (2014, Nov 1). proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics, 30(21), 3004-3011. https://doi.org/10.1093/bioinformatics/btu392

Haft, D. H., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., O'Neill, K., Li, W., Chitsaz, F., Derbyshire, M. K., Gonzales, N. R., Gwadz, M., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Zheng, C., Thibaud-Nissen, F., Geer, L. Y., Marchler-Bauer, A., & Pruitt, K. D. (2018). RefSeq: an update on prokaryotic genome annotation and curation. Nucleic acids research, 46(D1), D851-D860. https://doi.org/10.1093/nar/gkx1068

Hall, R. M., & Collis, C. M. (1995, Feb). Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination. Mol Microbiol, 15(4), 593-600. https://doi.org/10.1111/j.1365-2958.1995.tb02368.x

Hara-Kudo, Y., & Takatori, K. (2011, Oct). Contamination level and ingestion dose of foodborne pathogens associated with infections. Epidemiol Infect, 139(10), 1505-1510. https://doi.org/10.1017/s095026881000292x

Hartl, D. L., & Dykhuizen, D. E. (1984). The population genetics of Escherichia coli. Annual review of genetics, 18, 31-68. https://doi.org/10.1146/annurev.ge.18.120184.000335

Hasegawa, H., Suzuki, E., & Maeda, S. (2018, 2018-October-04). Horizontal Plasmid Transfer by Transformation in Escherichia coli: Environmental Factors and Possible Mechanisms [Mini Review]. Frontiers in Microbiology, 9(2365). https://doi.org/10.3389/fmicb.2018.02365

Hashempour-Baltork, F., Hosseini, H., Shojaee-Aliabadi, S., Torbati, M., Alizadeh, A. M., & Alizadeh, M. (2019). Drug Resistance and the Prevention Strategies in Food Borne Bacteria: An Update Review. Advanced pharmaceutical bulletin, 9(3), 335-347. https://doi.org/10.15171/apb.2019.041

Havelaar, A. H., Kirk, M. D., Torgerson, P. R., Gibb, H. J., Hald, T., Lake, R. J., Praet, N., Bellinger, D. C., de Silva, N. R., Gargouri, N., Speybroeck, N., Cawthorne, A., Mathers, C., Stein, C., Angulo, F. J., Devleesschauwer, B., & World Health Organization Foodborne Disease Burden Epidemiology Reference, G. (2015). World Health Organization Global Estimates and Regional Comparisons of the Burden of Foodborne Disease in 2010. PLoS medicine, 12(12), e1001923-e1001923. https://doi.org/10.1371/journal.pmed.1001923

Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C. G., Ohtsubo, E., Nakayama, K., Murata, T., Tanaka, M., Tobe, T., Iida, T., Takami, H., Honda, T., Sasakawa, C., Ogasawara, N., Yasunaga, T., Kuhara, S., Shiba, T., Hattori, M., & Shinagawa, H. (2001, Feb 28). Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res, 8(1), 11-22. https://doi.org/10.1093/dnares/8.1.11

He, X.-F., Zhang, H.-J., Cao, J.-G., Liu, F., Wang, J.-K., Ma, W.-J., & Yin, W. (2017). A novel method to detect bacterial resistance to disinfectants. Genes & diseases, 4(3), 163-169. https://doi.org/10.1016/j.gendis.2017.07.001

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, 107(1), 1-8. https://doi.org/10.1016/j.ygeno.2015.11.003

Heldenbrand, J. R., Baheti, S., Bockol, M. A., Drucker, T. M., Hart, S. N., Hudson, M. E., Iyer, R. K., Kalmbach, M. T., Kendig, K. I., Klee, E. W., Mattson, N. R., Wieben, E. D., Wiepert, M., Wildman, D. E., & Mainzer, L. S. (2019, 2019/11/08). Recommendations for performance optimizations when using GATK3.8 and GATK4. BMC Bioinformatics, 20(1), 557. https://doi.org/10.1186/s12859-019-3169-7

Hendrix, R. W. (2003, Oct). Bacteriophage genomics. Curr Opin Microbiol, 6(5), 506-511. https://doi.org/10.1016/j.mib.2003.09.004

Heredia, N., & García, S. (2018). Animals as sources of food-borne pathogens: A review. Animal nutrition (Zhongguo xu mu shou yi xue hui), 4(3), 250-255. https://doi.org/10.1016/j.aninu.2018.04.006

Ho Sui, S. J., Fedynak, A., Hsiao, W. W. L., Langille, M. G. I., & Brinkman, F. S. L. (2009). The association of virulence factors with genomic islands. PloS one, 4(12), e8094-e8094. https://doi.org/10.1371/journal.pone.0008094

Holder, M., & Lewis, P. O. (2003, Apr). Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet, 4(4), 275-284. https://doi.org/10.1038/nrg1044

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4378521/

Huang, K., Herrero-Fresno, A., Thøfner, I., Skov, S., & Olsen, J. E. (2019, Dec). Interaction Differences of the Avian Host-Specific Salmonella enterica Serovar Gallinarum, the Host-Generalist S. Typhimurium, and the Cattle Host-Adapted S. Dublin with Chicken Primary Macrophage. Infect Immun, 87(12). https://doi.org/10.1128/iai.00552-19

Hung, L.-W., Kim, H.-B., Murakami, S., Gupta, G., Kim, C.-Y., & Terwilliger, T. C. (2013). Crystal structure of AcrB complexed with linezolid at 3.5 Å resolution. Journal of Structural and Functional Genomics, 14(2), 71-75. https://doi.org/10.1007/s10969-013-9154-x

Hur, Y., Chalita, M., Ha, S.-m., Baek, I., & Chun, J. (2019). VCGIDB: A Database and Web Resource for the Genomic Islands from Vibrio cholerae. Pathogens, 8(4). https://doi.org/10.3390/pathogens8040261

Illumina. (2011). Quality Scores for Next-Generation Sequencing Retrieved 04/19/2021 from https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf

Illumina. (2013). An introduction to Illumina Next-Generation Sequencing Technology for Microbiologists. Retrieved 04/11/2021 from https://www.illumina.com/content/dam/illumina-marketing/documents/products/sequencing_introduction_microbiology.pdf

Illumina. (2015). Faster sequencing and data processing 2-channel SBS generates data faster than 4-channel SBS, while maintaining quality and accuracy. Retrieved 04/11/2021 from https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology/2-channel-sbs.html#:~:text=The%202%2Dchannel%20SBS%20method,that%20sets%20Illumina%20systems%20apart.

Ingle, D. J., Valcanis, M., Kuzevski, A., Tauschek, M., Inouye, M., Stinear, T., Levine, M. M., Robins-Browne, R. M., & Holt, K. E. (2016). In silico serotyping of E. coli from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages. Microbial Genomics, 2(7), e000064-e000064. https://doi.org/10.1099/mgen.0.000064

Inouye, M., Dashnow, H., Raven, L.-A., Schultz, M. B., Pope, B. J., Tomita, T., Zobel, J., & Holt, K. E. (2014, 2014/11/20). SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome medicine, 6(11), 90. https://doi.org/10.1186/s13073-014-0090-6

Institute of Medicine (US) Forum on Microbial Threats. (2006). Surveillance of the Food Supply. Retrieved 04/01/2021 from https://www.ncbi.nlm.nih.gov/books/NBK57083/

Jackson, B. R., Tarr, C., Strain, E., Jackson, K. A., Conrad, A., Carleton, H., Katz, L. S., Stroika, S., Gould, L. H., Mody, R. K., Silk, B. J., Beal, J., Chen, Y., Timme, R., Doyle, M., Fields, A., Wise, M., Tillman, G., Defibaugh-Chavez, S., Kucerova, Z., Sabol, A., Roache, K., Trees, E., Simmons, M., Wasilenko, J., Kubota, K., Pouseele, H., Klimke, W., Besser, J., Brown, E., Allard, M., & Gerner-Smidt, P. (2016). Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 63(3), 380-386. https://doi.org/10.1093/cid/ciw242

Jagadeesan, B., Gerner-Smidt, P., Allard, M. W., Leuillet, S., Winkler, A., Xiao, Y., Chaffron, S., Van Der Vossen, J., Tang, S., Katase, M., McClure, P., Kimura, B., Ching Chai, L., Chapman, J., & Grant, K. (2019, 2019/06/01/). The use of next generation sequencing for improving food safety: Translation into practice. Food Microbiology, 79, 96-115. https://doi.org/https://doi.org/10.1016/j.fm.2018.11.005

Jiang, X., Hall, A. B., Xavier, R. J., & Alm, E. J. (2019). Comprehensive analysis of chromosomal mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. PloS one, 14(12), e0223680. https://doi.org/10.1371/journal.pone.0223680

Jin, M., Liu, L., Wang, D.-n., Yang, D., Liu, W.-l., Yin, J., Yang, Z.-w., Wang, H.-r., Qiu, Z.-g., Shen, Z.-q., Shi, D.-y., Li, H.-b., Guo, J.-h., & Li, J.-w. (2020, 2020/07/01). Chlorine disinfection promotes the exchange of antibiotic resistance genes across bacterial genera by natural transformation. The ISME Journal, 14(7), 1847-1856. https://doi.org/10.1038/s41396-020-0656-9

Joensen, K. G., Scheutz, F., Lund, O., Hasman, H., Kaas, R. S., Nielsen, E. M., & Aarestrup, F. M. (2014, May). Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. J Clin Microbiol, 52(5), 1501-1510. https://doi.org/10.1128/jcm.03617-13

Johnson, A. P., & Woodford, N. (2013). Global spread of antibiotic resistance: the example of New Delhi metallo-β-lactamase (NDM)-mediated carbapenem resistance. Journal of Medical Microbiology, 62(4), 499-513. https://doi.org/https://doi.org/10.1099/jmm.0.052555-0

Johnson, C. M., & Grossman, A. D. (2015). Integrative and Conjugative Elements (ICEs): What They Do and How They Work. Annual review of genetics, 49, 577-601. https://doi.org/10.1146/annurev-genet-112414-055018

Johnson, T. J., Wannemuehler, Y. M., Johnson, S. J., Logue, C. M., White, D. G., Doetkott, C., & Nolan, L. K. (2007). Plasmid replicon typing of commensal and pathogenic Escherichia coli isolates. Applied and Environmental Microbiology, 73(6), 1976-1983. https://doi.org/10.1128/AEM.02171-06

Jones, C. M., Price, R. E., & Breidt, F. (2020). Escherichia coli O157:H7 Stationary-Phase Acid Resistance and Assessment of Survival in a Model Vegetable Fermentation System. Journal of Food Protection, 83(5), 745-753. https://doi.org/10.4315/JFP-19-463

Jurtz, V. I., Villarroel, J., Lund, O., Voldby Larsen, M., & Nielsen, M. (2016). MetaPhinder—Identifying Bacteriophage Sequences in Metagenomic Data Sets. PLoS One, 11(9), e0163111. https://doi.org/10.1371/journal.pone.0163111

Kamath, G. M., Shomorony, I., Xia, F., Courtade, T. A., & Tse, D. N. (2017, May). HINGE: long-read assembly achieves optimal repeat resolution. Genome Res, 27(5), 747-756. https://doi.org/10.1101/gr.216465.116

Katayama, Y., Ito, T., & Hiramatsu, K. (2000). A New Class of Genetic Element, Staphylococcus Cassette Chromosome mec, Encodes Methicillin Resistance in Staphylococcus aureus. Antimicrobial Agents and Chemotherapy, 44(6), 1549-1555. https://doi.org/10.1128/AAC.44.6.1549-1555.2000

Kim, M., Weigand, M. R., Oh, S., Hatt, J. K., Krishnan, R., Tezel, U., Pavlostathis, S. G., & Konstantinidis, K. T. (2018). Widely Used Benzalkonium Chloride Disinfectants Can Promote Antibiotic Resistance. Applied and Environmental Microbiology, 84(17), e01201-01218. https://doi.org/10.1128/AEM.01201-18

Kim, S.-O., & Kim, S.-S. (2021, 2021/02/06). Recent (2011–2017) foodborne outbreak cases in the Republic of Korea compared to the United States: a review. Food Science and Biotechnology. https://doi.org/10.1007/s10068-020-00864-x

Kimani, R. W., Muigai, A. W. T., Sang, W., Kiiru, J. N., & Kariuki, S. (2014). Virulence factors in environmental and clinical Vibrio cholerae from endemic areas in Kenya. African journal of laboratory medicine, 3(1), 41-41. https://doi.org/10.4102/ajlm.v3i1.41

Kleinheinz, K. A., Joensen, K. G., & Larsen, M. V. (2014). Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and E. coli virulence genes in bacteriophage and prophage nucleotide sequences. Bacteriophage, 4(1), e27943-e27943. https://doi.org/10.4161/bact.27943

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019, 2019/05/01). Assembly of long, error-prone reads using repeat graphs. Nature Biotechnology, 37(5), 540-546. https://doi.org/10.1038/s41587-019-0072-8

Komano, T., Yoshida, T., Narahara, K., & Furuya, N. (2000, 2000/03/01). The transfer region of IncI1 plasmid R64: similarities between R64 tra and Legionella icm/dot genes [https://doi.org/10.1046/j.1365-2958.2000.01769.x]. Molecular Microbiology, 35(6), 1348-1359. https://doi.org/https://doi.org/10.1046/j.1365-2958.2000.01769.x

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017, May). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res, 27(5), 722-736. https://doi.org/10.1101/gr.215087.116

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome research, 27(5), 722-736. https://doi.org/10.1101/gr.215087.116

Kudva, I. T., Evans, P. S., Perna, N. T., Barrett, T. J., Ausubel, F. M., Blattner, F. R., & Calderwood, S. B. (2002). Strains of Escherichia coli O157:H7 Differ Primarily by Insertions or Deletions, Not Single-Nucleotide Polymorphisms. Journal of Bacteriology, 184(7), 1873. https://doi.org/10.1128/JB.184.7.1873-1879.2002

Kurenbach, B., Grothe, D., Farías, M. E., Szewzyk, U., & Grohmann, E. (2002). The tra Region of the Conjugative Plasmid pIP501 Is

Organized in an Operon with the First Gene Encoding the Relaxase. Journal of Bacteriology, 184(6), 1801. https://doi.org/10.1128/JB.184.6.1801-1805.2002

Kuśmirek, W., & Nowak, R. (2018, 2018/07/18). De novo assembly of bacterial genomes with repetitive DNA regions by dnaasm application. BMC Bioinformatics, 19(1), 273. https://doi.org/10.1186/s12859-018-2281-4

Langmead, B., & Salzberg, S. L. (2012, 2012/04/01). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357-359. https://doi.org/10.1038/nmeth.1923

Laver, J. R., Hughes, S. E., & Read, R. C. (2015). Neisserial Molecular Adaptations to the Nasopharyngeal Niche. Adv Microb Physiol, 66, 323-355. https://doi.org/10.1016/bs.ampbs.2015.05.001

Lavigne, J.-P., Sotto, A., Nicolas-Chanoine, M.-H., Bouziges, N., Pagès, J.-M., & Davin-Regli, A. (2013). An adaptive response of Enterobacter aerogenes to imipenem: regulation of porin balance in clinical isolates. International Journal of Antimicrobial Agents, 41(2), 130-136. https://doi.org/https://doi.org/10.1016/j.ijantimicag.2012.10.010

Leclercq, R. (2002). Mechanisms of Resistance to Macrolides and Lincosamides: Nature of the Resistance Elements and Their Clinical Implications. Clinical Infectious Diseases, 34(4), 482-492. https://doi.org/10.1086/324626

Lee, B. S., Ban, O. H., Bang, W. Y., Chae, S. A., Oh, S., Park, C., Lee, M., Kim, S.-J., Yang, J., & Jung, Y. H. (2021, 2021/02/15). Safety assessment of Lactobacillus reuteri IDCC 3701 based on phenotypic and genomic analysis. Annals of Microbiology, 71(1), 10. https://doi.org/10.1186/s13213-021-01622-y

Leekitcharoenphon, P., Sørensen, G., Löfström, C., Battisti, A., Szabo, I., Wasyl, D., Slowey, R., Zhao, S., Brisabois, A., Kornschober, C., Kärssin, A., Szilárd, J., Černý, T., Svendsen, C. A., Pedersen, K., Aarestrup, F. M., & Hendriksen, R. S. (2019). Cross-Border Transmission of Salmonella Choleraesuis var. Kunzendorf in European Pigs and Wild Boar: Infection, Genetics, and Evolution [Original Research]. Frontiers in Microbiology, 10(179). https://doi.org/10.3389/fmicb.2019.00179

Leimbach, A., Hacker, J., & Dobrindt, U. (2013). E. coli as an all-rounder: the thin line between commensalism and pathogenicity. Current topics in microbiology and immunology, 358, 3-32. https://doi.org/10.1007/82_2012_303

Leplae, R., Hebrant, A., Wodak, S. J., & Toussaint, A. (2004). ACLAME: a CLAssification of Mobile genetic Elements. Nucleic acids research, 32(Database issue), D45-D49. https://doi.org/10.1093/nar/gkh084

Leplae, R., Lima-Mendez, G., & Toussaint, A. (2010, Jan). ACLAME: a CLAssification of Mobile genetic Elements, update 2010. Nucleic acids research, 38(Database issue), D57-61. https://doi.org/10.1093/nar/gkp938

Lepuschitz, S., Baron, S., Larvor, E., Granier, S. A., Pretzer, C., Mach, R. L., Farnleitner, A. H., Ruppitsch, W., Pleininger, S., Indra, A., & Kirschner, A. K. T. (2019). Phenotypic and Genotypic Antimicrobial Resistance Traits of Vibrio cholerae Non-O1/Non-O139 Isolated From a Large Austrian Lake Frequently Associated With Cases of Human Infection [Original Research]. Frontiers in Microbiology, 10(2600). https://doi.org/10.3389/fmicb.2019.02600

Leyer, G. J., Wang, L. L., & Johnson, E. A. (1995). Acid adaptation of Escherichia coli O157:H7 increases survival in acidic foods. Applied and Environmental Microbiology, 61(10), 3752-3755. https://doi.org/10.1128/AEM.61.10.3752-3755.1995

Li, C., Xiang, X., Huang, Y., Zhou, Y., An, D., Dong, J., Zhao, C., Liu, H., Li, Y., Wang, Q., Du, C., Messing, J., Larkins, B. A., Wu, Y., & Wang, W. (2020, 2020/01/07). Long-read sequencing reveals genomic structural variations that underlie creation of quality protein maize. Nature Communications, 11(1), 17. https://doi.org/10.1038/s41467-019-14023-2

Li, D., & Liu, S. (2019). Chapter 12 - Water Quality Monitoring in Aquaculture. In D. Li & S. Liu (Eds.), Water Quality Monitoring and Management (pp. 303-328). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-811330-1.00012-0

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics, 32(14), 2103-2110. https://doi.org/10.1093/bioinformatics/btw152

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18), 3094-3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., & Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. Briefings in Bioinformatics, 11(5), 473-483. https://doi.org/10.1093/bib/bbq015

Li, H., Wu, K., Ruan, C., Pan, J., Wang, Y., & Long, H. (2019, 2019/11/01). Cost-reduction strategies in massive genomics experiments. Marine Life Science & Technology, 1(1), 15-21. https://doi.org/10.1007/s42995-019-00013-2

Li, W., Pires, S. M., Liu, Z., Ma, X., Liang, J., Jiang, Y., Chen, J., Liang, J., Wang, S., Wang, L., Wang, Y., Meng, C., Huo, X., Lan, Z., Lai, S., Liu, C., Han, H., Liu, J.,

Fu, P., & Guo, Y. (2020, 2020/12/01/). Surveillance of foodborne disease outbreaks in China, 2003–2017. Food Control, 118, 107359. https://doi.org/https://doi.org/10.1016/j.foodcont.2020.107359

Lima-Mendez, G., Van Helden, J., Toussaint, A., & Leplae, R. (2008). Prophinder: a computational tool for prophage prediction in prokaryotic genomes. Bioinformatics, 24(6), 863-865. https://doi.org/10.1093/bioinformatics/btn043

Liptáková, A., Siegfried, L., Podracká, L., Sabol, M., Sehnálková, H., Bogyiová, E., Rosocha, J., Kmetová, M., Kerestesová, H., & Kotulová, D. (2002). Detection of Shiga toxins, intimin and enterohemolysin in Escherichia coli strains isolated from children in eastern Slovakia. Folia Microbiol (Praha), 47(2), 185-188. https://doi.org/10.1007/bf02817680

Liu, B., Zheng, D., Jin, Q., Chen, L., & Yang, J. (2019). VFDB 2019: a comparative pathogenomic platform with an interactive web interface. Nucleic acids research, 47(D1), D687-D692. https://doi.org/10.1093/nar/gky1080

Liu, M., Li, X., Xie, Y., Bi, D., Sun, J., Li, J., Tai, C., Deng, Z., & Ou, H.-Y. (2019). ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. Nucleic acids research, 47(D1), D660-D665. https://doi.org/10.1093/nar/gky1123

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020, 2020/10/01). Long-read human genome sequencing and its applications. Nature Reviews Genetics, 21(10), 597-614. https://doi.org/10.1038/s41576-020-0236-x

Lou, L., Zhang, P., Piao, R., & Wang, Y. (2019, 2019-July-31). Salmonella Pathogenicity Island 1 (SPI-1) and Its Complex Regulatory Network [Review]. Frontiers in Cellular and Infection Microbiology, 9(270). https://doi.org/10.3389/fcimb.2019.00270

Luna, V. A., King, D. S., Peak, K. K., Reeves, F., Heberlein-Larson, L., Veguilla, W., Heller, L., Duncan, K. E., Cannons, A. C., Amuso, P., & Cattani, J. (2006). Bacillus anthracis virulent plasmid pX02 genes found in large plasmids of two other Bacillus species. Journal of Clinical Microbiology, 44(7), 2367-2377. https://doi.org/10.1128/JCM.00154-06

Luo, Y., Van Nguyen, U., de la Fe Rodriguez, P. Y., Devriendt, B., & Cox, E. (2015, 2015/10/21). F4+ ETEC infection and oral immunization with F4 fimbriae elicits an IL-17-dominated immune response. Veterinary Research, 46(1), 121. https://doi.org/10.1186/s13567-015-0264-2

Maga, G. (2017). DNA Replication☆. In Reference Module in Biomedical Sciences. Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-801238-3.64155-7

Mageiros, L., Méric, G., Bayliss, S. C., Pensar, J., Pascoe, B., Mourkas, E., Calland, J. K., Yahara, K., Murray, S., Wilkinson, T. S., Williams, L. K., Hitchings, M. D., Porter, J., Kemmett, K., Feil, E. J., Jolley, K. A., Williams, N. J., Corander, J., & Sheppard, S. K. (2021, 2021/02/03). Genome evolution and the emergence of pathogenicity in avian Escherichia coli. Nature Communications, 12(1), 765. https://doi.org/10.1038/s41467-021-20988-w

Maiden, M. C. J. (1998). Horizontal Genetic Exchange, Evolution, and Spread of Antibiotic Resistance in Bacteria. Clinical Infectious Diseases, 27(Supplement_1), S12-S20. https://doi.org/10.1086/514917

Maiden, M. C. J., van Rensburg, M. J. J., Bray, J. E., Earle, S. G., Ford, S. A., Jolley, K. A., & McCarthy, N. D. (2013, 2013/10/01). MLST revisited: the gene-by-gene approach to bacterial genomics. Nature Reviews Microbiology, 11(10), 728-736. https://doi.org/10.1038/nrmicro3093

Malachowa, N., & DeLeo, F. R. (2010, Sep). Mobile genetic elements of Staphylococcus aureus. Cell Mol Life Sci, 67(18), 3057-3071. https://doi.org/10.1007/s00018-010-0389-4

Manchanda, N., Portwood, J. L., Woodhouse, M. R., Seetharam, A. S., Lawrence-Dill, C. J., Andorf, C. M., & Hufford, M. B. (2020). GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations. BMC Genomics, 21(1), 193. https://doi.org/10.1186/s12864-020-6568-2

Manning, S. D., Motiwala, A. S., Springman, A. C., Qi, W., Lacher, D. W., Ouellette, L. M., Mladonicky, J. M., Somsel, P., Rudrik, J. T., Dietrich, S. E., Zhang, W., Swaminathan, B., Alland, D., & Whittam, T. S. (2008, Mar 25). Variation in virulence among clades of Escherichia coli O157:H7 associated with disease outbreaks. Proceedings of the National Academy of Sciences of the United States of America, 105(12), 4868-4873. https://doi.org/10.1073/pnas.0710834105

Manning, S. D., Motiwala, A. S., Springman, A. C., Qi, W., Lacher, D. W., Ouellette, L. M., Mladonicky, J. M., Somsel, P., Rudrik, J. T., Dietrich, S. E., Zhang, W., Swaminathan, B., Alland, D., & Whittam, T. S. (2008). Variation in virulence among clades of Escherichia coli O157:H7 associated with disease outbreaks. Proceedings of the National Academy of Sciences of the United States of America, 105(12), 4868-4873. https://doi.org/10.1073/pnas.0710834105

Marcus, S. L., Brumell, J. H., Pfeifer, C. G., & Finlay, B. B. (2000, Feb). Salmonella pathogenicity islands: big virulence in small packages. Microbes Infect, 2(2), 145-156. https://doi.org/10.1016/s1286-4579(00)00273-2

Martens, E., & Demain, A. L. (2017, 2017/05/01). The antibiotic resistance crisis, with a focus on the United States. The Journal of Antibiotics, 70(5), 520-526. https://doi.org/10.1038/ja.2017.30

Menezes, F. G., Neves Sda, S., Sousa, O. V., Vila-Nova, C. M., Maggioni, R., Theophilo, G. N., Hofer, E., & Vieira, R. H. (2014, Sep-Oct). Detection of virulence genes in environmental strains of Vibrio cholerae from estuaries in northeastern Brazil. Rev Inst Med Trop Sao Paulo, 56(5), 427-432. https://doi.org/10.1590/s0036-46652014000500010

Mercante, J. W., Morrison, S. S., Desai, H. P., Raphael, B. H., & Winchell, J. M. (2016). Genomic Analysis Reveals Novel Diversity among the 1976 Philadelphia Legionnaires' Disease Outbreak Isolates and Additional ST36 Strains. PloS one, 11(9), e0164074. https://doi.org/10.1371/journal.pone.0164074

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., & Edwards, R. A. (2008, 2008/09/19). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics, 9(1), 386. https://doi.org/10.1186/1471-2105-9-386

Milani, R. V., Wilt, J. K., Entwisle, J., Hand, J., Cazabon, P., & Bohan, J. G. (2019). Reducing inappropriate outpatient antibiotic prescribing: normative comparison using unblinded provider reports. BMJ open quality, 8(1), e000351-e000351. https://doi.org/10.1136/bmjoq-2018-000351

Miller, W. J., & Capy, P. (2004). Mobile genetic elements as natural tools for genome evolution. Methods Mol Biol, 260, 1-20. https://doi.org/10.1385/1-59259-755-6:001

Million-Weaver, S., & Camps, M. (2014). Mechanisms of plasmid segregation: have multicopy plasmids been overlooked? Plasmid, 75, 27-36. https://doi.org/10.1016/j.plasmid.2014.07.002

Mohammed, M., & Thapa, S. (2020). Evaluation of WGS-subtyping methods for epidemiological surveillance of foodborne salmonellosis. One Health Outlook, 2(1), 13. https://doi.org/10.1186/s42522-020-00016-5

Morozova, O., Hirst, M., & Marra, M. A. (2009, 2009/09/01). Applications of New Sequencing Technologies for Transcriptome Analysis. Annual Review of

Genomics and Human Genetics, 10(1), 135-151. https://doi.org/10.1146/annurev-genom-082908-145957

Moss, E. L., Maghini, D. G., & Bhatt, A. S. (2020, 2020/06/01). Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nature Biotechnology, 38(6), 701-707. https://doi.org/10.1038/s41587-020-0422-6

Motta, S. S., Cluzel, P., & Aldana, M. (2015). Adaptive Resistance in Bacteria Requires Epigenetic Inheritance, Genetic Noise, and Cost of Efflux Pumps. PloS one, 10(3), e0118464. https://doi.org/10.1371/journal.pone.0118464

Münch, K., Münch, R., Biedendieck, R., Jahn, D., & Müller, J. (2019). Evolutionary model for the unequal segregation of high copy plasmids. PLOS Computational Biology, 15(3), e1006724. https://doi.org/10.1371/journal.pcbi.1006724

Năşcuţiu, A. M. (2010, Jan-Mar). [Viable non-culturable bacteria]. Bacteriol Virusol Parazitol Epidemiol, 55(1), 11-18. (Bacterii viabile necultivabile.)

National Academies Press. (2010). Sequence-Based Classification of Select Agents: A Brighter Line: Challenges of Predicting Pathogenicity from Sequence (Vol. 2). https://www.ncbi.nlm.nih.gov/books/NBK50869/

Nguyen, Y., & Sperandio, V. (2012). Enterohemorrhagic E. coli (EHEC) pathogenesis. Frontiers in Cellular and Infection Microbiology, 2, 90-90. https://doi.org/10.3389/fcimb.2012.00090

Nielsen, L. R., Schukken, Y. H., Gröhn, Y. T., & Ersbøll, A. K. (2004). Salmonella Dublin infection in dairy cattle: risk factors for becoming a carrier. Prev Vet Med, 65(1-2), 47-62. https://doi.org/10.1016/j.prevetmed.2004.06.010

Nolan, L. M., Turnbull, L., Katrib, M., Osvath, S. R., Losa, D., Lazenby, J. J., & Whitchurch, C. B. (2020). Pseudomonas aeruginosa is capable of natural transformation in biofilms. Microbiology, 166(10), 995-1003. https://doi.org/https://doi.org/10.1099/mic.0.000956

Novais, Â., Rodrigues, C., Branquinho, R., Antunes, P., Grosso, F., Boaventura, L., Ribeiro, G., & Peixe, L. (2012). Spread of an OmpK36-modified ST15 Klebsiella pneumoniae variant during an outbreak involving multiple carbapenem-resistant Enterobacteriaceae species and clones. European Journal of Clinical Microbiology & Infectious Diseases, 31(11), 3057-3063. https://doi.org/10.1007/s10096-012-1665-z

Nygaard, A. B., Tunsjø, H. S., Meisal, R., & Charnock, C. (2020, 2020/02/21). A preliminary study on the potential of Nanopore MinION and Illumina MiSeq 16S

rRNA gene sequencing to characterize building-dust microbiomes. Scientific Reports, 10(1), 3209. https://doi.org/10.1038/s41598-020-59771-0

O'Rourke, A., Beyhan, S., Choi, Y., Morales, P., Chan, A. P., Espinoza, J. L., Dupont, C. L., Meyer, K. J., Spoering, A., Lewis, K., Nierman, W. C., & Nelson, K. E. (2020). Mechanism-of-Action Classification of Antibiotics by Global Transcriptome Profiling. Antimicrobial agents and chemotherapy, 64(3), e01207-01219. https://doi.org/10.1128/aac.01207-19

Oakeson, K. F., Wagner, J. M., Rohrwasser, A., & Atkinson-Dunn, R. (2018). Whole-Genome Sequencing and Bioinformatic Analysis of Isolates from Foodborne Illness Outbreaks of Campylobacter jejuni and Salmonella enterica. Journal of Clinical Microbiology, 56(11), e00161-00118. https://doi.org/10.1128/JCM.00161-18

ONT. Products comparison. Retrieved 04/12/2021 from https://nanoporetech.com/products/comparison

Orlek, A., Phan, H., Sheppard, A. E., Doumith, M., Ellington, M., Peto, T., Crook, D., Walker, A. S., Woodford, N., Anjum, M. F., & Stoesser, N. (2017, 2017/05/01/). Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids. Plasmid, 91, 42-52. https://doi.org/https://doi.org/10.1016/j.plasmid.2017.03.002

Orlek, A., Stoesser, N., Anjum, M. F., Doumith, M., Ellington, M. J., Peto, T., Crook, D., Woodford, N., Walker, A. S., Phan, H., & Sheppard, A. E. (2017, 2017-February-09). Plasmid Classification in an Era of Whole-Genome Sequencing: Application in Studies of Antibiotic Resistance Epidemiology [Mini Review]. Frontiers in Microbiology, 8(182). https://doi.org/10.3389/fmicb.2017.00182

PacBio. (2021). Genomes vs. GenNNNes: The Difference between Contigs and Scaffolds in Genome Assemblies. Retrieved 04/16/2021 from https://www.pacb.com/blog/genomes-vs-gennnnes-difference-contigs-scaffolds-genome-assemblies/

Pallen, M. J., & Wren, B. W. (2007, Oct 18). Bacterial pathogenomics. Nature, 449(7164), 835-842. https://doi.org/10.1038/nature06248

Palmer, A. C., & Kishony, R. (2013). Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance. Nature Reviews Genetics, 14(4), 243-248. https://doi.org/10.1038/nrg3351

Partridge, S. R., Kwong, S. M., Firth, N., & Jensen, S. O. (2018). Mobile Genetic Elements Associated with Antimicrobial Resistance. Clinical Microbiology Reviews, 31(4), e00088-00017. https://doi.org/10.1128/CMR.00088-17

Passarelli-Araujo, H., Palmeiro, J. K., Moharana, K. C., Pedrosa-Silva, F., Dalla-Costa, L. M., & Venancio, T. M. (2019, 2019/10/01). Genomic analysis unveils important aspects of population structure, virulence, and antimicrobial resistance in Klebsiella aerogenes [https://doi.org/10.1111/febs.15005]. The FEBS Journal, 286(19), 3797-3810. https://doi.org/https://doi.org/10.1111/febs.15005

Pepper, I. L., & Gentry, T. J. (2015). Chapter 2 - Microorganisms Found in the Environment. In I. L. Pepper, C. P. Gerba, & T. J. Gentry (Eds.), Environmental Microbiology (Third Edition) (pp. 9-36). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-394626-3.00002-8

Petersen, R. F., Litrup, E., Larsson, J. T., Torpdahl, M., Sørensen, G., Müller, L., & Nielsen, E. M. (2011, 2011/06/01). Molecular Characterization of Salmonella Typhimurium Highly Successful Outbreak Strains. Foodborne Pathogens and Disease, 8(6), 655-661. https://doi.org/10.1089/fpd.2010.0683

Peterson, J. W. (1996). Bacterial Phatogenesis (B. S, Ed. 4th ed.). University of Texas Medical Branch at Galveston.

