Optimizing Quasi-metagenomic Sequencing for Strain-Level Detection and Genomic Surveillance of Foodborne Pathogens in Complex Food Matrices

by

Jillian L. Ryz

A Thesis submitted to the Faculty of Graduate and Postdoctoral Studies of

The University of Manitoba

in partial fulfilment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

Department of Medical Microbiology and Infectious Diseases

University of Manitoba

Winnipeg

Copyright © 2026 by Jillian L. Ryz

ACKNOWLEDGMENTS

I would like to sincerely thank my supervisors, Dr. Natalie Knox and Dr. Celine Nadon, for their guidance, mentorship, and support throughout this research. Their expertise, thoughtful feedback, and consistent encouragement played an important role in shaping both this work and my growth as a scientist. As accomplished women in science, they have also been incredibly inspiring role models. Their leadership, resilience, and dedication to advancing public health have made a lasting impression on me. I am truly grateful for the opportunity to learn from them and for the confidence they placed in me throughout this journey.

I am also grateful to my advisory committee members, Dr. Morag Graham, Dr. Lisa Lix, and Dr. Grant McClarty, for their guidance and thoughtful feedback throughout this process. Their diverse perspectives and insightful questions helped strengthen this research. I greatly appreciated the constructive and supportive nature of our discussions, which made the committee process a productive, encouraging and positive experience. Their expertise was truly invaluable to this work.

I would also like to thank my colleagues for their willingness to share ideas, troubleshoot challenges, and offer practical support along the way. Their collaboration and generosity with their time and knowledge made many of the day-to-day challenges of this work easier to navigate and helped create a positive and enjoyable working environment.

I extend my sincere appreciation to the PulseNet Canada Network for providing the bacterial strains used in this research and for their continued commitment to advancing foodborne disease surveillance. I am also thankful to the many individuals who helped shape the broader research project and secure the funding that made this work possible. Their efforts laid the foundation on which this thesis was built.

Most importantly, I thank my family for their patience, encouragement, and constant support throughout this journey. Their flexibility and belief in me made it possible to balance the many demands of this work and to keep moving forward when things felt difficult. I am especially grateful to my children, who inspire me every day to be the best person I can be and remind me to keep going, even when things are difficult. I could not have completed this thesis without the support of the many people who stood beside me along the way.

DEDICATION

To my children

May you always believe in yourselves,

and remember that your potential is limitless.

ABSTRACT

Foodborne disease surveillance relies heavily on whole-genome sequencing (WGS) of bacterial isolates to support outbreak detection and source attribution. However, culture-based workflows introduce delays that can limit timely public health response, particularly for perishable foods and complex matrices. This thesis evaluates quasi-metagenomics, defined as sequencing enriched food matrices without secondary isolate recovery, as a culture-independent approach for strain-level pathogen detection and characterization. The objective was to determine whether quasi-metagenomic sequencing can provide strain-resolved genomic information suitable for foodborne outbreak detection. This thesis therefore investigates the laboratory and analytical requirements needed to achieve outbreak-level genomic resolution directly from enriched food samples.

Chapter 3 establishes the bioinformatic foundation by benchmarking taxonomic classifiers, reference databases, and analytical parameters. Results demonstrated that a curated foodborne pathogen (FBP)-specific reference database used with KrakenUniq provided the most accurate and consistent taxonomic classification while reducing computational burden. Chapter 4 applies this optimized workflow to short-read quasi-metagenomic datasets generated from enriched food matrices artificially contaminated with *Listeria monocytogenes*, verotoxigenic *Escherichia coli* (VTEC), and *Shigella* spp. Selective enrichment had a greater influence on pathogen recovery than eukaryotic DNA depletion. When sufficient pathogen DNA was recovered, quasi-metagenomic sequencing enabled accurate pathogen detection, serotype prediction, genome assembly, and whole-genome multilocus sequence typing (wgMLST)-based outbreak characterization with genomic distance ranges comparable to isolate-derived WGS. To support standardized interpretation of quasi-metagenomics, a pathogen-presence likelihood metric was developed that integrates sequencing depth-normalized unique *k*-mer counts (unique KPMR; *k*-mers per million reads) with robust *z*-scores to distinguish true pathogen signal from background noise. Chapter 5 extends this workflow to Oxford Nanopore Technologies (ONT) long-read sequencing. Despite higher raw read error rates, long-read quasi-metagenomic sequencing supported accurate taxonomic classification, genome assembly, and outbreak characterization. Single nucleotide polymorphism (SNP)- and wgMLST-based analyses produced genomic distance ranges comparable to those obtained using short-read quasi-metagenomics and isolate-derived WGS.

Overall, this research establishes an end-to-end quasi-metagenomic workflow for strain-level pathogen detection and outbreak characterization directly from enriched food samples. These findings provide a strong methodological foundation for the continued development of culture-independent genomic approaches to support foodborne disease surveillance and outbreak response.

TABLE OF CONTENTS

ACKNOWLEDGMENTS 2

DEDICATION 3

ABSTRACT 4

LIST OF FIGURES 8

LIST OF TABLES 10

LIST OF APPENDICES 11

LIST OF ABBREVIATIONS 12

STATEMENT OF AUTHOR CONTRIBUTIONS 13

1.0 Chapter 1 - Introduction 16

1.1 Public health and economic burden of foodborne illness 16
1.2 High priority foodborne pathogens in Canada 17
1.3 Current frameworks for foodborne outbreak detection 18
1.4 Culture-based and culture-independent diagnostic approaches in foodborne surveillance 19
1.5 Metagenomics and quasi-metagenomics in foodborne pathogen detection 20
1.6 Challenges in food metagenomic pathogen detection 21
1.7 Long-read sequencing technologies in quasi-metagenomic surveillance 23
1.8 Challenges for long-read quasi-metagenomics 24
1.9 Aims and rationale 25

2.0 Chapter 2 - Methods 27

2.1 Study Datasets 27
2.2 Artificial contamination and enrichment 28
2.3 Eukaryotic DNA depletion 30
2.4 Genomic DNA extraction 30
2.5 Library preparation and sequencing 31
2.6 Reference sequence data 32
2.7 Tool selection 32
2.8 Reference database design and development 32
2.9 Data analysis 33

2.9.1 Data preprocessing and eukaryotic sequence removal 33
2.9.2 Taxonomic classification and pathogen identification 36
2.9.3 Pathogen-presence likelihood metric 38
2.9.4 Genome assembly and quality assessment 40
2.9.5 Pathogen verification 40
2.9.6 Outbreak analysis 40
2.9.7 Genome coverage 41
2.10 Statistical Analysis and Visualizations 41
3.0 Chapter 3 - Optimization of bioinformatic pipelines and reference database architecture for the detection of *Listeria monocytogenes* in complex food matrices 42
3.1 Rationale 42
3.2 Results 43
3.2.1 Provision of test datasets 43
3.2.2 Including an *in silico* eukaryotic DNA removal step can minimize false-positive classifications when using a taxon-specific reference database 44
3.2.3 Metagenomic classifier and analysis parameters influence the number of false-positive classifications when using a taxon-specific reference database 50
3.2.4 Detection of *Listeria monocytogenes* in RTE turkey deli meat and prepackaged spinach is challenging without culture enrichment 52
3.2.5 Fewer computational resources are required for classification using a taxon-specific reference database 53
3.3 Discussion 53
4.0 Chapter 4 - Validation of a quasi-metagenomic workflow for the rapid detection and subtyping of priority foodborne pathogens from enriched food matrices 58
4.1 Rationale 58
4.2 Results 59
4.2.1 Data quality 59
4.2.2 Saponin-based eukaryotic DNA depletion significantly decreases the percentage of eukaryotic sequences 61
4.2.3 Saponin-based eukaryotic DNA depletion does not significantly improve the pathogen signal when the enrichment period is sufficiently optimized 61
4.2.4 Species-specific reference databases for VTEC and *Shigella* species significantly improve pathogen signal 66
4.2.5 Assembly quality is consistent between wet-laboratory protocols 70

4.2.6 Genome coverage showed no statistically significant variation between wet laboratory protocols but was substantially improved with species-specific reference databases for verotoxigenic *Escherichia coli* (VTEC) and *Shigella* datasets. 73
4.2.7 Pathogen detection and characterization 78
4.2.8 Outbreak detection using short-read quasi-metagenomic datasets 88
4.3 Discussion 94
5.0 Chapter 5 - Implementation of long-read nanopore sequencing for rapid quasi-metagenomic pathogen surveillance and outbreak detection of *Listeria monocytogenes* in complex food matrices 98
5.1 Rationale 98
5.2 Results 99
5.2.1 General data quality 99
5.2.2 Data preprocessing 101
5.2.3 Saponin-based eukaryotic DNA depletion 101
5.2.4 RefSeq Masher provides a rapid preliminary screen of pre-processed datasets and can inform the selection of appropriate reference databases for subsequent taxonomic classification 101
5.2.5 Saponin-based depletion does not substantially improve pathogen signal under optimized enrichment conditions 103
5.2.6 Pathogen-presence likelihood metric 105
5.2.7 RefSeq Masher Matches provides dependable species-level verification for *Listeria* species. 108
5.2.8 Assembly quality is consistent between wet-laboratory protocols 108
5.2.9 Genome coverage 108
5.2.10 Quality assessment of long-read data in BioNumerics 111
5.2.11 Outbreak detection using long-read quasi-metagenomic datasets 113
5.3 Discussion 116
6.0 Chapter 6: Conclusion and future directions 119
6.1 Summary of key findings 119
6.2 Strengths and contributions 121
6.3 Limitations 122
6.4 Future directions 124
6.5 Concluding remarks 127
7.0 References 128
8.0 Appendices 146

LIST OF FIGURES

Figure 2.1. Overview of the bioinformatic workflows developed and applied across Chapters 3–5. 34

Figure 3.1. *Listeria monocytogenes* read recovery in artificially contaminated RTE turkey deli meat. 46

Figure 3.2. Average *Listeria monocytogenes* read recovery in artificially contaminated prepackaged spinach. 47

Figure 3.3. Comparison of precision, recall and F1 score across reference databases and classifiers. 49

Figure 3.4. Comparison of computational resource requirements for taxonomic classification. 54

Figure 4.1. Comparative analysis of raw sequencing metrics and DNA concentration 60

Figure 4.2. Impact of saponin-based depletion on eukaryotic DNA in enriched food samples. 62

Figure 4.3. KrakenUniq classification metrics for *Listeria* strains across various food matrices, including single-isolate (SI) and negative controls (NC). 63

Figure 4.4. KrakenUniq classification metrics for verotoxigenic *Escherichia coli* (VTEC) strains across various food matrices, including single-isolate (SI) and negative controls (NC) 65

Figure 4.5. KrakenUniq classification metrics for *Shigella* strains across a single food matrix, including single-isolate (SI) and negative controls (NC). 67

Figure 4.6. Comparative assembly quality metrics for *Listeria* strains in various food matrices 69

Figure 4.7. Comparative assembly quality metrics for verotoxigenic *Escherichia coli* (VTEC) strains in various food matrices 71

Figure 4.8. Comparative assembly quality metrics for *Shigella* strains in a single food matrix. 72

Figure 4.9. Genome coverage metrics for *Listeria* species across varied food matrices and protocols. 74

Figure 4.10. Comparative genome coverage metrics for verotoxigenic *Escherichia coli* (VTEC) across various food matrices and protocols. 75

Figure 4.11. Comparative genome coverage metrics for *Shigella* species across protocols. 77

Figure 4.12. Detection signal and within-sample enrichment in quasi-metagenomic datasets from food matrices spiked with *Listeria* strains 80

Figure 4.13. Detection signal and within-sample enrichment in quasi-metagenomic datasets from food matrices spiked with verotoxigenic *Escherichia coli* (VTEC) strains. 81

Figure 4.14. Detection signal and within-sample enrichment in quasi-metagenomic datasets from food matrices spiked with *Shigella* strains. 82

Figure 4.15. Most likely pathogen assignments across all quasi-metagenomic datasets as determined by the pathogen-presence likelihood metric. 83

Figure 4.16. Comparative assembly performance for *Listeria*-specific sequences extracted from quasi-metagenomic datasets..... 89

Figure 4.17. Comparative assembly performance for verotoxigenic *Escherichia coli* (VTEC)-specific sequences extracted from quasi-metagenomic datasets. 90

Figure 4.18. Comparative assembly performance for *Shigella*-specific sequences extracted from quasi-metagenomic datasets..... 91

Figure 4.19. Single-nucleotide variant (SNV) and whole-genome multilocus sequence typing (wgMLST) allele differences between single-isolate (SI) reference genomes and short-read quasi-metagenomic datasets across pathogen-food matrices 93

Figure 5.1. Comparison of long-read sequencing performance and DNA concentration across quasi-metagenomic Oxford Nanopore MinION datasets. 100

Figure 5.2. Effect of saponin-based depletion on eukaryotic DNA background in quasi-metagenomic Oxford Nanopore MinION datasets. 102

Figure 5.3. KrakenUniq classification metrics for *Listeria* strains in long-read quasi-metagenomic Oxford Nanopore MinION datasets. 104

Figure 5.4. Absolute detection signal and within-sample enrichment in long-read quasi-metagenomic datasets derived from food matrices spiked with *Listeria* strains. 106

Figure 5.5. Most likely pathogen assignments across long-read quasi-metagenomic datasets as determined by the pathogen-presence likelihood metric. 107

Figure 5.6. Comparative assembly quality metrics for *Listeria* strains in long-read quasi-metagenomic Oxford Nanopore MinION datasets..... 109

Figure 5.7. Comparative assembly coverage metrics for *Listeria* strains in long-read quasi-metagenomic Oxford Nanopore MinION datasets..... 110

Figure 5.8. Comparative assembly and subtyping performance for *Listeria* long-read quasi-metagenomic Oxford Nanopore MinION datasets..... 112

Figure 5.9. Whole genome multilocus sequence typing (wgMLST) allele differences between single-isolate references and long-read quasi-metagenomic Oxford Nanopore MinION *L. monocytogenes* datasets across food matrices. 115

LIST OF TABLES

Table 2.1. Summary of computational and analytical software. 35

Table 4.1. *In silico* serotyping performance for verotoxigenic *Escherichia coli* (VTEC) using quasi-metagenomic assemblies. 85

Table 4.2. *In silico* serotyping performance for *Shigella* strains using quasi-metagenomic assemblies..... 86

Table 4.3. RefSeq Masher Matches Results for *Listeria* strains using quasi-metagenomic assemblies. 87

Table 4.4. Comparison of genomic distance metrics for simulated outbreak detection between original outbreak isolates and the quasi-metagenomic datasets. 92

Table 5.1. Comparison of genomic distance metrics for outbreak detection of *Listeria monocytogenes* using long-read quasi-metagenomic Oxford Nanopore MinION datasets. 114

LIST OF APPENDICES

Appendix 2.1. Reference database features. 146

Appendix 2.2. Distribution of sequencing depth across samples. 147

Appendix 2.3. Class separation using depth-normalized unique *k*-mers 148

Appendix 2.4. Distribution of top-to-second unique KPMR dominance ratios across samples 149

Appendix 2.5. Distribution of *E. coli* to *Shigella* dominance ratios stratified by true spike organism 150

Appendix 4.1. Sequencing and taxonomic classification metrics for *Listeria* spp. short-read quasi-metagenomic datasets generated in Chapter 4 151

Appendix 4.2. Sequencing and taxonomic classification metrics for VTEC short-read quasi-metagenomic datasets generated in Chapter 4 153

Appendix 4.3. Sequencing and taxonomic classification metrics for *Shigella* spp. short-read quasi-metagenomic datasets generated in Chapter 4 155

Appendix 4.4. PulseNet Canada Guidelines for Assessing the Quality of Sequence Data in BioNumerics. 156

Appendix 5.1. Sequencing and taxonomic classification metrics for Listeria spp. long-read quasi-metagenomic datasets generated in Chapter 5. 159

LIST OF ABBREVIATIONS

BLAST – Basic Local Alignment Search Tool

CIDT – Culture Independent Diagnostic Test

DNA – Deoxyribonucleic acid

FBP – Foodborne pathogen

FN – False negative

FP – False positive

gDNA – genomic deoxyribonucleic acid

KPMR – *k*-mers per million reads

LCA – lowest common ancestor

NC – negative control

NCBI – National Center for Biotechnology Information

NSD – No saponin depletion

ONT – Oxford Nanopore Technologies

PNC – PulseNet Canada

RefSeq – Reference Sequence Database

RTE – ready-to-eat

SD – saponin depletion

SI – single isolate

SNV – single nucleotide variant

SNP – single nucleotide polymorphism

SNVPhyl – single nucleotide variant phylogenomic pipeline

TP – true positive

VTEC – verotoxigenic *Escherichia coli*

wgMLST – whole genome multilocus sequencing typing

WGS – whole genome sequencing

STATEMENT OF AUTHOR CONTRIBUTIONS

This thesis follows a manuscript-based (sandwich) format. Chapter 3 has been published as an original research article in *BMC Genomics* (https://doi.org/10.1186/s12864-023-09338-w). Chapters 4 and 5 are presented in manuscript format and are intended for submission to peer-reviewed journals. Final authorship and author order will be determined prior to submission in consultation with all contributing researchers and in accordance with journal authorship guidelines.

Chapter 3: Use of a taxon-specific reference database for accurate metagenomics-based pathogen detection of *Listeria monocytogenes* in turkey deli meat and spinach

Authors: Jillian Rumore, Matthew Walker, Franco Pagotto, Jessica D Forbes, Christy-Lynn Peterson, Andrea D Tyler, Morag Graham, Gary Van Domselaar, Celine Nadon, Aleisha Reimer, Natalie Knox

- All authors conceptualized the study. FP selected the *L. monocytogenes* strain and generated the spike-in datasets for the study. MW performed genomic DNA extraction, library preparation and sequencing. **Jillian Ryz (Rumore)** performed the analysis, interpretation and drafting of the manuscript and figures. MW, FP, JDF, CP, ADT, MG, GV, CN, AR, NK provided critical revision of the manuscript. All authors read and approved the final manuscript.

- All figure captions have been condensed in this thesis from their original form in the published manuscript version to align with thesis formatting requirements.

- Portions of this chapter were reproduced from the published manuscript with minor modifications to formatting and terminology to align with thesis formatting requirements (Section 1.0, Section 2.0, Section 3.0). In instances where the published manuscript referred to “host DNA,” this term was updated to “eukaryotic DNA” to more accurately reflect the origin of DNA in the food matrices analyzed.

Chapter 4: Validation of a Quasi-Metagenomic Workflow for the Rapid Detection and Subtyping of Priority Foodborne Pathogens from Enriched Food Matrices

Authors: Jillian Ryz, Matthew Walker, Morag Graham, Lisa Lix, Grant McClarty, Celine Nadon, Natalie Knox. *Additional authors to be determined prior to manuscript submission*.

- **Jillian Ryz (Rumore)** selected the strains for the study and designed the experimental framework comparing wet-laboratory depletion strategies and bioinformatic pipelines. She prepared the overnight enrichment cultures, performed the artificial contamination (spike-in) of food matrices, conducted saponin-based DNA depletion and genomic DNA extractions, and carried out all data analysis and interpretation. She also designed all figures and tables, and drafting of the manuscript.
- Matthew Walker performed library preparation and high-throughput sequencing of the genomic DNA extractions using the Illumina NextSeq platform. He also generated hybrid reference assemblies corresponding to the strains used in this study.
- Dr. Celine Nadon assisted with strain selection and facilitated access to priority pathogen strains through PulseNet Canada. She also provided critical revisions to the research objectives, and study design.
- Dr. Natalie Knox helped conceptualize the study, secured project funding, and contributed to refining the research objectives and study design.
- Members of the graduate advisory committee (Drs. Morag Graham, Lisa Lix, and Grant McClarty) provided guidance throughout the course of this research. Their feedback on the study design, analytical approach, and statistical framework, particularly the development and interpretation of the pathogen-presence likelihood metric, helped strengthen the overall rigor of the work.

Chapter 5: Implementation of Long-Read Nanopore Sequencing for Rapid Quasi-Metagenomic Pathogen Surveillance and Outbreak Detection of *Listeria monocytogenes* in Complex Food Matrices

Authors: Jillian Ryz, Matthew Walker, Morag Graham, Lisa Lix, Grant McClarty, Celine Nadon, Natalie Knox. *Additional authors to be determined prior to manuscript submission*.

- **Jillian Ryz (Rumore)** conceptualized the study and optimized the bioinformatic workflow developed in Chapter 4 for application to long-read sequencing. She also developed the pathogen-presence likelihood metric based on unique *k*-mer content and carried out all data analysis and interpretation. She also generated all figures and tables, and drafting of the manuscript.
- Matthew Walker performed library preparation and long-read sequencing on the Oxford Nanopore technology MinION using genomic DNA extracts prepared by Jillian Ryz (Rumore) from Chapter 4.
- Dr. Celine Nadon provided critical feedback on the epidemiological interpretation of the results and research objectives.
- Dr. Natalie Knox provided critical review of the bioinformatic workflow, secured funding and contributed to the conceptual development and refinement of the research objectives.
- Members of the graduate advisory committee (Drs. Morag Graham, Lisa Lix, and Grant McClarty) provided ongoing oversight and advice throughout this research. Their insights on the analytical framework and interpretation of results helped refine the methodological approach and strengthen the overall rigor of the study.

1.0 Chapter 1 - Introduction

1.1 Public health and economic burden of foodborne illness

Protecting the food supply and reducing the burden of foodborne illness remain important global public health priorities. Although food safety regulations, preventive control measures, and laboratory surveillance have advanced considerably, foodborne diseases continue to cause substantial illness worldwide. The World Health Organization estimates that foodborne pathogens are responsible for approximately 600 million illnesses and 420,000 deaths globally each year (1). Although the United States and Canada are widely regarded as having among the safest food supplies globally, the burden of foodborne disease remains considerable. In the United States, foodborne pathogens are estimated to cause 48 million illnesses, 128,000 hospitalizations, and 3,000 deaths each year (2). Similarly, foodborne illness continues to represent a significant public health concern in Canada, where approximately 4 million illnesses, 11,600 hospitalizations, and 238 deaths are reported annually (3).

Beyond health impacts, foodborne illness imposes substantial economic costs on healthcare systems, the food industry, and society. In the United States, the annual economic burden associated with major foodborne pathogens has been estimated to exceed USD 17 billion, accounting for medical costs, productivity losses, and premature mortality (4). Prior Canadian analyses similarly demonstrate the substantial economic burden associated with foodborne disease. These studies estimate that food-related illnesses generate approximately CAD 3.7 billion in annual costs, including direct healthcare expenditures as well as industry-related losses associated with product recalls, reduced consumer confidence, and litigation (5). Outbreaks involving high-severity pathogens and large-scale recalls often generate the greatest societal and financial impacts. For instance, the nationwide 2008 outbreak of *Listeria monocytogenes* linked to contaminated ready-to-eat (RTE) delicatessen meats resulted in estimated economic losses of approximately CAD 242 million (5). These findings highlight that foodborne illness continues to represent an important public health issue, with foodborne outbreaks continuing to occur in Canada despite sustained prevention and control efforts. The diversity and complexity of these outbreaks reflect the wide range of hazards capable of contaminating the food supply.

Foodborne contamination can arise from a broad range of biological and chemical hazards introduced at any stage of the farm-to-fork continuum, from primary production through to consumption. These hazards include viruses, bacteria, parasites, chemical agents, toxins, and allergens, each capable of causing adverse health effects following ingestion (3). Although many foodborne illnesses are self-limiting and predominantly characterized by generalized and non-specific gastrointestinal symptoms such as

nausea, abdominal cramping, and diarrhea, severe outcomes including fever, bacteremia, and renal failure remain clinically significant and occur disproportionately in persons with weakened immune systems (6).

1.2 High priority foodborne pathogens in Canada

The public health impact of foodborne pathogens is not evenly distributed across organisms. In Canada, *Listeria monocytogenes* and verotoxigenic *Escherichia coli* (VTEC) O157 contribute disproportionately to severe disease outcomes, accounting for approximately 10% of hospitalizations and 41% of deaths associated with domestically acquired infections (3). VTEC, including *E. coli* O157:H7 and clinically significant non-O157 serotypes such as O26, O103, O111, and O145, remains a leading cause of severe foodborne illness in North America (7–9). For these pathogens, clinical presentations range from hemorrhagic colitis to hemolytic uremic syndrome, a potentially fatal complication characterized by acute renal failure that occurs most frequently among children (10). Fresh and minimally processed foods have emerged as important vehicles of VTEC transmission. From 2009 to 2018, at least 40 VTEC outbreaks linked to leafy greens were identified across the United States and Canada, accounting for more than 1,200 confirmed illnesses (11). Canadian outbreak investigations have similarly documented recurrent associations between *E. coli* O157:H7 infections and lettuce products, particularly romaine lettuce (12). At the same time, ground beef remains a significant source of VTEC exposure, with Canadian outbreaks continuing to prompt large-scale recalls from beef-processing facilities (13). These findings highlight the persistent risk of VTEC contamination across both fresh produce and beef commodities.

Although comparatively rare, *Listeria monocytogenes* causes listeriosis, a severe invasive disease with one of the highest case-fatality rates among foodborne pathogens, estimated at 20–30% (14). Consequently, *L. monocytogenes* is the leading cause of death among foodborne illnesses of known etiology in Canada (3). The microorganism's ability to grow at refrigeration temperatures and persist within food-processing environments contributes to its recurrent association with RTE foods (15). Outbreak investigations have frequently implicated refrigerated RTE products, including deli meats, dairy products, and prepared foods (16–19), while more recent evidence points to an expanding range of vehicles, particularly minimally processed and prepackaged produce (20–22).

In contrast, *Shigella* spp. account for a smaller proportion of hospitalizations and deaths but remain an important public health concern due to their extremely low infectious dose, estimated at approximately 10–100 organisms (23). This low infectious dose facilitates rapid transmission within both community and foodborne settings (23). Although *Shigella* transmission has traditionally been associated with person-to-

person fecal–oral transmission and contaminated water, foodborne transmission is well documented and has been linked to numerous outbreaks involving fresh produce and RTE foods (24–26). The increasing global prevalence of multidrug-resistant *Shigella* strains presents an additional challenge for public health systems. Resistance to commonly used antimicrobial agents has been widely reported, complicating clinical management, limiting treatment options, and increasing the risk of prolonged transmission during outbreak events (27–29). Taken together, the epidemiological and clinical characteristics of high-priority foodborne pathogens in Canada highlight the critical importance of rapid, high-resolution surveillance and outbreak detection systems.

1.3 Current frameworks for foodborne outbreak detection

A foodborne illness outbreak is generally defined as the occurrence of two or more illnesses linked to a common food or beverage source within a defined period (30). Modern food distribution networks and increased international travel increase the likelihood that outbreaks may extend across wide geographic regions, including multiple jurisdictions and countries. In Canada, PulseNet Canada (PNC), established in 2001 and coordinated by the Public Health Agency of Canada, serves as the national laboratory surveillance network for foodborne disease. PNC links federal, provincial, and territorial public health laboratories through a centralized, virtual database that supports rapid interlaboratory comparison and information exchange (31). To facilitate global harmonization of outbreak detection and response, PNC operates within PulseNet International, a collaborative network encompassing 88 countries. Through routine surveillance activities, PNC monitors priority foodborne pathogens of public health importance, including *Vibrio* spp., *E. coli*, *Shigella* spp., *Salmonella* spp., and *L. monocytogenes*, to identify clusters of genetically related cases suggestive of common-source outbreaks (31).

A central element of foodborne outbreak detection involves high-resolution characterization of bacterial isolates through DNA “fingerprinting”. Currently, this is achieved using whole genome sequencing (WGS), which provides comprehensive genomic information for a given organism in a single assay (32). The adoption of WGS of cultured microorganisms has substantially improved the speed and discriminatory power of outbreak investigations, enabling more precise identification of contamination sources and transmission pathways. These advances have contributed to measurable improvements in public health response and have informed food safety policy decisions in Canada (18,20,22,33,34).

Despite its demonstrated utility, widespread implementation of WGS remains uneven globally. Infrastructure costs, operational expenses, bioinformatics capacity, and data management requirements

continue to limit routine adoption in many regions outside North America (35). In addition, WGS-based surveillance relies on successful recovery of the target organism. Pathogen isolation can be labor-intensive and time-consuming and may be particularly challenging for fastidious or non-culturable organisms, potentially delaying detection (36). These challenges are amplified in food samples, which are often heterogeneous, compositionally complex, and may contain pathogens at concentrations lower than those typically observed in clinical specimens, yet still sufficient to cause disease (37). Background microbiota may further complicate recovery through competitive interactions or the production of inhibitory compounds, as reported for *S. sonnei* (38). What is most interesting and certainly of the greatest concern, is the causative agent cannot be identified in ~60% of suspected food-related illnesses. This means that of the 4 million cases annually, ~ 2.4 million cases cannot be resolved with conventional, culture-based methods (3).

1.4 Culture-based and culture-independent diagnostic approaches in foodborne surveillance

Culture-enrichment methods remain fundamental to food microbiology and are widely regarded as the benchmark approach for enhancing pathogen detection sensitivity (37). These procedures are designed to facilitate recovery of stressed or damaged cells while increasing pathogen abundance to levels compatible with downstream detection and isolation. Standard enrichment protocols generally consist of two stages. The initial pre-enrichment step uses non-selective or mildly selective media to promote cellular resuscitation and alleviate matrix-associated inhibitory effects, thereby improving the recovery of sub-lethally injured organisms (39,40). This is followed by selective enrichment, where growth conditions are optimized to favor the target organism and restrict competing microbiota, increasing the probability of successful isolation. Although culture-based enrichment remains critical for recovering foodborne pathogens from food samples, the workflow is labour-intensive and inherently time-consuming because sequential enrichment, isolation, and confirmatory testing are required before genomic characterization can begin. These limitations have driven the development of molecular and sequencing-based approaches capable of accelerating pathogen detection and characterization.

Culture-independent diagnostic tests (CIDTs) have been widely adopted in frontline clinical and food microbiology laboratories and have significantly altered pathogen detection workflows. These methods, which commonly rely on nucleic acid amplification or antigen-based detection, allow rapid identification of enteric pathogens directly from clinical specimens or food matrices without requiring organism recovery (41,42). Their widespread adoption is largely driven by operational advantages,

including faster turnaround times, simplified laboratory workflows, reduced dependence on specialized expertise, and lower overall costs. (36). In addition, CIDTs frequently demonstrate greater analytical sensitivity than conventional culture methods, particularly when detecting fastidious or low-abundance organisms (36).

Despite these advantages, the increased use of CIDTs presents challenges for public health surveillance systems that rely on cultured isolates. Reduced recovery of viable organisms limits the ability to perform strain-level characterization, evaluate genomic relatedness among cases, and establish epidemiological links between clinical infections and potential food, environmental, or animal sources (43,44). These limitations may delay outbreak detection and prolong the period during which contaminated products remain in circulation. In response to these constraints, culture-independent approaches capable of retaining genomic resolution have received growing attention. Among these, metagenomics has emerged as a promising sequencing-based approach for pathogen detection and characterization.

1.5 Metagenomics and quasi-metagenomics in foodborne pathogen detection

Metagenomics involves sequencing all genetic material present within a sample without prior isolation or cultivation of the target organism (45). Unlike targeted CIDTs, metagenomic sequencing generates a comprehensive digital record that includes pathogens, background microbiota, and eukaryotic-derived DNA, allowing retrospective analysis and data sharing across laboratories (46). To date, most methodological developments in metagenomics have occurred within the clinical domain, where the approach has demonstrated increasing accuracy for diagnosing infectious diseases directly from stool, blood, and other clinical specimens (47–51).

Within the context of food safety, metagenomics has shown promise for characterizing microbial population dynamics during enrichment (52,53), detecting fastidious or non-culturable organisms (54,55), and identifying pathogens directly from contaminated foods (41,56,57). As a rapidly evolving laboratory tool, metagenomics has the potential to modernize food safety management by expanding both the speed and scope of contamination detection and outbreak investigations. However, achieving sufficient genomic coverage of low-abundance pathogens using direct metagenomic sequencing often requires large volumes of sequencing data, which can be financially and computationally impractical for many laboratory settings (50). Sequencing directly from primary enrichment cultures, commonly referred to as quasi-metagenomics, offers a practical compromise between sensitivity and analytical speed (58). By leveraging selective enrichment to increase bacterial pathogen abundance while bypassing microbiological isolation, quasi-

metagenomic approaches can shorten the time required for pathogen detection (58–63). Ultimately, this method has the potential to expedite pathogen detection and provide a result at least one week sooner than conventional methods. Although promising, multiple technical and analytical barriers continue to restrict the routine integration of quasi-metagenomic approaches into foodborne surveillance workflows.

1.6 Challenges in food metagenomic pathogen detection

One of the most significant challenges in quasi-metagenomic analysis arises from the overwhelming presence of non-target DNA within many sample types. In food matrices, eukaryotic DNA derived from plant or animal tissues often constitutes the majority of sequencing reads, thereby reducing the effective sequencing depth available for microbial targets. This imbalance can reduce analytical sensitivity when pathogens are present at low abundance and may introduce classification artifacts that complicate downstream interpretation (64). Several strategies have been proposed to address this challenge, including eukaryotic DNA depletion, selective lysis, and targeted enrichment (65,66). While these approaches may reduce the proportion of non-target DNA, they also introduce additional laboratory complexity and may alter the observed microbial community structure. Among these strategies, eukaryotic DNA depletion has received considerable attention, particularly in clinical sample types, where it has demonstrated promising results (66–69). However, its effectiveness and potential to introduce analytical bias remain less well characterized for complex food matrices.

Pathogen detection in quasi-metagenomic datasets typically relies on taxonomic classification algorithms that assign sequence reads or assembled contigs to reference genomes contained within curated databases (70). Consequently, the reliability of taxonomic inference depends strongly on the composition, structure, and curation of these reference databases (71,72) . Several factors have been shown to influence classification accuracy, including low-complexity genomic regions, contamination within publicly available genome assemblies, and incomplete representation of taxonomic diversity (73–76). These issues can lead to systematic classification errors, particularly false-positive (FP) assignments.

From a food safety perspective, a FP result occurs when a pathogen is incorrectly identified in a sample where it is absent. Such outcomes can have significant consequences, including unnecessary product recalls, economic losses, and avoidable food waste. Conversely, false-negative (FN) results occur when a pathogen is present but remains undetected by the analytical workflow. FN classifications frequently arise when reference databases lack suitable or closely related genomes. This issue is

exacerbated by historical sequencing priorities that have focused primarily on common human pathogens, resulting in underrepresentation of rare, emerging, or environmentally associated organisms (71,77).

Interpreting FN events presents a distinct challenge in quasi-metagenomics. Unlike FP errors, which can often be investigated through database inspection or reclassification, FN determinations are difficult to verify because the negative class may include organisms that are unknown, uncharacterized, or absent from existing reference collections (78). This “unknown unknowns” problem complicates performance assessment and limits the ability to accurately estimate detection sensitivity under real-world conditions. Consequently, method development and benchmarking studies have largely focused on evaluating positive classification outcomes, which are more readily measured and validated (78).

To reduce FP classifications, some researchers have advocated for comprehensive, taxon-agnostic reference databases that encompass multiple domains of life (79). While this strategy can reduce spurious matches resulting from incomplete reference representation, its practical implementation faces several limitations. First, genomic repositories do not fully capture the diversity or ecological distributions of organisms encountered in natural or food-associated environments (77,78). Second, large reference databases require substantial computational resources, often demanding tens to hundreds of gigabytes of memory, which may exceed the capabilities of routine diagnostic laboratories without access to high-performance computing infrastructure (32,77,78). These limitations have prompted interest in smaller, taxon-specific reference databases, which may improve classification specificity while reducing computational demands. However, the effectiveness of such databases for metagenomic pathogen detection in food remains incompletely characterized.

Analytical interpretation of quasi-metagenomic datasets derived from food matrices remains methodologically complex. A major limitation is the absence of a universally accepted end-to-end analytical workflow capable of supporting reliable pathogen detection, strain-level inference, and outbreak investigation across diverse food matrices. Although numerous bioinformatic pipelines and modular frameworks have been developed, many were originally designed for single-isolate sequencing or clinical metagenomics applications (80,81). As a result, they often rely on assumptions, including clonal input, relatively uniform sequencing coverage, and minimal background interference, that are rarely met in complex food-derived datasets. In addition, current analytical workflows frequently require substantial computational resources, including large reference databases, high-memory computing environments, and extensive parameter optimization to produce stable and reproducible results (82).

Beyond these technical challenges, an important practical limitation lies in how analytical outputs are translated into clear decision criteria for end users. Classification tools typically report read counts, *k*-mer statistics, or abundance estimates, yet they provide limited guidance on how these values should be interpreted when determining pathogen-presence or absence. A *k*-mer is a short DNA sequence of a defined length (*k*) generated by dividing sequencing reads into overlapping sequence fragments and can be used to identify taxon-specific sequence signatures (83). In practice, metagenomic workflows frequently use the number of reads assigned to a taxon as a primary indicator of pathogen-presence or abundance, although this approach can be influenced by sequencing depth and background DNA (47,84). However, read counts can be influenced by sequencing depth, library composition, genome size, and classifier behavior, making them susceptible to bias and inconsistent performance across datasets.

Previous studies have suggested that additional quantitative filters, including coverage distribution, abundance consistency, and breadth of genome representation, can improve specificity and reduce false-positive detections in complex metagenomic datasets (64,85,86). Nevertheless, these metrics are often applied independently or without a structured interpretive framework. The absence of standardized decision criteria can lead to inconsistent thresholds, reliance on single metrics, or subjective interpretation, particularly in borderline cases involving low-abundance organisms or closely related taxa. In the context of surveillance, such variability can complicate reproducibility and limit comparability across laboratories. These challenges underscore the need for robust analytical frameworks capable of integrating multiple lines of evidence to support consistent pathogen detection and interpretation.

Taken together, these limitations indicate that although quasi-metagenomics holds considerable promise for culture-independent foodborne pathogen surveillance, its broader implementation will require careful evaluation of sample composition, reference database design, and analytical performance. Continued methodological refinement and rigorous validation will be essential for establishing quasi-metagenomic sequencing as a dependable tool for public health surveillance and foodborne outbreak detection.

1.7 Long-read sequencing technologies in quasi-metagenomic surveillance

Beyond improvements in computational workflows and reference database design, recent advances in long-read sequencing technologies have created new opportunities for quasi-metagenomic applications in foodborne disease surveillance (59,87,88). Platforms developed by Oxford Nanopore Technologies (ONT) and Pacific Biosciences generate long sequence reads capable of spanning repetitive

regions, structural variants, and accessory genomic elements that are often fragmented when using short-read sequencing technologies (89). These capabilities are particularly relevant for pathogen genomics, where plasmids, prophages, genomic islands, and other mobile genetic elements play important roles in virulence, environmental persistence, and antimicrobial resistance. Within quasi-metagenomic workflows, the ability to recover longer contiguous sequences directly from enriched samples may improve genome reconstruction and enhance strain-level discrimination without requiring isolate recovery (90). In addition, the portability, flexible throughput, and rapid library preparation associated with certain long-read platforms enable near-real-time sequencing, making these technologies attractive for decentralized or time-sensitive surveillance scenarios (91).

1.8 Challenges for long-read quasi-metagenomics

Despite these advantages, the analytical characteristics of long-read data introduce challenges that are particularly relevant for quasi-metagenomic applications. Long-read sequencing platforms have historically exhibited higher per-base error rates than short-read technologies, although ongoing improvements in sequencing chemistry, base-calling algorithms, and consensus correction methods have substantially improved raw-read accuracy (92). Sequencing errors can propagate through downstream analyses, affecting taxonomic classification, *de novo* genome assembly, and variant detection, which are essential processes for strain-level characterization and outbreak detection. These challenges are further compounded by the fact that many taxonomic classification tools were originally developed and parameterized using short-read datasets. As a result, their performance may differ when applied to long-read data (93).

Classifier parameterization therefore plays a critical role in long-read analytical performance. Empirical studies have shown that shorter *k*-mer and minimizer lengths can improve classification sensitivity for error-prone reads by increasing tolerance to mismatches while preserving discriminatory signal (85,94). However, excessive reduction in *k*-mer length introduces trade-offs, including reduced taxonomic specificity and an increased likelihood of spurious matches within large reference databases (95). Accurate classification of long-read quasi-metagenomic data therefore requires careful balancing of error tolerance and taxonomic resolution through context-specific parameter optimization.

Quantitative interpretation of long-read quasi-metagenomic data also requires methodological adjustment. Abundance estimates based solely on read counts may be biased due to the substantial variability in read length and informational content associated with long-read sequencing. Alternative

approaches incorporating *k*-mer frequencies, coverage profiles, or length-normalized metrics have been proposed as more stable indicators of taxon representation (85,94). These approaches may be particularly valuable in foodborne surveillance contexts, where accurate estimation of pathogen signal can influence both risk assessment and epidemiological interpretation.

Overall, these considerations suggest that while long-read sequencing technologies offer important advantages for quasi-metagenomic surveillance, their effective application depends on analytical frameworks that account for platform-specific error profiles, classifier performance, and appropriate strategies for data interpretation.

1.9 Aims and rationale

This thesis follows a manuscript-based format and consists of a series of interrelated studies that evaluate the feasibility of quasi-metagenomic sequencing as a rapid, culture-independent framework for foodborne pathogen detection and high-resolution genomic characterization. The central hypothesis of this work is that quasi-metagenomic sequencing can generate strain-level genomic information comparable to conventional isolate-derived WGS while reducing the time required to produce actionable data for outbreak detection and public health response. This hypothesis reflects the potential of quasi-metagenomics to circumvent delays associated with pathogen isolation while maintaining the discriminatory resolution required for rapid and accurate pathogen characterization and outbreak detection.

Despite its promise, the application of quasi-metagenomic sequencing in food safety presents several methodological challenges. Laboratory workflows, together with the complexity of food matrices, can influence pathogen recovery, and taxonomic classification remains difficult in mixed microbial communities. Interpretation becomes further complicated when pathogens are present at low abundance or when closely related taxa share substantial genomic similarity. Addressing these challenges requires systematic evaluation of both laboratory and computational factors that influence analytical reliability and detection accuracy. Accordingly, this thesis is organized as a series of interconnected studies that examine methodological factors influencing quasi-metagenomic pathogen detection and characterization, with the goal of optimizing these factors to address the challenges that currently limit its broader application.

Chapter 3 focuses on the computational determinants of taxonomic classification performance, examining how reference database composition and classifier parameterization influence pathogen detection accuracy in metagenomic datasets. Particular emphasis is placed on *L. monocytogenes* detection within food matrices dominated by eukaryotic DNA, reflecting both its public health significance and the

analytical difficulties associated with identifying low-abundance pathogens in complex food matrices. This chapter evaluates whether curated pathogen-specific reference databases, coupled with optimized classifier parameters, can improve taxonomic precision while reducing computational burden and minimizing classification artifacts.

Building on the computational framework established in Chapter 3, Chapter 4 investigates how laboratory procedures and analytical strategies influence pathogen detection in quasi-metagenomic datasets derived from enriched food samples. Specifically, this chapter evaluates the effects of selective enrichment and saponin-based eukaryotic DNA depletion on pathogen-specific sequencing signal across multiple priority pathogens and food matrices. In parallel, the study examines how reference database design and classification decision criteria affect taxonomic resolution, particularly for closely related organisms such as *E. coli* (VTEC) and *Shigella* spp., which often produce ambiguous assignments when analyzed against broad reference databases. To address these challenges, curated pathogen-specific databases are assessed for their ability to improve classification stability. In addition, a pathogen-presence likelihood metric is introduced to formalize detection criteria by integrating depth-normalized unique *k*-mer counts (unique KPMR), robust *z*-score, and tie-resolution decision rules. Together, these components are evaluated to determine whether coordinated laboratory and analytical optimization can improve the consistency and reliability of pathogen detection across diverse pathogens and food matrices when using a quasi-metagenomic approach.

Chapter 5 extends this investigation to long-read sequencing technologies, assessing their suitability for strain-level genomic inference directly from enriched food samples. Long-read sequencing platforms offer several potential advantages for quasi-metagenomic applications, including improved genome reconstruction, enhanced resolution of structural variation, and improved recovery of plasmids and other mobile genetic elements. However, the distinct error profiles associated with long-read sequencing require careful analytical consideration. This chapter evaluates whether long-read quasi-metagenomic datasets, when analyzed using appropriate classification and phylogenomic approaches, can support reliable strain-level discrimination and outbreak clustering for *L. monocytogenes* without reliance on isolate recovery.

Collectively, these studies evaluate whether coordinated optimization of laboratory protocols, reference databases, analytical workflows and defined interpretation criteria can enable quasi-metagenomics to approach the strain-level resolution of isolate-based WGS while reducing the time required to generate actionable results for foodborne surveillance and outbreak detection.

2.0 Chapter 2 - Methods

This chapter consolidates experimental and computational methods used across the thesis to avoid redundancy. Methods developed and benchmarked in Chapter 3 underpin reference database construction, taxonomic classifier selection, and validation of quasi-metagenomic performance. These workflows are applied to short-read quasi-metagenomic datasets in Chapter 4 and long-read quasi-metagenomic datasets in Chapter 5. Where procedures are shared across chapters, they are described once here and explicitly mapped to their originating chapter(s).

2.1 Study Datasets

Since naturally contaminated samples were not available, all datasets used in this thesis were derived from controlled artificial contamination experiments. For Chapter 3, a total of six samples were prepared, consisting of three ready-to-eat (RTE) turkey deli meat samples and three prepackaged fresh spinach samples, with each food commodity represented by a high (10^4 CFU/mL), medium (10^3 CFU/mL), or low (10^2 CFU/mL) *Listeria monocytogenes* spike-in level. To generate these datasets, two food commodities previously implicated in foodborne disease outbreaks of *L. monocytogenes*, including RTE turkey deli meat and prepackaged fresh spinach, were artificially contaminated with a strain of *L. monocytogenes* (HPB5415; GCF_000712385.1) from the 2008 Canadian listeriosis outbreak (96). This strain was isolated from the implicated food product (i.e., RTE turkey deli meat) during the outbreak and was therefore selected to more closely resemble a true outbreak scenario. To streamline data analysis, the same outbreak strain (HPB5415) was also used to contaminate spinach. Following overnight growth, a single colony was inoculated into 200 mL of Brain Heart Infusion (BHI) broth and incubated overnight at 35 °C ± 2 °C to yield approximately 10^9 CFU/mL. For each food product, 25 g of food was added to 225 mL of UVM1 (primary enrichment broth) and stomached for 2 min at 260 rpm. Aliquots of food homogenate were then inoculated with serial dilutions of HPB5415 prepared in peptone water to achieve the target spike-in levels of 10^4, 10^3, and 10^2 CFU/mL. These concentrations were selected to represent a range of contamination levels relevant to foodborne exposure and are consistent with infectious doses reported for both healthy individuals and persons with weakened immune systems (97). Prior to inoculation, both food commodities were confirmed to be negative for *L. monocytogenes* by overnight culture on RAPID'L.Mono™ (Bio-Rad, Hercules, California).

To generate study datasets in Chapter 4 and Chapter 5**,** eight outbrea*k*-associated *Listeria monocytogenes* strains and three non-outbreak associated *Listeria* spp. strains previously isolated from

various foods were selected for the study. All isolates were streaked on to RAPID'L.Mono™ (RLM) agar (Bio-Rad, Hercules, CA), and incubated overnight at 35°C ± 2°C. For each strain, a single colony was inoculated into 200mL of Brain Heart Infusion (BHI) broth and grown overnight at 35°C ± 2°C to yield approximately 10^9 CFU/mL. Serial dilutions were prepared in peptone water at three concentrations categorized as high (10^3 CFU/mL), medium (10^2 CFU/mL) and low (10^1 CFU/mL). Following Health Canada's procedure for stressing microorganisms in artificially contaminated samples (Annex 4.2) serial dilutions of strains selected for artificial contamination of delicatessen (deli) meat were placed in a hot water bath at 55°C ± 2°C for 10 minutes (39). No stressing was required for strains selected for artificial contamination of raw, unprocessed foods. To determine the effects of the stressing treatment on each strain, 100 µL aliquots of each of the three dilutions was plated in triplicate on selective media, including PALCAM agar and RAPID'L.Mono™ (RLM) agar, as well as non-selective agar (BHIA) and incubated overnight at 35°C ± 2°C. Enumeration was performed to verify colony counts acquired on selective agar were 20% of the counts acquired on non-selective agar (e.g., 80% stressed cells) (39).

Exclusive to Chapter 4, outbreak-associated and non-outbreak-associated *E. coli* and *Shigella* strains were prepared using a standardized workflow. Four outbrea*k*-associated *E. coli* O157:H7 strains previously isolated from ground beef or leafy greens, along with two non-outbrea*k*-associated *E. coli* strains (*E. coli* O26:H11 and *E. coli* O103:H25), were streaked onto Rainbow Agar O157 (Biolog) and incubated overnight at 35°C ± 2°C. In parallel, two outbrea*k*-associated *Shigella sonnei* strains linked to the consumption of baby carrots and one non-outbrea*k*-associated *Shigella flexneri* strain were streaked onto Rainbow® Agar Shigella/Aeromonas (Biolog, Hayward, CA) and incubated under identical conditions (39).

For each strain, a single colony was inoculated into 200 mL of Brain Heart Infusion (BHI) broth and grown overnight at 35°C ± 2°C to achieve approximately 10^9 CFU/mL. Cultures were serially diluted in peptone water to generate three spike-in levels categorized as high (10^3 CFU/mL), medium (10^2 CFU/mL), and low (10^1 CFU/mL), which were used for downstream artificial contamination experiments in Chapter 4 (39).

2.2 Artificial contamination and enrichment

Artificial contamination experiments were conducted using food matrices and enrichment conditions specific to each target pathogen, following Health Canada reference methods. For *Listeria* spp., three ready-to-eat (RTE) deli meats (turkey, ham, and roast beef) previously implicated in outbreaks of *Listeria monocytogenes*, as well as a RTE prepackaged salad kit, were selected. Sample preparation

followed The Compendium of Analytical Methods MFLP-01 (39). For each matrix, 25 g of food was combined with 225 mL of UVM1 primary enrichment broth and homogenized by stomaching for 2 min at 260 rpm. Food homogenates were artificially contaminated with *Listeria* spp. to generate three spike-in levels: high (10^3 CFU/mL), medium (10^2 CFU/mL), and low (10^1 CFU/mL). Enrichments were incubated at 30°C ± 2°C for 28 h. Spike-in levels were verified by plating 100 µL aliquots onto RAPID'L.Mono™ (RLM) agar (Bio-Rad) and incubating overnight at 35°C ± 2°C. Negative controls were prepared in parallel and plated on RLM to confirm the absence of pre-existing *Listeria* contamination. For each food product, three spike-in samples and one negative control were prepared (Chapters 4 and 5).

For VTEC, foods commonly associated with outbreaks, including retail ground beef and prepackaged lettuce were artificially contaminated in accordance with The Compendium of Analytical Methods MFLP-52 (39). Twenty-five grams of each food were added to 225 mL of modified tryptone soya broth (mTSB) containing 100 mg/mL novobiocin and homogenized by stomaching for 2 min at 260 rpm. Antimicrobial supplements including vancomycin (1 mg/mL) and cefsulodin (0.3 mg/mL) were added to achieve final concentrations of 10 µg/mL and 3 µg/mL, respectively. Homogenates were spiked with VTEC strains at high (10^3 CFU/mL), medium (10^2 CFU/mL), and low (10^1 CFU/mL) levels and incubated at 42°C ± 2°C for 24 h. Negative controls were prepared and plated on Rainbow Agar O157 (Biolog) to confirm the absence of background VTEC contamination. Three spike-in samples and one negative control were prepared per food matrix (Chapter 4).

For *Shigella* spp., artificial contamination followed The Compendium of Analytical Methods MFLP-25 (39). Twenty-five grams of baby carrots were added to 225 mL of *Shigella* enrichment broth containing 0.5 µg/mL novobiocin and held at room temperature for 10 min with periodic agitation. Homogenates were spiked with *Shigella* strains at high (10^3 CFU/mL), medium (10^2 CFU/mL), and low (10^1 CFU/mL) levels and incubated at 42°C ± 2°C for 24 h in anaerobic jars containing CO_2 generator bags. Spike-in levels were confirmed by plating 100 µL aliquots onto Rainbow® Agar Shigella/Aeromonas (Biolog) and incubating overnight at 35°C ± 2°C (24 h ± 2 h). Negative controls were included to verify the absence of pre-existing *Shigella* contamination. Following enrichment, 100 µL of each spike-in culture and the corresponding negative control were plated on selective media and incubated overnight at 35°C ± 2°C to confirm pathogen growth in spiked samples and the absence of growth in controls (Chapter 4).

2.3 Eukaryotic DNA depletion

Eukaryotic DNA depletion was performed in Chapter 4 using the saponin-based method developed by Charalampous et al., with minor modifications (67). This approach exploits differences in membrane composition between eukaryotic and bacterial cells. Saponin selectively lyses cholesterol-rich eukaryotic cells, releasing DNA that is subsequently degraded by DNase treatment, while microbial cells remain largely intact. This reduces eukaryotic-derived DNA and enriches pathogen-derived nucleic acids for downstream sequencing and genomic analyses (67). Briefly, 400 µL of the overnight enriched sample was added to a 1.5 mL microfuge tube and centrifuged at 8000g for 10 minutes. The supernatant was removed, and the pellet was re-suspended in 350 µL of Phosphate-buffered saline (PBS) to which 100 µL of 10% saponin was added followed by incubation at room temperature for 10 minutes. Next, 350 µL of UltraPure DNAse and RNAse free water was added and incubated for an additional 30 seconds at room temperature, after which 12 µL of 5M NaCl was added, mixed, and then centrifuged at 6000g for 5 minutes. The supernatant was removed, and the pellet resuspended in 100 µL of PBS to which 100 µL of HL-SAN buffer (5.5M NaCl and 100 mM $MgCl_2$) and 10 µL of HL-SAN DNase was added, vortexed, and then incubated at 37°C ± 2°C for 15 minutes at 800 rpm. To wash the sample, 800 µL of PBS was added followed by centrifugation at 6000g for 3 minutes. The supernatant was then discarded, and the pellet was washed again with 1 mL of PBS and centrifuged at 6000g for another 3 minutes. The supernatant was removed, and the pellet was used in the workflow for gDNA extraction.

For non-eukaryotic DNA depleted samples created in Chapter 4, a 200 µL aliquot of the 24-hour enriched sample was transferred to a 1.7 mL microcentrifuge tube and centrifuged at 8000g for 10 minutes. The supernatant was discarded, and the pellet was used in the workflow for gDNA extraction.