Podolsky, S. H. (2018, 2018/10/23). The evolving response to antibiotic resistance (1945–2018). Palgrave Communications, 4(1), 124. https://doi.org/10.1057/s41599-018-0181-x

Poirel, L., Bonnin, R. A., & Nordmann, P. (2012). Genetic support and diversity of acquired extended-spectrum β-lactamases in Gram-negative rods. Infection, Genetics and Evolution, 12(5), 883-893. https://doi.org/https://doi.org/10.1016/j.meegid.2012.02.008

Pomposiello Pablo, J., Bennik Marjon, H. J., & Demple, B. (2001). Genome-Wide Transcriptional Profiling of theEscherichia coli Responses to Superoxide Stress and Sodium Salicylate. Journal of Bacteriology, 183(13), 3890-3902. https://doi.org/10.1128/JB.183.13.3890-3902.2001

Price, S. B., Wright, J. C., DeGraves, F. J., Castanie-Cornet, M.-P., & Foster, J. W. (2004). Acid resistance systems required for survival of Escherichia coli O157:H7 in the bovine gastrointestinal tract and in apple cider are different. Applied and Environmental Microbiology, 70(8), 4792-4799. https://doi.org/10.1128/AEM.70.8.4792-4799.2004

Prieto, A., Urcola, I., Blanco, J., Dahbi, G., Muniesa, M., Quirós, P., Falgenhauer, L., Chakraborty, T., Hüttener, M., & Juárez, A. (2016, 2016/05/12). Tracking bacterial virulence: global modulators as indicators. Scientific Reports, 6(1), 25973. https://doi.org/10.1038/srep25973

Puente, J. L., & Finlay, B. B. (2001). CHAPTER 9 - Pathogenic Escherichia coli. In E. A. Groisman (Ed.), Principles of Bacterial Pathogenesis (pp. 387-456). Academic Press. https://doi.org/https://doi.org/10.1016/B978-012304220-0/50010-8

Puttamreddy, S., Carruthers, M. D., Madsen, M. L., & Minion, F. C. (2008, Aug). Transcriptome analysis of organisms with food safety relevance. Foodborne Pathog Dis, 5(4), 517-529. https://doi.org/10.1089/fpd.2008.0112

Quainoo, S., Coolen, J. P. M., van Hijum, S. A. F. T., Huynen, M. A., Melchers, W. J. G., van Schaik, W., & Wertheim, H. F. L. (2017). Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. Clinical Microbiology Reviews, 30(4), 1015. https://doi.org/10.1128/CMR.00016-17

Quijada, N. M., Hernández, M., & Rodríguez-Lázaro, D. (2020). Chapter Seven - High-throughput sequencing and food microbiology. In F. Toldrá (Ed.), Advances in Food and Nutrition Research (Vol. 91, pp. 275-300). Academic Press. https://doi.org/https://doi.org/10.1016/bs.afnr.2019.10.003

Rabinovitz, B. C., Gerhardt, E., Tironi Farinati, C., Abdala, A., Galarza, R., Vilte, D. A., Ibarra, C., Cataldi, A., & Mercado, E. C. (2012, 2012/06/01/). Vaccination of pregnant cows with EspA, EspB, γ-intimin, and Shiga toxin 2 proteins from Escherichia coli O157:H7 induces high levels of specific colostral antibodies that are transferred to newborn calves. Journal of Dairy Science, 95(6), 3318-3326. https://doi.org/https://doi.org/10.3168/jds.2011-5093

Rabsch, W. (2007). Salmonella typhimurium phage typing for pathogens. Methods Mol Biol, 394, 177-211. https://doi.org/10.1007/978-1-59745-512-1_10

Rawlings, D. E., & Tietze, E. (2001). Comparative biology of IncQ and IncQ-like plasmids. Microbiology and molecular biology reviews : MMBR, 65(4), 481-496. https://doi.org/10.1128/MMBR.65.4.481-496.2001

Read, T. D., Peterson, S. N., Tourasse, N., Baillie, L. W., Paulsen, I. T., Nelson, K. E., Tettelin, H., Fouts, D. E., Eisen, J. A., Gill, S. R., Holtzapple, E. K., Okstad, O. A., Helgason, E., Rilstone, J., Wu, M., Kolonay, J. F., Beanan, M. J., Dodson, R. J., Brinkac, L. M., Gwinn, M., DeBoy, R. T., Madpu, R., Daugherty, S. C., Durkin, A. S., Haft, D. H., Nelson, W. C., Peterson, J. D., Pop, M., Khouri, H. M., Radune, D., Benton, J. L., Mahamoud, Y., Jiang, L., Hance, I. R., Weidman, J. F.,

Berry, K. J., Plaut, R. D., Wolf, A. M., Watkins, K. L., Nierman, W. C., Hazen, A., Cline, R., Redmond, C., Thwaite, J. E., White, O., Salzberg, S. L., Thomason, B., Friedlander, A. M., Koehler, T. M., Hanna, P. C., Kolstø, A. B., & Fraser, C. M. (2003, May 1). The genome sequence of Bacillus anthracis Ames and comparison to closely related bacteria. Nature, 423(6935), 81-86. https://doi.org/10.1038/nature01586

Reis-Cunha, J. L., Bartholomeu, D. C., Manson, A. L., Earl, A. M., & Cerqueira, G. C. (2019). ProphET, prophage estimation tool: A stand-alone prophage sequence prediction tool with self-updating reference database. PloS one, 14(10), e0223364. https://doi.org/10.1371/journal.pone.0223364

Ribot, E. M., Freeman, M., Hise, K. B., & Gerner-Smidt, P. (2019). PulseNet: Entering the Age of Next-Generation Sequencing. Foodborne Pathogens and Disease, 16(7), 451-456. https://doi.org/10.1089/fpd.2019.2634

Rice, P. A., & Baker, T. A. (2001, 2001/04/01). Comparative architecture of transposase and integrase complexes. Nature Structural Biology, 8(4), 302-307. https://doi.org/10.1038/86166

Roberts, S. C., & Zembower, T. R. (2021). Global increases in antibiotic consumption: a concerning trend for WHO targets. The Lancet Infectious Diseases, 21(1), 10-11. https://doi.org/10.1016/S1473-3099(20)30456-4

Robinson, D. A., Turner, J. S., Facklam, R. R., Parkinson, A. J., Breiman, R. F., Gratten, M., Steinhoff, M. C., Hollingshead, S. K., Briles, D. E., & Crain, M. J. (1999). Molecular Characterization of a Globally Distributed Lineage of Serotype 12F Streptococcus pneumoniae Causing Invasive Disease. The Journal of Infectious Diseases, 179(2), 414-422. https://doi.org/10.1086/314589

Sabbagh, P., Rajabnia, M., Maali, A., & Ferdosi-Shahandashti, E. (2021). Integron and its role in antimicrobial resistance: A literature review on some bacterial pathogens. Iranian Journal of Basic Medical Sciences, 24(2), 136-142. https://doi.org/10.22038/ijbms.2020.48905.11208

Sabino, R. (2020). Exposure to Fungi in Health Care Facilities. In Reference Module in Life Sciences. Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-809633-8.21034-0

Sabino, Y. N. V., Santana, M. F., Oyama, L. B., Santos, F. G., Moreira, A. J. S., Huws, S. A., & Mantovani, H. C. (2019, 2019/11/20). Characterization of antibiotic resistance genes in the species of the rumen microbiota. Nature Communications, 10(1), 5252. https://doi.org/10.1038/s41467-019-13118-0

Sahl, J. W., Lemmer, D., Travis, J., Schupp, J. M., Gillece, J. D., Aziz, M., Driebe, E. M., Drees, K. P., Hicks, N. D., Williamson, C. H. D., Hepp, C. M., Smith, D. E., Roe, C., Engelthaler, D. M., Wagner, D. M., & Keim, P. (2016). NASP: an accurate, rapid method for the identification of SNPs in WGS datasets that supports flexible input and output formats. Microbial Genomics, 2(8), e000074-e000074. https://doi.org/10.1099/mgen.0.000074

Sarowska, J., Futoma-Koloch, B., Jama-Kmiecik, A., Frej-Madrzak, M., Ksiazczyk, M., Bugla-Ploskonska, G., & Choroszy-Krol, I. (2019). Virulence factors, prevalence and potential transmission of extraintestinal pathogenic Escherichia coli isolated from different sources: recent reports. Gut Pathog, 11, 10. https://doi.org/10.1186/s13099-019-0290-0

Scallan, E., Hoekstra, R. M., Angulo, F. J., Tauxe, R. V., Widdowson, M.-A., Roy, S. L., Jones, J. L., & Griffin, P. M. (2011). Foodborne illness acquired in the United States--major pathogens. Emerging infectious diseases, 17(1), 7-15. https://doi.org/10.3201/eid1701.p11101

Scallan, E., Hoekstra, R. M., Angulo, F. J., Tauxe, R. V., Widdowson, M. A., Roy, S. L., Jones, J. L., & Griffin, P. M. (2011). Foodborne illness acquired in the United States--major pathogens. Emerging infectious diseases, 17(1), 7-15. https://doi.org/10.3201/eid1701.p11101

Scharff, R. L., Besser, J., Sharp, D. J., Jones, T. F., Peter, G.-S., & Hedberg, C. W. (2016, 2016/05/01/). An Economic Evaluation of PulseNet: A Network for Foodborne Disease Surveillance. American Journal of Preventive Medicine, 50(5, Supplement 1), S66-S73. https://doi.org/https://doi.org/10.1016/j.amepre.2015.09.018

Scheutz, F., Teel, L. D., Beutin, L., Piérard, D., Buvens, G., Karch, H., Mellmann, A., Caprioli, A., Tozzoli, R., Morabito, S., Strockbine, N. A., Melton-Celsa, A. R., Sanchez, M., Persson, S., & O'Brien, A. D. (2012). Multicenter Evaluation of a Sequence-Based Protocol for Subtyping Shiga Toxins and Standardizing Stx Nomenclature. Journal of Clinical Microbiology, 50(9), 2951-2963. https://doi.org/10.1128/jcm.00860-12

Schirone, M., Visciano, P., Tofalo, R., & Suzzi, G. (2019). Editorial: Foodborne Pathogens: Hygiene and Safety. Frontiers in Microbiology, 10, 1974-1974. https://doi.org/10.3389/fmicb.2019.01974

Schmieder, R. (2013). PRINSEQ. Retrieved 04/19/2021 from http://prinseq.sourceforge.net/index.html

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics, 27(6), 863-864. https://doi.org/10.1093/bioinformatics/btr026

Seemann, T. (2014, Jul 15). Prokka: rapid prokaryotic genome annotation. Bioinformatics, 30(14), 2068-2069. https://doi.org/10.1093/bioinformatics/btu153

Sharma, A. K., Dhasmana, N., Dubey, N., Kumar, N., Gangwal, A., Gupta, M., & Singh, Y. (2017). Bacterial Virulence Factors: Secreted for Survival. Indian journal of microbiology, 57(1), 1-10. https://doi.org/10.1007/s12088-016-0625-1

Sharma-Kuinkel, B. K., Rude, T. H., & Fowler, V. G., Jr. (2016). Pulse Field Gel Electrophoresis. Methods in molecular biology (Clifton, N.J.), 1373, 117-130. https://doi.org/10.1007/7651_2014_191

Shintani, M., Sanchez, Z. K., & Kimbara, K. (2015). Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. Frontiers in Microbiology, 6, 242-242. https://doi.org/10.3389/fmicb.2015.00242

Shore Anna, C., Deasy Emily, C., Slickers, P., Brennan, G., O'Connell, B., Monecke, S., Ehricht, R., & Coleman David, C. (2011). Detection of Staphylococcal Cassette Chromosome mec Type XI Carrying Highly Divergent mecA, mecI, mecR1, blaZ, and ccr Genes in Human Clinical Isolates of Clonal Complex 130 Methicillin-Resistant Staphylococcus aureus. Antimicrobial Agents and Chemotherapy, 55(8), 3765-3773. https://doi.org/10.1128/AAC.00187-11

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31(19), 3210-3212. https://doi.org/10.1093/bioinformatics/btv351

Smith, T., Wolff, K. A., & Nguyen, L. (2013). Molecular biology of drug resistance in Mycobacterium tuberculosis. Current topics in microbiology and immunology, 374, 53-80. https://doi.org/10.1007/82_2012_279

Snyder, L., & Champness, W. (2007). Molecular genetics of bacteria. In D. C. A. P. Washington (Ed.). Washington, D.C: ASM Press.

Song, W., Sun, H.-X., Zhang, C., Cheng, L., Peng, Y., Deng, Z., Wang, D., Wang, Y., Hu, M., Liu, W., Yang, H., Shen, Y., Li, J., You, L., & Xiao, M. (2019). Prophage Hunter: an integrative hunting tool for active prophages. Nucleic acids research, 47(W1), W74-W80. https://doi.org/10.1093/nar/gkz380

Spellberg, B., & Gilbert, D. N. (2014). The Future of Antibiotics and Resistance: A Tribute to a Career of Leadership by John Bartlett. Clinical Infectious Diseases, 59(suppl_2), S71-S75. https://doi.org/10.1093/cid/ciu392

Strathdee, S. A., Davies, S. C., & Marcelin, J. R. (2020). Confronting antimicrobial resistance beyond the COVID-19 pandemic and the 2020 US election. The Lancet, 396(10257), 1050-1053. https://doi.org/10.1016/S0140-6736(20)32063-8

Sun, Y., Hu, X., Guo, D., Shi, C., Zhang, C., Peng, X., Yang, H., & Xia, X. (2019, 2019/06/01). Disinfectant Resistance Profiles and Biofilm Formation Capacity of Escherichia coli Isolated from Retail Chicken. Microbial Drug Resistance, 25(5), 703-711. https://doi.org/10.1089/mdr.2018.0175

Surachat, K., Kantachote, D., Deachamag, P., & Wonglapsuwan, M. (2021). Genomic Insight into Pediococcus acidilactici HN9, a Potential Probiotic Strain Isolated from the Traditional Thai-Style Fermented Beef Nhang. Microorganisms, 9(1), 50. https://www.mdpi.com/2076-2607/9/1/50

Tamber, S., & Hancock, R. E. W. (2003). On the mechanism of solute uptake in Pseudomonas. Frontiers in Bioscience-Landmark, 8(6), 472-483. https://doi.org/10.2741/1075

Tang, S., Orsi, R. H., Luo, H., Ge, C., Zhang, G., Baker, R. C., Stevenson, A., & Wiedmann, M. (2019). Assessment and Comparison of Molecular Subtyping and Characterization Methods for Salmonella. Frontiers in Microbiology, 10, 1591-1591. https://doi.org/10.3389/fmicb.2019.01591

Tanner, J. R., & Kingsley, R. A. (2018). Evolution of Salmonella within Hosts. Trends in Microbiology, 26(12), 986-998. https://doi.org/10.1016/j.tim.2018.06.001

Timme, R. E., Lafon, P. C., Balkey, M., Adams, J. K., Wagner, D., Carleton, H., Strain, E., Hoffmann, M., Sabol, A., Rand, H., Lindsey, R., Sheehan, D., Baugher, J. D., & Trees, E. (2020, 2020/11/19). Gen-FS coordinated proficiency test data for genomic foodborne pathogen surveillance, 2017 and 2018 exercises. Scientific Data, 7(1), 402. https://doi.org/10.1038/s41597-020-00740-7

Timme, R. E., Rand, H., Sanchez Leon, M., Hoffmann, M., Strain, E., Allard, M., Roberson, D., & Baugher, J. D. (2018). GenomeTrakr proficiency testing for foodborne pathogen surveillance: an exercise from 2015. Microbial Genomics, 4(7). https://doi.org/https://doi.org/10.1099/mgen.0.000185

Timme, R. E., Rand, H., Shumway, M., Trees, E. K., Simmons, M., Agarwala, R., Davis, S., Tillman, G. E., Defibaugh-Chavez, S., Carleton, H. A., Klimke, W. A., & Katz, L. S. (2017). Benchmark datasets for phylogenomic pipeline validation,

applications for foodborne pathogen surveillance. PeerJ, 5, e3893. https://doi.org/10.7717/peerj.3893

Timme, R. E., Sanchez Leon, M., & Allard, M. W. (2019). Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback. Methods Mol Biol, 1918, 201-212. https://doi.org/10.1007/978-1-4939-9000-9_17

To, C. Z., & Bhunia, A. K. (2019, 2019-May-07). Three Dimensional Vero Cell-Platform for Rapid and Sensitive Screening of Shiga-Toxin Producing Escherichia coli [Original Research]. Frontiers in Microbiology, 10(949). https://doi.org/10.3389/fmicb.2019.00949

Tolar, B., Joseph, L. A., Schroeder, M. N., Stroika, S., Ribot, E. M., Hise, K. B., & Gerner-Smidt, P. (2019). An Overview of PulseNet USA Databases. Foodborne Pathogens and Disease, 16(7), 457-462. https://doi.org/10.1089/fpd.2019.2637

Tolar, B., Joseph, L. A., Schroeder, M. N., Stroika, S., Ribot, E. M., Hise, K. B., & Gerner-Smidt, P. (2019). An Overview of PulseNet USA Databases. Foodborne pathogens and disease, 16(7), 457-462. https://doi.org/10.1089/fpd.2019.2637

Uelze, L., Grützke, J., Borowiak, M., Hammerl, J. A., Juraschek, K., Deneke, C., Tausch, S. H., & Malorny, B. (2020, 2020/02/18). Typing methods based on whole genome sequencing data. One Health Outlook, 2(1), 3. https://doi.org/10.1186/s42522-020-0010-1

University Of Maryland Medical Center. (1998). Intestinal Bug Likely Killed Alexander The Great. ScienceDaily. www.sciencedaily.com/releases/1998/06/980622061325.htm

Urmi, U. L., Nahar, S., Rana, M., Sultana, F., Jahan, N., Hossain, B., Alam, M. S., Mosaddek, A. S. M., McKimm, J., Rahman, N. A. A., Islam, S., & Haque, M. (2020). Genotypic to Phenotypic Resistance Discrepancies Identified Involving β-Lactamase Genes, blaKPC, blaIMP, blaNDM-1, and blaVIM in Uropathogenic Klebsiella pneumoniae. Infection and drug resistance, 13, 2863-2875. https://doi.org/10.2147/IDR.S262493

Utturkar, S. M., Klingeman, D. M., Hurt, R. A., & Brown, S. D. (2017, 2017-July-18). A Case Study into Microbial Genome Assembly Gap Sequences and Finishing Strategies [Original Research]. Frontiers in Microbiology, 8(1272). https://doi.org/10.3389/fmicb.2017.01272

van Asten, A. J. A. M., & van Dijk, J. E. (2005). Distribution of "classic" virulence factors among Salmonella spp. FEMS Immunology & Medical Microbiology, 44(3), 251-259. https://doi.org/10.1016/j.femsim.2005.02.002

Van Boeckel, T. P., Gandra, S., Ashok, A., Caudron, Q., Grenfell, B. T., Levin, S. A., & Laxminarayan, R. (2014). Global antibiotic consumption 2000 to 2010: an analysis of national pharmaceutical sales data. The Lancet Infectious Diseases, 14(8), 742-750. https://doi.org/10.1016/S1473-3099(14)70780-7

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014, 2014/09/01/). Ten years of next-generation sequencing technology. Trends in Genetics, 30(9), 418-426. https://doi.org/https://doi.org/10.1016/j.tig.2014.07.001

van Hoek, A. H. A. M., Mevius, D., Guerra, B., Mullany, P., Roberts, A. P., & Aarts, H. J. M. (2011). Acquired antibiotic resistance genes: an overview. Frontiers in Microbiology, 2, 203-203. https://doi.org/10.3389/fmicb.2011.00203

Vaser, R., & Šikić, M. (2020). Raven: a de novo genome assembler for long reads. bioRxiv, 2020.2008.2007.242461. https://doi.org/10.1101/2020.08.07.242461

Ventola, C. L. (2015). The antibiotic resistance crisis: part 1: causes and threats. P & T : a peer-reviewed journal for formulary management, 40(4), 277-283. https://pubmed.ncbi.nlm.nih.gov/25859123

Ventola, C. L. (2015, Apr). The antibiotic resistance crisis: part 1: causes and threats. P t, 40(4), 277-283.

Villa, L., García-Fernández, A., Fortini, D., & Carattoli, A. (2010). Replicon sequence typing of IncF plasmids carrying virulence and resistance determinants. Journal of Antimicrobial Chemotherapy, 65(12), 2518-2529. https://doi.org/10.1093/jac/dkq347

Wadapurkar, R. M., & Vyas, R. (2018). Computational analysis of next generation sequencing data and its applications in clinical oncology. Informatics in Medicine Unlocked, 11, 75-82. https://doi.org/https://doi.org/10.1016/j.imu.2018.05.003

Wang, Y. (2017, 2017/05/01/). Spatial distribution of high copy number plasmids in bacteria. Plasmid, 91, 2-8. https://doi.org/https://doi.org/10.1016/j.plasmid.2017.02.005

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., Gillespie, J. J., Gough, R., Hix, D., Kenyon, R., Machi, D., Mao, C., Nordberg, E. K., Olson, R., Overbeek, R., Pusch, G. D., Shukla, M., Schulman, J., Stevens, R. L., Sullivan, D. E., Vonstein, V., Warren, A., Will, R., Wilson, M. J. C., Yoo, H. S., Zhang, C., Zhang, Y., & Sobral, B. W. (2013). PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic acids research, 42(D1), D581-D591. https://doi.org/10.1093/nar/gkt1099

Weirather, J., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X., Buck, D., & Au, K. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis [version 2; peer review: 2 approved]. F1000Research, 6(100). https://doi.org/10.12688/f1000research.10571.2

Wentz, T. G., Hu, L., Hammack, T. S., Brown, E. W., Sharma, S. K., & Allard, M. W. (2019). Next Generation Sequencing for the Detection of Foodborne Microbial Pathogens. In S. K. Singh & J. H. Kuhn (Eds.), Defense Against Biological Attacks: Volume II (pp. 311-337). Springer International Publishing. https://doi.org/10.1007/978-3-030-03071-1_14

Whiley, H., & Ross, K. (2015). Salmonella and eggs: from production to plate. International journal of environmental research and public health, 12(3), 2543-2556. https://doi.org/10.3390/ijerph120302543

WHO. (2009). PulseNet International. INFOSAN Information Note No. 4/2009.

WHO. (2018). Salmonella (non-typhoidal). https://www.who.int/news-room/fact-sheets/detail/salmonella-(non-typhoidal)

WHO. (2020). Food Safety. https://doi.org/https://www.who.int/news-room/fact-sheets/detail/food-safety

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLOS Computational Biology, 13(6), e1005595. https://doi.org/10.1371/journal.pcbi.1005595

Wick, R., & Holt, K. (2021). Benchmarking of long-read assemblers for prokaryote whole genome sequencing [version 4; peer review: 4 approved]. F1000Research, 8(2138). https://doi.org/10.12688/f1000research.21782.4

Williams, T. L., & Moret, B. M. E. (2003, 12-12 March 2003). An investigation of phylogenetic likelihood methods. Third IEEE Symposium on Bioinformatics and Bioengineering, 2003. Proceedings.,

Wong, C. S., Jelacic, S., Habeeb, R. L., Watkins, S. L., & Tarr, P. I. (2000). The risk of the hemolytic-uremic syndrome after antibiotic treatment of Escherichia coli O157:H7 infections. The New England journal of medicine, 342(26), 1930-1936. https://doi.org/10.1056/NEJM200006293422601

Woolhouse, M., Ward, M., van Bunnik, B., & Farrar, J. (2015). Antimicrobial resistance in humans, livestock and the wider environment. Philosophical transactions of the

Royal Society of London. Series B, Biological sciences, 370(1670), 20140083-20140083. https://doi.org/10.1098/rstb.2014.0083

Wright, G. D. (2005). Bacterial resistance to antibiotics: Enzymatic degradation and modification. Advanced Drug Delivery Reviews, 57(10), 1451-1470. https://doi.org/https://doi.org/10.1016/j.addr.2005.04.002

Wu, G., Carter, B., Mafura, M., Liebana, E., Woodward, M. J., & Anjum, M. F. (2008). Genetic Diversity among Escherichia coli O157:H7 Isolates and Identification of Genes Linked to Human Infections. Infection and Immunity, 76(2), 845. https://doi.org/10.1128/IAI.00956-07

Xavier, B. B., Das, A. J., Cochrane, G., De Ganck, S., Kumar-Singh, S., Aarestrup, F. M., Goossens, H., & Malhotra-Kumar, S. (2016). Consolidating and Exploring Antibiotic Resistance Gene Data Resources. Journal of Clinical Microbiology, 54(4), 851. https://doi.org/10.1128/JCM.02717-15

Xu, J., Kiesel, B., Kallies, R., Jiang, F.-L., Liu, Y., & Maskow, T. (2018). A fast and reliable method for monitoring of prophage-activating chemicals. Microbial biotechnology, 11(6), 1112-1120. https://doi.org/10.1111/1751-7915.13042

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics, 13(5), 329-342. https://doi.org/10.1038/nrg3174

Yang, X., Sun, H., Fan, R., Fu, S., Zhang, J., Matussek, A., Xiong, Y., & Bai, X. (2020, 2020/02/24). Genetic diversity of the intimin gene (eae) in non-O157 Shiga toxin-producing Escherichia coli strains in China. Scientific Reports, 10(1), 3275. https://doi.org/10.1038/s41598-020-60225-w

Yoon, S. H., Park, Y. K., & Kim, J. F. (2015, Jan). PAIDB v2.0: exploration and analysis of pathogenicity and resistance islands. Nucleic acids research, 43(Database issue), D624-630. https://doi.org/10.1093/nar/gku985

Yoshimura, D., Kajitani, R., Gotoh, Y., Katahira, K., Okuno, M., Ogura, Y., Hayashi, T., & Itoh, T. (2019, May). Evaluation of SNP calling methods for closely related bacterial isolates and a novel high-accuracy pipeline: BactSNP. Microb Genom, 5(5). https://doi.org/10.1099/mgen.0.000261

Zerbino, D. R. (2010). Using the Velvet de novo assembler for short-read sequencing technologies. Current protocols in bioinformatics, Chapter 11, Unit-11.15. https://doi.org/10.1002/0471250953.bi1105s31

Zhang, R., & Zhang, C. T. (2006, May). The impact of comparative genomics on infectious disease research. Microbes Infect, 8(6), 1613-1622. https://doi.org/10.1016/j.micinf.2005.11.019

Zhang, Y., Zhang, Z., Zhang, H., Zhao, Y., Zhang, Z., & Xiao, J. (2020). PADS Arsenal: a database of prokaryotic defense systems related genes. Nucleic acids research, 48(D1), D590-D598. https://doi.org/10.1093/nar/gkz916

Zheng, L.-L., Li, Y.-X., Ding, J., Guo, X.-K., Feng, K.-Y., Wang, Y.-J., Hu, L.-L., Cai, Y.-D., Hao, P., & Chou, K.-C. (2012). A comparison of computational methods for identifying virulence factors. PloS one, 7(8), e42517-e42517. https://doi.org/10.1371/journal.pone.0042517

Zhou, K., Aertsen, A., & Michiels, C. W. (2014, Jan). The role of variable DNA tandem repeats in bacterial adaptation. FEMS Microbiol Rev, 38(1), 119-141. https://doi.org/10.1111/1574-6976.12036

Zwietering, M. H., Jacxsens, L., Membré, J.-M., Nauta, M., & Peterz, M. (2016, 2016/02/01/). Relevance of microbial finished product testing in food safety management. Food Control, 60, 31-43. https://doi.org/https://doi.org/10.1016/j.foodcont.2015.07.002

CHAPTER III

RAPID IDENTIFICATION AND MOLECULAR CHARACTERIZATION OF *ESCHERICHIA COLI* ISOLATES FROM FOOD AND ENVIRONMENT THROUGH NANOPORE SEQUENCING

ABSTRACT

Whole genome sequencing is becoming the tool for various applications, thereby this study was conducted to evaluate the performance of Nanopore sequencing for rapid identification and molecular characterization of *E. coli*. Eleven *E. coli* isolates obtained from pecan orchards were sequenced using MinION and Illumina NextSeq 500. As MinION allows real-time reads analysis, the reads were time-based subsampled to determine the earliest identification turnaround time for each isolate. Species level identification was achieved at 15 mins of sequencing run. In 16 hours, complete antigenic profile and variants of the virulence genes *eae* and *stx* were detected from assemblies obtained from the subsampled reads. Additionally, comparisons of the Nanopore-based assemblies against hybrid assemblies from the combined total reads from MinION and Illumina showed that the best values of continuity and annotated features were obtained in just 4 and 16 hours of sequencing run, respectively ($p < 0.05$). A stringent BLASTn search (percentage of identity of 95% and query coverage of 85 %) against the Comprehensive Antibiotic Resistance Database (CARD) and Virulence Factor

Database (VFDB) using the time-based assembled reads revealed that neither these datasets were sufficient to generate results significantly similar to those obtained from the hybrid assemblies. Nevertheless, an average of 87.5% and 78.3% of the hits acquired from the hybrid assemblies using the CARD and VFDB, respectively, were achieved with the assembles obtained after 4 hours with no significant changes were observed after 4 hours compared to the full datasets ($p < 0.05$). Finally, phylogeny analysis results obtained from assemblies created with reads produced in 3 hours of sequencing run, were significantly similar to those of the results with hybrid genomes ($p < 0.05$). These results demonstrated that Nanopore can offer an effective sequencing platform for the rapid identification of *E. coli* isolates with pathogenic potential, with additional certain capabilities for their characterization.

Key words: Whole genome sequencing, Nanopore, MinION, Illumina, Shiga toxin-producing *E. coli*

INTRODUCTION

The high genetic plasticity of *Escherichia coli* is why we can find a wide variety of *E. coli* strains with mechanisms to thrive and colonize different ecological niches, thus allowing this particular microorganism to inhabit the large intestine of humans and other animals as a small but constant part of their normal microbiota (Gordon & Cowling, 2003; Proença et al., 2017; Russo & Johnson, 2009). Despite the majority of *E. coli* strains have acquired defensive functions for the adaptation to a specific host either to avoid being recognized by the host's immune system or to modulate it (Ho Sui et al., 2009; Sokurenko et al., 1998), some of them harbor additional offensive elements which increase their level of pathogenicity (Chen et al., 2005), among these including genes

encoding adhesins, invasins and toxins (Chen et al., 2005; Ho Sui et al., 2009). Apart from these virulence factors (VFs), there are other genes or clusters of genes that contribute to greater survival of these microorganisms under unfavorable conditions, providing them resistance to antibiotics (Ventola, 2015), or making some strains more prevalent in harsh environments (Adzitey et al., 2020; Eltai et al., 2018; Li & Gänzle, 2016). In reality, these elements can jump between bacteria (horizontal transfer of mobile genetic elements), making these traits increase in places where there is a strong selective pressure working as a driving force of bacterial evolution, such as the constant use of antibiotics in livestock and poultry (Martin et al., 2015; Shea, 2003; Woods et al., 2020). As most VFs are often carried by plasmids or prophages, VFs are the main elements of horizontal gene transfer (Gyles & Boerlin, 2013; Ho Sui et al., 2009; Rankin et al., 2011).

The U.S. public health agencies are in charge of surveillance programs of foodborne pathogens including Shiga-toxin producing *E. coli* (STEC), which seek to trace back to the origin of the outbreak as soon as possible to take prompt measures to reduce the number of people affected by the contaminated food (Scallan et al., 2011). The use of whole genome sequencing (WGS) to subtype foodborne bacterial pathogens in outbreak surveillance began as a pilot project, which elucidated the higher discrimination resolution that can be achieved with this technology compared to pulsed-field gel electrophoresis (PFGE) (Jackson et al., 2016). Thereby, PulseNet, i.e., the national molecular subtyping network for foodborne disease surveillance, has adopted WGS as the primary subtyping tool for surveillance of *Listeria*, *Salmonella*, *E. coli*, *Shigella*, and *Campylobacter* in 2019 (Tolar et al., 2019). Consequently, apart from superior

discrimination for the identification of pathogens, it is now also possible to obtain greater and more detailed information regarding the virulence profile, resistance to antibiotics, serotypes, and even other characteristics that can be taken into consideration for disinfection programs (Forbes et al., 2017). Currently, PulseNet and GenomeTrakr (i.e., an international genomic reference database of mostly food and environmental isolates from foodborne pathogens) use the Illumina sequencing platform MiSeq (Sekse et al., 2017; Timme et al., 2020; Timme et al., 2019), which yields short paired-end reads with a maximum length of 300bp (Wentz et al., 2019), and a median error rate of 0.473% (Stoler & Nekrutenko, 2021). However, despite short reads have high accuracy, they can also pose difficulties when assembling complete genomes, thus introducing ambiguities that cause difficulties in understanding the correct organization of genomes, a feature also known as synteny (Margos et al., 2017; Orlek et al., 2017; Sharma et al., 2019). Such limitation can be disadvantageous when determining whether genes are co-regulated or transmissible, such as the case of genes located within mobile genetic elements (Gyles & Boerlin, 2013; Ho Sui et al., 2009).

Currently, there are sequencing platforms that produce longer reads but of lower accuracy than short reads technologies (Taylor et al., 2019). However, with the continuous evolution of bioinformatics and the chemistry implemented in these platforms, the accuracy of long reads sequencers have been improving thus giving rise to devices that, apart from providing the benefits of reads that span more nucleotides, also seek to lower the costs of sequencing and reduce the turnaround time in pathogens detection (Feng et al., 2015; Taylor et al., 2019). MinION (Oxford Nanopore Technologies, Oxford, UK) is a portable device that seeks to comply with these

characteristics, allowing organisms to be sequenced in a simpler setting and also allowing the generated reads to be analyzed in real time (Taylor et al., 2019). Thereby in this study, we aimed to evaluate the performance of Nanopore sequencing platform, for the rapid identification and molecular characterization of *E. coli* isolated from food and the environment.

MATERIALS AND METHODS

Bacterial isolates

The *E. coli* isolates used in this study were isolated from environmental samples collected from pecan orchards (pecans, soil and animal feces) located in Oklahoma as part of the doctoral dissertation conducted by (Diaz-Proano, 2019). Bacterial isolates were grown following the methodology used by the FDA Bacteriological Analytical Manual (FDA-BAM) (Andrews et al., 2018) with minor modifications. A general procedure was used to enrich the samples prior to individual isolation. In brief, 10 g of each soil and fecal sample were added to 90 mL of Universal pre-enrichment broth (UPB) (Becton-Dickinson, Sparks, Maryland) and stomached in a filter bag (Whirlpak) using a Seward Stomacher® 400 (Seward, London, United Kingdom) circulator for 1 min at 230 RPM. For pecan samples, 25 g of pecan were added to 225 mL of UPB in a filter bag (Whirlpak) and massaged by hand for 1 min. Suspensions were incubated for 24 h at 37°C, thereafter the enriched samples were streaked in parallel onto CHROMagar STEC (CHROMagar, Paris, France) and Rainbow agar O157 (Biolog, Hayward, California), and incubated for 24 h at 37°C.

The presence of the specific of STEC genes (*uidA*, *stx1*, *stx2*, and *eae*) was used to detect and confirm the presence of STEC by multiplex PCR of DNA extracted at two stages using the boiling method described by Kawasaki et al. (2005). In the first stage, DNA was obtained from 1 ml of 24-hour enrichment broth, and secondly, DNA was extracted from up to ten colonies from CHROMagar STEC and Rainbow agar O15. Primer pairs were chosen according to target genes described in the literature and 16S rRNA gene was used as an internal control (Diaz-Proano, 2019). Purified and confirmed STEC were transferred to Tryptic Soy Agar (TSA) (Becton-Dickinson, Sparks, Maryland) and stored at 4°C.

DNA extraction and whole genome sequencing

As part of Diaz-Proano (2019) doctoral dissertation, isolates were cultured in 5 mL tryptic soy broth (TSB, Difco, Sparks, MD) at 37°C for 18-20 h. Following overnight incubation, cells were harvested by centrifugation at 12000 rpm for 3 min and re-suspended in 1X buffered peptone water (BPW). DNA extraction was performed using the DNeasy 96 blood and tissue kit (Qiagen, Valencia, CA) according to the manufacturer's recommendation for Gram-negative bacteria and high-throughput applications. DNA samples were then cleaned using a DNA clean up and concentrator kit (Zymo Research). The quality of the DNA was determined using NanoDrop 1000 - OD 260/280 and OD 260/230 - (Thermo Scientific, Rockford, IL), and the concentration was determined using the Qubit 3 fluorometer with double-stranded DNA BR assay kit (Life Technologies, Grand Island, NY) according to each manufacturer's instructions. WGS was performed using Illumina and Oxford Nanopore Technologies (ONT) platforms. For Illumina sequencing, libraries were prepared using the Nextera XT DNA sample

preparation kit with the NextSeq®500 high output kit (2*150 bp paired-end reads) (Illumina, Inc., San Diego, CA) and sequenced at the Oklahoma State University Center for Genomics and Proteomics facility (Stillwater, Oklahoma). Whereas, Nanopore sequencing libraries were prepared using the Rapid Barcoding Sequencing kit (SQK-RBK004) and run on the MinION sequencing system (ONT, Oxford, UK) following the standard 48 h 1D sequencing protocol in the MinKNOW software (ONT, Oxford, UK) using three different FLO-MIN106 R.9.4.1 flowcells (Appendix 1).