2.4 Genomic DNA extraction

Three commercial kits demonstrating good performance in two comprehensive benchmarking studies involving clinical specimens were selected and assessed in Chapter 3: 1) QIAamp® Fast DNA Stool Mini Kit (Qiagen, Valencia, California), 2) DNeasy® PowerSoil Kit (Qiagen), and 3) Quick-DNA Fecal/Soil Microbe Miniprep Kit (Zymo Research, Irvine, CA) (98,99). Genomic DNA (gDNA) extraction was carried out according to the manufacturer's instructions; however, an initial bead-beating step (PowerBead Tubes 2 ml, Glass 0.1 mm; Qiagen) was performed using the Vortex-Genie mixer at maximum speed for ten minutes before gDNA extraction using the QIAamp® Fast DNA Stool Mini Kit (Qiagen). Datasets were processed in duplicate for each extraction kit. gDNA concentration was quantitated using the

Qubit Fluorometer (ThermoFisher, Waltham, MA) and diluted to 0.2 ng/µL. To align with current wet-laboratory practices for extraction and sequencing of priority pathogens for PulseNet Canada (PNC), all samples included in Chapter 4 and Chapter 5 were prepped using the Masterpure Complete DNA and RNA Purification Kit according to internal standard operating procedure DED-PR-350 with minor modifications. The gDNA was re-suspended in 50 µL of Elution Buffer, incubated at 37°C ± 2°C for 3.5 hours shaking at 900 rpm and then stored at 4°C until further processing. Technical replicates were not performed for Chapters 4 or 5, as the primary objective was to assess detection performance and strain-level resolution across different matrices and enrichment conditions, rather than to measure technical variability. Reproducibility was supported through standardized laboratory and bioinformatic workflows, along with internal validation approaches including cross-method comparisons (single isolate sequencing vs. quasi-metagenomics vs. long-read). This is consistent with routine public health genomics workflows, where robust quality control and standardized pipelines are used in place of technical replication to ensure data reliability.

2.5 Library preparation and sequencing

Genomic DNA concentration was quantitated using the Qubit Fluorometer (ThermoFisher, Waltham, MA) and diluted to 0.2 ng/µL. Libraries were prepped and assessed for quality according to the manufacturer's instructions for the Nextera XT DNA Preparation Library Kit (Illumina, San Diego, CA). To maximize pathogen recovery and avoid bias in fragment lengths, size selection was not performed. For samples investigated in chapter 3, sequencing was performed on the Illumina MiSeq using the Illumina MiSeq Reagent Kit v3 (600-cycle).

In Chapter 4, approximately 30-32 libraries were multiplexed per run to achieve a total of five runs using the Illumina NextSeq 550 and Illumina NextSeq 500/550 Mid Output v2 kit (300 cycles). The same DNA preparations for *Listeria* samples were also used for sequencing on the Oxford Nanopore Technologies MinION in Chapter 5. Prior to library preparation, DNA concentration was measured using the Qubit fluorometer. Sequencing libraries were prepared by multiplexing 12 DNA extracts on the R9.4.1 (FLO-MIN106) flowcell using the Rapid Barcoding Sequencing (SQK-RBK004) kit according to the manufacturer's protocol. Each sequencing run was performed on a MinION (MinKNOW v24.02.6) for 24 hours for a total of 12 runs (Chapter 5). For Chapter 4, a total of 134 samples were sequenced, including 66 *Listeria* spp. samples (33 NSD and 33 SD) and 8 negative controls (4 NSD and 4 SD), 36 VTEC samples (18 NSD and 18 SD) and 4 negative controls (2 NSD and 2 SD), and 18 *Shigella* spp. samples (9

NSD and 9 SD) and 2 negative controls (1 NSD and 1 SD). A total of 74 samples were sequenced for Chapter 5, comprising 66 *Listeria* spp. samples (33 NSD and 33 SD) and 8 negative controls (4 NSD and 4 SD).

2.6 Reference sequence data

Complete genomes for bacteria (n = 17,215), archaea (n = 351), viruses (n = 9,507), and human (GRCh38.p13; GCF_ 000,001,405.39) were downloaded from the NCBI RefSeq Database Release 98 (January 2020) using the krakenuniq-download script (94,100). Low-complexity sequence masking of all complete genomes was performed using DustMasker (v1.1.0) with default parameters (95). Contaminant sequence databases, UniVec and EMVEC, were also downloaded using the krakenuniq-download script (94). All reference genomes were screened for completion and contamination using CheckM (v1.2.2; now referred to as CheckM1) and for cross-kingdom contamination using conterminator (v1.c74b5) (Chapter 3) (101–103).

2.7 Tool selection

Three popular *k*-mer based taxonomic classifiers, Centrifuge (v1.0.4_beta)(104), Kraken 2 (v2.1.1) (95), and KrakenUniq (v0.6) (94), were selected for the study based on availability, usability, and adoption. All tools are freely available, well documented, widely adopted by the scientific community, and actively maintained and updated by developers. Additionally, these classifiers support custom database construction and therefore, were evaluated in Chapter 3. Based on the results of the evaluation, KrakenUniq was selected for preforming taxonomic classification in Chapter 4 and Chapter 5.

2.8 Reference database design and development

All taxonomic classifiers depend on a pre-computed reference database composed of previously sequenced genomes. To support the analyses in Chapter 3, two custom reference databases were constructed using identical genome content and taxonomic structure across all classifiers, thereby minimizing potential confounding effects arising from differences among pre-built (default) databases. Complete bacterial, archaeal, viral, and human genomes were retrieved from the RefSeq repository described in Section 2.1, including established contaminant databases, and used to generate a universal (taxon-agnostic) reference database for each of the three metagenomic classifiers.

A subset of complete genomes from the original RefSeq download, corresponding to twenty of the top thirty foodborne bacterial and viral pathogens in Canada, was used to create a foodborne pathogen (FBP)-specific database for each of the three metagenomic classifiers (3). Additional enteric pathogens, including other *Listeria* and *Helicobacter* species, were also included in the FBP-specific reference database based on interest within PNC. The human genome (GCF_000001405.40, GCF_009914755.1) and contaminant databases UniVEC and EMVEC were also included in the FBP-specific reference database to reduce FP classifications as previously recommended (71). To assess reference genome quality and detect reference sequence contamination both the FBP-specific and universal reference databases were screened for contamination using checkM (https://doi.org/10.1186/s12864-023-09338-w; Additional file 6) and for cross kingdom contamination using conterminator (https://doi.org/10.1186/s12864-023-09338-w; Additional file 7). Reference databases were constructed according to the developers' guidelines using default parameters. Owing to the large size of the dataset, the complete list of reference genomes included in the universal and FBP-specific databases is provided in Additional file 7 of the original published manuscript (https://doi.org/10.1186/s12864-023-09338-w). Database size, build time, and memory requirements are detailed in Appendix 2.1. Only the FBP-specific reference database developed for KrakenUniq was used for conducting taxonomic classification in Chapter 4 and Chapter 5.

Using a subset of the data from the FBP-specific reference database described in Chapter 3, two additional reference databases were created, one containing only complete genomes for *E. coli* (EC-specific), and the other containing only complete genomes for *Shigella* spp (SH-specific). The human genome and contaminant databases were also included in both species-specific reference databases. All databases were built according to KrakenUniq developer's guidelines (Chapter 4).

2.9 Data analysis

An overview of the bioinformatic workflows developed and applied across the three research chapters is presented in Figure 2.1, with a comprehensive summary of the bioinformatic software and tools provided in Table 2.1.

2.9.1 Data preprocessing and eukaryotic sequence removal

For all chapters, the general quality of the raw fastq files was assessed using fastqc (v0.11.8) with default parameters (105). Adaptor removal and read trimming were performed using fastp (v0.23.2) with a stringent qualified quality Phred value (-q) adjusted to 20 (default 15) (106).

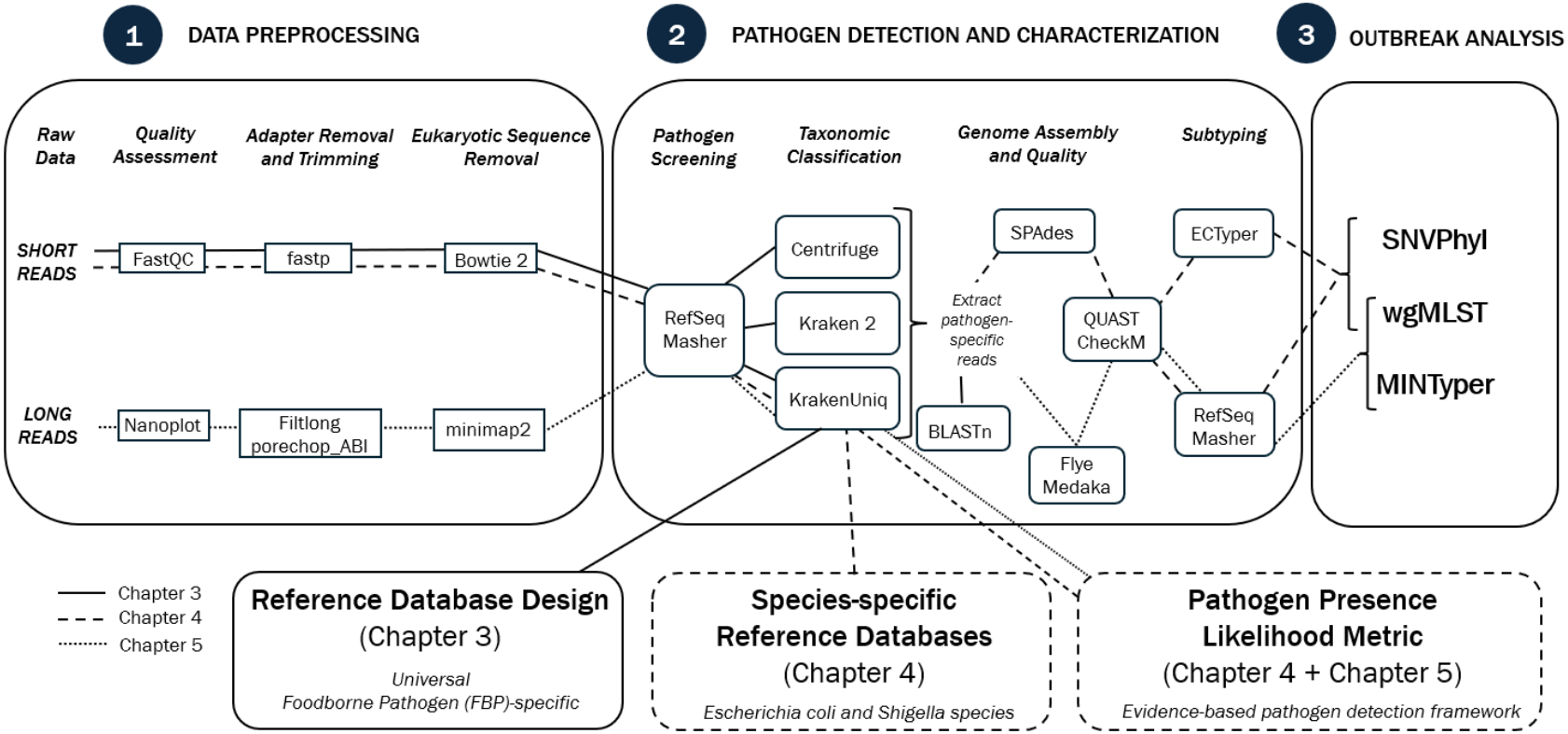


Figure 2.1. Overview of the bioinformatic workflows developed and applied across Chapters 3–5. The workflow is organized into three major stages: (1) data preprocessing, including quality assessment, adapter removal and trimming, and eukaryotic sequence removal; (2) pathogen detection and characterization, including pathogen screening, taxonomic classification, genome assembly and quality assessment, and subtyping; and (3) outbreak analysis using SNP- and allele-based approaches. FBP, foodborne pathogen; wgMLST, whole-genome multilocus sequence typing; SNP, single nucleotide polymorphism.

Table 2.1. Summary of computational and analytical software.

FUNCTIONAL CATEGORY	SOFTWARE TOOL	VERSION	PRIMARY APPLICATION	THESIS CHAPTER(S)
Reference Database Preparation	DustMasker	1.1.0	Low-complexity sequence masking	3
	CheckM	1.2.2	Screening reference genomes for completion and contamination	3
	Conterminator	1.c74b5	Screening for cross-kingdom contamination	3
Data Preprocessing	FastQC	0.11.8	General quality assessment of raw FASTQ files	3, 4, 5
	fastp	0.23.2	Adapter removal and stringent quality trimming (q=20)	3, 4, 5
	Guppy	6.5.7	Basecalling with super-accurate mode for long-read data	5
	NanoPlot	1.41.0	Extraction of long-read length and quality metrics	5
	porechop_ABI	0.5.0	Removal of residual adapters from long-reads	5
	filtlong	0.2.1	Filtering long-reads to retain top 90% quality	5
Eukaryotic Sequence Removal	Bowtie 2	2.4.4	Short-read alignment against food matrix reference indexes	3, 4
	minimap2	2.24 / 2.28	Long-read alignment for eukaryotic filtering and coverage analysis	5
	SAMtools	1.17	Extraction of unmapped reads (Flags -f 12, -F 256 or -f 4)	3, 4, 5
Taxonomic Classification	Centrifuge	1.0.4_beta	K-mer based classification benchmarking	3
	Kraken 2	2.1.1	K-mer/Minimizer based classification benchmarking	3
	KrakenUniq	0.6	Primary taxonomic classifier for all chapters	3, 4, 5
	Pavian	-	Visualization of classification results	3
Pathogen Verification	RefSeq Masher	0.1.2	contains (preliminary screen) and matches (Listeria verification)	4, 5
	ECTyper	1.0.0	In silico speciation and serotyping for VTEC/Shigella	4
Genome Assembly	SPAdes	3.15.3	Short-read genome assembly (--careful mode)	4
	Flye	2.9.3	Long-read de novo assembly	5
	Medaka	1.11.3	Polishing of long-read assemblies	5
	Unicycler	0.4.8.0	Generation of hybrid reference assemblies	4, 5
Assembly Quality	QUAST	5.0.2	Assessment of assembly quality metrics	3, 4, 5
	CheckM	1.2.2	Assessing completeness and contamination of assemblies	3, 4, 5
	Qualimap 2	2.2.2d	Evaluation of read mapping and coverage depth	4, 5
Outbreak Detection	BioNumerics	7.6	wgMLST allele identification	4, 5
	SNVPhyl	1.1	Core genome SNV analysis for short-read data	4
	MINTyper	1.1.2	SNP identification for long-read and mixed datasets	5
Statistics / Visualizations	R / RStudio	4.5.1 / 2025.05.1	Statistical analysis (z-scores, Shapiro-Wilk, T-tests)	3, 4, 5
	ggplot2	4.0.0	Generation of research figures and visualizations	3, 4, 5

Specific to Chapter 3, eukaryotic sequences were filtered using Bowtie 2 (v2.4.4) with custom-built indexes, one for the spinach genome (GCF_002007265.1) and the other for the turkey genome (GCF_000146605.3) (107). Two unplaced scaffolds previously identified as contaminated in the spinach genome (RefSeq identifiers NW_018932190.1 and NW_018932355.1) were masked before building the index (102). Only primary alignments with both reads unmapped were included in the eukaryotic sequence removed datasets using the SAM flags -f 12, -F 256 (108).

In Chapter 4, eukaryotic-derived sequences (attributed to the respective food commodity genomes) were filtered *in silico* using Bowtie 2 (v2.4.4) against custom-built reference indexes (107). Two distinct composite indexes were constructed to address the specific matrices investigated: a meat-specific index containing reference genomes for turkey (GCF_000146605.3), chicken (GCF_016699485.2, GCF_016700215.2), bovine (GCF_002263795.2), and swine (GCF_000003025.6), and a produce-specific index comprising the genomes of carrot (GCF_001625215.1) and lettuce (GCF_002870075.3). Using samtools (v1.17), primary alignments with both reads unmapped were extracted using the SAM flags -f 12, -F 256, sorted and then converted to fastq files using bam2fq (108).

Exclusive to Chapter 5, basecalling was performed using Guppy (v6.5.7) with trimming mode and super accurate base calling enabled (109). Read length and quality metrics per sample were extracted using NanoPlot (v1.41.0) (110). Additional adapter trimming was performed with porechop_ABI (v0.5.0) to remove all remaining adapters (111). Low quality reads were filtered using filtlong (v0.2.1) with --*keep_percent* 90 to retain the best 90% of reads as previously recommended (112). Eukaryotic DNA (attributed to the food matrix) was filtered with minimap2 (v2.24) using the same references as previously described for Chapter 4 datasets (113). Using samtools (v1.17), alignments were extracted using the SAM flag -f 4, sorted and then converted to fastq files using bam2fq (108).

2.9.2 Taxonomic classification and pathogen identification

Since an in-depth evaluation of DNA extraction methods for the detection of *L. monocytogenes* in foods was not an objective of this research, all Chapter 3 analyses were conducted using datasets generated with the Zymo Research Quick-DNA Fecal/Soil Microbe Miniprep Kit. This approach simplified the analysis and minimized the potential for confounding effects associated with variation in DNA extraction efficiency. Although multiple commercially available extraction kits were available, these methods are highly standardized, kit-based workflows that are widely used for microbial genomic applications. Furthermore, a preliminary pilot assessment conducted prior to this work did not identify substantial differences in

pathogen detection or downstream analytical performance among the extraction methods evaluated. While these findings were not formally published, they suggest that DNA extraction methodology was unlikely to be a major source of variation. Consequently, restricting the analysis to a single extraction method allowed the effects of the bioinformatic variables under investigation to be assessed more directly in Chapter 3.

To identify reads mapping to *L. monocytogenes* in Chapter 3, all datasets were subjected to taxonomic classification using the universal and FBP-specific reference databases across all three taxonomic classifiers with default parameters. To assess whether modifications to analysis parameters could improve classification precision and recall for the reference databases, the sequence threshold label (*k*) was adjusted from five to one with and without adjusting the minimum length of partial hits *(--min-hitlen*) parameter from a default of 22 bp to 31 bp for datasets classified with Centrifuge. For datasets classified with Kraken 2 or KrakenUniq, a confidence score threshold (*--confidence*) of 0.1 was applied. The proportion of *k*-mers mapping to a taxon is referred to as the confidence score. A user-defined confidence score (default = 0) determines whether a taxonomic classification is accepted, with only classifications exceeding the threshold being reported (82). All classification results were loaded into Pavian (v1.0) (114). To account for the heterogeneity in assigned taxonomy ID within a species, all reads classified as *L. monocytogenes* (taxID 1639), including children taxon (e.g. *L. monocytogenes* strains), were extracted from the respective fastq files using the *krakenuniq-extract-reads* script with the -t option (94). Extracted reads were aligned to the reference genome (HPB5415) using Bowtie 2 (v2.4.4) to verify the number of true positive and false positive reads (107). In Chapter 3, reads classified as *L. monocytogenes* and uniquely aligning to the reference genome with a mapping quality score of 42 using Bowtie 2 (v2.4.4) were categorized as true positive (TP), whereas reads that were classified as *L. monocytogenes* but did not uniquely align to the reference genome were defined as FP. Reads uniquely aligning to the *L. monocytogenes* reference genome in the raw datasets (ground truth) but not classified as *L. monocytogenes* were defined as FN.

To confirm the presence of the target pathogen within each quasi-metagenomic dataset in Chapter 4 and Chapter 5, an initial screening step was performed using RefSeq Masher *contains* (v0.1.2), which identifies NCBI RefSeq genomes represented in the input sequences by comparison against a precomputed database comprising 54,925 RefSeq genomes (115). Datasets were taxonomically classified using KrakenUniq (v0.6) and the foodborne pathogen (FBP)-specific reference database with default parameters. Confidence filtering, although generally recommended, was not applied because, at the time of the study, enabling confidence thresholds in KrakenUniq suppressed reporting of unique *k*-mer counts in

the output. As unique *k*-mer counts were central to downstream analyses and pathogen detection in this study, retention of this metric was prioritized over confidence-based filtering.

2.9.3 Pathogen-presence likelihood metric

In Chapter 4, raw *k*-mer counts obtained from KrakenUniq reports were normalized to sequencing depth to ensure valid comparisons across samples. Sequencing depth was defined as the total number of reads assigned to the root taxonomic node, representing the true read count per sample. Normalization was performed using unique *k*-mers per million reads (KPMR), which adjusts for differences in library size and reduces depth-related bias (94). To avoid unstable normalization at low read counts, samples with fewer than 500,000 root reads were excluded, which equated to approximately 10% of samples (Appendix 2.2). To account for differences in sequencing depth, the minimum number of root reads was decreased to 10,000 for the long-read quasi-metagenomic datasets generated in Chapter 5; only 12% (9/74) of samples fell below this threshold. Normalized unique *k*-mer values (KPMR) were then transformed using the log1p function, y = log(1+x), to reduce skewness and stabilize variability (86).

Robust *z*-scores were calculated as (x - median(x)) / (1.4826 x MAD(x)), where x represents the log-transformed KPMR value for a given taxon and MAD is the median absolute deviation of all taxa within the sample. This approach measures how much a taxon's normalized KPMR differs from the overall taxonomic profile of the sample. The MAD was multiplied by 1.4826 so that it approximates the standard deviation for normally distributed data, allowing the resulting scores to be interpreted similarly to conventional *z*-scores. Because the median and MAD are less sensitive to outliers than the mean and standard deviation, robust *z*-scores provide a stable measure of taxonomic enrichment in sparse and highly skewed metagenomic datasets (116,117). Positive values indicate taxa with higher-than-expected normalized KPMR values relative to the sample background, whereas negative values indicate lower-than-expected abundances. Consequently, taxa with large positive robust *z*-scores are more likely to represent meaningful biological signal rather than random variation or background noise.

To suppress background noise, taxa reported for each sample in the KrakenUniq output were required to meet minimum thresholds for both raw read counts and unique *k*-mer support before being included in downstream analysis. Specifically, taxa were required to have ≥1,000 raw reads, aligning with the minimum read threshold used in the PNC isolate-based WGS quality workflow (80) , and ≥1,000 unique *k*-mers, consistent with KrakenUniq recommendations for confident taxonomic assignment (94). After applying these filters, detected taxa were ranked according to unique KPMR to identify the most likely

pathogen in each sample. Unique KPMR was selected as the primary detection metric because it consistently distinguished true pathogen signal from background noise (Appendix 2.3). When the two highest-ranking taxa exhibited comparable signal strength (top-to-second ratio ≤ 1.2), a robust *z*-score was used as a secondary discriminator to prioritize the more strongly enriched taxon. The 1.2 threshold, corresponding to a 20% difference in unique KPMR, was derived empirically from the observed ratio distribution in the Chapter 4 datasets (Appendix 2.4). At this cutoff, fewer than 1% of samples fell within the tie range, indicating that magnitude-based ranking alone was sufficient for pathogen identification in most cases.

In Chapter 5, the highest KPMR value exceeded the second-highest value by a clear margin in all long-read quasi-metagenomic datasets that met the minimum sequencing depth requirement. None of the samples fell within the predefined tie threshold of 1.2 established in Chapter 4. As a result, the separation between competing taxa was sufficient for ranking, and the tie-resolution rule was not required.

Consistent with the close genomic relationship between *Shigella* spp. and *E. coli*, co-detection was frequently observed in *Shigella*-spiked samples in Chapter 4. Shared genomic content often resulted in *E. coli* ranking higher, even when clear *Shigella* signal was present. To reduce this bias, an override rule was introduced in Chapter 4: when *Shigella* unique KPMR reached at least 20% of the competing *E. coli* value, the highest-ranking *Shigella* species was designated as the most likely pathogen. Fewer than 8% of samples fell below this threshold (Appendix 2.5). As Chapter 5 focused exclusively on *Listeria* spp., this override rule was not required.

To assess whether sufficient genomic information was retained for strain-level resolution and outbreak detection, pathogen-specific reads corresponding to the true spike organism were extracted for each sample in Chapters 4 and 5. Reads were retrieved from the preprocessed FASTQ files using the *krakenuniq-extract-reads* script with the -t option, which extracts reads assigned to the specified taxID, including all descendant taxa within the taxonomy hierarchy (94). If *E. coli* or *Shigella* was identified as the top pathogen, the dataset(s) were further analyzed using the EC-specific or SH-specific reference database, respectively. Pathogen specific reads were then extracted using the results obtained from the species-specific reference database. Since KrakenUniq uses the lowest common ancestor (LCA) approach for classification, all pathogen reads were extracted at the genus level. Briefly, the LCA is a more conservative approach that tends to spread the taxonomic level of the classifications from the more specific to the more general to reduce the risk of false positive classifications (94).

2.9.4 Genome assembly and quality assessment

Extracted pathogen-specific reads were assembled using SPAdes (v3.15.3) for short-read data in Chapter 4 with assembly and error correction mode and options --*careful* and --*cov-cutoff* 1 for *Listeria* spp. datasets and --*cov-cutoff* auto for VTEC and *Shigella* spp. datasets (118). The assembler and corresponding parameters were specifically selected for analysis in Chapter 4 to mimic the pathogen-specific denovo assembly pipeline in the BioNumerics platform (bioMerieux, Durham, NC).

In Chapter 5, long-read datasets were assembled using flye (v2.9.3) with --*asm-coverage* 50 to account for the higher coverage (119). Long-read assemblies were subjected to one round of polishing using medaka (v1.11.3) and model (-m) r941_min_sup_g507 (120).

All assemblies were assessed for quality using quast (v5.0.2) (121). The completeness and contamination of the assemblies was also assessed using checkM (v1.2.2) (101). All quality metrics obtained for the datasets generated in Chapter 4 and Chapter 5 were compared to the quality metrics obtained for the corresponding single isolate (SI) reference strain. To note, single isolate (SI) reference strains were previously sequenced using both the Illumina MiSeq and ONT MinION sequencing technologies to generate hybrid reference assemblies using Unicycler with default parameters (v0.4.8.0) (122).

2.9.5 Pathogen verification

To characterize high-quality VTEC and *Shigella* assemblies, *in silico* speciation and serotyping were conducted using ECTyper (v1.0.0) in Chapter 4 (123). For the verification of *Listeria* datasets in Chapters 4 and 5, RefSeq Masher matches (v0.1.2) was employed to identify the closest NCBI RefSeq genome representative for each query sequence (115).

2.9.6 Outbreak analysis

Two fundamentally different, complementary analytical approaches implemented by the PNC network were used to perform outbreak detection, whole genome multi-locus sequencing typing (wgMLST) and single nucleotide variant phylogenomic (SNVPhyl) analysis.

The primary method, wgMLST is based on allelic differences, which are determined by assessing sequence variations in the coding regions of select genes (or “loci”) incorporated into pre-defined schema. This assessment was conducted using the BioNumerics (BN) platform (v7.6) for outbreak datasets only with PNC-specific settings (124). The *L. monocytogenes* schema contains 4,804 loci, while the *Escherichia*

coli/Shigella schema contains 17,380 loci. Extracted pathogen-specific reads were imported into BN and analyzed using two algorithms for allele identification, assembly-free and assembly- based, as previously described (124). Due to the limitations of BN, allele identification could only be performed on imported polished long-read data assemblies using the assembly-based method (Chapter 5).

The secondary method, SNVPhyl, is based on single nucleotide variants in the shared core genome, which represent regions that are present in all isolates and the reference (125). To align with PNC standards, the SNVPhyl pipeline (v1.1) was executed in Galaxy for each outbreak dataset. For short-read quasi-metagenomic datasets generated in Chapter 4, the SNV density filtering parameters were adjusted to a sliding window size of 20 bp with a density threshold of two SNVs. The single-isolate (SI) hybrid assemblies described in Section 2.8.4 were used as reference genomes. Since SNVPhyl was not designed for long-read sequence data, MINTyper (v.1.1.2) was used to identify single-nucleotide polymorphisms (SNPs) in the long-read datasets analyzed in Chapter 5 (126). For these analyses, the same SI reference genomes used in Chapter 4 were also applied in Chapter 5. In addition, MINTyper supports the analysis of mixed short- and long-read sequence data, enabling direct comparison between sequencing technologies.

2.9.7 Genome coverage

Extracted pathogen-specific reads were assessed for coverage against previously generated hybrid references as previously described in Section 2.10.5 (Chapter 4 and Chapter 5). However, GCF_002213585.1 (strain 08-6056) and GCF_002214045.1 (strain 10-0814) were already available as complete genomes in NCBI RefSeq. In Chapter 4, Bowtie 2 (v2.4.4) was used to generate alignments for short-read datasets, while minimap2 (v2.28) was used to generate alignments for long-read datasets. Sorted BAM files were evaluated using qualimap 2 (v2.2.2d) for percent reads mapped and mean coverage depth (107,113,127).

2.10 Statistical Analysis and Visualizations

To assess classifier performance, datasets generated in Chapter 3 were averaged across replicates prior to calculating precision, recall, and F1-scores, which are standard metrics commonly used in taxonomic classifier benchmarking studies (78,98,128). For this study, precision (TP / (TP+FP)) is defined as the proportion of samples identified as positive for *L. monocytogenes* that are truly positive, whereas recall (TP / (TP+FN)) is the proportion of *L. monocytogenes* positive samples that are correctly

identified by the method. The F1-score (2 x ((Precision x Recall) / (Precision + Recall))) is the harmonic mean of precision and recall and summarizes the overall performance of the classifier by balancing the proportion of positive *L. monocytogenes* detections that are correct (precision) with the proportion of true positive samples that are successfully detected (recall).

For Chapters 4 and 5, the Shapiro–Wilk test was used to assess normality prior to statistical analysis due to the small sample size of each pathogen group ($n < 30$; Section 2.1). For data meeting parametric assumptions in Chapter 4 and Chapter 5, paired *t*-tests were utilized to compare means, whereas the Wilcoxon signed-rank test was applied to non-parametric datasets (129). Statistical significance was defined by a *p*-value threshold of ≥ 0.05. All statistical computations were performed in RStudio (v2025.05.1+513) using the R programming language (v4.5.1) (130,131). Figures were also generated in RStudio using the R programming language and ggplot2 R package (v4.0.0) (Chapter 3, Chapter 4, Chapter 5) (130,132).

3.0 Chapter 3 - Optimization of bioinformatic pipelines and reference database architecture for the detection of *Listeria monocytogenes* in complex food matrices

3.1 Rationale

Chapter 3 serves as the foundational component of this thesis, establishing the bioinformatic framework required for high-resolution quasi-metagenomics. While metagenomic sequencing has emerged as a powerful tool for pathogen detection, its accuracy remains highly dependent on the architecture of the reference database and the specific parameters of the classification algorithm. The transition from isolate-based WGS to culture-independent methods requires careful evaluation to ensure reliable pathogen detection within the complex microbial backgrounds of food commodities.

A critical challenge addressed in this chapter is the selection of an appropriate reference database. Public health laboratories must frequently navigate the trade-off between universal databases, which encompass the full breadth of known genomic diversity, and taxon-specific (FBP-specific) databases, which are curated to include only relevant pathogens and common contaminants. While universal databases minimize the risk of missing unexpected organisms, they require substantial memory (RAM) and are prone to "noise" from distantly related environmental sequences. Conversely, taxon-specific databases offer significantly faster processing speeds and lower computational overhead, yet they may exacerbate false-positive classifications if the target pathogen shares conserved sequences with unrepresented organisms in the sample.

Using artificially contaminated ready-to-eat (RTE) meat and produce, this chapter systematically benchmarks three widely adopted *k*-mer-based classifiers: Centrifuge, Kraken 2, and KrakenUniq. By maintaining consistent reference sequences across both universal and FBP-specific database designs, this study examines the impact of database architecture and *k*-mer thresholds on classification precision. These findings define the essential "dry-laboratory" quality control measures needed to reduce false-positive signals attributed to complex food matrices. The systematic evaluation of classification performance in this chapter resulted in the selection of KrakenUniq as the primary taxonomic classifier and the foodborne pathogen–specific (FBP-specific) reference database as the optimized resource for taxonomic classification in subsequent chapters. These methodologies and validation benchmarks were peer-reviewed and published in BMC Genomics (https://doi.org/10.1186/s12864-023-09338-w) and provide the methodological foundation for the taxonomically complex investigations presented in Chapter 4 and Chapter 5.

3.2 Results

3.2.1 Provision of test datasets

An average of 1,025,280 paired-end (PE) reads were generated for RTE turkey deli meat datasets using the Qiagen QIAamp® Fast DNA Stool Mini Kit (Qiagen, Valencia, California), which was 8–11% higher than the average number of PE reads generated for datasets processed with the Qiagen DNeasy® PowerSoil Kit (Qiagen, Valencia, California) (average 948,890 PE reads) or the Zymo Research Quic*k*-DNA Fecal/Soil Microbe Miniprep Kit (Zymo Research, Irvine, CA) (average 923,429 PE reads). As for prepackaged spinach datasets, 6% more PE reads were generated when using the Qiagen DNeasy® PowerSoil Kit (average 964,671) compared to Qiagen QIAamp® Fast DNA Stool Mini Kit (average 902,795 PE reads), and 75% more compared to the Zymo Research Quic*k*-DNA Fecal/Soil Microbe Miniprep Kit (average 551,484 PE reads). Although the lowest number of PE reads was generated for datasets extracted with the Zymo Research Quic*k*-DNA Fecal/Soil Microbe Miniprep Kit across both foods types, the average proportion of reads classified as *L. monocytogenes* was highest for these datasets (https://doi.org/10.1186/s12864-023-09338-w; Additional file 1). This finding was likely attributed to the more efficient lysis of Gram-positive bacteria compared to the other commercial kits. Since an in-depth evaluation of the DNA extraction kits for the detection of *L. monocytogenes* in foods was not an objective of this pilot study, we conducted our evaluation using datasets extracted with the Zymo Research Quic*k*-DNA Fecal/Soil Microbe Miniprep Kit to simplify the analysis and eliminate a potential confounder associated with differences in DNA extraction efficiency.

3.2.2 Including an *in silico* eukaryotic DNA removal step can minimize false-positive classifications when using a taxon-specific reference database

Compared to the *L. monocytogenes* genome (~ 3.0 million bp), the turkey genome (*Melagris gallopavo*) is around 1000 times larger (~ 1.1 billion bp) (133), while the spinach genome (*Spinacea oleracea*) is around 100 times larger (~ 990 million bp) (134). Therefore, the primary food matrix can introduce a large quantity of "eukaryotic-derived" DNA that can obscure the detection of low-level pathogens. As expected, the proportion of FP reads in RTE turkey deli meat raw datasets varied across metagenomic classifiers but was generally higher for low spike-in raw datasets than the medium and high spike-in raw datasets (Figure 3.1, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2). Across both reference databases, 56 FP reads were detected in RTE turkey deli meat raw datasets classified with Centrifuge compared to ≤ 11 FP reads when classified with Kraken 2 and ≤ 1 FP read when classified with KrakenUniq (Figure 3.1, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2). Including a data quality filtering step to remove adaptor contamination, low-quality bases, and short reads using fastp did not substantially impact the number of FP classifications for either reference database. However, the preprocessing of raw datasets is a prerequisite step to improve data quality for downstream analysis and is highly recommended. On average, FP reads were reduced by < 8% (≤ 5 reads) in RTE turkey deli meat datasets classified with either Centrifuge or Kraken 2 for both the FBP-specific and universal reference databases (Figure 3.1, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2). A single FP classification was resolved in the medium spike-in RTE turkey deli meat dataset when KrakenUniq was used. Including an *in silico* eukaryotic sequence removal step to low-quality read filtered datasets reduced FP reads by 90% or more (~ 50 reads) in RTE turkey deli meat datasets across both reference databases when classified with either Centrifuge or Kraken 2 (Figure 3.1, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2). The FP reads were investigated using the Basic Local Alignment Search Tool (BLAST) and the National Center for Biotechnology Information (NCBI) nucleotide (nt) database (135,136). BLAST analysis revealed that FP reads predominately aligned to the turkey genome (*M. gallopavo*; GCF_000146605.3) or a close descendant thereof; this was not unexpected since, on average, > 96% of the total reads were eukaryotic-derived (https://doi.org/10.1186/s12864-023-09338-w; Additional file 3).

Reference database contamination due to sequence mislabeling can lead to misclassifications in metagenomics analyses. Therefore, to rule out cross-kingdom contamination between the turkey genome and the *L. monocytogenes* genomes included in the reference databases, conterminator was used to perform an all-against-all sequence comparison (102). The conterminator analysis did not identify

contamination between the two taxa (https://doi.org/10.1186/s12864-023-09338-w; Additional file 4). This finding was not surprising given recent efforts to remove contaminated sequences from the turkey genome, which was once considered the most contaminated genome in the Reference Sequence Database (RefSeq) (102).

Like the RTE turkey deli meat datasets, the proportion of FP classifications in prepackaged spinach datasets varied across metagenomic classifiers and was consistently higher for low spike-in datasets (Figure 3.2, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2). A higher proportion of FP classifications was observed in datasets classified with the FBP-specific reference database compared to the universal reference database, particularly when using Centrifuge or Kraken 2. On average, 30 FP reads were identified in prepackaged spinach datasets when classified with Centrifuge or Kraken 2 using the FBP-specific reference database compared to ≤ 10 FP reads for datasets classified with Centrifuge and ≤ 3 FP reads for datasets classified with Kraken 2 using the universal reference database (Figure. 3.2, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2). Regardless of the reference database used, fewer than five FP reads were detected in unfiltered prepackaged spinach datasets when classified with KrakenUniq.

Filtering low-quality reads with fastp led to a maximum 21% reduction (maximum 7 reads) in FP reads for prepackaged spinach datasets classified with Centrifuge and the FBP-specific reference database compared to a 16% reduction (maximum 2 reads) when using the universal reference database (Figure 3.2, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2). Across the same datasets, a 10% reduction in FP reads (3 reads) was observed when using Kraken 2 and the FBP-specific reference database. No change in FP reads occurred in the medium and low spike-in datasets classified with Kraken 2 and the universal database, whereas a 33% decrease in FP reads occurred in the high spike-in dataset (1 read). When classified with KrakenUniq, FP reads were reduced by 59% (2 reads) in the high spike-in dataset across both reference databases (Figure 3.2, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2). A more substantial decrease in FP reads was observed when the taxonomic classification was performed on preprocessed (i.e., low-quality read filtered and *in silico* removal of eukaryotic DNA) prepackaged spinach datasets, particularly when using Kraken 2 and the FBP-specific reference database. In these datasets, FP reads were reduced by an average of 85% (23 reads) compared to 21% (2 reads) when classified with Centrifuge across both reference databases. The number of TP reads did not change for datasets classified with Kraken 2 or KrakenUniq and was consistent across reference databases (Figure 3.2, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2).

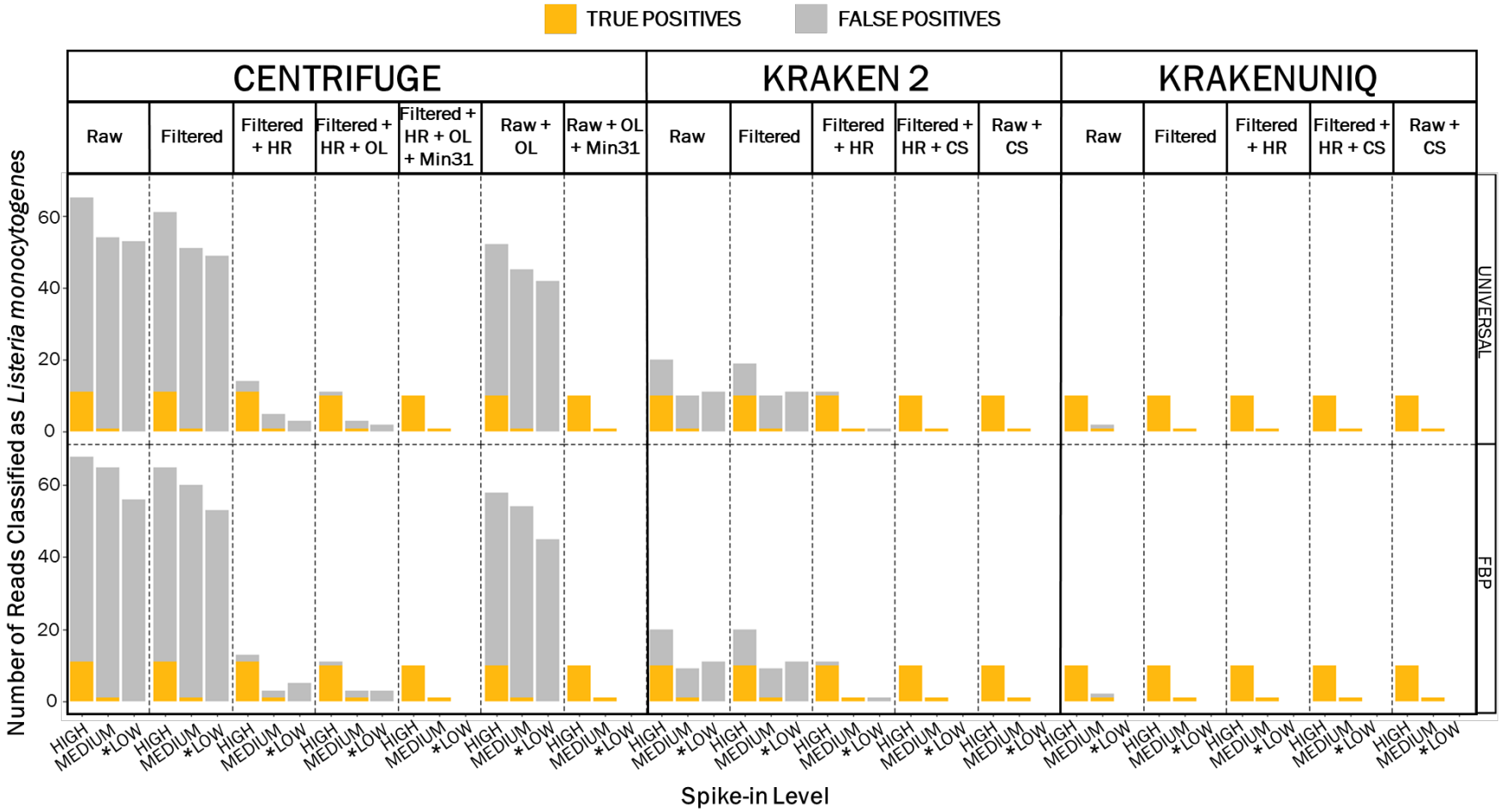


Figure 3.1. *Listeria monocytogenes* read recovery in artificially contaminated RTE turkey deli meat. Turkey samples were spiked at three levels (10^4, 10^3, and 10^2 CFU/mL) with gDNA extracted in duplicate using the Zymo Research Quick-DNA Fecal/Soil Microbe Miniprep Kit. Values represent read counts averaged across replicates, classifiers (Centrifuge, Kraken 2, KrakenUniq), and reference databases, including a taxon-agnostic "UNIVERSAL" RefSeq database and a "FBP" foodborne pathogen-specific database. Data processing treatments included raw datasets (Raw), quality filtering via fastp (Filtered), and *in silico* eukaryotic removal via Bowtie 2 (HR). Algorithmic modifications consisted of adjusting the Centrifuge sequence threshold to *k*=1 (OL) to mimic Kraken-style classification, increasing the minimum partial hit length to 31 bp (Min31), and applying a 0.1 confidence score threshold (CS) for Kraken 2 and KrakenUniq. An asterisk indicates that no true positive (TP) reads were detected for that specific parameter combination.

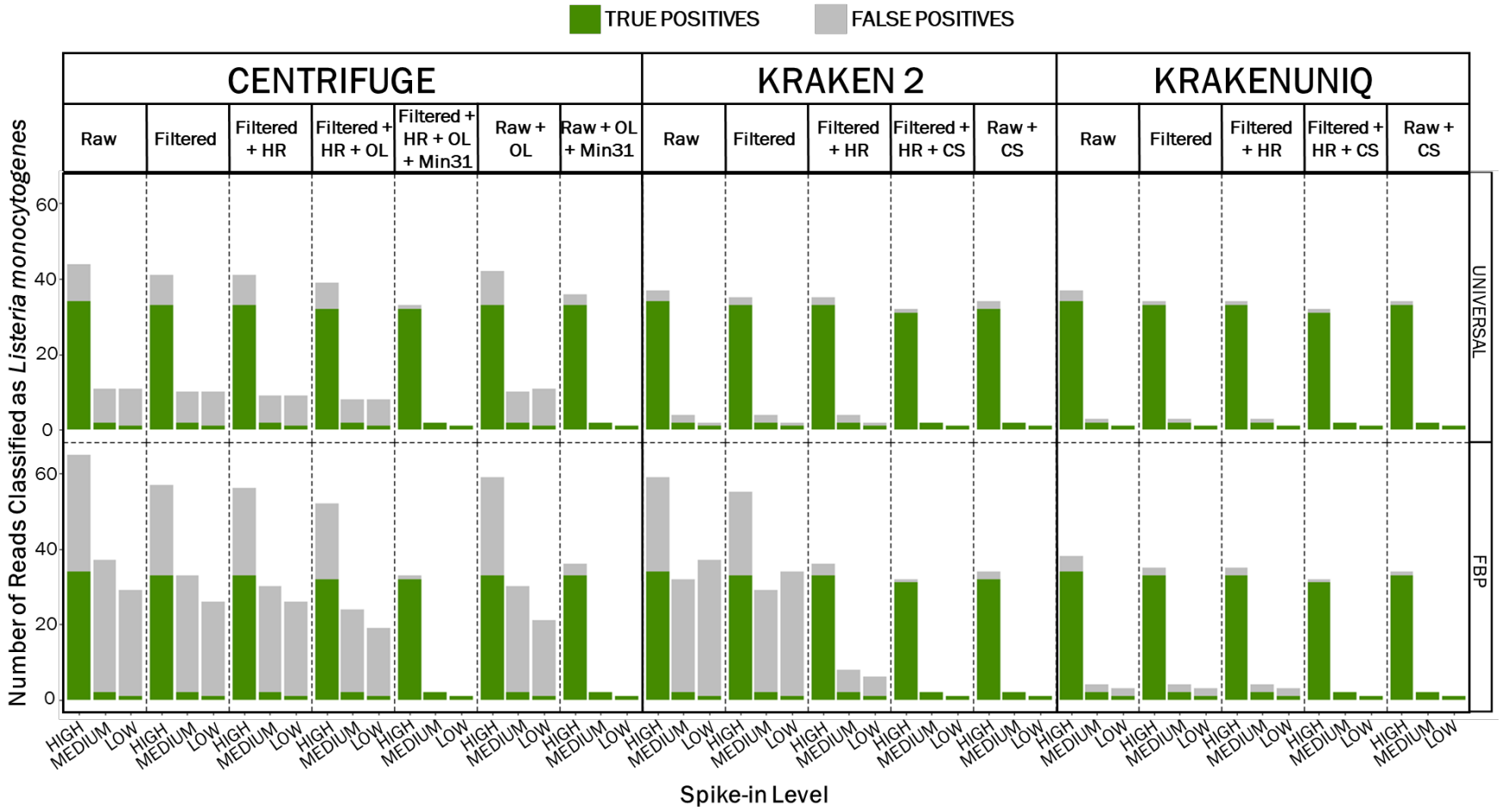


Figure 3.2. Average *Listeria monocytogenes* read recovery in artificially contaminated prepackaged spinach. Prepackaged spinach samples were spiked at three levels (10^4, 10^3, and 10^2 CFU/mL) with gDNA extracted in duplicate using the Zymo Research Quick-DNA Fecal/Soil Microbe Miniprep Kit. Values represent read counts averaged across replicates, classifiers (Centrifuge, Kraken 2, KrakenUniq), and reference databases, including a taxon-agnostic "UNIVERSAL" RefSeq database and a "FBP" foodborne pathogen-specific database. Data processing treatments included raw datasets (Raw), quality filtering via fastp (Filtered), and *in silico* eukaryotic removal via Bowtie 2 (HR). Algorithmic modifications consisted of adjusting the Centrifuge sequence threshold to *k*=1 (OL) to mimic Kraken-style classification, increasing the minimum partial hit length to 31 bp (Min31), and applying a 0.1 confidence score threshold (CS) for Kraken 2 and KrakenUniq.

Unlike the RTE turkey deli meat datasets, which consisted almost entirely of eukaryotic DNA (> 96%), less than 7% of the total sequenced PE reads aligned to the spinach genome (https://doi.org/10.1186/s12864-023-09338-w; Additional file 3). Therefore, it was unsurprising that including a eukaryotic-removal step generally had a lower impact on the number of FP reads for these datasets. BLAST analysis against the *nt* database revealed that FP reads detected in prepackaged spinach datasets classified with the universal reference database primarily aligned to the lettuce genome (*Lactuca sativa*), which we suspect was better represented in the NCBI nt database compared to the spinach genome (*Spinacia oleracea)*. Though several FP reads detected in prepackaged spinach datasets classified with the FBP-specific reference database also aligned to the lettuce genome, the majority of FP reads aligned to the soil-ubiquitous genera *Pseudomonas*, which is known to colonize spinach (137,138). When using the universal reference database, ≥ 50% of the total reads were classified as *Pseudomonas* spp. in the prepackaged spinach datasets across all classifiers. Since these genera were not included in the FBP-specific reference database, true *Pseudomonas* spp. reads were associated with an identical or similar region also present in a distantly related taxon in the database (i.e., *L. monocytogenes*), leading to misclassification (79).

Using conterminator, cross-kingdom contamination was predicted between the *L. monocytogenes* reference genomes and the lettuce genome (GCF_002870075.2), specifically, the chloroplast genome (NC_007578.1) and ribosomal RNA (rRNA) regions in *L. monocytogenes* (https://doi.org/10.1186/s12864-023-09338-w; Additional file 4). Due to the high similarity between plastid and bacterial rRNAs, erroneous integration of bacterial sequences into chloroplast ribosomal RNA is not uncommon (139).

Preprocessing the data improved classification precision across both reference databases (Figure 3.3, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2). For RTE turkey deli meat datasets, the precision increased from < 0.52 in raw datasets to > 0.95 for preprocessed datasets when using Kraken 2 and was equivalent across reference databases. Data preprocessing did not affect the precision for high and low spike-in RTE turkey deli meat datasets classified with KrakenUniq as it was already optimal (1.0), but it did improve the precision from 0.67 to 1.0 for the medium spike-in RTE turkey deli meat dataset (Figure 3.3, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2). This finding was consistent across reference databases. Although a noticeable improvement in precision was observed, preprocessed RTE turkey deli meat datasets classified with Centrifuge still had the lowest precision (< 0.85) across both reference databases. When using the universal reference database, the precision increased from 0.92 to

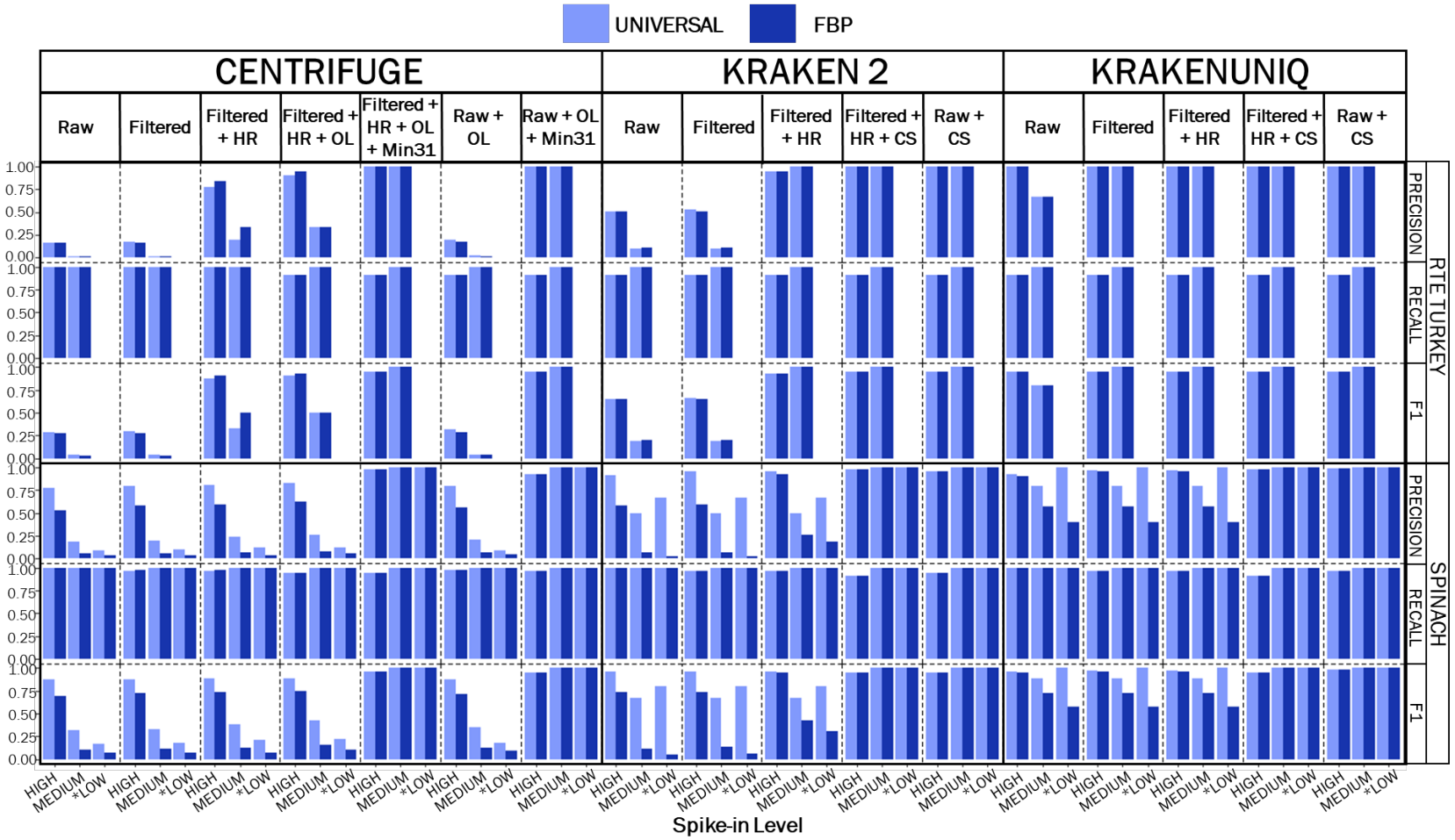


Figure 3.3. Comparison of precision, recall and F1 score across reference databases and classifiers. Samples were artificially contaminated at three levels (high: 10^4; medium: 10^3; low: 10^2 CFU/mL) and extracted in duplicate using the Zymo Quick-DNA Fecal/Soil Kit. Metrics represent averages across replicates for Raw (no processing), Filtered (fastp quality-trimmed), and Filtered + HR (fastp trimmed plus Bowtie 2 eukaryotic-removed) datasets. Reference databases include "UNIVERSAL" (taxon-agnostic RefSeq) and "FBP" (foodborne pathogen-specific). *No *L. monocytogenes* reads were detected in low spike-in RTE turkey deli meat datasets.

0.96 for high spike-in prepackaged spinach datasets classified with Kraken 2 and from 0.93 to 0.97 when classified with KrakenUniq. No change in precision was observed in medium and low spike-in datasets across the same classifiers and universal reference database (Figure 3.3, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2).

When using the FBP-specific reference database, a more substantial increase in precision from 0.58 to 0.93 was observed for preprocessed prepackaged spinach datasets classified with Kraken 2 compared to a minor increase in precision from 0.91 to 0.96 when using KrakenUniq. No change in precision was observed across the same datasets when classified with Centrifuge, regardless of the reference database. Recall remained high across reference databases for all classifiers (> 0.90), and the F1 score was highest for datasets classified with the universal reference database (Figure 3.3, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2).

3.2.3 Metagenomic classifier and analysis parameters influence the number of false-positive classifications when using a taxon-specific reference database

Unlike Kraken 2 and KrakenUniq, which assign a single taxonomic label per read using the lowest common ancestor approach (LCA), Centrifuge, by default, assigns up to five taxonomic labels per read using the Burrows-Wheeler transform (BWT) and Ferragina-Mazini (FM) index (104). To emulate the behavior of Kraken 2 and KrakenUniq, Centrifuge's sequence threshold label (*k*) was adjusted from five (*k*=5) to one (*k* = 1) to assign a single taxonomic label per read using the LCA (104). After applying this parameter to low-quality read filtered and *in silico* eukaryotic-removed datasets, a minor improvement in Centrifuge's FP reads was observed. For high spike-in RTE turkey deli meat datasets classified with Centrifuge and *k* = 1, FP reads decreased by 33–100% (≤ 3 reads) across both reference databases and the precision increased to > 0.75 (Figure 3.3, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2). Unfortunately, the precision for medium and low spike-in RTE turkey deli meat datasets remained low (≤ 0.33). For prepackaged spinach datasets classified with Centrifuge (*k* = 1) and the universal reference database, FP reads decreased by 15% or less (1 read) compared to ≤ 27% (≤ 7 reads) when using the FBP-specific reference database (Figure 3.2, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2). As a result, the precision did not noticeably improve and was low (< 0.65) across both reference databases for nearly all prepackaged spinach datasets except for the high spike-in dataset, where the precision was approximately 0.83 (Figure 3.3, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2). The recall decreased slightly in high-spike in datasets for both RTE turkey deli meat (0.91) and

prepackaged spinach (0.94) datasets classified with Centrifuge and the previously described parameters but remained unchanged in medium and low spike-in datasets across both sample types (Figure 3.3, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2).

In contrast to Kraken 2 and KrakenUniq, which use a single default *k*-mer length of 35 bp and 31 bp, respectively, Centrifuge makes use of both large (i.e., ≥ 31 bp) and small (i.e., 20–25 bp) *k*-mers to achieve a more desirable trade-off between sensitivity and precision (104). Since Centrifuge can assign taxonomic identity based on exact *k*-mer matches of at least 22 bp, we suspected the low precision was attributed to these smaller *k*-mer matches. To confirm our suspicions, we adjusted the minimum length of partial hits *(--min-hitlen*) parameter from a default of 22 bp to 31 bp and re-analyzed the data with $k = 1$ to assign a single taxonomic label per read using the LCA approach. Using the modified Centrifuge parameters, the precision (≥ 0.98) and recall (> 0.90) were found to be equivalent across reference databases and nearly identical to Kraken 2 and KrakenUniq (Figure 3.3, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2). Even when applying the same modified Centrifuge parameters to raw datasets, high precision (> 0.92) and recall (> 0.90) were still achieved for both reference databases (Figure 3.3, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2).

Different from KrakenUniq, where all *k*-mer information is indexed in a large reference database, Kraken 2 only indexes a small fraction of this information in the form of minimizers, which is a representative sequence of a group of highly similar *k*-mers (95). Since less information is stored in Kraken-2-indexed databases, more FP classifications (although minimal) are possible and were evident in this study. To determine whether classification precision could be further improved for the FBP-specific reference database, we adjusted the confidence score threshold (*--confidence*) from a default of zero to 0.1. This means that at least 10% of the read's *k*-mer evidence must support the lowest taxonomic rank assigned; otherwise, the read is assigned to a higher taxonomic rank meeting the specified *k*-mer threshold (95). In this study, a threshold of 0.1 was sufficient to resolve nearly all FP classifications using Kraken 2 or KrakenUniq and increased classification precision (≥ 0.98) across both reference databases (Figure 3.3, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2). A single FP read remained in the high spike-in prepackaged spinach dataset but was resolved by increasing the confidence score threshold to 0.2. When a confidence score threshold of 0.1 was applied to raw datasets, a high level of precision was also attained (> 0.98), albeit slightly lower for datasets classified with Kraken 2 (≥ 0.95) (Figure 3.3, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2).

While adjusting the analysis parameters improved classification precision across reference databases and classifiers, a slight loss in sensitivity and recall was observed. On average, ≤ 6% of TP classifications (≤ 2 reads) were lost to read reclassifications to a higher taxonomic rank (e.g., *Listeria* species or higher) because of insufficient *k*-mer evidence (confidence threshold of 0.1) to support a taxonomic label of *L. monocytogenes*. The F1 score remained high (≥ 0.95) and was equivalent across reference databases for each classifier (Figure 3.3, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2).

3.2.4 Detection of *Listeria monocytogenes* in RTE turkey deli meat and prepackaged spinach is challenging without culture enrichment

Overall, the number of TP reads was equivalent across reference databases for each classifier (Figures 3.1, Figure 3.2, https://doi.org/10.1186/s12864-023-09338-w; Additional file 2). An average of ten TP reads were detected in high spike-in RTE turkey deli meat datasets compared to a single read in medium spike-in and no reads in low spike-in datasets. Although slightly higher, an average of 31 TP reads were detected in high spike-in prepackaged spinach datasets compared to two reads and one read in medium and low spike-in datasets, respectively. These findings highlight the need for culture enrichment, the first step in pathogen detection and recovery from foods, to increase the target pathogen to levels sufficient for food safety monitoring and outbreak response activities, especially in foods with low contamination levels. Interestingly, recent evidence suggests that using read counts for pathogen detection can be misleading, particularly for detecting low pathogen levels (94).

To address this concern, the developers of Kraken 2 recently incorporated distinct-counting estimation of minimizers, a feature leveraged from KrakenUniq (described as unique *k*-mers) that aims to improve pathogen detection accuracy (140). In essence, distinct *k*-minimizers are used as a proxy for breadth of genome coverage. For example, taxonomic classifications with a higher number of distinct *k*-minimizers indicate the reads are more evenly distributed across the genome, suggesting the pathogen is likely present. In contrast, a low number of distinct *k*-minimizers suggests the reads are concentrated in a single or very few locations in the genome and likely a false-positive identification (94). Since accurate detection of *L. monocytogenes* represents an enormous public benefit and < 0.1% of the total sequenced reads were classified as *L. mono*cytogenes in our study datasets, we wanted to explore the usefulness of distinct *k*-minimizers and how they compared between reference databases for preprocessed datasets. We

specifically focused on distinct *k*-minimizers reported by Kraken 2 as the number of unique *k*-mers is absent from the KrakenUniq report file when confidence scoring is applied and cannot be assessed.

In our assessment of Kraken 2 distinct *k*-minimizer metric, approximately 775 *k*-min were identified in the high spike-in RTE turkey deli meat dataset compared to over 2,300 *k*-min in the high spike-in prepackaged spinach dataset. This was consistent across reference databases. In the medium and low spike-in datasets, less than 160 *k*-min were detected across both sample types (https://doi.org/10.1186/s12864-023-09338-w; Additional file 5). While the application of distinct *k*-minimizers for pathogen detection is promising, this feature is experimental and requires further validation with well-defined negative controls to determine the minimum number of distinct *k*-minimizers, corresponding to an empirical threshold for genome coverage breadth, necessary to confidently support a positive pathogen detection result (141).

3.2.5 Fewer computational resources are required for classification using a taxon-specific reference database

On average, memory usage and runtime were lower when using the FBP-specific database than the universal database, especially when using Kraken 2 for taxonomic classification (Figure 3.4). Performing taxonomic classification on the same dataset immediately following the first run improved runtime across both reference databases for all classification software (data not shown). This finding was consistent with a previous study conducted by Ye *et al.* and is attributed to efficient database caching, which maps the database files into memory (78). Adjusting the analysis parameters had a negligible impact on memory usage and runtime and, therefore, was not reported.

3.3 Discussion

Reference databases perform well when taxa in a dataset are genetically distinct from one other and genetically similar to genomes in the reference database (142). However, when there is insufficient diversity in the reference database, especially in the case of taxon-specific databases, it can lead to a considerable amount of non-specific read classification, which can be further exacerbated if even a small fraction of contamination is present in the reference genomes (71,143). This problem is particularly acute when using short reads (<1000 bp). Several studies have emphasized the presence of source eukaryotic DNA in genomes deposited in large-scale sequence repositories as contaminants, including RefSeq and GenBank, which are two key resources used for building reference databases (71,73–76). If not dealt with

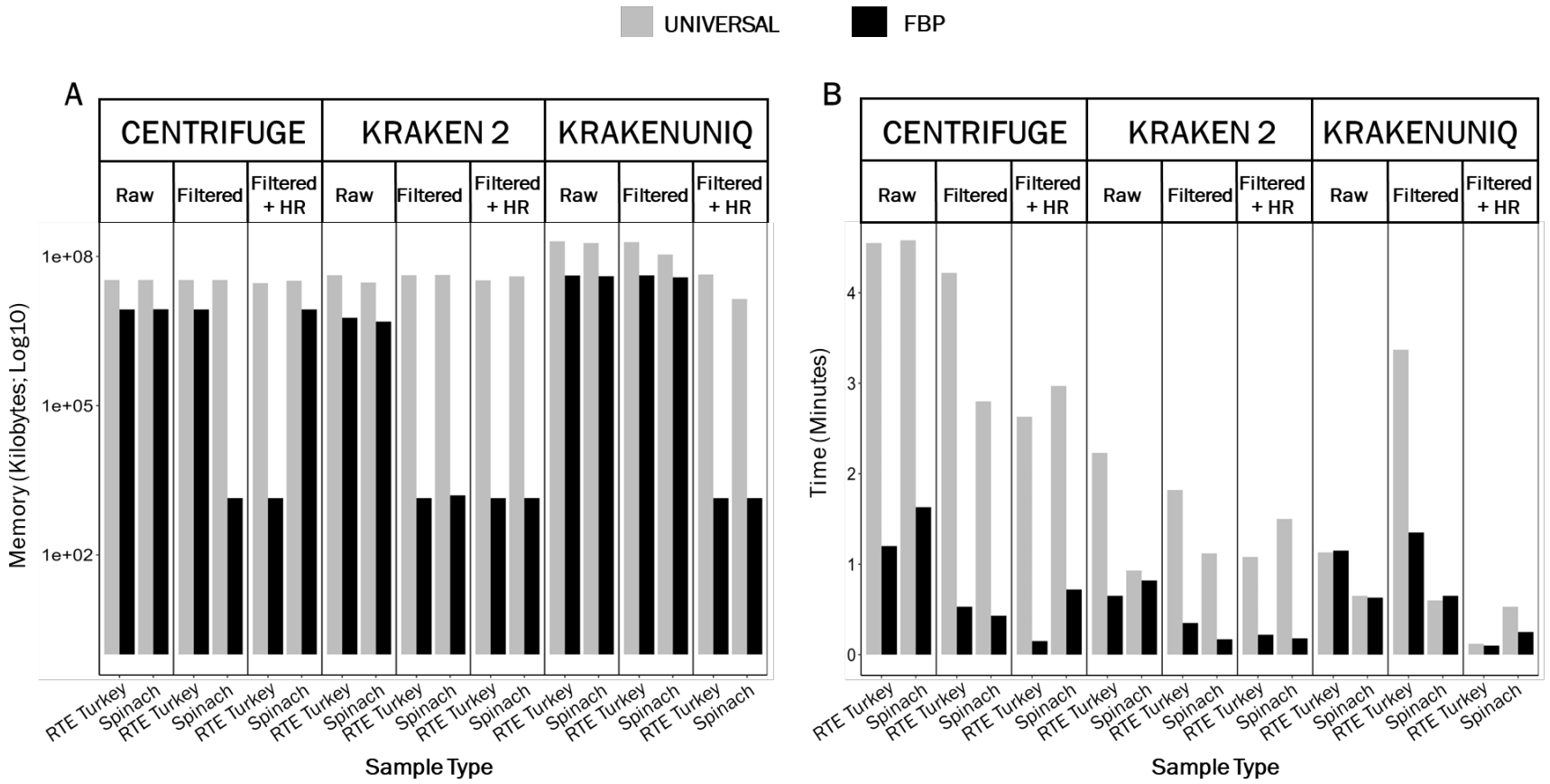


Figure 3.4. Comparison of computational resource requirements for taxonomic classification. Computational performance was evaluated based on [A] peak memory usage (kilobytes) and [B] processing time (minutes) using 8 cores to analyze RTE turkey and spinach datasets spiked with 10^4 CFU/mL of *L. monocytogenes*. Metrics were compared across three data processing stages: Raw (unfiltered), fastp (quality-filtered), and fastp + HR (Bowtie 2 eukaryotic-removed). Total read counts per matrix are provided for each stage to correlate sequence depth with compute resource consumption. Classification was performed using both a "UNIVERSAL" (taxon-agnostic) and "FBP" (foodborne pathogen-specific) reference database. *KrakenUniq databases were preloaded via the *--preload* option prior to classification.

appropriately, eukaryotic DNA can become particularly challenging for accurate pathogen detection, especially when only a limited fraction of the DNA within a sample represents the pathogen of interest, as is often the case with contaminated foods. In this study, we show that *in silico* removal of eukaryotic DNA, and to a lesser extent, low-quality read filtering was valuable for improving classification precision when conducting metagenomics-based pathogen detection on datasets with elevated eukaryotic content using a FBP-specific reference, such as RTE turkey deli meat. Despite equivalent spike-in volumes across both sample types, we presume the high eukaryotic DNA content coupled with a lower microbial load in the RTE turkey deli meat datasets led to fewer TP classifications across both reference databases compared to prepackaged spinach datasets. Eukaryotic DNA contamination is not unique to foods and is also problematic for pathogen detection in clinical specimens (66,67,144). Fortunately, eukaryotic-read removal *in silico* can improve pathogen detection, as demonstrated in this study. However, unwanted sequencing of eukaryotic reads incurs costly and time-consuming computational eukaryotic-read subtraction (145). To overcome this challenge, several commercial-based and in-house wet-laboratory methods have been developed and evaluated for eukaryotic DNA depletion in various clinical specimens, with some capable of removing up to 99.9% of eukaryotic DNA (66,67,144). Although promising, Ganda *et al.* demonstrated that these methods are not guaranteed to work with all sample types, particularly foods, and must be validated according to the pathogen and matrix (65).

In this study, quasi-metagenomic based pathogen detection of *L. monocytogenes* was more reliable in foods containing higher pathogen levels approaching the lower limit of detection of traditional, non-enrichment culture-based methods (~ 10^4 to 10^5 CFU/ml) in food (146). While a 48-h enrichment period is traditionally applied for detecting and isolating *L. monocytogenes* in foods, evidence suggests that direct sequencing of overnight primary enrichment cultures (quasi-metagenomics) can generate sufficient genomic information for source tracking of *L. monocytogenes* (58,60). If properly optimized to balance the cost (time) with the benefit (sufficient genomic data), such an approach could speed up and simplify workflows for FBP surveillance and outbreak detection, especially with a taxon-specific reference database.

The diagnostic value and impending role of metagenomics as a surveillance and outbreak detection tool have led to a drastic increase in classification tools and, correspondingly, benchmarking studies (78,98,128,147–149). While valuable, it is important to acknowledge that most of these studies were conducted using existing pre-computed default reference databases. Until recently, they did not consider using reference databases with consistent references and taxonomy across classifiers (78). Because of this, it is difficult to say whether differences in classification performance were attributed to the

reference database or the classification algorithm. Therefore, to eliminate any confounding effects that may arise from differences in default database composition, using reference databases with consistent reference sequences and taxonomy across classifiers is essential when conducting a comparative analysis—this was a major strength of this study.

Another common trend of benchmarking studies is the tendency to treat bioinformatics software as "black box" devices, wherein users commonly and blindly apply default parameters. However, careful consideration of analysis parameters in the context of the research question should be considered, as default parameters do not always provide the most optimal result. By design, Centrifuge, in its default state, makes use of both large (≥ 31 bp) and small (22 bp) *k*-mers to achieve a more desirable trade-off between sensitivity and precision. Short *k*-mer matches can be problematic when the genomes in the dataset do not have a close genetic match in the reference database, as in the case of a taxon-specific reference database. This is because shorter *k*-mers are less likely to be unique to a specific taxon; thus, spurious hits to multiple genomes are more likely to ensue. When classified with Centrifuge, nearly all FP classifications identified in preprocessed datasets generated from artificially contaminated prepackaged spinach samples arose from short *k*-mer matches. This scenario also explains why highly abundant bacterial sequences in the prepackaged spinach datasets (i.e., *Pseudomonas* spp.) contributed to a substantial amount of FP classifications and could not be resolved even when the sequence label threshold was adjusted from the default of five down to one. Only when the minimum length of partial hits was adjusted from 22 to 31 bp was an equivalent precision to Kraken 2 and KrakenUniq achieved for the FBP-specific reference database. Likewise, applying a confidence score threshold of 0.1 to datasets classified with Kraken 2 or KrakenUniq was necessary to achieve high classification precision when using the FBP-specific reference database (≥ 0.98). Interestingly, when the aforementioned analysis parameters were applied to the raw datasets classified with the FBP-specific reference database, high classification precision (≥ 0.92) and recall (> 0.90) were still achieved across all classifiers, though slightly lower than preprocessed datasets. This study's findings indicate that removing non-informative reads and adjusting analysis parameters prior to metagenomics-based pathogen detection analysis is beneficial when using a taxon-specific reference database and will reduce FP classifications and improve the true pathogen signal. Since optimization of the confidence score threshold was not conducted in this study, robust parameter testing is necessary to determine the most suitable threshold when using a taxon-specific reference database.

Ultimately, the classification approach implemented by Kraken 2 and KrakenUniq helped drastically reduce the risk of FP classifications by assigning shared sequences to the LCA among the set of matching

taxa. While the LCA approach can be extremely useful for resolving misclassifications, the approach also tends to spread the taxonomic level of the classifications from the more specific to the more general. Therefore, as public sequence repositories grow (i.e., RefSeq), it will be important to continuously re-evaluate how classification performance is impacted by changes in the distribution of reference genomes per taxon (150). Although this study focused on three popular metagenomic classifiers that can generate custom reference databases, many classification tools offer similar features that were not considered but can be assessed using this study as the experimental model.

Speed and memory requirements are often critical factors in analyzing large-scale datasets, especially when conducting metagenomics-based pathogen detection with large reference databases, which can require a significant amount of computational resources (i.e., RAM) depending on the classification software (78). A recent survey of laboratories in low and middle-income countries participating in PulseNet International, a global network comprising 88 countries to track foodborne diseases, found that only 28% of laboratories have access to local high-performance computing (35). This limited access highly restricts the type of data analysis that can be conducted in those laboratories. We demonstrate that equivalent precision can be attained using a smaller reference database when the appropriate quality control measures, classification software, and analysis parameters are applied. Although the highest level of precision and recall for metagenomic-based detection of *L. monocytogenes* was achieved when using KrakenUniq across both reference databases, the most optimal trade-off between classification performance and computational efficiency was achieved when using Kraken 2 and the FBP-specific reference database. Overall, this approach is less computationally demanding and more attainable for the broader scientific and food safety communities. Unfortunately, the advantages of a lower memory footprint offered by the smaller Kraken 2 reference database come at the cost of slightly lower classification performance (95).

At a minimum, applying a confidence-scoring threshold can improve classification precision when using Kraken 2. However, in scenarios that require very high precision, for example, in an outbreak response, where even a few FP can be detrimental, KrakenUniq may be preferred over Kraken 2 (151). As a result, recent improvements to the KrakenUniq software now enable "database chunking" (151). By loading the reference database in chunks according to the available memory, this feature enables users to perform taxonomic classification using KrakenUniq on virtually any modern computer. However, the trade-off with "database chunking" is the much slower classification speeds, especially for large datasets (151).

A taxon-specific reference database enables reliable metagenomic detection of *L. monocytogenes* in RTE turkey deli meat and prepackaged spinach when supported by appropriate quality control procedures, classification tools, and analytical parameters. This strategy reduces computational requirements, potentially expanding accessibility and supporting broader adoption of metagenomics-based pathogen detection for food safety, surveillance, and outbreak investigations in Canada.

4.0 Chapter 4 - Validation of a quasi-metagenomic workflow for the rapid detection and subtyping of priority foodborne pathogens from enriched food matrices

4.1 Rationale

The bioinformatic optimizations established in Chapter 3 provided the computational foundation for the quasi-metagenomic workflow evaluated in Chapter 4. Benchmarking identified KrakenUniq in combination with the foodborne pathogen (FBP)-specific reference database as the most suitable taxonomic classifier and reference database combination for downstream analyses. Under the optimized workflow, KrakenUniq achieved pooled performance metrics of up to 98.9% precision, 95.9% recall, and a 97.4% F1-score when confidence filtering was applied, demonstrating highly accurate *L. monocytogenes* detection. Although confidence filtering was not implemented in the present chapter because, at the time of this study, it suppressed reporting of unique *k*-mer counts required for pathogen-presence likelihood analyses, the optimized preprocessing workflow maintained excellent performance without confidence filtering, achieving 96.9% precision, 95.9% recall, and a 96.4% F1-score. Collectively, these findings supported the selection of KrakenUniq and the FBP-specific reference database as the analytical foundation for the quasi-metagenomic workflow. Building on this foundation, Chapter 4 investigated whether the optimized quasi-metagenomic workflow could support strain-level characterization and outbreak detection within the PNC surveillance network.

The practical value of quasi-metagenomics for public health surveillance depends on its ability to perform reliably across a range of bacterial species and complex food matrices. Accordingly, Chapter 4 evaluates the workflow using three priority foodborne pathogens in Canada: *L. monocytogenes*, VTEC O157:H7, and *Shigella sonnei*. Additional *Listeria*, VTEC, and *Shigella* species were incorporated to ensure accurate differentiation among closely related taxa.

A persistent challenge in quasi-metagenomic and metagenomic datasets is the high proportion of eukaryotic DNA, which can obscure pathogen signal and often requires deep sequencing to achieve sufficient microbial coverage. To address this limitation, Chapter 4 evaluates a saponin-based eukaryotic

DNA depletion method as a wet-laboratory approach to complement the in-silico filtering strategies developed in Chapter 3. While saponin-based depletion has proven effective in clinical specimens, particularly those with high host DNA content (e.g., cerebrospinal fluid), its suitability for food matrices remains largely unexplored. In this study, depleted and non-depleted samples were compared across several food matrices, including delicatessen meats, ground beef, and raw produce, to assess how removal of eukaryotic DNA influences pathogen read recovery and the quality of downstream genome assemblies.

Chapter 4 also introduces the pathogen-presence likelihood metric to improve consistency in pathogen detection for quasi-metagenomic datasets. Detection criteria based solely on read counts or raw *k*-mer counts can be strongly influenced by sequencing depth, background variability, and classifier-specific behavior. The likelihood framework was therefore developed to incorporate multiple features of the sequencing signal, including depth-normalized unique *k*-mer evidence and within-sample enrichment patterns (*z*-score), into an integrated decision framework. This approach provides a more stable representation of pathogen-specific signal and improves interpretation under conditions of uneven coverage or variable sequencing depth.

In addition to pathogen detection, this chapter assesses whether quasi-metagenomic data can provide the genomic resolution required for outbreak investigations using established PNC analytical methods, including whole-genome multilocus sequence typing (wgMLST) and single nucleotide variant (SNV) phylogenomics. By benchmarking these results against conventional isolate-derived WGS, Chapter 4 assesses whether quasi-metagenomics can generate the high-resolution genomic evidence required for outbreak detection while reducing the time needed to obtain actionable results.

4.2 Results

4.2.1 Data quality

General data quality metrics for all datasets are summarized in Figure 4.1. When saponin-based eukaryotic DNA depletion (SD) was applied, the average DNA concentration was 25.49 ng/µl (min=1.82; max=68.18) for *Listeria* datasets, slightly lower than the 27.40 ng/µl (min=1.33; max=59.83) observed for VTEC datasets, but substantially higher than the 9.00 ng/µl (min=5.03; max=13.10) measured for *Shigella* datasets. Without saponin-based depletion (NSD), DNA concentrations were markedly lower: 15.73 ng/µl (min=0.40; max=51.97) for *Listeria*, 12.28 ng/µl (min=0.91; max=30.02) for VTEC, and 4.68 ng/µl (min=2.65; max=7.29) for *Shigella* datasets. These differences were statistically significant across all pathogens (*Listeria*: $p<0.001$; VTEC: $p<0.001$; *Shigella*: $p=0.007$).

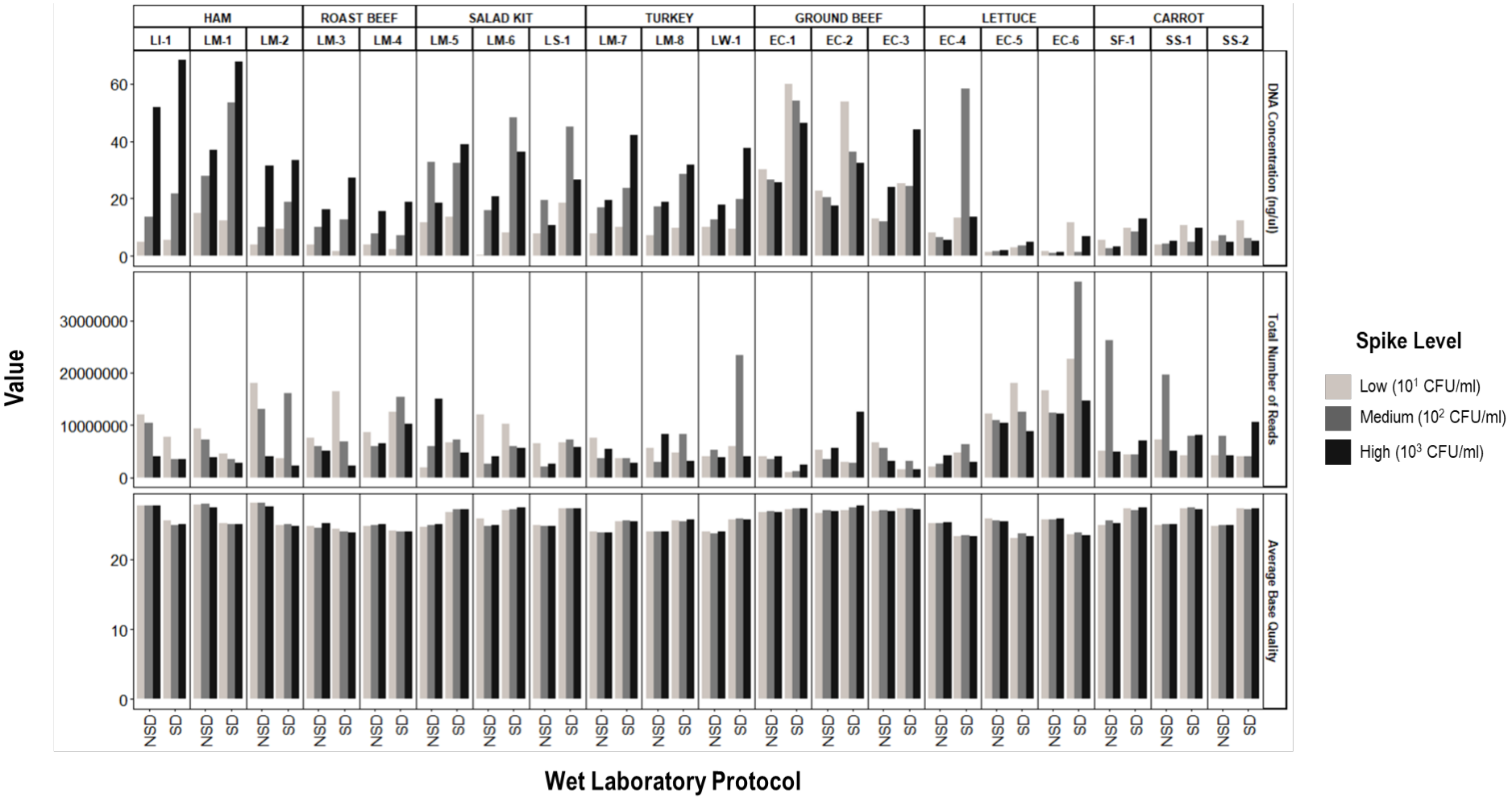


Figure 4.1. Comparative analysis of raw sequencing metrics and DNA concentration. Data are presented for three performance indicators—DNA concentration (ng/µL, top panel), total number of reads (middle panel), and average base quality (bottom panel)—across various food matrices (Ham, Roast Beef, Salad Kit, Turkey, Ground Beef, Lettuce, and Carrot) and three pathogen spike-in levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). The x-axis differentiates between two wet-laboratory protocols: no saponin-based eukaryotic DNA depletion (NSD) and saponin-based eukaryotic DNA depletion (SD). Datasets are identified by species code: *Listeria* species (LI = *L. innocua*; LM-1, 2, 3, 4, 5, 6, 7, 8 = *L. monocytogenes*; LS = *L. seeligeri*; LW = *L. welshimeri*), *Escherichia coli* (EC-1, 2, 5, 6 = *E. coli* O157:H7; EC-3 = *E. coli* O26:H11; EC-4 = *E. coli* O103:H25), and *Shigella* species (SF-1 = *S. flexneri*; SS-1, 2 = *S. sonnei*).

Total raw read counts also varied by protocol. For NSD datasets, average read numbers were 6,727,168 (min=1,852,520; max=18,044,540) for *Listeria*, 6,954,040 (min=2,116,136; max=16,534,908) for VTEC, and 9,409,348 (min=4,248,978; max=26,181,722) for *Shigella* datasets (Figure 4.1). SD datasets generated slightly higher averages for *Listeria* (6,996,968; min=2,274,960; max=23,282,244) and VTEC (8,751,933; min=1,105,410; max=37,382,186), and a lower average for *Shigella* (6,091,391; min=3,991,104; max=10,652,750) (Figure 4.1). Differences in total raw reads between protocols were not statistically significant. Across all datasets, mean base quality scores were ≥25 (≥99.6% base-call accuracy) regardless of protocol (Figure 4.1), with no significant differences between SD and NSD workflows. After pre-processing with fastp, fewer than 5% of reads were removed for low quality across both protocols (Appendix 4.1-Appendix 4.3).

4.2.2 Saponin-based eukaryotic DNA depletion significantly decreases the percentage of eukaryotic sequences

In foods artificially contaminated with *Listeria* strains, a higher percentage of eukaryotic reads originating from the animal or plant source of the food commodity was observed in NSD datasets compared to SD datasets (Figure 4.2). The highest proportions were detected in roast beef datasets (57%–24%), whereas salad kits contained less than 4%. When SD was applied, eukaryotic sequences were reduced to less than 1%. Across all *Listeria*-contaminated foods, this reduction was statistically significant (p<0.001).

In foods artificially contaminated with VTEC strains, the percentage of eukaryotic reads in NSD datasets was already low (<0.5%) (Figure 4.2); however, statistically significant differences between wet-laboratory protocols were detected for both ground beef (p<0.001) and lettuce (p=0.02). Similarly, in carrots spiked with *Shigella* strains, eukaryotic sequences were low without depletion (<0.05%) and further decreased to <0.01% with SD (Figure 4.2). A statistically significant difference between NSD and SD protocols was observed (p=0.02).

4.2.3 Saponin-based eukaryotic DNA depletion does not significantly improve the pathogen signal when the enrichment period is sufficiently optimized

Despite a statistically significant decrease in the percentage of eukaryotic sequences between NSD and SD datasets across all pathogens, the pathogen signal did not significantly improve. The average proportion of reads classified as *Listeria* was similar across protocols, with 96.4% (min=75.8%,

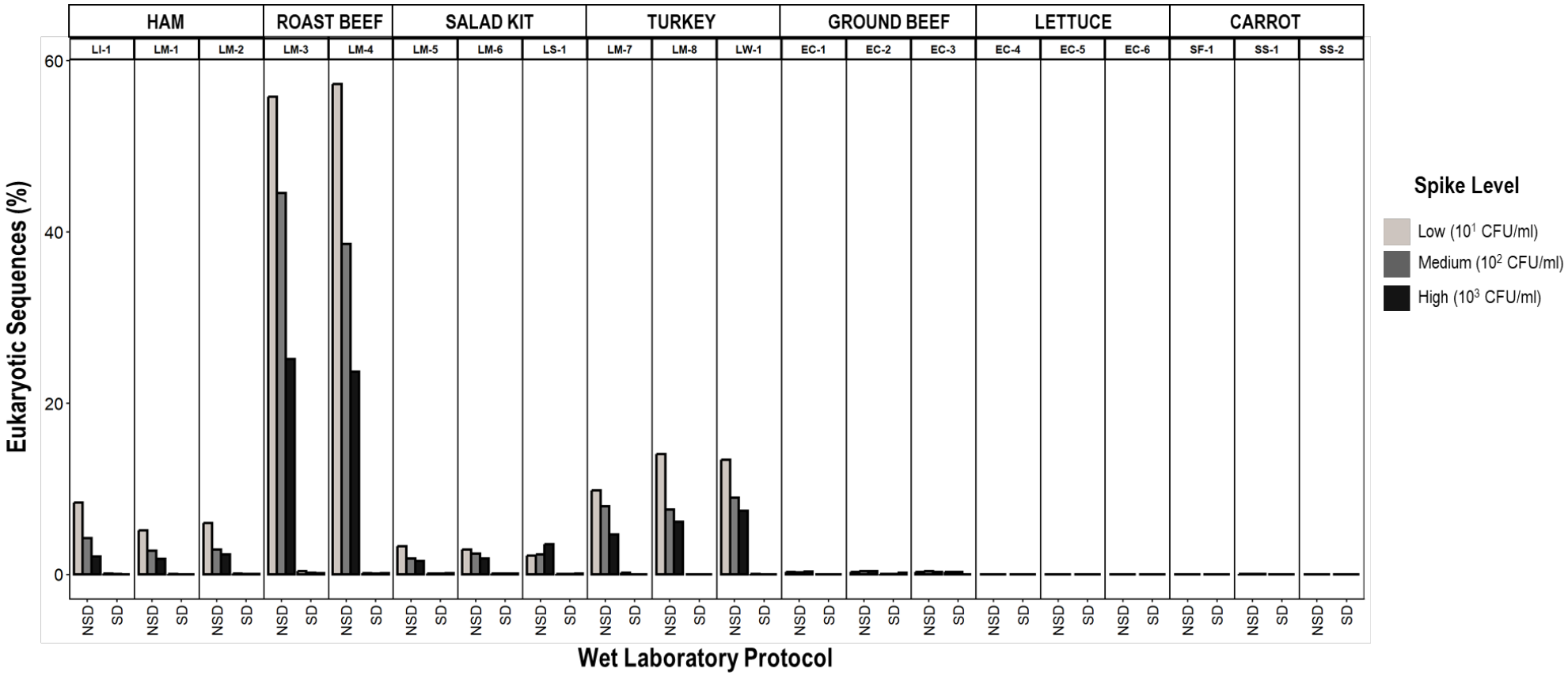


Figure 4.2. Impact of saponin-based depletion on eukaryotic DNA in enriched food samples. This figure illustrates the percentage of total sequencing reads identified as eukaryotic DNA in artificially contaminated food samples, comparing the performance of no saponin-based depletion (NSD) and saponin-based eukaryotic DNA depletion (SD) protocols**.** Data are presented across various food matrices (Ham, Roast Beef, Salad Kit, Turkey, Ground Beef, Lettuce, and Carrot) and three pathogen spike-in levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). The x-axis differentiates between two wet-laboratory protocols: no saponin-based eukaryotic DNA depletion (NSD) and saponin-based eukaryotic DNA depletion (SD). Datasets are identified by species code: *Listeria* species (LI = *L. innocua*; LM-1, 2, 3, 4, 5, 6, 7, 8 = *L. monocytogenes*; LS = *L. seeligeri*; LW = *L. welshimeri*), *Escherichia coli* (EC-1, 2, 5, 6 = *E. coli* O157:H7; EC-3 = *E. coli* O26:H11; EC-4 = *E. coli* O103:H25), and *Shigella* species (SF-1 = *S. flexneri*; SS-1, 2 = *S. sonnei*).

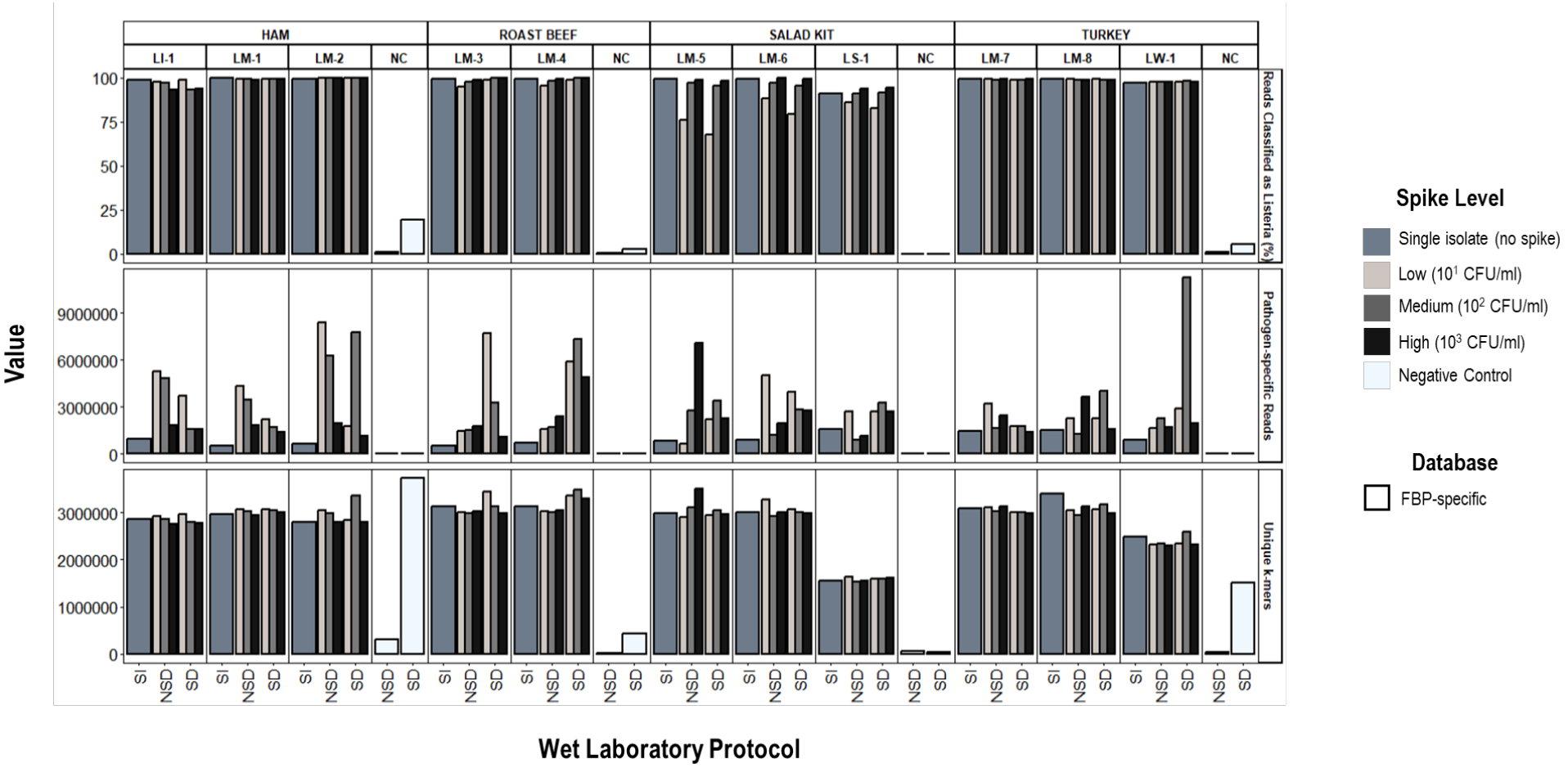


Figure 4.3. KrakenUniq classification metrics for *Listeria* strains across various food matrices, including single-isolate (SI) and negative controls (NC). This figure illustrates the impact of wet-laboratory interventions on taxonomic classification by comparing the percentage of reads classified (top panel), pathogen-specific read counts (middle panel), and unique *k*-mer counts (bottom panel). Data are stratified by two wet-laboratory protocols, non-saponin-depleted (NSD) and saponin-depleted (SD) and compared against single-isolate (SI) reference strains. Pathogen detection was evaluated across four food commodities (Ham, Roast Beef, Salad Kit, and Turkey) at three contamination levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). Negative controls (NC) were included to assess baseline taxonomic noise. Datasets are identified by species code: LI = *L. innocua*; LM-1 through 8 = *L. monocytogenes*; LS = *L. seeligeri*; and LW = *L. welshimeri*. All classifications were performed using the optimized Foodborne Pathogen-specific (FBP-specific) reference database.

max=99.9%) in NSD datasets and 96.1% (min=67.6%, max=99.9%) in SD datasets (p=0.327) (Figure 4.3). Following taxonomic classification with KrakenUniq and the FBP-specific reference database, an average of 2,903,878 paired-end pathogen-specific reads (min=652,007; max=8,349,048) were assigned at the genus level for *L. monocytogenes* NSD datasets (Figure 4.3). With SD, this increased to an average of 3,171,875 reads (min=1,079,124; max=7,744,742). The number of unique *k*-mers was similar between protocols, averaging 3,043,428 (min=2,787,067; max=3,492,969) in NSD datasets and 3,079,636 (min=2,788,152; max=3,471,465) in SD datasets (Figure 4.3). No statistically significant differences in pathogen-specific reads (p=0.657) or unique *k*-mers (p= 0.442) were detected. As expected, datasets spiked with other *Listeria* species yielded fewer pathogen-specific reads and unique *k*-mers across both protocols, yet no significant differences were observed (pathogen-specific reads: p=0.385; unique *k*-mers: p=0.25).

The percentage of reads classified and the unique *k*-mer counts for SI reference strains were comparable to those recovered from the NSD and SD quasi-metagenomic protocols (Figure 4.3). Although not unexpected, there was a considerable difference in the pathogen-specific read counts acquired by design between SI reference strains and the more complex datasets (NSD and SD) and is likely attributed to differences in total sequencing effort. This pattern reflects the substantially greater sequencing depths applied to quasi-metagenomic samples to offset matrix-associated background signal, including eukaryotic DNA and non-target microbial taxa, as discussed in Section 4.2.6. Based on the stability of the unique *k*-mer counts, these reads likely represent redundant coverage rather than new genomic information (Figure 4.3).

In the negative controls, the SD ham matrix showed a higher number of pathogen-specific unique *k*-mers than the NSD counterpart despite containing relatively few pathogen-specific reads. Low-level assignments to *L. monocytogenes* were supported by BLAST analysis; however, all negative control samples were culture-negative for the pathogen of interest (Section 2.2), indicating that the signal was unlikely to represent viable *L. monocytogenes*. Instead, it was interpreted as low-level background signal that may reflect environmental DNA, reagent or laboratory contamination, or low-level cross-sample carryover that become detectable at high sequencing depth. This interpretation is further supported by the absence of *L. monocytogenes* among the top ten matches identified by RefSeq Masher (contains) using the preprocessed quasi-metagenomic datasets (Section 4.2.7.1). Collectively, these findings emphasize the importance of analytical thresholds for distinguishing true pathogen detections from low-level background signal.

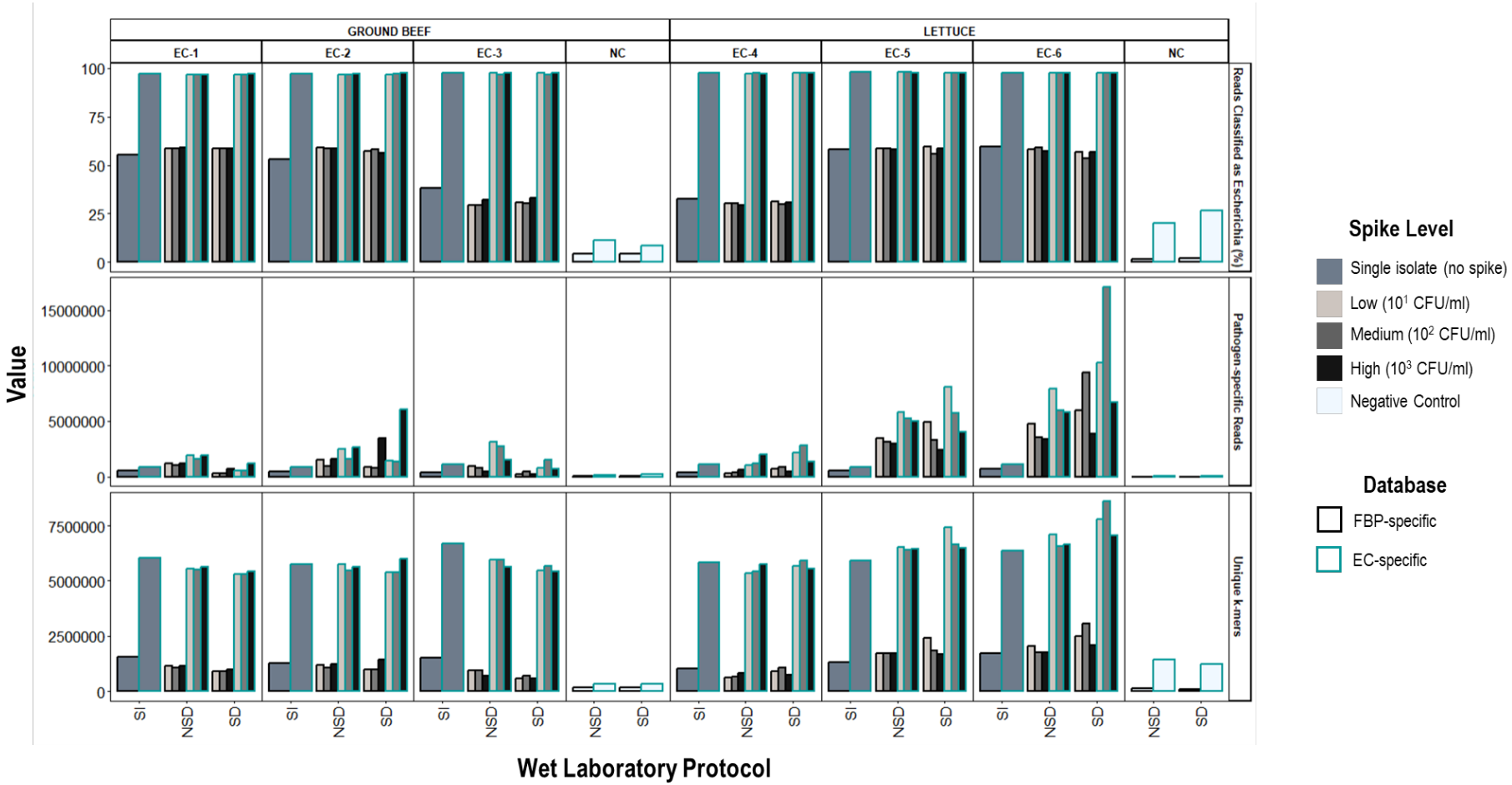


Figure 4.4. KrakenUniq classification metrics for verotoxigenic *Escherichia coli* (VTEC) strains across various food matrices, including single-isolate (SI) and negative controls (NC). This figure illustrates the impact of wet-laboratory interventions on taxonomic classification by comparing the percentage of reads classified (top panel), pathogen-specific read counts (middle panel), and unique *k*-mer counts (bottom panel). Data are stratified by two wet-laboratory protocols, non-saponin-depleted (NSD) and saponin-depleted (SD) and compared against single-isolate (SI) reference strains. Pathogen detection was evaluated across two food commodities (Lettuce and Ground Beef) at three contamination levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). Negative controls (NC) were included to assess baseline taxonomic noise. Datasets are identified by species code: EC-1, 2, 5, 6 (*E. coli* O157:H7); EC-3 (*E. coli* O26:H11); and EC-4 (*E. coli* O103:H25) SI = single-isolate reference strain and NC = negative control. All classifications were performed using the optimized Foodborne Pathogen-specific (FBP-specific) and *Escherichia coli* specific (EC-specific) reference databases.

4.2.4 Species-specific reference databases for VTEC and *Shigella* species significantly improve pathogen signal

In food matrices artificially contaminated with VTEC O157 and analyzed via the FBP-specific database, the proportion of reads identified as *E. coli* remained low and stable across protocols; NSD datasets yielded a mean of 49.1% (range: 29.1%–59.3%), while SD datasets produced 48.6% (range: 29.8%–59.5%) (p=0.229; Figure 9). Although the average number of pathogen-specific reads and unique *k*-mers increased in SD datasets compared to NSD from 2,392,156 to 3,032,665 reads and 1,458,364 to 1,653,666 *k*-mers, respectively, neither change was statistically significant (p=0.57 and p=0.87; Figure 4.4). For non-O157 VTEC strains (O26:H11 and O103:H25), pathogen-specific reads and unique *k*-mers were generally lower and showed a non-significant decrease between NSD and SD protocols (p=0.605 and p=0.825; Figure 4.4).

Transitioning to the EC-specific reference database substantially enhanced the pathogen signal compared to the FBP-specific database (Figure 4.4) with an average increase of 1,515,389 and 4,723,431, for the pathogen-specific reads and unique *k*-mers, respectively (p<0.001). Comparatively, significant increases were observed for SD datasets, with pathogen-specific reads rising by an average of 1,833,808 and unique *k*-mers by 4,783,165 (p<0.001). Subsequently, the proportion of reads classified as *E. coli* rose to 97.3% in NSD and 97.4% in SD datasets (Figure 4.4). For the VTEC SI reference strains, the proportion of classified reads and the number of unique *k*-mers recovered remained comparable across datasets, irrespective of wet-laboratory protocol. Notably, both metrics were consistently higher when taxonomic classification was performed using the EC-specific reference database relative to the broader FBP-specific database.

Negative controls analyzed with the FBP-specific database showed high specificity, with background signals remaining below 5% classified reads in ground beef and lettuce, irrespective of depletion protocol (Figure 4.4). In contrast, NC samples analyzed with the EC-specific reference database displayed a clear elevation in baseline signal: ground beef datasets exceeded 10% classified reads, and lettuce SD datasets approached 27%. This increase likely reflects the database's enhanced sensitivity to conserved *E. coli* genomic regions, which can also capture commensal *E. coli* and closely related members of the Enterobacteriaceae. This was supported by RefSeq Masher analysis: pathogen-specific reads from ground beef NC samples were most similar to *Atlantibacter hermannii* (~97.5% identity; 265/400 shared hashes for NSD NC and 97.3%; 259/400 for SD NC). In contrast, reads from lettuce NC primarily matched

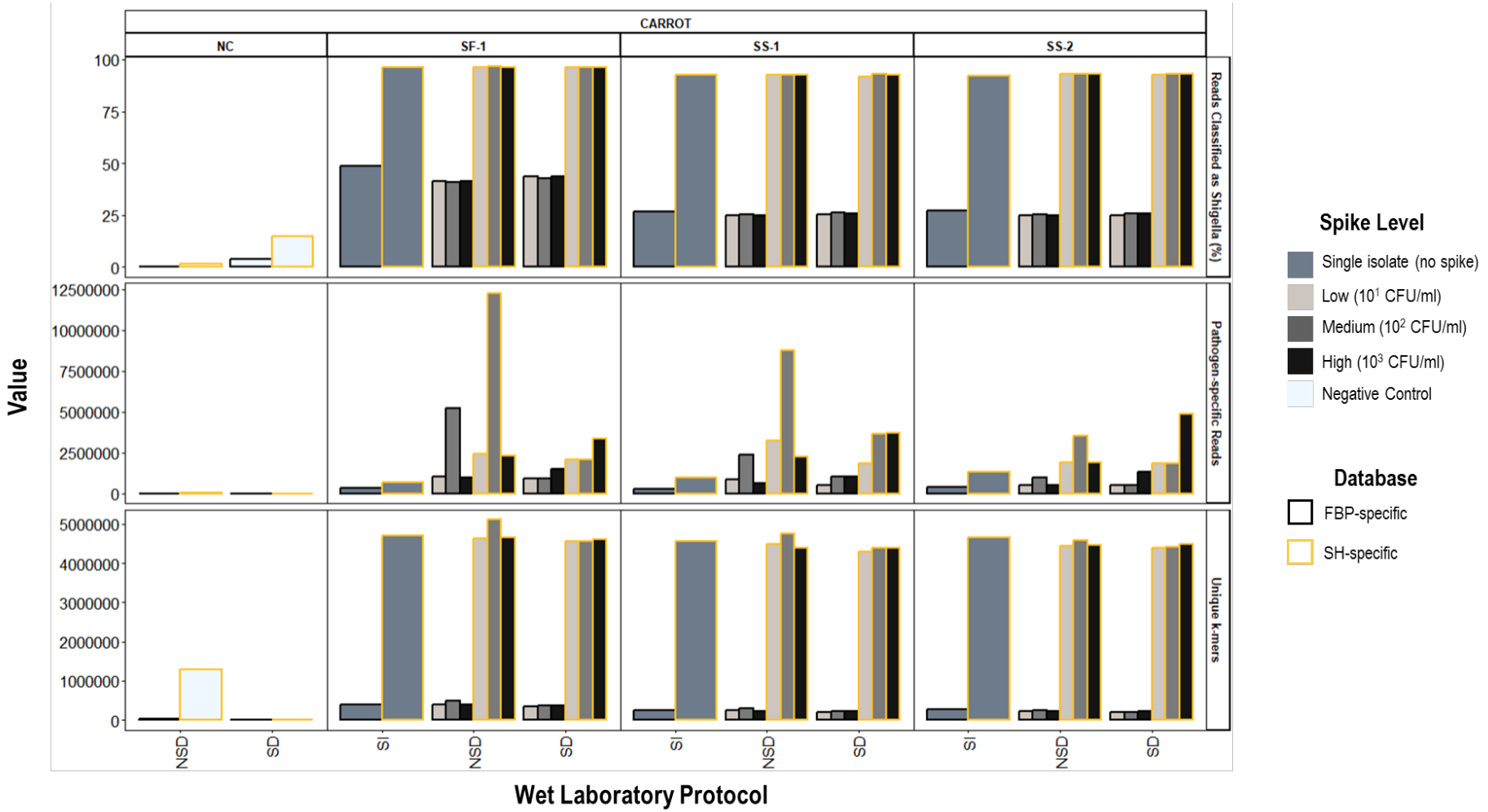


Figure 4.5. KrakenUniq classification metrics for *Shigella* strains across a single food matrix, including single-isolate (SI) and negative controls (NC). This figure illustrates the impact of wet-laboratory interventions on taxonomic classification by comparing the percentage of reads classified (top panel), pathogen-specific read counts (middle panel), and unique *k*-mer counts (bottom panel). Data are stratified by two wet-laboratory protocols, non-saponin-depleted (NSD) and saponin-depleted (SD) and compared against single-isolate (SI) reference strains. Pathogen detection was evaluated in a single food commodity (Carrots) at three contamination levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). Negative controls (NC) were included to assess baseline taxonomic noise. Datasets are identified by species code: (SF-1 = *S. flexneri*; SS-1, 2 = *S. sonnei*). All classifications were performed using the optimized Foodborne Pathogen-specific (FBP-specific) and *Shigella species* specific (SH-specific) reference databases.