Analysis of whole genome sequencing data

Initial data processing

For Illumina reads, Trimmomatic (version 0.32) (Bolger et al., 2014) was used to remove barcodes and to trim the sequences with a window size of 4 and a Q score cutoff of 20 (Del Fabbro et al., 2013), and FastQC (version 0.11.9) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc) was used for quality control. MinION reads were basecalled with Guppy (version 3.0.3) (ONT, Oxford, UK) and first filtered by EPI2ME (version 2.48) (ONT, Oxford, UK) with a Q score cutoff of 7. Later, "Passed" reads were demultiplexed with Porechop (version 0.2.4) (https://github.com/rrwick/Porechop) for further additional filtering with Filtlong (version 0.2.0) (https://github.com/rrwick/Filtlong) using a Q score cutoff of 9 (Tyler et al., 2018; Wick et al., 2017a). To analyze the Nanopore sequencing data over time, filtered reads were subsampled using a custom Perl script at intervals from start of the sequencing: at 15, 30, 60, 120, 240, 480, 960 and 1500 min (Taylor et al., 2019).

Species identification, typing and virulence gene detection using clean raw reads of nanopore sequencing

All bioinformatics analyzes were carried out on a Dual Intel Xeon Gold 6130 High Performance Cluster (HPC) with 2.8 GB max memory per core. In order to verify the presence of genes specific only to *E. coli* in Nanopore reads, a species confirmation step was performed using Kraken2 (version 2.0.7) (Wood & Salzberg, 2014) in default mode (kmer size of 35 and minimizer length of 31) with a confidence threshold of 0.05 (Ye et al., 2019) against a custom RefSeq database (i.e., containing Archaeal, bacterial, viral and plasmids reference sequences) created on January 2020. For the identification of serotypes, sequencing reads were analyzed through Serotypefinder (version 2.0.1) (Katrine G. Joensen et al., 2015) with a minimum gene coverage of 60% and minimum identity percentage of 90%. Additionally, a custom database containing different variants for the major virulence factors (VFs) *eae* and *stx* genes was constructed from reference genomes obtained from Genbank (Bai et al., 2018; Fu et al., 2018; Joensen et al., 2014; Ooka et al., 2012). Due to a high redundancy in the generated database, KMA (version 1.2.21) (Clausen et al., 2018) was used to map Nanopore reads against VFs variants sequences with the settings "-mrs 0.75 -gapopen -5 -gapextend of -1 –penalty -3 -reward 1 –e 1.0".

Genome assembly

Nanopore reads were assembled with Flye (version 2.6) (Kolmogorov et al., 2019) in nano-raw mode, with an expected genome size of 5.5Mb, and asm coverage (reduced coverage for initial disjointig assembly) of 50. Furthermore, in order to get the most out of the data obtained from this technology, a polishing step was carried out using Rebaler (version 0.2.0) (https://github.com/rrwick/Rebaler) a pipeline that uses minimap2

to align long reads to an already assembled genome and Racon (version 1.0.) (https://github.com/isovic/racon) for creating consensus sequences. Finally, an extra polishing step with Medaka (version 1.4.1) (https://github.com/nanoporetech/medaka), a tool that can create consensus sequences from Nanopore sequencing data using neural networks applied to a "read pileup" against a draft sequence using a variety of trained models, the "r941_min_fast_g303" model was used for this step. All the final assemblies generated were subjected to a preliminary quality assessment through QUAST (version 5.0.2) (Gurevich et al., 2013) and BUSCO (version 4.1.4) (Mathieu Seppey et al., 2019) against the "enterobacterales_odb10" database.

For further comparisons, hybrid assemblies were generated using the reads from both Illumina and Nanopore for each isolate (Appendix 2). Two approaches were used and compared to select the best assembly for each isolate in terms of continuity (i.e., N50, L50, and the total number of contigs), and their BUSCO score. In the first approach, Illumina paired-end reads were mapped against the genomes generated directly from the aforementioned process with the complete dataset of Nanopore reads, thereafter samtools (version 1.10) (Danecek et al., 2021) was used to sort and index the aligned reads, and finally, an extra polishing step was included using Pilon (version 1.23) (Walker et al., 2014) with the setting "–changes" for which a custom bash script performed multiple runs until no further changes left in the polished genomes. In the second approach, assemblies obtained first from Illumina reads and later bridged with long reads were obtained using the Unicycler pipeline (version 0.4.8) (Wick et al., 2017b) in default mode, with the aforementioned Bowtie2, samtools, and Pilon versions, as well as, Blast+ (version 2.10.1) (Camacho et al., 2009). Nanopore and Illumina reads were mapped

against the best hybrid assembly obtained for each isolate using minimap2 and Bowtie2, respectively, then mapped reads indexed via samtools were used to calculate the average coverage for both reads dataset using mosdepth (version 0.3.1) (https://github.com/brentp/mosdepth) with the parameters "-n --fast-mode --by 500". Additionally, the average nucleotide identity (ANI) from all generated assemblies from Nanopore reads against the best hybrid assemblies was calculated using OrthoANI (version 1.2) (Lee et al., 2016; Yoon et al., 2017) for each isolate.

Features annotation

Annotation of all assembled genomes was performed with Prokka pipeline (version 1.14.6) (Seemann, 2014) in default mode. To identify VFs and antibiotic resistance (AR) genes, assembled genomes were aligned against the Virulence Factors Database (VFDB) core dataset (VFDB_setA retrieved on January 9, 2020) (Chen et al., 2005), ResFinder database (retrieved on December 16, 2020) (Zankari et al., 2012), and Comprehensive Antibiotic Resistance Database (CARD) protein homologs dataset for acquired resistance genes (retrieved on March 10, 2020) (Alcock et al., 2020) using Blastn with an E-value cutoff of 1e-6. The obtained hits were filtered using a custom Python script with a minimum percentage of identity of 95%, and minimum gene coverage of 85%, as well as overlapping hits, were evaluated and removed based on an in-house strategy. From this search, two sets of genes of AR and VFs detected for each assembly were generated, thereafter the two sets obtained from each genome created from the Nanopore subsampled data were compared based on similarity with the two sets obtained from the best hybrid assemblies for each species respectively using cd-hit-est-2d with the parameters "-c 0.9 -n 8 -r 0 -G 1 -g 1 -b 20 -l 10 -s 0.0 -aL 0.0 -aS 0.0 -s2 1.0 -S2

0 -T 4 -M 32000 ". The pairs of similar genes generated by this comparison were counted and normalized based on the total number of genes present in the set of AR and VFs genes of each best hybrid assembly, respectively. A one-sample Wilcoxon signed rank test with Benjamini-Hochberg correction for multiple tests was used to check if the normalized values were close to 1, where $p > 0.05$ indicates that the set of genes analyzed is significantly similar to the set of genes of the best hybrid assembly for a given isolate.

Furthermore, the process of serotype identification and identification of allelic variants of the *stx1*, *stx2* and *eae* genes was repeated using the generated assemblies as inputs, although this time Blastn was used for the later analysis with the same parameters and filtering as the search described above. For the detection of mobile genetic elements, replicons were identified using PlasmidFinder (version 2.0.1) (Carattoli & Hasman, 2020) with a minimum percentage of identity of 85%, and a minimum gene coverage of 70%. Whereas, prophages prediction was performed using ProphET (Reis-Cunha et al., 2019) using the GFF files obtained from the annotation process as an input. Clustered regularly interspaced short palindromic repeats (CRISPRs) were predicted with CRISPRCASidentifier (version 1.1.0) (Padilha et al., 2020) in default mode.

Finally, biosynthetic gene clusters (BGCs) were predicted using Antismash (version 5.1.2) (Blin et al., 2019) with default parameters; in which interleaved, chemical hybrid, and neighboring clusters were divided into individual clusters for their quantification.

Phylogenetic analysis

A matrix of core single nucleotide polymorphisms (SNPs) was generated among the 11 isolates using kSNP (version 3.1) (Gardner et al., 2015) with the assemblies generated from the subsampled reads as well as the best hybrid assemblies, in which first Kchooser was run twice for each dataset to find the best kmer size. The core SNPs were used as an input for the construction of the maximum likelihood phylogenies using RAxML (version 8.2.11) (Stamatakis, 2014) with the GTRCAT model, a Lewis ascertainment bias correction and 1000 bootstrap replicates. The resulting phylogenetic trees were rooted in the midpoint using NJplot (version 2.3) (Perrière & Gouy, 1996) and formatted using Figtree (version 1.4.4) (https://github.com/rambaut/figtree).

Statistical analyses

Genomic continuity comparison

A matrix containing the estimated length of the genome, the total number of contigs, the length of the largest contig, N50 and L50 for all the assemblies generated overtime was normalized using the Min-Max scale method, which was applied to grouped values with respect to the identity of the isolate from which they were obtained using the following equation:

$$X_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where X_{norm} is the normalized value, x_{max} and x_{min} are the maximum and minimum values, respectively, of a particular metric evaluated in the assemblies from the same isolate, while x is a value to be normalized that is in this same set of values as x_{max} and x_{mins} The normalized dataset was dimensional reduced by principle component analysis

(PCA) through the prcomp function in R (version 4.0.5), thereafter Kruskal Wallis and Dunn's Multiple Comparison Test with Benjamini-Hochberg correction was used to compare the values of PC1 obtained from each time interval, additionally the p values were corrected with respect to the weight of the variance covered by the PC1 (Fachada et al., 2016).

Genomic features comparison

A matrix containing the GC-content, the BUSCO score, the ANI, as well as the number of annotated genes, tRNA, CRISPRs, plasmids, prophages and BGCs for all the assemblies generated over time and from the best hybrid assemblies was normalized using the Min-Max scale method with the equation (1), which was applied to grouped values with respect to the identity of the isolate from which they were obtained. The normalized dataset was dimensional reduced by PCA and the aforementioned multivariate analysis was performed.

Statistical comparison of phylogenetic trees

The Kendall-Colijin test (Kendall & Colijn, 2016) described by Katz et al. (2017) was used to compare the topologies of the phylogenies generated from Nanopore subsampled reads with the phylogeny created from the hybrid assemblies using the R libraries treespace (Jombart et al., 2017) and phytools (Revell, 2012). A lambda value of 0, as well as, 100,000 random trees as a background distribution were employed for all pairwise tree comparisons. For which a Z-test was calculated and a $p < 0.05$ indicates that the pair of compared trees are significantly similar (Katz et al., 2017).

RESULTS AND DISCUSSION

Evaluation of sequencing reads yield

For Illumina, the average of the median length of the paired-end reads of all isolates was 149.3. The total bases ranged from 274.8 Mbps to 571.49 Mbps, obtaining mean depth values between 50.26x to 103.53x after filtering (Table 7). For Nanopore, the average of the median length of the reads of all isolates was 2,472.36. The total bases ranged from 108.42 Mbps to 3,333.3 Mbps, obtaining mean depth values between 17.72x to 663.34x after filtering (Table 8). The isolates with the lowest and highest mean depth values were G4M0F1_1 and G5BLF3_8 in Illumina, and G1M4F3_31 and G5M2P3_1 in Nanopore, respectively (Table 7, 8). The total number of subsampled reads that were generated at different stages of the sequencing run can be seen in table 8. From the subsampled Nanopore reads, we could observe how the differences in mean depth values remained proportional among barcoded samples over sequencing time (Table 9; Figure 1B). Additionally, through a paired Wilcoxon Signer-Rank test between the mean depth values from Illumina and Nanopore, we obtained that values from Nanopore were significantly similar to those generated by Illumina ($\alpha = 0.05$) (Figure 1A),

Table 7. Summary for Illumina sequencing of *E. coli* isolates (Paired-end reads).

Isolate	Number of raw reads	Total Bases (Mbps)	Mean length	Median length	Total reads after filtering	Mean depth
G1BLF1_5	3359124	453.69	138.11	150	3318448	80.3
G1BLF2_1	2711004	372.79	139.79	151	2683006	65.85
G1M0S3_4	3272080	443.70	138.39	151	3236848	77.27
G1M4F3_31	3140466	404.78	131.87	150	3095500	71.26
G4M0F1_1	2252238	274.80	123.83	139	2224566	50.26
G4M0F2_14	2521252	345.27	139.23	151	2495648	63.73
G5BLF1_1	3894406	523.72	137.78	150	3850210	93.01
G5BLF3_3	3464496	475.90	140.47	151	3410792	85.44
G5BLF3_8	4253992	571.49	138.01	150	4210078	103.53

G5M2P3_1	3867316	522.21	138.22	150	3833358	90.79
G5M4F2_1	2964582	407.88	140.39	151	2929626	69.17

Table 8. Summary for Nanopore sequencing of *E. coli* isolates and number of subsampled reads according to the time of their generation.

Isolate	Number of raw reads	Total bases (Mbps)	Mean length	Median length	Total reads after filtering	Mean depth
G1BLF1_5	120219	114.06	1318.95	239	86481	20.23
G1BLF2_1	364001	1650.47	5500.02	3503	300085	302.57
G1M0S3_4	470060	961.07	3020.98	1439	318130	162.92
G1M4F3_31	103280	108.42	1483.15	1018	73098	17.72
G4M0F1_1	422032	754.54	2879.48	2160	262040	140.1
G4M0F2_14	180835	863.66	5910.24	3410	146129	172.54
G5BLF1_1	358747	1509.82	5275.07	3241	286219	271.89
G5BLF3_3	88747	397.27	5402.53	3241	73534	73.18
G5BLF3_8	314605	532.58	3153.33	3058	168895	90.48
G5M2P3_1	706879	3333.30	5980.79	3612	557334	663.34
G5M4F2_1	83745	285.72	4190.73	2275	68178	50.73

Time	15 mins	30 mins	60 mins	120 mins	240 mins	480 mins	960 mins	1500 mins	Full
Number of reads	23251	51611	113899	243037	479229	876746	1490878	1941622	2340123

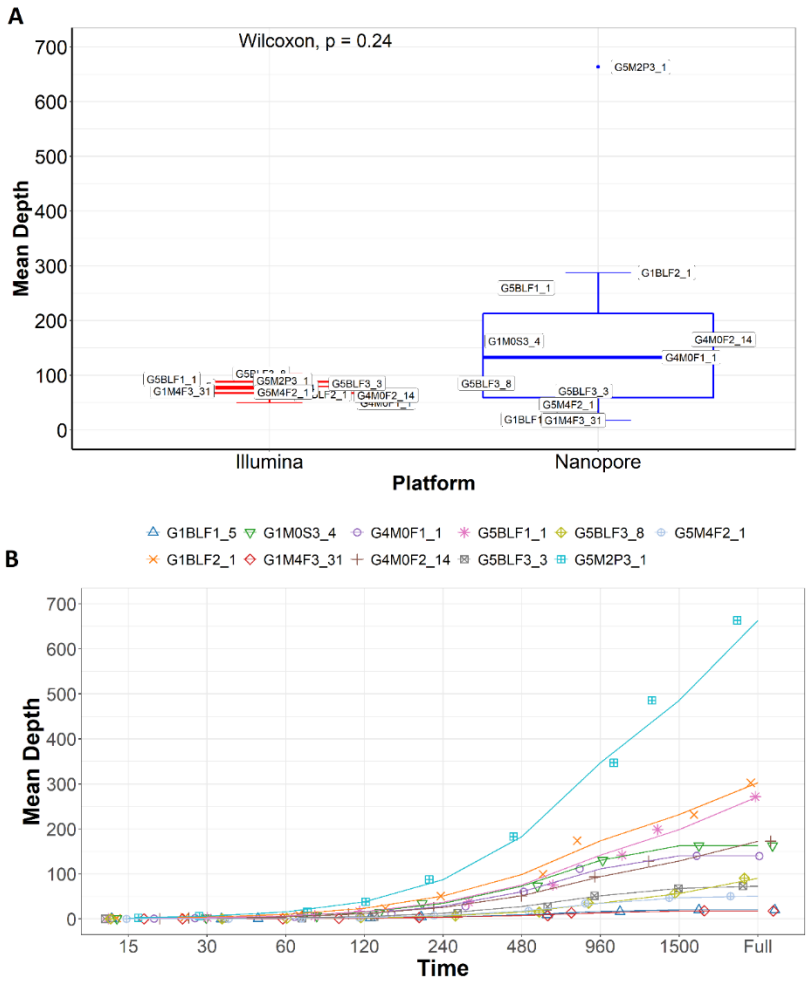

Figure 1. Mean depth of (A) the complete filtered reads set obtained from Illumina and Nanopore sequencing, and (B) the subsampled filtered Nanopore reads.

Table 9. Mean depth of the filtered Nanopore subsampled reads set.

Isolate	15 mins	30 mins	60 mins	120 mins	240 mins	480 mins	960 mins	1500 mins	Full
G1BLF1_5	0.14	0.34	0.78	1.96	4.45	9.15	16.45	20.23	20.23
G1BLF2_1	1.8	4.35	10.03	23.65	51.12	99.12	173.75	231.73	302.57
G1M0S3_4	1.02	2.55	6.25	15.03	34.86	72.94	130.54	162.92	162.92
G1M4F3_31	0.11	0.26	0.61	1.52	3.58	7.58	13.9	17.72	17.72
G4M0F1_1	0.74	1.93	4.88	11.92	27.92	60.45	111.36	140.1	140.1
G4M0F2_14	0.76	1.95	4.97	11.82	25.89	51.53	93.74	128.8	172.54
G5BLF1_1	1.14	2.7	6.71	16.28	37.11	75.37	141.55	198.64	271.89
G5BLF3_3	0.34	0.85	2.11	5.2	12.78	27.43	51.42	68	73.18
G5BLF3_8	0.19	0.49	1.2	2.86	6.7	15.32	35.17	56.32	90.48
G5M2P3_1	2.43	6.14	15.71	38.17	87.63	183.05	347.12	485.41	663.34
G5M4F2_1	0.17	0.45	1.29	3.22	8	17.75	34.46	46.9	50.73

Despite identical protocols for the DNA extraction , either for Illumina or for Nanopore platform, were followed across each isolate (Appendix 1), there are variations in the total number of reads generated between different barcoded samples as well as their mean depths, the value that refers to the average of the number of unique reads that cover a given nucleotide or region in a reference genome, for which we used the best hybrid assembly for each species as explained below (Table 7, 8) (Sims et al., 2014). Yielding different numbers of reads with variable mean depth values is considered common when sequencing pools of genomic DNA from different individuals (Kanamori et al., 2017; Muller et al., 2019), nonetheless the greater dispersion in the mean depth values obtained from the Nanopore reads (Figure 1A, B; Table 9) could be the result of the different degree of DNA fragmentation of the libraries prepared for Nanopore sequencing (Table 8), which might occur due to the implementation of an engineered transposase in the library preparation kits for the tagmentation process (Kia et al., 2017; Muller et al., 2019), as well as to mechanical stress during DNA extraction (Anchordoquy & Molina, 2007; Cinque et al., 2010), or even the result of how DNA was

stored (Röder et al., 2010; Ross et al., 1990). Wherein, although the same DNA was used to generate the Illumina reads, the concentrations of each extracted DNA were readjusted to equimolar values by the center of Genomics and Proteomics from Oklahoma State University, and not used in equal mass concentrations as recommended in the Rapid Barcoding Sequencing (SQK-RBK004) protocol for Nanopore sequencing. Pools containing equimolar concentrations are also used when preparing pools of DNA obtained from distant bacteria since they can have significant differences in genomic size, which limits reaching an acceptable coverage for all the microorganisms present in the pool (Muller et al., 2019; Salipante et al., 2015); however, performing this readjustment is limited to laboratories possessing the necessary equipment to precisely calculate the fragment size distribution of each sample included in the pool (Anand et al., 2016). Even though G1BLF1_5 and G1M4F3_31 had mean depths less than 45x in Nanopore reads, the recommended value for scaffolding bacterial genomes with this platform (Karlsson et al., 2015); we continued with the analyses to determine to what extent the data generated was sufficient for the identification and molecular characterization of our isolates.

Species iidentification, serotyping and major VFs allelic variants determination

Specific reads for species other than *E. coli* could be detected in only the Illumina reads set from the G4M0F1_1 isolate (1.05% classified as *Herbaspirillum huttiense*), but not in the Nanopore reads yielded from the same isolate (Figure 2A, B), wherein an average of 26.33% and 10.50% Illumina and Nanopore reads, respectively, were classified as *E. coli*. From the 15 mins subsampled Nanopore reads, it was possible to identify *E. coli* as the unique species in all the isolates (Figure 2C). According to the results we obtained using the Nanopore reads for serotype identification, from 240 mins

(4 h.) of the sequencing process, the O-antigens for all isolates could be predicted, yet the H-antigen of 3 isolates could not be identified even with the full set of Nanopore reads (Figure 3), while in the Illumina reads only 2 isolates presented this problem (Figure 3). Whereas, in the VFs allelic variants determination, the allelic variants of 3 isolates could not be identified not even with the full set of Nanopore reads (Figure 4); nonetheless for the isolate G5BLF1_1, the variant of the intimin gene could be identified neither with the reads of Illumina nor in the detection step that was performed further with the best hybrid assemblies.

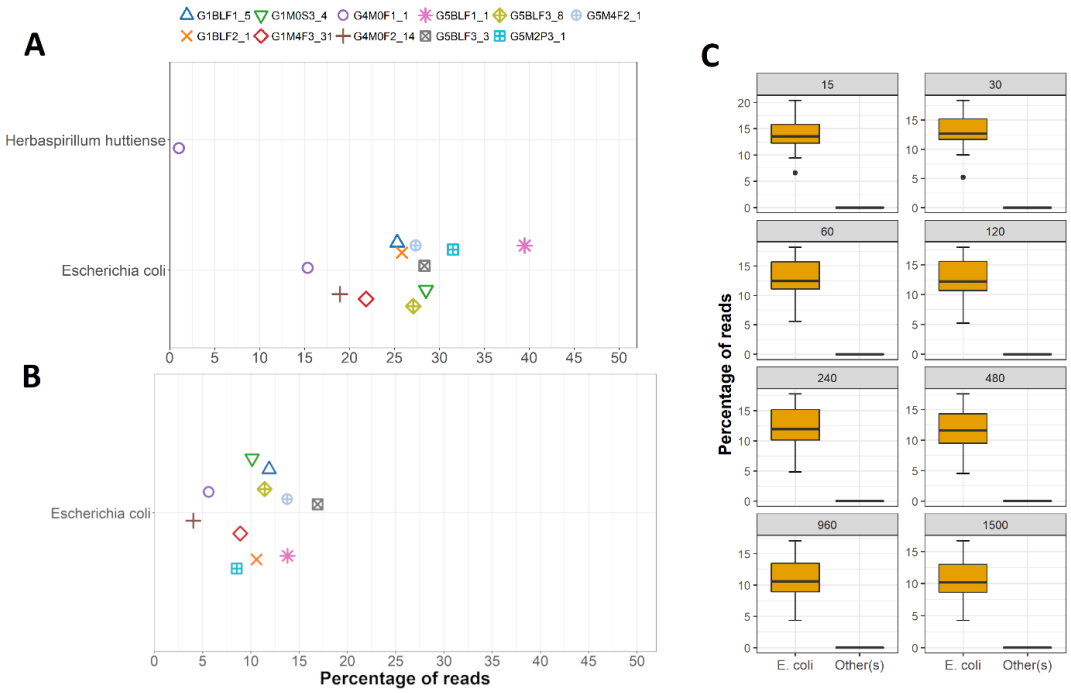

Figure 2. Percentage of reads classified to species level by Kraken2 in the filtered reads set from (A) Illumina and (B) Nanopore, as well as for (C) the subsampled filtered Nanopore reads.

		G1BLF1_5	G1BLF2_1	G1M0S3_4	G1M4F3_31	G4M0F1_1	G4M0F2_14	G5BLF1_1	G5BLF3_3	G5BLF3_8	G5M2P3_1	G5M4F2_1
Antigenic profile	**O**	168	168	108	178	8	91	157	130	109	188	103
	H	8	8	9	19	7	21	7	11	48	20	2
15 min.	**O**	-	-	-	-	-	-	-	-	+	-	-
	H	-	-	-	-	-	-	-	-	-	-	-
30 min.	**O**	-	-	-	-	-	-	-	-	+	+	-
	H	-	+	-	-	-	-	-	-	-	-	-
60 min.	**O**	+	-	-	-	-	+	+	+	+	+	-
	H	-	+	-	-	-	-	+	-	-	-	-
120 min.	**O**	+	-	+	-	-	+	+	+	+	+	+
	H	-	+	-	-	-	-	+	-	-	-	-
240 min.	**O**	+	+	+	+	+	+	+	+	+	+	+
	H	-	+	-	-	+	+	+	+	-	+	+
480 min.	**O**	+	+	+	+	+	+	+	+	+	+	+
	H	-	+	-	-	+	+	+	+	-	+	+
960 min.	**O**	+	+	+	+	+	+	+	+	+	+	+
	H	+	+	-	-	+	+	+	+	-	+	+
1500 min.	**O**	+	+	+	+	+	+	+	+	+	+	+
	H	+	+	-	-	+	+	+	+	-	+	+
Full	**O**	+	+	+	+	+	+	+	+	+	+	+
	H	+	+	-	-	+	+	+	+	-	+	+
Illumina	**O**	+	+	+	+	+	+	+	+	+	+	+
	H	+	+	-	+	+	+	+	+	-	+	+

Figure 3. A heat-map showing the antigenic profile of the *E. coli* isolates and from which set of reads it could be detected. O, O-antigen or somatic antigen; H, H-antigen or flagellar antigen; Full, the complete filtered Nanopore reads set.

		G1BLF1_5	G1BLF2_1	G1M0S3_4	G1M4F3_31	G4M0F1_1	G4M0F2_14	G5BLF1_1	G5BLF3_3	G5BLF3_8	G5M2P3_1	G5M4F2_1
Allelic variants	stx1	-	-	-	Sd*	-	-	-	Sd*	Sd*	-	A
	stx2	D	D	-	-	-	-	A	-	-	-	A
	eae	-	-	β1	-	-	-	N/A	-	-	-	ε1
15 min.	stx1	-	-	-	-	-	-	-	-	-	-	-
	stx2	-	-	-	-	-	-	-	-	-	-	-
	eae	-	-	-	-	-	-	-	-	-	-	-
30 min.	stx1	-	-	-	-	-	-	-	-	-	-	-
	stx2	-	-	-	-	-	-	+	-	-	-	-
	eae	-	-	-	-	-	-	-	-	-	-	-
60 min.	stx1	-	-	-	-	-	-	-	+	-	-	-
	stx2	-	+	-	-	-	-	+	-	-	-	-
	eae	-	-	-	-	-	-	-	-	-	-	-
120 min.	stx1	-	-	-	-	-	-	-	+	-	-	-
	stx2	-	+	-	-	-	-	+	-	-	-	-
	eae	-	-	-	-	-	-	-	-	-	-	-
240 min.	stx1	-	-	-	-	-	-	-	+	-	-	-
	stx2	-	+	-	-	-	-	+	-	-	-	-
	eae	-	-	-	-	-	-	-	-	-	-	-
480 min.	stx1	-	-	-	-	-	-	-	+	+	-	+
	stx2	-	+	-	-	-	-	+	-	-	-	+
	eae	-	-	-	-	-	-	-	-	-	-	-
960 min.	stx1	-	-	-	-	-	-	-	+	+	-	+
	stx2	-	+	-	-	-	-	+	-	-	-	+
	eae	-	-	-	-	-	-	-	-	-	-	-
1500 min.	stx1	-	-	-	-	-	-	-	+	+		+
	stx2	-	+	-	-	-	-	+	-	-	-	+
	eae	-	-	+	-	-	-	-	-	-	-	+
Full	stx1	-	-	-	-	-	-	-	+	+		+
	stx2	-	+	-	-	-	-	+	-	-	-	+
	eae	-	-	+	-	-	-	-	-	-	-	+
Illumina	stx1	-	-	-	+	-	-	-	+	+	-	+
	stx2	+	+	-	-	-	-	+	-	-	-	+
	eae	-	-	+	-	-	-	-	-	-	-	+

Figure 4. A heat-map showing allelic variants for the major VFs stx1, stx2 and eae identified in the *E. coli* isolates and from which set of reads each one could be detected. *stx1*, Shiga-toxin 1 gene; *stx2*, Shiga-toxin 2 gene; *eae*, intimin gene; Full, the complete filtered Nanopore reads set; Sd*, *stx1* variant from *Shigella dysenteriae*; N/A, allelic variant not identified despite the confirmed presence of the gene via multiplex PCR.

For the species identification, we used Kraken 2, i.e., a taxonomic classification system, which matches a list of k-mer within a DNA read query sequence to the lowest common ancestor of all genomes containing the given k-mer (Wood et al., 2019). The reads classified as a species other than *E. coli* in the isolate G4M0F1_1 with the Illumina reads (Figure 2A) could come from contamination present in the flowcells used during the Illumina sequencing process (Laurence et al., 2014), or even sequences acquired by horizontal gene transfer that is not classified in Nanopore due to its lower accuracy compared to Illumina (Tyler et al., 2018). However, another characteristic to note is that a lower portion of Nanopore reads was classified at the species level (Fig 2A, B) when compared to these of Illumina. As for long reads, such as these generated by Nanopore sequencing, , Kraken 2 needs to find a greater number of specific k-mers for that particular species to assign the read to that taxon, hence this decreases the percentage of Nanopore reads that can be found solely for *E. coli* and not other species in the same genera, a reasonable trade-off to increase the accuracy of this species classifier tool and avoid false positives as can happen with Illumina short reads (Leidenfrost et al., 2020; Pearman et al., 2020). As Walsh et al. (2018) described, the detection of species using this tool does not depend on the mean depth and this is reflected in the fact that *E. coli* was identified as the unique species in all the isolates using the 15 mins subsampled Nanopore reads (Figure 2C). It is worth mentioning that the average run time of this tool with all full reads sets was 2 s using 1 core (2.8 Gb).

Subtyping of *E. coli* based on O- and H-antigens agglutination tests with specific antisera was considered a gold standard technique to identify strains related to serious diseases or pathotypes for many decades (Fratamico et al., 2016; Kauffmann, 1947), but

this could take a long time and sometimes there were cross-reactions between variants of the same antigen (Fratamico et al., 2016; Lacher et al., 2014). Hand in hand with the development of PCR and sequencing techniques, new methods were developed based on the genotypic differences of the genes related to these antigens (Fratamico et al., 2016; Katrine G. Joensen et al., 2015), with the O polymerase and flippase (Wzx / Wzy) antigen genes being the most common cluster between serotypes (DebRoy et al., 2016; K. G. Joensen et al., 2015), in addition to the pathway mediated by the enzymes ABC permease transporter and ATP-binding cassette transporter (ABC) (Wzm / Wzt) (DebRoy et al., 2016; Greenfield & Whitfield, 2012), whereas for the H-antigen different allelic variants for the fliC gene, as well as, other less recurrent genes, such as flkA, fllA, flmA or flnA, are responsible for the production of flagellin, the structural subunit of bacterial flagella (Tominaga, 2004; Wang et al., 2003). This led to the generation of databases that store detailed information on these genes, thus opening the way to new typing strategies that involve the use of WGS (Carattoli & Hasman, 2020; Katrine G. Joensen et al., 2015; Salipante et al., 2015); which despite the fact that there are currently methods with higher resolution, e.g., the identification of wgSNPs or wgMLST (Gardner et al., 2015; Kingry et al., 2016; Miro et al., 2020), it can still be used rapidly to understand the relationship that an isolate can have with strains involved in an outbreak (Fratamico et al., 2016), since certain serotypes are often involved in severe diseases, such as O157:H7 or O103:H21, strains that have been commonly related to the enterohemorrhagic *E. coli* (EHEC) pathotype. As shown in Figure 3, we were only able to detect the O-antigen, but not the H-antigen, from all our isolates using the reads from both Nanopore and Illumina platforms, this impediment might be caused by the larger size of the allelic variants for

the genes involved with the H-antigen (Figure 5) and/or a greater number of genes part of the synthesis cluster of the O-antigen. Additionally, we can see the effect of mean depth in the isolates G1BLF1_5 and G1BLF2_1 which presented the same antigenic profile but had different reads yield (Table 8, Figure 5). For the isolate G1BLF1_5 (mean depth of 20.23x), we could not detect the H-antigen until 960 mins (16 h.) of sequencing, yet for the isolate G1BLF2_1 (mean depth of 302.57x), both antigens were detected in the 240 mins (4 h.) subsampled reads; which, unlike the species identification, serotype detection relies on a specific genes search, hence the greater the mean depth obtained, the greater the probability of detecting the desired targets.

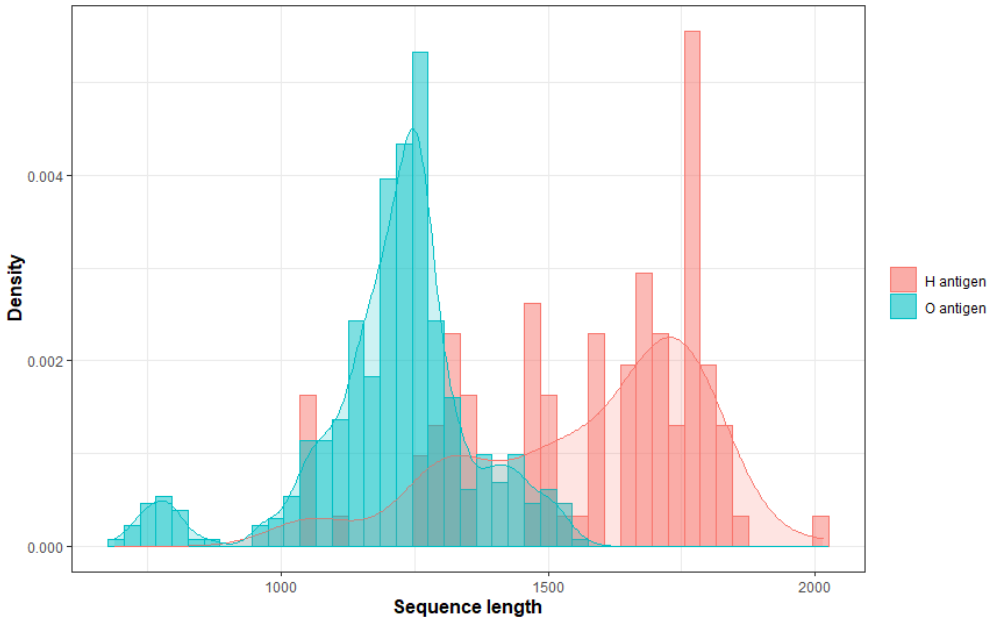

Figure 5. Distribution of gene length in the O- and H-antigen database used by Serotypefinder (retrieved in October 2020).