E. coli, with 94.5% identity and <150/400 shared hashes for the SD NC, and 98.4% identity with 310/400 shared hashes for the NSD NC. Importantly, unique *k*-mer counts in all NC samples remained low relative to true positives, indicating that these signals are unlikely to represent meaningful pathogen signal (Figure 4.4).

For *Shigella* datasets analyzed with the FBP-specific database, a general decreasing trend in both pathogen-specific reads and unique *k*-mers was observed following saponin-based depletion (Figure 4.5). In *S. sonnei* datasets, pathogen-specific reads declined from 974,189 to 820,053 (p=0.64), while unique *k*-mers decreased from 240,123 to 212,529 (p=0.11). A similar non-significant trend occurred for *S. flexneri*, where pathogen-specific reads decreased from 2,408,041 in NSD datasets to 1,124,050 in SD datasets (p=0.48), and unique *k*-mers were higher in NSD (mean=418,441) than in SD (mean=366,578) datasets (p=0.29). As observed with VTEC, the overall proportion of reads classified as *Shigella* remained low, though a statistically significant increase was noted from NSD (30.4%) to SD (31.5%) datasets (p=0.002; Figure 4.5). Similar to the findings for VTEC, the SH-specific reference database substantially improved detection metrics (Figure 4.5). Pathogen-specific reads increased by an average of 2,842,343 in NSD datasets (p=0.004) and 1,885,766 in SD datasets (p=0.004). Unique *k*-mer counts also rose significantly, with increases of 4,314,117 in NSD (p=0.004) and 4,192,149 in SD datasets (p<0.001). Consequently, the proportion of reads classified as *Shigella* reached a mean of 94.0% across both NSD (range: 92.6%–96.6%) and SD (range: 91.8%–96.4%) protocols (Figure 4.5). Similarly, for the *Shigella* SI reference strains, the proportion of classified reads and the number of unique *k*-mers were comparable across datasets, irrespective of wet-laboratory protocol. As anticipated, both metrics were consistently higher when analyses were conducted using the SH-specific reference database relative to the FBP-specific database.

Negative control results in the carrot matrix further validate the specificity of the analytical pipeline, though results varied between reference databases. When utilizing the FBP-specific database, background signal remained negligible across both wet-laboratory protocols, with the percentage of reads classified as *Shigella* and unique *k*-mer counts staying near baseline. Conversely, the SH-specific database detected a higher proportion of classified reads in the negative controls, particularly in the NSD protocol, where background unique *k*-mer counts reached approximately 1.3×10^6. Although RefSeq Masher indicated approximately 94.0% identity to *Shigella* and *E. coli*, the number of shared hashes was substantially lower (<150/400) than that observed in true positive samples (Section 4.2.7.1). Despite the increased background signal associated with the more sensitive species-specific database, these matches are unlikely to represent true pathogen signal.

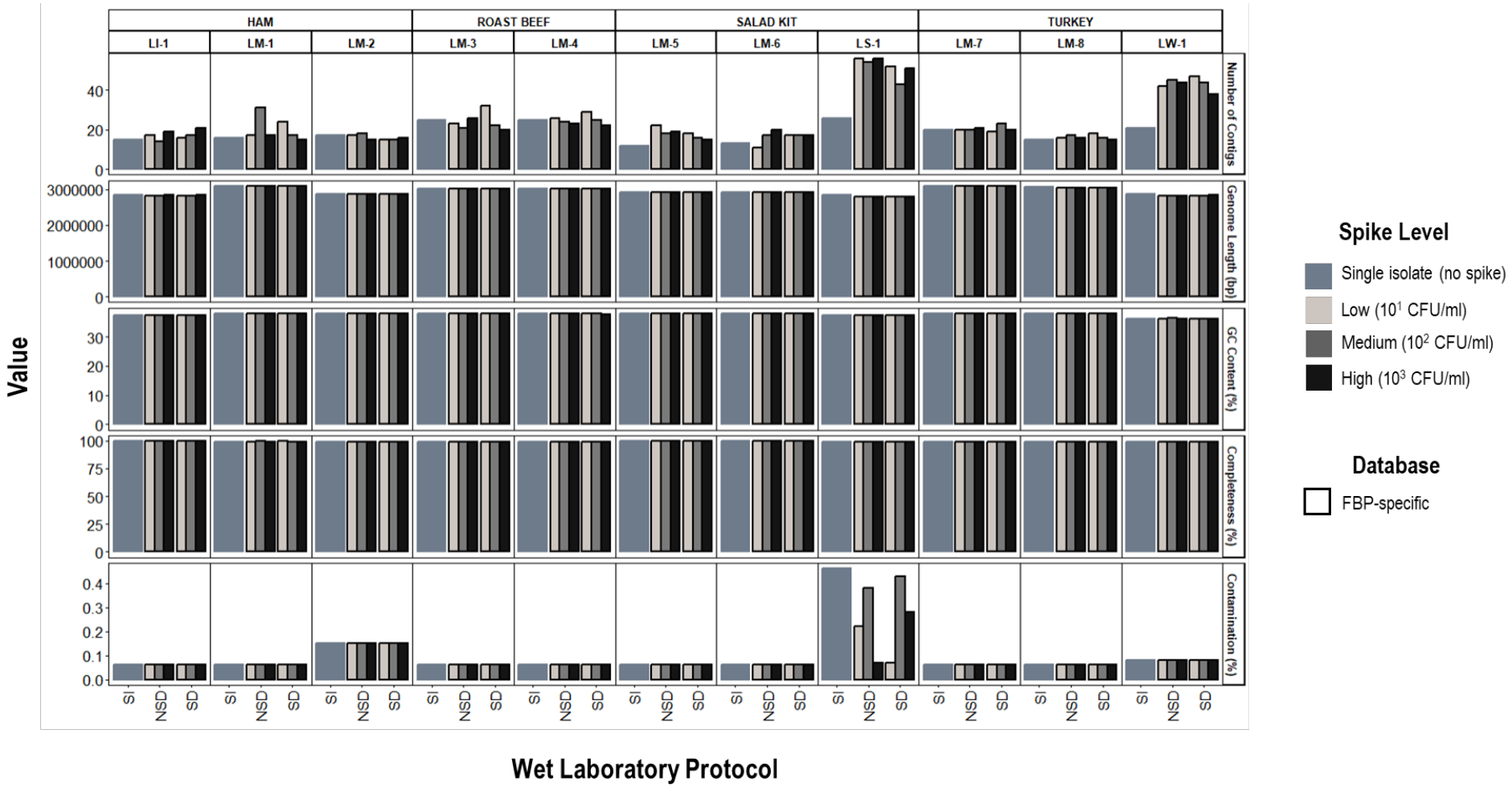


Figure 4.6. Comparative assembly quality metrics for *Listeria* strains in various food matrices. This figure evaluates *de novo* assembly performance by comparing quasi-metagenomic assemblies generated from reads extracted using the FBP-specific reference database against single isolate (SI) reference assemblies. Assessment parameters include QUAST-derived Number of Contigs, Genome Length (bp), GC Content (%) and CheckM-derived Completeness (%) and Contamination (%). Data are stratified by two wet-laboratory protocols, non-saponin-depleted (NSD) and saponin-depleted (SD) and evaluated across four food commodities (Ham, Roast Beef, Salad Kit and Turkey) at three contamination levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). Datasets are identified by species code: *Listeria* species (LI = *L. innocua*; LM-1 through 8 = *L. monocytogenes*; LS = *L. seeligeri*; LW = *L. welshimeri*).

4.2.5 Assembly quality is consistent between wet-laboratory protocols

For all datasets, assemblies generated from pathogen-specific reads demonstrated quality metrics comparable to the assemblies of the corresponding SI reference assemblies, regardless of wet-laboratory protocol (Figures 4.6-4.8). For *L. monocytogenes*, the mean number of contigs in the NSD assemblies (20; min=11, max=31) and SD datasets (19; min=15, max=32) closely matched those of the single isolate reference strains (mean=18; min=12, max=25) (Figure 4.6). Genome lengths were also consistent between NSD (mean=2,993,472 bp; min=2,859,791; max=3,086,345) and SD assemblies (mean=2,993,495 bp; min=2,859,574; max=3,086,129) and aligned with the SI reference assembly values. The GC content remained stable at 37.8% across all assemblies. Completeness (99.4%) and contamination (0.07%) were also consistent between wet-laboratory protocols and SI reference assemblies.

For datasets spiked with other *Listeria* species (Figure 4.6), NSD assemblies averaged 36 contigs (min=16, max=52) and a genome length of 2,827,598 bp (min=2,798,825; max=2,895,804), while SD assemblies averaged 37 contigs (min=17, max=56) and a genome length of 2,834,475 bp (min=2,793,588; max=2,927,155). GC content was consistent between protocols (37.0%). Completeness remained uniform at 99.4% (min=99.2%, max=99.8%), although contamination differed slightly, averaging 0.26% (min=0.04%, max=1.68%) in NSD datasets and 0.12% (min=0.04%, max=0.43%) in SD datasets. For VTEC assemblies, reads extracted using the EC-specific reference database produced quality metrics more consistent with the corresponding reference strains than assemblies generated using the FBP-specific database (Figure 4.7). For VTEC O157, the average number of contigs decreased from >500 to 170 (min=136; max=227) in NSD assemblies and to 193 (min=152; max=221) in SD datasets, comparable to the 179 contigs observed for the single-isolate reference assembly. Genome lengths also improved, reaching 5,394,079 bp (min=5,326,462; max=5,463,729) in NSD assemblies and 5,387,304 bp (min=5,315,583; max=5,459,566) in SD assemblies, approaching the SI reference strain mean of 5,423,895 bp. GC content remained unchanged at 50.3% for NSD, SD, and SI assemblies. For both wet-laboratory protocols, completeness increased from >95% to 99.4%, matching the value for the SI reference assembly (99.6%), while contamination decreased from ≥0.72% to ≤0.20%, aligning with the SI reference assembly (0.18%).

For VTEC O26:H11, the average number of contigs decreased from >1,000 to 242 (min=236; max=252) in NSD assemblies and 271 (min=244; max=292) in SD assemblies, consistent with SI reference strain (261 contigs) (Figure 4.7). For VTEC O103:H25, contig counts decreased to 225 (min=222; max=230) in NSD assemblies and 237 (min=231; max=240) in SD assemblies, aligning with the SI

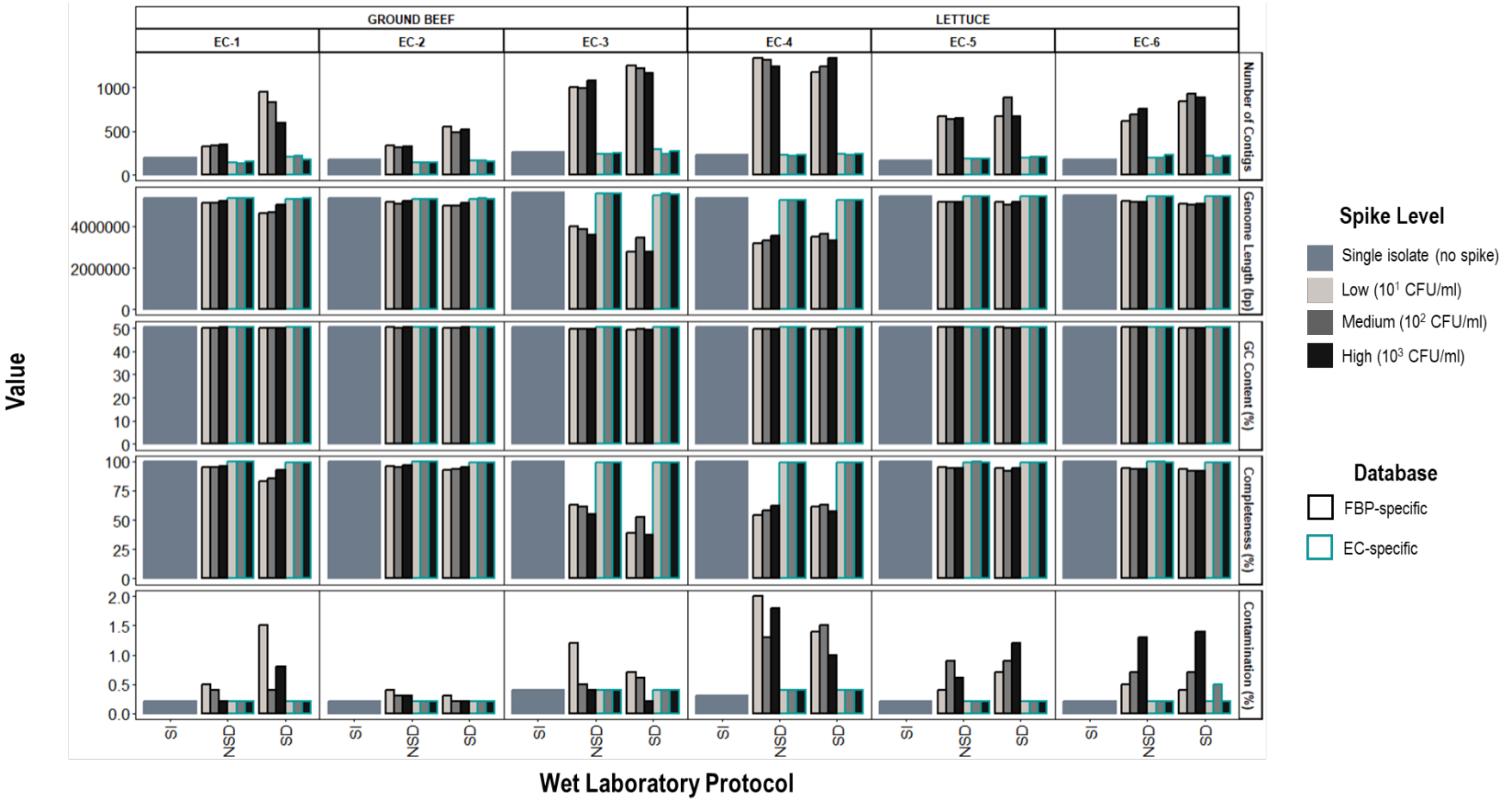


Figure 4.7. Comparative assembly quality metrics for verotoxigenic *Escherichia coli* (VTEC) strains in various food matrices. This figure evaluates *de novo* assembly performance by comparing quasi-metagenomic assemblies generated from reads extracted using the FBP-specific reference database and *Escherichia coli* (EC) specific reference database against single isolate (SI) reference assemblies. Assessment parameters include QUAST-derived Number of Contigs, Genome Length (bp), GC Content (%) and CheckM-derived Completeness (%) and Contamination (%). Data are stratified by two wet-laboratory protocols, non-saponin-depleted (NSD) and saponin-depleted (SD) and evaluated across two food commodities (Ground Beef and Lettuce) at three contamination levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). Datasets are identified by species code: EC-1, 2, 5, 6 (*E. coli* O157:H7); EC-3 (*E. coli* O26:H11); and EC-4 (*E. coli* O103:H25).

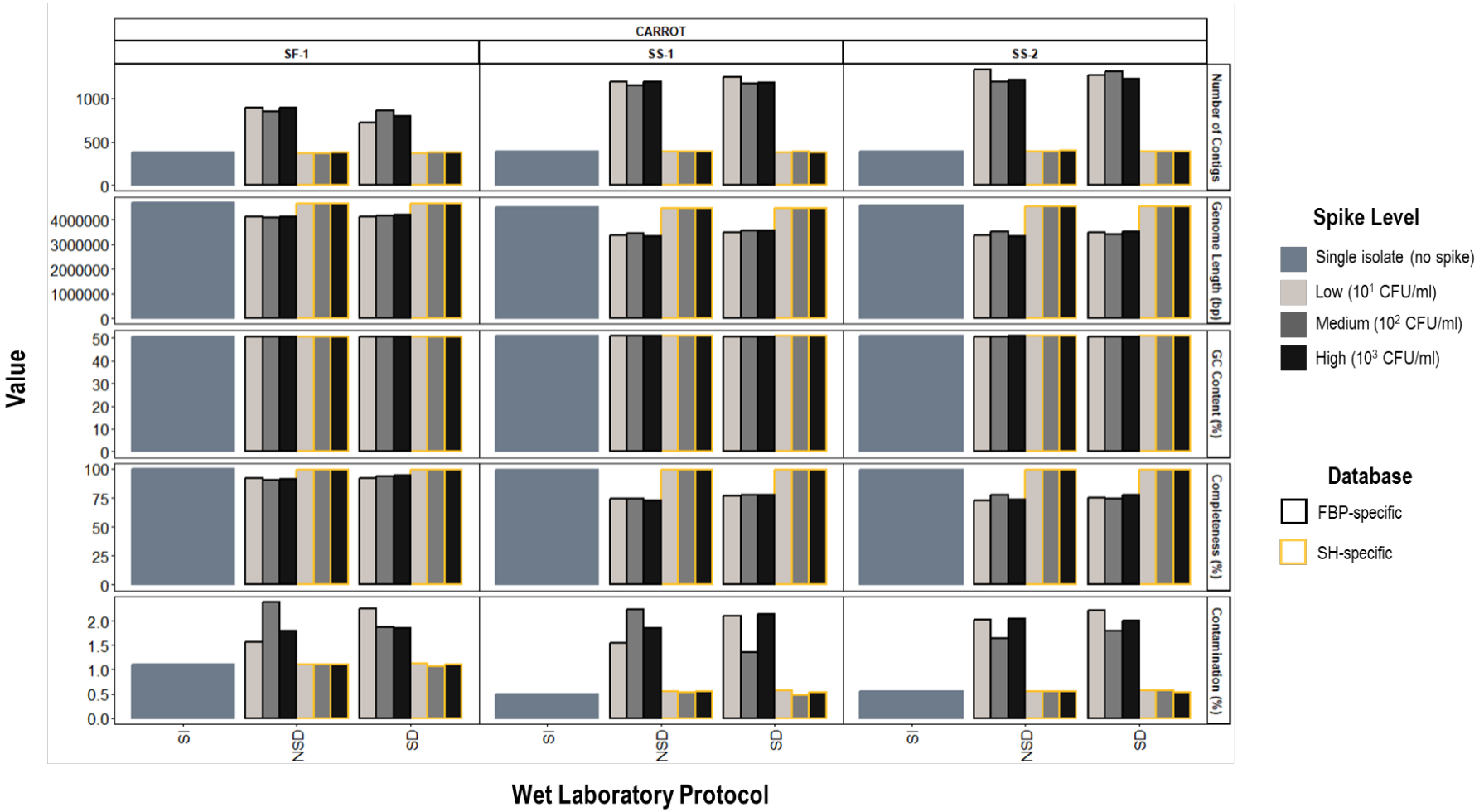


Figure 4.8. Comparative assembly quality metrics for *Shigella* strains in a single food matrix. This figure evaluates *de novo* assembly performance by comparing quasi-metagenomic assemblies generated from reads extracted using the FBP-specific reference database and *Shigella* (SH) specific reference database against single isolate (SI) reference assemblies. Assessment parameters include QUAST-derived Number of Contigs, Genome Length (bp), GC Content (%) and CheckM-derived Completeness (%) and Contamination (%). Data are stratified by two wet-laboratory protocols, non-saponin-depleted (NSD) and saponin-depleted (SD) and evaluated across a single food commodity (Ground Beef and Lettuce) at three contamination levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). Datasets are identified by species code: (SF-1 = *S. flexneri*; SS-1, 2 = *S. sonnei*).

reference assembly (225 contigs). Genome lengths also improved. For VTEC O26:H11, lengths increased to 5,578,898 bp (min=5,577,147; max=5,580,159) in NSD assemblies and 5,545,424 bp (min=5,504,133; max=5,572,972) in SD assemblies, consistent with the SI reference assembly length (5,622,824 bp). For VTEC O103:H25, genome lengths increased to 5,280,348 bp (min=5,279,966; max=5,280,630) in NSD assemblies and 5,278,405 bp (min=5,277,583; max=5,279,022) in SD assemblies. GC content remained stable at 50.4%. Completeness increased from <65% to 99.3%, and contamination decreased from ~2.0% to 0.4% across both protocols.

For *Shigella* assemblies generated using reads extracted with the SH-specific database, quality metrics more closely matched the corresponding single isolates than those generated using the FBP-specific database (Figure 4.8). For *S. sonnei*, contig counts decreased from >1,000 to 389 (min=385; max=395) in NSD assemblies and 387 (min=382; max=391) in SD assemblies, comparable to the single-isolate reference assemblies (mean=385). Genome lengths increased to 4,482,262 bp (min=4,437,626; max=4,528,961) in NSD assemblies and 4,483,645 bp (min=4,440,492; max=4,527,212) in SD assemblies. GC content remained constant at 50.8%. Completeness increased from <80% to 99.2%, and contamination decreased from ~2.0% to 0.55%, aligning more closely with the SI reference assemblies. For *S. flexneri*, contig counts decreased from <750 to 372 (min=368; max=378) in NSD assemblies and 373 (min=371; max=376) in SD assemblies. Genome lengths increased to 4,638,646 bp (min=4,636,781; max=4,642,305) in NSD datasets and 4,642,978 bp (min=4,642,566; max=4,643,317) in SD assemblies, approaching the SI reference genome length (4,676,637 bp). GC content remained stable at 50.4%. Completeness increased from <95% to 99.5%, and contamination decreased from ~2.0% to 1.1% across both wet-laboratory protocols.

4.2.6 Genome coverage showed no statistically significant variation between wet laboratory protocols but was substantially improved with species-specific reference databases for verotoxigenic *Escherichia coli* (VTEC) and *Shigella* datasets.

Following taxonomic classification with KrakenUniq using the FBP-specific reference database, an average of 87.3% of pathogen-specific reads (min=72.0%; max=100%) in NSD *L. monocytogenes* datasets mapped to the corresponding single-isolate reference genome, with a mean coverage depth of 70.3X (min=38.0X; max=108X) (Figure 4.9). Comparable results were observed for SD datasets, with 88.5% of pathogen-specific reads mapped (min=73.1%; max=99.9%) and a mean coverage depth of 73.6X (min=44.4X; max=110.2X). No statistically significant differences were detected in the percentage of

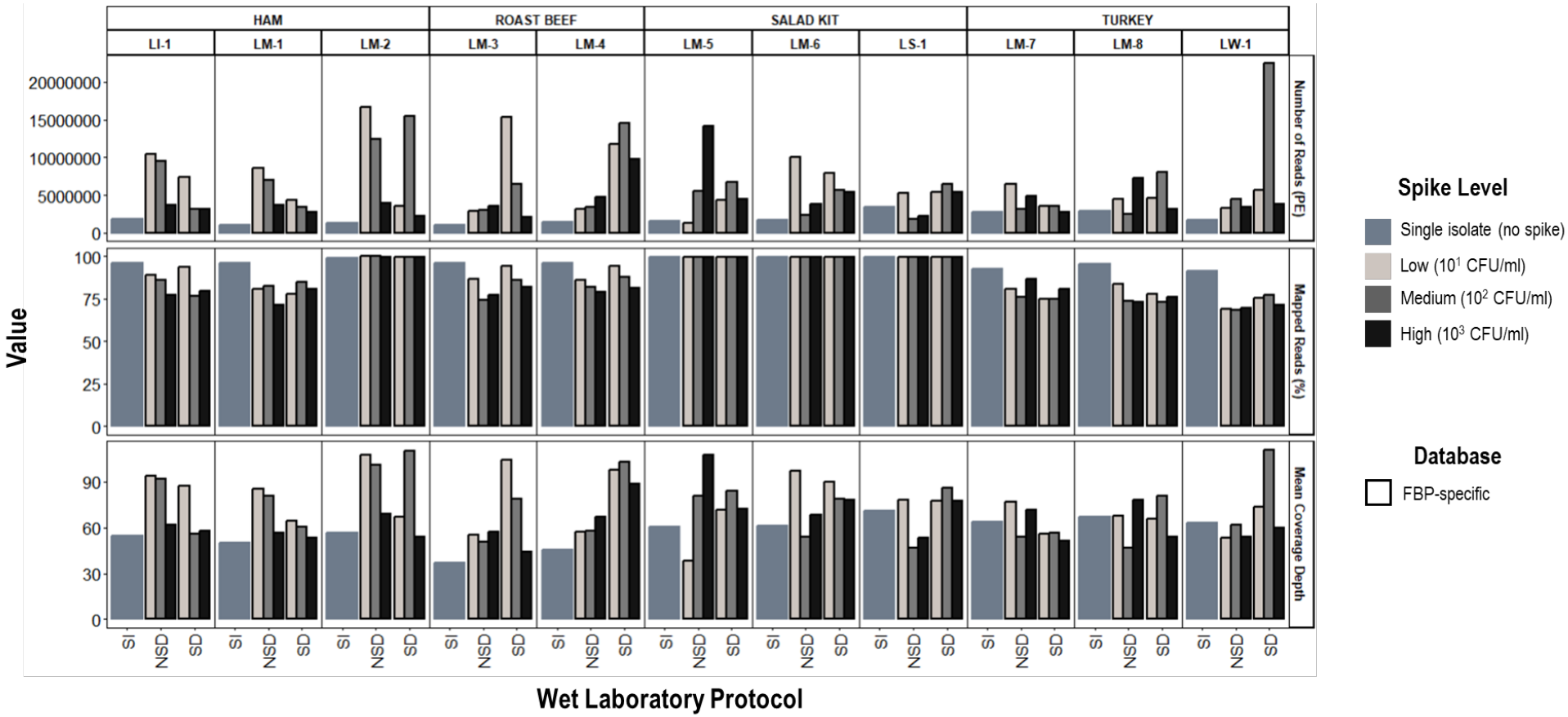


Figure 4.9. Genome coverage metrics for *Listeria* species across varied food matrices and protocols. This figure evaluates the quality of sequence data used for assembly by comparing quasi-metagenomic datasets—comprising reads extracted via the FBP-specific reference database—against single isolate (SI) reference controls. Performance is assessed across non-saponin-depleted (NSD) and saponin-depleted (SD) protocols for three contamination levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL) in four food commodities: Ham, Roast Beef, Salad Kit, and Turkey. Panels display Qualimap-derived metrics including the Number of Reads (PE) successfully assigned to the target pathogen, the percentage of Mapped Reads (%) aligning to the corresponding reference genome, and the Mean Coverage Depth. Datasets are identified by species code: *Listeria* species (LI = *L. innocua*; LM-1 through 8 = *L. monocytogenes*; LS = *L. seeligeri*; LW = *L. welshimeri*).

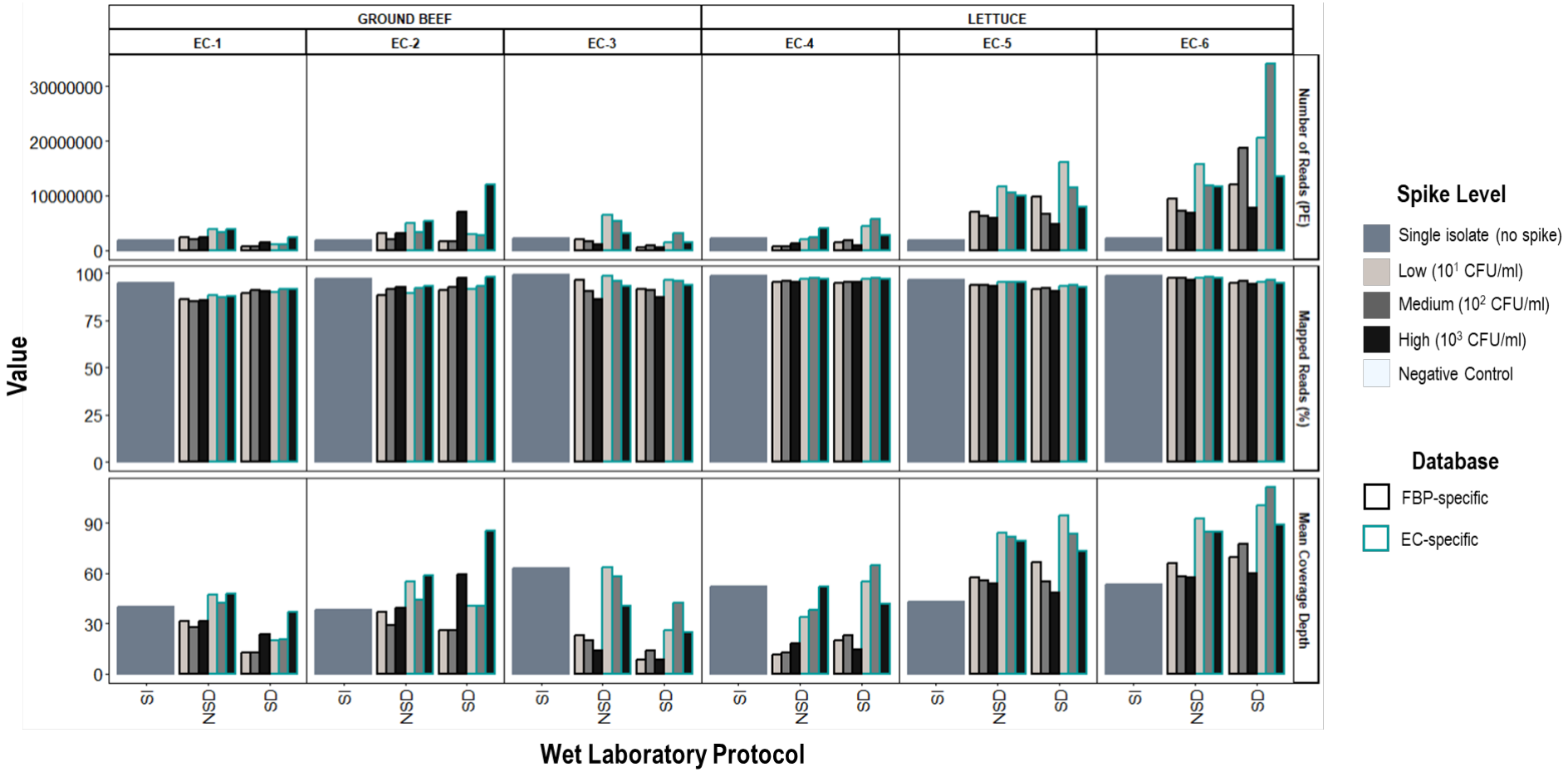


Figure 4.10. Comparative genome coverage metrics for verotoxigenic *Escherichia coli* (VTEC) across various food matrices and protocols. This figure evaluates the quality of sequence data used for assembly by comparing quasi-metagenomic datasets comprising reads extracted via the Foodborne pathogen (FBP)-specific reference database and *Escherichia coli* (EC)- specific reference database against single isolate (SI) reference controls. Performance is assessed across non-saponin-depleted (NSD) and saponin-depleted (SD) protocols for three contamination levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL) in two food commodities Ground Beef and Lettuce. Panels display Qualimap-derived metrics including the Number of Reads (PE) successfully assigned to the target pathogen, the percentage of Mapped Reads (%) aligning to the corresponding reference genome, and the Mean Coverage Depth. Datasets are identified by species code: EC-1, 2, 5, 6 (*E. coli* O157:H7); EC-3 (*E. coli* O26:H11); and EC-4 (*E. coli* O103:H25).

mapped reads (p=0.192) or mean coverage depth (p=0.54) between wet-laboratory protocols. For datasets spiked with other *Listeria* species, an average of 84.2% of reads mapped in NSD datasets (min=68.6%; max=99.7%) and 85.9% in SD datasets (min=71.4%; max=99.86%) (p=0.341). Mean coverage depth also did not statistically differ between NSD (mean=66.1X; min=46.9X; max=94.0X) and SD datasets (mean=76.2X; min=56.2X; max=111.2X) (p=0.272). Overall, mapped-read percentages and coverage depths were generally consistent with the corresponding SI reference assemblies (Figure 4.9).

For VTEC-specific reads extracted using the EC-specific reference database, an average of 93.0% of reads (min=87.2%; max=97.8%) mapped in NSD *E. coli* O157:H7 datasets, compared with 93.4% (min=89.9%; max=97.9%) in SD datasets (p=0.791) (Figure 4.10). Mean coverage depth also did not statistically differ, with 66.9X (min=42.6X; max=92.9X) in NSD datasets and 66.3X (min=19.8X; max=111.5X) in SD datasets (p=0.907). Across both protocols, datasets processed with the EC-specific reference database consistently demonstrated higher mapped-read percentages and coverage depths than those processed with the FBP-specific database and were more comparable to the SI reference assemblies. Statistically significant differences between EC-specific and FBP-specific reference databases were observed for NSD datasets (percent mapped reads p<0.001; mean coverage depth p<0.001) and SD datasets (percent mapped reads p<0.001; mean coverage depth p<0.001).

For VTEC non-O157 datasets, 96.4% of reads mapped in NSD datasets (min=92.9%; max=98.5%) and 96.2% in SD datasets (min=93.9%; max=97.2%) (p=0.66) (Figure 4.10). Mean coverage depths of 47.7X (min=33.7X; max=63.8X) in NSD datasets and 42.4X (min=24.8X; max=64.6X) in SD datasets were not statistically different (p=0.62). As with O157:H7, datasets processed with the EC-specific database demonstrated significantly higher mapped-read percentages (NSD p=0.0313; SD p=0.0066) and mean coverage depths (NSD p<0.001; SD p<0.001) compared to those processed with the FBP-specific database and better aligned with SI metrics.

For *S. sonnei* datasets, an average of 93.9% of SH-specific reads mapped in NSD datasets (min=91.6%; max=95.7%), compared with 92.2% in SD datasets (min=88.8%; max=96.0%) (p=0.195) (Figure 4.11). Mean coverage depth similarly did not differ between NSD (mean=66.5X; min=51.9X; max=97.6X) and SD datasets (mean=62.5X; min=49.7X; max=80.7X) (p=0.666). For *S. flexneri*, NSD datasets showed 94.6% of reads mapped (min=92.6%; max=96.0%) with a mean coverage depth of 75.9X (min=59.2X; max=107.8X), while SD datasets showed slightly lower mapped-read percentages (92.3%; min=90.9%; max=93.1%) and mean depths (60.5X; min=55.6X; max=69.8X). A statistically significant difference was detected between wet-laboratory protocols for percent mapped reads (p=0.0245), but not for

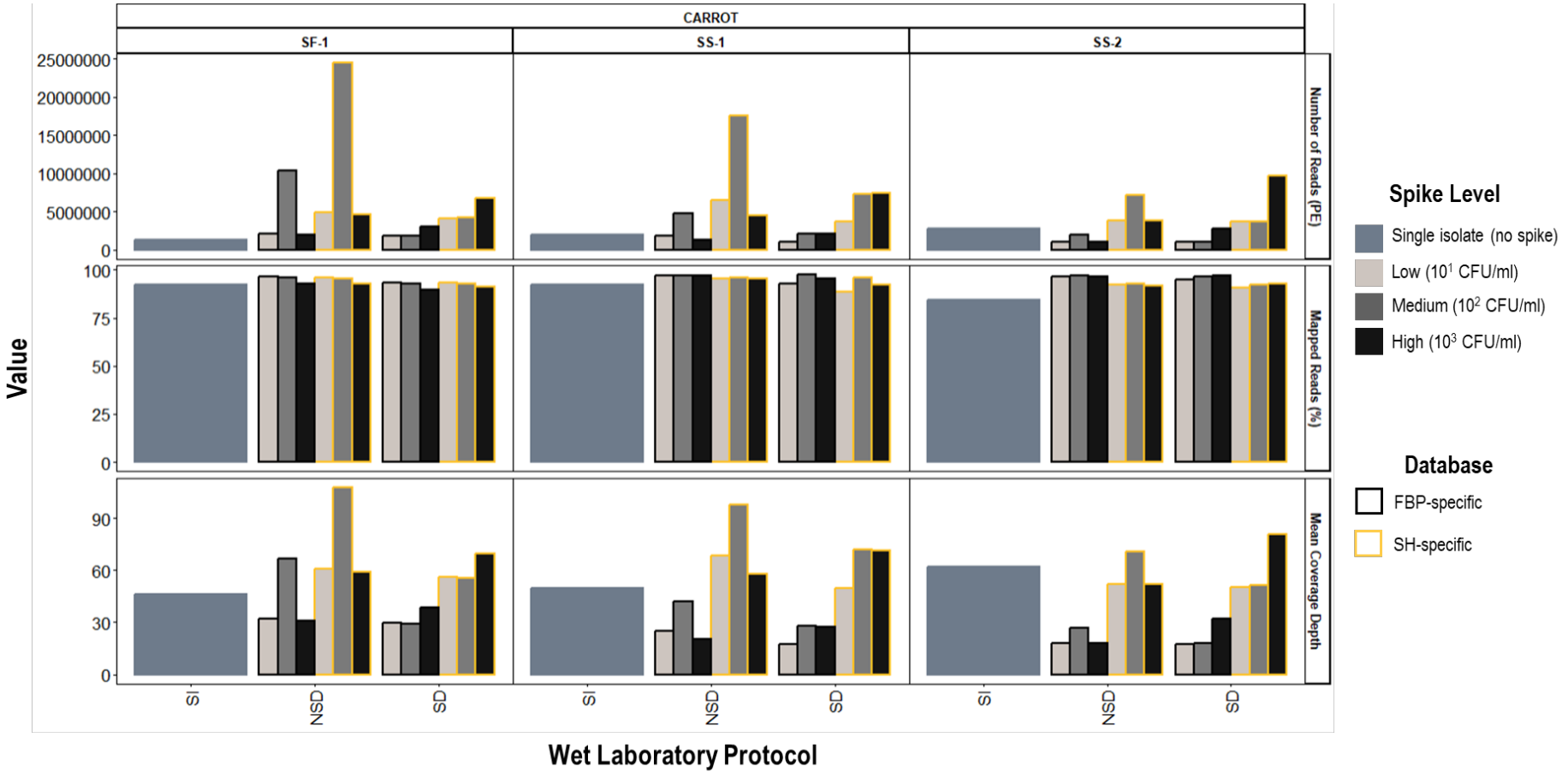


Figure 4.11. Comparative genome coverage metrics for *Shigella* species across protocols. This figure evaluates the quality of sequence data used for assembly by comparing quasi-metagenomic datasets comprising reads extracted via the Foodborne pathogen (FBP)-specific reference database and *Shigella* (SH)- specific reference database against single isolate (SI) reference controls. Performance is assessed across non-saponin-depleted (NSD) and saponin-depleted (SD) protocols for three contamination levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL) in a single food commodity: Carrot. Panels display Qualimap-derived metrics including the Number of Reads (PE) successfully assigned to the target pathogen, the percentage of Mapped Reads (%) aligning to the corresponding reference genome, and the Mean Coverage Depth. Datasets are identified by species code: (SF-1 = *S. flexneri*; SS-1, 2 = *S. sonnei*).

mean depth (p=0.499). Consistent with the VTEC datasets, the choice of reference database significantly affected genome coverage metrics for *Shigella* datasets. For *S. sonnei*, percent mapped reads (NSD p=0.00575; SD p=0.0313) and mean coverage depth (NSD p<0.001; SD p<0.001) differed significantly between reference databases (Figure 4.11). For *S. flexneri*, percent mapped reads did not differ between reference databases for NSD (p=0.104) or SD datasets (p=0.686), although mean coverage depth differed significantly for both NSD (p=0.0158) and SD datasets (p=0.00341). As expected, the SH-specific reference database yielded metrics most consistent with the corresponding SI reference strains (Figure 4.11).

4.2.7 Pathogen detection and characterization

4.2.7.1 RefSeq Masher provides a rapid preliminary screen of pre-processed datasets and can inform the selection of appropriate reference databases for subsequent taxonomic classification.

Using RefSeq Masher (contains) as a rapid screening tool for preprocessed quasi-metagenomic datasets, *L. monocytogenes* was consistently identified among the top ten hits, with >380/400 shared hashes and an average percent identity ≥99%. In datasets spiked with other *Listeria* species, including *L. innocua* and *L. welshimeri*, the target organism also appeared within the top ten results, with >325/400 shared hashes and ≥98% identity. By comparison, *L. seeligeri* datasets showed fewer shared hashes (>250/400) and a lower average percent identity (~97%) across samples. Notably, *L. seeligeri* (or any other *Listeria* species) did not appear among the top ten hits in low spike-in datasets under any wet-laboratory protocol; instead, only *Pseudomonas* species were detected.

In the preprocessed VTEC quasi-metagenomic datasets, *E. coli* was consistently detected within the top ten hits, with ≥398/400 shared hashes and ≥99% identity. The *Shigella* spike-in datasets produced top ten matches to both *Shigella* species and *E. coli*, each showing 100% identity and 400/400 shared hashes, which reflects the extensive genomic similarity between these organisms.

Importantly, none of the spike-in pathogens were detected among the top ten hits in the corresponding NC samples. Collectively, these results indicate that preliminary screening with RefSeq Masher can help guide reference database selection for subsequent taxonomic classification. This is particularly useful when analyzing closely related organisms such as *E. coli* and *Shigella*, where overlapping genomic signatures may lead to ambiguous assignments. In such cases, species-focused reference databases (e.g., EC-specific or SH-specific) may be necessary to improve taxonomic resolution.

4.2.7.2 Pathogen-presence likelihood metric for accurate pathogen identification

Figures 4.12–4.14 summarize the pathogen-presence likelihood metric, integrating unique KPMR and robust *z*-scores across all quasi-metagenomic datasets. In *Listeria*-spiked matrices (Figure 4.12), *Listeria* spp. consistently produced high unique KPMR values under both NSD and SD workflows, while non-target taxa remained low. Signal intensity was comparable between laboratory protocols, indicating that SD had little effect on pathogen-specific *k*-mer recovery, consistent with earlier observations (Section 4.2.3). Robust *z*-scores followed the same pattern, with noticeable enrichment of *Listeria* taxa above background. A greater dispersion in *z*-score values was observed in salad kit samples compared with deli meats (ham, roast beef, turkey), likely reflecting the higher microbial diversity and heterogeneity of fresh produce. In all cases, the highest unique KPMR corresponded to the true spike strain, and magnitude-based ranking alone was sufficient for pathogen identification. A similar pattern was observed for VTEC in ground beef and lettuce (Figure 4.13). VTEC strains consistently demonstrated high unique KPMR values across spike levels and wet laboratory protocols (Figure 4.13A) compared to non-target taxa, which showed markedly lower absolute signal. Within-sample robust *z*-scores (Figure 4.13B) confirmed enrichment of VTEC strains, however, the separation in unique KPMR values was generally sufficient to identify the correct pathogen without reliance on *z*-score discrimination.

In contrast, *Shigella*-spiked carrot datasets (Figure 4.14) highlighted the challenge of resolving closely related taxa. Both *Shigella* spp. and *E. coli* displayed similarly high unique KPMR values, reflecting extensive genomic overlap (Figure 4.14A). Robust *z*-scores showed enrichment of *S. flexneri* and *E. coli* in SF-1 samples and enrichment of *S. sonnei* and *E. coli* in SS-2 samples (Figure 4.14B). In nearly all samples, *E. coli* exhibited higher *z*-scores than the corresponding *Shigella* strain. To address this, an override rule was applied: when *Shigella* unique KPMR reached at least 20% of the competing *E. coli* value, the highest-ranking *Shigella* species was reported as the most likely pathogen. This adjustment preserved detection accuracy while improving species-level attribution for *Shigella* strains.

Overall, application of the pathogen-presence likelihood metric supported accurate pathogen identification across all quasi-metagenomic datasets (Figure 4.15). Additionally, the framework addressed challenges associated with closely related taxa, including differentiation of VTEC from *S. sonnei* and *S. flexneri*, as well as discrimination among species within the *Listeria* genus. In contrast, none of the negative control (NC) samples across the evaluated food matrices met minimum coverage requirements for metric generation (Figures 4.12–4.14), indicating no detectable pathogen signal under the defined criteria used in this study.

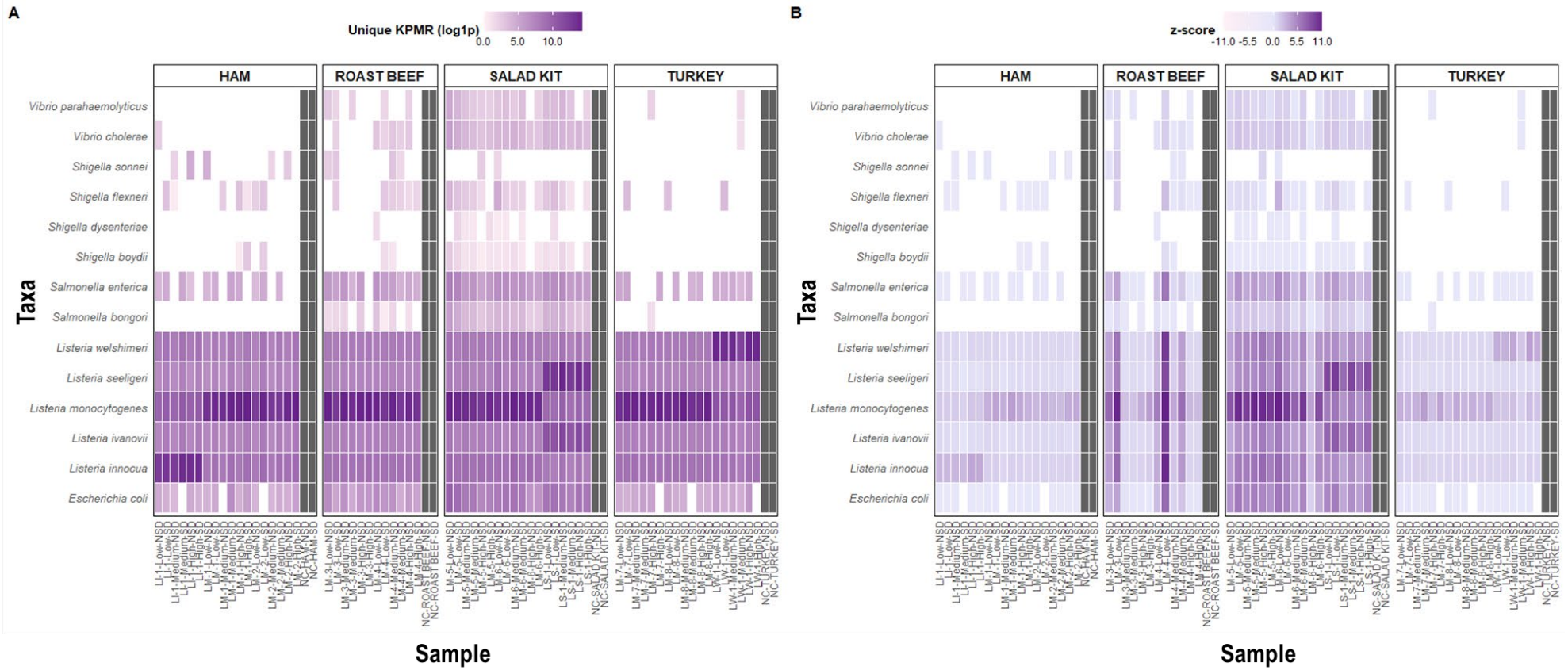


Figure 4.12. Detection signal and within-sample enrichment in quasi-metagenomic datasets from food matrices spiked with *Listeria* strains. Heatmaps display depth-normalized unique *k*-mer counts (KPMR; log1p-transformed; panel A) and corresponding within-sample robust *z*-scores (panel B). In panel A, increasing purple intensity represents higher absolute abundance of pathogen-specific unique *k*-mers (total detection signal). In panel B, a diverging color scale centered at zero illustrates relative enrichment within each sample, with darker purple indicating taxa enriched above the sample-specific background. Datasets were generated from artificially contaminated ham, roast beef, salad kit, and turkey matrices at low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL) concentrations under two laboratory workflows: no saponin depletion (NSD) and saponin depletion (SD). Negative controls (NC) were included to assess baseline signal. Grey tiles denote samples that failed minimum sequencing depth requirements, while white tiles indicate absence of detected signal. Strain abbreviations are as follows: LI = *L. innocua*; LM-1 through LM-8 = *L. monocytogenes*; LS = *L. seeligeri*; and LW = *L. welshimeri*. Metrics were derived from KrakenUniq outputs using the optimized Foodborne Pathogen–specific reference database.

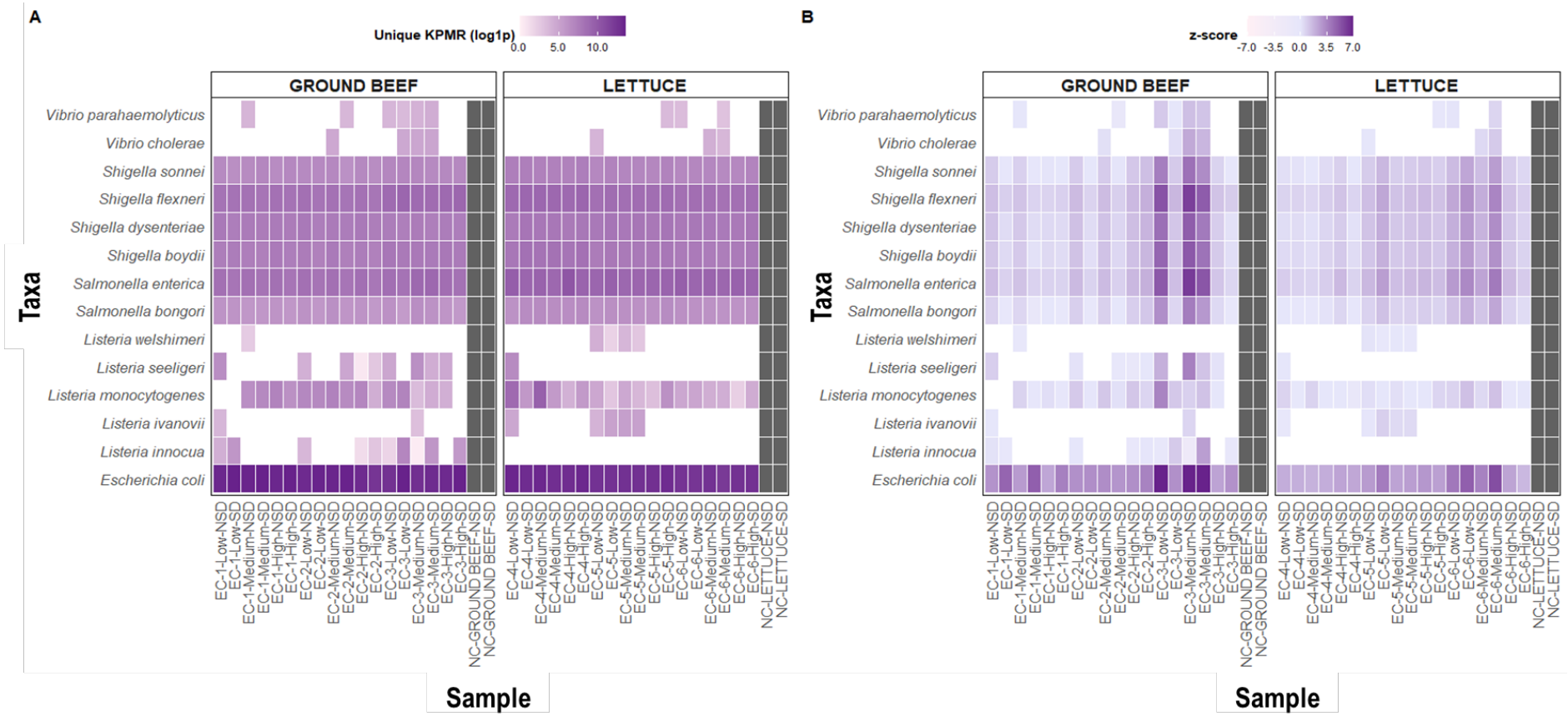


Figure 4.13. Detection signal and within-sample enrichment in quasi-metagenomic datasets from food matrices spiked with verotoxigenic *Escherichia coli* (VTEC) strains. Heatmaps display depth-normalized unique *k*-mer counts (KPMR; log1p-transformed; panel A) and corresponding within-sample robust *z*-scores (panel B). In panel A, increasing purple intensity represents higher absolute abundance of pathogen-specific unique *k*-mers (total detection signal). In panel B, a diverging color scale centered at zero illustrates relative enrichment within each sample, with darker purple indicating taxa enriched above the sample-specific background. Datasets were generated from artificially contaminated ground beef and lettuce matrices at low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL) concentrations under two laboratory workflows: no saponin depletion (NSD) and saponin depletion (SD). Negative controls (NC) were included to assess baseline signal. Grey tiles denote samples that failed minimum sequencing depth requirements, while white tiles indicate absence of detected signal. Strains include EC-1, EC-2, EC-5, and EC-6 (*E. coli* O157:H7), EC-3 (*E. coli* O26:H11), and EC-4 (*E. coli* O103:H25). Metrics were derived from KrakenUniq outputs using the optimized Foodborne Pathogen–specific reference database.

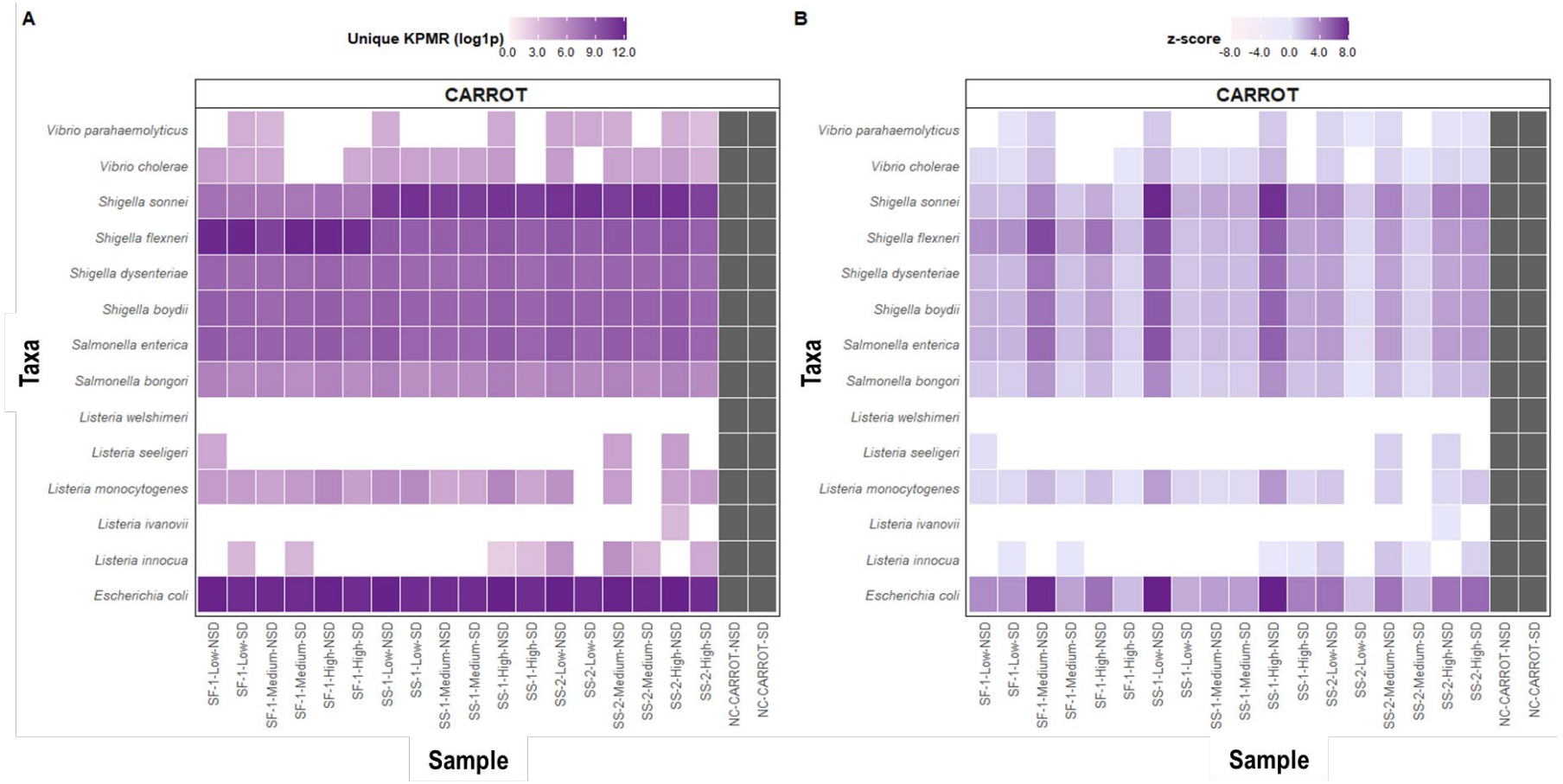


Figure 4.14. Detection signal and within-sample enrichment in quasi-metagenomic datasets from food matrices spiked with *Shigella* strains. Heatmaps display depth-normalized unique *k*-mer counts (KPMR; log1p-transformed; panel A) and corresponding within-sample robust *z*-scores (panel B). In panel A, increasing purple intensity represents higher absolute abundance of pathogen-specific unique *k*-mers (total detection signal). In panel B, a diverging color scale centered at zero illustrates relative enrichment within each sample, with darker purple indicating taxa enriched above the sample-specific background. Datasets were generated from artificially contaminated ground beef and lettuce matrices at low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL) concentrations under two laboratory workflows: no saponin depletion (NSD) and saponin depletion (SD). Negative controls (NC) were included to assess baseline signal. Grey tiles denote samples that failed minimum sequencing depth requirements, while white tiles indicate absence of detected signal. Strains include SF-1 (*S. flexneri*) and SS-1, 2 (*S. sonnei*). Metrics were derived from KrakenUniq outputs using the optimized Foodborne Pathogen–specific reference database.

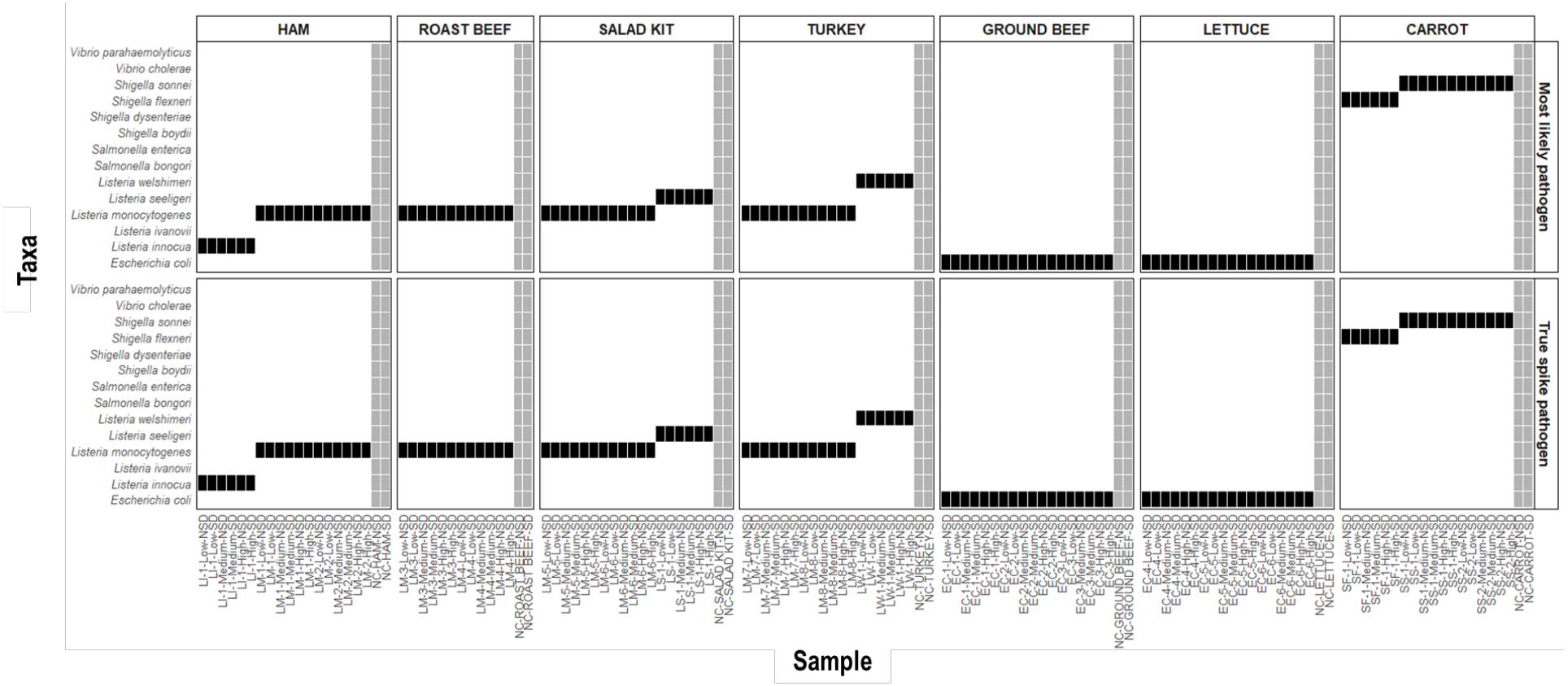


Figure 4.15. Most likely pathogen assignments across all quasi-metagenomic datasets as determined by the pathogen-presence likelihood metric. Black tiles indicate the taxon designated as the most likely pathogen for each sample following application of the pathogen-presence likelihood metric, which incorporates unique KPMR thresholds, robust within-sample *z*-score scaling, and predefined tie-resolution rules. The true spike pathogen is shown for comparison in the lower panel. Samples are stratified by food matrix (ham, roast beef, salad kit, turkey, ground beef, lettuce, and carrot) and include low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL) contamination levels processed under both no saponin depletion (NSD) and saponin depletion (SD) protocols. Samples failing minimum sequencing depth requirements are indicated in grey.

4.2.7.3 ECTyper delivers robust *in silico* serotype prediction and speciation for verotoxigenic *Escherichia coli* (VTEC) and *Shigella*, while RefSeq Masher Matches provides dependable species-level verification for *Listeria* species.

Serotype prediction of VTEC assemblies using ECTyper demonstrated that the FBP-specific reference database accurately identified O-antigens across all evaluated serotypes (O157:H7, O103:H25, and O26:H11). However, performance was substantially reduced for H-antigen prediction (Table 4.1). Specifically, the FBP-specific workflow failed to produce *in silico* H-antigen predictions for the majority of non-O157 assemblies. In contrast, the EC-specific database achieved 100% accuracy for both O and H antigens across all 36 assemblies, regardless of the wet-laboratory protocol (NSD or SD) utilized (Table 4.1). These results indicate that a species-specific read extraction workflow is essential for achieving high-resolution subtyping for VTEC.

In silico species prediction for *Shigella* assemblies using ECTyper demonstrated perfect accuracy with 100% of strains correctly identified as either *S. sonnei* or *S. flexneri* (Table 4.2). Unlike the serotype results for VTEC assemblies, both the FBP-specific and SH-specific reference databases performed identically, yielding 18 correct predictions. Furthermore, this consistency was maintained regardless of the wet-laboratory protocol, as both non-saponin-depleted (NSD) and saponin-depleted (SD) samples achieved species-level classification (Table 4.2). These findings indicate that for *Shigella*, a broad-spectrum reference database provides sufficient genomic depth for reliable *in silico* species identification.

For *Listeria* assemblies, the RefSeq Masher analysis correctly reported the target pathogen as the top result in 100% of the *Listeria* spike-in datasets. For all species tested, including *L. monocytogenes*, *L. innocua*, *L. seeligeri*, and *L. welshimeri*, the minimum number of shared hashes remained high and consistent between NSD and SD protocols (Table 4.3). Specifically, *L. monocytogenes* assemblies showed the highest shared hash counts (ranging from 354 to 390), compared to only 175 shared hashes for *L. seeligeri*.

4.2.7.4 Data quality assessment using BioNumerics

To assess whether the sequence data was of sufficient quality for cluster detection using the pathogen-specific wgMLST schemas implemented in the commercial BioNumerics platform, all datasets were evaluated in accordance with PNC guidelines for sequence quality assessment (Appendix 4.4). All quality metrics generated for *L. monocytogenes* pathogen-specific reads fully satisfied PNC criteria (Figure 4.16).

Table 4.1. *In silico* serotyping performance for verotoxigenic *Escherichia coli* (VTEC) using quasi-metagenomic assemblies.

			O antigen				H antigen			
Reference Database	Wet-Lab	Serotype	Perfect Match	Ambiguous Match	Incorrect Prediction	No in silico prediction	Perfect Match	Ambiguous Match	Incorrect Prediction	No in silico prediction
FBP-specific	NSD	O157:H7	12	0	0	0	12	0	0	0
		O103:H25	3	0	0	0	0	0	0	3
		O26:H11	3	0	0	0	1	0	0	2
	SD	O157:H7	12	0	0	0	12	0	0	0
		O103:H25	3	0	0	0	0	0	0	3
		O26:H11	3	0	0	0	0	0	0	3
EC-specific	NSD	O157:H7	12	0	0	0	12	0	0	0
		O103:H25	3	0	0	0	3	0	0	0
		O26:H11	3	0	0	0	3	0	0	0
	SD	O157:H7	12	0	0	0	12	0	0	0
		O103:H25	3	0	0	0	3	0	0	0
		O26:H11	3	0	0	0	3	0	0	0

*This table summarizes the accuracy of ECTyper in predicting O and H antigens from *Escherichia*-specific reads extracted via the FBP-specific and EC-specific reference databases. Reads were classified using KrakenUniq (v0.6), assembled using shovill with SPAdes (v1.1.0), and evaluated across non-saponin-depleted (NSD) and saponin-depleted (SD) protocols.

Table 4.2. *In silico* serotyping performance for *Shigella* strains using quasi-metagenomic assemblies.

Reference Database	Wet-Lab	Species	ECTyper Species Prediction	
			Correct	Incorrect
FBP-specific	NSD	S. sonnei	6	0
		S. flexneri	3	0
	SD	S. sonnei	6	0
		S. flexneri	3	0
SH-specific	NSD	S. sonnei	6	0
		S. flexneri	3	0
	SD	S. sonnei	6	0
		S. flexneri	3	0

*This table summarizes the accuracy of ECTyper in species prediction for *Shigella* reads extracted via the FBP-specific and SH-specific reference databases. Reads were classified using KrakenUniq (v0.6), assembled using shovill with SPAdes (v1.1.0), and evaluated across non-saponin-depleted (NSD) and saponin-depleted (SD) protocols.

Table 4.3. RefSeq Masher Matches Results for *Listeria* strains using quasi-metagenomic assemblies.

Reference Database	Pathogen	Food	Wet-Lab	Minimum Number of Shared Hashes	Pathogen Reported as Top Result?
FBP-specific	*L. monocytogenes*	Ham	NSD	382	Yes
			SD	382	Yes
		Roast Beef	NSD	385	Yes
			SD	385	Yes
		Salad Kit	NSD	354	Yes
			SD	354	Yes
		Turkey	NSD	390	Yes
			SD	390	Yes
	L. innocua	Ham	NSD	285	Yes
			SD	285	Yes
	L. seeligeri	Salad Kit	NSD	175	Yes
			SD	175	Yes
	L. welshimeri	Turkey	NSD	301	Yes
			SD	300	Yes

**Listeria*-specific reads were extracted following taxonomic classification with KrakenUniq (v0.6) and assembled using shovill with spades (v1.1.0).

For the additional *Listeria* species examined, the quality metrics were consistent with those of the corresponding reference strains and met accepted thresholds for average sequence quality, number of contigs, and genome length. However, the number of consensus allele calls (NrConsensus), representing alleles concordant between assembly-free and assembly-based algorithms and the percent core, defined as the proportion of loci classified as core genome targets, were substantially below the accepted guidelines (Figure 4.16). This outcome was expected, as the wgMLST scheme integrated into BioNumerics is specifically optimized for *L. monocytogenes* and not for other *Listeria* species, which are not routinely included in PNC surveillance. Nonetheless, the ability to differentiate these *Listeria* species from *L. monocytogenes* remains relevant from clinical, epidemiological, and food-processing perspectives, particularly for environmental monitoring and contamination-source tracking.

In contrast to the *Listeria* datasets, only pathogen-specific reads extracted using the EC-specific and SH-specific reference databases met the required BioNumerics quality thresholds for VTEC (Figure 4.17) and *Shigella* (Figure 4.18), respectively. Based on these findings, pathogen-specific reads generated using the EC-specific and SH-specific reference databases were subsequently used for outbreak detection analyses.

4.2.8 Outbreak detection using short-read quasi-metagenomic datasets

4.2.8.1 Whole-genome multilocus sequence typing (wgMLST)

Only datasets that passed the BioNumerics data quality assessment and were linked to known outbreaks were evaluated using wgMLST and SNVPhyl. This included *L. monocytogenes*, VTEC O157:H7, and *Shigella sonnei*. For nearly all outbreaks examined, the number of wgMLST alleles detected in the original outbreak isolates was fully concordant with the number detected from pathogen-specific reads extracted from the quasi-metagenomic datasets (Table 4.4). One additional allele was identified in the VTEC outbreak associated with lettuce and in the *Shigella* outbreak associated with carrots (Table 4.4); however, these values remained within the allele ranges considered related according to PNC guidelines (152).

To further assess concordance, allele differences between each quasi-metagenomic dataset and its corresponding SI reference strain were examined. For outbreaks involving *L. monocytogenes*, no wgMLST allele differences were observed except for strain LM-2, for which the quasi-metagenomic dataset differed by a single allele (Figure 4.19).

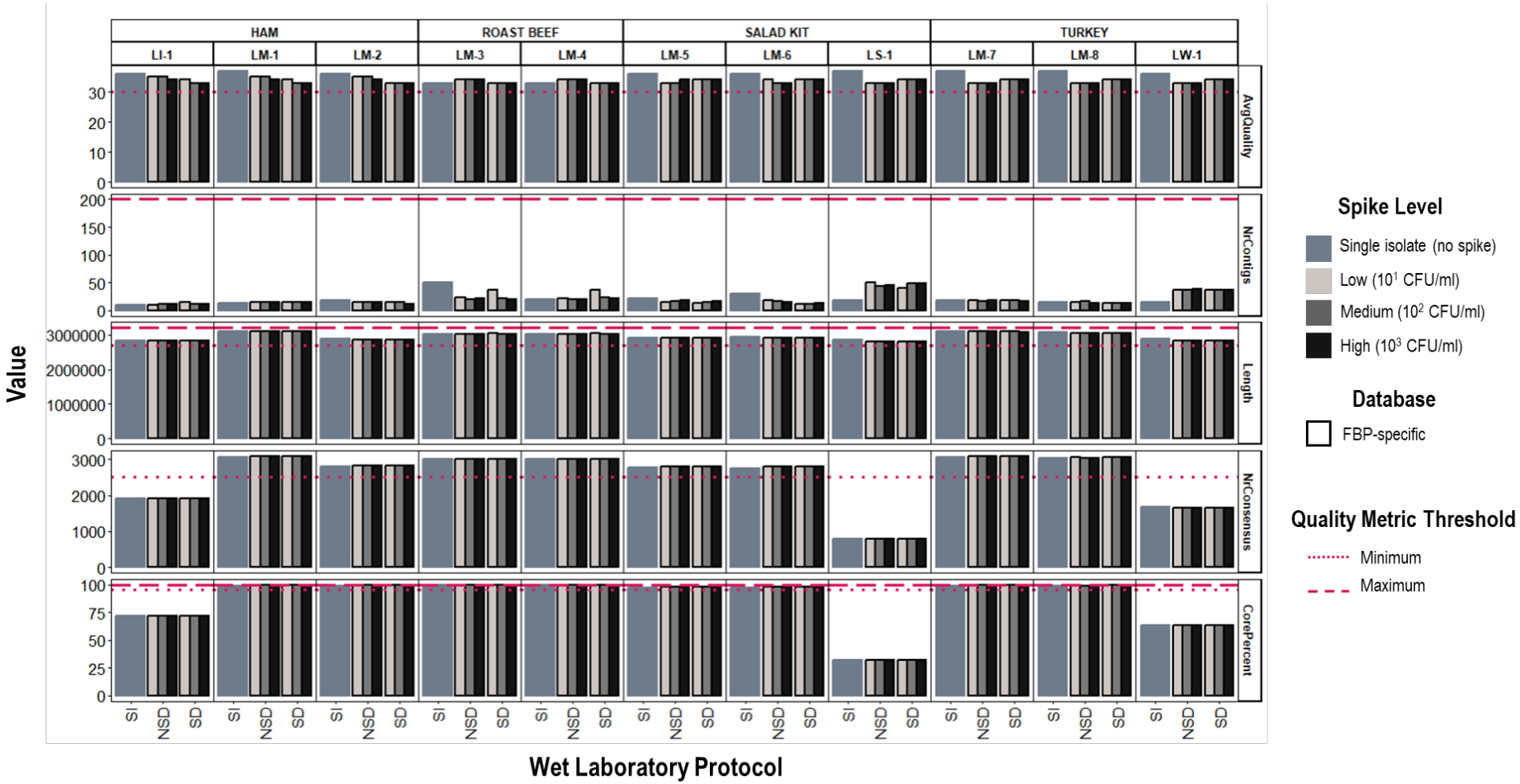


Figure 4.16. Comparative assembly performance for *Listeria*-specific sequences extracted from quasi-metagenomic datasets. Metrics were evaluated across non-saponin-depleted (NSD) and saponin-depleted (SD) protocols at three contamination levels, low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL), and compared against single-isolate (SI) reference strains. Assembly and subtyping metrics were generated in BioNumerics (v7.6) using *Listeria*-specific reads extracted from Ham, Roast Beef, Salad Kit, and Turkey matrices using the foodborne pathogen (FBP-specific) reference database. Datasets are identified by strain codes: *Listeria* species (LI = *L. innocua*; LM-1 through 8 = *L. monocytogenes*; LS = *L. seeligeri*; LW = *L. welshimeri*). Data quality indicators include: AvgQuality (mean Phred quality scores of raw reads), NrContigs (total number of contigs in the SPAdes assembly), Length (total assembly size in base pairs), NrConsensus (count of identical allele calls between independent identification algorithms), and CorePercent (percentage of identified core wgMLST loci). PulseNet Canada organism-specific thresholds for assessing the quality of sequence data are indicated by dotted (minimum threshold) and dashed (maximum threshold) lines.

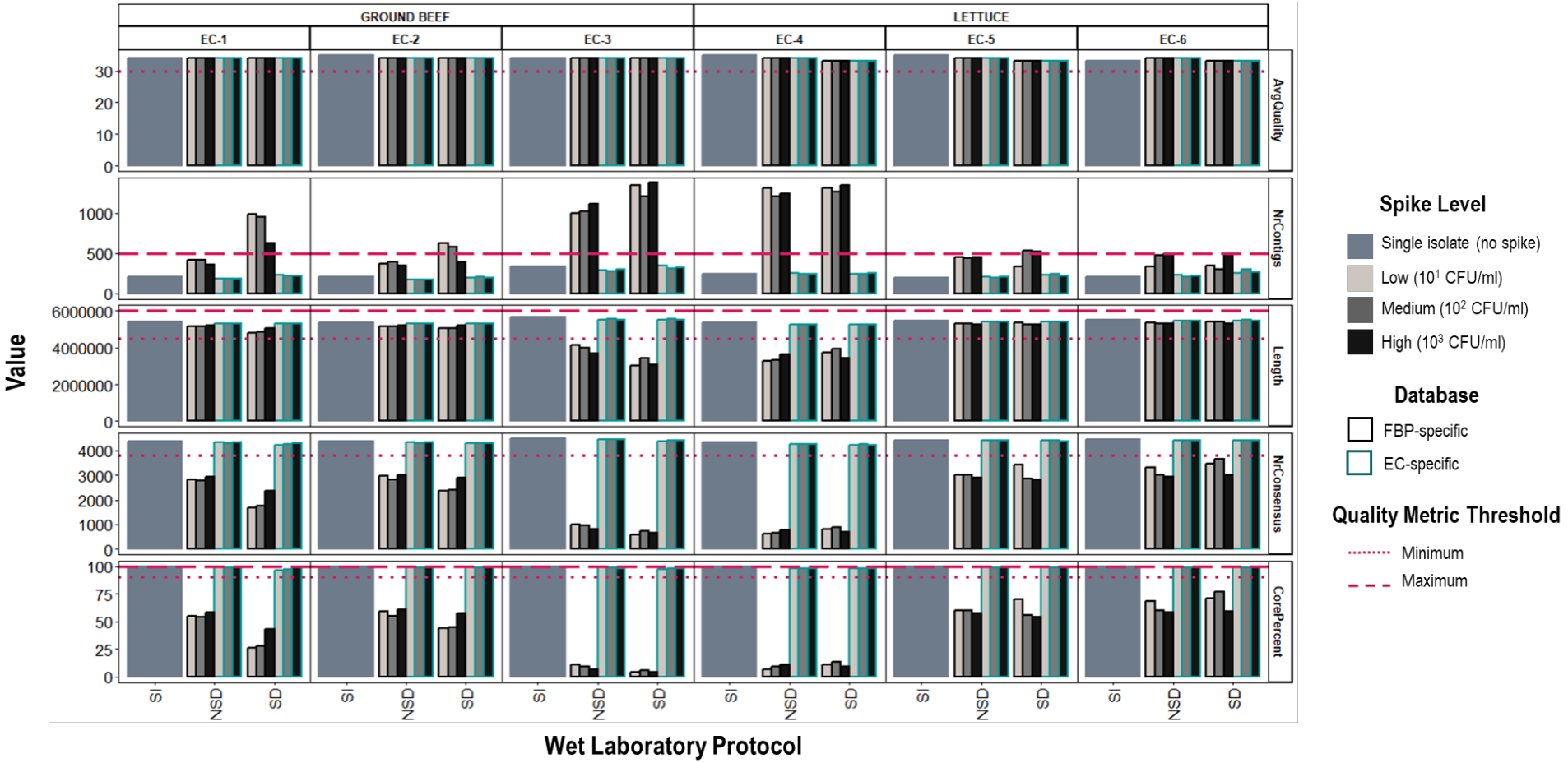


Figure 4.17. Comparative assembly performance for verotoxigenic *Escherichia coli* (VTEC)-specific sequences extracted from quasi-metagenomic datasets. Metrics were evaluated across non-saponin-depleted (NSD) and saponin-depleted (SD) protocols at three contamination levels, low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL), and compared against single-isolate (SI) reference strains. Assembly and subtyping metrics were generated in BioNumerics (v7.6) using pathogen-specific reads extracted from ground beef and lettuce matrices using either the foodborne pathogen (FBP-specific) or *Escherichia coli* (EC-specific) reference database. Datasets are identified by strain codes: EC-1, 2, 5, 6 (O157:H7); EC-3 (O26:H11); and EC-4 (O103:H25). Data quality indicators include: AvgQuality (mean Phred quality scores of raw reads), NrContigs (total number of contigs in the SPAdes assembly), Length (total assembly size in base pairs), NrConsensus (count of identical allele calls between independent identification algorithms), and CorePercent (percentage of identified core wgMLST loci). PulseNet Canada organism-specific thresholds for assessing the quality of sequence data are indicated by dotted (minimum threshold) and dashed (maximum threshold) lines.

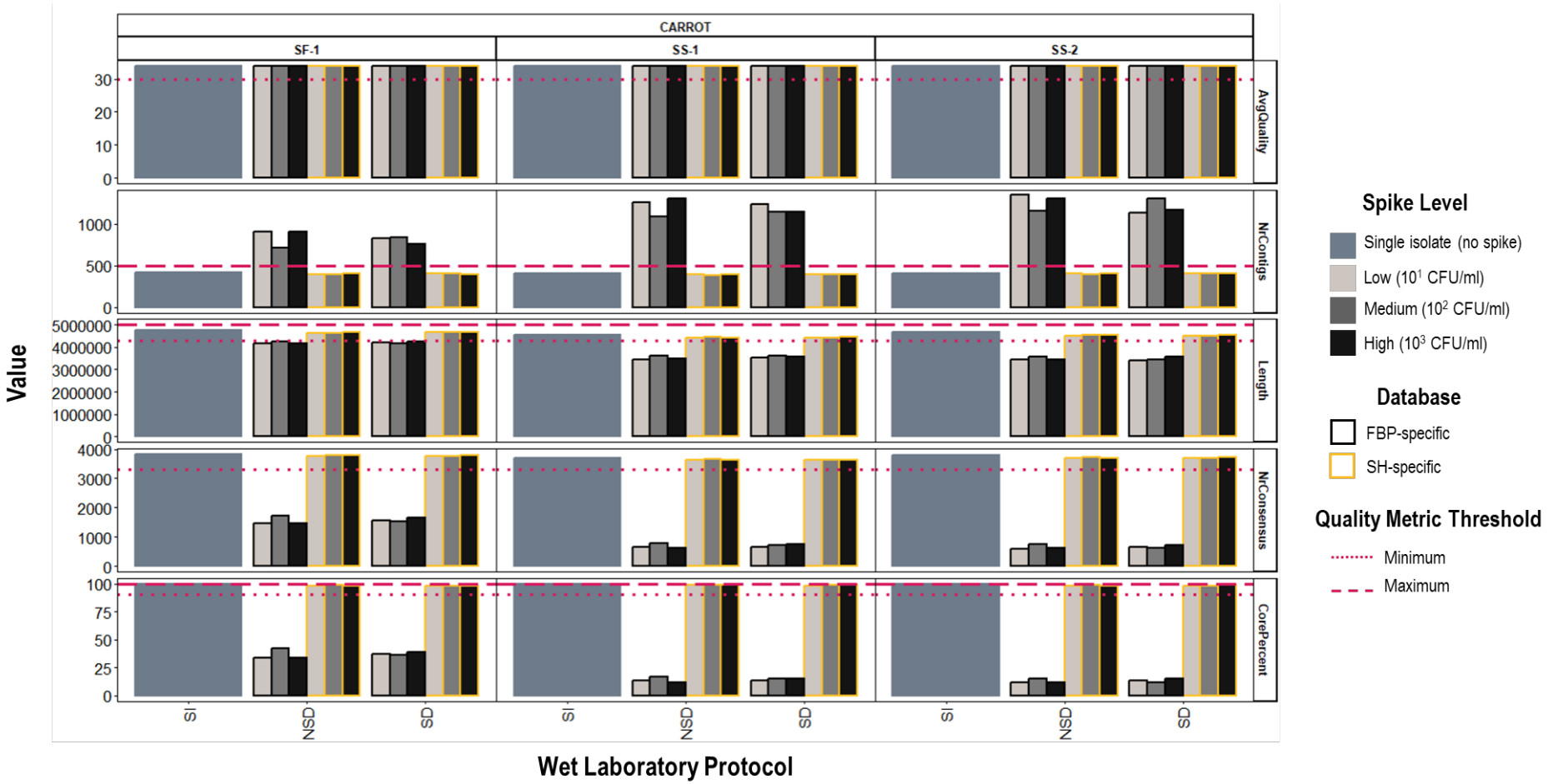


Figure 4.18. Comparative assembly performance for *Shigella*-specific sequences extracted from quasi-metagenomic datasets. Metrics were evaluated across non-saponin-depleted (NSD) and saponin-depleted (SD) protocols at three contamination levels, low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL), and compared against single-isolate (SI) reference strains. Assembly and subtyping metrics were generated in BioNumerics (v7.6) using *Shigella*-specific reads extracted from Carrots using either the foodborne pathogen (FBP-specific) or *Shigella* (SH-specific) reference database. Datasets are identified by species code: (SF-1 = *S. flexneri*; SS-1, 2 = *S. sonnei*). Data quality indicators include: AvgQuality (mean Phred quality scores of raw reads), NrContigs (total number of contigs in the SPAdes assembly), Length (total assembly size in base pairs), NrConsensus (count of identical allele calls between independent identification algorithms), and CorePercent (percentage of identified core wgMLST loci). PulseNet Canada organism-specific thresholds for assessing the quality of sequence data are indicated by dotted (minimum threshold) and dashed (maximum threshold) lines.

Table 4.4. Comparison of genomic distance metrics for simulated outbreak detection between original outbreak isolates and the quasi-metagenomic datasets.

Pathogen	Study Strains	Implicated Food Product	Total Number of Isolates (Clinical/ Food/ Environmental)	ORIGINAL OUTBREAK		QUASI-METAGENOMIC DATASET	
				SNVs (SNVPhyl)	wgMLST alleles (BioNumerics)	SNVs (SNVPhyl)	wgMLST alleles (BioNumerics)
L. monocytogenes	LM-2	Ham*	3 (1/1/0)	0	0-1	0	0-1
	LM-3, LM-4	Roast Beef	28 (4/24/0)	0-11	0-10	0-11	0-10
	LM-5, LM-6	Salad Kit	19 (11/8/0)	0-4	0-2	0-4	0-2
	LM-7, LM-8	Turkey	65 (54/9/2)	0-6	0-5	0-6	0-5
E. coli O157:H7	EC-1, EC-2	Ground Beef	53 (29/24/0)	0-2	0-7	0-3	0-7
	EC-5, EC-6	Lettuce	31 (31/0/0)	0-12	0-13	0-12	0-14
S. sonnei	SS-1, SS-2	Carrot	14 (14/0/0)	0-2	0-6	0-2	0-7

*Metrics were evaluated using whole-genome multilocus sequence typing (wgMLST) and single nucleotide variant (SNV) phylogenomics to assess the resolution of the quasi-metagenomic datasets across three pathogens. *Analysis ran with non-outbreak isolate LM-1; >200 wgMLST alleles and >6000 SNVs from outbreak isolates.

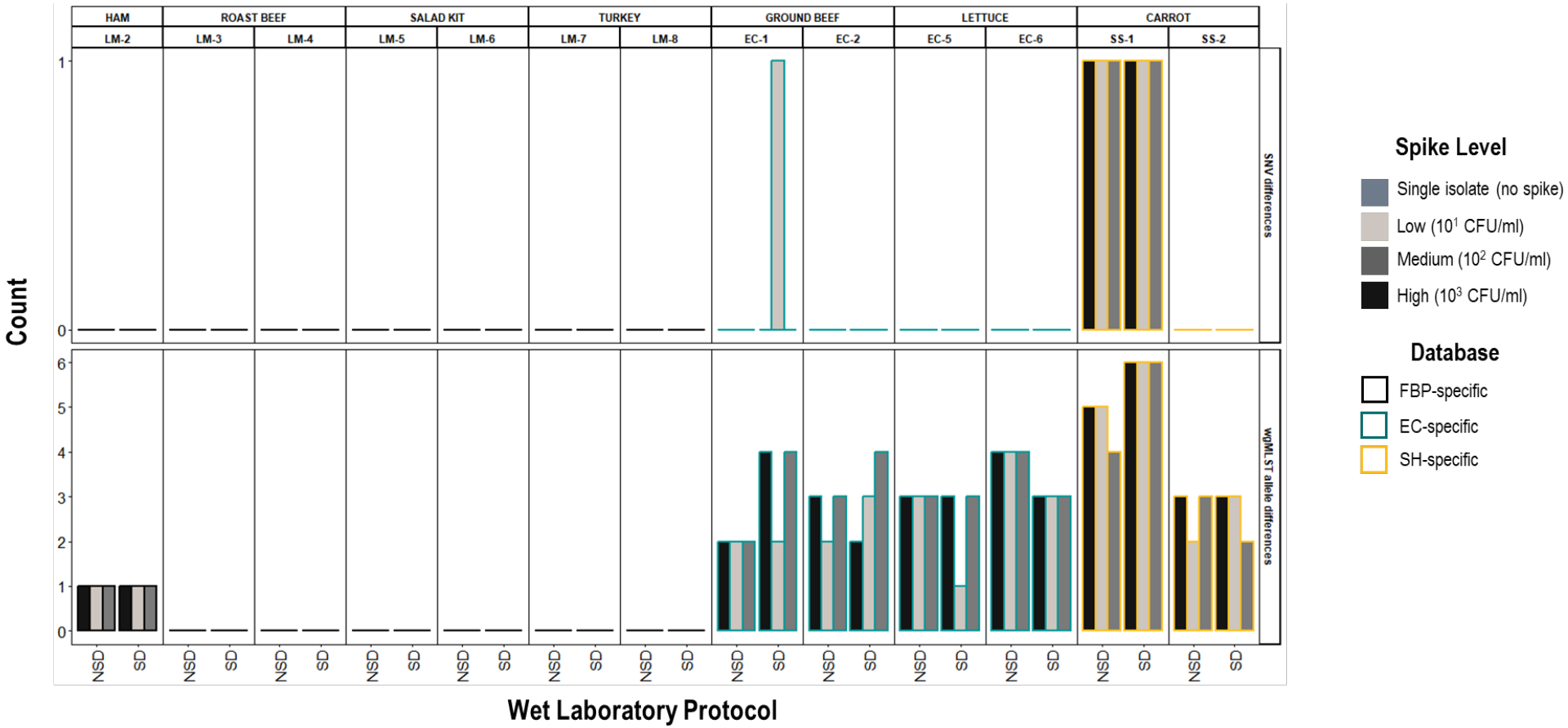


Figure 4.19. Single-nucleotide variant (SNV) and whole-genome multilocus sequence typing (wgMLST) allele differences between single-isolate (SI) reference genomes and short-read quasi-metagenomic datasets across pathogen-food matrices. Genomic variation was evaluated using SNVPhyl (v1.1) and wgMLST in BioNumerics (v7.6), with SNV differences shown in the upper panel and wgMLST allele differences in the lower panel. Quasi-metagenomic datasets were generated using non–saponin-depleted (NSD) and saponin-depleted (SD) workflows at three contamination levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL), and compared with the corresponding single isolate (SI) reference genomes. Datasets are labeled by strain code (LM-1–8 = *L. monocytogenes*; EC-1, EC-2, EC-5, and EC-6 = *E. coli* O157:H7; SS-1 and SS-2 = *S. sonnei*).

In comparison, up to four allele differences were identified for outbrea*k*-associated VTEC O157:H7 datasets (Figure 4.19), and up to six allele differences were detected for *S. sonnei* datasets (Figure 4.19). Overall, the allele differences fell within acceptable thresholds based on current PNC guidelines (144).

4.2.8.2 Single nucleotide variant phylogenomic pipeline (SNVPhyl)

For *L. monocytogenes* outbreak strains, no SNV differences were observed between the quasi-metagenomic datasets and their corresponding previously sequenced SI reference strains (Figure 4.19). In contrast, up to one SNV was detected for outbrea*k*-associated VTEC O157:H7 and *S. sonnei* datasets (Figure 4.19). Importantly, all SNV differences fell within the acceptable thresholds defined by current PNC guidelines (152).

All datasets met the minimum requirement of ≥10X coverage across at least 80% of the reference genome and were therefore considered suitable for genomic relatedness inference under PNC guidelines. The number of SNVs identified between the original outbreak isolates and the pathogen-specific reads extracted from the quasi-metagenomic datasets were identical and consistent with the wgMLST allele ranges, with one exception. For the VTEC outbreak associated with ground beef, a single additional SNV was detected in the quasi-metagenomic dataset (0–3 SNVs) compared to the original outbreak isolates (0–2 SNVs) (Table 4.4). This difference was negligible, as it remained within the SNV ranges considered related according to PNC guidelines (152).

4.3 Discussion

This study demonstrates that quasi-metagenomic sequencing can recover strain-level genomic information directly from enriched food matrices and, when paired with curated reference databases and appropriate analytical workflows, can achieve a resolution comparable to whole-genome sequencing (WGS) of pure isolates. The central objective was to determine whether this strategy can accelerate the generation of actionable genomic data without compromising subtyping accuracy. The results support this aim: across *L. monocytogenes*, VTEC, and *Shigella* spp., quasi-metagenomic assemblies consistently captured the features required for foodborne pathogen surveillance, in line with recent reports suggesting that quasi-metagenomic sequencing can deliver high-resolution results earlier than traditional culture-based approaches (58–61).

Overall, saponin depletion increased the relative proportion of microbial DNA and reduced eukaryotic background, consistent with previous reports demonstrating that eukaryotic depletion can

improve microbial read depth in metagenomic datasets (65,153,154). Though not unexpected, produce samples yielded fewer eukaryotic reads, which is consistent with the reduced susceptibility of plant cells to lysis. This likely reflects two factors: the presence of rigid cell walls that limit efficient cell disruption, and the limited effectiveness of saponin-based depletion methods, which target cholesterol-rich membranes. Since plant membranes are enriched in phytosterols (e.g., sitosterol, stigmasterol) rather than cholesterol, they are less susceptible to saponin-mediated lysis (67,155). For plant cells that did lyse, membrane integrity was likely compromised by extensive processing of the food product (e.g., prepackaged salad kits), which typically involves cutting, washing with sanitizers, and mechanical drying steps that can disrupt plant tissue structure and cellular integrity (156,157).

Although saponin-based host DNA depletion can improve the recovery of pathogen-derived sequences, its effectiveness appears to be highly matrix-dependent and may vary according to host DNA burden, microbial load, and sample composition. In the present study, once enrichment conditions were optimized, reductions in eukaryotic reads did not consistently translate into increased pathogen-specific read counts or unique *k*-mer recovery compared with non-depleted (NSD) samples. Similar observations have been reported in clinical studies, where enrichment has been shown to exert a greater influence on pathogen signal than eukaryotic DNA depletion alone (158). Collectively, these findings indicate that when enrichment effectively amplifies the target organism, eukaryotic DNA depletion contributes only marginally to pathogen-specific sequence recovery (65,159,160). The limited benefits observed suggest that saponin depletion may be most advantageous in matrices where pathogens remain at low abundance despite enrichment, eukaryotic DNA represents a substantial proportion of the total sequencing output, and background microbial populations are relatively low, allowing removal of eukaryotic DNA to meaningfully increase the relative representation of the target pathogen. Furthermore, modest reductions in pathogen signal observed in some *Shigella* datasets indicate that the additional chemical and mechanical processing steps associated with depletion may contribute to partial biomass loss. This effect may be more pronounced for Gram-negative organisms, which tend to be more susceptible to cell disruption during pre-extraction depletion procedures (153). Overall, these findings highlight the importance of evaluating eukaryotic DNA depletion strategies on a pathogen- and matrix-specific basis, as the benefits and potential biases of depletion are unlikely to be uniform across different organisms and food commodities. Accordingly, a tiered approach may be most appropriate, whereby host DNA depletion is applied selectively to matrices with high eukaryotic-to-pathogen ratios or where enrichment alone does not provide sufficient pathogen signal for downstream genomic analyses.

In contrast to the comparatively limited effects of the laboratory workflow, reference database composition exerted a substantial influence on classification accuracy and downstream subtyping performance. This was especially apparent for samples spiked with VTEC or *Shigella* spp. For these samples, species-specific databases (EC-specific/ SH-specific) consistently outperformed the broader FBP-specific reference database, yielding higher proportions of pathogen-specific reads, improved assembly completeness, and increased concordance with SI reference genomes. These findings are consistent with the close evolutionary relationship between *E. coli* and *Shigella*, which has been well documented. Genomic analyses demonstrate extensive sequence similarity between the two groups, with *Shigella* often considered a pathogenic lineage within the broader *E. coli* species complex, which complicates species-level discrimination in molecular and *k*-mer–based classification approaches (161,162). Overall, the improvements associated with curated, pathogen-focused databases observed in this study indicate that restricting the reference space can reduce classification noise and enhance pathogen-specific signal recovery. This observation aligns with prior work identifying reference database structure and curation as primary determinants of taxonomic classification performance, particularly when resolving closely related taxa (163–167). For *L. monocytogenes*, however, the broader FBP database was sufficient, reflecting the clearer phylogenetic separation of this species from closely related taxa.

Across all pathogens, assemblies generated from pathogen-specific reads supported reliable routine subtyping. This included accurate identification of VTEC serotypes (O157:H7 and non-O157), concordant *Shigella* species assignments, and consistent *L. monocytogenes* serotype predictions. These findings are consistent with previous studies showing that enriched metagenomic data can enable strain-level subtyping when pathogen coverage is sufficient (58–60,90).

The central aim of this study was to evaluate whether the proposed workflow could reproduce outbreak-associated genomic clustering with a level of resolution comparable to single-isolate based WGS analysis. Across all pathogens examined, wgMLST allele profiles derived from quasi-metagenomic datasets fell within established allele-difference thresholds used for outbreak detection within the PNC surveillance network and consistently grouped with their corresponding cultured isolate genomes in phylogenomic analyses. This concordance indicates that, when pathogen abundance is sufficient, quasi-metagenomic datasets capture enough of the core and accessory genome to enable high-resolution genomic characterization. These findings align with prior work demonstrating that enriched metagenomic sequencing can achieve SNV- and wgMLST-level discrimination when target pathogens dominate the microbial community (58,60).

The pathogen-presence likelihood metric developed in this study was informed by Lu et al. (2022), who demonstrated that *z*-score normalization can distinguish true signal from background variation in metagenomic data (85). In this study, however, robust *z*-scores were calculated from depth-normalized, transformed unique *k*-mer counts rather than raw read counts (85). Unlike read counts, which scale with per base sequencing depth, unique *k*-mer counts reflect genomic breadth and plateau once regions are sufficiently sampled, making them more stable across datasets and less sensitive to artifacts such as ambiguous multi-mapping reads or repetitive elements (94).

Unlike previous approaches, *z*-scores were used as a supporting metric in this study rather than the primary criterion for pathogen detection (85). Since *z*-scores reflect relative differences within a sample, they can be high even when absolute signal is low. This is especially true in low-complexity or low-count samples, where small differences can appear as strong outliers, increasing the risk of overinterpreting a weak signal. Accordingly, primary detection was based on unique KPMR thresholds, with *z*-scores used to confirm enrichment relative to the sample background. However, in practice, unique KPMR alone was typically sufficient to identify the true pathogen, as most samples exhibited strong, unambiguous signal. The inclusion of *z*-scores provided limited additional benefit under these conditions but may be more useful in scenarios involving low-abundance pathogens or complex backgrounds, where confirmation of relative enrichment within a sample can help resolve more ambiguous signals (168).

Application of the metric to *E. coli* and *Shigella* datasets revealed a common limitation when analyzing closely related organisms: elevated signals were often detected for both taxa within the same sample. This pattern did not reflect true co-occurrence, but rather shared genomic content leading to taxonomic ambiguity. As a result, sequence-based methods alone have limited ability to distinguish VTEC from *Shigella* spp., highlighting a well-recognized diagnostic challenge (169). To address this issue, the interpretation framework was modified to include biologically informed decision rules for resolving closely ranked signals. When *E. coli* appeared as the dominant classification, but *Shigella* signals fell within the defined ratio threshold, the highest-ranking *Shigella* taxon in the sample was reported as the most likely pathogen. This adjustment reflects both the known genomic relationship between these organisms and the greater public health consequences of misclassification, which can impact outbreak detection and response (170). The need for this rule highlights an important limitation of detection metrics when applied to pathogens with very high genomic similarity. In such cases, accurate interpretation requires consideration of classifier behavior, database composition, and organism biology (94,95,98,150). Despite these

challenges, the pathogen-presence likelihood metric provides a structured and reproducible approach for distinguishing true pathogen signal from background noise.

Overall, these results support the use of quasi-metagenomics as a practical complement to culture-based WGS in food safety surveillance. When combined with optimized enrichment, curated databases, and well-defined analytical thresholds, this approach can deliver near-real-time, strain-level genomic data. This advantage is particularly relevant for high-risk or perishable foods, where delays in culture-based workflows can limit timely response.

5.0 Chapter 5 - Implementation of long-read nanopore sequencing for rapid quasi-metagenomic pathogen surveillance and outbreak detection of *Listeria monocytogenes* in complex food matrices

5.1 Rationale

Building on the short-read quasi-metagenomic validation described in Chapter 4, this chapter evaluates the use of Oxford Nanopore Technologies MinION sequencing platform as a strategy to further reduce time to result in foodborne outbreak investigations. Illumina (short-read) platforms generate highly accurate sequence data, but operational factors such as sample batching and fixed run durations can limit responsiveness in time-sensitive public health settings. In practice, these logistical constraints often represent a greater barrier to rapid decision-making than analytical performance. The MinION platform offers a portable, real-time sequencing alternative that allows immediate data acquisition and flexible run durations. This capability is particularly relevant for foodborne investigations, where actionable genomic information may be required before conventional sequencing runs are completed.

A key objective of this chapter is to adapt the analytical framework developed in Chapter 4 to the characteristics of long-read sequencing data. This includes replacing short-read alignment tools such as Bowtie2 with long-read–optimized mappers, including minimap2, for removal of eukaryotic sequences, as well as incorporating nanopore-specific basecalling and read-filtering tools such as Guppy and filtlong. These modifications account for the distinct error profiles and read-length distributions associated with nanopore sequencing. The study therefore evaluates whether long-read quasi-metagenomic datasets can achieve the same strain-level discrimination demonstrated with short-read data, despite the higher per-read error rates historically associated with this platform.

This chapter also examines the phylogenomic utility of long-read data for outbreak investigation using MINTyper, a pipeline designed to accommodate long-read alignments and enable comparison across sequencing technologies. By applying this workflow to the same *L. monocytogenes* enrichment

preparations generated in Chapter 4, a direct comparison between short- and long-read quasi-metagenomic approaches is possible. This design allows evaluation of whether the phylogenetic resolution achievable with the MinION platform is sufficient for outbreak-related genomic analyses and helps determine its suitability for foodborne pathogen surveillance.

5.2 Results

5.2.1 General data quality

General data quality metrics for all *Listeria* long-read datasets are summarized in Figure 5.1. For *L. monocytogenes* NSD datasets, DNA concentrations (mean=16.9 ng/µL; min=0.1 ng/µL; max=68.9 ng/µL), total bases (mean=827,293,005; min=5,723,908; max=2,238,945,802), total reads (mean=106,071; min=6179; max=327,950), average read length (mean=7,825; min=926; max=17,318), and average base quality (mean=13.29; min=12; max=14) were generally lower than those observed for SD datasets. For SD, DNA concentration averaged 31.8 ng/µL (min=0 ng/µL; max=72.3 ng/µL), total bases 850,127,476 (min=23,952,652; max=1,910,649,825), total reads 110,858 (min=9893; max=249,756), average read length 8,181 (min=2052; max=17,318), and average base quality 13.54 (min=13; max=14). Except for DNA concentration (p=0.007), none of these differences were statistically significant. Two datasets (low spike-in SD LM-3 and low spike-in NSD LM-6) had extremely low DNA concentrations (<0.2 ng/µL). Although identical DNA preparations were used for both the Illumina NextSeq 550 and ONT MinION sequencing, the time interval between sequencing runs (>4 months) may have resulted in some degradation of the stored DNA. This, combined with the generally greater input DNA requirements and reliance on high-quality, intact template DNA for ONT library preparation compared with Illumina workflows, may have contributed to the lower sequencing depth observed in some long-read datasets. However, the potential impact of these factors was not directly assessed in the current study.

For non-monocytogenes *Listeria* NSD datasets, DNA concentration averaged 14.0 ng/µL (min=5.0 ng/µL; max=32.8 ng/µL), total bases 771,184,905 (min=394,539,961; max=1,502,132,082), total reads 104,195 (min=42,070; max=211,603), average read length 8,337 (min=4,273; max=11,365), and average base quality 13.56 (min=13.0; max=14.0) (Figure 5.1). For SD datasets, DNA concentration averaged 33.3 ng/µL (min=16.3 ng/µL; max=71.3 ng/µL), total bases 920,401,300 (min=391,944,888; max=2,261,066,695), total reads 110,980 (min=27,576; max=265,879), average read length 9,282 (min=6,089; max=14,213), and average base quality 13.4 (min=13.0; max=14.0). No significant differences were detected between protocols.

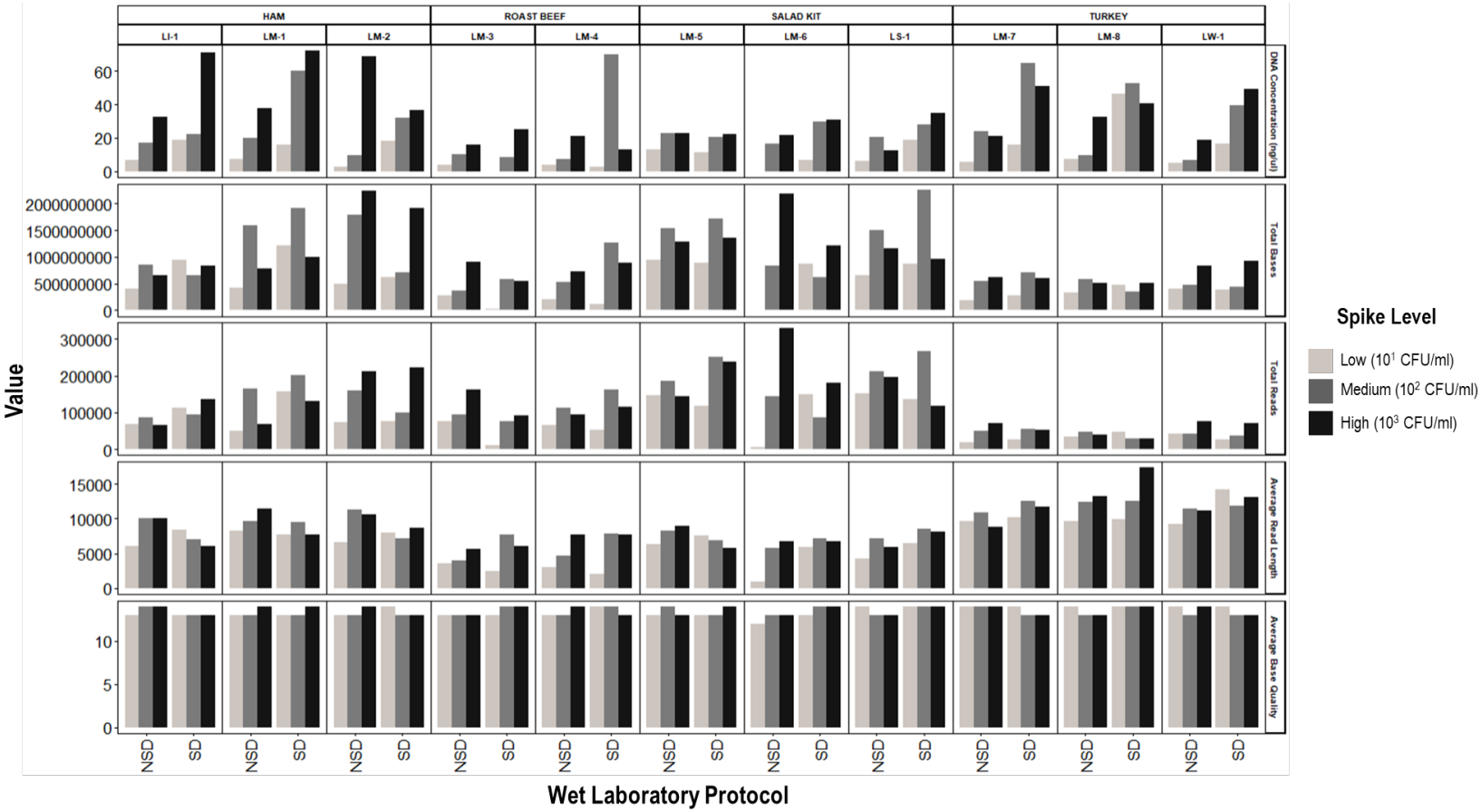


Figure 5.1. Comparison of long-read sequencing performance and DNA concentration across quasi-metagenomic Oxford Nanopore MinION datasets. Metrics shown include DNA concentration (ng/µL), total bases sequenced, read count, mean read length, and mean base quality. Samples were generated from multiple food matrices (ham, roast beef, salad kit, turkey) at three pathogen spike-in levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). The x-axis indicates wet-laboratory processing with no saponin-based eukaryotic DNA depletion (NSD) versus saponin-based depletion (SD). Datasets are labeled using *Listeria* species codes: LI = *Listeria innocua*; LM-1 to LM-8 = *Listeria monocytogenes*; LS = *Listeria seeligeri*; LW = *Listeria welshimeri*.

5.2.2 Data preprocessing

After filtlong filtering (retaining 90% of the highest-quality reads) and porechop ABI adapter trimming, an average of 49.0% of reads (min=35.6%; max=61.2%) were removed from SD datasets and 51.2% (min=38.3%; max=76.0%) from NSD datasets (Appendix 5.1).

5.2.3 Saponin-based eukaryotic DNA depletion

NSD datasets generally contained a higher proportion of eukaryotic reads than SD datasets (Figure 5.2). The highest eukaryotic content occurred in the NSD roast beef dataset LM-4 (7.6%), while all other datasets were below 2% aside from the salad kit datasets. In those NSD salad kit samples, eukaryotic reads ranged from 3.5–4.3%. When SD was applied, eukaryotic content fell below 1%, although salad kit datasets were similar between protocols (3–6%). Statistically significant differences between protocols were observed only for ham (p=0.0143) and turkey (p=0.00391). Fewer eukaryotic reads were observed in quasi-metagenomic long-read datasets compared to short-read data. This likely reflects the preference of long-read sequencing for high–molecular weight DNA, resulting in reduced retention of fragmented eukaryotic DNA compared to short-read workflows (Section 4.2.2).

5.2.4 RefSeq Masher provides a rapid preliminary screen of pre-processed datasets and can inform the selection of appropriate reference databases for subsequent taxonomic classification

Across all *L. monocytogenes* spike-in datasets, RefSeq Masher (contains) detected the target pathogen as the top hit with ≥384/400 shared hashes and ≥99% identity. One exception was the low spike-in NSD dataset LM-6, which yielded 154/400 shared hashes and 94% identity, likely due to its extremely low DNA concentration (Figure 5.1). For samples spiked with other *Listeria* species, the pathogen of interest had fewer shared hashes but similar identity values (≥99%). Maximum values were 369/400 for *L. welshimeri* (minimum 361/400), 360/400 for *L. innocua* (minimum 354/400), and 337/400 for *L. seeligeri* (minimum 295/400). The correct species was the top hit for all *L. seeligeri* and *L. innocua* datasets, but not for *L. welshimeri*, for which *L. monocytogenes* consistently ranked first.