Among the pathotypes associated with *E. coli* foodborne illnesses, EHEC is responsible for producing more severe illnesses, i.e., bloody diarrhea, hemolytic uremic syndrome (HUS), and/or hospitalizations (Melton-Celsa, 2014; Panel et al., 2020). These strains owe their name to the presence of genes that encode Shiga toxin (Stx), one of the

most potent bacterial toxins known (Panel et al., 2020). The Stxs consist of two major subunits, an A subunit that joins non-covalently to a pentamer of five identical B subunits (Melton-Celsa, 2014). The A subunit of the toxin affects the eukaryotic ribosome and halts protein synthesis in target cells. The function of the B pentamer is to bind to the cellular receptor globotriaosylceramide (Gb3) of the host, found primarily on endothelial cells (Melton-Celsa, 2014). Subtypes of each toxin have been also identified, thereby, the Stx1a and Stx2a variants are called prototypes and there are three variants of Stx1 (Stx1a, Stx1c, and Stx1d (Melton-Celsa, 2014; Scheutz et al., 2012), and eight variants of Stx2 (Stx2a, Stx2b, Stx2c, Stx2d, Stx2e, Stx2f, Stx2g, and Stx2h) (Bai et al., 2018; Melton-Celsa, 2014; Scheutz et al., 2012). Stx1 variants produce mild effects compared to Stx2 variants, of which the Stx2a, Stx2c, and Stx2d have been more often associated with the hemolytic uremic syndrome (Bielaszewska et al., 2006; Persson et al., 2007), on the other hand, Stx2e, Stx2f, Stx2g and Stx2h are associated with animal STEC, of which only Stx2e produces disease in them (Bai et al., 2018; Feng & Reddy, 2013). In addition to Stx, the *eae* gene has been identified as a possible factor for high pathogenicity. This gene encodes a protein named intimin, an outer membrane protein expressed by enteric bacterial pathogens capable of inducing intestinal attachment-and-effacement lesions, a fundamental step in invading host cells (Hartland et al., 1999). Among E. coli serotypes that were most frequently reported in global dysentery and HUS cases, O157: H7 and O145: H28 serotypes are associated with the eae-γ1 subtype, whereas O26: H11 often carries *eae*-β1, O103: H2 and O121: H19 harbor *eae*-ε, and O111: H8 harbors *eae*-θ subtype (Bibbal et al., 2014; Ito et al., 2007). Thereby, *eae* and *stx* genes subtyping can be a valuable tool for risk assessment and prediction of disease outcome. For this reason,

we built two custom databases containing the allelic variants detected for *eae, and stx1-stx2* genes present in reference genomes of *E. coli* from Genbank. With a direct search using reads against these databases, we aimed to generate quick results that could represent a prompt response at the time of being identified in products or food production facilities. As shown in Figure 4, the VFs determinants search was affected by the underrepresentation of reads and low mean depth from the isolates G1BLF1_5 and G1M4F3_31 during the Nanopore sequencing process (20.23x and 17.72x, respectively), as we have already observed in serotype identification, mean depth can limit the search for specific genes using raw sequencing reads, especially if these reads are low precision like those of Nanopore with an error rate that ranges between 5% and 15% (Rang et al., 2018). For the search performed in the G5BLF1_1 reads, we could not detect its *eae* variant even in the further analysis we performed using the best hybrid genomes (see below); in this case, the reason was related to the great limitation of the use of databases, which are restricted to the information they contain and cannot be used directly to find variants not included in them, hindering the identification of novel variants and requiring constant database curations (Rhee, 2005). The prediction of serotypes and the determination of allelic variants of these major VFs using reads took an average of 10 s using 1 core (2.8 Gb).

Continuity of the Genome Assemblies

The biggest advantage of using long reads, such as Nanopore reads, is that highly repetitive regions of the genome of different sizes can be covered, allowing continuous genomes to be assembled (Kolmogorov et al., 2019; Leidenfrost et al., 2020). For this reason, we decided to evaluate at what point in the sequencing process using Nanopore, it

was possible to obtain continuity metrics significantly similar to each other; comparing the estimated genome size, the number of contigs, the size of the longest contig, N50 (i.e., the sequence length of the shortest contig that covers at least half the genome) and the L50 (i.e., the smallest number of contigs whose length sum constitutes half the genome) of the assemblies obtained from the subsampled reads in comparison of the assemblies obtained from the full Nanopore reads set (Kolmogorov et al., 2019). These values were also essential at the time of choosing the best hybrid assemblies from the two approaches performed, whereby it could be evidenced that the isolates with a mean depth lower than 45x, G1BLF1_5 and G1M4F3_31 with values of 20.23x and 17.72x (Table 9), respectively, had more contigs and presented a smaller size in their longest contig (Table 10), which proves that during the creation of the hybrid assemblies, the long reads were responsible to lead to more continuous assemblies as expected.

Table 10. Assembly metrics for the best hybrid assemblies obtained for each *E. coli* isolate.

Isolate	Number of contigs	Largest contig	Total length	GC (%)	N50	L50
G1BLF1_5	43	1347791	5368849	50.69	518023	4
G1BLF2_1	5	5314189	5487282	50.71	5314189	1
G1M0S3_4	25	4359963	5654877	50.45	4359963	1
G1M4F3_31	51	854945	5292468	50.45	384227	5
G4M0F1_1	7	2558830	5089896	50.57	2558830	1
G4M0F2_14	2	4685726	4891637	50.64	4685726	1
G5BLF1_1	3	5447391	5553524	50.47	5447391	1
G5BLF3_3	9	4983366	5500181	50.8	4983366	1
G5BLF3_8	9	2904903	5512584	50.55	2904903	1

G5M2P3_1	2	4602735	4719183	50.81	4602735	1
G5M4F2_1	38	4063509	5664540	50.64	4063509	1

For the comparison of the subsampled reads, due to the varied mean depth that we obtained for each isolate, we were able to obtain assemblies for all of our isolates from 60 mins, so we hereafter used this set of reads as the starting point for the subsequent comparisons. From 240 mins (4 h.), we already had assemblies significantly similar to those obtained with the full reads set ($\alpha = 0.05$) (Figure 5A, Table 11). Furthermore, taking into consideration that in the loading plots obtained from a transformation of a dataset by PCA, the correlation between the different variables graphed as vectors or arrows can be understood through the cosine of the angles between them (Økelsrud et al., 2016; Zitko, 1994), we could also infer that the length of the longest contig, the N50 and the total estimated size of the genome correlate with each other positively, but negatively with the L50 and the number of contigs (Figure 5B). Expected result, since there is a greater number of contigs when the assembly is more fragmented (Thrash et al., 2020); finally, the color and direction of the arrows in this graph allowed us to determine that both the N50 and the length of the longest contig were the variables that contributed most positively to the PC1 values in the multivariate comparison (Figure 5B).

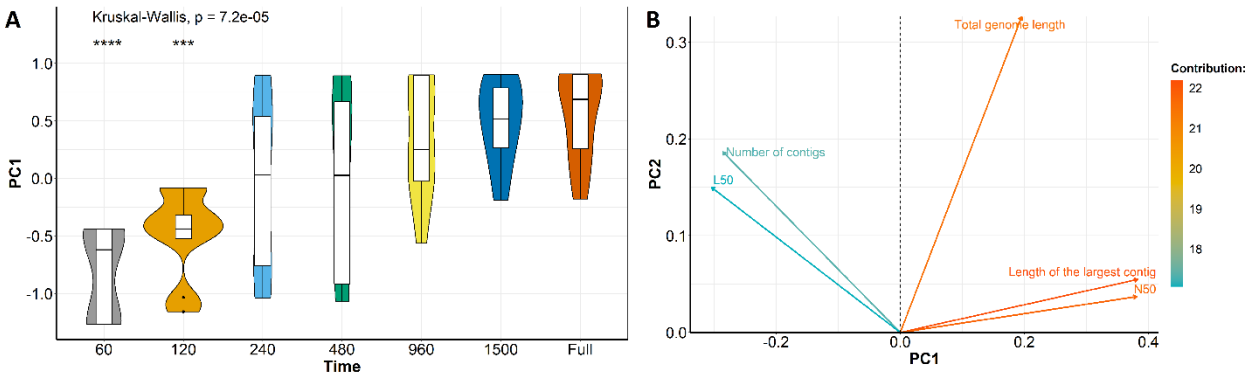

Figure 6. Multivariate comparisons based on continuity between assemblies created from subsampled filtered Nanopore reads. (A) Comparison between PC1 values associated with continuity of the assemblies obtained from the subsampled reads against the full set of filtered Nanopore reads with respect to time. (B) Contribution plot of the variables used for the multivariate analysis. The colors in the arrows represent the contribution weight for each variable, with a gradient from light blue (low contribution) to red (high contribution). PC1 and PC2 explain the 70% and 18.3% of the total variance of the complete data set, respectively. ****: p < 1E-03. ***: 1E-03 < p < 1E-02.

Table 11. *p*-values obtained from the Dunn's Multiple Comparison between the PC1 values associated with continuity of the assemblies obtained from the subsampled reads against the full set of filtered Nanopore reads with respect to time.

	Z	***p*-values**	**Corrected *p*-values**
60 mins	-4.22	2.4645E-05	7.5005E-04
120 mins	-3.30	9.4983E-04	7.2270E-03
240 mins	-2.04	4.0883E-02	1.3825E-01
480 mins	-1.81	6.9795E-02	2.1242E-01
960 mins	-0.80	4.2527E-01	7.1906E-01
1500 mins	-0.42	6.7129E-01	1

Annotation of genome assemblies

In Table 12, we can see the number of features that we were able to identify in the best hybrid assemblies. These features were employed in the multivariate comparison against the assemblies created from the subsampled Nanopore reads, whereby, from 960 mins (16 h.), we were able to obtain results significantly similar to those obtained from the best hybrid assemblies ($\alpha = 0.05$) (Figure 7A, Table 13). All the variables used had a positive correlation and contributed positively to PC1 (Figure 7B). Furthermore, after failing to identify all the serotypes and allelic variants of the major VFs described above with the filtered Nanopore reads (Figure 3, 4), we performed similar searches using the assemblies as input (see Materials and Methods), managing to detect all antigens and variants available even from subsampled reads that were obtained in previous points in comparison to the analyses performed with solely reads (Table 14).

Table 12. Number of annotated features for the best hybrid assemblies obtained for each *E. coli* isolate. rRNA, ribosomal RNA; tmRNA, transfer-messenger RNA; tRNA, transfer RNA; CRISPRs, clustered regularly interspaced short palindromic repeats; BGCs, biosynthetic gene clusters.

Isolate	Genes	Repeated regions	rRNA	tmRNA	tRNA	Plasmids	CRISPRs	BGCs	Prophages
G1BLF1_5	5243	2	6	1	87	3	8	4	10
G1BLF2_1	5386	2	22	1	93	4	7	4	10
G1M0S3_4	5560	2	22	1	96	2	12	3	13
G1M4F3_31	5102	2	3	1	82	5	8	4	9
G4M0F1_1	4877	2	18	1	87	3	8	3	5
G4M0F2_14	4927	2	22	1	86	2	9	1	8

G5BLF1_1	5537	1	22	1	103	2	12	6	18
G5BLF3_3	5474	2	22	1	89	7	10	3	13
G5BLF3_8	5607	2	23	1	99	2	9	4	17
G5M2P3_1	4654	2	22	1	92	2	8	4	5
G5M4F2_1	5667	1	22	1	102	1	9	5	21

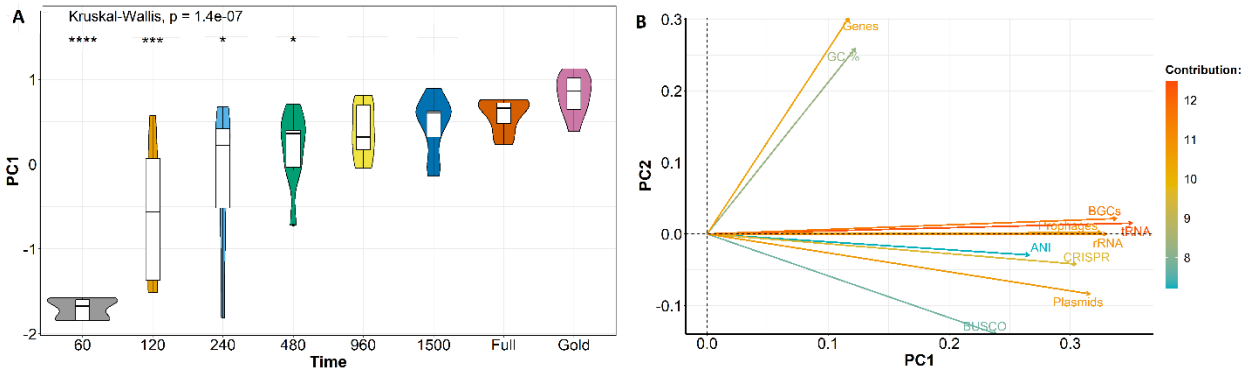

Figure 7. Multivariate comparisons based on genomic features between assemblies created from subsampled filtered Nanopore reads and best hybrid assemblies (Gold). (A) Comparison between PC1 values associated with genomic features of the assemblies obtained from the subsampled reads against the best hybrid assemblies with respect to time. (B) Contribution plot of the variables used for the multivariate analysis. The colors in the arrows represent the contribution weight for each variable, with a gradient from light blue (low contribution) to red (high contribution). PC1 and PC2 explain the 60.1% and 14.1% of the total variance of the complete data set, respectively. ****: p < 1E-04. ***: 1E-04 < p < 1E-03. **: 1E-03 < p < 1E-02. *: 1E-02 < p < 5E-02.

Table 13. *p*-values obtained from the Dunn's Multiple Comparison Test between the PC1 values associated with genomic features of the assemblies obtained from the subsampled reads against the best hybrid assemblies with respect to time.

	Z	*p*-values	Corrected *p*-values
60 mins	-5.71	1.1287E-08	5.2671E-07
120 mins	-4.12	3.7538E-05	5.8393E-04
240 mins	-3.21	1.3280E-03	1.0329E-02
480 mins	-2.76	5.7917E-03	3.0031E-02
960 mins	-2.08	3.7197E-02	1.3353E-01
1500 mins	-1.71	8.6911E-02	2.7039E-01
Full	-1.08	2.7960E-01	6.2134E-01

Table 14. Time in which the antigenic profile and allelic variants of the major VFs (*stx1*, *stx2* and *eae*) genes could be detected from assemblies produced from subsampled reads. "-" was placed in the isolates that do not harbor any of the major VFs analyzed.

Isolate	Antigenic profile	Major VFs
G1BLF1_5	480 mins	960 mins
G1BLF2_1	60 mins	60 mins
G1M0S3_4	60 mins	60 mins
G1M4F3_31	480 mins	960 mins
G4M0F1_1	60 mins	-
G4M0F2_14	120 mins	-
G5BLF1_1	60 mins	60 mins
G5BLF3_3	240 mins	120 mins
G5BLF3_8	480 mins	240 mins
G5M2P3_1	60 mins	-
G5M4F2_1	240 mins	120 mins

As discussed earlier, the assembly of sequencing reads can provide a greater understanding of genomic synteny. Quality that is important when evaluating the risk posed by the presence of certain strains harboring transmissible VFs, such as those already mentioned, or AR genes, genetic elements that are often transferred through plasmids or phages to neighbor bacteria (Ho Sui et al., 2009). Moreover, an assembly

with the minimum number of gaps can facilitate the identification and prediction of complex gene clusters, such as the biosynthetic gene clusters (BGCs) involved in the formation of biofilms by means of aryl polyene synthesis found in *E. coli* (Cimermancic et al., 2014; Johnston et al., 2020). Or even the detection of Clustered regularly interspaced short palindromic repeats (CRISPR), i.e., short, highly conserved DNA repeats separated by unique sequences of similar length, which have been used for subtyping, identification, and detection of STEC in epidemiological studies (Delannoy et al., 2016; Shariat & Dudley, 2014). All the genetic elements already mentioned were found in the eleven isolates tested (Table 12), making this ideal for the comparison performed between the best hybrid assemblies and the assemblies constructed from the subsampled reads. In this comparison, apart from the properties already described, we also included the ANI values and BUSCO scores, which indicate how identical the assembly generated is to a reference genome (Chen et al., 2020) and how complete this same genome is based on highly conserved housekeeping orthologs present in the family Enterobacteriaceae (M. Seppey et al., 2019), respectively.

The complete detection of antigens for the identification of serotypes and the identification of the allelic variants in the assemblies obtained from the subsampled Nanopore reads is another example of how the continuity can improve the identification capacity of essential features (Table 14); noteworthy, apart from obtaining continuous assemblies, the extra polishing steps implemented in custom assembly pipelines can improve the quality of the information obtained from Nanopore sequencing platform (Taylor et al., 2019; Wick et al., 2019), that is why we recommend the implementation of assembly strategies in conjunction with polishing tools when features annotation is

needed in assemblies obtained from Nanopore reads, such as the workflow we used. Yet an average of 60 mins was necessary to obtain the assembled genomes from full Nanopore reads with the implemented bioinformatics pipeline using 32 cores (89.6 Gb).

AR and VFs genes annotation

By searching for AR genes in the best hybrid assemblies (Figure 9A), as expected from *E. coli (*Kwak et al., 2015; Sharma et al., 2018*)*, we were able to elucidate a widespread variety of resistance genes to various antibiotics with a similar distribution among the isolates tested. However, there are two isolates (G1BLF1_5 and G1BLF2_1) in which AR additional genes against sulfone and sulfonamide antibiotics were detected (Figure 9A). Since these isolates were isolated from the same fecal sample at the same orchard, we believe that the recent administration of this type of antibiotics as treatment for a sick animal could have caused a selective pressure for these genes to persist in these two isolates (Ma et al., 2021).

Whereas through the performed VFs search, it was possible to identify VFs with a varied function, which may be related to the high presence of plasmids and prophages, essential components from *E. coli*'s mobilome that contribute to the high plasticity of this species (Table 12) (Delannoy et al., 2017; Mbelle et al., 2019). Among the detected VFs, adherence factors were present in all isolates in conjunction with secretion systems, and non-Locus of Enterocyte Effacement (LEE) encoded Type three secretion systems (TTSS) effectors (Figure 9B), genes that may be related to the survival of the organisms within the host as either commensals or pathogens (Frömmel et al., 2013; Govindarajan et al., 2020; Ritchie & Waldor, 2005).

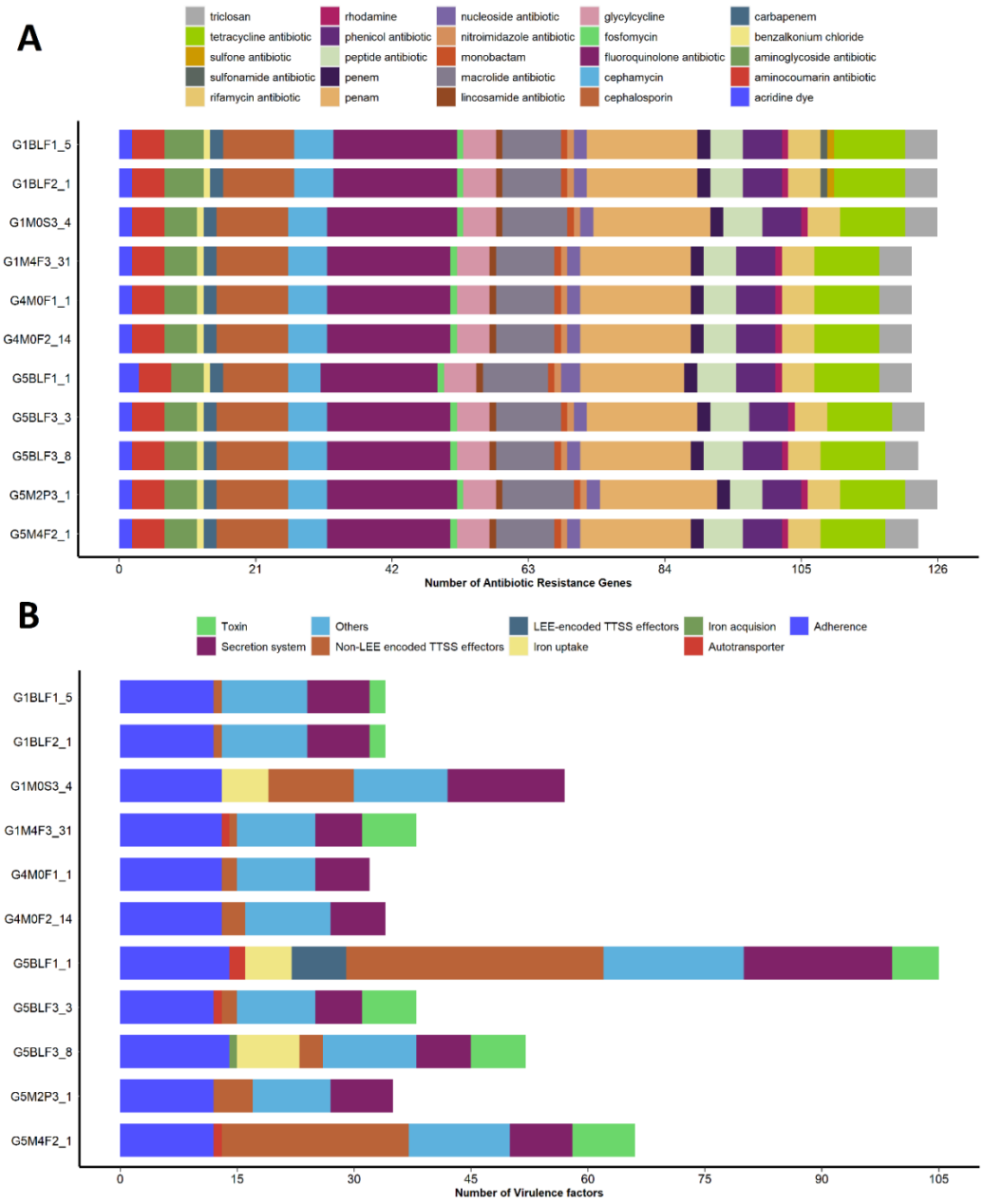

Figure 8. Number of gene ontologies associated with (A) the AR and (B) the VFs genes identified in the best hybrid genomes (gene ontology was analyzed using the aro_index.tsv file from the CARD and the intra-genera VFs comparison tables from the VFDB for AR and VFs, respectively). LEE: Locus of Enterocyte Effacement. TTSS: Type three secretion system.

The comparison of these searches carried out in the genomes of the subsampled reads was not enough to obtain results significantly similar to the results obtained from our best hybrid assemblies ($\alpha = 0.05$) (Table 15), although an average of 87.5% and 78.3% of the total hits found for the ARs genes and the VFs, respectively, were determined in the searches performed with the genomes created using Nanopore reads, whereby from 480 mins no significant differences were obtained between the hits detected from these genomes (Figure 9 A, B).

Table 15. *p*-values obtained from the one-sample Wilcoxon signed rank test of the identical hits ratio. $P > 0.05$ indicates that the results of the assemblies from the subsampled reads at that time are significantly similar to the best hybrid assemblies.

	AR genes		***Virulence factors***	
	***p*-values**	**Corrected *p*-values**	***p*-values**	**Corrected *p*-values**
60 mins	1.8422E-03	9.9863E-03	1.8851E-03	2.6879E-03
120 mins	2.9608E-03	9.9863E-03	1.8938E-03	2.6879E-03
240 mins	7.1331E-03	9.9863E-03	4.8828E-04	1.7090E-03
480 mins	4.5150E-03	9.9863E-03	4.8828E-04	1.7090E-03
960 mins	1.1127E-02	1.1247E-02	2.9608E-03	2.9608E-03
1500 mins	7.0145E-03	9.9863E-03	2.9608E-03	2.9608E-03
Full	1.1247E-02	1.1247E-02	1.9199E-03	2.6879E-03

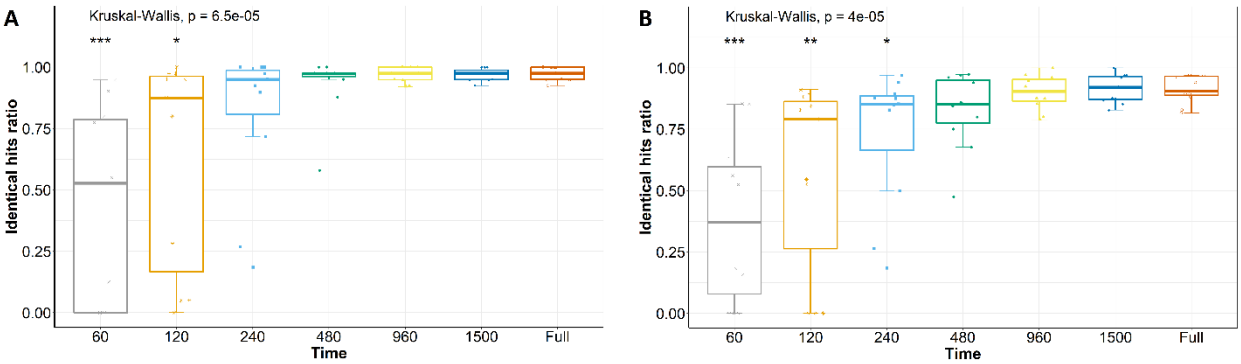

Figure 9. Identical hits ratio comparison between the genes obtained from assemblies created using the subsampled filtered Nanopore reads. (A) Comparison of the hits obtained from the search for AR genes.

(B) Comparison of the hits obtained from the search for VFs. ***: 1E-03 < *p*. **: 1E-03 *p* < 1E-02. *: 1E-02 < *p*-values < 5E-02.

Phylogenetic inference

The variety of serotypes identified and the genetic differences reflected in the results of the phylogenetic analysis carried out from the best hybrid assemblies are consistent with the fact that the populations of *E. coli* present in agriculture are highly diverse and dynamic (Figure 3, 10) (Doane et al., 2007; Marchant & Moreno, 2013; Naganandhini et al., 2015). From the subsampled reads, we could notice that from the reads subsampled at 240 mins, we began to have enough data in the core SNPs matrices to carry out the phylogenetic analyzes, so we decided to subsample the reads that were generated at 180 mins (3 h.), from which it was possible to perform the pertinent phylogenetic analysis and its subsequent comparison with the trees generated at later minutes. In the comparisons made, from 180 mins, we could obtain phylogenies significantly similar to those obtained from the best hybrid genomes ($\alpha = 0.05$) (Table 16); however, it is worth mentioning that we only took into account the topology of the trees and not the length of the branches using a lambda value of 0 in the distances calculations among trees when using the Kendall-Colijin test (Kendall & Colijn, 2016). Therefore, we believe that these results would not represent analyses where it is desired to quantify the amount of evolutionary divergence between isolates (Paradis, 2016). In addition, it took 5 hours on average to carry out the phylogenetic analyzes with the kSNP3-RAxML strategy using 16 cores (44.8 Gb).

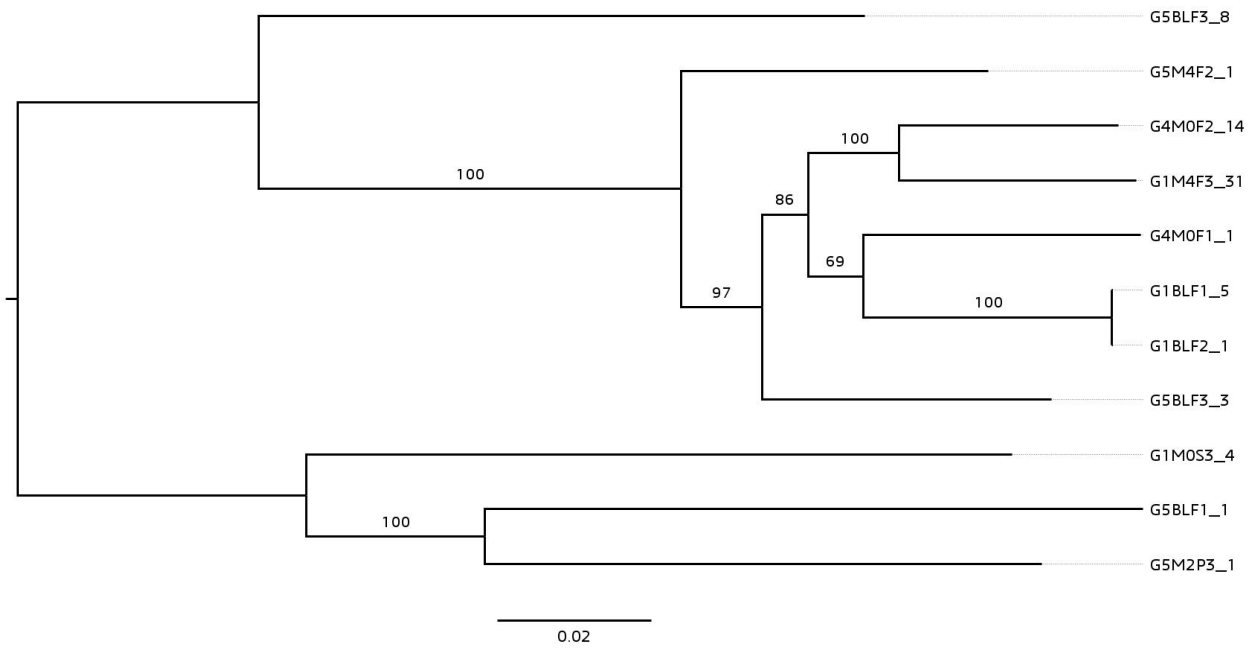

Figure 10. A maximum likelihood tree constructed using RAxML based on the core SNPs dataset of the best hybrid assemblies for the 11 *E. coli* isolates.

Table 16. *p*-values obtained from the Kendall-Colijin test between the topologies of the core SNPs phylogenetic trees generated from the subsampled filtered Nanopore reads and the best hybrid assemblies.

	***p*-values**	**Corrected *p*-values**
180 mins	7.42E-06	4.45E-05
240 mins	1.53E-11	9.17E-11
480 mins	3.88E-13	2.33E-12
960 mins	5.09E-15	3.06E-14
1500 mins	3.36E-22	2.02E-21
Full	3.36E-22	2.02E-21

SNP phylogenetics plays a role in outbreak monitoring, forensic investigations, inference of lineage evolution, and identification of mutations linked to phenotypes such as AR (Gardner & Hall, 2013); since unlike the phylogeny constructed from species-

specific protein-coding genes, e.g., cgMLST, this analysis includes several intergenetic regions that can harbor essential information on the evolutionary events to which a species has undergone (Davis et al., 2015; Li et al., 2016; Schürch et al., 2018). Traditionally, reference genomes close to the isolates to be analyzed are first chosen, to which the reads of a group of isolates are mapped and the relevant phylogenetic analyzes are performed preferably from the core SNPs that are identified in this step (Oakeson et al., 2018; Wentz et al., 2019). However, the choice of a reference genome limits its use when it is necessary to analyze isolates not directly related, even between organisms of the same species with different serotypes (Oakeson et al., 2018), valuable information from regions not present in the chosen reference genome can be omitted. However, by using kSNP3 we were able to overcome this limitation since this tool detects SNPs by directly comparing k-mers generated from the analyzed genomes, and does not require the establishment of a reference genome (Gardner et al., 2015).

CONCLUSIONS

Although in an outbreak it is necessary to use pathogen identification techniques with a low turnaround time, the Nanopore sequencing reads obtained from 3 hours in this study could be used to obtain phylogenetic analyzes with a higher resolution than other traditional techniques. And even at 16 hours, it can offer certain capabilities to detect fundamental elements in case it is required to characterize molecularly a group of isolates and identify potential reservoirs of pathogenic *E. coli*.

REFERENCES

Adzitey, F., Assoah-Peprah, P., Teye, G. A., Somboro, A. M., Kumalo, H. M., & Amoako, D. G. (2020). Prevalence and Antimicrobial Resistance of Escherichia coli Isolated from Various Meat Types in the Tamale Metropolis of Ghana. *International Journal of Food Science, 2020*, 8877196. https://doi.org/10.1155/2020/8877196

Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A. V., Cheng, A. A., Liu, S., Min, S. Y., Miroshnichenko, A., Tran, H. K., Werfalli, R. E., Nasir, J. A., Oloni, M., Speicher, D. J., Florescu, A., Singh, B., Faltyn, M., Hernandez-Koutoucheva, A., Sharma, A. N., Bordeleau, E., Pawlowski, A. C., Zubyk, H. L., Dooley, D., Griffiths, E., Maguire, F., Winsor, G. L., Beiko, R. G., Brinkman, F. S. L., Hsiao, W. W. L., Domselaar, G. V., & McArthur, A. G. (2020). CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res, 48*(D1), D517-d525. https://doi.org/10.1093/nar/gkz935

Anand, S., Mangano, E., Barizzone, N., Bordoni, R., Sorosina, M., Clarelli, F., Corrado, L., Martinelli Boneschi, F., D'Alfonso, S., & De Bellis, G. (2016). Next Generation Sequencing of Pooled Samples: Guideline for Variants' Filtering. *Scientific Reports, 6*(1), 33735. https://doi.org/10.1038/srep33735

Anchordoquy, T. J., & Molina, M. C. (2007). Preservation of DNA. *Cell Preservation Technology, 5*(4), 180-188. https://doi.org/10.1089/cpt.2007.0511

Andrews, W. H., Wang, H., Jacobson, A., & Hammack, T. (2018). Bacteriological Analytical Manual Chapter 5 Salmonella. In. https://www.fda.gov/food/foodscienceresearch/laboratorymethods/ucm070149.htm

Bai, X., Fu, S., Zhang, J., Fan, R., Xu, Y., Sun, H., He, X., Xu, J., & Xiong, Y. (2018). Identification and pathogenomic analysis of an Escherichia coli strain producing a novel Shiga toxin 2 subtype. *Sci Rep, 8*(1), 6756. https://doi.org/10.1038/s41598-018-25233-x

Bibbal, D., Loukiadis, E., Kérourédan, M., Garam, C. P. d., Ferré, F., Cartier, P., Gay, E., Oswald, E., Auvray, F., & Brugère, H. (2014). Intimin Gene (eae) Subtype-Based Real-Time PCR Strategy for Specific Detection of Shiga Toxin-Producing Escherichia coli Serotypes O157:H7, O26:H11, O103:H2, O111:H8, and O145:H28 in Cattle Feces. *Applied and Environmental Microbiology, 80*(3), 1177-1184. https://doi.org/doi:10.1128/AEM.03161-13

Bielaszewska, M., Friedrich, A. W., Aldick, T., Schürk-Bulgrin, R., & Karch, H. (2006). Shiga toxin activatable by intestinal mucus in Escherichia coli isolated from humans: predictor for a severe clinical outcome. *Clin Infect Dis, 43*(9), 1160-1167. https://doi.org/10.1086/508195

Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S. Y., Medema, M. H., & Weber, T. (2019). antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research, 47*(W1), W81-W87. https://doi.org/10.1093/nar/gkz310

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. https://doi.org/10.1093/bioinformatics/btu170

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics, 10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Carattoli, A., & Hasman, H. (2020). PlasmidFinder and In Silico pMLST: Identification and Typing of Plasmid Replicons in Whole-Genome Sequencing (WGS). *Methods Mol Biol, 2075*, 285-294. https://doi.org/10.1007/978-1-4939-9877-7_20

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., & Jin, Q. (2005). VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res, 33*(Database issue), D325-328. https://doi.org/10.1093/nar/gki008

Chen, Z., Erickson, D. L., & Meng, J. (2020). 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

Cimermancic, P., Medema, M. H., Claesen, J., Kurita, K., Wieland Brown, L. C., Mavrommatis, K., Pati, A., Godfrey, P. A., Koehrsen, M., Clardy, J., Birren, B. W., Takano, E., Sali, A., Linington, R. G., & Fischbach, M. A. (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell, 158*(2), 412-421. https://doi.org/10.1016/j.cell.2014.06.034

Cinque, L., Ghomchi, Y., Chen, Y., Bensimon, A., & Baigl, D. (2010). Protection of Human Genomic DNA from Mechanical Stress by Reversible Folding Transition. *ChemBioChem, 11*(3), 340-343. https://doi.org/https://doi.org/10.1002/cbic.200900734

Clausen, P. T. L. C., Aarestrup, F. M., & Lund, O. (2018). Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics, 19*(1), 307. https://doi.org/10.1186/s12859-018-2336-6

Dallman, T. J., Byrne, L., Launders, N., Glen, K., Grant, K. A., & Jenkins, C. (2015). The utility and public health implications of PCR and whole genome sequencing for the detection and investigation of an outbreak of Shiga toxin-producing Escherichia coli serogroup O26:H11. *Epidemiol Infect, 143*(8), 1672-1680. https://doi.org/10.1017/s0950268814002696

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience, 10*(2). https://doi.org/10.1093/gigascience/giab008

Davis, S., Pettengill, J. B., Luo, Y., Payne, J., Shpuntoff, A., Rand, H., & Strain, E. (2015). CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. *PeerJ Computer Science, 1*, e20. https://doi.org/10.7717/peerj-cs.20

DebRoy, C., Fratamico, P. M., Yan, X., Baranzoni, G., Liu, Y., Needleman, D. S., Tebbs, R., O'Connell, C. D., Allred, A., Swimley, M., Mwangi, M., Kapur, V., Raygoza Garay, J. A., Roberts, E. L., & Katani, R. (2016). Comparison of O-Antigen Gene Clusters of All O-Serogroups of Escherichia coli and Proposal for Adopting a New Nomenclature for O-Typing. *PLoS One, 11*(1), e0147434. https://doi.org/10.1371/journal.pone.0147434

Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One, 8*(12), e85024. https://doi.org/10.1371/journal.pone.0085024

Delannoy, S., Beutin, L., & Fach, P. (2016). Improved traceability of Shiga-toxin-producing Escherichia coli using CRISPRs for detection and typing. *Environ Sci Pollut Res Int, 23*(9), 8163-8174. https://doi.org/10.1007/s11356-015-5446-y

Delannoy, S., Mariani-Kurkdjian, P., Webb, H. E., Bonacorsi, S., & Fach, P. (2017). The Mobilome; A Major Contributor to Escherichia coli stx2-Positive O26:H11

Strains Intra-Serotype Diversity. *Frontiers in Microbiology*, *8*, 1625-1625. https://doi.org/10.3389/fmicb.2017.01625

Diaz-Proano, C. (2019). *Prevalence, molecular characterization and inactivation of foodborne pathogens on native pecans* Oklahoma State University].