No target pathogen was detected in the negative controls, except for the SD ham and turkey samples, which exhibited 223/400 (96.4% identity) and 211/400 (96.1% identity) shared hashes for *L. monocytogenes*, respectively. These values were lower than those observed in the true-positive spike samples, and all NC samples were culture-negative for the pathogen of interest (Section 2.2). Although these findings suggest that the observed signal may not represent viable *L. monocytogenes*, its origin was

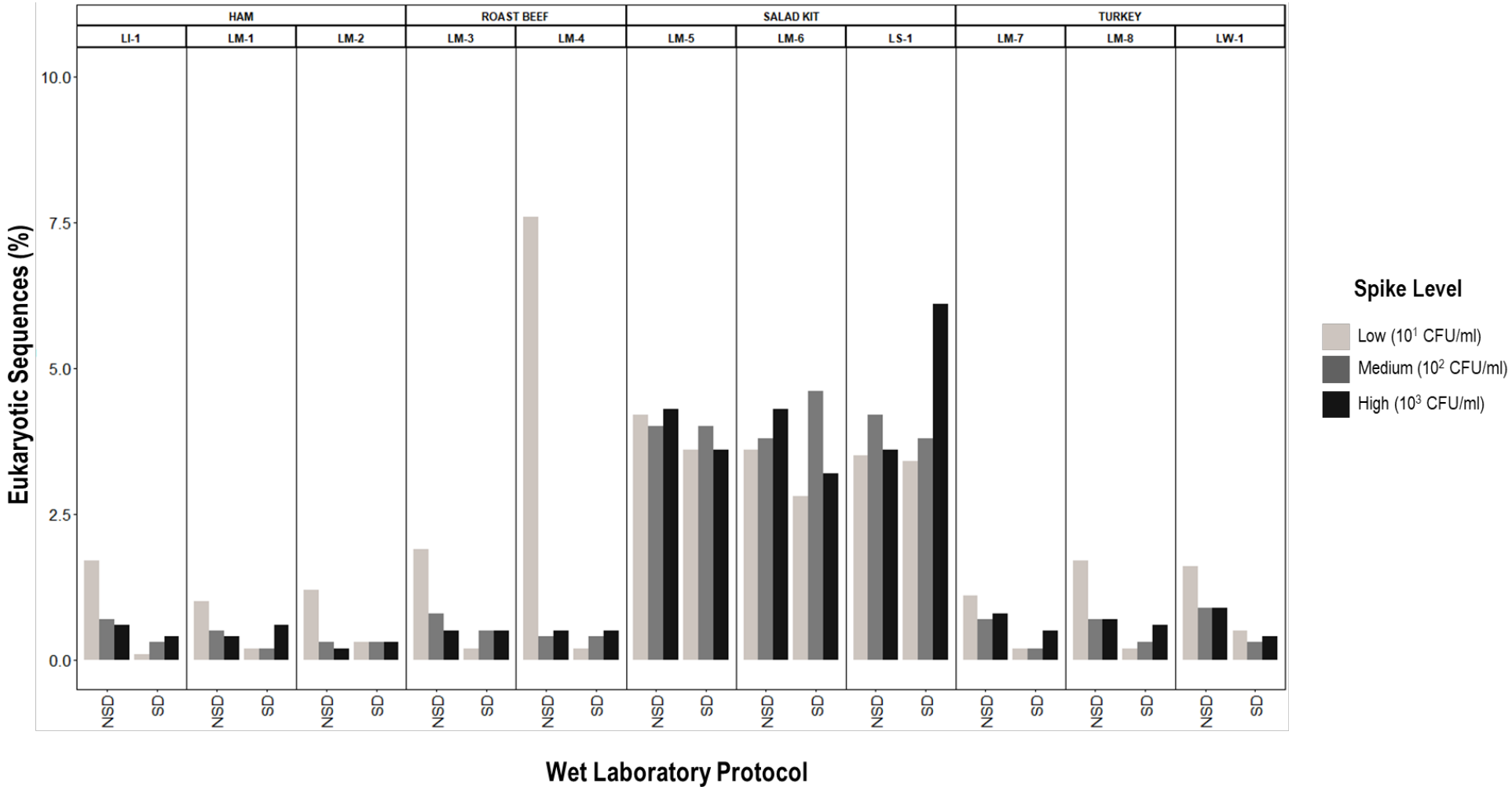


Figure 5.2. Effect of saponin-based depletion on eukaryotic DNA background in quasi-metagenomic Oxford Nanopore MinION datasets. Shown is the percentage of total sequencing reads classified as eukaryotic DNA for artificially contaminated food samples processed with no saponin-based depletion (NSD) or saponin-based eukaryotic DNA depletion (SD). Results are presented across multiple food matrices (ham, roast beef, salad kit, turkey) and three pathogen spike-in levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). Datasets are labeled using *Listeria* species codes: LI = *Listeria innocua*; LM-1 to LM-8 = *Listeria monocytogenes*; LS = *Listeria seeligeri*; LW = *Listeria welshimeri*.

not formally investigated in this study. Potential explanations include trace environmental or reagent-associated DNA, low-level cross-sample carryover, or other sources of background sequencing signal.

5.2.5 Saponin-based depletion does not substantially improve pathogen signal under optimized enrichment conditions

KrakenUniq classification using the FBP-specific reference database showed that 97.47% of reads on average (min=65.8%; max=99.9%) were assigned to *Listeria* in NSD *L. monocytogenes* datasets and 98.0% (min=79.2%; max=99.9%) in SD datasets (p=0.553) (Figure 5.3). For other *Listeria* species, 98.6% (min=92.7%; max=99.9%) and 98.8% (min=93.9%; max=99.9%) of reads were assigned to *Listeria* in NSD and SD datasets, respectively (p=0.294).

Pathogen-specific read counts averaged 47,973 for NSD (min=1,387; max=154,478) and 54,137 for SD (min=3,995; max=120,035). Unique *k*-mers averaged 4,030,066 in NSD (min=1,313,638; max=5,235,514) and 3,976,501 in SD datasets (min=2,851,915; max=4,741,801). None of these differences were significant (p=0.345 and p=0.317). For other *Listeria* species, pathogen-specific reads averaged 51,677 (min=23,863; max=105,505) for NSD and 52,220 (min=15,484; max=127,261) for SD datasets, again with no significant difference (p=0.91). Unique *k*-mer counts were likewise comparable (p=0.496).

The metrics for the single-isolate (SI) reference strains were largely comparable to those of the corresponding quasi-metagenomic datasets across all food matrices (Figure 5.3). For both SI and quasi-metagenomic workflows (NSD and SD), the proportion of reads classified as Listeria remained consistently near 100%. In addition, unique *k*-mer recovery in the quasi-metagenomic datasets closely matched that observed in the SI references, indicating that both NSD and SD protocols effectively preserved the genomic signal of the target pathogen despite the complexity of the food matrices.

In contrast, negative controls (NC) exhibited substantially lower proportions of reads classified as *Listeria*, with minimal pathogen-specific read counts and markedly reduced unique *k*-mer recovery across all matrices. Since NC samples were confirmed to be culture negative for the pathogens of interest (Section 2.2), the low-level signals detected are likely attributed to trace DNA from environmental or reagent contamination, or low-level cross-sample carryover rather than true pathogen-presence.

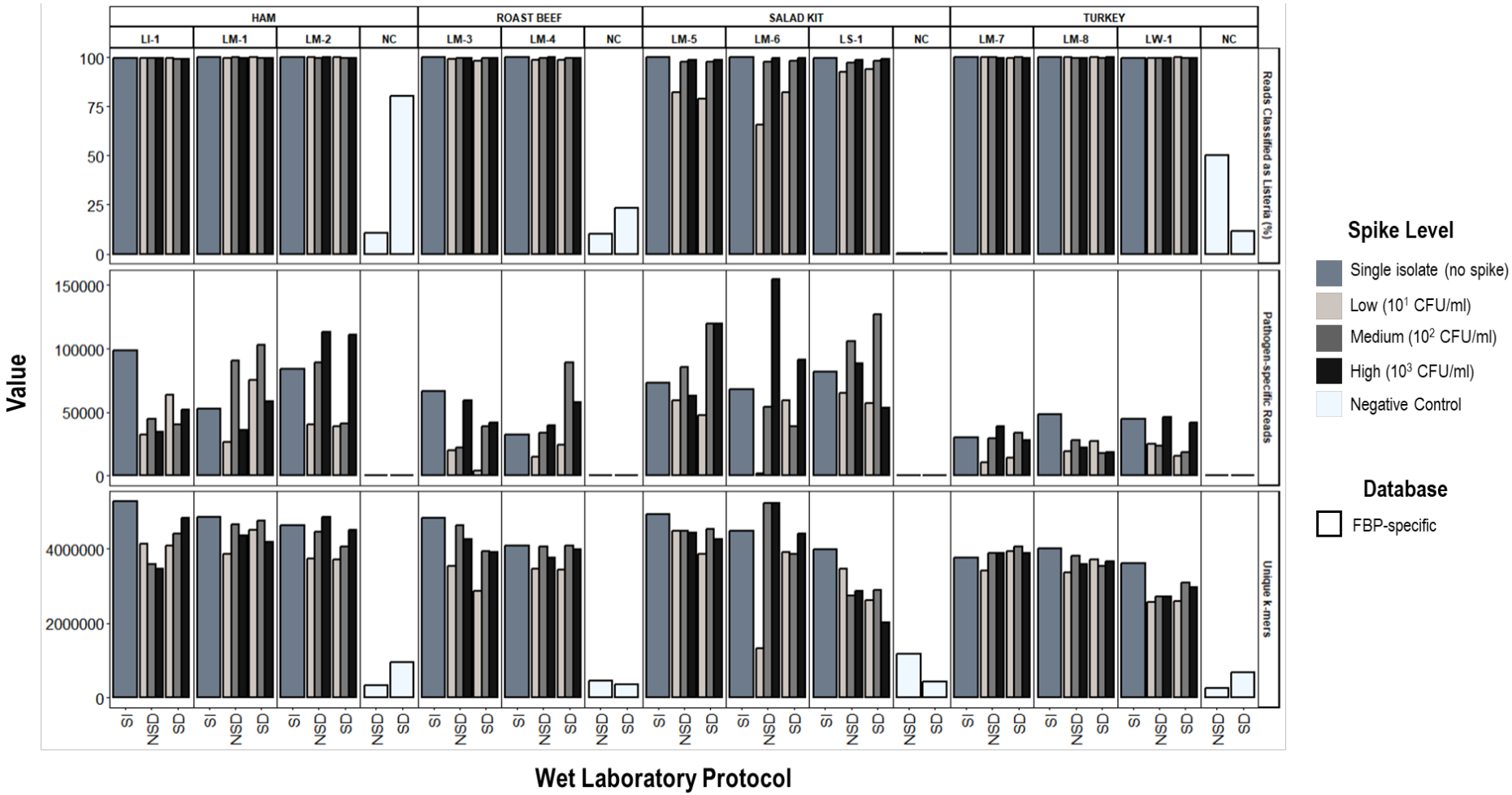


Figure 5.3. KrakenUniq classification metrics for *Listeria* strains in long-read quasi-metagenomic Oxford Nanopore MinION datasets. The figure summarizes the effects of wet-laboratory workflows on taxonomic classification by comparing the percentage of reads classified (top panel), pathogen-specific read counts (middle panel), and unique *k*-mer counts (bottom panel). Quasi-metagenomic datasets generated using non–saponin-depleted (NSD) and saponin-depleted (SD) protocols are shown alongside single-isolate (SI) reference strains. Results are presented across four food matrices (ham, roast beef, salad kit, turkey) and three contamination levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). Negative controls (NC) are included to assess baseline taxonomic noise. Datasets are labeled using *Listeria* species codes: LI = *Listeria innocua*; LM-1 to LM-8 = *Listeria monocytogenes*; LS = *Listeria seeligeri*; LW = *Listeria welshimeri*. All classifications were performed using the optimized foodborne pathogen–specific reference database.

5.2.6 Pathogen-presence likelihood metric

The long-read quasi-metagenomic datasets showed consistent detection of the spiked *Listeria* strains across food matrices and contamination levels (Figure 5.4). In the unique KPMR heatmap, *L. monocytogenes* and related *Listeria* taxa produced the strongest signal (Figure 5.4A). The corresponding *z*-score heatmap (Figure 5.4B) also demonstrated within-sample enrichment of the spiked *Listeria* taxa relative to background organisms. Compared with deli meat matrices (ham, roast beef, and turkey), salad kit samples exhibited greater taxonomic diversity and broader background signal across multiple non-target taxa, reflecting the more complex microbial communities associated with plant-derived foods.

Despite this increased background diversity, the spiked *Listeria* strains remained clearly distinguishable from non-target organisms across all matrices. Two samples with previously noted low DNA concentrations (low-spike LM-3-SD and low-spike LM-6-NSD) did not meet the minimum sequencing depth threshold and were excluded from downstream interpretation. Except for the NSD NC salad kit sample, none of the NC samples across the evaluated food matrices met the minimum coverage required for metric generation (Figures 4.12–4.14).

Strong agreement between the pathogen assigned by the likelihood framework and the true spike organism was observed across all quasi-metagenomic long-read samples (Figure 5.5). For the deli meat matrices (ham, roast beef, and turkey), the predicted pathogen consistently matched the spiked *Listeria* species across spike levels and wet-laboratory conditions, indicating that the decision framework correctly prioritized the dominant taxon when pathogen signal was present. Similar agreement was observed for the salad kit samples, although the higher microbial background associated with the plant matrix resulted in greater taxonomic diversity and more dispersed signal across non-target taxa. Despite this increased complexity, the metric correctly identified the spiked *Listeria* taxa in nearly all samples.

One exception occurred in the NC NSD salad kit sample, where Salmonella enterica was identified as the most likely pathogen (Figure 5.5), with 1,583 reads and approximately 3,057 unique *k*-mers. S. enterica was also detected in the NC SD salad kit sample (1,151 reads; 2,636 unique *k*-mers), although this dataset did not meet the minimum sequencing depth threshold (<10,000 root reads) required for evaluation by the pathogen-presence likelihood metric (Section 2.9.3). These findings highlight the potential value of further threshold refinement to improve discrimination of low-level signals in complex plant-derived matrices.

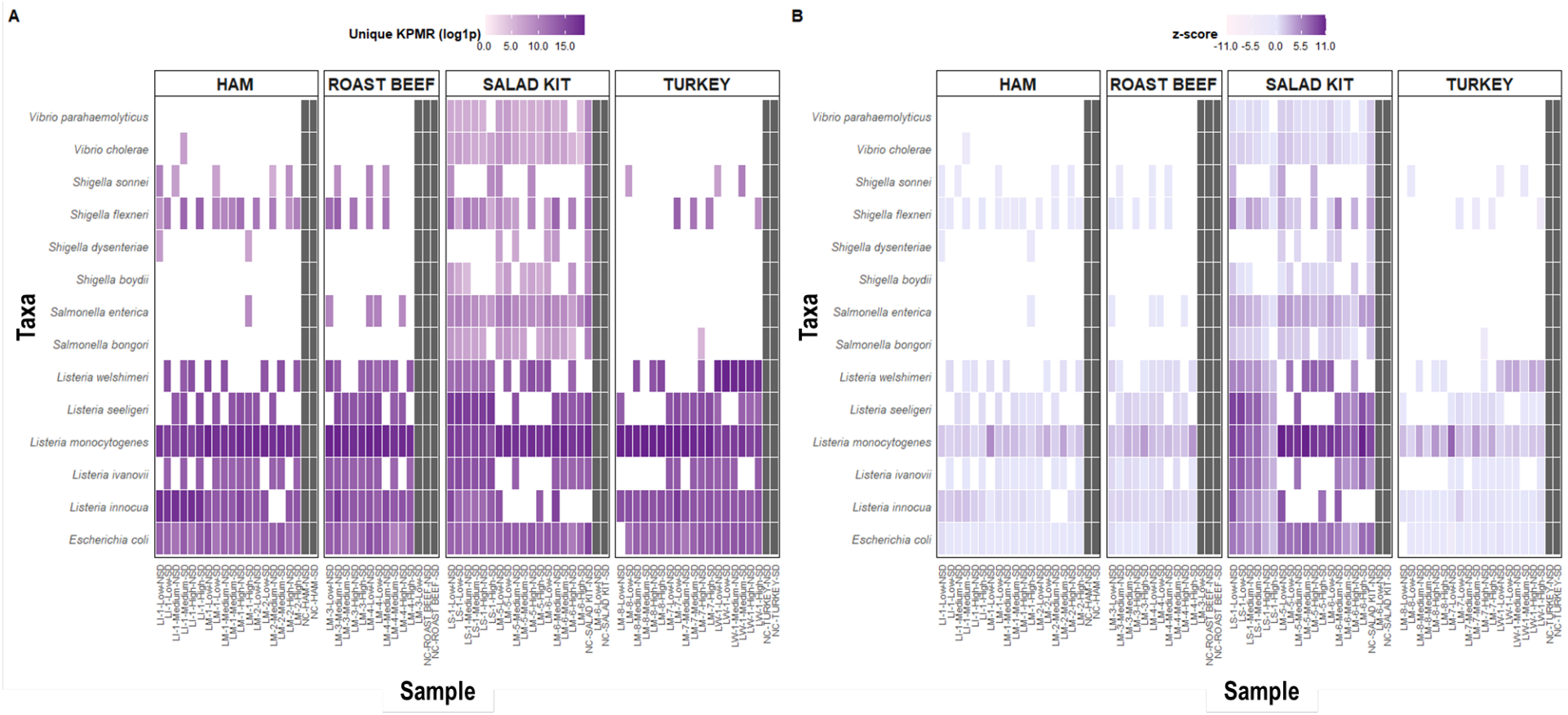


Figure 5.4. Absolute detection signal and within-sample enrichment in long-read quasi-metagenomic datasets derived from food matrices spiked with *Listeria* strains. Heatmaps display depth-normalized unique *k*-mer counts (KPMR; log1p-transformed; panel A) and corresponding within-sample robust *z*-scores (panel B). In panel A, increasing purple intensity represents higher absolute abundance of pathogen-specific unique *k*-mers (total detection signal). In panel B, a diverging color scale centered at zero illustrates relative enrichment within each sample, with darker purple indicating taxa enriched above the sample-specific background. Datasets were generated from artificially contaminated ham, roast beef, salad kit, and turkey matrices at low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL) concentrations under two laboratory workflows: no saponin depletion (NSD) and saponin depletion (SD). Negative controls (NC) were included to assess baseline signal. Grey tiles denote samples that failed minimum sequencing depth requirements, while white tiles indicate absence of detected signal. Strain abbreviations are as follows: LI = *L. innocua*; LM-1 through LM-8 = *L. monocytogenes*; LS = *L. seeligeri*; and LW = *L. welshimeri*. Metrics were derived from KrakenUniq outputs using the optimized Foodborne Pathogen–specific reference database.

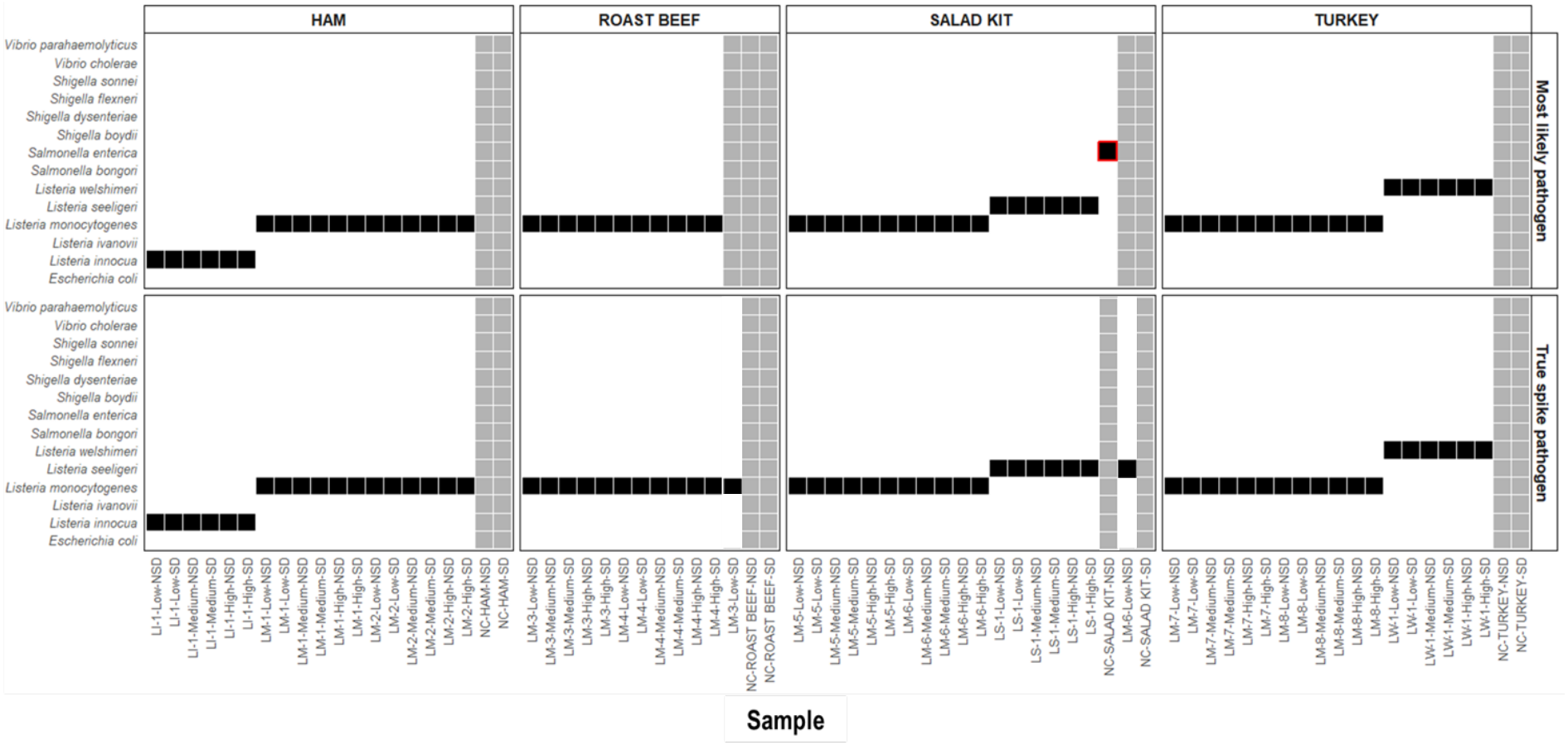


Figure 5.5. Most likely pathogen assignments across long-read quasi-metagenomic datasets as determined by the pathogen-presence likelihood metric. Black tiles indicate the taxon designated as the most likely pathogen for each sample following application of the integrated framework, which incorporates unique KPMR magnitude thresholds, robust within-sample *z*-score scaling, and predefined tie-resolution rules. The true spike pathogen is shown for comparison in the lower panel. A red outline indicates a mismatch between most likely pathogen and the true spike pathogen. Samples are stratified by food matrix (ham, roast beef, salad kit, turkey, ground beef, lettuce, and carrot) and include low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL) contamination levels processed under both no saponin depletion (NSD) and saponin depletion (SD) protocols. Samples failing minimum sequencing depth requirements are indicated in grey. Strain abbreviations are as follows: LI = *L. innocua*; LM-1 through LM-8 = *L. monocytogenes*; LS = *L. seeligeri*; and LW = *L. welshimeri*.

5.2.7 RefSeq Masher Matches provides dependable species-level verification for *Listeria* species.

For *L. monocytogenes*, the target species was consistently identified by RefSeq Masher (matches) as the top hit, with ≥350/400 matching hashes in both NSD and SD datasets. Two exceptions were observed: low spike datasets SD LM-3 (328/400) and NSD LM-6 (4/400), both of which were associated with extremely low DNA concentrations (Figure 5.1). In samples spiked with other *Listeria* species, ≥174/400 matching hashes were recovered across all protocols.

5.2.8 Assembly quality is consistent between wet-laboratory protocols

Assembly quality for *L. monocytogenes* pathogen-specific reads was consistent between protocols, except for SD LM-3 and NSD LM-6 (Figure 5.6). NSD datasets averaged three contigs (min=1; max=6) and SD datasets averaged five (min=1; max=44) (p=0.590). Genome lengths did not differ significantly (NSD mean=2,900,822 bp; SD mean=3,009,148 bp; p=0.884). GC content was similar across protocols (NSD mean=38.0%; SD mean=38.0%; p=0.347). Completeness was slightly higher for SD (mean=98.6%) than NSD (mean=95.1%), but not significantly (p=0.191). Contamination also did not differ significantly (p=0.041). For other *Listeria* species, no significant differences were found in contigs (p=1), genome length (p=0.343), GC content (p=1), completeness (p=0.075), or contamination (p=0.789) (Figure 5.6).

Across most strains and food matrices, quasi-metagenomic assemblies closely matched the SI references in genome length, GC content, and completeness, indicating that strain-level genomic representation was preserved despite the added complexity of the food matrix. Although quasi-metagenomic datasets occasionally exhibited modest increases in contig number or low-level contamination relative to SI assemblies, these differences were minor and did not substantially impact overall assembly concordance (Figure 5.6).

5.2.9 Genome coverage

In NSD *L. monocytogenes* datasets, an average of 96.4% of pathogen-specific reads mapped to the reference (min=92.5%; max=100%) (Figure 5.7). SD datasets showed similar results (mean=96.7%; min=93.7%; max=100%). Total reads and percent mapped reads did not differ significantly (p=0.345 and p=0.103). Mean coverage depth was also comparable (NSD mean=227.9X; SD mean=231.5X; p=0.747). Low spike datasets SD LM-3 and NSD LM-6 produced <4,000 reads and <10X coverage, as expected based on DNA concentration.

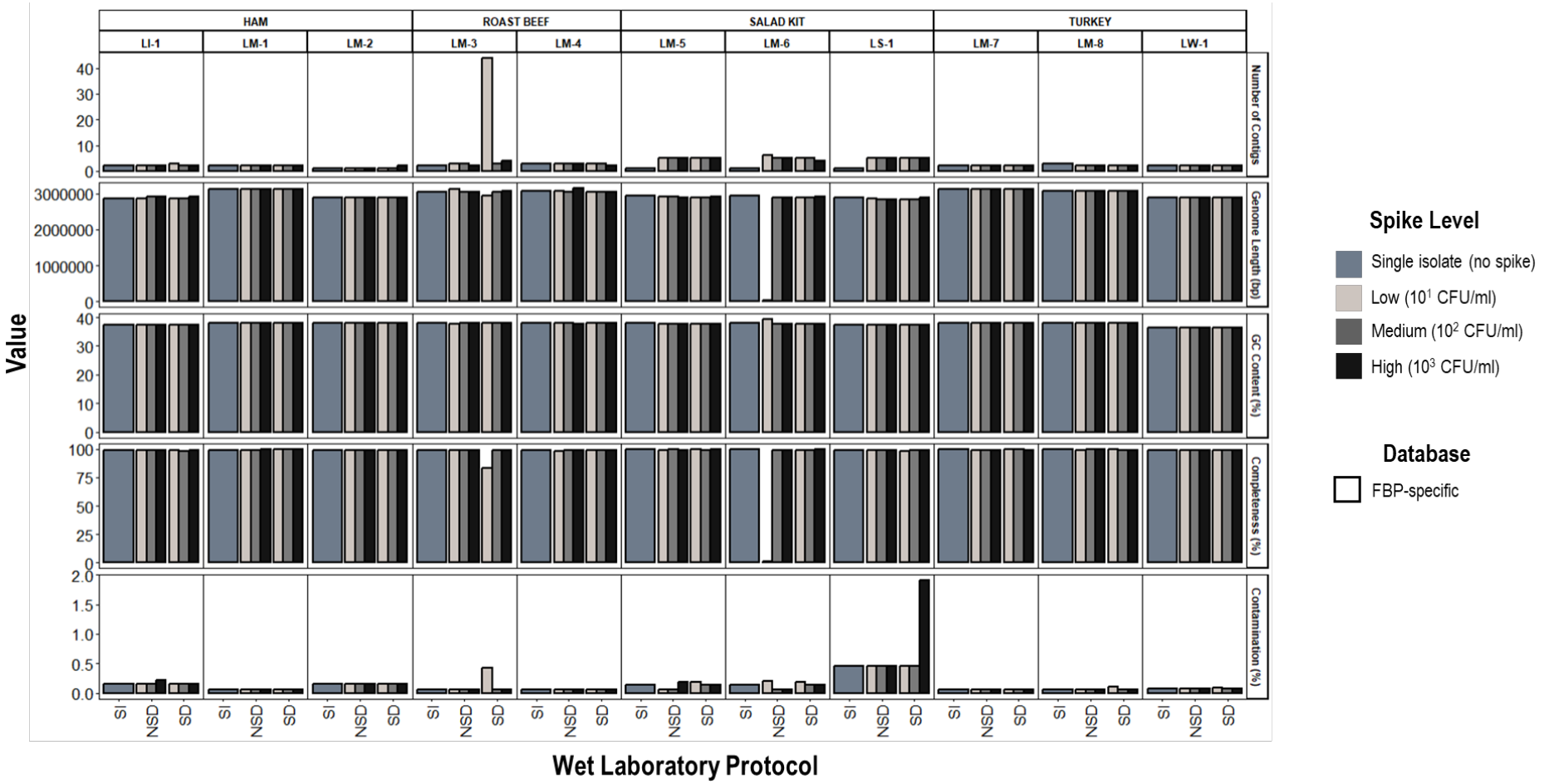


Figure 5.6. Comparative assembly quality metrics for *Listeria* strains in long-read quasi-metagenomic Oxford Nanopore MinION datasets. This figure summarizes de novo assembly performance by comparing quasi-metagenomic assemblies generated from pathogen-specific reads (extracted using the FBP-specific reference database) with corresponding single-isolate (SI) reference assemblies. Assembly quality was evaluated using QUAST-derived metrics, including number of contigs, total genome length (bp), and GC content (%), together with CheckM-derived completeness (%) and contamination (%). Results are stratified by wet-laboratory workflow, non–saponin-depleted (NSD) and saponin-depleted (SD), and shown across four food matrices (Ham, Roast Beef, Salad Kit, and Turkey) at three contamination levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). Datasets are labeled by *Listeria* species code (LI = *L. innocua*; LM-1–8 = *L. monocytogenes*; LS = *L. seeligeri*; LW = *L. welshimeri*).

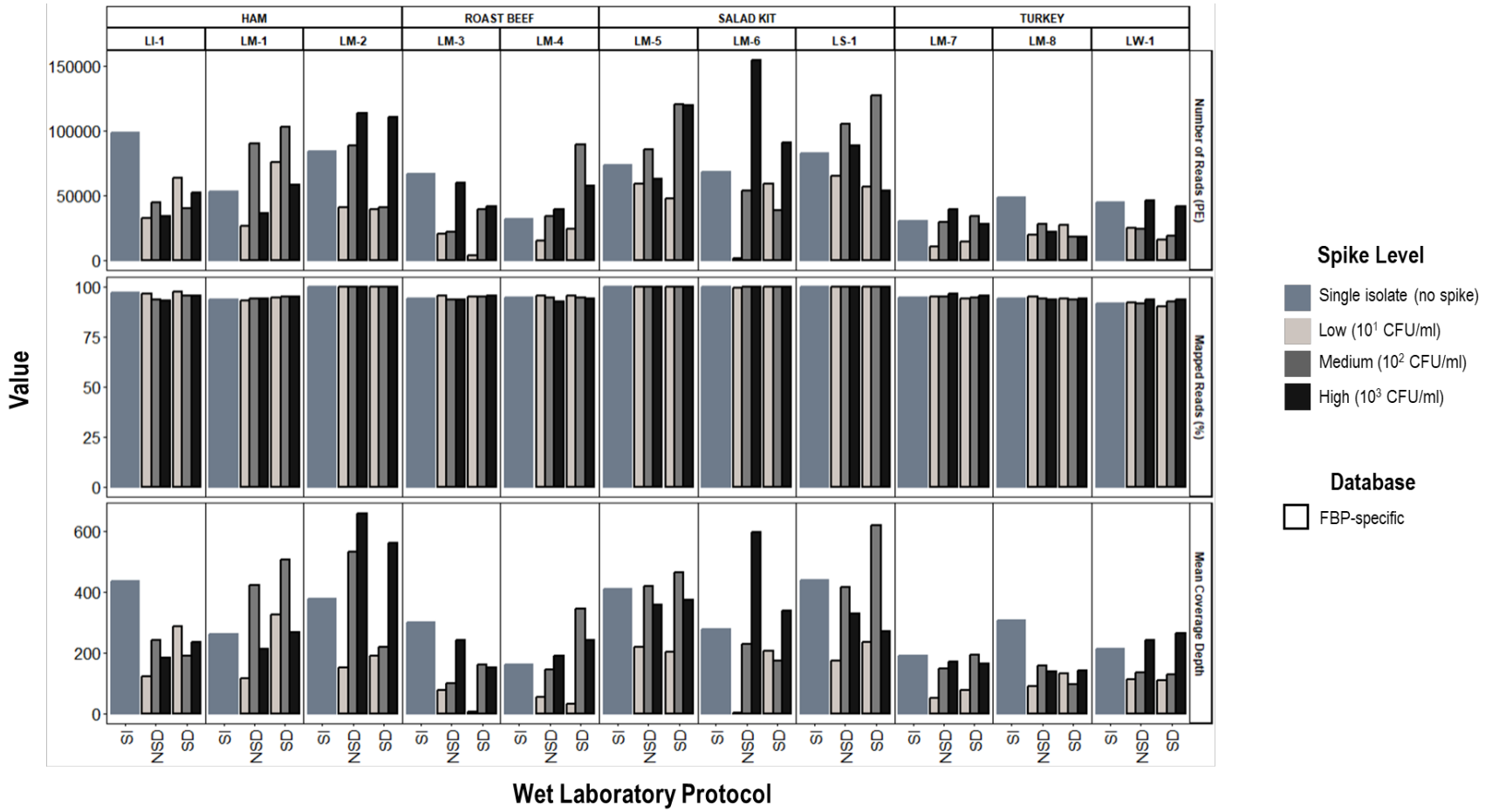


Figure 5.7. Comparative assembly coverage metrics for *Listeria* strains in long-read quasi-metagenomic Oxford Nanopore MinION datasets. This figure summarizes genome coverage characteristics by comparing quasi-metagenomic assemblies generated from pathogen-specific reads (extracted using the FBP-specific reference database) with corresponding single-isolate (SI) reference assemblies. Coverage performance is evaluated across two wet-laboratory workflows, non–saponin-depleted (NSD) and saponin-depleted (SD), at three contamination levels—low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL)—within four food matrices (Ham, Roast Beef, Salad Kit, and Turkey). Panels display Qualimap-derived metrics, including the number of reads assigned to the target pathogen, the percentage of mapped reads (%) aligning to the reference genome, and mean coverage depth. Datasets are labeled by Listeria species code (LI = L. innocua; LM-1–8 = *L. monocytogenes*; LS = *L. seeligeri*; LW = *L. welshimeri*).

For other *Listeria* species, no significant differences were observed across any coverage measure. The SI reference strains showed coverage profiles that were largely comparable to those of the corresponding quasi-metagenomic datasets generated under both NSD and SD workflows for each food matrix. Although SI references typically demonstrated higher mean coverage depth, likely owing to the absence of matrix-derived background, the quasi-metagenomic datasets maintained sufficient genome coverage to preserve the pathogen signal.

Across nearly all samples, detection of *L. monocytogenes* and other *Listeria* species was achievable under both wet-laboratory protocols. Due to insufficient long-read data, SD LM-3 (low-spike) and NSD LM-6 (low-spike) were excluded from further analysis.

5.2.10 Quality assessment of long-read data in BioNumerics

Since BioNumerics v7.6.3 is not designed for long-read data, quality metrics were derived from imported assemblies described in Section 5.2.8. Only *L. monocytogenes* assemblies were assessed, as the schema in BioNumerics is species-specific. Low-spike datasets SD LM-3 and NSD LM-6 were excluded due to low data quality. In the absence of PNC guidelines for long-read quality assessment, the metrics currently implemented for short-read quasi-metagenomic data were applied as a proxy (Appendix 4.4). However, average base quality (AvgQuality) could not be calculated for the long-read datasets because this metric is derived from raw sequencing reads, whereas the long-read analyses in BioNumerics were performed on assembled genomes and not on raw read data.

Overall, the long-read assemblies resembled the corresponding single-isolate assemblies, except for LM-5 and LM-6, which had much lower N50 values (NSD mean=1,490,541 bp; SD mean=1,492,723 bp) than the single-isolate reference assemblies (LM-5 mean=2,939,906 bp; LM-6 mean=2,939,886 bp). This likely reflects the greater number of contigs in the quasi-metagenomic datasets (NSD mean=5; SD mean=5) compared to single-isolate assemblies (one contig) (Figure 5.8). Given the substantial differences in read length between short- and long-read data, N50 and contig counts (NrContigs) are not directly comparable to current PNC criteria (Appendix 4.4). Other metrics, including genome length, consensus allele calls, and core percent, were generally consistent with accepted thresholds. One dataset (medium-spike NSD LM-2) did not meet the 95% core threshold (core=91%) but was retained for outbreak detection due to the exploratory nature of this work.

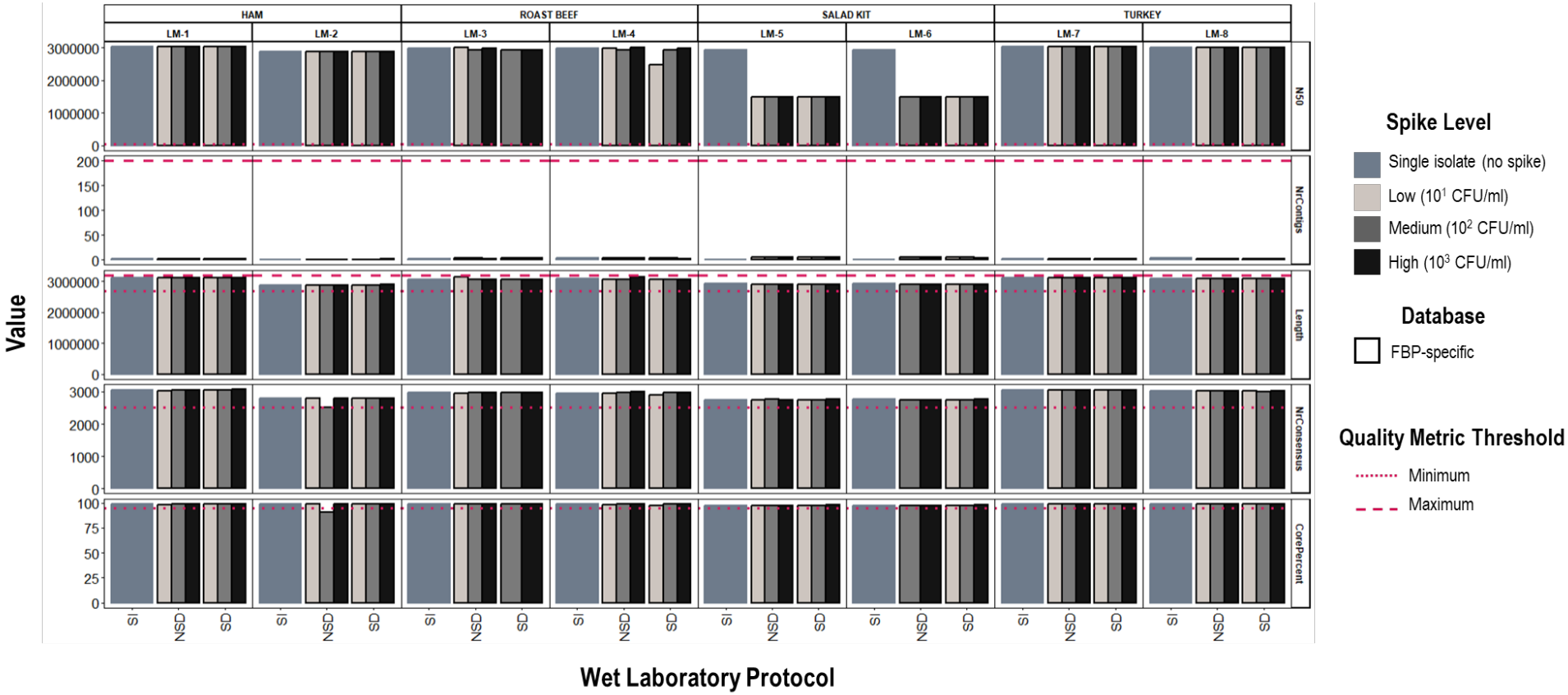


Figure 5.8. Comparative assembly and subtyping performance for *Listeria* long-read quasi-metagenomic Oxford Nanopore MinION datasets. Assembly and subtyping metrics were evaluated for quasi-metagenomic datasets generated using non–saponin-depleted (NSD) and saponin-depleted (SD) workflows at three contamination levels—low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL)—and compared with corresponding single-isolate (SI) reference strains. Metrics were generated in BioNumerics (v7.6) using *Listeria*-specific reads extracted from Ham, Roast Beef, Salad Kit, and Turkey matrices with the foodborne pathogen–specific (FBP-specific) reference database. Data quality indicators include AvgQuality (mean Phred quality score of raw reads), NrContigs (number of contigs in the de novo assembly), Length (total assembly size in base pairs), NrConsensus (number of concordant allele calls across identification algorithms), and CorePercent (percentage of core wgMLST loci identified). Datasets are labeled by *Listeria* strain code (LM-1–8 = *L. monocytogenes*). PulseNet Canada organism-specific quality thresholds are indicated by dotted (minimum) and dashed (maximum) reference lines.

5.2.11 Outbreak detection using long-read quasi-metagenomic datasets

Table 5.1 summarizes the reported genomic distance metrics for outbreak detection of *L. monocytogenes* based on long-read quasi-metagenomic datasets generated with the Oxford Nanopore (MinION) platform.

5.2.11.1 Whole-genome multilocus sequence typing (wgMLST)

For most outbreaks, fewer than ten wgMLST allele differences were observed between outbreak isolates and the corresponding quasi-metagenomic long-read datasets, consistent with results obtained from short-read quasi-metagenomic analyses in Chapter 4 as well as from single-isolate sequencing associated with the original outbreaks.

Although wgMLST allele differences remained within thresholds considered related by PNC (152), the salad kit and turkey outbreaks exhibited slightly broader allele ranges compared with both the short-read quasi-metagenomic datasets and the original outbreak isolates. For the salad kit outbreak, a wgMLST range of 0–8 alleles was observed, driven primarily by three datasets, NSD low-spike LM-5, NSD high-spike LM-5, and NSD medium-spike LM-6, which differed by six to seven alleles from the SI reference strain. Similarly, the turkey outbreak displayed a wgMLST range of 0–9 alleles, largely attributable to the NSD low-spike LM-7 dataset, which differed by seven alleles from the SI reference strain and accounted for much of the observed increase in diversity (Figure 5.9).

A more pronounced deviation was observed for the roast beef outbreak, where inclusion of long-read datasets resulted in wgMLST allele differences ranging from 0 to 35. This expanded range was driven primarily by LM-4, with the low-spike NSD LM-4 differing by 13 alleles and the low-spike SD LM-4 differing by 29 alleles from the reference strain (Figure 5.9). When these two datasets were excluded, the wgMLST range narrowed to 0–12 alleles. As both datasets met PNC sequence quality assessment guidelines (Section 5.2.10; Appendix 4.4), the observed discrepancies may reflect basecalling or sequencing error associated with MinION long-read data, rather than true genomic differences. For all remaining datasets, wgMLST allele differences between long-read quasi-metagenomic and single-isolate assemblies did not exceed seven alleles (Figure 5.9).

Table 5.1. Comparison of genomic distance metrics for outbreak detection of *Listeria monocytogenes* using long-read quasi-metagenomic Oxford Nanopore MinION datasets.

Pathogen	Study Strains	Implicated Food Product	Total Number of isolates (Clinical/ Food/ Environmental)	Original Outbreak			Quasi-metagenomics				
							ILLUMINA Short Reads (Chapter 4)			MinION Long Reads (Chapter 5)	
				SNVs (SNVPhyl)	**SNPs** (MINTyper)	**wgMLST alleles** (BioNumerics)	**SNVs** (SNVPhyl)	**SNPs** (MINTyper)	**wgMLST alleles** (BioNumerics)	**SNPs** (MINTyper)	**wgMLST alleles** (BioNumerics)
L. monocytogenes	LM-2	Ham*	3 (1/1/0)	0	0	0-1	0	0	0-1	0	0-5
	LM-3, LM-4	Roast Beef	28 (4/24/0)	0-11	0-11	0-10	0-11	0-11	0-10	0-11	0-35
	LM-5, LM-6	Salad Kit	19 (11/8/0)	0-4	0-4	0-2	0-4	0-4	0-2	0-4	0-8
	LM-7, LM-8	Turkey	65 (54/9/2)	0-6	0-5	0-5	0-6	0-5	0-5	0-5	0-9

*Genomic distances between outbreak reference isolates and their corresponding long-read quasi-metagenomic assemblies were evaluated using whole-genome multilocus sequence typing (wgMLST) in BioNumerics (v7.6) and the single-nucleotide variant phylogenomics pipeline (SNVPhyl) to assess outbreak-level resolution for *Listeria monocytogenes*. Inclusion of the non-outbreak isolate LM-1 resulted in >200 wgMLST allele differences and >5,000 SNVs/SNPs relative to outbreak-associated isolates.

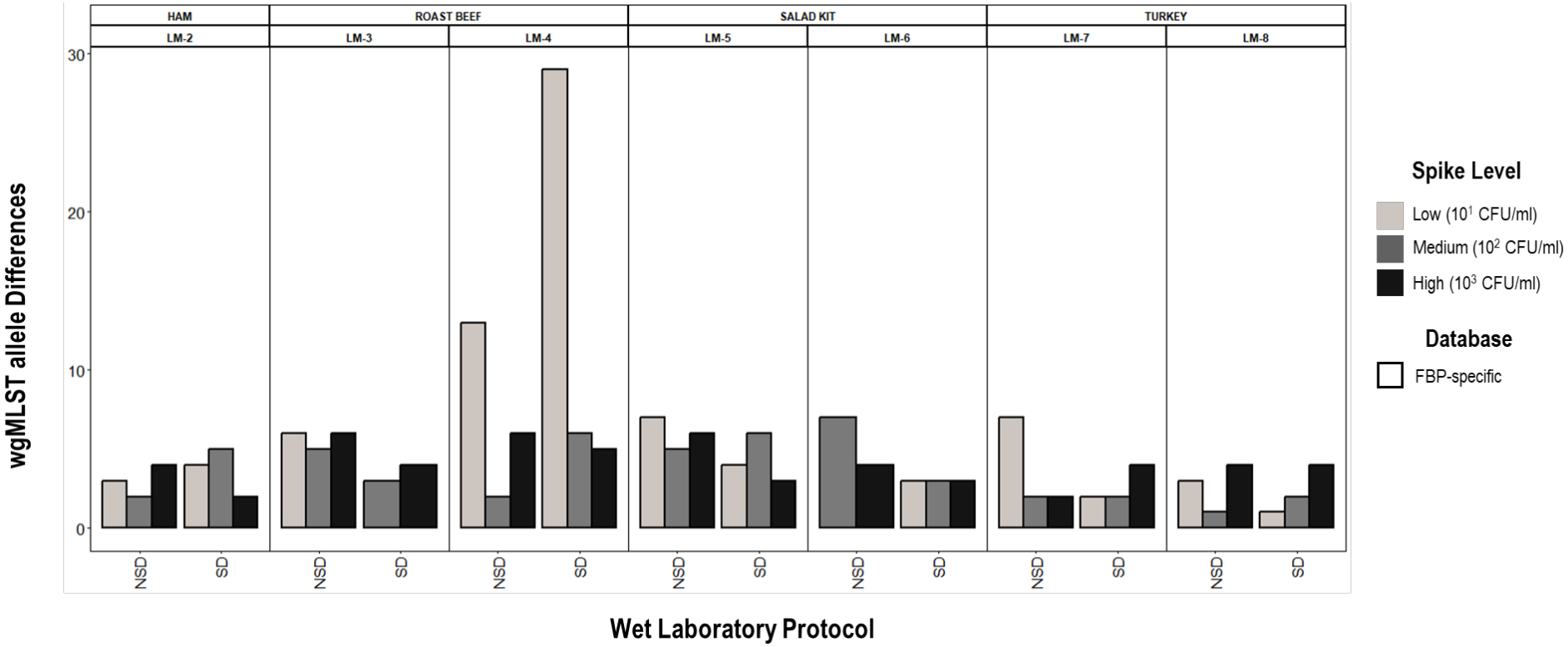


Figure 5.9. Whole genome multilocus sequence typing (wgMLST) allele differences between single-isolate references and long-read quasi-metagenomic Oxford Nanopore MinION *L. monocytogenes* datasets across food matrices. Genomic differences were evaluated using wgMLST in BioNumerics (v7.6) for non–saponin-depleted (NSD) and saponin-depleted (SD) long-read quasi-metagenomic datasets across four food commodities (Ham, Roast Beef, Salad Kit and Turkey). Results are shown for three contamination levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL), and are compared with the corresponding single-isolate (SI) reference genome. Datasets are labeled by *Listeria monocytogenes* strain code (LM-1–8).

5.2.11.2 SNP analysis

SNP-based results for the long-read datasets were concordant with those previously obtained for the original outbreak isolates and the short-read quasi-metagenomic datasets (Table 5.1). Because SNVPhyl does not support long-read data, MINTyper was used for SNP analysis, and both the outbreak isolates and quasi-metagenomic datasets were reanalyzed to ensure methodological consistency. Across all outbreaks, SNP distances were nearly identical between datasets, with one exception. For the outbreak linked to turkey deli meat, MINTyper reported a SNP range of 0–5 compared with a range of 0–6 obtained using SNVPhyl (Table 5.1). This minor difference likely reflects the more stringent base-filtering criteria applied by MINTyper and further supports the interpretation that the broader wgMLST allele range observed for pathogen-specific reads extracted from roast beef matrices is attributable to base-calling errors within the allele sequences rather than true genomic differences.

5.3 Discussion

This study evaluated whether quasi-metagenomics, when paired with long-read sequencing using the MinION platform, can support rapid detection and strain-level characterization of *L. monocytogenes* and other *Listeria* spp. in enriched food matrices. The results demonstrate that, under optimized conditions, long-read quasi-metagenomics can reproduce outbreak-level resolution comparable to WGS from isolates, thereby supporting its potential role in foodborne pathogen surveillance.

Overall data quality from MinION sequencing was adequate for downstream analysis, despite notable variability in DNA concentration across samples. As expected, the SD protocol generally yielded higher DNA concentrations than the NSD workflow, but most other quality metrics, including total bases, read length, and read accuracy, were similar between protocols. Two samples with extremely low DNA concentrations performed poorly across all analyses, illustrating the well-recognized sensitivity of long-read sequencing to DNA degradation and input quality (171,172). Apart from these outliers, both protocols produced datasets suitable for taxonomic and genomic characterization.

The SD protocol reduced the proportion of eukaryotic sequences in several matrices, although the effect was matrix-dependent and less pronounced in fresh produce samples, consistent with observations in Chapter 4. Despite these reductions, SD did not significantly increase pathogen-specific reads or unique *k*-mer counts. Similar patterns were observed in the Illumina short-read datasets described in Chapter 4 and align with previous studies showing that when enrichment yields high pathogen abundance, eukaryotic depletion strategies may have only a modest effect on microbial composition (159,173). Adaptive sampling

offers another potential strategy for improving sequencing efficiency in long-read quasi-metagenomic workflows. By selectively rejecting non-target reads or preferentially retaining pathogen-associated sequences during sequencing, this approach can increase the proportion of microbial DNA without requiring additional wet-laboratory processing (174–177). Although this approach was not examined in the present study, it warrants consideration as a potential refinement for long-read quasi-metagenomic workflows.

Application of the pathogen-presence likelihood metric to long-read quasi-metagenomic datasets produced results consistent with those observed for short-read data, with the true pathogen ranked highest across matrices and contamination levels. This indicates that the framework performs reliably despite differences between sequencing platforms, including the longer read lengths and higher base-level error rates associated with nanopore sequencing. The metric combines magnitude-based evidence from depth-normalized unique *k*-mer counts with enrichment measured through robust *z*-scores to distinguish pathogen signal from background noise in complex food matrices. Although enrichment metrics such as *z*-scores have been used in clinical metagenomics to separate true signal from background contamination (178,179), their use in long-read workflows remains limited, where interpretation typically relies on read counts or relative abundance thresholds. These results show that enrichment-based interpretation can support reliable signal detection in long-read quasi-metagenomic datasets.

At the same time, applying the framework to long-read data highlights the importance of establishing appropriate filtering thresholds before metric evaluation. Differences in read structure and sequencing error profiles can influence both read classification and unique k-mer recovery, thereby affecting the level of background signal detected within a sample. One exception was observed in the salad kit negative controls, where a low-level *S. enterica* signal was detected despite the absence of intentional inoculation. Although this signal was substantially weaker than those observed in spiked samples and did not interfere with accurate identification of the target pathogen, it highlights the challenge of distinguishing true pathogen signal from low-level background classifications in complex plant-derived matrices. Similar low-level background classifications have been described in metagenomic studies and are commonly attributed to environmental or reagent-derived DNA, laboratory contamination, sequencing artefacts, and analytical challenges associated with complex microbial communities (180,181). Collectively, these findings underscore the need to further refine minimum thresholds for sequencing depth, raw read counts, unique *k*-mer counts, and unique KPMR to improve suppression of background signal while maintaining sensitivity for true pathogen detection in long-read quasi-metagenomic datasets.

Assembly metrics for *L. monocytogenes* pathogen-specific reads were broadly comparable between wet-laboratory protocols. Differences observed in contig counts, genome length, GC content, and completeness were not statistically significant, and assemblies generated from quasi-metagenomic datasets closely approximated those derived from SI references. This concordance suggests that quasi-metagenomic data can preserve strain-level genomic resolution despite the compositional complexity of food matrices (90). Although some quasi-metagenomic assemblies showed modest increases in contig number or low-level contamination relative to isolate genomes, these features are typical of assemblies derived from microbiologically complex datasets and reflect the challenges of reconstructing genomes from mixed sequencing data rather than analytical failure (92,163). Genome coverage analyses support this interpretation, as both NSD and SD datasets showed high mapping proportions and similar mean coverage depths. Lower coverage in low-spike datasets was expected because fewer pathogen-specific reads were recovered, reflecting sampling variability at low DNA input rather than limitations of the experimental protocols.

The ability of long-read quasi-metagenomics to recover outbreak-level genomic relationships was evaluated using wgMLST and SNP-based analyses. For most outbreak datasets, allele and SNP distances between long-read quasi-metagenomic assemblies and corresponding SI genomes fell within interpretive thresholds commonly applied in foodborne pathogen outbreak investigations (152). The roast beef outbreak was the primary exception, with larger wgMLST distances potentially reflecting residual sequencing and base-calling errors, particularly indels, which can inflate allele differences. This may have been influenced by the older ONT R9.4.1 flow cell chemistry used in this study, although the underlying cause was not directly investigated. Notably, these assemblies passed standard isolate genome QC criteria, suggesting that conventional QC metrics may not capture all errors affecting allele-based analyses. In contrast, MINTyper SNP analysis produced clustering patterns nearly identical to those obtained from short-read quasi-metagenomes and corresponding outbreak isolates, indicating that overall strain-level relationships were preserved. Collectively, these findings support the potential of long-read sequencing for outbreak-level resolution, particularly with robust SNP-based approaches (126,183).

In summary, this study demonstrates that long-read quasi-metagenomics can deliver rapid, strain-level pathogen characterization comparable to conventional single-isolate WGS when supported by optimized enrichment, curated references, and standardized analytical workflows. As long-read accuracy continues to improve and databases expand, quasi-metagenomics represents a viable complement to

existing foodborne pathogen surveillance frameworks, with the potential to shorten analytical timelines and enable earlier access to genomic information for outbreak detection and response.

6.0 Chapter 6: Conclusion and future directions

This concluding chapter synthesizes the findings of the preceding studies to evaluate the potential of quasi-metagenomics as a framework for detecting and characterizing priority foodborne pathogens. By integrating optimized laboratory workflows with systematically evaluated bioinformatic approaches, this thesis provides evidence that strain-level pathogen characterization is feasible directly from enriched food samples under the conditions evaluated. This approach has the potential to reduce reliance on pure-culture isolation while generating the genomic resolution needed to support strain-level discrimination and outbreak detection.

6.1 Summary of key findings

The overarching objective of this thesis was to determine whether quasi-metagenomics can achieve the taxonomic and genomic resolution required for public health surveillance and outbreak detection of priority foodborne pathogens. Collectively, the results presented in Chapters 3, 4, and 5 demonstrate that quasi-metagenomics can achieve resolution comparable to conventional isolate-based WGS when combined with curated reference databases, optimized analytical thresholds, and pathogen-informed interpretation.

Chapter 3 established the foundational bioinformatics architecture required for reliable taxonomic classification of complex datasets. Benchmarking of Centrifuge, Kraken 2, and KrakenUniq showed that classifier performance is strongly dependent on reference database composition. Although universal databases offer broad taxonomic coverage, this work demonstrated that carefully curated, taxon-specific databases can achieve equivalent precision when combined with *in-silico* eukaryotic sequence removal and optimized confidence thresholds. Under these conditions, precision values of ≥0.98 were achieved, effectively offsetting the computational challenges traditionally attributed to larger, memory-intensive databases.

Building on this framework, Chapter 4 evaluated how laboratory workflow and computational design affected pathogen recovery from complex food matrices. Although saponin-based depletion reduced eukaryotic DNA in certain high-biomass samples, this reduction did not consistently enhance pathogen signal once selective enrichment was optimized. These results indicate that enrichment conditions have a

greater impact on downstream sequencing performance than eukaryotic DNA depletion alone. The chapter also showed that reference database composition plays a central role in resolving closely related pathogens, particularly *E. coli* and *Shigella* spp. Given the high genomic similarity between these taxa, classification against broader reference databases, such as the FBP-specific database, frequently produced unstable or ambiguous assignments. In contrast, curated species-specific databases improved classification and reduced background signal, resulting in clearer pathogen identification. These findings emphasize that reference database selection is a critical determinant of classification accuracy when analyzing closely related organisms. To address limitations in interpreting quasi-metagenomic data, Chapter 4 also introduced a pathogen-presence likelihood metric supported by explicit decision rules. Rather than relying solely on read counts, which are sensitive to sequencing depth and classifier behavior, the metric incorporated depth-normalized unique *k*-mer counts (KPMR) together with a within-sample enrichment statistic (robust *z*-score). Unique KPMR provided a more stable representation of pathogen-specific signal by reducing sensitivity to redundant coverage. In this study, unique KPMR was sufficient to identify the true pathogen in all cases, while the *z*-score functioned as a secondary discriminator and may be most useful in scenarios where closely ranked taxa require additional resolution. However, overlapping signals between VTEC and *Shigella* highlighted the need for further refinement, leading to the introduction of ratio-based decision rules to manage classification ambiguity. Together, these findings demonstrate that accurate pathogen detection in quasi-metagenomic datasets depends on coordinated optimization of reference database design, classifier behavior, and interpretive criteria.

Chapter 5 extended the quasi-metagenomic framework to long-read sequencing using Oxford Nanopore Technologies. Despite higher raw error rates, long-read data supported robust outbreak detection when analyzed using SNP-based phylogenomic pipelines. While wgMLST performance varied among long-read quasi-metagenomic assemblies, tools such as MINTyper consistently recovered outbreak-associated clustering for *L. monocytogenes*. Application of the pathogen-presence likelihood metric developed in Chapter 4 further demonstrated reliable pathogen detection, correctly identifying the true spike organism across all samples. These findings suggest that long-read sequencing, when paired with appropriate analytical strategies, has the potential to meet the requirements of foodborne pathogen surveillance networks, while offering practical advantages in portability and turnaround time.

Overall, the results presented in Chapters 3–5 demonstrate the feasibility of quasi-metagenomics for strain-level pathogen detection in complex food matrices and form the basis for the methodological contributions described in the following section.

6.2 Strengths and contributions

This thesis makes several contributions to the advancement of quasi-metagenomics for foodborne pathogen surveillance and outbreak detection. A central contribution is the development and application of a pathogen-presence likelihood metric designed to support structured interpretation of quasi-metagenomic data. The metric integrates magnitude-based evidence from depth-normalized unique *k*-mer counts (KPMR) with enrichment-based evidence derived from robust within-sample *z*-scores. Primary detection is based on depth-normalized unique *k*-mer counts, which reflect genome representation rather than total read abundance. Robust *z*-scores are applied as a secondary measure to evaluate relative enrichment within each sample when closely ranked taxa require additional resolution. The metric also incorporates explicit decision rules to address classification challenges associated with closely related pathogens, particularly VTEC and *Shigella* spp. Due to their extensive genomic similarity, both taxa can produce comparable unique *k*-mer signals within the same dataset, leading to ambiguous results. To reduce this effect, a ratio-based override rule was implemented: when *Shigella* unique KPMR reached a defined proportion of the competing VTEC signal, the highest-ranking *Shigella* species was reported as the most likely pathogen. To our knowledge, the integration of rule-based interpretation within a quasi-metagenomic detection metric to resolve VTEC–*Shigella* ambiguity has not been formally described. Together, magnitude thresholds, enrichment metrics, and targeted decision rules provide a clear, statistically grounded approach for distinguishing true pathogen signal from background noise, particularly in complex food matrices.

Another key strength of this research was the development and systematic evaluation of a curated foodborne pathogen-specific reference database for quasi-metagenomic pathogen detection. Rather than relying on default classifier databases, which vary substantially in content, curation status, and taxonomic representation, this work developed a standardized reference database containing curated foodborne pathogen genomes and applied it consistently across all taxonomic classifiers evaluated. This controlled design eliminated database-related confounders by ensuring that each classifier was benchmarked against an identical reference database, minimizing variability introduced by differences in genome inclusion, redundancy, or taxonomic structure. As a result, observed performance differences could be attributed more confidently to classifier algorithms rather than inconsistencies in database composition. Beyond enabling robust benchmarking, the foodborne pathogen-specific reference database improved pathogen detection accuracy and established a standardized analytical resource that formed the foundation for the downstream workflows developed and evaluated throughout this thesis.

Finally, and most importantly, this research establishes an end-to-end quasi-metagenomic workflow designed for foodborne pathogen surveillance. The framework integrates optimized wet-laboratory protocols, including enrichment and host DNA depletion strategies, with curated taxon-specific and enhanced species-specific reference databases tailored to priority foodborne pathogens. These laboratory and database components are coupled with a structured analytical workflow and an explicit decision framework, including the pathogen-presence likelihood metric, to support consistent interpretation of complex sequencing outputs. Importantly, the workflow was developed to accommodate both short-read and long-read sequencing technologies, allowing analytical strategies to be adapted to the characteristics of each platform. By demonstrating that epidemiologically meaningful conclusions can be reached across sequencing platforms when analyses are appropriately tailored, this thesis provides a practical and flexible roadmap for food safety laboratories operating under varying infrastructure, computational, and time constraints.

6.3 Limitations

Several methodological limitations should be considered when interpreting the findings of this research. A key limitation of this thesis is that naturally contaminated food samples were unavailable, necessitating the use of controlled spiking experiments. While controlled spiking experiments do not fully represent the complexity of natural contamination events, they provide a more realistic assessment than purely in silico datasets by incorporating the native food microbiome, background signal arising from laboratory reagents and sequencing workflows, and potential low-level contamination. As such, these datasets capture many of the analytical challenges associated with distinguishing true pathogen signals from background noise. Hence, the controlled experimental design was intentionally selected to support method development, optimization, and benchmarking. By using known pathogen identities, inoculum levels, food matrices, and enrichment conditions, it was possible to systematically evaluate the effects of reference database architecture, host DNA depletion, and bioinformatic workflow performance.

Despite these advantages, controlled spiking experiments cannot fully reproduce the conditions encountered in naturally contaminated foods, where pathogens may occur at low and variable abundances, be sublethally injured, or be heterogeneously distributed within the food matrix. Complex background microbiota may further influence pathogen enrichment, recovery, and subsequent sequencing-based analyses (53,59,62). Consequently, the generalizability of performance observed under controlled experimental conditions to routine food safety surveillance should be interpreted with caution. Additional

validation using naturally contaminated foods will therefore be required to establish performance under real-world conditions. The scope of this study was also limited to a small number of priority foodborne pathogens, including *L. monocytogenes*, VTEC, and *Shigella* spp. Although these pathogens are of major public health importance, analytical performance may differ for pathogens with different genomic characteristics or growth requirements. Accordingly, the broader applicability of the workflow to other pathogens and food matrices will require further evaluation.

Sequencing depth represents an additional practical limitation. The ability to detect and characterize pathogens in quasi-metagenomic datasets depends on the proportion of pathogen-derived DNA recovered during enrichment and sequencing. When pathogen abundance remains low relative to background microbiota or eukaryotic DNA, sequencing depth may be insufficient to recover complete or highly contiguous genomes. Minimum thresholds for raw reads, sequencing depth, and unique KPMR were applied to reduce background noise and improve interpretability. However, these thresholds may require further refinement across sequencing platforms, food matrices, and pathogen concentrations. This consideration is particularly relevant for long-read sequencing technologies, where differences in read length, sequencing yield, and base-level error profiles can influence taxonomic classification and genome reconstruction (184). Additionally, the detection thresholds and decision rules proposed in this work were established using controlled spike-in datasets. While this approach enabled systematic calibration of classifier performance, it does not fully capture the complexity of naturally contaminated samples, which often exhibit greater background microbial diversity, variable pathogen abundance, and matrix-specific effects that may influence detection performance.

One example is the selection of the confidence threshold used for taxonomic classification with Kraken 2 and KrakenUniq. A confidence threshold of 0.1 was applied in Chapter 3; however, at the time the analytical framework was developed, no studies had systematically evaluated the impact of confidence score thresholds on pathogen detection performance in quasi-metagenomic or foodborne pathogen sequencing datasets. Consequently, this threshold was selected as a pragmatic starting point to assess its effect on reducing spurious taxonomic assignments while maintaining detection sensitivity. Since completion of this work, subsequent benchmarking studies have generally supported the use of low-to-moderate confidence thresholds as an effective balance between sensitivity and specificity (82,185). Although this data was not available during framework development, the findings retrospectively support the selection of a 0.1 confidence threshold as a reasonable starting point for evaluating pathogen detection performance. Nevertheless, the optimal confidence threshold is likely to vary depending on the target

pathogen, sequencing platform, reference database composition, and complexity of the food matrix. Therefore, establishing application-specific confidence thresholds will be an important step toward standardizing quasi-metagenomic workflows for routine foodborne pathogen surveillance.

Finally, long-read sequencing in this study (Chapter 5) was performed using Oxford Nanopore R9.4.1 flow cells, which represented the standard chemistry available when the experiments were initiated. Although this platform remains robust and capable of generating high-throughput sequencing using well-established protocols, the per-read error rate is higher than that of more recent chemistries (186,187). Insertion and deletion errors, particularly within homopolymer regions, may have affected genome assembly accuracy and variant interpretation, potentially influencing strain-level resolution in this study.

These limitations underscore the importance of optimizing both laboratory and analytical frameworks to ensure reliable pathogen detection and characterization, while also highlighting key opportunities for future methodological refinement and continued evaluation of quasi-metagenomic workflows for foodborne pathogen surveillance and outbreak detection.

6.4 Future directions

The adoption of quasi-metagenomics represents an important shift in food safety surveillance, moving from culture-dependent workflows toward sequence-based, data-driven approaches. By eliminating the need for secondary isolation prior to sequencing, quasi-metagenomic workflows can shorten turnaround times by approximately 24–48 hours. Earlier access to genomic data can facilitate faster epidemiological linkage between cases, food sources, and production environments, ultimately supporting more timely public health interventions. In addition to detecting target pathogens, quasi-metagenomic analyses may identify related non-target organisms, such as *Listeria innocua* and *Listeria welshimeri*. The detection of these species may provide additional context regarding environmental contamination and conditions that support *Listeria* persistence within food processing environments (188).

Several areas of future work will be important for advancing the application of quasi-metagenomics for foodborne pathogen surveillance. A key priority will be the validation of the proposed workflow using naturally contaminated samples. While controlled spike-in experiments provide a structured framework for benchmarking analytical performance, they cannot fully replicate the biological complexity of natural contamination events. Naturally contaminated samples capture variability in pathogen distribution, background microbiota, and matrix composition (53). Future studies should focus on validating the proposed quasi-metagenomic framework using naturally contaminated food samples representing a

broader range of matrices, contamination levels, and environmental conditions. Such samples would better capture the variability encountered in routine food safety surveillance, including differences in pathogen abundance, distribution within the food matrix, and interactions with background microbiota. Evaluating workflow performance under these conditions will provide a more realistic assessment of detection sensitivity, pathogen recovery, and the ability to resolve outbreak-associated strains, while helping to establish operational performance characteristics and determine the generalizability of the framework for routine foodborne pathogen surveillance and outbreak response.

In parallel, additional work is needed to improve both the operational scalability and robustness of the analytical workflow. Although the multi-step framework developed in Chapter 3 and evaluated in subsequent chapters was necessary for method optimization and benchmarking, routine implementation will require streamlined and automated bioinformatic pipelines that minimize user intervention, improve reproducibility, and reduce opportunities for workflow failure while supporting scalable and consistent generation of results across laboratory settings (189). Future studies should therefore evaluate pipeline reliability, computational resource requirements, failure rates, and opportunities for workflow simplification. In addition, although the framework demonstrated robust detection of the target pathogens evaluated in this study, its performance should be assessed using emerging, rare, or intentionally excluded pathogen genomes to quantify the impact of incomplete reference databases on detection sensitivity and taxonomic resolution. Such evaluations would provide a more realistic assessment of workflow performance under routine surveillance conditions, where novel strains and previously uncharacterized variants may be encountered.

Further work is also needed to establish the minimum sequencing depth required for both reliable pathogen detection and downstream genomic characterization. While relatively low sequencing depths may be sufficient for pathogen detection, more extensive genome coverage is typically needed to enable reliable serotyping, subtyping, phylogenetic analysis, and outbreak-level genomic characterization (190). As a result, sequencing depth should be optimized based on the specific objectives of the downstream analyses. Since sequencing depth requirements are likely to vary with pathogen abundance, food matrix, sequencing platform, and background microbial complexity, subsampling analyses could be used to systematically determine the minimum read depth needed to achieve reliable pathogen detection and downstream genomic characterization.

Future studies should also incorporate technical replicates to complement the biological replicates evaluated in this work. Although technical replication was beyond the scope of this study, it would enable

analytical variability associated with DNA extraction, library preparation, and sequencing to be quantified, providing additional evidence of workflow reproducibility and supporting future standardization for routine public health implementation. Together, validation using naturally contaminated samples, assessment of database robustness, optimization of analytical thresholds across diverse foodborne pathogens and food matrices, evaluation of technical reproducibility, determination of sequencing depth requirements, and development of streamlined bioinformatic pipelines will be essential for refining analytical performance and translating quasi-metagenomic approaches into practical tools for routine foodborne pathogen surveillance and outbreak response. These efforts will help establish robust classification parameters, improve detection sensitivity and specificity, reduce false-positive classifications, and support reliable implementation across both short- and long-read sequencing platforms.

Continued methodological development will also be important. Since the completion of this study, Oxford Nanopore sequencing technology has undergone substantial improvements in flow cell architecture and pore chemistry. Newer chemistries provide higher raw read accuracy and improved resolution of homopolymer regions, and further gains can be achieved through updated basecalling algorithms and high-accuracy models (191). These developments are expected to reduce systematic sequencing errors, improve consensus genome quality, and strengthen strain-level characterization (192). Evaluating these improvements will be crucial for evaluating the suitability of long-read sequencing for outbreak detection quasi-metagenomic workflows.

Further work is also necessary to refine the analytical thresholds used in the pathogen-presence likelihood metric. In this study, thresholds were empirically derived from spike-in datasets and may not fully reflect real-world contamination scenarios. Importantly, requirements for simple pathogen detection differ from those needed for genomic characterization and outbreak resolution. While detection is a critical first step, effective surveillance also requires sufficient genomic data for strain-level discrimination (31,193). Future studies should focus on optimizing the minimum unique KPMR and sequencing depth needed not only for confident detection but also for reliable genome characterization and outbreak analysis. These thresholds will likely vary by pathogen, food matrix, and sequencing platform. Establishing clear criteria that distinguish detection-level from characterization-level data will be essential for translating quasi-metagenomic approaches into routine surveillance applications.

Finally, emerging analytical approaches may further improve interpretation of complex quasi-metagenomic datasets. In particular, machine learning–based classification frameworks may offer improved discrimination among closely related taxa, such as *Shigella* spp. and *E. coli*, especially when pathogen

abundance is low or genomic signatures overlap (194,195). Continued integration of improved sequencing technologies, curated reference databases, parameter optimization and advanced analytical approaches will be critical for realizing the full potential of quasi-metagenomics in foodborne pathogen surveillance.

6.5 Concluding remarks

This thesis was motivated by a key challenge in foodborne pathogen surveillance: the dependence on isolate recovery before genomic characterization can begin. The findings provide evidence that quasi-metagenomics can help address this limitation by enabling pathogen detection and strain-level genomic characterization directly from enriched food samples under the conditions evaluated.

By integrating optimized laboratory workflows with curated reference databases and structured interpretation through the pathogen-presence likelihood metric, this work establishes an end-to-end methodological framework for analyzing quasi-metagenomic data from complex food matrices. Importantly, the findings highlight that successful application requires more than pathogen detection alone; sequencing must generate sufficient genomic information to support reliable strain-level discrimination and outbreak analysis. The extension of this framework across both Illumina short-read and Oxford Nanopore long-read sequencing further demonstrates its adaptability across sequencing platforms.

Collectively, this work provides a methodological foundation for the continued development of quasi-metagenomics as a complementary approach to conventional isolate-based WGS, with the potential to reduce the time required to obtain actionable genomic information. Further evaluation using naturally contaminated samples, broader pathogen and matrix combinations, and evolving sequencing technologies will be important to establish performance under real-world conditions and determine how quasi-metagenomic approaches can be most effectively integrated into routine foodborne pathogen surveillance.

7.0 References

1. Mathers CD. History of global burden of disease assessment at the World Health Organization. Archives of Public Health. 2020 Aug 24;78(1):77. doi:10.1186/s13690-020-00458-3

2. Scallan Walter EJ, Cui Z, Tierney R, Griffin PM, Hoekstra RM, Payne DC, et al. Foodborne Illness Acquired in the United States-Major Pathogens, 2019. Emerg Infect Dis. 2025 Apr;31(4):669–77. doi:10.3201/eid3104.240913 PubMed PMID: 40133035; PubMed Central PMCID: PMC11950263.

3. Thomas MK, Murray R, Flockhart L, Pintar K, Fazil A, Nesbitt A, et al. Estimates of Foodborne Illness–Related Hospitalizations and Deaths in Canada for 30 Specified Pathogens and Unspecified Agents. Foodborne Pathog Dis. 2015 Oct 1;12(10):820–7. doi:10.1089/fpd.2015.1966 PubMed PMID: 26259128; PubMed Central PMCID: PMC4601674.

4. Hoffmann S, White AE, McQueen RB, Ahn JW, Gunn-Sandell LB, Scallan Walter EJ. Economic Burden of Foodborne Illnesses Acquired in the United States. Foodborne Pathogens and Disease. 2025 Jan 1;22(1):4–14. doi:10.1089/fpd.2023.0157

5. Thomas MK, Vriezen R, Farber JM, Currie A, Schlech W, Fazil A. Economic Cost of a Listeria monocytogenes Outbreak in Canada, 2008. Foodborne Pathogens and Disease. 2015 Nov 19;12(12):966–71. doi:10.1089/fpd.2015.1965

6. Switaj TL, Winter KJ, Christensen S. Diagnosis and Management of Foodborne Illness. AFP. 2015 Sep 1;92(5):358–65.

7. Zhang H, Yamamoto E, Murphy J, Carrillo C, Locas A. Shiga Toxin–Producing *Escherichia coli* (STEC) and STEC-Associated Virulence Genes in Raw Ground Pork in Canada. Journal of Food Protection. 2021 Nov 1;84(11):1956–64. doi:10.4315/JFP-21-147

8. Catford A, Kouamé V, Martinez-Perez A, Gill A, Buenaventura E, Couture H, et al. Risk Profile on Non-O157 Verotoxin-Producing Escherichia coli in Produce, Beef, Milk and Dairy Products in Canada. International Food Risk Analysis Journal. 2014 Aug 10;4:1–25. doi:10.5772/59208

9. Gould LH, Mody RK, Ong KL, Clogher P, Cronquist AB, Garman KN, et al. Increased recognition of non-O157 Shiga toxin-producing Escherichia coli infections in the United States during 2000-2010: epidemiologic features and comparison with E. coli O157 infections. Foodborne Pathog Dis. 2013 May;10(5):453–60. doi:10.1089/fpd.2012.1401 PubMed PMID: 23560425.

10. Freedman SB, Kar NCAJ van de, Tarr PI. Shiga Toxin–Producing Escherichia coli and the Hemolytic–Uremic Syndrome. New England Journal of Medicine. 2023 Oct 11;389(15):1402–14. doi:10.1056/NEJMra2108739

11. Marshall KE, Hexemer A, Seelman SL, Fatica MK, Blessington T, Hajmeer M, et al. Lessons Learned from a Decade of Investigations of Shiga Toxin–Producing Escherichia coli Outbreaks Linked to

Leafy Greens, United States and Canada. Emerg Infect Dis. 2020 Oct;26(10):2319–28. doi:10.3201/eid2610.191418 PubMed PMID: 32946367; PubMed Central PMCID: PMC7510726.

12. Coulombe G, Catford A, Martinez-Perez A, Buenaventura E. Outbreaks of Escherichia coli O157:H7 Infections Linked to Romaine Lettuce in Canada from 2008 to 2018: An Analysis of Food Safety Context. J Food Prot. 2020 Aug 1;83(8):1444–62. doi:10.4315/JFP-20-029 PubMed PMID: 32297933.

13. Currie A, Honish L, Cutler J, Locas A, Lavoie MC, Gaulin C, et al. Outbreak of Escherichia coli O157:H7 Infections Linked to Mechanically Tenderized Beef and the Largest Beef Recall in Canada, 2012. Journal of Food Protection. 2019 Sep 1;82(9):1532–8. doi:10.4315/0362-028X.JFP-19-005

14. Listeriosis [Internet]. [cited 2026 Feb 21]. Available from: https://www.who.int/news-room/fact-sheets/detail/listeriosis

15. Osek J, Lachtara B, Wieczorek K. Listeria monocytogenes – How This Pathogen Survives in Food-Production Environments? Frontiers in Microbiology [Internet]. 2022 [cited 2023 Mar 27];13. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2022.866462

16. McIntyre L, Wilcott L, Naus M. Listeriosis outbreaks in British Columbia, Canada, caused by soft ripened cheese contaminated from environmental sources. Biomed Res Int. 2015;2015:131623. doi:10.1155/2015/131623 PubMed PMID: 25918702; PubMed Central PMCID: PMC4396127.

17. Hanson H, Whitfield Y, Lee C, Badiani T, Minielly C, Fenik J, et al. Listeria monocytogenes Associated with Pasteurized Chocolate Milk, Ontario, Canada - Volume 25, Number 3—March 2019 - Emerging Infectious Diseases journal - CDC [Internet]. doi:10.3201/eid2503.180742

18. Currie A, Farber JM, Nadon C, Sharma D, Whitfield Y, Gaulin C, et al. Multi-Province Listeriosis Outbreak Linked to Contaminated Deli Meat Consumed Primarily in Institutional Settings, Canada, 2008. Foodborne Pathog Dis. 2015 Aug;12(8):645–52. doi:10.1089/fpd.2015.1939 PubMed PMID: 26258258.

19. Conrad AR, Tubach S, Cantu V, Webb LM, Stroika S, Moris S, et al. Listeria monocytogenes Illness and Deaths Associated With Ongoing Contamination of a Multiregional Brand of Ice Cream Products, United States, 2010-2015. Clin Infect Dis. 2023 Jan 6;76(1):89–95. doi:10.1093/cid/ciac550 PubMed PMID: 35797187; PubMed Central PMCID: PMC10201537.

20. Self JL, Conrad A, Stroika S, Jackson A, Whitlock L, Jackson KA, et al. Multistate Outbreak of Listeriosis Associated with Packaged Leafy Green Salads, United States and Canada, 2015–2016. Emerg Infect Dis. 2019 Aug;25(8):1461–8. doi:10.3201/eid2508.180761 PubMed PMID: 31310227; PubMed Central PMCID: PMC6649349.

21. Garner D, Kathariou S. Fresh Produce–Associated Listeriosis Outbreaks, Sources of Concern, Teachable Moments, and Insights. Journal of Food Protection. 2016 Jan 1;79(2):337–44. doi:10.4315/0362-028X.JFP-15-387

22. Pereira E, Conrad A, Tesfai A, Palacios A, Kandar R, Kearney A, et al. Multinational Outbreak of Listeria monocytogenes Infections Linked to Enoki Mushrooms Imported from the Republic of Korea 2016-

2020. J Food Prot. 2023 Jul;86(7):100101. doi:10.1016/j.jfp.2023.100101 PubMed PMID: 37169291; PubMed Central PMCID: PMC10947956.

23. Hendrick J, Raqib R, Noor Z, Faruque ASG, Haque R, Petri WA. Shigellosis. The Lancet. 2025 Oct 4;406(10511):1508–19. doi:10.1016/S0140-6736(25)01033-5 PubMed PMID: 40915309.

24. Reller ME, Nelson JM, Mølbak K, Ackman DM, Schoonmaker-Bopp DJ, Root TP, et al. A Large, Multiple-Restaurant Outbreak of Infection with Shigella flexneri Serotype 2a Traced to Tomatoes. Clinical Infectious Diseases. 2006 Jan 15;42(2):163–9. doi:10.1086/498900

25. Nygren BL, Schilling KA, Blanton EM, Silk BJ, Cole DJ, Mintz ED. Foodborne outbreaks of shigellosis in the USA, 1998–2008. Epidemiol Infect. 2013 Feb;141(2):233–41. doi:10.1017/S0950268812000222 PubMed PMID: 22361246; PubMed Central PMCID: PMC4610123.

26. Kozak GK, MacDonald D, Landry L, Farber JM. Foodborne outbreaks in Canada linked to produce: 2001 through 2009. J Food Prot. 2013 Jan;76(1):173–83. doi:10.4315/0362-028X.JFP-12-126 PubMed PMID: 23317873.

27. Gaudreau C, Pilon PA, Cornut G, Marchand-Senecal X, Bekal S. Shigella flexneri with Ciprofloxacin Resistance and Reduced Azithromycin Susceptibility, Canada, 2015. Emerg Infect Dis. 2016 Nov;22(11):2016–8. doi:10.3201/eid2211.160883 PubMed PMID: 27767915; PubMed Central PMCID: PMC5088001.

28. Neemuchwala A, Zittermann S, Johnson K, Middleton D, Stapleton PJ, Ravirajan V, et al. Whole genome sequencing of increased number of azithromycin-resistant Shigella flexneri 1b isolates in Ontario. Sci Rep. 2023 Oct 3;13:16582. doi:10.1038/s41598-023-36733-w PubMed PMID: 37789081; PubMed Central PMCID: PMC10547750.

29. Stefanovic A, Matic N, Ritchie G, Lowe CF, Leung V, Hull M, et al. Multidrug-Resistant Shigella sonnei Bacteremia among Persons Experiencing Homelessness, Vancouver, British Columbia, Canada. Emerg Infect Dis. 2023 Aug;29(8):1668–71. doi:10.3201/eid2908.230323 PubMed PMID: 37486309; PubMed Central PMCID: PMC10370870.

30. Weight of evidence - factors to consider for appropriate and timely action in a foodborne illness oubreak investigation. Ottawa: Health Canada; 2011.

31. Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, et al. PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Euro Surveill. 2017 Jun 8;22(23). doi:10.2807/1560-7917.ES.2017.22.23.30544 PubMed PMID: 28662764; PubMed Central PMCID: PMC5479977.

32. Lynch T, Petkau A, Knox N, Graham M, Van Domselaar G. A Primer on Infectious Disease Bacterial Genomics. Clin Microbiol Rev. 2016 Oct;29(4):881–913. doi:10.1128/CMR.00001-16 PubMed PMID: 28590251; PubMed Central PMCID: PMC5010755.

33. Smith CR, Bond H, Kearney A, Chau K, Chui L, Gerrie M, et al. Fermenting a place in history: The first outbreak of Escherichia coli O157 associated with kimchi in Canada. Epidemiol Infect. 2023 Jun 8;151:e106. doi:10.1017/S0950268823000882 PubMed PMID: 37288513; PubMed Central PMCID: PMC10369422.

34. Morton V, Kandar R, Kearney A, Hamel M, Nadon C. Transition to Whole Genome Sequencing Surveillance: The Impact on National Outbreak Detection and Response for Listeria monocytogenes, Salmonella, Shiga Toxin–Producing Escherichia coli, and Shigella Clusters in Canada, 2015–2021. Foodborne Pathogens and Disease. 2024 Nov 1;21(11):689–97. doi:10.1089/fpd.2024.0041

35. Davedow T, Carleton H, Kubota K, Palm D, Schroeder M, Gerner-Smidt P, et al. PulseNet International Survey on the Implementation of Whole Genome Sequencing in Low and Middle-Income Countries for Foodborne Disease Surveillance. Foodborne Pathogens and Disease. 2022 May;19(5):332–40. doi:10.1089/fpd.2021.0110

36. Imdad A, Retzer F, Thomas LS, McMillian M, Garman K, Rebeiro PF, et al. Impact of Culture-Independent Diagnostic Testing on Recovery of Enteric Bacterial Infections. Clin Infect Dis. 2018 Jun 1;66(12):1892–8. doi:10.1093/cid/cix1128 PubMed PMID: 29293941.

37. Gill A. The Importance of Bacterial Culture to Food Microbiology in the Age of Genomics. Front Microbiol. 2017;8:777. doi:10.3389/fmicb.2017.00777 PubMed PMID: 28507541; PubMed Central PMCID: PMC5410609.

38. McMahon TC, Kingombe CB, Mathews A, Seyer K, Wong A, Blais BW, et al. Microbial Antagonism in Food-Enrichment Culture: Inhibition of Shiga Toxin-Producing Escherichia coli and Shigella Species. Front Microbiol. 2022;13:880043. doi:10.3389/fmicb.2022.880043 PubMed PMID: 35814680; PubMed Central PMCID: PMC9259949.

39. Canada H. Official Methods for the Microbiological Analysis of Foods [policies;guidance] [Internet]. 2013 [cited 2026 Feb 21]. Available from: https://www.canada.ca/en/health-canada/services/food-nutrition/research-programs-analytical-methods/analytical-methods/compendium-methods/official-methods-microbiological-analysis-foods-compendium-analytical-methods.html

40. Enrichment Culture - an overview | ScienceDirect Topics [Internet]. [cited 2026 Feb 21]. Available from: https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/enrichment-culture

41. Ogunremi D, Dupras AA, Naushad S, Gao R, Duceppe MO, Omidi K, et al. A New Whole Genome Culture-Independent Diagnostic Test (WG-CIDT) for Rapid Detection of Salmonella in Lettuce. Front Microbiol. 2020;11. doi:10.3389/fmicb.2020.00602

42. Qin X, Klein EJ, Galanakis E, Thomas AA, Stapp JR, Rich S, et al. Real-Time PCR Assay for Detection and Differentiation of Shiga Toxin-Producing Escherichia coli from Clinical Samples. J Clin Microbiol. 2015 Jul;53(7):2148–53. doi:10.1128/JCM.00115-15 PubMed PMID: 25926491; PubMed Central PMCID: PMC4473226.

43. McAdam AJ. Unforeseen Consequences: Culture-Independent Diagnostic Tests and Epidemiologic Tracking of Foodborne Pathogens. Journal of Clinical Microbiology. 2017 Jul 1;55(7):1978–9. doi:10.1128/JCM.00678-17 PubMed PMID: 28468857.

44. Cukrovany AE, Wroblewski D, Wirth SE, Thompson LM, Saylors AL, Connors JA, et al. Shiga Toxin-Producing Escherichia coli Testing in New York 2011-2022 Reveals Increase in Non-O157 Identifications. Foodborne Pathog Dis. 2025 Apr;22(4):273–80. doi:10.1089/fpd.2023.0152 PubMed PMID: 38452173.

45. Wylezich C, Papa A, Beer M, Höper D. A Versatile Sample Processing Workflow for Metagenomic Pathogen Detection. Sci Rep. 2018 Aug 30;8(1):13108. doi:10.1038/s41598-018-31496-1 PubMed PMID: 30166611; PubMed Central PMCID: PMC6117295.

46. Forbes JD, Knox NC, Ronholm J, Pagotto F, Reimer A. Metagenomics: The Next Culture-Independent Game Changer. Front Microbiol. 2017;8:1069. doi:10.3389/fmicb.2017.01069 PubMed PMID: 28725217; PubMed Central PMCID: PMC5495826.

47. Miller S, Naccache SN, Samayoa E, Messacar K, Arevalo S, Federman S, et al. Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. Genome Res. 2019 May;29(5):831–42. doi:10.1101/gr.238170.118 PubMed PMID: 30992304; PubMed Central PMCID: PMC6499319.

48. Hasan MR, Sundararaju S, Tang P, Tsui KM, Lopez AP, Janahi M, et al. A metagenomics-based diagnostic approach for central nervous system infections in hospital acute care setting. Sci Rep. 2020 Jul 8;10(1):11194. doi:10.1038/s41598-020-68159-z PubMed PMID: 32641704; PubMed Central PMCID: PMC7343800.

49. Ruppé E, Lazarevic V, Girard M, Mouton W, Ferry T, Laurent F, et al. Clinical metagenomics of bone and joint infections: a proof of concept study. Sci Rep. 2017 Aug 10;7(1):7718. doi:10.1038/s41598-017-07546-5 PubMed PMID: 28798333; PubMed Central PMCID: PMC5552814.

50. Peterson CL, Alexander D, Chen JCY, Adam H, Walker M, Ali J, et al. Clinical Metagenomics Is Increasingly Accurate and Affordable to Detect Enteric Bacterial Pathogens in Stool. Microorganisms. 2022 Feb 15;10(2):441. doi:10.3390/microorganisms10020441 PubMed PMID: 35208895; PubMed Central PMCID: PMC8880012.

51. Forbes JD, Knox NC, Peterson CL, Reimer AR. Highlighting Clinical Metagenomics for Enhanced Diagnostic Decision-making: A Step Towards Wider Implementation. Computational and Structural Biotechnology Journal. 2018 Jan 1;16:108–20. doi:10.1016/j.csbj.2018.02.006

52. Margot H, Stephan R, Tasara T. Mungo bean sprout microbiome and changes associated with culture based enrichment protocols used in detection of Gram-negative foodborne pathogens. Microbiome. 2016 Sep 6;4(1):48. doi:10.1186/s40168-016-0193-y PubMed PMID: 27600392; PubMed Central PMCID: PMC5012049.

53. Ottesen A, Ramachandran P, Reed E, White JR, Hasan N, Subramanian P, et al. Enrichment dynamics of Listeria monocytogenes and the associated microbiome from naturally contaminated ice cream linked to a listeriosis outbreak. BMC Microbiol. 2016 Nov 16;16(1):275. doi:10.1186/s12866-016-0894-1 PubMed PMID: 27852235; PubMed Central PMCID: PMC5112668.

54. Aw TG, Wengert S, Rose JB. Metagenomic analysis of viruses associated with field-grown and retail lettuce identifies human and animal viruses. Int J Food Microbiol. 2016 Apr 16;223:50–6. doi:10.1016/j.ijfoodmicro.2016.02.008 PubMed PMID: 26894328.

55. Kawai T, Sekizuka T, Yahata Y, Kuroda M, Kumeda Y, Iijima Y, et al. Identification of Kudoa septempunctata as the causative agent of novel food poisoning outbreaks in Japan by consumption of Paralichthys olivaceus in raw fish. Clin Infect Dis. 2012 Apr;54(8):1046–52. doi:10.1093/cid/cir1040 PubMed PMID: 22281845.

56. Escobar-Zepeda A, Sanchez-Flores A, Quirasco Baruch M. Metagenomic analysis of a Mexican ripened cheese reveals a unique complex microbiota. Food Microbiol. 2016 Aug;57:116–27. doi:10.1016/j.fm.2016.02.004 PubMed PMID: 27052710.

57. Leonard SR, Mammel MK, Lacher DW, Elkins CA. Strain-Level Discrimination of Shiga Toxin-Producing Escherichia coli in Spinach Using Metagenomic Sequencing. PLOS ONE. 2016 Dec 8;11(12):e0167870. doi:10.1371/journal.pone.0167870

58. Ottesen A, Ramachandran P, Chen Y, Brown E, Reed E, Strain E. Quasimetagenomic source tracking of Listeria monocytogenes from naturally contaminated ice cream. BMC Infectious Diseases. 2020 Jan 29;20(1):83. doi:10.1186/s12879-019-4747-z

59. Hyeon JY, Li S, Mann DA, Zhang S, Li Z, Chen Y, et al. Quasimetagenomics-Based and Real-Time-Sequencing-Aided Detection and Subtyping of Salmonella enterica from Food Samples. Appl Environ Microbiol. 2018 Feb 15;84(4):e02340-17. doi:10.1128/AEM.02340-17 PubMed PMID: 29196295.

60. Wagner E, Fagerlund A, Langsrud S, Møretrø T, Jensen MR, Moen B. Surveillance of Listeria monocytogenes: Early Detection, Population Dynamics, and Quasimetagenomic Sequencing during Selective Enrichment. Appl Environ Microbiol. 87(24):e01774-21. doi:10.1128/AEM.01774-21 PubMed PMID: 34613762; PubMed Central PMCID: PMC8612253.

61. Buytaers FE, Saltykova A, Mattheus W, Verhaegen B, Roosens NHC, Vanneste K, et al. Application of a strain-level shotgun metagenomics approach on food samples: resolution of the source of a Salmonella food-borne outbreak. Microb Genom. 2021 Apr 7;7(4):000547. doi:10.1099/mgen.0.000547 PubMed PMID: 33826490; PubMed Central PMCID: PMC8208685.

62. Buytaers FE, Saltykova A, Denayer S, Verhaegen B, Vanneste K, Roosens NHC, et al. A Practical Method to Implement Strain-Level Metagenomics-Based Foodborne Outbreak Investigation and Source Tracking in Routine. Microorganisms. 2020 Aug 5;8(8):1191. doi:10.3390/microorganisms8081191 PubMed PMID: 32764329; PubMed Central PMCID: PMC7463776.

63. Kocurek B, Ramachandran P, Grim CJ, Morin P, Howard L, Ottesen A, et al. Application of quasimetagenomics methods to define microbial diversity and subtype Listeria monocytogenes in dairy and seafood production facilities. Microbiol Spectr. 11(6):e01482-23. doi:10.1128/spectrum.01482-23 PubMed PMID: 37812012; PubMed Central PMCID: PMC10714831.

64. Bradford LM, Carrillo C, Wong A. Managing false positives during detection of pathogen sequences in shotgun metagenomics datasets. BMC Bioinformatics. 2024 Dec 3;25:372. doi:10.1186/s12859-024-05952-x PubMed PMID: 39627685; PubMed Central PMCID: PMC11613480.

65. Ganda E, Beck KL, Haiminen N, Silverman JD, Kawas B, Cronk BD, et al. DNA Extraction and Host Depletion Methods Significantly Impact and Potentially Bias Bacterial Detection in a Biological Fluid. mSystems. 2021 Jun 29;6(3):e0061921. doi:10.1128/mSystems.00619-21 PubMed PMID: 34128697; PubMed Central PMCID: PMC8574158.

66. Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome. 2018 Feb 27;6(1):42. doi:10.1186/s40168-018-0426-3

67. Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nature Biotechnology. 2019 Jul;37(7):7. doi:10.1038/s41587-019-0156-5

68. Hasan MR, Rawat A, Tang P, Jithesh PV, Thomas E, Tan R, et al. Depletion of Human DNA in Spiked Clinical Specimens for Improvement of Sensitivity of Pathogen Detection by Next-Generation Sequencing. J Clin Microbiol. 2016 Apr;54(4):919–27. doi:10.1128/JCM.03050-15 PubMed PMID: 26763966; PubMed Central PMCID: PMC4809942.

69. Kok NA, Peker N, Schuele L, de Beer JL, Rossen JWA, Sinha B, et al. Host DNA depletion can increase the sensitivity of Mycobacterium spp. detection through shotgun metagenomics in sputum. Frontiers in Microbiology [Internet]. 2022 [cited 2023 May 24];13. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2022.949328

70. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. Brief Bioinformatics. 2017 Sep 23. doi:10.1093/bib/bbx120 PubMed PMID: 29028872.

71. Lu J, Salzberg SL. Removing contaminants from databases of draft genomes. PLOS Computational Biology. 2018 Jun 25;14(6):e1006277. doi:10.1371/journal.pcbi.1006277

72. Lindner BG, Gerhardt K, Feistel DJ, Rodriguez-R LM, Hatt JK, Konstantinidis KT. A user's guide to the bioinformatic analysis of shotgun metagenomic sequence data for bacterial pathogen detection. International Journal of Food Microbiology. 2024 Jan 30;410:110488. doi:10.1016/j.ijfoodmicro.2023.110488

73. Longo MS, O'Neill MJ, O'Neill RJ. Abundant Human DNA Contamination Identified in Non-Primate Genome Databases. PLOS ONE. 2011 Feb 16;6(2):e16410. doi:10.1371/journal.pone.0016410

74. Kryukov K, Imanishi T. Human Contamination in Public Genome Assemblies. PLOS ONE. 2016 Sep 9;11(9):e0162424. doi:10.1371/journal.pone.0162424

75. Breitwieser FP, Pertea M, Zimin A, Salzberg SL. Human contamination in bacterial genomes has created thousands of spurious proteins. Genome Res. 2019 May 7;gr.245373.118. doi:10.1101/gr.245373.118 PubMed PMID: 31064768.

76. Merchant S, Wood DE, Salzberg SL. Unexpected cross-species contamination in genome sequencing projects. PeerJ. 2014;2:e675. doi:10.7717/peerj.675 PubMed PMID: 25426337; PubMed Central PMCID: PMC4243333.

77. Jiao JY, Liu L, Hua ZS, Fang BZ, Zhou EM, Salam N, et al. Microbial dark matter coming to light: challenges and opportunities. National Science Review. 2021 Mar 1;8(3):nwaa280. doi:10.1093/nsr/nwaa280

78. Ye SH, Siddle KJ, Park DJ, Sabeti PC. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell. 2019 Aug;178(4):779–94. doi:10.1016/j.cell.2019.07.010

79. R. Marcelino V, Holmes EC, Sorrell TC. The use of taxon-specific reference databases compromises metagenomic classification. BMC Genomics. 2020 Feb 27;21(1):184. doi:10.1186/s12864-020-6592-2

80. phac-nml/mikrokondo [Nextflow] [Internet]. National Microbiology Laboratory; 2026 [cited 2026 Feb 21]. Available from: https://github.com/phac-nml/mikrokondo

81. Petit RA, Read TD. Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial Genomes. mSystems. 2020 Aug 4;5(4):e00190-20. doi:10.1128/mSystems.00190-20

82. Wright RJ, Comeau AM, Langille MGI. From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. Microb Genom. 2023 Mar 3;9(3):000949. doi:10.1099/mgen.0.000949 PubMed PMID: 36867161; PubMed Central PMCID: PMC10132073.

83. Salzberg SL, Wood DE. Releasing the Kraken. Frontiers in Bioinformatics [Internet]. 2021 [cited 2023 May 29];1. Available from: https://www.frontiersin.org/articles/10.3389/fbinf.2021.808003

84. Ebinger A, Fischer S, Höper D. A theoretical and generalized approach for the assessment of the sample-specific limit of detection for clinical metagenomics. Comput Struct Biotechnol J. 2020 Dec 26;19:732–42. doi:10.1016/j.csbj.2020.12.040 PubMed PMID: 33552445; PubMed Central PMCID: PMC7822954.

85. Lu J, Rincon N, Wood DE, Breitwieser FP, Pockrandt C, Langmead B, et al. Metagenome analysis using the Kraken software suite. Nat Protoc. 2022 Dec;17(12):12. doi:10.1038/s41596-022-00738-y

86. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol. 2014 Apr;10(4):e1003531. doi:10.1371/journal.pcbi.1003531 PubMed PMID: 24699258; PubMed Central PMCID: PMC3974642.

87. Jaudou S, Deneke C, Tran ML, Schuh E, Goehler A, Vorimore F, et al. A step forward for Shiga toxin-producing Escherichia coli identification and characterization in raw milk using long-read metagenomics. Microbial Genomics. 2022;8(11):000911. doi:10.1099/mgen.0.000911

88. Forghani F, Li S, Zhang S, Mann DA, Deng X, den Bakker HC, et al. Salmonella enterica and Escherichia coli in Wheat Flour: Detection and Serotyping by a Quasimetagenomic Approach Assisted by Magnetic Capture, Multiple-Displacement Amplification, and Real-Time Sequencing. Applied and Environmental Microbiology. 2020 Jun 17;86(13):e00097-20. doi:10.1128/AEM.00097-20

89. Olson ND, Treangen TJ, Hill CM, Cepeda-Espinoza V, Ghurye J, Koren S, et al. Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. Brief Bioinformatics. 2017 Aug 7. doi:10.1093/bib/bbx098 PubMed PMID: 28968737.

90. Commichaux S, Javkar K, Ramachandran P, Nagarajan N, Bertrand D, Chen Y, et al. Evaluating the accuracy of Listeria monocytogenes assemblies from quasimetagenomic samples using long and short reads. BMC Genomics. 2021 Dec;22(1):389. doi:10.1186/s12864-021-07702-2

91. Buytaers FE, Saltykova A, Denayer S, Verhaegen B, Vanneste K, Roosens NHC, et al. Towards Real-Time and Affordable Strain-Level Metagenomics-Based Foodborne Outbreak Investigations Using Oxford Nanopore Sequencing Technologies. Frontiers in Microbiology. 2021;12:3372. doi:10.3389/fmicb.2021.738284

92. Espinosa E, Bautista R, Larrosa R, Plata O. Advancements in long-read genome sequencing technologies and algorithms. Genomics. 2024 May 1;116(3):110842. doi:10.1016/j.ygeno.2024.110842

93. Marić J, Križanović K, Riondet S, Nagarajan N, Šikić M. Comparative analysis of metagenomic classifiers for long-read sequencing datasets. BMC Bioinformatics. 2024 Jan 11;25:15. doi:10.1186/s12859-024-05634-8 PubMed PMID: 38212694; PubMed Central PMCID: PMC10782538.

94. Breitwieser FP, Baker DN, Salzberg SL. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome Biology. 2018 Nov 16;19(1):198. doi:10.1186/s13059-018-1568-0

95. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biology. 2019 Nov 28;20(1):257. doi:10.1186/s13059-019-1891-0

96. Pightling AW, Pagotto F. Genome Sequence of Listeria monocytogenes Strain HPB5415, Collected during a 2008 Listeriosis Outbreak in Canada. Genome Announc. 2015 Jun 11;3(3). doi:10.1128/genomeA.00637-15 PubMed PMID: 26067972; PubMed Central PMCID: PMC4463536.

97. Pouillot* R, Hoelzer K, Chen Y, Dennis SB. Listeria monocytogenes Dose Response Revisited—Incorporating Adjustments for Variability in Strain Virulence and Host Susceptibility. Risk Analysis. 2015;35(1):90–108. doi:10.1111/risa.12235

98. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, et al. Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. Nat Methods. 2017 Nov;14(11):1063–71. doi:10.1038/nmeth.4458 PubMed PMID: 28967888; PubMed Central PMCID: PMC5903868.

99. Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, et al. Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol. 2017 Nov;35(11):1069–76. doi:10.1038/nbt.3960 PubMed PMID: 28967887.

100. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016 Jan 4;44(D1):D733-745. doi:10.1093/nar/gkv1189 PubMed PMID: 26553804; PubMed Central PMCID: PMC4702849.

101. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research. 2015 Jul;25(7):1043. doi:10.1101/gr.186072.114 PubMed PMID: 25977477.

102. Steinegger M, Salzberg SL. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. Genome Biology. 2020 May 12;21(1):115. doi:10.1186/s13059-020-02023-1

103. Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nat Methods. 2023 Aug;20(8):1203–12. doi:10.1038/s41592-023-01940-w PubMed PMID: 37500759.

104. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016 Oct 17. doi:10.1101/gr.210641.116 PubMed PMID: 27852649.

105. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. [cited 2019 Feb 25]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

106. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018 Sep 1;34(17):i884–90. doi:10.1093/bioinformatics/bty560

107. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357–9. doi:10.1038/nmeth.1923 PubMed PMID: 22388286; PubMed Central PMCID: PMC3322381.

108. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9. doi:10.1093/bioinformatics/btp352

109. Oxford Nanopore Technologies [Internet]. 2024 [cited 2026 Feb 21]. Guppy. Available from: https://nanoporetech.com/software/other/guppy

110. De Coster W, Rademakers R. NanoPack2: population-scale evaluation of long-read sequencing data. Bioinformatics. 2023 May 1;39(5):btad311. doi:10.1093/bioinformatics/btad311

111. Bonenfant Q, Noé L, Touzet H. Porechop_ABI: discovering unknown adapters in Oxford Nanopore Technology sequencing reads for downstream trimming. Bioinformatics Advances. 2023 Jan 1;3(1):vbac085. doi:10.1093/bioadv/vbac085

112. Wick RR, Judd LM, Holt KE. Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing. PLOS Computational Biology. 2023 Mar 2;19(3):e1010905. doi:10.1371/journal.pcbi.1010905

113. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094–100. doi:10.1093/bioinformatics/bty191 PubMed PMID: 29750242; PubMed Central PMCID: PMC6137996.

114. Breitwieser FP, Salzberg SL. Pavian: Interactive analysis of metagenomics data for microbiomics and pathogen identification. bioRxiv. 2016 Oct 31. doi:10.1101/084715

115. RefSeq Masher [Jupyter Notebook] [Internet]. National Microbiology Laboratory; 2022 [cited 2022 May 18]. Available from: https://github.com/phac-nml/refseq_masher

116. Huber PJ, Ronchetti EM. Robust Statistics. John Wiley & Sons; 2011. 322 p.

117. Rousseeuw PJ, Croux C. Alternatives to the Median Absolute Deviation. Journal of the American Statistical Association. 1993 Dec 1;88(424):1273–83. doi:10.1080/01621459.1993.10476408

118. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes De Novo Assembler. Current Protocols in Bioinformatics. 2020;70(1):e102. doi:10.1002/cpbi.102

119. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019 May;37(5):540–6. doi:10.1038/s41587-019-0072-8

120. nanoporetech/medaka [Python] [Internet]. Oxford Nanopore Technologies; 2026 [cited 2026 Feb 21]. Available from: https://github.com/nanoporetech/medaka

121. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013 Apr 15;29(8):1072–5. doi:10.1093/bioinformatics/btt086

122. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLOS Computational Biology. 2017 Jun 8;13(6):e1005595. doi:10.1371/journal.pcbi.1005595

123. Bessonov K, Laing C, Robertson J, Yong I, Ziebell K, Gannon VPJ, et al. ECTyper: in silico Escherichia coli serotype and species prediction from raw and assembled whole-genome sequence data.

Microb Genom. 2021 Dec;7(12):000728. doi:10.1099/mgen.0.000728 PubMed PMID: 34860150; PubMed Central PMCID: PMC8767331.

124. Rumore J, Tschetter L, Kearney A, Kandar R, McCormick R, Walker M, et al. Evaluation of whole-genome sequencing for outbreak detection of Verotoxigenic Escherichia coli O157:H7 from the Canadian perspective. BMC Genomics. 2018 Dec 4;19(1):870. doi:10.1186/s12864-018-5243-3 PubMed PMID: 30514209; PubMed Central PMCID: PMC6278084.

125. Petkau A, Mabon P, Sieffert C, Knox NC, Cabral J, Iskander M, et al. SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. Microb Genom. 2017 Jun 8;3(6):e000116. doi:10.1099/mgen.0.000116 PubMed PMID: 29026651; PubMed Central PMCID: PMC5628696.

126. Hallgren MB, Overballe-Petersen S, Lund O, Hasman H, Clausen PTLC. MINTyper: an outbreak-detection method for accurate and rapid SNP typing of clonal clusters with noisy long reads. Biol Methods Protoc. 2021 Apr 21;6(1):bpab008. doi:10.1093/biomethods/bpab008 PubMed PMID: 33981853; PubMed Central PMCID: PMC8106442.

127. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016 Jan 15;32(2):292–4. doi:10.1093/bioinformatics/btv566 PubMed PMID: 26428292; PubMed Central PMCID: PMC4708105.

128. McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biology. 2017 Sep 21;18(1):182. doi:10.1186/s13059-017-1299-7

129. Mishra P, Singh U, Pandey CM, Mishra P, Pandey G. Application of Student's t-test, Analysis of Variance, and Covariance. Ann Card Anaesth. 2019;22(4):407–11. doi:10.4103/aca.ACA_94_19 PubMed PMID: 31621677; PubMed Central PMCID: PMC6813708.

130. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2025. Available from: https://www.R-project.org/

131. Posit Team. RStudio: Integrated Development Environment for R [Internet]. Boston, MA: Posit Software, PBC; 2025. Available from: https://posit.co/

132. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org

133. Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, Blomberg LA, et al. Multi-Platform Next-Generation Sequencing of the Domestic Turkey (Meleagris gallopavo): Genome Assembly and Analysis. PLOS Biology. 2010 Sep 7;8(9):e1000475. doi:10.1371/journal.pbio.1000475

134. Xu C, Jiao C, Sun H, Cai X, Wang X, Ge C, et al. Draft genome of spinach and transcriptome diversity of 120 Spinacia accessions. Nat Commun. 2017 May 24;8:15275. doi:10.1038/ncomms15275 PubMed PMID: 28537264; PubMed Central PMCID: PMC5458060.

135. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec 15;10:421. doi:10.1186/1471-2105-10-421 PubMed PMID: 20003500; PubMed Central PMCID: PMC2803857.

136. Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2019 Jan 8;47(Database issue):D23–8. doi:10.1093/nar/gky1069 PubMed PMID: 30395293; PubMed Central PMCID: PMC6323993.