Doane, C. A., Pangloli, P., Richards, H. A., Mount, J. R., Golden, D. A., & Draughon, F. A. (2007). Occurrence of Escherichia coli O157:H7 in diverse farm environments. *J Food Prot, 70*(1), 6-10. https://doi.org/10.4315/0362-028x-70.1.6

Eltai, N. O., Yassine, H. M., Al Thani, A. A., Abu Madi, M. A., Ismail, A., Ibrahim, E., & Alali, W. Q. (2018). Prevalence of antibiotic resistant Escherichia coli isolates from fecal samples of food handlers in Qatar. *Antimicrobial Resistance & Infection Control, 7*(1), 78. https://doi.org/10.1186/s13756-018-0369-2

Fachada, N., Rodrigues, J., Lopes, V., & Martins, R. (2016). micompr: An R Package for Multivariate Independent Comparison of Observations. *The R Journal, 8*(2), 405-420.

Feng, P. C. H., & Reddy, S. (2013). Prevalences of Shiga toxin subtypes and selected other virulence factors among Shiga-toxigenic Escherichia coli strains isolated from fresh produce. *Applied and Environmental Microbiology, 79*(22), 6917-6923. https://doi.org/10.1128/AEM.02455-13

Feng, Y., Zhang, Y., Ying, C., Wang, D., & Du, C. (2015). Nanopore-based Fourth-generation DNA Sequencing Technology. *Genomics, Proteomics & Bioinformatics, 13*(1), 4-16. https://doi.org/https://doi.org/10.1016/j.gpb.2015.01.009

Forbes, J. D., Knox, N. C., Ronholm, J., Pagotto, F., & Reimer, A. (2017). Metagenomics: The Next Culture-Independent Game Changer [Review]. *Frontiers in Microbiology, 8*(1069). https://doi.org/10.3389/fmicb.2017.01069

Fratamico, P. M., DebRoy, C., Liu, Y., Needleman, D. S., Baranzoni, G. M., & Feng, P. (2016). Advances in Molecular Serotyping and Subtyping of Escherichia coli. *Frontiers in Microbiology, 7*, 644-644. https://doi.org/10.3389/fmicb.2016.00644

Frömmel, U., Lehmann, W., Rödiger, S., Böhm, A., Nitschke, J., Weinreich, J., Groß, J., Roggenbuck, D., Zinke, O., Ansorge, H., Vogel, S., Klemm, P., Wex, T., Schröder, C., Wieler, L. H., & Schierack, P. (2013). Adhesion of human and animal Escherichia coli strains in association with their virulence-associated genes and phylogenetic origins. *Applied and Environmental Microbiology, 79*(19), 5814-5829. https://doi.org/10.1128/AEM.01384-13

Gardner, S. N., & Hall, B. G. (2013). When Whole-Genome Alignments Just Won't Work: kSNP v2 Software for Alignment-Free SNP Discovery and Phylogenetics of Hundreds of Microbial Genomes. *PLoS One, 8*(12), e81760. https://doi.org/10.1371/journal.pone.0081760

Gardner, S. N., Slezak, T., & Hall, B. G. (2015). kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics, 31*(17), 2877-2878. https://doi.org/10.1093/bioinformatics/btv271

Gordon, D. M., & Cowling, A. (2003). The distribution and genetic structure of Escherichia coli in Australian vertebrates: host and geographic effects. *Microbiology (Reading), 149*(Pt 12), 3575-3586. https://doi.org/10.1099/mic.0.26486-0

Govindarajan, D. K., Viswalingam, N., Meganathan, Y., & Kandaswamy, K. (2020). Adherence patterns of Escherichia coli in the intestine and its role in pathogenesis. *Medicine in Microecology, 5*, 100025. https://doi.org/https://doi.org/10.1016/j.medmic.2020.100025

Greenfield, L. K., & Whitfield, C. (2012). Synthesis of lipopolysaccharide O-antigens by ABC transporter-dependent pathways. *Carbohydr Res, 356*, 12-24. https://doi.org/10.1016/j.carres.2012.02.027

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics, 29*(8), 1072-1075. https://doi.org/10.1093/bioinformatics/btt086

Gyles, C., & Boerlin, P. (2013). Horizontally Transferred Genetic Elements and Their Role in Pathogenesis of Bacterial Disease. *Veterinary Pathology, 51*(2), 328-340. https://doi.org/10.1177/0300985813511131

Hartland, E. L., Batchelor, M., Delahay, R. M., Hale, C., Matthews, S., Dougan, G., Knutton, S., Connerton, I., & Frankel, G. (1999). Binding of intimin from enteropathogenic Escherichia coli to Tir and to host cells. *Mol Microbiol, 32*(1), 151-158. https://doi.org/10.1046/j.1365-2958.1999.01338.x

Ho Sui, S. J., Fedynak, A., Hsiao, W. W. L., Langille, M. G. I., & Brinkman, F. S. L. (2009). The association of virulence factors with genomic islands. *PLoS One, 4*(12), e8094-e8094. https://doi.org/10.1371/journal.pone.0008094

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4378521/

Ito, K., Iida, M., Yamazaki, M., Moriya, K., Moroishi, S., Yatsuyanagi, J., Kurazono, T., Hiruta, N., & Ratchtrachenchai, O. A. (2007). Intimin types determined by heteroduplex mobility assay of intimin gene (eae)-positive Escherichia coli strains. *J Clin Microbiol*, *45*(3), 1038-1041. https://doi.org/10.1128/jcm.01103-06

Jackson, B. R., Tarr, C., Strain, E., Jackson, K. A., Conrad, A., Carleton, H., Katz, L. S., Stroika, S., Gould, L. H., Mody, R. K., Silk, B. J., Beal, J., Chen, Y., Timme, R., Doyle, M., Fields, A., Wise, M., Tillman, G., Defibaugh-Chavez, S., Kucerova, Z., Sabol, A., Roache, K., Trees, E., Simmons, M., Wasilenko, J., Kubota, K., Pouseele, H., Klimke, W., Besser, J., Brown, E., Allard, M., & Gerner-Smidt, P. (2016). Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation. *Clinical Infectious Diseases*, *63*(3), 380-386. https://doi.org/10.1093/cid/ciw242

Joensen, K. G., Tetzschner, A. M. M., Iguchi, A., Aarestrup, F. M., Scheutz, F., & Carroll, K. C. (2015). Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data. *Journal of Clinical Microbiology*, *53*(8), 2410-2426. https://doi.org/doi:10.1128/JCM.00008-15

Joensen, K. G., Tetzschner, A. M., Iguchi, A., Aarestrup, F. M., & Scheutz, F. (2015). Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data. *J Clin Microbiol*, *53*(8), 2410-2426. https://doi.org/10.1128/jcm.00008-15

Jombart, T., Kendall, M., Almagro-Garcia, J., & Colijn, C. (2017). treespace: Statistical exploration of landscapes of phylogenetic trees. *Mol Ecol Resour*, *17*(6), 1385-1392. https://doi.org/10.1111/1755-0998.12676

Kanamori, H., Parobek, C. M., Juliano, J. J., Duin, D. v., Cairns, B. A., Weber, D. J., & Rutala, W. A. (2017). A Prolonged Outbreak of KPC-3-Producing Enterobacter cloacae and Klebsiella pneumoniae Driven by Multiple Mechanisms of Resistance Transmission at a Large Academic Burn Center. *Antimicrobial Agents and Chemotherapy*, *61*(2), e01516-01516. https://doi.org/doi:10.1128/AAC.01516-16

Karlsson, E., Lärkeryd, A., Sjödin, A., Forsman, M., & Stenberg, P. (2015). Scaffolding of a bacterial genome using MinION nanopore sequencing. *Scientific Reports*, *5*, 11996-11996. https://doi.org/10.1038/srep11996

Katz, L. S., Griswold, T., Williams-Newkirk, A. J., Wagner, D., Petkau, A., Sieffert, C., Van Domselaar, G., Deng, X., & Carleton, H. A. (2017). A Comparative Analysis of the Lyve-SET Phylogenomics Pipeline for Genomic Epidemiology of

Foodborne Pathogens [Methods]. *Frontiers in Microbiology, 8*(375). https://doi.org/10.3389/fmicb.2017.00375

Kauffmann, F. (1947). The serology of the coli group. *J Immunol, 57*(1), 71-100.

Kawasaki, S., Horikoshi, N., Okada, Y., Takeshita, K., Sameshima, T., & Kawamoto, S. (2005). Multiplex PCR for simultaneous detection of Salmonella spp., Listeria monocytogenes, and Escherichia coli O157:H7 in meat samples. *J Food Prot, 68*(3), 551-556. https://doi.org/10.4315/0362-028x-68.3.551

Kendall, M., & Colijn, C. (2016). Mapping Phylogenetic Trees to Reveal Distinct Patterns of Evolution. *Molecular Biology and Evolution, 33*(10), 2735-2743. https://doi.org/10.1093/molbev/msw124

Kia, A., Gloeckner, C., Osothprarop, T., Gormley, N., Bomati, E., Stephenson, M., Goryshin, I., & He, M. M. (2017). Improved genome sequencing using an engineered transposase. *BMC Biotechnology, 17*(1), 6-6. https://doi.org/10.1186/s12896-016-0326-1

Kingry, L. C., Rowe, L. A., Respicio-Kingry, L. B., Beard, C. B., Schriefer, M. E., & Petersen, J. M. (2016). Whole genome multilocus sequence typing as an epidemiologic tool for Yersinia pestis. *Diagnostic Microbiology and Infectious Disease, 84*(4), 275-280. https://doi.org/10.1016/j.diagmicrobio.2015.12.003

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology, 37*(5), 540-546. https://doi.org/10.1038/s41587-019-0072-8

Kwak, Y. K., Colque, P., Byfors, S., Giske, C. G., Möllby, R., & Kühn, I. (2015). Surveillance of antimicrobial resistance among Escherichia coli in wastewater in Stockholm during 1 year: does it reflect the resistance trends in the society? *Int J Antimicrob Agents, 45*(1), 25-32. https://doi.org/10.1016/j.ijantimicag.2014.09.016

Lacher, D. W., Gangiredla, J., Jackson, S. A., Elkins, C. A., & Feng, P. C. (2014). Novel microarray design for molecular serotyping of shiga toxin- producing Escherichia coli strains isolated from fresh produce. *Appl Environ Microbiol, 80*(15), 4677-4682. https://doi.org/10.1128/aem.01049-14

Laurence, M., Hatzis, C., & Brash, D. E. (2014). Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS One, 9*(5), e97876-e97876. https://doi.org/10.1371/journal.pone.0097876

Lee, I., Ouk Kim, Y., Park, S. C., & Chun, J. (2016). OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *Int J Syst Evol Microbiol, 66*(2), 1100-1103. https://doi.org/10.1099/ijsem.0.000760

Leidenfrost, R. M., Pöther, D.-C., Jäckel, U., & Wünschiers, R. (2020). Benchmarking the MinION: Evaluating long reads for microbial profiling. *Scientific Reports, 10*(1), 5125. https://doi.org/10.1038/s41598-020-61989-x

Li, H., & Gänzle, M. (2016). Some Like It Hot: Heat Resistance of Escherichia coli in Food. *Frontiers in Microbiology, 7*, 1763-1763. https://doi.org/10.3389/fmicb.2016.01763

Li, H., Achour, I., Bastarache, L., Berghout, J., Gardeux, V., Li, J., Lee, Y., Pesce, L., Yang, X., Ramos, K. S., Foster, I., Denny, J. C., Moore, J. H., & Lussier, Y. A. (2016). Integrative genomics analyses unveil downstream biological effectors of disease-specific polymorphisms buried in intergenic regions. *npj Genomic Medicine, 1*(1), 16006. https://doi.org/10.1038/npjgenmed.2016.6

Ma, Y., Chen, J., Fong, K., Nadya, S., Allen, K., Laing, C., Ziebell, K., Topp, E., Carroll, L. M., Wiedmann, M., Delaquis, P., & Wang, S. (2021). Antibiotic Resistance in Shiga Toxigenic Escherichia coli Isolates from Surface Waters and Sediments in a Mixed Use Urban Agricultural Landscape. *Antibiotics (Basel, Switzerland), 10*(3), 237. https://doi.org/10.3390/antibiotics10030237

Marchant, M., & Moreno, M. A. (2013). Dynamics and diversity of Escherichia coli in animals and system management of the manure on a commercial farrow-to-finish pig farm. *Applied and Environmental Microbiology, 79*(3), 853-859. https://doi.org/10.1128/AEM.02866-12

Margos, G., Hepner, S., Mang, C., Marosevic, D., Reynolds, S. E., Krebs, S., Sing, A., Derdakova, M., Reiter, M. A., & Fingerle, V. (2017). Lost in plasmids: next generation sequencing and the complex genome of the tick-borne pathogen Borrelia burgdorferi. *BMC Genomics, 18*(1), 422. https://doi.org/10.1186/s12864-017-3804-5

Martin, M. J., Thottathil, S. E., & Newman, T. B. (2015). Antibiotics Overuse in Animal Agriculture: A Call to Action for Health Care Providers. *American journal of public health, 105*(12), 2409-2410. https://doi.org/10.2105/AJPH.2015.302870

Mbelle, N. M., Feldman, C., Osei Sekyere, J., Maningi, N. E., Modipane, L., & Essack, S. Y. (2019). The Resistome, Mobilome, Virulome and Phylogenomics of Multidrug-Resistant Escherichia coli Clinical Isolates from Pretoria, South Africa. *Scientific Reports, 9*(1), 16457. https://doi.org/10.1038/s41598-019-52859-2

Melton-Celsa, A. R. (2014). Shiga Toxin (Stx) Classification, Structure, and Function. *Microbiology spectrum, 2*(4), 10.1128/microbiolspec.EHEC-0024-2013-2013. https://doi.org/10.1128/microbiolspec.EHEC-0024-2013

Miro, E., Rossen, J. W. A., Chlebowicz, M. A., Harmsen, D., Brisse, S., Passet, V., Navarro, F., Friedrich, A. W., & García-Cobos, S. (2020). Core/Whole Genome Multilocus Sequence Typing and Core Genome SNP-Based Typing of OXA-48-Producing Klebsiella pneumoniae Clinical Isolates From Spain [Original Research]. *Frontiers in Microbiology, 10*(2961). https://doi.org/10.3389/fmicb.2019.02961

Muller, B. H., Mollon, P., Santiago-Allexant, E., Javerliat, F., & Kaneko, G. (2019). In-depth comparison of library pooling strategies for multiplexing bacterial species in NGS. *Diagn Microbiol Infect Dis, 95*(1), 28-33. https://doi.org/10.1016/j.diagmicrobio.2019.04.014

Naganandhini, S., Kennedy, Z. J., Uyttendaele, M., & Balachandar, D. (2015). Persistence of Pathogenic and Non-Pathogenic Escherichia coli Strains in Various Tropical Agricultural Soils of India. *PLoS One, 10*(6), e0130038. https://doi.org/10.1371/journal.pone.0130038

Oakeson, K. F., Wagner, J. M., Rohrwasser, A., & Atkinson-Dunn, R. (2018). Whole-Genome Sequencing and Bioinformatic Analysis of Isolates from Foodborne Illness Outbreaks of Campylobacter jejuni and Salmonella enterica. *Journal of Clinical Microbiology, 56*(11), e00161-00118. https://doi.org/10.1128/JCM.00161-18

Økelsrud, A., Lydersen, E., & Fjeld, E. (2016). Biomagnification of mercury and selenium in two lakes in southern Norway. *Science of The Total Environment, 566-567*, 596-607. https://doi.org/https://doi.org/10.1016/j.scitotenv.2016.05.109

Orlek, A., Stoesser, N., Anjum, M. F., Doumith, M., Ellington, M. J., Peto, T., Crook, D., Woodford, N., Walker, A. S., Phan, H., & Sheppard, A. E. (2017). Plasmid Classification in an Era of Whole-Genome Sequencing: Application in Studies of Antibiotic Resistance Epidemiology [Mini Review]. *Frontiers in Microbiology, 8*(182). https://doi.org/10.3389/fmicb.2017.00182

Padilha, V. A., Alkhnbashi, O. S., Shah, S. A., de Carvalho, A. C. P. L. F., & Backofen, R. (2020). CRISPRcasIdentifier: Machine learning for accurate identification and classification of CRISPR-Cas systems. *Gigascience, 9*(6). https://doi.org/10.1093/gigascience/giaa062

Panel, E. B., Koutsoumanis, K., Allende, A., Alvarez-Ordóñez, A., Bover-Cid, S., Chemaly, M., Davies, R., De Cesare, A., Herman, L., Hilbert, F., Lindqvist, R., Nauta, M., Peixe, L., Ru, G., Simmons, M., Skandamis, P., Suffredini, E., Jenkins, C., Monteiro Pires, S., Morabito, S., Niskanen, T., Scheutz, F., da Silva Felício, M. T., Messens, W., & Bolton, D. (2020). Pathogenicity assessment of Shiga toxin-producing Escherichia coli (STEC) and the public health risk posed by contamination of food with STEC. *EFSA Journal, 18*(1), e05967. https://doi.org/https://doi.org/10.2903/j.efsa.2020.5967

Paradis, E. (2016). The distribution of branch lengths in phylogenetic trees. *Molecular Phylogenetics and Evolution, 94*, 136-145. https://doi.org/https://doi.org/10.1016/j.ympev.2015.08.010

Pearman, W. S., Freed, N. E., & Silander, O. K. (2020). Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads. *BMC Bioinformatics, 21*(1), 220-220. https://doi.org/10.1186/s12859-020-3528-4

Perrière, G., & Gouy, M. (1996). WWW-query: An on-line retrieval system for biological sequence banks. *Biochimie, 78*(5), 364-369. https://doi.org/https://doi.org/10.1016/0300-9084(96)84768-7

Persson, S., Olsen, K. E. P., Ethelberg, S., & Scheutz, F. (2007). Subtyping Method for Escherichia coli Shiga Toxin (Verocytotoxin) 2 Variants and Correlations to Clinical Manifestations. *Journal of Clinical Microbiology, 45*(6), 2020-2024. https://doi.org/doi:10.1128/JCM.02591-06

Proença, J. T., Barral, D. C., & Gordo, I. (2017). Commensal-to-pathogen transition: One-single transposon insertion results in two pathoadaptive traits in Escherichia coli -macrophage interaction. *Scientific Reports, 7*(1), 4504. https://doi.org/10.1038/s41598-017-04081-1

Rang, F. J., Kloosterman, W. P., & de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome biology, 19*(1), 90. https://doi.org/10.1186/s13059-018-1462-9

Rankin, D. J., Rocha, E. P. C., & Brown, S. P. (2011). What traits are carried on mobile genetic elements, and why? *Heredity, 106*(1), 1-10. https://doi.org/10.1038/hdy.2010.24

Reis-Cunha, J. L., Bartholomeu, D. C., Manson, A. L., Earl, A. M., & Cerqueira, G. C. (2019). ProphET, prophage estimation tool: A stand-alone prophage sequence

prediction tool with self-updating reference database. *PLoS One, 14*(10), e0223364. https://doi.org/10.1371/journal.pone.0223364

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution, 3*(2), 217-223. https://doi.org/https://doi.org/10.1111/j.2041-210X.2011.00169.x

Rhee, S. Y. (2005). Bioinformatics. Current Limitations and Insights for the Future. *Plant Physiology, 138*(2), 569-570. https://doi.org/10.1104/pp.104.900153

Ritchie, J. M., & Waldor, M. K. (2005). The locus of enterocyte effacement-encoded effector proteins all promote enterohemorrhagic Escherichia coli pathogenicity in infant rabbits. *Infection and Immunity, 73*(3), 1466-1474. https://doi.org/10.1128/IAI.73.3.1466-1474.2005

Röder, B., Frühwirth, K., Vogl, C., Wagner, M., & Rossmanith, P. (2010). Impact of Long-Term Storage on Stability of Standard DNA for Nucleic Acid-Based Methods. *Journal of Clinical Microbiology, 48*(11), 4260-4262. https://doi.org/doi:10.1128/JCM.01230-10

Ross, K. S., Haites, N. E., & Kelly, K. F. (1990). Repeated freezing and thawing of peripheral blood and DNA in suspension: effects on DNA yield and integrity. *Journal of Medical Genetics, 27*(9), 569-570. https://doi.org/10.1136/jmg.27.9.569

Russo, T. A., & Johnson, J. R. (2009). Chapter 48 - Extraintestinal Pathogenic Escherichia coli. In A. D. T. Barrett & L. R. Stanberry (Eds.), *Vaccines for Biodefense and Emerging and Neglected Diseases* (pp. 939-961). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-369408-9.00048-2

Salipante, S. J., SenGupta, D. J., Cummings, L. A., Land, T. A., Hoogestraat, D. R., Cookson, B. T., & Tang, Y.-W. (2015). Application of Whole-Genome Sequencing for Bacterial Strain Typing in Molecular Epidemiology. *Journal of Clinical Microbiology, 53*(4), 1072-1079. https://doi.org/doi:10.1128/JCM.03385-14

Scallan, E., Griffin, P. M., Angulo, F. J., Tauxe, R. V., & Hoekstra, R. M. (2011). Foodborne illness acquired in the United States--unspecified agents. *Emerging infectious diseases, 17*(1), 16-22. https://doi.org/10.3201/eid1701.091101p2

Scheutz, F., Teel, L. D., Beutin, L., Piérard, D., Buvens, G., Karch, H., Mellmann, A., Caprioli, A., Tozzoli, R., Morabito, S., Strockbine, N. A., Melton-Celsa, A. R., Sanchez, M., Persson, S., & O'Brien, A. D. (2012). Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and standardizing Stx

nomenclature. *J Clin Microbiol, 50*(9), 2951-2963. https://doi.org/10.1128/jcm.00860-12

Schürch, A. C., Arredondo-Alonso, S., Willems, R. J. L., & Goering, R. V. (2018). Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. *Clinical Microbiology and Infection, 24*(4), 350-354. https://doi.org/https://doi.org/10.1016/j.cmi.2017.12.016

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics, 30*(14), 2068-2069. https://doi.org/10.1093/bioinformatics/btu153

Sekse, C., Holst-Jensen, A., Dobrindt, U., Johannessen, G. S., Li, W., Spilsberg, B., & Shi, J. (2017). High Throughput Sequencing for Detection of Foodborne Pathogens [Review]. *Frontiers in Microbiology, 8*(2029). https://doi.org/10.3389/fmicb.2017.02029

Sekse, C., Holst-Jensen, A., Dobrindt, U., Johannessen, G. S., Li, W., Spilsberg, B., & Shi, J. (2017). High Throughput Sequencing for Detection of Foodborne Pathogens [Review]. *Frontiers in Microbiology, 8*(2029). https://doi.org/10.3389/fmicb.2017.02029

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. In M. Kollmar (Ed.), *Gene Prediction: Methods and Protocols* (pp. 227-245). Springer New York. https://doi.org/10.1007/978-1-4939-9173-0_14

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol, 1962*, 227-245. https://doi.org/10.1007/978-1-4939-9173-0_14

Shariat, N., & Dudley, E. G. (2014). CRISPRs: molecular signatures used for pathogen subtyping. *Appl Environ Microbiol, 80*(2), 430-439. https://doi.org/10.1128/aem.02790-13

Sharma, C., Rokana, N., Chandra, M., Singh, B. P., Gulhane, R. D., Gill, J. P. S., Ray, P., Puniya, A. K., & Panwar, H. (2018). Antimicrobial Resistance: Its Surveillance, Impact, and Alternative Management Strategies in Dairy Animals [Review]. *Frontiers in Veterinary Science, 4*(237). https://doi.org/10.3389/fvets.2017.00237

Sharma, V. K., Akavaram, S., Schaut, R. G., & Bayles, D. O. (2019). Comparative genomics reveals structural and functional features specific to the genome of a foodborne Escherichia coli O157:H7. *BMC Genomics, 20*(1), 196-196. https://doi.org/10.1186/s12864-019-5568-6

Shea, K. M. (2003). Antibiotic Resistance: What Is the Impact of Agricultural Uses of Antibiotics on Children's Health? *Pediatrics, 112*(Supplement 1), 253-258. https://pediatrics.aappublications.org/content/pediatrics/112/Supplement_1/253.full.pdf

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet, 15*(2), 121-132. https://doi.org/10.1038/nrg3642

Sokurenko, E. V., Chesnokova, V., Dykhuizen, D. E., Ofek, I., Wu, X. R., Krogfelt, K. A., Struve, C., Schembri, M. A., & Hasty, D. L. (1998). Pathogenic adaptation of Escherichia coli by natural variation of the FimH adhesin. *Proceedings of the National Academy of Sciences of the United States of America, 95*(15), 8922-8926. https://doi.org/10.1073/pnas.95.15.8922

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30*(9), 1312-1313. https://doi.org/10.1093/bioinformatics/btu033

Stoler, N., & Nekrutenko, A. (2021). Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics, 3*(1). https://doi.org/10.1093/nargab/lqab019

Struelens, M. J., Palm, D., & Takkinen, J. (2011). Enteroaggregative, Shiga toxin-producing Escherichia coli O104:H4 outbreak: new microbiological findings boost coordinated investigations by European public health laboratories. *Euro Surveill, 16*(24). https://doi.org/10.2807/ese.16.24.19890-en

Taylor, T. L., Volkening, J. D., DeJesus, E., Simmons, M., Dimitrov, K. M., Tillman, G. E., Suarez, D. L., & Afonso, C. L. (2019). Rapid, multiplexed, whole genome and plasmid sequencing of foodborne pathogens using long-read nanopore technology. *Scientific Reports, 9*(1), 16350. https://doi.org/10.1038/s41598-019-52424-x

Thrash, A., Hoffmann, F., & Perkins, A. (2020). Toward a more holistic method of genome assembly assessment. *BMC Bioinformatics, 21*(4), 249. https://doi.org/10.1186/s12859-020-3382-4

Timme, R. E., Lafon, P. C., Balkey, M., Adams, J. K., Wagner, D., Carleton, H., Strain, E., Hoffmann, M., Sabol, A., Rand, H., Lindsey, R., Sheehan, D., Baugher, J. D., & Trees, E. (2020). Gen-FS coordinated proficiency test data for genomic foodborne pathogen surveillance, 2017 and 2018 exercises. *Scientific Data, 7*(1), 402. https://doi.org/10.1038/s41597-020-00740-7

Timme, R. E., Sanchez Leon, M., & Allard, M. W. (2019). Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback. *Methods Mol Biol, 1918*, 201-212. https://doi.org/10.1007/978-1-4939-9000-9_17

Timme, R. E., Sanchez Leon, M., & Allard, M. W. (2019). Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback. *Methods Mol Biol, 1918*, 201-212. https://doi.org/10.1007/978-1-4939-9000-9_17

Tolar, B., Joseph, L. A., Schroeder, M. N., Stroika, S., Ribot, E. M., Hise, K. B., & Gerner-Smidt, P. (2019). An Overview of PulseNet USA Databases. *Foodborne pathogens and disease, 16*(7), 457-462. https://doi.org/10.1089/fpd.2019.2637

Tominaga, A. (2004). Characterization of six flagellin genes in the H3, H53 and H54 standard strains of Escherichia coli. *Genes Genet Syst, 79*(1), 1-8. https://doi.org/10.1266/ggs.79.1

Tyler, A. D., Mataseje, L., Urfano, C. J., Schmidt, L., Antonation, K. S., Mulvey, M. R., & Corbett, C. R. (2018). Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. *Scientific Reports, 8*(1), 10931. https://doi.org/10.1038/s41598-018-29334-5

Ventola, C. L. (2015). The antibiotic resistance crisis: part 1: causes and threats. *P & T : a peer-reviewed journal for formulary management, 40*(4), 277-283. https://pubmed.ncbi.nlm.nih.gov/25859123

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS One, 9*(11), e112963. https://doi.org/10.1371/journal.pone.0112963

Walsh, A. M., Crispie, F., O'Sullivan, O., Finnegan, L., Claesson, M. J., & Cotter, P. D. (2018). Species classifier choice is a key consideration when analysing low-complexity food microbiome data. *Microbiome, 6*(1), 50. https://doi.org/10.1186/s40168-018-0437-0

Wang, L., Rothemund, D., Curd, H., & Reeves, P. R. (2003). Species-wide variation in the Escherichia coli flagellin (H-antigen) gene. *J Bacteriol, 185*(9), 2936-2943. https://doi.org/10.1128/jb.185.9.2936-2943.2003

Wentz, T. G., Hu, L., Hammack, T. S., Brown, E. W., Sharma, S. K., & Allard, M. W. (2019). Next Generation Sequencing for the Detection of Foodborne Microbial Pathogens. In S. K. Singh & J. H. Kuhn (Eds.), *Defense Against Biological*

Attacks: Volume II (pp. 311-337). Springer International Publishing. https://doi.org/10.1007/978-3-030-03071-1_14

Wick, R. R., Judd, L. M., & Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome biology*, *20*(1), 129. https://doi.org/10.1186/s13059-019-1727-y

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017a). Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics, 3*(10). https://doi.org/https://doi.org/10.1099/mgen.0.000132

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017b). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology, 13*(6), e1005595. https://doi.org/10.1371/journal.pcbi.1005595

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology, 15*(3), R46-R46. https://doi.org/10.1186/gb-2014-15-3-r46

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome biology, 20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Woods, L. C., Gorrell, R. J., Taylor, F., Connallon, T., Kwok, T., & McDonald, M. J. (2020). Horizontal gene transfer potentiates adaptation by reducing selective constraints on the spread of genetic variation. *Proceedings of the National Academy of Sciences, 117*(43), 26868-26875. https://doi.org/10.1073/pnas.2005331117

Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell, 178*(4), 779-794. https://doi.org/https://doi.org/10.1016/j.cell.2019.07.010

Yoon, S. H., Ha, S. M., Lim, J., Kwon, S., & Chun, J. (2017). A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie Van Leeuwenhoek, 110*(10), 1281-1286. https://doi.org/10.1007/s10482-017-0844-4

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M., & Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *The Journal of antimicrobial chemotherapy, 67*(11), 2640-2644. https://doi.org/10.1093/jac/dks261

Zitko, V. (1994). Principal component analysis in the evaluation of environmental data. *Marine Pollution Bulletin, 28*(12), 718-722. https://doi.org/https://doi.org/10.1016/0025-326X(94)90329-8

CHAPTER IV

RAPID IDENTIFICATION AND MOLECULAR CHARACTERIZATION OF SALMONELLA ISOLATES FROM PECAN ORCHARDS THROUGH NANOPORE SEQUENCING

ABSTRACT

Whole genome sequencing is becoming the tool for various applications, thereby this study was conducted to evaluate the performance of Nanopore sequencing for rapid identification and molecular characterization of of *Salmonella*. Ten isolates obtained from pecan orchards were sequenced using MinION and Illumina NextSeq 500. As MinION allows real-time reads analysis, the reads were time-based subsampled to determine the earliest identification turnaround time for each isolate. Species level identification was achieved at 15 mins of sequencing run. In 25 hours, complete sets of *Salmonella* pathogenicity islands were detected from assemblies obtained from the subsampled reads. Additionally, comparisons of the Nanopore-based assemblies against hybrid assemblies from the combined reads of MinION and Illumina showed that the best values of continuity and annotated features were obtained in just 8 and 25 hours of sequencing run, respectively ($p < 0.05$). Whereas, using a stringent BLASTn search (percentage of identity of 95 % and query coverage of 85 %) against

the Comprehensive Antibiotic Resistance Database (CARD), we could find significantly similar results to those obtained from the hybrid assemblies, oppositely, the hits obtained from the search against the Virulence Factor Database (VFDB) were not sufficient to generate results significantly similar, nevertheless, it was possible to obtain an average of 96.37% from the hits acquired from the hybrid assemblies using VFDB, where no significant changes were observed after 16 hours compared to the full datasets ($p < 0.05$). Finally, phylogeny analysis obtained from assemblies created with reads produced in 3 hours of sequencing process, were significantly similar to those of the results with hybrid genomes ($p < 0.05$). These results demonstrated that Nanopore can offer an effective sequencing platform for the rapid identification of Nontyphoidal *Salmonella* isolates, with certain capabilities for their characterization.

Key words: Whole genome sequencing, Nanopore, MinION, Illumina, Nontyphoidal *Salmonella*

INTRODUCTION

Nontyphoidal serotypes of *Salmonella enterica* (henceforth referred to as *Salmonella*) are responsible for causing salmonellosis, a disease contracted by consuming improperly cooked animal derived food, contaminated fresh produces and water, or direct exposure to reptiles, amphibians, and other infected animals, including humans (Acheson & Hohmann, 2001; Kurtz et al., 2017). This reportable disease causes an estimated 1.2 million infections and 450 deaths annually in the U.S., generating an average annual expenditure of $ 0.5 to $ 2.3 billion (Frenzen et al., 1999; Scallan et al., 2011; Sher et al., 2021). Although the implementation of guidelines for the handling, transport, and storage of food has been efficient in controlling the incidence of

salmonellosis attributed to the consumption of eggs (i.e., the main source of *Salmonella* Enteritidis - SE) (Wright et al., 2016) as well as the widespread practice of vaccinating chickens against *Salmonella* Typhimurium (Desin et al., 2013; Dórea et al., 2010), a positive trend in the increase in cases of SE attributed to alternative sources (Chai et al., 2012; Sher et al., 2021), or an increase in cases of other less recurrent strains such as Newport or Javiana (Boore et al., 2015; Sher et al., 2021) have made Nontyphoidal *Salmonella* a recurrent foodborne pathogen. Hence, it is essential to maintain a constant characterization and comparison of the strains that inhabit different niches with incident strains in outbreaks, in order to identify if this indicates an increased presence of the host, increase in human exposure to the source of contamination, or increased contamination from anthropogenic activities (Chai et al., 2012; Gast et al., 2004; Sher et al., 2021).

Currently, *Salmonella* outbreak surveillance is carried out with Whole Genome Sequencing (WGS) in the U.S. (Tolar et al., 2019), since despite having a longer turnaround time than its predecessor (i.e., PFGE) (Rounds et al., 2020), it can offer a much more complete level of resolution, providing insights into the strain's serotype, virulence, pathogenicity, resistance to antibiotics, or even elucidating the subtype to which it belongs (Brown et al., 2019; Delgado-Suárez et al., 2018; Ibrahim & Morin, 2018). Several studies have shown the level of sensitivity and specificity that WGS could provide for the study of *Salmonella* (Chen, Kuang, et al., 2020; Pornsukarom et al., 2018)*;* for instance, Banerji et al. (2020) obtained a level of concordance of >99% when combining results from serotype prediction and multi-locus sequence typing (MLST) for genome based *Salmonella* serotyping, as well as, in the study carried out by Cooper et al. (2020) which apart from obtaining a high level of concordance in the prediction of

serotypes using WGS data, the accuracy of their antimicrobial resistance profile prediction ranged from 94.7% to 99%. PulseNet and GenomeTrakr, i.e. main foodborne pathogen surveillance networks in the U.S., carry out the collection of sequencing data through MiSeq, an Illumina-based platform (Timme et al., 2019), yet Nanopore technology is posed as a new contender that is making its way since, unlike Illumina, it yields long reads, which do not require to be generated through expensive equipment, as well as, highly trained personnel to use the sequencing devices it offers, and additionally, it provides the possibility of analyzing sequencing reads in real time (Chen, Kuang, et al., 2020; Leidenfrost et al., 2020; Tyler et al., 2018). Nonetheless, these benefits have a tradeoff related to the lower precision that Nanopore offers, since its error rate ranges between 5 to 15% (Rang et al., 2018) while Illumina only ~1% (Stoler & Nekrutenko, 2021), making it necessary to identify the scope that this platform offers for surveillance of foodborne pathogens. Therefore, in this study, we conducted the performance evaluation of MinION, a device based on Nanopore sequencing technology, for the rapid identification and molecular characterization of *Salmonella* isolated from pecan orchards.

MATERIALS AND METHODS

Bacterial isolates

The *Salmonella* isolates used in this study were isolated from environmental samples collected from pecan orchards (soil and animal feces) located in Oklahoma as part of the doctoral dissertation conducted by Diaz-Proano (2019). Bacterial isolates were grown following the methodology used by the FDA Bacteriological Analytical Manual (FDA-BAM) (Andrews et al., 2018) with minor modifications. A general procedure was

used to enrich the samples before individual isolation. In brief, 10 g of each soil and fecal sample were added to 90 mL of Universal pre-enrichment broth (UPB) (Becton-Dickinson, Sparks, Maryland) and stomached in a filter bag (Whirlpak) using a Seward Stomacher® 400 (Seward, London, United Kingdom) circulator for 1 min at 230 RPM. Suspensions were incubated for 24 h at 42°C, 0.1 mL and 1 mL of the pre-enriched samples were selectively enriched in 9.9 mL of Rappaport-Vassiliadis (RV) (Benton-Dickinson, Sparks, Maryland) and 9 mL of Tetrathionate (TT) (Benton-Dickinson, Sparks, Maryland) broth, respectively. Selective broths were incubated at 37°C for 24 h followed by streaking onto Xylose lysine desoxycholate (XLD) (Benton-Dickinson, Sparks, Maryland) agar and incubation at 37°C for 24 h.

The presence of the specific *Salmonella* gene *invA* was used to detect and confirm the presence of *Salmonella* by PCR of DNA extracted at two stages using the boiling method described by Kawasaki et al. (2005). In the first stage, DNA was obtained from 1 ml of 24-hour enrichment broth, and secondly, DNA was extracted from five typical colonies from XLD plates. Primer pairs were chosen according to the target gene described in the literature and the 16S rRNA gene was used as an internal control (Diaz-Proano, 2019). Purified and confirmed *Salmonella* was transferred to Tryptic Soy Agar (TSA) (Becton-Dickinson, Sparks, Maryland) and stored at 4°C.

DNA extraction and whole genome sequencing

As part of Diaz-Proano (2019) doctoral dissertation, isolates were cultured in 5 mL tryptic soy broth (TSB, Difco, Sparks, MD) at 37°C for 18-20 h. Following overnight incubation, cells were harvested by centrifugation at 12000 rpm for 3 min and re-

suspended in 1X buffered peptone water (BPW). DNA extraction was performed using the DNeasy 96 blood and tissue kit (Qiagen, Valencia, CA) according to the manufacturer's recommendation for Gram-negative bacteria and high-throughput applications. DNA samples were then cleaned using a DNA clean up and concentrator kit (Zymo Research). The quality of the DNA was determined using NanoDrop 1000 - OD 260/280 and OD 260/230 - (Thermo Scientific, Rockford, IL), and the concentration was determined using the Qubit 3 fluorometer with a double-stranded DNA BR assay kit (Life Technologies, Grand Island, NY) according to each manufacturer's instructions. WGS was performed using Illumina and Oxford Nanopore Technologies (ONT) platforms. For Illumina sequencing, libraries were prepared using the Nextera XT DNA sample preparation kit with the NextSeq®500 high output kit (2*150 bp paired-end reads) (Illumina, Inc., San Diego, CA). Whereas, ONT sequencing libraries were prepared using the Rapid Barcoding Sequencing kit (SQK-RBK004) and run on the MinION sequencing system (ONT, Oxford, UK) following the standard 48 h 1D sequencing protocol in the MinKNOW software (ONT, Oxford, UK) using one FLO-MIN106 R.9.4.1 flowcell (Appendix 3).

Analysis of whole genome sequencing data

Initial data processing

For Illumina reads, Trimmomatic (version 0.32) (Bolger et al., 2014) was used to remove barcodes and to trim the sequences with a window size of 4 and a Q score cutoff of 20 (Del Fabbro et al., 2013), and FastQC (version 0.11.9) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc) was used for quality

control. MinION reads were base called with Guppy (version 3.0.3) (Oxford Nanopore Technologies) and first filtered by EPI2ME (version 2.48) (Oxford Nanopore Technologies) with a Q score cutoff of 7. Later, "Passed" reads were demultiplexed with Porechop (version 0.2.4) (https://github.com/rrwick/Porechop) for further additional filtering with Filtlong (version 0.2.0) (https://github.com/rrwick/Filtlong) using a Q score cutoff of 9 (Tyler et al., 2018; Wick et al., 2017a). To analyze the MinION sequencing data over time, filtered reads were subsampled using a custom Perl script at intervals from the start of the sequencing: at 15, 30, 60, 120, 240, 480, 960, and 1500 min (Taylor et al., 2019).

Species identification, typing, and pathogenicity island detection

All bioinformatics analyzes were carried out on a Dual Intel Xeon Gold 6130 High Performance Cluster (HPC) with 2.8 GB max memory per core. The presence of specific reads for *Salmonella* complex species was verified with Kraken 2 (version 2.0.7) (Wood & Salzberg, 2014) in default mode (kmer size of 35 and minimizer length of 31) with a confidence threshold of 0.05 (Ye et al., 2019) against a custom RefSeq database (i.e., containing Archaeal, bacterial, viral and plasmids sequences) created on January 2020. For the identification of serotypes, SeqSero2 (version 1.2.1) (Zhang et al., 2019) was used with the "k" workflow. Additionally, sequencing reads were analyzed with SPIfinder (Roer et al., 2016) in order to identify *Salmonella* pathogenicity islands with a minimum percentage of identity of 95%, and a minimum gene coverage of 70%.

Genome assembly

MinION reads were assembled with Flye (version 2.6) (Kolmogorov et al., 2019) in nano-raw mode, with an expected genome size of 5.5Mb, and asm coverage (reduced

coverage for initial disjointig assembly) of 50. Furthermore, in order to get the most out of the data obtained from this technology, a polishing step was carried out using Rebaler (version 0.2.0) (https://github.com/rrwick/Rebaler), a pipeline that uses minimap2 to align long reads to an already assembled genome and Racon (version 1.0.) (https://github.com/isovic/racon) for creating consensus sequences. Finally, an extra polishing step with Medaka (version 1.4.1) (https://github.com/nanoporetech/medaka) a tool that can create consensus sequences from nanopore sequencing data using neural networks applied to a "read pileup" against a draft sequence using a variety of trained models, the "r941_min_fast_g303" model was used for this step. All the final assemblies generated were subjected to a preliminary quality assessment through QUAST (version 5.0.2) (Gurevich et al., 2013) and BUSCO (version 4.1.4) (Seppey et al., 2019) against the "enterobacterales_odb10" database.

For further comparisons, due to the lack of reference genomes available for these particular isolates, hybrid assemblies were generated using the reads from both Illumina and MinION for each isolate (Appendix 4). Two approaches were used and compared to select the best assembly for each isolate in terms of continuity (i.e., N50, L50, and the total number of contigs), and their BUSCO score. In the first approach, Illumina paired-end reads were mapped against the genomes generated directly from the aforementioned process with the complete dataset of MinION reads, thereafter samtools (version 1.10) (Danecek et al., 2021) was used to sort and index the aligned reads, and finally, an extra polishing step was included using Pilon (version 1.23) (Walker et al., 2014) with the setting "–changes" for which a custom bash script performed multiple runs until no changes were done in the polished genomes. Whereas, in the second approach,

assemblies obtained first from Illumina reads and later bridged with long reads were obtained using the Unicycler pipeline (version 0.4.8) (Wick et al., 2017b) in default mode, with the aforementioned Bowtie2, samtools, and Pilon versions, as well as, Blast+ (version 2.10.1) (Camacho et al., 2009). MinION and Illumina reads were mapped against the best hybrid assembly obtained for each isolate using minimap2 and Bowtie2, respectively, then mapped reads indexed via samtools were used to calculate the average coverage for both reads dataset using mosdepth (version 0.3.1) (https://github.com/brentp/mosdepth) with the parameters "-n --fast-mode --by 500". Additionally, the average nucleotide identity (ANI) from all generated assemblies from MinION reads against the best hybrid assemblies was calculated using OrthoANI (version 1.2) (Lee et al., 2016; Yoon et al., 2017) for each isolate.

Features annotation

Annotation of all the assembled genomes was performed with Prokka pipeline (version 1.14.6) (Seemann, 2014) in default mode. To identify VFs and antibiotic resistance (AR) genes, assembled genomes were aligned against the Virulence Factors Database (VFDB) core dataset (VFDB_setA retrieved on January 9, 2020) (Chen et al., 2005), ResFinder database (retrieved on December 16, 2020) (Zankari et al., 2012), and Comprehensive Antibiotic Resistance Database (CARD) protein homologs dataset for acquired resistance genes (retrieved on March 10, 2020) (Alcock et al., 2020) using Blastn with an E-value cutoff of 1e-6. The obtained hits were filtered using a custom Python script with a minimum percentage of identity of 95%, and minimum gene coverage of 85%, as well as overlapping hits, were evaluated and removed based on an in-house strategy. From this search, two sets of genes of AR and VFs detected for each

assembly were generated, thereafter the two sets obtained from each genome created from the MinION subsampled data were compared based on similarity with the two sets obtained from the best hybrid assemblies for each species respectively using cd-hit-est-2d with the parameters "-c 0.9 -n 8 -r 0 -G 1 -g 1 -b 20 -l 10 -s 0.0 -aL 0.0 -aS 0.0 -s2 1.0 -S2 0 -T 4 -M 32000 ". The pairs of similar genes generated by this comparison were counted and normalized based on the total number of genes present in the set of AR and VFs genes of each best hybrid assembly, respectively. A one-sample Wilcoxon signed rank test with Benjamini-Hochberg correction for multiple tests was used to check if the normalized values were close to 1, where $p > 0.05$ indicates that the set of genes analyzed is significantly similar to the set of genes of the best hybrid assembly for a given isolate.

For the detection of mobile genetic elements, replicons were identified using PlasmidFinder (version 2.0.1) (Carattoli & Hasman, 2020) with a minimum percentage of identity of 85%, and a minimum gene coverage of 70%. Whereas, prophages prediction was performed using ProphET (Reis-Cunha et al., 2019) using the GFF files obtained from the annotation process as an input. Clustered regularly interspaced short palindromic repeats (CRISPRs) were predicted with CRISPRCASidentifier (version 1.1.0) (Padilha et al., 2020) in default mode. Additionally,

Finally, biosynthetic gene clusters (BGCs) were predicted using Antismash (version 5.1.2) (Blin et al., 2019) with default parameters; in which interleaved, chemical hybrid, and neighboring clusters were divided into individual clusters for their quantification.

Phylogenetic analysis

A matrix of core single nucleotide polymorphisms (SNPs) was generated among the 10 isolates using kSNP (version 3.1) (Gardner et al., 2015) with the assemblies generated from the subsampled reads as well as the best hybrid assemblies, in which first Kchooser was run twice for each dataset to find the best kmer size. The core SNPs were used as an input for the construction of the maximum likelihood phylogenies using RAxML (version 8.2.11) (Stamatakis, 2014) with the GTRCAT model, a Lewis ascertainment bias correction, and 1000 bootstrap replicates. The resulting phylogenetic trees were rooted in the midpoint using NJplot (version 2.3) (Perrière & Gouy, 1996) and formatted using Figtree (version 1.4.4) (https://github.com/rambaut/figtree).

Statistical analyses

Genomic continuity comparison

A matrix containing the estimated length of the genome, the total number of contigs, the length of the largest contig, N50 and L50 for all the assemblies generated over time was normalized using the Min-Max scale method, which was applied to grouped values with respect to the identity of the isolate from which they were obtained using the following equation:

$$X_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where X_{norm} is the normalized value, x_{max} and x_{min} are the maximum and minimum values, respectively, of a particular metric evaluated in the assemblies from the same isolate, while x is a value to be normalized that is in this same set of values as x_{max} and x_{mins}

The normalized dataset was dimensionally reduced by principal component analysis (PCA) through the prcomp function in R (version 4.0.5), thereafter Kruskal Wallis and Dunn's Multiple Comparison Test with Benjamini-Hochberg correction was used to comparing the values of PC1 obtained from each time interval, additionally, the p values were corrected with respect to the weight of the variance covered by the PC1 (Fachada et al., 2016).

Genomic features comparison

A matrix containing the GC-content, the BUSCO score, the ANI, as well as the number of annotated genes, tRNA, tmRNA, rRNA, repeated regions, CRISPRs, and prophages for all the assemblies generated over time and from the best hybrid assemblies was normalized using the Min-Max scale method with the equation (1), which was applied to grouped values with respect to the identity of the isolate from which they were obtained. The normalized dataset was dimensional reduced by PCA and the aforementioned multivariate analysis was performed.

Statistical comparison of phylogenetic trees

The Kendall-Colijin test (Kendall & Colijn, 2016) described by Katz et al. (2017) was used to compare the topologies of the phylogenies generated from MinION subsampled reads with the phylogeny created from the hybrid assemblies using the R libraries treespace (Jombart et al., 2017) and phytools (Revell, 2012). A lambda value of 0 and 100,000 random trees as a background distribution were employed for all pairwise tree comparisons. For which a Z-test was calculated and a $p < 0.05$ indicates that the pair of compared trees are significantly similar (Katz et al., 2017).

RESULTS AND DISCUSSION

Evaluation of sequencing reads yield

For Illumina, the average median length of the paired-end reads of all isolates was 137.4. The total bases ranged from 381.53 Mbps to 565.46 Mbps, obtaining mean depth values between 75.68x to 115.65x after filtering (Table 17). For Nanopore, the average median length of the reads of all isolates was 4,034.75. The total bases ranged from 50.91 Mbps to 1002.51 Mbps, obtaining mean depth values between 10.72x to 209.96x after filtering (Table 18). While the isolate G5_25BLS1_1 had the highest reads yield on Illumina, the opposite occurred in Nanopore, where it obtained the lowest yield; while the lowest yield of Illumina was from the isolate G2M0S1_2 and the highest yield of Nanopore was G4BLF1_2 (Table 17, 18). The total number of subsampled reads that were generated at different stages of the sequencing run can be seen in table 18.The differences among the mean depth values in Nanopore sequenced isolates were even represented in the subsampled reads based on time (Table 19; Figure 11B). Additionally, through a paired Wilcoxon Signer-Rank test between the mean depth values from Illumina and Nanopore, we obtained that values from Nanopore were significantly lower than those generated by Illumina ($\alpha = 0.05$) (Figure 11A),

Table 17. Summary for Illumina sequencing of *Salmonella* isolates (Paired-end reads).

Isolate	Number of raw reads	Total Bases (Mbps)	Mean length	Median length	Total reads after filtering	Mean depth
G1BLS3_3	2920038	397.00	138.68	151	2887218	78.03
G2BLF1_3	3496024	413.30	121.90	137	3457512	85.67
G2M0S1_2	3288192	381.53	119.33	130	3255562	75.68
G2M4F3_3	3421988	413.34	123.87	140	3384982	85.46
G4BLF1_2	3946342	439.27	115.37	122	3907290	89.12
G4BLS2_11	3671052	409.85	115.38	122	3634522	81.37

G5_20BLS1_1	4038074	476.76	122.40	136	3996366	98.63
G5_25BLS1_1	4815770	565.46	121.70	134	4764928	115.65
HNG1S2_2	3582062	486.95	139.17	151	3532166	102.15
HNG1S2_3	3058082	414.53	138.37	151	3023924	86.79

Table 18. Summary for Nanopore sequencing of *Salmonella* isolates and number of subsampled reads according to the time of their generation.

Isolate	Number of raw reads	Total Bases (Mbps)	Mean length	Median length	Total reads after filtering	Mean depth
G1BLS3_3	131646	357.57	3482.54	2453	102676	71.65
G2BLF1_3	29495	205.53	8819.87	5706	23303	43.62
G2M0S1_2	26713	114.32	5405.41	3158	21149	23.33
G2M4F3_3	17311	96.33	7044.57	4345.5	13674	20.4
G4BLF1_2	243766	1002.51	5247.22	3460	191055	209.96
G4BLS2_11	10933	64.92	7393.85	4395.5	8780	13.28
G5_20BLS1_1	29251	154.52	6618.00	3835	23349	32.91
G5_25BLS1_1	7749	50.91	8328.76	4359.5	6112	10.72
HNG1S2_2	24937	172.52	8733.18	5282	19755	36.93
HNG1S2_3	15291	88.72	7233.23	3353	12265	18.96

Time	15 mins	30 mins	60 mins	120 mins	240 mins	480 mins	960 mins	1500 mins	Full
Number of reads	3481	8326	20095	48679	105260	200706	337642	422052	422118

Table 19. Mean depth of the filtered Nanopore subsampled reads set.

Isolate	15 mins	30 mins	60 mins	120 mins	240 mins	480 mins	960 mins	1500 mins	Full
G1BLS3_3	0.53	1.31	3.35	8.43	18.79	36.06	58.6	71.64	71.65
G2BLF1_3	0.19	0.64	1.7	4.22	9.65	19.88	34.82	43.62	43.62
G2M0S1_2	0.12	0.37	0.92	2.5	5.66	11.21	19.01	23.33	23.33
G2M4F3_3	0.17	0.43	1	2.36	5.23	10.24	16.7	20.4	20.4
G4BLF1_2	0.86	2.21	6.16	16.94	41.98	90.06	164.39	209.94	209.96
G4BLS2_11	0.11	0.23	0.73	1.68	3.61	6.82	10.96	13.28	13.28
G5_20BLS1_1	0.16	0.45	1.26	3.41	7.81	15.34	26.29	32.9	32.91
G5_25BLS1_1	0.08	0.27	0.67	1.45	3.02	5.53	8.82	10.72	10.72
HNG1S2_2	0.27	0.71	1.77	4.1	9.14	17.85	29.92	36.93	36.93
HNG1S2_3	0.18	0.38	0.86	2.24	4.86	9.51	15.55	18.95	18.96

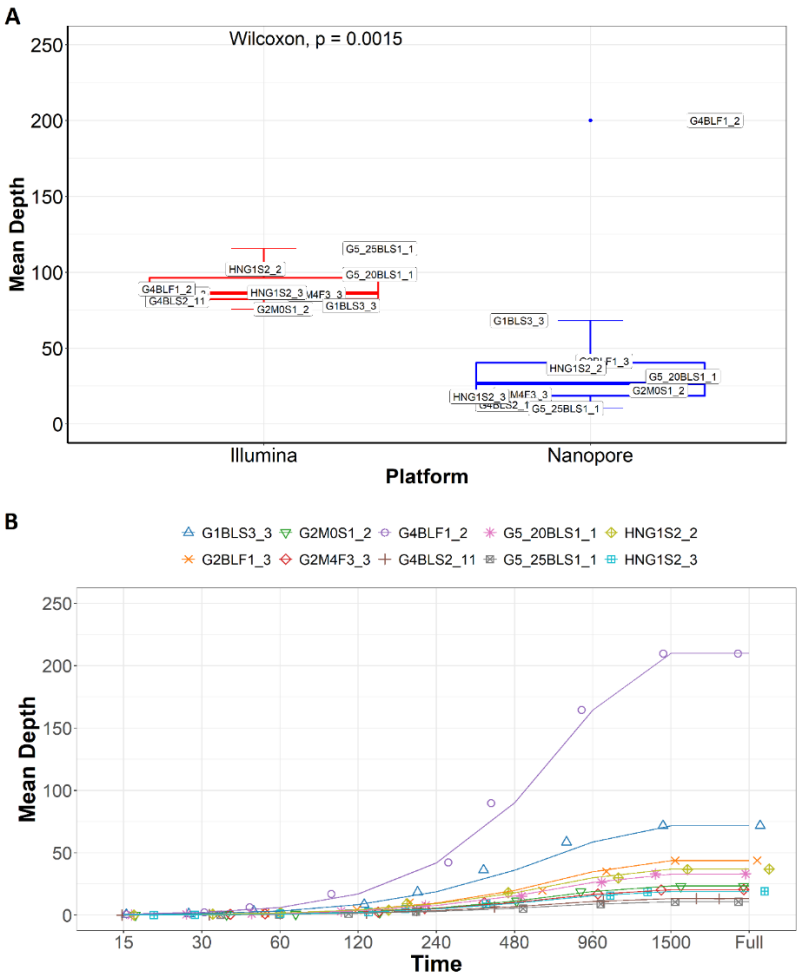

Figure 11. Mean depth of (A) the complete filtered reads set obtained from Illumina and Nanopore sequencing, and (B) the subsampled filtered Nanopore reads.

As we explained in the previous chapter, it is common to obtain a variable amount of reads between different barcoded samples when multiplexing during the library preparation step of the sequencing process (Table 17, 18, Figure 11A); this phenomenon was also reflected in all the subsampled reads over time (Table 19, Figure 11B), this may be largely caused by the difference in sequenced fragment sizes as explained above. The difference between the mean depth values from Nanopore in comparison to Illumina occurred because, unlike the previous chapter, in this case, we barcoded 12 isolates to

sequence them using one flowcell with the maximum number of barcodes present in the Rapid Barcoding Sequencing kit (SQK-RBK004), of which only 10 were chosen for further analysis due to contamination of the remaining 2 (data not shown). Additionally, after 1500 mins (25 h.) we could see that no more data was generated (Table 19, Figure 11B), which could have been caused by the early clogging of the pores of the flowcell used, a problem generally attributed to the use of the Rapid Barcoding Sequencing Kit (Maghini et al., 2021) (see https://community.nanoporetech. com/contaminants). Despite the limitations presented, we aimed to optimize the use of resources while testing the limits of this platform in the identification and characterization of the leading foodborne bacterial pathogen in the U.S. optimizing the use of a sequencing flowcell with MinION (Eng et al., 2015).

Species identification, serotyping, and specific pathogenic islands determination

No species other than Salmonella were identified in the reads from both platforms (Figure 12A, B), classifying an average of 25.53% and 5.28% reads specifically for *Salmonella* in Illumina and Nanopore, respectively. This step took an average of 2 s per isolate using 1 core (2.8 Gb). Additionally, we could confirm the identity of the isolates from the 15 mins subsampled filtered reads (Figure 12C). As the result of the serotype prediction performed on the best hybrid assemblies, we were able to obtain the antigenic profiles and the serotype name based on the Kauffman White Scheme of all the analyzed isolates (Table 20). The serotypes found were Bareilly (n = 3), Litchfield (n = 4), Newport (n = 2), and Muenchen (n = 1). Comparing these results and the serotype prediction performed using the subsampled reads as input, we demonstrated that 480

mins (8 h.) of Nanopore sequencing were enough to determine the serotype of all our isolates without requiring prior reads assembly. It took an average time of 5 s with 1 core (2.8Gb) for the detection of serotypes per isolate using the full Nanopore filtered reads. However, the identification of *Salmonella* pathogenicity islands (SPIs) carried out directly in both filtered Nanopore and Illumina reads, was not similar to the results obtained from the same analysis using the best hybrid genomes as input (data not shown). When taking a look at the lengths of the sequences in the spifinder database (retrieved on June 28, 2021), we could see that the hits that could not be detected with the reads came from sequences of up to 133,638 nucleotides, we decided thereby to include the same analysis after obtaining all the assemblies from the subsampled reads (see below).

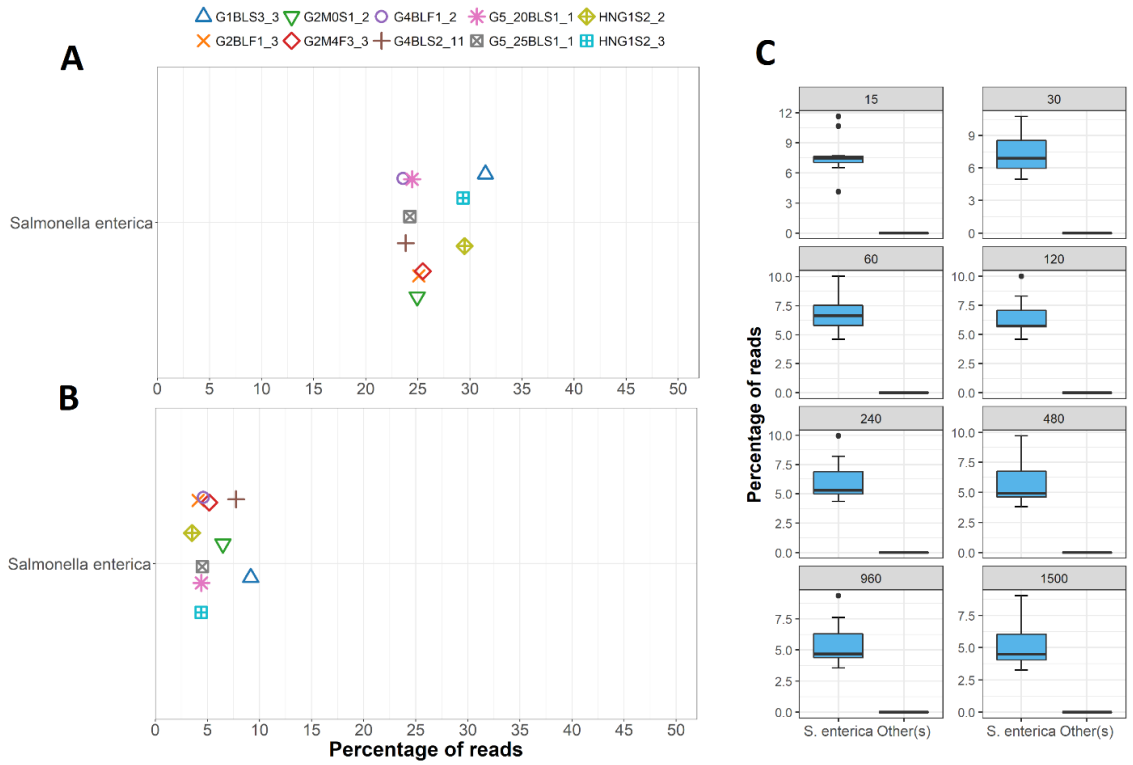

Figure 12. Percentage of reads classified to species level by Kraken 2 in the filtered reads set from (A) Illumina and (B) Nanopore, as well as for (C) the subsampled filtered Nanopore reads.

Table 20. The antigenic profile of the *Salmonella* isolates and from which set of reads it could be detected. The antigenic profile of *Salmonella* is composed of: O, O-antigen or somatic antigen; and two H-antigens or flagellar antigens.

Isolate	15 mins	30 mins	60 mins	120 mins	240 mins	480 mins	960 mins	1500 mins	Antigenic profile	Serotype
G1BLS3_3	-:-:-	-:-:-	7:-:-	7:y:1,5	7:y:1,5	7:y:1,5	7:y:1,5	7:y:1,5	7:y:1,5	Bareilly
G2BLF1_3	-:-:-	-:-:-	-:-:-	-:l,v:-	8:l,v:1,2	8:l,v:1,2	8:l,v:1,2	8:l,v:1,2	8:l,v:1,2	Litchfield
G2M0S1_2	-:-:-	-:-:1,5	-:-:1,5	-:y:1,5	7:y:1,5	7:y:1,5	7:y:1,5	7:y:1,5	7:y:1,5	Bareilly
G2M4F3_3	-:-:-	-:-:-	8:-:-	8:-:1,2	8:l,v:1,2	8:l,v:1,2	8:l,v:1,2	8:l,v:1,2	8:l,v:1,2	Litchfield
G4BLF1_2	-:-:-	-:-:1,2	-:e,h:1,2	8:e,h:1,2	8:e,h:1,2	8:e,h:1,2	8:e,h:1,2	8:e,h:1,2	8:e,h:1,2	Newport
G4BLS2_11	-:-:-	-:-:-	-:-:-	7:-:-	7:y:1,5	7:y:1,5	7:y:1,5	7:y:1,5	7:y:1,5	Bareilly
G5_20BLS1_1	-:-:-	-:-:-	-:-:-	-:-:-	8:d:1,2	8:d:1,2	8:d:1,2	8:d:1,2	8:d:1,2	Muenchen
G5_25BLS1_1	-:-:-	-:-:-	-:-:-	-:e,h:1,2	8:e,h:1,2	8:e,h:1,2	8:e,h:1,2	8:e,h:1,2	8:e,h:1,2	Newport
HNG1S2_2	-:-:-	-:-:-	-:-:1,2	-:-:1,2	-:l,v:1,2	8:l,v:1,2	8:l,v:1,2	8:l,v:1,2	8:l,v:1,2	Litchfield
HNG1S2_3	-:-:-	-:-:-	-:-:-	-:l,v:1,2	-:l,v:1,2	8:l,v:1,2	8:l,v:1,2	8:l,v:1,2	8:l,v:1,2	Litchfield

Again, the percentage of filtered reads of Nanopore assigned to *Salmonella enterica* is lower than that of Illumina, a possible effect of the higher error rate of Nanopore with respect to Illumina, ~10% and ~1%, respectively (Chandak et al., 2020; Tyler et al., 2018), or as mentioned before the specificity of this analysis benefits from the larger size of the reads since the greater the size of a read, the greater number of k-mers specific to a species will be required to classify that read to that specific taxon (Leidenfrost et al., 2020; Pearman et al., 2020). And as in the previous chapter, we were able to detect the specific species in the 15 mins subsampled reads.

The species *Salmonella enterica* has six subspecies with more than 2600 serotypes that differ from each other based on their O-antigen, as well as, their H-antigen in which depending on the strain, an additional H-antigen may be present as a result of flagellar phase variation produced by different expression levels of the invertase *hin* and the *fliC* repressor gene *fljA* (Andino & Hanning, 2015; Barco et al., 2014; Brenner et al., 2000; Crump & Wain, 2017). Traditional serological techniques for the identification of the antigenic profile of *Salmonella* relied on the availability of more than 150 specific antisera and well-trained personnel to correctly interpret the results (Diep et al., 2019; Wattiau et al., 2008), whereby autoagglutination or loss of antigen can lead to false positives (Wattiau et al., 2008). The appearance of techniques based on the detection of loci or genes related to the production of different combinations of antigens in *Salmonella* gave rise to bioinformatic tools that (Herrera-León et al., 2007), through the use of databases previously generated from different isolates, seek to predict the antigenic profile through WGS, for instance, SeqSero, a k-mer based algorithm for rapid serotype prediction from raw reads or genome assemblies (Zhang et al., 2019), which showed 98% concordance with serotyping reported from routine use on 520 isolates (20 serotypes) (S. Banerji et al., 2020). All of the serotypes identified in the isolates tested are multi-host and have been reported as responsible for different outbreaks in the U.S. (CDC, 2008, 2016, 2020; Chapple et al., 2017; Hoffmann et al., 2016). Several studies have shown high precision when determining serotypes using *in silico tools* with the raw data generated from WGS (Sangeeta Banerji et al., 2020; Ibrahim & Morin, 2018; Mohammed & Thapa, 2020; Xu et al., 2020), herein, we achieved to serotype the isolates tested even at 480 mins (8 h.) of sequencing time with Nanopore, which reduces the time and

computational resources required to subtype different *Salmonella* serotypes (Taylor et al., 2019). In the previous chapter, we had to assemble the sequences to detect all the antigens of the *E. coli* serotypes, this could be due to the fact that the size of the sequences for the identification of the H antigen used by SeqSero is shorter compared to the used for *E. coli* (Figure 13), requiring less query coverage to determine specific antigenic profiles.

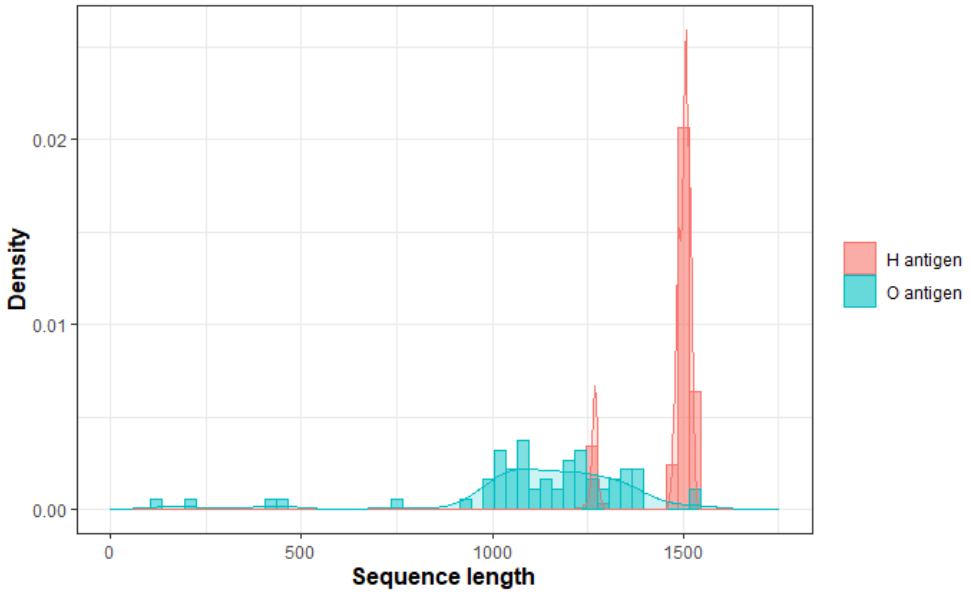

Figure 13. Distribution of gene length in the O- and H-antigen database used by SeqSero (retrieved in October 2020).

Continuity of the Genome Assemblies

In spite of the low mean depth values obtained from the Nanopore sequencing with values below 45x, i.e., the recommended value for scaffolding of bacterial genomes (Goldstein et al., 2019; Karlsson et al., 2015), except for the isolate G4BLF1_2 (Table 18), high values of N50 and a low number of contigs were obtained from the assessment

of the best hybrid genomes thus demonstrating high continuity which is mostly the effect of Nanopore long reads (Table 21). For comparison of the subsampled reads, we were able to obtain assemblies for all of our isolates starting at 120 mins (2 h.) due to the unequal mean depth values, hence the assemblies generated from the set of subsampled reads at 120 mins was used as the starting point for the subsequent comparisons. Wherein, after 480 mins (8 h.), we could obtain assemblies significantly similar to those produced from the full set of reads ($\alpha = 0.05$) (Figure 14A, Table 22).

Table 21. Assembly metrics for the best hybrid assemblies obtained for each *Salmonella* isolate.

Isolate	Number of contigs	Largest contig	Total length	GC (%)	N50	L50
G1BLS3_3	3	4808067	4986936	52.07	4808067	1
G2BLF1_3	1	4722110	4722110	52.25	4722110	1
G2M0S1_2	2	4837849	4909251	52.09	4837849	1
G2M4F3_3	1	4722253	4722253	52.25	4722253	1
G4BLF1_2	1	4784938	4784938	52.2	4784938	1
G4BLS2_11	2	4823151	4894546	52.01	4823151	1
G5_20BLS1_1	1	4699744	4699744	52.13	4699744	1
G5_25BLS1_1	2	4739838	4741384	52.25	4739838	1
HNG1S2_2	1	4675899	4675899	52.27	4675899	1
HNG1S2_3	1	4675843	4675843	52.27	4675843	1

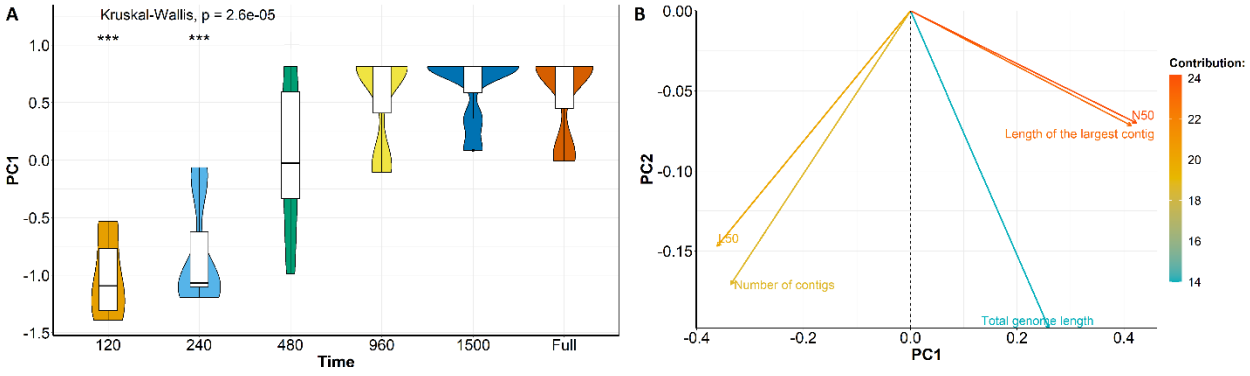

Figure 14. Multivariate comparisons based on continuity between assemblies created from subsampled filtered Nanopore reads. (A) Comparison between PC1 values associated with continuity of the assemblies

obtained from the subsampled reads against the full set of filtered Nanopore reads with respect to time. (B) Contribution plot of the variables used for the multivariate analysis. The colors in the arrows represent the contribution weight for each variable, with a gradient from light blue (low contribution) to red (high contribution). PC1 and PC2 explain 80% and 11.9% of the total variance of the complete data set, respectively. ***: *p*-values < 1E-02.

Table 22. *p*-values obtained from the Dunn's Multiple Comparison between the PC1 values associated with continuity of the assemblies obtained from the subsampled reads against the full set of filtered Nanopore reads with respect to time.

	Z	p-values	Corrected p-values
120 mins	-3.70	2.125E-04	2.017E-03
240 mins	-3.38	7.331E-04	3.480E-03
480 mins	-1.68	9.343E-02	1.774E-01
960 mins	-0.41	6.789E-01	9.916E-01
1500 mins	0.09	9.306E-01	1

Goldstein et al. (2019) demonstrated that low complexity genomes can be assembled with low reads yield as well as low mean depth values, wherein the factors that contribute to increasing genomic complexity are an elevated GC content (Benjamini & Speed, 2012; Goldstein et al., 2019), as well as, the presence of mobile genetic elements (Bohlin et al., 2017; Hayek, 2013). For which, the analyzed assemblies showed an expected GC content for the species *S. enterica* (~52.2%) which is considered intermediate GC content (Papanikolaou et al., 2009), and that in subsequent analyzes we were able to show a low prevalence of mobile genetic elements (see below) thus contributing to the low complexity of the assembled genomes and explaining the reason for their high degree of continuity.

By observing the loading plots (Figure 14B), we can confirm how the N50, the total length of the genome, and the length of the largest contig are positively correlated as a

function of the angles of the vectors that represent them; in addition, these parameters are negatively correlated with the L50 and the total number of contigs, which are related to the degree of fragmentation of the assemblies (Thrash et al., 2020). Furthermore, in the loading plots (Figure 14B), we were also able to corroborate that the total estimated genome size is the parameter that contributed the least variability in the analysis, which shows that the length of the readings and the mean depth obtained from the Nanopore sequencing were sufficient to achieve assemblies with sizes that were close to those obtained from the full set of reads from early stages of sequencing thanks to the low complexity of the analyzed genomes (Fachada et al., 2016; Goldstein et al., 2019; Zitko, 1994).

Features annotation in genome assemblies

In Table 23, we can see the features annotated in the best hybrid genomes of the *Salmonella* isolates. From which, only G1BLS3_3, G2M0S1_2, G4BLS2_11, G5_25BLS1_1, and G2_25BLS1_1 have plasmids. However, multiple prophages could be predicted in all the isolates tested. Additionally, several CRISPRs could be detected in the 10 isolates, and, we were also able to find certain Biosynthetic gene clusters (BGCs) in these assemblies, among which we found non-ribosomal peptide synthetases (NRPSs), thiopeptide, bacteriocins, and oligosaccharides. In addition, genes, repeated regions, rRNA, tmRNA, and tRNA were also annotated in all the assembled genomes. These features, except for the number of BGCs, as well as, the number of plasmids that showed a variance equal to 0 in 6 isolates (data not shown), were employed in the multivariate comparison against the assemblies created from the subsampled Nanopore reads. Finding

that from 960 mins (16 h.), we were able to obtain results significantly similar to those obtained from the best hybrid assemblies ($\alpha = 0.05$) (Figure 17A, Table 23). Wherein, except for the GC content that contributed less than other variables to the total variation, all the variables analyzed had a positive correlation and contributed negatively to PC1 (Figure 15B).

Finally, from the SPIs search we carried out with the assemblies, we could obtain all hits in the subsampled reads of the times shown in Table 25.

Table 23. Number of annotated features for the best hybrid assemblies obtained for each *Salmonella* isolate. rRNA, ribosomal RNA; tmRNA, transfer-messenger RNA; tRNA, transfer RNA; CRISPRs, clustered regularly interspaced short palindromic repeats; BGCs, biosynthetic gene clusters.

Isolate	Genes	Repeated regions	rRNA	tmRNA	tRNA	Plasmids	CRISPRs	BGCs	Prophages
G1BLS3_3	4834	1	22	1	86	1	8	2	6
G2BLF1_3	4464	2	22	1	85	0	9	2	5
G2M0S1_2	4746	1	22	1	87	1	8	2	7
G2M4F3_3	4477	2	22	1	85	0	9	2	5
G4BLF1_2	4545	2	22	1	85	0	8	2	5
G4BLS2_11	4728	1	22	1	87	1	8	2	6
G5_20BLS1_1	4456	1	22	1	84	0	7	2	4
G5_25BLS1_1	4491	2	22	1	85	1	8	2	4
HNG1S2_2	4415	2	22	1	85	0	9	2	4
HNG1S2_3	4413	2	22	1	85	0	9	2	4

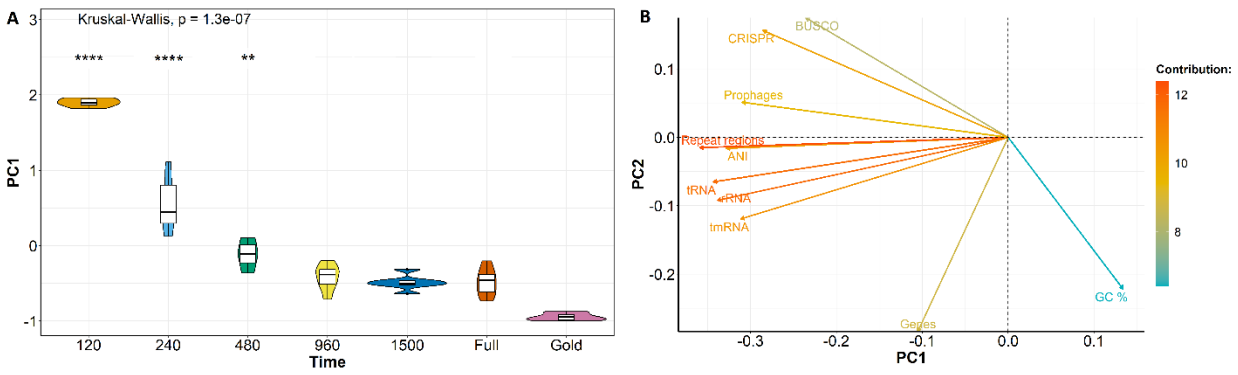

Figure 15. Multivariate comparisons based on genomic features between assemblies created from subsampled filtered Nanopore reads and best hybrid assemblies (Gold). (A) Comparison between PC1 values associated with genomic features of the assemblies obtained from the subsampled reads against the best hybrid assemblies with respect to time. (B) Contribution plot of the variables used for the multivariate analysis. The colors in the arrows represent the contribution weight for each variable, with a gradient from light blue (low contribution) to red (high contribution). PC1 and PC2 explain the 65.2% and 16.7% of the total variance of the complete data set, respectively. ****: p < 1E-04. ***: 1E-04 < p < 1E-03. **: 1E-03 < p < 1E-02. *: 1E-02 < p < 5E-02.

Table 24. *p*-values obtained from the Dunn's Multiple Comparison Test between the PC1 values associated with genomic features of the assemblies obtained from the subsampled reads against the best hybrid assemblies with respect to time.

	Z	*p*-values	Corrected *p*-values
120 mins	5.50	3.811E-08	1.239E-06
240 mins	4.58	4.588E-06	7.457E-05
480 mins	3.50	4.689E-04	3.049E-03
960 mins	2.19	2.863E-02	9.307E-02
1500 mins	1.66	9.596E-02	2.080E-01
Full	1.81	6.961E-02	1.741E-01

Table 25. Pathogenic islands specific for *Salmonella* detected from the assemblies of the analyzed isolates. C63PI: centisome 63 pathogenic island. CS54: centisome 54. SPI: *Salmonella* Pathogenic Island

Isolate	*Salmonella* Pathogenic Island	Time
G1BLS3_3	C63PI,CS54,SPI-1,SPI-13,SPI-14,SPI-2,SPI-3,SPI-4,SPI-5	120 mins
G2BLF1_3	C63PI,CS54,SPI-1,SPI-13,SPI-14,SPI-2,SPI-3,SPI-4,SPI-5	240 mins
G2M0S1_2	SPI-1,SPI-13,SPI-14,SPI-2,SPI-3,SPI-4,SPI-5	240 mins
G2M4F3_3	SPI-1,SPI-13,SPI-14,SPI-2,SPI-3,SPI-4,SPI-5	480 mins
G4BLF1_2	C63PI,CS54,SPI-1,SPI-13,SPI-14,SPI-2,SPI-3,SPI-4,SPI-5	240 mins
G4BLS2_11	C63PI,CS54,SPI-1,SPI-13,SPI-14,SPI-2,SPI-3,SPI-4,SPI-5	1500 mins
G5_20BLS1_1	C63PI,CS54,SPI-1,SPI-13,SPI-14,SPI-2,SPI-3,SPI-4,SPI-5	480 mins
G5_25BLS1_1	SPI-1,SPI-13,SPI-14,SPI-2,SPI-3,SPI-4,SPI-5	1500 mins
HNG1S2_2	C63PI,CS54,SPI-1,SPI-13,SPI-14,SPI-2,SPI-3,SPI-4,SPI-5	240 mins
HNG1S2_3	C63PI,CS54,SPI-1,SPI-13,SPI-14,SPI-2,SPI-3,SPI-4,SPI-5	240 mins

In these isolates, we found a reduced number of plasmids, limiting the use of this variable to compare all isolates. Yet we could find an average of 5 prophages, a value close to 5.29, i.e., the average number found by Bobay et al. (2013) from 21 analyzed genomes of *S. enterica*. Prophages encompass the largest horizontal gene transfer mechanisms in *Salmonella*, contributing genes in up to 5% of the total genomic content (Wahl et al., 2019). Prophages can be present in a dormant state and be vertically transmitted and induced under stress conditions, such as DNA damage or upon entering the gut of a host (Kim et al., 2014; Wahl et al., 2019); or they may also undergo spontaneous induction, which can provide phenotypic advantages to the strain that harbors them (Bossi et al., 2003). Whereas, the presence of CRISPRs in all our isolates is clear evidence of the adaptive immune response of bacteria against certain phages, and can provide key insights for phylogenetic inferences focused on the accessory genome present in different *Salmonella* serotypes to understand the horizontal gene transfer

events to which they have been subjected (Gupta et al., 2019; Kushwaha et al., 2020), as well as, the presence of certain CRISPRs can be associated with regulatory mechanisms of biosynthetic pathways present in *Salmonella*, such as the regulation of quorum sensing through CRISPR-cas3 for the production of bacterial biofilm (Cui et al., 2020; Kushwaha et al., 2020). On the other hand, the exploration of BGCs can lead to a better understanding of the chemical communication used by this pathogen, opening the door to the detection of targets for the decontamination of these isolates through disinfectants or specific antimicrobials (Gulick, 2017), as well as, identifying strains capable of generating biofilms and attaching to several food matrixes, which can prevent eventual cross-contamination between food batches in a manufacturing plant or supermarket, thus reducing the risk of an outbreak occurring (Galié et al., 2018; Wang et al., 2013). Despite having found BGCs distributed in our isolates, we could not use them as a variable in our multivariate analysis since 6 isolates had no variability. Indeed, this could be the result of obtaining assemblies with a size similar to that of the best hybrid assemblies with subsampled reads from early sequencing stages thus benefiting the prediction of BGCs through the compositional analysis performed by Antismash, which is favored by the continuity of these assemblies (Blin et al., 2019; Goldstein et al., 2019). Finally, we included the ANI values and BUSCO scores in the genomic features multivariate analysis, which indicate how identical the assembly generated is to a reference genome (Chen, Erickson, et al., 2020) and how complete this same assembly is based on highly conserved housekeeping orthologs present in the family Enterobacteriaceae (Seppey et al., 2019), respectively.

Salmonella pathogenic islands (SPI) have been the object of study because they are genomic regions that harbor virulence genes associated with the pathogenicity of *Salmonella* (Lyu et al., 2021). At present, 24 SPIs have been identified, but not all have been experimentally validated (Lerminiaux et al., 2020). SPI-1 and SPI-2 are related to the pathogenicity of Nontyphoidal *Salmonella* (Suez et al., 2013), the first of these plays a fundamental role in the invasion of host cells and the regulation of the host's immune response (L. Lou et al., 2019), and the latter is involved in intracellular survival and replication (Abd El Ghany et al., 2016; L. Lou et al., 2019; Lyu et al., 2021). As shown in Table 25, the isolates with the lowest mean depth values (G4BLS2_11 and G5_25BLS1_1 with 13.28x and 10.72x, respectively) were the bottleneck for obtaining the SPIs from the best hybrid genomes, and in spite of the high continuity that all assemblies presented, low mean depth values provide insufficient information to improve the quality of the sequences, hence the data generated at 1500 mins (25 h.) was necessary to improve precision when detecting long elements (Leidenfrost et al., 2020). An average of 33 mins was necessary to obtain the assembled genomes from full Nanopore reads with the implemented bioinformatics pipeline using 32 cores (89.6 Gb).

AR and VFs genes annotation

In the searches we carried out in the best hybrid assemblies, we were able to find that all isolates harbor genes that could make them multidrug resistant (Figure 16A). According to the report of the greatest threats regarding AR generated in 2019 in the U.S. (CDC, 2019), Nontyphoidal *Salmonella* is considered a serious threat, which could be

explained since the situation has been aggravating in recent years as antibiotic resistant clones of this species are frequently implicated as etiological agents in outbreaks that lead to a greater number of affected patients with severe Salmonellosis (Nair et al., 2018). The essential first and second line drugs to treat this disease are ceftriaxone (cephalosporin type), ciprofloxacin (fluoroquinolone type), ampicillin (penam or penicillin type), azithromycin (macrolide type), and trimethoprim-sulfamethoxazole (sulfonamide) (CDC, 2019; Nair et al., 2018). Whereby the detection of these resistance genes towards these drugs is of great vitality to find a way to counteract the presence of this pathogen in reservoirs, including food animals (Nair et al., 2018; Souza et al., 2020; Zhao et al., 2009). Among the resistance patterns predicted, we can find the antibiotics used for the treatment of severe Salmonellosis already mentioned (Figure 16A). Nonetheless, it is important to emphasize that the results obtained provide a broad panorama regarding the potential resistance mechanisms the studied isolates may have, however in some cases genotypic results may not be reflected in the phenotypic tests due to possible regulatory mechanisms that do not favor the expression of the genes found, and the presence of mutations or genes not previously described that confer resistance to these or different antibiotics (Hendriksen et al., 2019; Köser et al., 2014). For this reason, it would be important to be able to carry out antimicrobial sensitivity tests for these isolates in the event that they represent a threat to the owner of the orchard from which they were isolated (Ellington et al., 2017).

On the other hand, the set of VFs detected in the best hybrid assemblies (Figure 16B) in conjunction with the presence of SPI-1 and SPI-2 in all isolates (Table 25) could favor an

intracellular invasion of host cells (Lerminiaux et al., 2020; Lixin Lou et al., 2019; Yeom et al., 2020).

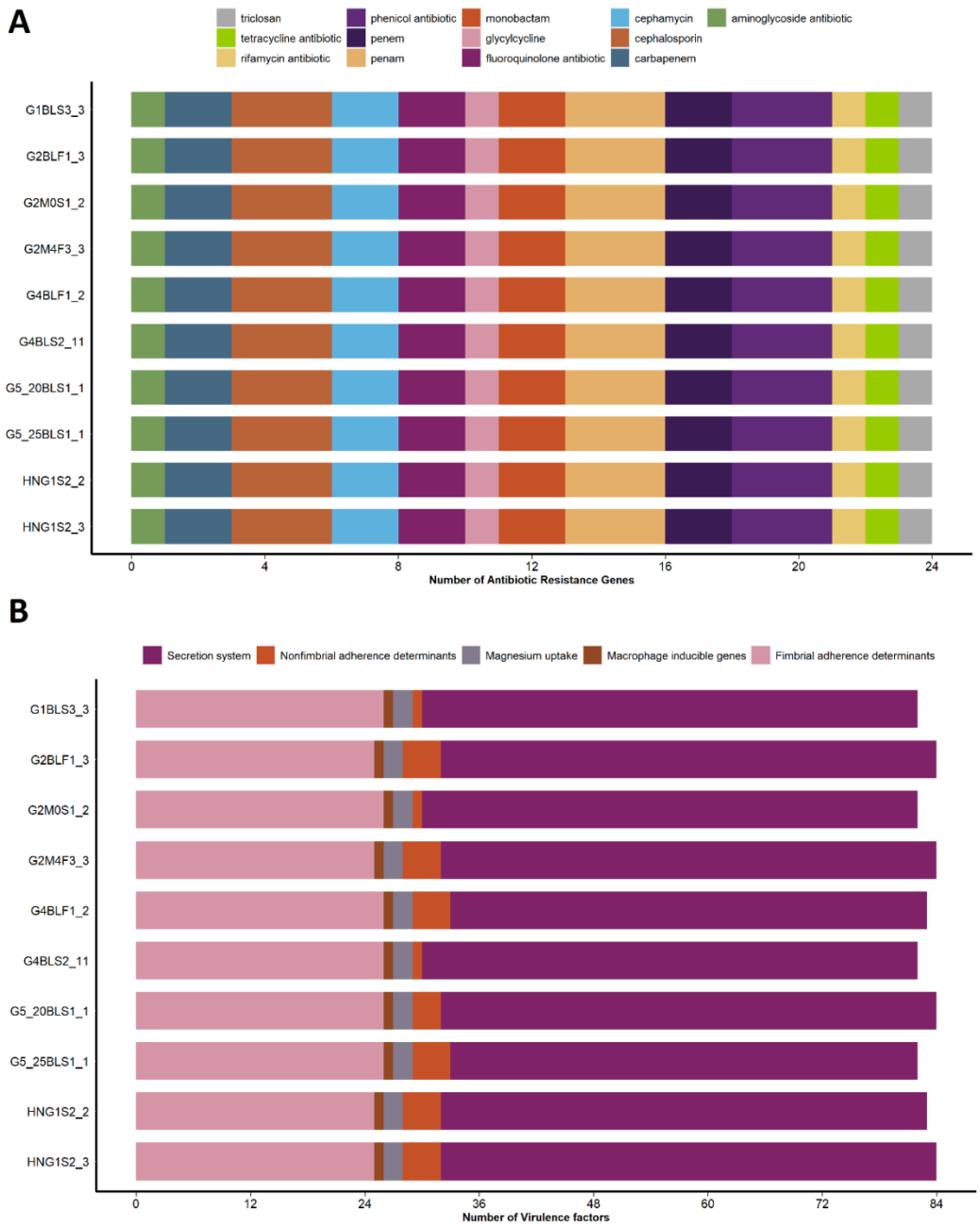

Figure 16. Number of gene ontologies associated with (A) the AR and (B) the VFs genes identified in the best hybrid genomes (gene ontology was analyzed using the aro_index.tsv file from the CARD and the intra-genera VFs comparison tables from the VFDB for AR and VFs, respectively).

The comparison of the hits obtained from the stringent AR search carried out in the genomes of the subsampled reads against the hits obtained from the best hybrid genomes was enough to obtain results significantly similar from reads obtained at 480 mins (8 h.) ($\alpha = 0.05$) (Table 26, Figure 17A), although the hits obtained from the stringent VFs search was not enough to be significantly similar to the results obtained using the best hybrid assemblies ($\alpha = 0.05$) (Table 26, Figure 17B). Nonetheless, an average of 96,37% of total hits found for the VFs hits was determined in the searches performed with the assemblies created using the full set of Nanopore reads for each isolate, whereby from 960 mins (16 h.) no significant differences were obtained between the results from these assemblies (Figure 17B). With the use of Nanopore sequencing accompanied by bioinformatic tools that seek to get the most out of this technology, we were able to capture a large portion of information from our isolates in 1500 mins (25 h.), and despite being able to extend the turnaround time it is worth noting that 12 isolates were sequenced in one flowcell, which implies a greater cost benefit and even portability that the use of MinION can offer (Bull et al., 2020).

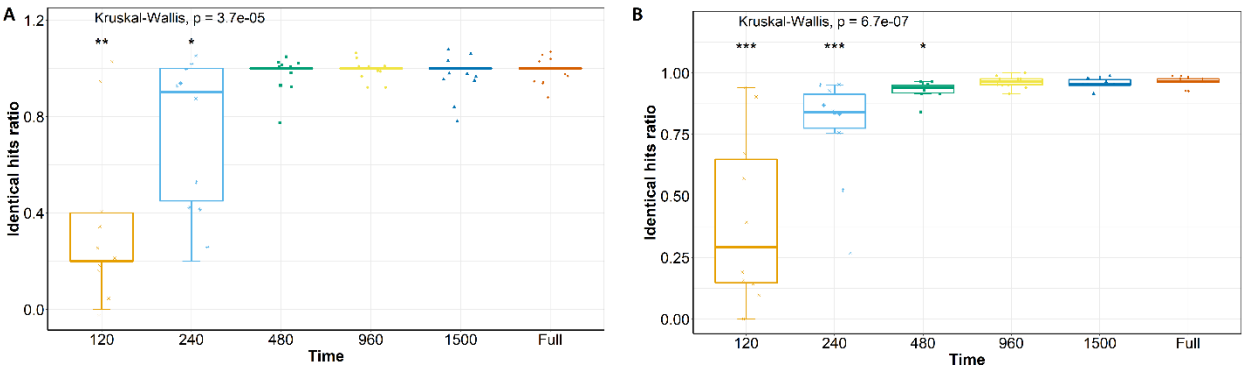

Figure 17. Identical hits ratio comparison between the genes obtained from assemblies created using the subsampled filtered Nanopore reads. (A) Comparison of the hits obtained from the

search for AR genes. (B) Comparison of the hits obtained from the search for VFs. **: 1E-02 < p. *: 1E-02 < p < 5E-02.

Table 26. *p*-values obtained from the one-sample Wilcoxon signed rank test of the identical hits ratio. $P > 0.05$ indicates that the results of the assemblies from the subsampled reads at that time are significantly similar to the best hybrid assemblies.

	AR genes		***Virulence factors***	
	***p*-values**	**Corrected *p*-values**	***p*-values**	**Corrected *p*-values**
120 mins	6.435E-03	3.861E-02	9.766E-04	3.514E-03
240 mins	2.895E-02	4.686E-02	2.929E-03	3.514E-03
480 mins	5.000E-01	6.000E-01	2.929E-03	3.514E-03
960 mins	1	1	4.545E-03	4.545E-03
1500 mins	1	1	2.913E-03	3.514E-03
Full	1	1	2.913E-03	3.514E-03

Phylogenetic inference

From the phylogenetic analysis of the best hybrid assemblies (Figure 18), we could see that the isolates whose serotype was previously predicted, present low genetic variability and grouped in clades that match the four different serotypes identified (Table 20). During the core SNPs matrices generation from the assemblies obtained using the subsampled reads, we could notice that from the reads subsampled at 240 mins (4 h.), we began to have enough data to carry out the phylogenetic analyzes, thereby we decided to subsample the reads that were generated at 180 mins (3 h.) from which it was possible to perform the pertinent phylogenetic analysis and its subsequent comparison with the trees generated at later minutes. In the comparisons made, from 180 mins (3 h.), we could obtain phylogenies significantly similar to those obtained from the best hybrid genomes ($\alpha = 0.05$) (Table 27). However, as mentioned in the previous chapter, we only took into

account the topology of the trees and not the length of the branches using a lambda value of 0 in the distances calculations among trees when using the Kendall-Colijin test (Kendall & Colijn, 2016). Therefore, we believe that these results would not represent analyses where it is desired to quantify the amount of evolutionary divergence between isolates (Paradis, 2016). It took 3 hours on average to carry out the phylogenetic analyzes with the kSNP3-RAxML strategy using 16 cores (44.8 Gb).

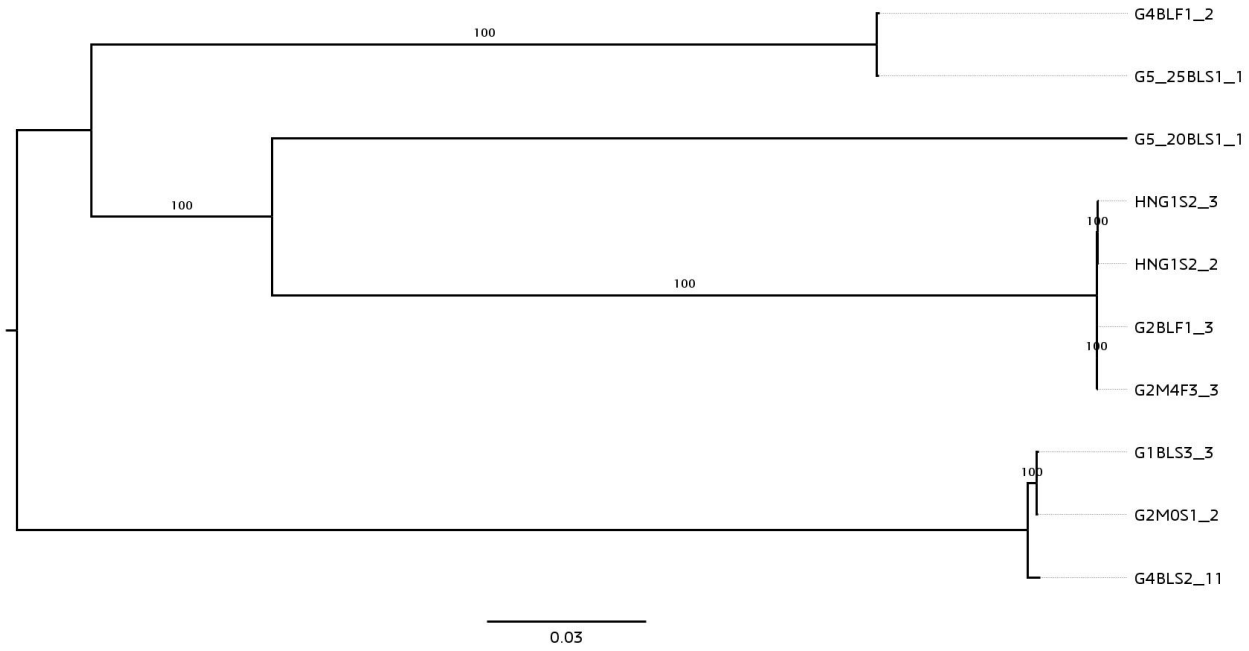

Figure 18. A maximum likelihood tree constructed using RAxML based on the core SNPs dataset of the best hybrid assemblies for the 10 *Salmonella* isolates.

Table 27. *p*-values obtained from the Kendall-Colijin test between the topologies of the core SNPs phylogenetic trees generated from the subsampled filtered Nanopore reads and the best hybrid assemblies.

	p-values	***Corrected p-values***
180 mins	3.34E-04	2.00E-03
240 mins	5.00E-03	3.00E-02
480 mins	6.31E-11	3.78E-10
960 mins	6.31E-11	3.78E-10
1500 mins	5.58E-08	3.35E-07
Full	5.58E-08	3.35E-07

Despite the high level of agreement with the results obtained of serotyping that can be obtained with subtyping tools based on the genes involved in the synthesis of the O antigen and the H antigens (S. Banerji et al., 2020), some strains still give rise to discrepancies between the tests carried out *in vivo* compared to the results obtained *in silico*, and this is usually due to the presence of complex groups of Salmonella, or a lower proportion of novel *Salmonella* (Chattaway et al., 2019). As mentioned in the previous chapter, SNP phylogenetic trees can offer key insights when understanding the origin of an outbreak or identifying potential reservoirs of pathogenic bacteria (Kingry et al., 2016; Lindsey et al., 2016; Schürch et al., 2018). Chattaway et al. (2021), through an extensive analysis of sequence data generated from all *Salmonella enterica* isolates referred from England and Wales to the *Salmonella* Reference Unit over 5 years, proposed to remove the need for antibody based serotyping and make the transition to classification based on phylogenetic methods from WGS data. Herein, we obtained isolates with the same serotype grouped in the same unique clades, a result consistent with the findings obtained from other phylogenetic analyzes that demonstrated a high correlation between the antigenic profile of *Salmonella* and phylogenetic analyzes (Achtman et al., 2012; Alikhan

et al., 2018). Through the time-based comparison using the assemblies created from the subsampled Nanopore reads yielded from one flowcell and using the maximum number of barcodes included in the Nanopore's Rapid Barcoding Sequencing kit (SQK-RBK004), we could produce the necessary information to infer phylogenetic trees with significant information in the topology that is present through the bioinformatic pipeline that we used.

CONCLUSIONS

Exemplifying a real life scenario where it is required to lower costs and optimize the use of resources, we show that Nanopore sequencing reads obtained from 3 hours in this study could be used to obtain phylogenetic analyzes with a higher resolution than other traditional techniques. And even at 25 hours, it can offer certain capabilities to detect fundamental elements in case it is required to characterize molecularly a group of isolates and identify potential reservoirs of Nontyphoidal *Salmonella*. With future advances in the chemistry of this technology, as well as bioinformatic advances, it is expected that the lower precision of Nanopore will be surpassed and it will become a fundamental tool for the wide-range characterization of strains that can represent a danger to the public health systems.

REFERENCES

Abd El Ghany, M., Shi, X., Li, Y., Ansari, H. R., Hill-Cawthorne, G. A., Ho, Y. S., Naeem, R., Pickard, D., Klena, J. D., Xu, X., Pain, A., & Hu, Q. (2016). Genomic and Phenotypic Analyses Reveal the Emergence of an Atypical Salmonella enterica Serovar Senftenberg Variant in China. *J Clin Microbiol, 54*(8), 2014-2022. https://doi.org/10.1128/jcm.00052-16

Acheson, D., & Hohmann, E. L. (2001). Nontyphoidal Salmonellosis. *Clinical Infectious Diseases, 32*(2), 263-269. https://doi.org/10.1086/318457

Achtman, M., Wain, J., Weill, F.-X., Nair, S., Zhou, Z., Sangal, V., Krauland, M. G., Hale, J. L., Harbottle, H., Uesbeck, A., Dougan, G., Harrison, L. H., Brisse, S., & the, S. e. M. s. g. (2012). Multilocus Sequence Typing as a Replacement for Serotyping in Salmonella enterica. *PLOS Pathogens, 8*(6), e1002776. https://doi.org/10.1371/journal.ppat.1002776

Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A. V., Cheng, A. A., Liu, S., Min, S. Y., Miroshnichenko, A., Tran, H. K., Werfalli, R. E., Nasir, J. A., Oloni, M., Speicher, D. J., Florescu, A., Singh, B., Faltyn, M., Hernandez-Koutoucheva, A., Sharma, A. N., Bordeleau, E., Pawlowski, A. C., Zubyk, H. L., Dooley, D., Griffiths, E., Maguire, F., Winsor, G. L., Beiko, R. G., Brinkman, F. S. L., Hsiao, W. W. L., Domselaar, G. V., & McArthur, A. G. (2020). CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res, 48*(D1), D517-d525. https://doi.org/10.1093/nar/gkz935

Alikhan, N.-F., Zhou, Z., Sergeant, M. J., & Achtman, M. (2018). A genomic overview of the population structure of Salmonella. *PLOS Genetics, 14*(4), e1007261. https://doi.org/10.1371/journal.pgen.1007261

Andino, A., & Hanning, I. (2015). Salmonella enterica: survival, colonization, and virulence differences among serovars. *TheScientificWorldJournal, 2015*, 520179-520179. https://doi.org/10.1155/2015/520179

Andrews, W. H., Wang, H., Jacobson, A., & Hammack, T. (2018). Bacteriological Analytical Manual Chapter 5 Salmonella. In. https://www.fda.gov/food/foodscienceresearch/laboratorymethods/ucm070149.htm

Banerji, S., Simon, S., Tille, A., Fruth, A., & Flieger, A. (2020). Genome-based Salmonella serotyping as the new gold standard. *Sci Rep, 10*(1), 4333. https://doi.org/10.1038/s41598-020-61254-1

Banerji, S., Simon, S., Tille, A., Fruth, A., & Flieger, A. (2020). Genome-based Salmonella serotyping as the new gold standard. *Scientific Reports, 10*(1), 4333. https://doi.org/10.1038/s41598-020-61254-1

Barco, L., Longo, A., Lettini, A. A., Cortini, E., Saccardin, C., Minorello, C., Olsen, J. E., & Ricci, A. (2014). Molecular Characterization of "Inconsistent" Variants of Salmonella Typhimurium Isolated in Italy. *Foodborne Pathogens and Disease, 11*(6), 497-499. https://doi.org/10.1089/fpd.2013.1714

Benjamini, Y., & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research, 40*(10), e72-e72. https://doi.org/10.1093/nar/gks001

Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S. Y., Medema, M. H., & Weber, T. (2019). antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research, 47*(W1), W81-W87. https://doi.org/10.1093/nar/gkz310

Bobay, L.-M., Rocha, E. P. C., & Touchon, M. (2013). The adaptation of temperate bacteriophages to their host genomes. *Molecular Biology and Evolution, 30*(4), 737-751. https://doi.org/10.1093/molbev/mss279

Bohlin, J., Eldholm, V., Pettersson, J. H. O., Brynildsrud, O., & Snipen, L. (2017). The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. *BMC Genomics, 18*(1), 151. https://doi.org/10.1186/s12864-017-3543-7

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. https://doi.org/10.1093/bioinformatics/btu170

Boore, A. L., Hoekstra, R. M., Iwamoto, M., Fields, P. I., Bishop, R. D., & Swerdlow, D. L. (2015). Salmonella enterica Infections in the United States and Assessment of Coefficients of Variation: A Novel Approach to Identify Epidemiologic

Characteristics of Individual Serotypes, 1996-2011. *PLoS One, 10*(12), e0145416-e0145416. https://doi.org/10.1371/journal.pone.0145416

Bossi, L., Fuentes, J. A., Mora, G., & Figueroa-Bossi, N. (2003). Prophage contribution to bacterial population dynamics. *Journal of Bacteriology, 185*(21), 6467-6471. https://doi.org/10.1128/JB.185.21.6467-6471.2003

Brenner, F. W., Villar, R. G., Angulo, F. J., Tauxe, R., & Swaminathan, B. (2000). Salmonella nomenclature. *Journal of Clinical Microbiology, 38*(7), 2465-2467. https://doi.org/10.1128/JCM.38.7.2465-2467.2000

Brown, E., Dessai, U., McGarry, S., & Gerner-Smidt, P. (2019). Use of Whole-Genome Sequencing for Food Safety and Public Health in the United States. *Foodborne pathogens and disease, 16*(7), 441-450. https://doi.org/10.1089/fpd.2019.2662

Bull, R. A., Adikari, T. N., Ferguson, J. M., Hammond, J. M., Stevanovski, I., Beukers, A. G., Naing, Z., Yeang, M., Verich, A., Gamaarachchi, H., Kim, K. W., Luciani, F., Stelzer-Braid, S., Eden, J.-S., Rawlinson, W. D., van Hal, S. J., & Deveson, I. W. (2020). Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nature Communications, 11*(1), 6272. https://doi.org/10.1038/s41467-020-20075-6

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics, 10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Carattoli, A., & Hasman, H. (2020). PlasmidFinder and In Silico pMLST: Identification and Typing of Plasmid Replicons in Whole-Genome Sequencing (WGS). *Methods Mol Biol, 2075*, 285-294. https://doi.org/10.1007/978-1-4939-9877-7_20

CDC. (2008). *Multistate Outbreak of Salmonella Litchfield Infections Linked to Cantaloupes (FINAL UPDATE)*. https://www.cdc.gov/salmonella/2008/cantaloupes-4-2-2008.html

CDC. (2016). *Multistate Outbreak of Salmonella Infections Linked to Alfalfa Sprouts from One Contaminated Seed Lot (Final Update)*. https://www.cdc.gov/salmonella/muenchen-02-16

CDC. (2019). *Antibiotic Resistance Threats in the United States*.

CDC. (2020). *Outbreak of Salmonella Newport Infections Linked to Onions*. https://www.cdc.gov/salmonella/newport-07-20

Chai, S. J., White, P. L., Lathrop, S. L., Solghan, S. M., Medus, C., McGlinchey, B. M., Tobin-D'Angelo, M., Marcus, R., & Mahon, B. E. (2012). Salmonella enterica serotype Enteritidis: increasing incidence of domestically acquired infections. *Clin Infect Dis, 54 Suppl 5*, S488-497. https://doi.org/10.1093/cid/cis231

Chandak, S., Neu, J., Tatwawadi, K., Mardia, J., Lau, B., Kubit, M., Hulett, R., Griffin, P., Wootters, M., Weissman, T., & Ji, H. (2020, 4-8 May 2020). Overcoming High Nanopore Basecaller Error Rates for DNA Storage via Basecaller-Decoder Integration and Convolutional Codes. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

Chapple, W., Martell, J., Wilson, J. S., & Matsuura, D. T. (2017). A Case Report of Salmonella muenchen Enteritis Causing Rhabdomyolysis and Myocarditis in a Previously Healthy 26-Year-Old Man. *Hawai'i journal of medicine & public health : a journal of Asia Pacific Medicine & Public Health, 76*(4), 106-109. https://pubmed.ncbi.nlm.nih.gov/28428924

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5395706/

Chattaway, M. A., Dallman, T. J., Larkin, L., Nair, S., McCormick, J., Mikhail, A., Hartman, H., Godbole, G., Powell, D., Day, M., Smith, R., & Grant, K. (2019). The Transformation of Reference Microbiology Methods and Surveillance for Salmonella With the Use of Whole Genome Sequencing in England and Wales [Original Research]. *Frontiers in Public Health, 7*(317). https://doi.org/10.3389/fpubh.2019.00317

Chattaway, M. A., Langridge, G. C., & Wain, J. (2021). Salmonella nomenclature in the genomic era: a time for change. *Scientific Reports, 11*(1), 7494. https://doi.org/10.1038/s41598-021-86243-w

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., & Jin, Q. (2005). VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res, 33*(Database issue), D325-328. https://doi.org/10.1093/nar/gki008

Chen, Z., Erickson, D. L., & Meng, J. (2020). 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

Chen, Z., Kuang, D., Xu, X., González-Escalona, N., Erickson, D. L., Brown, E., & Meng, J. (2020). Genomic analyses of multidrug-resistant Salmonella Indiana, Typhimurium, and Enteritidis isolates using MinION and MiSeq sequencing

technologies. *PLoS One, 15*(7), e0235641. https://doi.org/10.1371/journal.pone.0235641

Cooper, A. L., Low, A. J., Koziol, A. G., Thomas, M. C., Leclair, D., Tamber, S., Wong, A., Blais, B. W., & Carrillo, C. D. (2020). Systematic Evaluation of Whole Genome Sequence-Based Predictions of Salmonella Serotype and Antimicrobial Resistance [Original Research]. *Frontiers in Microbiology, 11*(549). https://doi.org/10.3389/fmicb.2020.00549

Crump, J. A., & Wain, J. (2017). Salmonella. In S. R. Quah (Ed.), *International Encyclopedia of Public Health (Second Edition)* (pp. 425-433). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-803678-5.00394-5

Cui, L., Wang, X., Huang, D., Zhao, Y., Feng, J., Lu, Q., Pu, Q., Wang, Y., Cheng, G., Wu, M., & Dai, M. (2020). CRISPR-cas3 of Salmonella Upregulates Bacterial Biofilm Formation and Virulence to Host Cells by Targeting Quorum-Sensing Systems. *Pathogens, 9*(1). https://doi.org/10.3390/pathogens9010053

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience, 10*(2). https://doi.org/10.1093/gigascience/giab008

Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One, 8*(12), e85024. https://doi.org/10.1371/journal.pone.0085024

Delgado-Suárez, E. J., Selem-Mojica, N., Ortiz-López, R., Gebreyes, W. A., Allard, M. W., Barona-Gómez, F., & Rubio-Lozano, M. S. (2018). Whole genome sequencing reveals widespread distribution of typhoidal toxin genes and VirB/D4 plasmids in bovine-associated nontyphoidal Salmonella. *Scientific Reports, 8*(1), 9864. https://doi.org/10.1038/s41598-018-28169-4

Desin, T. S., Köster, W., & Potter, A. A. (2013). Salmonella vaccines in poultry: past, present and future. *Expert Rev Vaccines, 12*(1), 87-96. https://doi.org/10.1586/erv.12.138

Diaz-Proano, C. (2019). *Prevalence, molecular characterization and inactivation of foodborne pathogens on native pecans* Oklahoma State University].

Diep, B., Barretto, C., Portmann, A.-C., Fournier, C., Karczmarek, A., Voets, G., Li, S., Deng, X., & Klijn, A. (2019). Salmonella Serotyping; Comparison of the

Traditional Method to a Microarray-Based Method and an in silico Platform Using Whole Genome Sequencing Data [Methods]. *Frontiers in Microbiology10*(2554). https://doi.org/10.3389/fmicb.2019.02554

Dórea, F. C., Cole, D. J., Hofacre, C., Zamperini, K., Mathis, D., Doyle, M. P., Lee, M. D., & Maurer, J. J. (2010). Effect of Salmonella vaccination of breeder chickens on contamination of broiler chicken carcasses in integrated poultry operations. *Applied and Environmental Microbiology, 76*(23), 7820-7825. https://doi.org/10.1128/AEM.01320-10

Ellington, M. J., Ekelund, O., Aarestrup, F. M., Canton, R., Doumith, M., Giske, C., Grundman, H., Hasman, H., Holden, M. T. G., Hopkins, K. L., Iredell, J., Kahlmeter, G., Köser, C. U., MacGowan, A., Mevius, D., Mulvey, M., Naas, T., Peto, T., Rolain, J. M., Samuelsen, Ø., & Woodford, N. (2017). The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. *Clin Microbiol Infect, 23*(1), 2-22. https://doi.org/10.1016/j.cmi.2016.11.012

Eng, S.-K., Pusparajah, P., Ab Mutalib, N.-S., Ser, H.-L., Chan, K.-G., & Lee, L.-H. (2015). Salmonella: A review on pathogenesis, epidemiology and antibiotic resistance. *Frontiers in Life Science, 8*(3), 284-293. https://doi.org/10.1080/21553769.2015.1051243

Fachada, N., Rodrigues, J., Lopes, V., & Martins, R. (2016). micompr: An R Package for Multivariate Independent Comparison of Observations. *The R Journal, 8*(2), 405-420.

Frenzen, P., Riggs, T., Buzby, J., Breuer, T., Roberts, T., Voetsch, D., & Reddy, S. (1999). Salmonella Cost Estimate Updated Using FoodNet Data.

Galié, S., García-Gutiérrez, C., Miguélez, E. M., Villar, C. J., & Lombó, F. (2018). Biofilms in the Food Industry: Health Aspects and Control Methods [Review]. *Frontiers in Microbiology, 9*(898). https://doi.org/10.3389/fmicb.2018.00898

Gardner, S. N., Slezak, T., & Hall, B. G. (2015). kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics, 31*(17), 2877-2878. https://doi.org/10.1093/bioinformatics/btv271

Gast, R. K., Guard-Bouldin, J., & Holt, P. S. (2004). Colonization of Reproductive Organs and Internal Contamination of Eggs After Experimental Infection of Laying Hens with Salmonella heidelberg

and Salmonella enteritidis. *Avian Diseases, 48*(4), 863-869, 867. https://doi.org/10.1637/7204-05050R

Goldstein, S., Beka, L., Graf, J., & Klassen, J. L. (2019). 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

Gulick, A. M. (2017). Nonribosomal peptide synthetase biosynthetic clusters of ESKAPE pathogens. *Natural product reports, 34*(8), 981-1009. https://doi.org/10.1039/c7np00029d

Gupta, S. K., Sharma, P., McMillan, E. A., Jackson, C. R., Hiott, L. M., Woodley, T., Humayoun, S. B., Barrett, J. B., Frye, J. G., & McClelland, M. (2019). Genomic comparison of diverse Salmonella serovars isolated from swine. *PLoS One, 14*(11), e0224518-e0224518. https://doi.org/10.1371/journal.pone.0224518

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics, 29*(8), 1072-1075. https://doi.org/10.1093/bioinformatics/btt086

Hayek, N. (2013). Lateral transfer and GC content of bacterial resistant genes. *Frontiers in Microbiology, 4*, 41-41. https://doi.org/10.3389/fmicb.2013.00041

Hendriksen, R. S., Bortolaia, V., Tate, H., Tyson, G. H., Aarestrup, F. M., & McDermott, P. F. (2019). Using Genomics to Track Global Antimicrobial Resistance [Review]. *Frontiers in Public Health, 7*(242). https://doi.org/10.3389/fpubh.2019.00242

Herrera-León, S., Ramiro, R., Arroyo, M., Díez, R., Usera, M. A., & Echeita, M. A. (2007). Blind comparison of traditional serotyping with three multiplex PCRs for the identification of Salmonella serotypes. *Research in Microbiology, 158*(2), 122-127. https://doi.org/https://doi.org/10.1016/j.resmic.2006.09.009

Hoffmann, M., Luo, Y., Monday, S. R., Gonzalez-Escalona, N., Ottesen, A. R., Muruvanda, T., Wang, C., Kastanis, G., Keys, C., Janies, D., Senturk, I. F., Catalyurek, U. V., Wang, H., Hammack, T. S., Wolfgang, W. J., Schoonmaker-Bopp, D., Chu, A., Myers, R., Haendiges, J., Evans, P. S., Meng, J., Strain, E. A., Allard, M. W., & Brown, E. W. (2016). Tracing Origins of the Salmonella Bareilly Strain Causing a Food-borne Outbreak in the United States. *The Journal of Infectious Diseases, 213*(4), 502-508. https://doi.org/10.1093/infdis/jiv297

Ibrahim, G. M., & Morin, P. M. (2018). Salmonella Serotyping Using Whole Genome Sequencing [Original Research]. *Frontiers in Microbiology, 9*(2993). https://doi.org/10.3389/fmicb.2018.02993

Jombart, T., Kendall, M., Almagro-Garcia, J., & Colijn, C. (2017). treespace: Statistical exploration of landscapes of phylogenetic trees. *Mol Ecol Resour, 17*(6), 1385-1392. https://doi.org/10.1111/1755-0998.12676

Karlsson, E., Lärkeryd, A., Sjödin, A., Forsman, M., & Stenberg, P. (2015). Scaffolding of a bacterial genome using MinION nanopore sequencing. *Scientific Reports, 5*, 11996-11996. https://doi.org/10.1038/srep11996

Katz, L. S., Griswold, T., Williams-Newkirk, A. J., Wagner, D., Petkau, A., Sieffert, C., Van Domselaar, G., Deng, X., & Carleton, H. A. (2017). A Comparative Analysis of the Lyve-SET Phylogenomics Pipeline for Genomic Epidemiology of Foodborne Pathogens [Methods]. *Frontiers in Microbiology, 8*(375). https://doi.org/10.3389/fmicb.2017.00375

Kawasaki, S., Horikoshi, N., Okada, Y., Takeshita, K., Sameshima, T., & Kawamoto, S. (2005). Multiplex PCR for simultaneous detection of Salmonella spp., Listeria monocytogenes, and Escherichia coli O157:H7 in meat samples. *J Food Prot, 68*(3), 551-556. https://doi.org/10.4315/0362-028x-68.3.551

Kendall, M., & Colijn, C. (2016). Mapping Phylogenetic Trees to Reveal Distinct Patterns of Evolution. *Molecular Biology and Evolution, 33*(10), 2735-2743. https://doi.org/10.1093/molbev/msw124

Kim, S., Ryu, K., Biswas, D., & Ahn, J. (2014). Survival, prophage induction, and invasive properties of lysogenic Salmonella Typhimurium exposed to simulated gastrointestinal conditions. *Arch Microbiol, 196*(9), 655-659. https://doi.org/10.1007/s00203-014-1005-z

Kingry, L. C., Rowe, L. A., Respicio-Kingry, L. B., Beard, C. B., Schriefer, M. E., & Petersen, J. M. (2016). Whole genome multilocus sequence typing as an epidemiologic tool for Yersinia pestis. *Diagnostic Microbiology and Infectious Disease, 84*(4), 275-280. https://doi.org/10.1016/j.diagmicrobio.2015.12.003

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology, 37*(5), 540-546. https://doi.org/10.1038/s41587-019-0072-8

Köser, C. U., Ellington, M. J., & Peacock, S. J. (2014). Whole-genome sequencing to control antimicrobial resistance. *Trends Genet, 30*(9), 401-407. https://doi.org/10.1016/j.tig.2014.07.003

Kurtz, J. R., Goggins, J. A., & McLachlan, J. B. (2017). Salmonella infection: Interplay between the bacteria and host immune system. *Immunology letters, 190*, 42-50. https://doi.org/10.1016/j.imlet.2017.07.006

Kushwaha, S. K., Bhavesh, N. L. S., Abdella, B., Lahiri, C., & Marathe, S. A. (2020). The phylogenomics of CRISPR-Cas system and revelation of its features in Salmonella. *Scientific Reports, 10*(1), 21156. https://doi.org/10.1038/s41598-020-77890-6

Lee, I., Ouk Kim, Y., Park, S. C., & Chun, J. (2016). OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *Int J Syst Evol Microbiol, 66*(2), 1100-1103. https://doi.org/10.1099/ijsem.0.000760

Leidenfrost, R. M., Pöther, D.-C., Jäckel, U., & Wünschiers, R. (2020). Benchmarking the MinION: Evaluating long reads for microbial profiling. *Scientific Reports, 10*(1), 5125. https://doi.org/10.1038/s41598-020-61989-x

Lerminiaux, N. A., MacKenzie, K. D., & Cameron, A. D. S. (2020). Salmonella Pathogenicity Island 1 (SPI-1): The Evolution and Stabilization of a Core Genomic Type Three Secretion System. *Microorganisms, 8*(4). https://doi.org/10.3390/microorganisms8040576

Lindsey, R. L., Pouseele, H., Chen, J. C., Strockbine, N. A., & Carleton, H. A. (2016). Implementation of Whole Genome Sequencing (WGS) for Identification and Characterization of Shiga Toxin-Producing Escherichia coli (STEC) in the United States [Original Research]. *Frontiers in Microbiology*, *7*(766). https://doi.org/10.3389/fmicb.2016.00766

Lou, L., Zhang, P., Piao, R., & Wang, Y. (2019). Salmonella Pathogenicity Island 1 (SPI-1) and Its Complex Regulatory Network [Review]. *Frontiers in Cellular and Infection Microbiology, 9*(270). https://doi.org/10.3389/fcimb.2019.00270

Lou, L., Zhang, P., Piao, R., & Wang, Y. (2019). Salmonella Pathogenicity Island 1 (SPI-1) and Its Complex Regulatory Network. *Front Cell Infect Microbiol, 9*, 270. https://doi.org/10.3389/fcimb.2019.00270

Lyu, N., Feng, Y., Pan, Y., Huang, H., Liu, Y., Xue, C., Zhu, B., & Hu, Y. (2021). Genomic Characterization of Salmonella enterica Isolates From Retail Meat in

Beijing, China [Original Research]. *Frontiers in Microbiology, 12*(784). https://doi.org/10.3389/fmicb.2021.636332

Maghini, D. G., Moss, E. L., Vance, S. E., & Bhatt, A. S. (2021). Improved high-molecular-weight DNA extraction, nanopore sequencing and metagenomic assembly from the human gut microbiome. *Nature Protocols, 16*(1), 458-471. https://doi.org/10.1038/s41596-020-00424-x

Mohammed, M., & Thapa, S. (2020). Evaluation of WGS-subtyping methods for epidemiological surveillance of foodborne salmonellosis. *One Health Outlook, 2*(1), 13. https://doi.org/10.1186/s42522-020-00016-5

Nair, D., Venkitanarayanan, K., & Kollanoor Johny, A. (2018). Antibiotic-Resistant Salmonella in the Food Supply and the Potential Role of Antibiotic Alternatives for Control. *Foods (Basel, Switzerland), 7*(10), 167. https://doi.org/10.3390/foods7100167

Padilha, V. A., Alkhnbashi, O. S., Shah, S. A., de Carvalho, A. C. P. L. F., & Backofen, R. (2020). CRISPRcasIdentifier: Machine learning for accurate identification and classification of CRISPR-Cas systems. *Gigascience, 9*(6). https://doi.org/10.1093/gigascience/giaa062

Papanikolaou, N., Trachana, K., Theodosiou, T., Promponas, V. J., & Iliopoulos, I. (2009). Gene socialization: gene order, GC content and gene silencing in Salmonella. *BMC Genomics, 10*, 597-597. https://doi.org/10.1186/1471-2164-10-597

Paradis, E. (2016). The distribution of branch lengths in phylogenetic trees. *Molecular Phylogenetics and Evolution, 94*, 136-145. https://doi.org/https://doi.org/10.1016/j.ympev.2015.08.010

Pearman, W. S., Freed, N. E., & Silander, O. K. (2020). Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads. *BMC Bioinformatics, 21*(1), 220-220. https://doi.org/10.1186/s12859-020-3528-4

Perrière, G., & Gouy, M. (1996). WWW-query: An on-line retrieval system for biological sequence banks. *Biochimie, 78*(5), 364-369. https://doi.org/https://doi.org/10.1016/0300-9084(96)84768-7

Pornsukarom, S., van Vliet, A. H. M., & Thakur, S. (2018). Whole genome sequencing analysis of multiple Salmonella serovars provides insights into phylogenetic relatedness, antimicrobial resistance, and virulence markers across humans, food

animals and agriculture environmental sources. *BMC Genomics, 19*(1), 801-801. https://doi.org/10.1186/s12864-018-5137-4

Rang, F. J., Kloosterman, W. P., & de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome biology, 19*(1), 90. https://doi.org/10.1186/s13059-018-1462-9

Reis-Cunha, J. L., Bartholomeu, D. C., Manson, A. L., Earl, A. M., & Cerqueira, G. C. (2019). ProphET, prophage estimation tool: A stand-alone prophage sequence prediction tool with self-updating reference database. *PLoS One, 14*(10), e0223364. https://doi.org/10.1371/journal.pone.0223364

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution, 3*(2), 217-223. https://doi.org/https://doi.org/10.1111/j.2041-210X.2011.00169.x

Roer, L., Hendriksen, R. S., Leekitcharoenphon, P., Lukjancenko, O., Kaas, R. S., Hasman, H., Aarestrup, F. M., & Eisen, J. (2016). Is the Evolution of Salmonella enterica subsp. enterica Linked to Restriction-Modification Systems? *mSystems, 1*(3), e00009-00016. https://doi.org/doi:10.1128/mSystems.00009-16

Rounds, J. M., Taylor, A. J., Eikmeier, D., Nichols, M. M., Lappi, V., Wirth, S. E., Boxrud, D. J., Smith, K. E., & Medus, C. (2020). Prospective Salmonella Enteritidis surveillance and outbreak detection using whole genome sequencing, Minnesota 2015-2017. *Epidemiology and infection, 148*, e254-e254. https://doi.org/10.1017/S0950268820001272

Scallan, E., Hoekstra, R. M., Angulo, F. J., Tauxe, R. V., Widdowson, M. A., Roy, S. L., Jones, J. L., & Griffin, P. M. (2011). Foodborne illness acquired in the United States--major pathogens. *Emerging infectious diseases, 17*(1), 7-15. https://doi.org/10.3201/eid1701.p11101

Schürch, A. C., Arredondo-Alonso, S., Willems, R. J. L., & Goering, R. V. (2018). Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. *Clinical Microbiology and Infection, 24*(4), 350-354. https://doi.org/https://doi.org/10.1016/j.cmi.2017.12.016

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics, 30*(14), 2068-2069. https://doi.org/10.1093/bioinformatics/btu153

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. In M. Kollmar (Ed.), *Gene Prediction:*

Methods and Protocols (pp. 227-245). Springer New York. https://doi.org/10.1007/978-1-4939-9173-0_14

Sher, A. A., Mustafa, B. E., Grady, S. C., Gardiner, J. C., & Saeed, A. M. (2021). Outbreaks of foodborne Salmonella enteritidis in the United States between 1990 and 2015: An analysis of epidemiological and spatial-temporal trends. *International Journal of Infectious Diseases, 105*, 54-61. https://doi.org/10.1016/j.ijid.2021.02.022

Souza, A. I. S., Saraiva, M. M. S., Casas, M. R. T., Oliveira, G. M., Cardozo, M. V., Benevides, V. P., Barbosa, F. O., Freitas Neto, O. C., Almeida, A. M., & Berchieri, A. J. (2020). High occurrence of β-lactamase-producing Salmonella Heidelberg from poultry origin. *PLoS One, 15*(3), e0230676. https://doi.org/10.1371/journal.pone.0230676

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30*(9), 1312-1313. https://doi.org/10.1093/bioinformatics/btu033

Stoler, N., & Nekrutenko, A. (2021). Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics, 3*(1). https://doi.org/10.1093/nargab/lqab019

Suez, J., Porwollik, S., Dagan, A., Marzel, A., Schorr, Y. I., Desai, P. T., Agmon, V., McClelland, M., Rahav, G., & Gal-Mor, O. (2013). Virulence gene profiling and pathogenicity characterization of non-typhoidal Salmonella accounted for invasive disease in humans. *PLoS One, 8*(3), e58449. https://doi.org/10.1371/journal.pone.0058449

Taylor, T. L., Volkening, J. D., DeJesus, E., Simmons, M., Dimitrov, K. M., Tillman, G. E., Suarez, D. L., & Afonso, C. L. (2019). Rapid, multiplexed, whole genome and plasmid sequencing of foodborne pathogens using long-read nanopore technology. *Scientific Reports, 9*(1), 16350. https://doi.org/10.1038/s41598-019-52424-x

Thrash, A., Hoffmann, F., & Perkins, A. (2020). Toward a more holistic method of genome assembly assessment. *BMC Bioinformatics, 21*(4), 249. https://doi.org/10.1186/s12859-020-3382-4

Timme, R. E., Sanchez Leon, M., & Allard, M. W. (2019). Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback. *Methods Mol Biol, 1918*, 201-212. https://doi.org/10.1007/978-1-4939-9000-9_17

Tolar, B., Joseph, L. A., Schroeder, M. N., Stroika, S., Ribot, E. M., Hise, K. B., & Gerner-Smidt, P. (2019). An Overview of PulseNet USA Databases. *Foodborne pathogens and disease, 16*(7), 457-462. https://doi.org/10.1089/fpd.2019.2637

Tyler, A. D., Mataseje, L., Urfano, C. J., Schmidt, L., Antonation, K. S., Mulvey, M. R., & Corbett, C. R. (2018). Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. *Scientific Reports, 8*(1), 10931. https://doi.org/10.1038/s41598-018-29334-5

Wahl, A., Battesti, A., & Ansaldi, M. (2019). Prophages in Salmonella enterica: a driving force in reshaping the genome and physiology of their bacterial host? *Molecular Microbiology, 111*(2), 303-316. https://doi.org/10.1111/mmi.14167

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS One, 9*(11), e112963. https://doi.org/10.1371/journal.pone.0112963

Wang, H., Ding, S., Wang, G., Xu, X., & Zhou, G. (2013). In situ characterization and analysis of Salmonella biofilm formation under meat processing environments using a combined microscopic and spectroscopic approach. *International Journal of Food Microbiology, 167*(3), 293-302. https://doi.org/https://doi.org/10.1016/j.ijfoodmicro.2013.10.005

Wattiau, P., Weijers, T., Andreoli, P., Schliker, C., Veken, H. V., Maas, H. M. E., Verbruggen, A. J., Heck, M. E. O. C., Wannet, W. J., Imberechts, H., & Vos, P. (2008). Evaluation of the Premi®Test Salmonella, a commercial low-density DNA microarray system intended for routine identification and typing of Salmonella enterica. *International Journal of Food Microbiology, 123*(3), 293-298. https://doi.org/https://doi.org/10.1016/j.ijfoodmicro.2008.01.006

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017a). Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics, 3*(10). https://doi.org/https://doi.org/10.1099/mgen.0.000132

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017b). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology, 13*(6), e1005595. https://doi.org/10.1371/journal.pcbi.1005595

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology, 15*(3), R46-R46. https://doi.org/10.1186/gb-2014-15-3-r46

Wright, A. P., Richardson, L., Mahon, B. E., Rothenberg, R., & Cole, D. J. (2016). The rise and decline in Salmonella enterica serovar Enteritidis outbreaks attributed to egg-containing foods in the United States, 1973-2009. *Epidemiol Infect, 144*(4), 810-819. https://doi.org/10.1017/s0950268815001867

Xu, F., Ge, C., Luo, H., Li, S., Wiedmann, M., Deng, X., Zhang, G., Stevenson, A., Baker, R. C., & Tang, S. (2020). Evaluation of real-time nanopore sequencing for Salmonella serotype prediction. *Food Microbiology, 89*, 103452. https://doi.org/https://doi.org/10.1016/j.fm.2020.103452

Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell, 178*(4), 779-794. https://doi.org/https://doi.org/10.1016/j.cell.2019.07.010

Yeom, J., Shao, Y., & Groisman, E. A. (2020). Small proteins regulate Salmonella survival inside macrophages by controlling degradation of a magnesium transporter. *Proceedings of the National Academy of Sciences, 117*(33), 20235. https://doi.org/10.1073/pnas.2006116117

Yoon, S. H., Ha, S. M., Lim, J., Kwon, S., & Chun, J. (2017). A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie Van Leeuwenhoek, 110*(10), 1281-1286. https://doi.org/10.1007/s10482-017-0844-4

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M., & Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *The Journal of antimicrobial chemotherapy, 67*(11), 2640-2644. https://doi.org/10.1093/jac/dks261

Zhang, S., den Bakker Hendrik, C., Li, S., Chen, J., Dinsmore Blake, A., Lane, C., Lauer, A. C., Fields Patricia, I., Deng, X., & Dudley Edward, G. (2019). SeqSero2: Rapid and Improved Salmonella Serotype Determination Using Whole-Genome Sequencing Data. *Applied and Environmental Microbiology, 85*(23), e01746-01719. https://doi.org/10.1128/AEM.01746-19

Zhao, S., Blickenstaff, K., Glenn, A., Ayers, S. L., Friedman, S. L., Abbott, J. W., & McDermott, P. F. (2009). β-Lactam Resistance in <i>Salmonella</i> Strains Isolated from Retail Meats in the United States by the National Antimicrobial Resistance Monitoring System between 2002 and 2006. *Applied*

and Environmental Microbiology, 75(24), 7624-7630. https://doi.org/doi:10.1128/AEM.01158-09

Zitko, V. (1994). Principal component analysis in the evaluation of environmental data. *Marine Pollution Bulletin, 28*(12), 718-722. https://doi.org/https://doi.org/10.1016/0025-326X(94)90329-8

APPENDICES

Appendix 1. Flowcells ID used for the sequencing of *E. coli* isolates in MinION and DNA purity values.

Isolate	Flowcell ID	260/280	260/230	DNA Concentration (ng/µl)
G1BLF1_5	FAK59422	1.92	1.3	25
G1BLF2_1	FAK63170	1.95	1.61	25
G1M0S3_4	FAK59422	1.9	1.44	25
G1M4F3_31	FAK59422	2.13	1.44	25
G4M0F1_1	FAK59422	2.12	1.96	25
G4M0F2_14	FAK63170	1.76	1.19	25
G5BLF1_1	FAK63170	1.8	1.33	25
G5BLF3_3	FAK35917	1.94	1.63	25
G5BLF3_8	FAK63170	1.76	1.12	25
G5M2P3_1	FAK63170	1.78	1.06	25
G5M4F2_1	FAK35917	1.85	1.16	25

Apendix 2. Metrics considered for the selection of the best hybrid assembly for *E. coli* isolates.

	N50		Number of contigs		Length of the largest contig	
Isolate	**Pilon**	**Unicycler**	**Pilon**	**Unicycler**	**Pilon**	**Unicycler**
G1BLF1_5	376405	518023	33	43	1085267	1347791
G1BLF2_1	5314189	5314081	5	6	5314189	5314081
G1M0S3_4	1349801	4359963	10	25	2489441	4359963
G1M4F3_31	97204	384227	70	51	391573	854945
G4M0F1_1	2558830	2502980	7	14	2558830	2502980
G4M0F2_14	4602735	751343	2	32	4602735	994287
G5BLF1_1	5447391	5399332	3	8	5447391	5399332
G5BLF3_3	4983366	4879966	9	9	4983366	4879966
G5BLF3_8	2904903	2886145	9	15	2904903	2886145
G5M2P3_1	4685726	2789679	2	15	4685726	2789679
G5M4F2_1	1420559	4063509	13	38	1546743	4063509

Appendix 3. Flowcells ID used for the sequencing of *Salmonella* isolates in MinION and DNA purity values.

***Salmonella* isolates**	**Flowcell ID**	**260/280**	**260/230**	**Concentration (ng/µl)**
G1BLS3_3	FAK35809	1.85	2.16	25
G2BLF1_3	FAK35809	2.04	1.85	25
G2M0S1_2	FAK35809	1.96	2.02	25
G2M4F3_3	FAK35809	2.02	2.23	25
G4BLF1_2	FAK35809	2.00	2.02	25
G4BLS2_11	FAK35809	1.94	2.02	25
G5_20BLS1_1	FAK35809	2.02	2.01	25
G5_25BLS1_1	FAK35809	1.96	1.63	25
HNG1S2_2	FAK35809	1.87	1.36	25
HNG1S2_3	FAK35809	1.69	1.06	25

Appendix 4. Metrics considered for the selection of the best hybrid assemblies for *Salmonella* isolates.

	N50		**Number of contigs**		**Length of the largest contig**	
Isolate	**Pilon**	**Unicycler**	**Pilon**	**Unicycler**	**Pilon**	**Unicycler**
G1BLS3_3	4808067	4807632	3	3	4808067	4807632
G2BLF1_3	4722067	4722110	1	1	4722067	4722110
G2M0S1_2	4837849	4686920	2	3	4837849	4686920
G2M4F3_3	4722253	4722116	1	1	4722253	4722116
G4BLF1_2	4784938	4784827	1	1	4784938	4784827
G4BLS2_11	1391112	4823151	4	2	2055110	4823151
G5_20BLS1_1	4699744	4699412	1	1	4699744	4699412
G5_25BLS1_1	1188584	4739838	4	2	1638776	4739838
HNG1S2_2	4675899	4675794	1	1	4675899	4675794
HNG1S2_3	4675843	4675824	1	1	4675843	4675824

VITA

Nicolas Javier Lopez Guerra

Candidate for the Degree of

Master of Science

Thesis: EVALUATION OF IDENTIFICATION AND MOLECULAR CHARACTERIZATION OF FOODBORNE PATHOGENS BY WHOLE GENOME SEQUENCING THROUGH ILLUMINA AND NANOPORE

Major Field: Food Science

Biographical:

Education:

Completed the requirements for the Master of Science in Food Science at Oklahoma State University, Stillwater, Oklahoma in July, 2021

Completed the requirements for the Bachelor of Science in Biotechnology Engineering at Army Forces University-ESPE, Sangolqui, Ecuador in 2019.

Experience:

Graduate Research Assistant in the Institute of Biosecurity and Microbial Forensics, Oklahoma State University, from August 2019 to July 2021.

Instructor and organizer of theoretical-practical training courses in the area of Biology and Biotechnology, from July 2018 to July 2019

Professional Memberships:

International Association for Food Protection 2020- Present