137. Jarvis KG, Daquigan N, White JR, Morin PM, Howard LM, Manetas JE, et al. Microbiomes Associated With Foods From Plant and Animal Sources. Front Microbiol. 2018;9. doi:10.3389/fmicb.2018.02540

138. Leonard SR, Mammel MK, Lacher DW, Elkins CA. Application of Metagenomic Sequencing to Food Safety: Detection of Shiga Toxin-Producing Escherichia coli on Fresh Bagged Spinach. Drake HL, editor. Appl Environ Microbiol. 2015 Dec 1;81(23):8183–91. doi:10.1128/AEM.02601-15

139. Robinson AJ, Daligault HE, Kelliher JM, LeBrun ES, Chain PSG. Multiple Cases of Bacterial Sequence Erroneously Incorporated Into Publicly Available Chloroplast Genomes. Front Genet. 2021;12:821715. doi:10.3389/fgene.2021.821715 PubMed PMID: 35096026; PubMed Central PMCID: PMC8793683.

140. GitHub [Internet]. [cited 2021 Jan 21]. DerrickWood/kraken2. Available from: https://github.com/DerrickWood/kraken2

141. Simner PJ, Miller HB, Breitwieser FP, Pinilla Monsalve G, Pardo CA, Salzberg SL, et al. Development and Optimization of Metagenomic Next-Generation Sequencing Methods for Cerebrospinal Fluid Diagnostics. J Clin Microbiol. 2018 Aug 27;56(9):e00472-18. doi:10.1128/JCM.00472-18 PubMed PMID: 29976594; PubMed Central PMCID: PMC6113476.

142. Méric G, Wick RR, Watts SC, Holt KE, Inouye M. Correcting index databases improves metagenomic studies. bioRxiv. 2019 Jul 23;712166. doi:10.1101/712166

143. Martí JM. Recentrifuge: Robust comparative analysis and contamination removal for metagenomics. PLOS Computational Biology. 2019 Apr 8;15(4):e1006967. doi:10.1371/journal.pcbi.1006967

144. Street TL, Barker L, Sanderson ND, Kavanagh J, Hoosdally S, Cole K, et al. Optimizing DNA Extraction Methods for Nanopore Sequencing of Neisseria gonorrhoeae Directly from Urine Samples. J Clin Microbiol. 2020 Feb 24;58(3):e01822-19. doi:10.1128/JCM.01822-19 PubMed PMID: 31852766; PubMed Central PMCID: PMC7041563.

145. Diao Z, Han D, Zhang R, Li J. Metagenomics next-generation sequencing tests take the stage in the diagnosis of lower respiratory tract infections. Journal of Advanced Research. 2021 Sep 29. doi:10.1016/j.jare.2021.09.012

146. Law JWF, Ab Mutalib NS, Chan KG, Lee LH. An insight into the isolation, enumeration, and molecular detection of Listeria monocytogenes in food. Frontiers in Microbiology [Internet]. 2015 [cited 2022 Mar 17];6. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2015.01227

147. Peabody MA, Van Rossum T, Lo R, Brinkman FSL. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. BMC Bioinformatics. 2015 Nov 4;16(1):362. doi:10.1186/s12859-015-0788-5

148. Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. Scientific Reports. 2016 Jan 18;6:19233. doi:10.1038/srep19233

149. Gardner PP, Watson RJ, Morgan XC, Draper JL, Finn RD, Morales SE, et al. Identifying accurate metagenome and amplicon software via a meta-analysis of sequence to taxonomy benchmarking studies. PeerJ. 2019 Jan 4;7:e6160. doi:10.7717/peerj.6160

150. Nasko DJ, Koren S, Phillippy AM, Treangen TJ. RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. Genome Biol. 2018 30;19(1):165. doi:10.1186/s13059-018-1554-6 PubMed PMID: 30373669; PubMed Central PMCID: PMC6206640.

151. Pockrandt C, Zimin AV, Salzberg SL. Metagenomic classification with KrakenUniq on low-memory computers. Journal of Open Source Software. 2022 Dec 28;7(80):4908. doi:10.21105/joss.04908

152. Interpretation criteria for WGS data – Outbreak Toolkit [Internet]. [cited 2026 Apr 30]. Available from: https://outbreaktools.ca/tools/interpretation-criteria-for-pfge-and-mlva/

153. Wang C, Zhang L, Kan C, He J, Liang W, Xia R, et al. Benefits and challenges of host depletion methods in profiling the upper and lower respiratory microbiome. NPJ Biofilms Microbiomes. 2025 Jul 7;11(1):130. doi:10.1038/s41522-025-00762-2 PubMed PMID: 40624015; PubMed Central PMCID: PMC12234985.

154. Duarte V da S, Porcellato D. Host DNA depletion methods and genome-centric metagenomics of bovine hindmilk microbiome. mSphere. 2024 Jan 30;9(1):e0047023. doi:10.1128/msphere.00470-23 PubMed PMID: 38054728; PubMed Central PMCID: PMC10826364.

155. Zhang X, Lin K, Li Y. Highlights to phytosterols accumulation and equilibrium in plants: Biosynthetic pathway and feedback regulation. Plant Physiology and Biochemistry. 2020 Oct 1;155:637–49. doi:10.1016/j.plaphy.2020.08.021

156. Cuggino SG, Possas A, Posada-Izquierdo GD, Theumer MG, Pérez-Rodríguez F. Unveiling Fresh-Cut Lettuce Processing in Argentine Industries: Evaluating Salmonella Levels Using Predictive Microbiology Models. Foods. 2023 Jan;12(21):3999. doi:10.3390/foods12213999

157. Nemati V. Advances in Lettuce postharvest processing: Implications for microbiological safety and storage quality. Journal of Agriculture and Food Research. 2025 Jun 1;21:101824. doi:10.1016/j.jafr.2025.101824

158. Maguire M, Ramachandran P, Tallent S, Mammel MK, Brown EW, Allard MW, et al. Precision metagenomics sequencing for food safety: hybrid assembly of Shiga toxin-producing Escherichia coli in enriched agricultural water. Front Microbiol. 2023;14:1221668. doi:10.3389/fmicb.2023.1221668 PubMed PMID: 37720160; PubMed Central PMCID: PMC10500926.

159. Marquet M, Zöllkau J, Pastuschek J, Viehweger A, Schleußner E, Makarewicz O, et al. Evaluation of microbiome enrichment and host DNA depletion in human vaginal samples using Oxford Nanopore's adaptive sequencing. Sci Rep. 2022 Mar 7;12(1):4000. doi:10.1038/s41598-022-08003-8 PubMed PMID: 35256725; PubMed Central PMCID: PMC8901746.

160. Lazarevic V, Gaïa N, Girard M, Mauffrey F, Ruppé E, Schrenzel J. Effect of bacterial DNA enrichment on detection and quantification of bacteria in an infected tissue model by metagenomic next-generation sequencing. ISME COMMUN. 2022 Dec 26;2(1):122. doi:10.1038/s43705-022-00208-2

161. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018;36(10):996–1004. doi:10.1038/nbt.4229 PubMed PMID: 30148503.

162. Parks DH, Chuvochina M, Reeves PR, Beatson SA, Hugenholtz P. Reclassification of Shigella species as later heterotypic synonyms of Escherichia coli in the Genome Taxonomy Database [Internet]. bioRxiv; 2021 [cited 2026 Mar 6]. p. 2021.09.22.461432. Available from: https://www.biorxiv.org/content/10.1101/2021.09.22.461432v1 doi:10.1101/2021.09.22.461432

163. Martí JM, Kok CR, Thissen JB, Mulakken NJ, Avila-Herrera A, Jaing CJ, et al. Addressing the dynamic nature of reference data: a new nucleotide database for robust metagenomic classification. mSystems. 2025 Apr 22;10(4):e0123924. doi:10.1128/msystems.01239-24 PubMed PMID: 40111052; PubMed Central PMCID: PMC12013259.

164. Kobus R, Abuín JM, Müller A, Hellmann SL, Pichel JC, Pena TF, et al. A big data approach to metagenomics for all-food-sequencing. BMC Bioinformatics. 2020 Mar 12;21(1):102. doi:10.1186/s12859-020-3429-6 PubMed PMID: 32164527; PubMed Central PMCID: PMC7069206.

165. Rumore J, Walker M, Pagotto F, Forbes JD, Peterson CL, Tyler AD, et al. Use of a taxon-specific reference database for accurate metagenomics-based pathogen detection of Listeria monocytogenes in turkey deli meat and spinach. BMC Genomics. 2023 Jun 27;24(1):361. doi:10.1186/s12864-023-09338-w PubMed PMID: 37370007; PubMed Central PMCID: PMC10303765.

166. Smith RH, Glendinning L, Walker AW, Watson M. Investigating the impact of database choice on the accuracy of metagenomic read classification for the rumen microbiome. Anim Microbiome. 2022 Nov 18;4(1):57. doi:10.1186/s42523-022-00207-7 PubMed PMID: 36401288; PubMed Central PMCID: PMC9673341.

167. Xu R, Rajeev S, Salvador LCM. The selection of software and database for metagenomics sequence analysis impacts the outcome of microbial profiling and pathogen detection. PLoS One.

2023;18(4):e0284031. doi:10.1371/journal.pone.0284031 PubMed PMID: 37027361; PubMed Central PMCID: PMC10081788.

168. Grundy BS, Parikh H, Jacob S, Banura P, Moore CC, Liu J, et al. Pathogen Detection Using Metagenomic Next-Generation Sequencing of Plasma Samples from Patients with Sepsis in Uganda. Microbiology Spectrum. 2023 Jan 10;11(1):e04312-22. doi:10.1128/spectrum.04312-22

169. Devanga Ragupathi NK, Muthuirulandi Sethuvel DP, Inbanathan FY, Veeraraghavan B. Accurate differentiation of *Escherichia coli* and *Shigella* serogroups: challenges and strategies. New Microbes and New Infections. 2018 Jan 1;21:58–62. doi:10.1016/j.nmni.2017.09.003

170. van den Beld MJC, Reubsaet FAG, Pijnacker R, Harpal A, Kuiling S, Heerkens EM, et al. A Multifactorial Approach for Surveillance of Shigella spp. and Entero-Invasive Escherichia coli Is Important for Detecting (Inter)national Clusters. Front Microbiol. 2020 Oct 19;11. doi:10.3389/fmicb.2020.564103

171. Kruasuwan W, Sawatwong P, Jenjaroenpun P, Wankaew N, Arigul T, Yongkiettrakul S, et al. Comparative evaluation of commercial DNA isolation approaches for nanopore-only bacterial genome assembly and plasmid recovery. Sci Rep. 2024 Nov 12;14(1):27672. doi:10.1038/s41598-024-78066-2

172. Zhang L, Chen T, Wang Y, Zhang S, Lv Q, Kong D, et al. Comparison Analysis of Different DNA Extraction Methods on Suitability for Long-Read Metagenomic Nanopore Sequencing. Front Cell Infect Microbiol. 2022;12:919903. doi:10.3389/fcimb.2022.919903 PubMed PMID: 35837476; PubMed Central PMCID: PMC9273838.

173. Wu-Woods NJ, Barlow JT, Trigodet F, Shaw DG, Romano AE, Jabri B, et al. Microbial-enrichment method enables high-throughput metagenomic characterization from host-rich samples. Nat Methods. 2023 Nov;20(11):1672–82. doi:10.1038/s41592-023-02025-4

174. Martin S, Heavens D, Lan Y, Horsfield S, Clark MD, Leggett RM. Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples. Genome Biol. 2022 Jan 24;23(1):11. doi:10.1186/s13059-021-02582-x PubMed PMID: 35067223; PubMed Central PMCID: PMC8785595.

175. Ong CT, Ross EM, Boe-Hansen GB, Turni C, Hayes BJ, Tabor AE. Technical note: overcoming host contamination in bovine vaginal metagenomic samples with nanopore adaptive sequencing. J Anim Sci. 2022 Jan 1;100(1):skab344. doi:10.1093/jas/skab344 PubMed PMID: 34791313; PubMed Central PMCID: PMC8722758.

176. Ulrich JU, Epping L, Pilz T, Walther B, Stingl K, Semmler T, et al. Nanopore adaptive sampling effectively enriches bacterial plasmids. mSystems. 2024 Mar 19;9(3):e0094523. doi:10.1128/msystems.00945-23 PubMed PMID: 38376263; PubMed Central PMCID: PMC10949517.

177. Lin Y, Dai Y, Zhang S, Guo H, Yang L, Li J, et al. Application of nanopore adaptive sequencing in pathogen detection of a patient with Chlamydia psittaci infection. Frontiers in Cellular and Infection

Microbiology [Internet]. 2023 [cited 2023 Apr 28];13. Available from: https://www.frontiersin.org/articles/10.3389/fcimb.2023.1064317

178. Chiu CY, Miller SA. Clinical metagenomics. Nature Reviews Genetics. 2019 Jun;20(6):6. doi:10.1038/s41576-019-0113-7

179. Wilson MR, Sample HA, Zorn KC, Arevalo S, Yu G, Neuhaus J, et al. Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. N Engl J Med. 2019 Jun 13;380(24):2327–40. doi:10.1056/NEJMoa1803396 PubMed PMID: 31189036; PubMed Central PMCID: PMC6764751.

180. Wang J, den Bakker HC, Denes TG. A critical review of metagenomic approaches for foodborne pathogen surveillance. Critical Reviews in Food Science and Nutrition. 2025 Dec 12;65(33):8488–501. doi:10.1080/10408398.2025.2503453 PubMed PMID: 40478725.

181. Buytaers FE, Verhaegen B, Van Nieuwenhuysen T, Roosens NHC, Vanneste K, Marchal K, et al. Strain-level characterization of foodborne pathogens without culture enrichment for outbreak investigation using shotgun metagenomics facilitated with nanopore adaptive sampling. Front Microbiol. 2024 Mar 1;15. doi:10.3389/fmicb.2024.1330814

182. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017 Sep;35(9):9. doi:10.1038/nbt.3935

183. Luan T, Commichaux S, Hoffmann M, Jayeola V, Jang JH, Pop M, et al. Benchmarking short and long read polishing tools for nanopore assemblies: achieving near-perfect genomes for outbreak isolates. BMC Genomics. 2024 Jul 8;25(1):679. doi:10.1186/s12864-024-10582-x

184. Govender KN, Eyre DW. Benchmarking taxonomic classifiers with Illumina and Nanopore sequence data for clinical metagenomic diagnostic applications. Microbial Genomics. 2022;8(10):000886. doi:10.1099/mgen.0.000886

185. Liu Y, Ghaffari MH, Ma T, Tu Y. Impact of database choice and confidence score on the performance of taxonomic classification using Kraken2. aBIOTECH. 2024 Dec 1;5(4):465–75. doi:10.1007/s42994-024-00178-0

186. Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods. 2022 Jul;19(7):823–6. doi:10.1038/s41592-022-01539-7

187. Ni Y, Liu X, Simeneh ZM, Yang M, Li R. Benchmarking of Nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing. Comput Struct Biotechnol J. 2023 Mar 24;21:2352–64. doi:10.1016/j.csbj.2023.03.038 PubMed PMID: 37025654; PubMed Central PMCID: PMC10070092.

188. Townsend A, Strawn LK, Chapman BJ, Dunn LL. A Systematic Review of Listeria Species and Listeria monocytogenes Prevalence, Persistence, and Diversity throughout the Fresh Produce Supply Chain. Foods. 2021 Jun;10(6):1427. doi:10.3390/foods10061427

189. Libuit KG, Doughty EL, Otieno JR, Ambrosio F, Kapsak CJ, Smith EA, et al. Accelerating bioinformatics implementation in public health. Microb Genom. 2023 Jul 10;9(7):mgen001051. doi:10.1099/mgen.0.001051 PubMed PMID: 37428142; PubMed Central PMCID: PMC10438813.

190. Treichel NS, Pauvert C, Séneca J, Pjevac P, Berry D, Penders J, et al. Benchmarking of shotgun sequencing depth reveals the potential and limitations of shallow metagenomics and strain-level analysis. Nat Microbiol. 2026 May;11(5):1233–44. doi:10.1038/s41564-026-02334-2

191. Oehler JB, Burns K, Warner J, Schmitz U. Long-Read Sequencing for the Rapid Response to Infectious Diseases Outbreaks. Curr Clin Micro Rpt. 2025 May 15;12(1):10. doi:10.1007/s40588-025-00247-y

192. Lohde M, Wagner GE, Dabernig-Heinz J, Viehweger A, Braun SD, Monecke S, et al. Accurate bacterial outbreak tracing with Oxford Nanopore sequencing and reduction of methylation-induced errors. Genome Research. 2024 Nov 20;34(11):2039–47. doi:10.1101/gr.278848.123 PubMed PMID: 39443152.

193. Pijnacker R, van den Beld MJC, Ullrich A, Ceyssens PJ, van Cauteren D, Jore S, et al. Implementing whole genome sequencing for foodborne pathogen surveillance: insights and recommendations based on expert experiences. Front Microbiol. 2025 Dec 18;16. doi:10.3389/fmicb.2025.1707621

194. Sharma S, Narahari HP, Raman K. Harnessing machine learning for metagenomic data analysis: trends and applications. mSystems. 2025 Oct 7;10(11):e01642-24. doi:10.1128/msystems.01642-24

195. Gao Y, Liu M. Application of machine learning based genome sequence analysis in pathogen identification. Front Microbiol. 2024 Oct 2;15. doi:10.3389/fmicb.2024.1474078

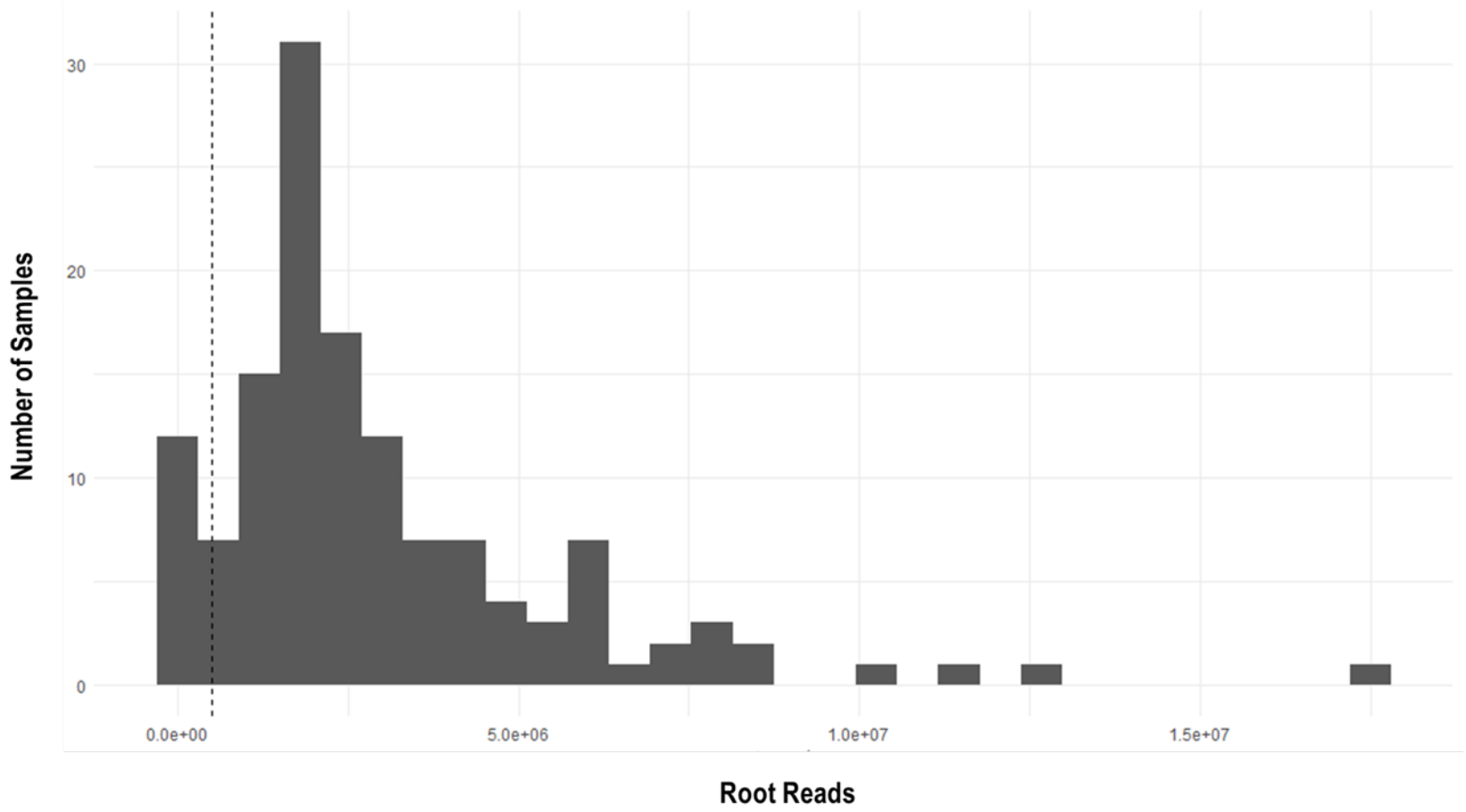
8.0 Appendices

Appendix 2.1. Reference database features.

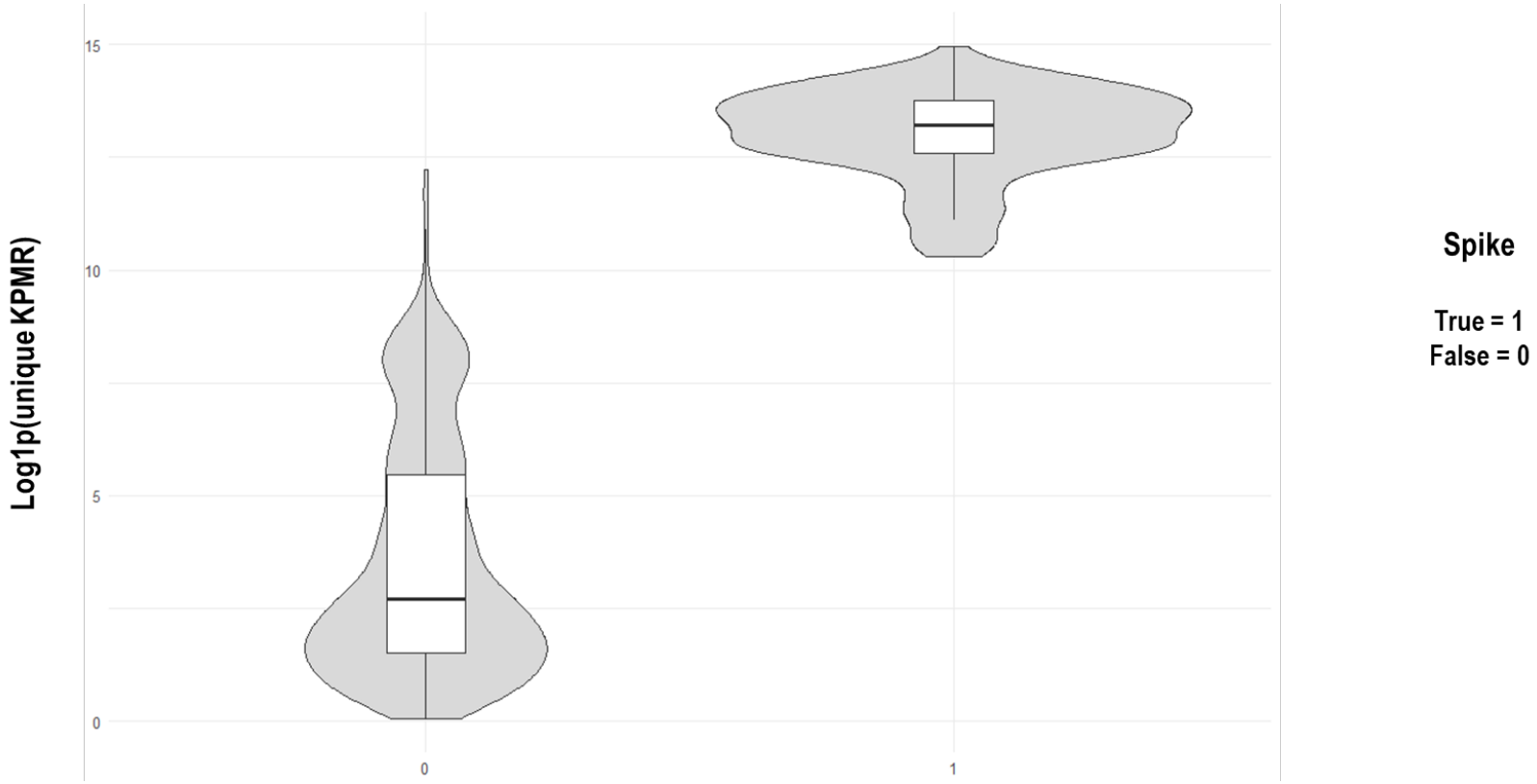
Classifier	Centrifuge		Kraken 2		KrakenUniq	
Features*/Database	**Universal**	**FBP**	**Universal**	**FBP**	**Universal**	**FBP**
Database Size	38.8 GB	9.8 GB	48.3 GB	6.2 GB	337 GB	52 GB
Build Time	22h28m	4h33m	2h19m	35m48s	14h58m	3h48m
Memory	278.29 GB	79.14 GB	46.03 GB	9.57 GB	307.40 GB	61.27 GB

*All values are approximate and were measured using 16 threads. Universal reference database contains 27,073 reference genomes + human genome (GRCh38.p13) + contaminant databases UniVec and EMVEC. Foodborne pathogen (FBP) specific reference database contains 3,804 reference genomes + human genome (GRCh38.p13) + contaminant databases UniVec and EMVEC.

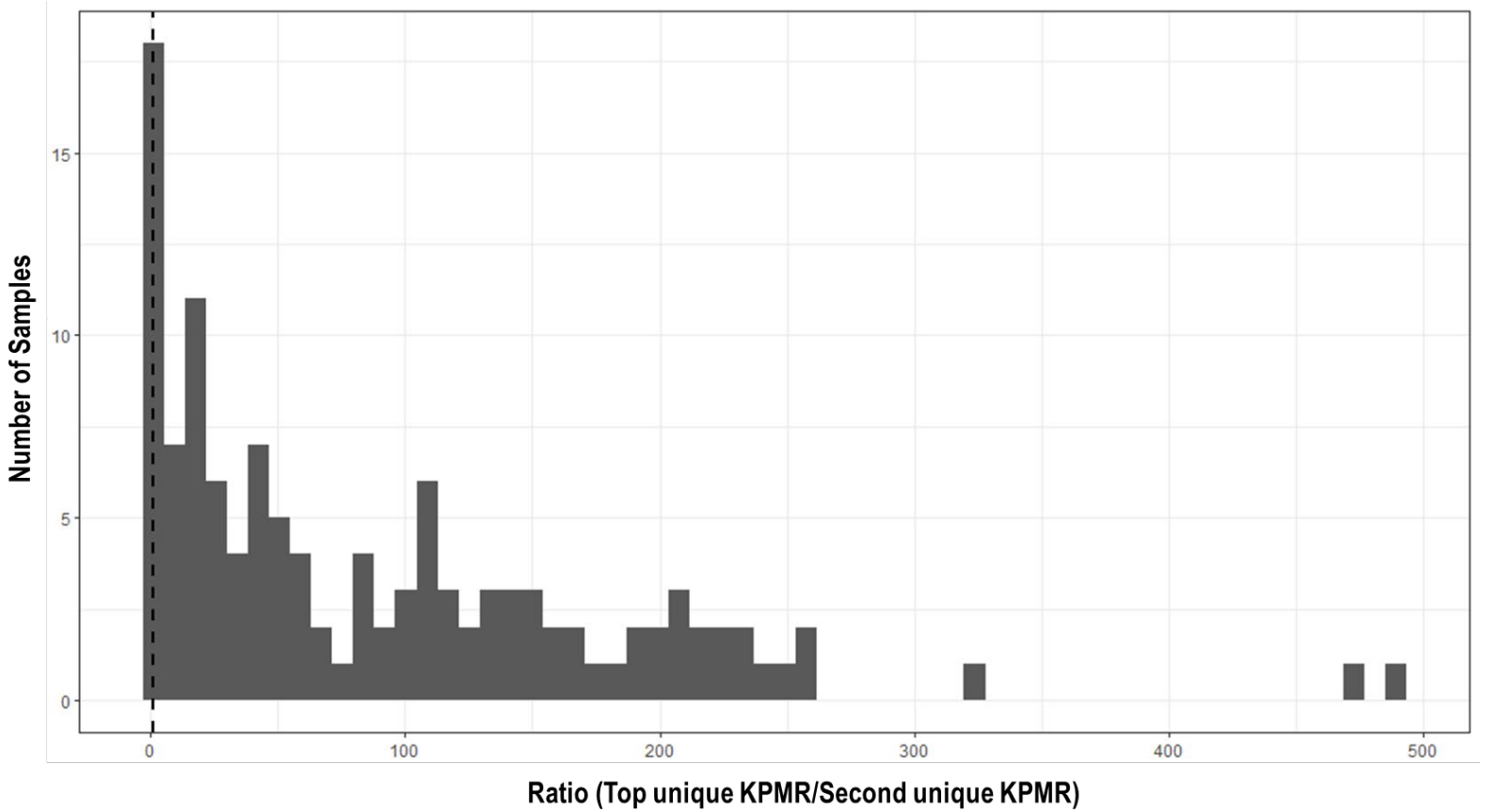
Appendix 2.2. Distribution of sequencing depth across samples. Histogram of total root-level read counts across all short-read quasi-metagenomic dataset samples (n=134). Most samples cluster within a moderate sequencing depth range, with a smaller number of high-depth outliers extending into the upper tail of the distribution. The vertical dashed line indicates the minimum sequencing depth threshold of 500,000 root reads applied for downstream analyses, illustrating that most samples exceed the required coverage.



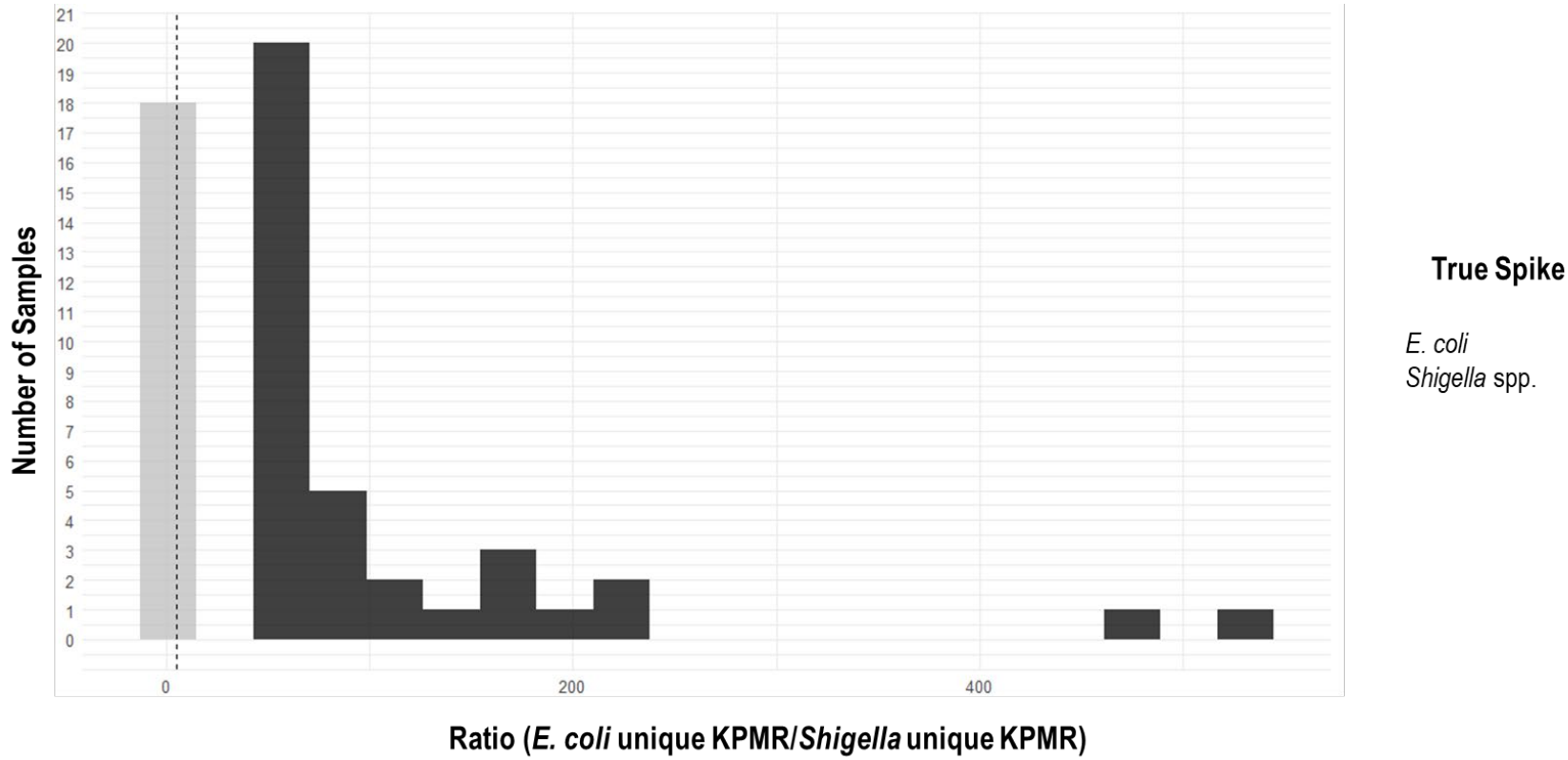
Appendix 2.3. Class separation using depth-normalized unique *k*-mers. Violin and boxplots show the distribution of log1p-transformed normalized unique k-mer counts (UKPMR) for negative samples (truth = 0) and pathogen-positive samples (truth = 1). Positive samples exhibit markedly higher normalized unique *k*-mer values with minimal overlap relative to negatives, indicating strong discrimination between pathogen signal and background noise. The separation demonstrates that depth normalization preserves biologically meaningful differences while reducing variability attributable to sequencing depth.



Appendix 2.4. Distribution of top-to-second unique KPMR dominance ratios across samples. Histogram depicts the ratio of the highest to second-highest depth-normalized unique *k*-mer support (KPMR) among taxa within each sample. Lower ratios indicate limited separation between competing taxa, whereas higher ratios reflect clear dominance of a single organism. The vertical dashed line denotes the selected ambiguity threshold (tie_ratio = 1.20); less than 1% of samples fall below this boundary, indicating that magnitude-based ranking is stable in most cases.



Appendix 2.5. Distribution of *E. coli* to *Shigella* dominance ratios stratified by true spike organism. Histogram showing the ratio of depth-normalized unique *k*-mers (KPMR) for *E. coli* relative to *Shigella* in samples where both taxa were detected. Samples are colored by the true spike organism. *Shigella*-spiked samples cluster at low ratio values and consistently fall below the dominance threshold (dashed vertical line), indicating that when *Shigella* unique KPMR reaches at least 20% of the competing *E. coli* value, *Shigella* is identified as the most likely pathogen. In contrast, *E. coli*–spiked samples show substantially higher ratios, reflecting strong dominance of *E. coli* signal. The clear separation between spike groups supports the use of a ratio-based override rule to resolve competition between closely related taxa while preserving specificity for true *E. coli*–dominant samples.



Appendix 4.1. Sequencing and taxonomic classification metrics for *Listeria* spp. short-read quasi-metagenomic datasets generated in Chapter 4. Performance is compared across two laboratory workflows: no saponin-based eukaryotic DNA depletion (NSD) and saponin depletion (SD). Samples were generated from four food matrices (ham, roast beef, salad kit, turkey) at three spike levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). Taxonomic results are reported for the foodborne pathogen (FBP)-specific database. Sample identifiers include *Listeria* species (LI = *L. innocua*; LM-1–8 = *L. monocytogenes*; LS = *L. seeligeri*; LW = *L. welshimeri*). Negative control (NC) samples were included to establish baseline taxonomic background. PE indicates paired-end sequencing. ER indicates eukaryotic reads. Single-isolate (SI) reference genomes were previously sequenced and were not evaluated using the preprocessing workflow developed in this study; these entries are indicated as NA in the table.

Strain	Food	Pathogen	Wet Lab	Spike Level	Classifier	Database	DNA Concentration	Total Reads R1 R2	Total Bases	Avg Base Quality	Raw Reads PE	fastp	Percent fastp	fastp HostRemoved	Eukaryotic Reads	Percent ER	Number of Reads Available for Classification (PE)	Classified reads	Percent classified	reads clade	taxReads	kmers	taxID	rank	taxName
LI-1	HAM	Listeria innocua	NSD	High	KrakenUniq	FBP-specific	51.966	4080956	612822778	27.665	2040478	2013943	1.30043059	1971180	42763	2.0957	1971180	1896918	93.41	1841329	316047	3E+06	1637	genus	Listeria
LI-1	HAM	Listeria innocua	SD	High	KrakenUniq	FBP-specific	68.181	3418160	513400150	25.08	1709080	1657126	3.03988111	1657033	93	0.0054	1657033	1597604	93.93	1556484	248231	3E+06	1637	genus	Listeria
LI-1	HAM	Listeria innocua	NSD	Low	KrakenUniq	FBP-specific	4.942	11975226	1.788E+09	27.72	5987613	5899639	1.46926663	5398531	501108	8.3691	5398531	5342909	97.41	5258496	650297	3E+06	1637	genus	Listeria
LI-1	HAM	Listeria innocua	SD	Low	KrakenUniq	FBP-specific	5.491	7788654	1.165E+09	25.52	3894327	3779585	2.94638843	3775153	4432	0.1138	3775153	3753022	98.56	3720935	371463	3E+06	1637	genus	Listeria
LI-1	HAM	Listeria innocua	NSD	Medium	KrakenUniq	FBP-specific	13.492	10496188	1.572E+09	27.77	5248094	5174331	1.40551979	4951659	222672	4.2429	4951659	4890819	96.83	4794592	644632	3E+06	1637	genus	Listeria
LI-1	HAM	Listeria innocua	SD	Medium	KrakenUniq	FBP-specific	21.695	3545464	532353301	24.915	1772732	1712883	3.37608843	1712080	803	0.0453	1712080	1647586	93.27	1596935	288515	3E+06	1637	genus	Listeria
LI-1	HAM	Listeria innocua	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	960406	NA	NA	NA	NA	NA	960406	951814	98.64	947346	125245	3E+06	1637	genus	Listeria
LM-1	HAM	Listeria monocytogenes	NSD	High	KrakenUniq	FBP-specific	36.906	3877230	582072498	27.5	1938615	1909323	1.51097562	1874271	35052	1.8081	1874271	1873378	98.76	1850969	287019	3E+06	1637	genus	Listeria
LM-1	HAM	Listeria monocytogenes	SD	High	KrakenUniq	FBP-specific	67.504	2862676	429975008	25.025	1431338	1387711	3.04798727	1387685	26	0.0018	1387685	1387211	99.19	1376378	156795	3E+06	1637	genus	Listeria
LM-1	HAM	Listeria monocytogenes	NSD	Low	KrakenUniq	FBP-specific	14.826	9316314	1.397E+09	27.835	4658157	4593252	1.39336223	4354801	238451	5.119	4354801	4349697	99.23	4321440	516531	3E+06	1637	genus	Listeria
LM-1	HAM	Listeria monocytogenes	SD	Low	KrakenUniq	FBP-specific	12.371	4544404	682070891	25.135	2272202	2200116	3.17251723	2198564	1552	0.0683	2198564	2197595	99.08	2178423	288800	3E+06	1637	genus	Listeria
LM-1	HAM	Listeria monocytogenes	NSD	Medium	KrakenUniq	FBP-specific	28.035	7296634	1.094E+09	27.93	3648317	3602063	1.26781746	3500650	101413	2.7797	3500650	3498496	99.23	3473609	373527	3E+06	1637	genus	Listeria
LM-1	HAM	Listeria monocytogenes	SD	Medium	KrakenUniq	FBP-specific	53.565	3580430	537516123	25.065	1790215	1733354	3.17621068	1733066	288	0.0161	1733066	1732393	99.35	1721792	174765	3E+06	1637	genus	Listeria
LM-1	HAM	Listeria monocytogenes	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	507636	NA	NA	NA	NA	NA	507636	507535	99.72	506221	17376	3E+06	1637	genus	Listeria
LM-2	HAM	Listeria monocytogenes	NSD	High	KrakenUniq	FBP-specific	31.596	4090974	614209671	27.595	2045487	2017140	1.38583135	1969838	47302	2.3125	1969838	1968760	99.92	1968213	66219	3E+06	1637	genus	Listeria
LM-2	HAM	Listeria monocytogenes	SD	High	KrakenUniq	FBP-specific	33.475	2344748	352276236	24.81	1172374	1132815	3.37426453	1132314	501	0.0427	1132314	1131853	99.93	1131548	37716	3E+06	1637	genus	Listeria
LM-2	HAM	Listeria monocytogenes	NSD	Low	KrakenUniq	FBP-specific	3.834	18044540	2.688E+09	28.06	9022270	8903079	1.32107552	8362044	541035	5.9967	8362044	8350810	99.84	8349048	316797	3E+06	1637	genus	Listeria
LM-2	HAM	Listeria monocytogenes	SD	Low	KrakenUniq	FBP-specific	9.382	3633448	545419793	24.975	1816724	1758940	3.18067026	1756943	1997	0.1099	1756943	1756053	99.92	1755534	61210	3E+06	1637	genus	Listeria
LM-2	HAM	Listeria monocytogenes	NSD	Medium	KrakenUniq	FBP-specific	10.119	13021214	1.941E+09	28.065	6510607	6427966	1.26932865	6237789	190177	2.921	6237789	6233232	99.91	6231966	234007	3E+06	1637	genus	Listeria
LM-2	HAM	Listeria monocytogenes	SD	Medium	KrakenUniq	FBP-specific	18.851	16018196	2.405E+09	25	8009098	7753432	3.19219468	7749978	3454	0.0431	7749978	7746521	99.93	7744742	278448	3E+06	1637	genus	Listeria
LM-2	HAM	Listeria monocytogenes	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	645653	NA	NA	NA	NA	NA	645653	643151	98.99	639141	19920	3E+06	1637	genus	Listeria
NC	HAM	Negative_control	NSD	NC	KrakenUniq	FBP-specific	0.922	26790160	3.921E+09	26.53	13395080	1.3E+07	3.77596849	243711	12645575	94.4046	243711	13092	0.8535	2080	185	300111	1637	genus	Listeria
NC	HAM	Negative_control	SD	NC	KrakenUniq	FBP-specific	0.332	1410856	205990023	25.635	705428	681731	3.35923723	190142	491589	69.6866	190142	75360	19.3	36701	2728	4E+06	1637	genus	Listeria
LM-3	ROAST BEEF	Listeria monocytogenes	NSD	High	KrakenUniq	FBP-specific	16.224	5010674	749930998	25.155	2505337	2431765	2.93661092	1802039	629726	25.1354	1802039	1783085	98.86	1781424	281576	3E+06	1637	genus	Listeria
LM-3	ROAST BEEF	Listeria monocytogenes	SD	High	KrakenUniq	FBP-specific	27.351	2274960	341228736	23.865	1137480	1083413	4.75322643	1081611	1802	0.1584	1081611	1079385	99.77	1079124	149039	3E+06	1637	genus	Listeria
LM-3	ROAST BEEF	Listeria monocytogenes	NSD	Low	KrakenUniq	FBP-specific	3.899	7651668	1.14E+09	24.74	3825834	3668377	4.11562551	1534386	2133991	55.7785	1534386	1462669	95	1457594	160902	3E+06	1637	genus	Listeria
LM-3	ROAST BEEF	Listeria monocytogenes	SD	Low	KrakenUniq	FBP-specific	1.821	16352420	2.431E+09	24.33	8176210	7833529	4.19119617	7800806	32723	0.4002	7800806	7684692	98.46	7680438	599639	3E+06	1637	genus	Listeria
LM-3	ROAST BEEF	Listeria monocytogenes	NSD	Medium	KrakenUniq	FBP-specific	10.117	6048802	903233812	24.58	3024401	2895902	4.24874215	1549263	1346639	44.5258	1549263	1515990	97.65	1512803	264281	3E+06	1637	genus	Listeria
LM-3	ROAST BEEF	Listeria monocytogenes	SD	Medium	KrakenUniq	FBP-specific	12.819	6883358	1.031E+09	24	3441679	3286091	4.52070051	3279232	6859	0.1993	3279232	3269215	99.67	3268348	386139	3E+06	1637	genus	Listeria
LM-3	ROAST BEEF	Listeria monocytogenes	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	533397	NA	NA	NA	NA	NA	533397	531263	99.41	530262	44145	3E+06	1637	genus	Listeria
LM-4	ROAST BEEF	Listeria monocytogenes	NSD	High	KrakenUniq	FBP-specific	15.552	6526034	976945889	25.11	3263017	3166859	2.94690466	2394022	772837	23.6847	2394022	2373166	99.04	2371128	356144	3E+06	1637	genus	Listeria
LM-4	ROAST BEEF	Listeria monocytogenes	SD	High	KrakenUniq	FBP-specific	18.773	10256282	1.539E+09	23.97	5128141	4907590	4.30079828	4900666	6924	0.135	4900666	4890140	99.76	4889089	693445	3E+06	1637	genus	Listeria
LM-4	ROAST BEEF	Listeria monocytogenes	NSD	Low	KrakenUniq	FBP-specific	4.053	8685514	1.294E+09	24.81	4342757	4167045	4.0460933	1681113	2485932	57.2432	1681113	1609847	95.41	1603956	187057	3E+06	1637	genus	Listeria
LM-4	ROAST BEEF	Listeria monocytogenes	SD	Low	KrakenUniq	FBP-specific	2.195	12500270	1.86E+09	24.14	6250135	5969913	4.48345516	5960552	9361	0.1498	5960552	5878679	98.57	5875304	471225	3E+06	1637	genus	Listeria
LM-4	ROAST BEEF	Listeria monocytogenes	NSD	Medium	KrakenUniq	FBP-specific	7.894	5966828	891333285	24.9	2983414	2877744	3.5419154	1727333	1150411	38.5602	1727333	1697082	98.09	1694358	229584	3E+06	1637	genus	Listeria
LM-4	ROAST BEEF	Listeria monocytogenes	SD	Medium	KrakenUniq	FBP-specific	7.055	15330704	2.296E+09	23.955	7665352	7336941	4.28435641	7328679	8262	0.1078	7328679	7307368	99.69	7305631	811858	3E+06	1637	genus	Listeria
LM-4	ROAST BEEF	Listeria monocytogenes	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	723704	NA	NA	NA	NA	NA	723704	720433	99.36	719095	61576	3E+06	1637	genus	Listeria
NC	ROAST BEEF	Negative_control	NSD	NC	KrakenUniq	FBP-specific	2.522	9465438	1.404E+09	24.125	4732719	4457009	5.82561525	117889	4339120	91.6834	117889	8903	0.3978	469	113	20629	1637	genus	Listeria
NC	ROAST BEEF	Negative_control	SD	NC	KrakenUniq	FBP-specific	0.36	6230854	896550062	24.31	3115427	2970240	4.66026005	1876039	1094201	35.122	1876039	92184	2.462	46192	12683	421287	1637	genus	Listeria

Appendix 4.1. Sequencing and taxonomic classification metrics for *Listeria* spp. short-read quasi-metagenomic datasets generated in Chapter 4. Performance is compared across two laboratory workflows: no saponin-based eukaryotic DNA depletion (NSD) and saponin depletion (SD). Samples were generated from four food matrices (ham, roast beef, salad kit, turkey) at three spike levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). Taxonomic results are reported for the foodborne pathogen (FBP)-specific database. Sample identifiers include *Listeria* species (LI = *L. innocua*; LM-1–8 = *L. monocytogenes*; LS = *L. seeligeri*; LW = *L. welshimeri*). Negative control (NC) samples were included to establish baseline taxonomic background. PE indicates paired-end sequencing. ER indicates eukaryotic reads. Single-isolate (SI) reference genomes were previously sequenced and were not evaluated using the preprocessing workflow developed in this study; these entries are indicated as NA in the table (Continued).

Strain	Food	Pathogen	Wet Lab	Spike Level	Classifier	Database	DNA Concentration	Total Reads R1 R2	Total Bases	Avg Base Quality	Raw Reads PE	fastp	Percent fastp	fastp HostRemoved	Eukaryotic Reads	Percent ER	Number of Reads Available for Classification (PE)	Classified reads	Percent classified	reads clade	taxReads	kmers	taxID	rank	taxName
LM-5	SALAD KIT	Listeria monocytogenes	NSD	High	KrakenUniq	FBP-specific	18.371	14987542	2.25E+09	25.1	7493771	7262929	3.08045175	7144538	118391	1.5799	7144538	7056145	98.72	7052777	272161	3E+06	1637	genus	Listeria
LM-5	SALAD KIT	Listeria monocytogenes	SD	High	KrakenUniq	FBP-specific	38.976	4701068	706532907	27.22	2350534	2317364	1.4111687	2313705	3659	0.1557	2313705	2269919	98.04	2268408	83764	3E+06	1637	genus	Listeria
LM-5	SALAD KIT	Listeria monocytogenes	NSD	Low	KrakenUniq	FBP-specific	11.59	1852520	278210192	24.65	926260	890902	3.81728672	860284	30618	3.3056	860284	658116	75.79	652007	26091	3E+06	1637	genus	Listeria
LM-5	SALAD KIT	Listeria monocytogenes	SD	Low	KrakenUniq	FBP-specific	13.806	6603730	992149086	26.84	3301865	3221162	2.44416413	3217670	3492	0.1058	3217670	2207646	67.6	2175124	84242	3E+06	1637	genus	Listeria
LM-5	SALAD KIT	Listeria monocytogenes	NSD	Medium	KrakenUniq	FBP-specific	32.574	5998192	901158704	24.89	2999096	2900828	3.27658735	2845011	55817	1.8611	2845011	2771126	97.31	2768498	101721	3E+06	1637	genus	Listeria
LM-5	SALAD KIT	Listeria monocytogenes	SD	Medium	KrakenUniq	FBP-specific	32.489	7159794	1.077E+09	27.245	3579897	3527195	1.47216526	3523024	4171	0.1165	3523024	3372761	95.59	3367741	121766	3E+06	1637	genus	Listeria
LM-5	SALAD KIT	Listeria monocytogenes	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	837433	NA	NA	NA	NA	NA	837433	837013	99.46	832907	31303	3E+06	1637	genus	Listeria
LM-6	SALAD KIT	Listeria monocytogenes	NSD	High	KrakenUniq	FBP-specific	20.645	4093238	614787841	24.985	2046619	1985229	2.99958126	1947237	37992	1.8563	1947237	1941290	99.67	1940862	73738	3E+06	1637	genus	Listeria
LM-6	SALAD KIT	Listeria monocytogenes	SD	High	KrakenUniq	FBP-specific	36.196	5577804	838231567	27.385	2788902	2753412	1.27254382	2750726	2686	0.0963	2750726	2734429	99.38	2733651	100959	3E+06	1637	genus	Listeria
LM-6	SALAD KIT	Listeria monocytogenes	NSD	Low	KrakenUniq	FBP-specific	0.395	12061846	1.783E+09	25.89	6030923	5854163	2.93089466	5678085	176078	2.9196	5678085	5018376	88.04	4999117	210609	3E+06	1637	genus	Listeria
LM-6	SALAD KIT	Listeria monocytogenes	SD	Low	KrakenUniq	FBP-specific	8.091	10253200	1.54E+09	27.11	5126600	5024868	1.98439512	5021224	3644	0.0711	5021224	4009782	79.2	3976992	151269	3E+06	1637	genus	Listeria
LM-6	SALAD KIT	Listeria monocytogenes	NSD	Medium	KrakenUniq	FBP-specific	15.89	2589694	388686286	24.805	1294847	1248701	3.56381874	1217531	31170	2.4072	1217531	1181136	96.9	1179804	44922	3E+06	1637	genus	Listeria
LM-6	SALAD KIT	Listeria monocytogenes	SD	Medium	KrakenUniq	FBP-specific	48.217	6047550	908967608	27.24	3023775	2980744	1.42308869	2977943	2801	0.0926	2977943	2846576	95.44	2842262	103285	3E+06	1637	genus	Listeria
LM-6	SALAD KIT	Listeria monocytogenes	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	887330	NA	NA	NA	NA	NA	887330	886772	99.39	881896	35555	3E+06	1637	genus	Listeria
LS-1	SALAD KIT	Listeria seeligeri	NSD	High	KrakenUniq	FBP-specific	10.902	2595262	389626134	24.78	1297631	1254425	3.32960603	1208588	45837	3.5324	1208588	1134534	93.82	1133851	32284	2E+06	1637	genus	Listeria
LS-1	SALAD KIT	Listeria seeligeri	SD	High	KrakenUniq	FBP-specific	26.703	5849504	879113760	27.325	2924752	2888234	1.2485845	2885552	2682	0.0917	2885552	2715569	94.05	2713949	77044	2E+06	1637	genus	Listeria
LS-1	SALAD KIT	Listeria seeligeri	NSD	Low	KrakenUniq	FBP-specific	7.917	6564326	985700622	24.865	3282163	3172587	3.33853011	3100698	71889	2.1903	3100698	2679242	86.12	2670184	76615	2E+06	1637	genus	Listeria
LS-1	SALAD KIT	Listeria seeligeri	SD	Low	KrakenUniq	FBP-specific	18.466	6688612	1.005E+09	27.275	3344306	3288363	1.67278353	3286226	2137	0.0639	3286226	2736223	82.84	2722420	78740	2E+06	1637	genus	Listeria
LS-1	SALAD KIT	Listeria seeligeri	NSD	Medium	KrakenUniq	FBP-specific	19.437	2127716	319568355	24.75	1063858	1026361	3.52462453	1001796	24565	2.309	1001796	911380	90.84	910022	24975	2E+06	1637	genus	Listeria
LS-1	SALAD KIT	Listeria seeligeri	SD	Medium	KrakenUniq	FBP-specific	44.952	7205688	1.083E+09	27.26	3602844	3551888	1.41432713	3550415	1473	0.0409	3550415	3261797	91.77	3258147	73401	2E+06	1637	genus	Listeria
LS-1	SALAD KIT	Listeria seeligeri	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	1710271	NA	NA	NA	NA	NA	1710271	1556041	90.85	1553704	82960	2E+06	1637	genus	Listeria
NC	SALAD KIT	Negative_control	NSD	NC	KrakenUniq	FBP-specific	1.892	23722996	3.554E+09	24.565	11861498	1.1E+07	5.02539393	10055945	1209466	10.1966	10055945	307177	0.005937	597	203	51695	1637	genus	Listeria
NC	SALAD KIT	Negative_control	SD	NC	KrakenUniq	FBP-specific	1.552	18563076	2.782E+09	26.475	9281538	8918567	3.9106773	8912277	6290	0.0678	8912277	284827	0.003793	338	139	29434	1637	genus	Listeria
LM-7	TURKEY	Listeria monocytogenes	NSD	High	KrakenUniq	FBP-specific	19.361	5488870	822691907	23.79	2744435	2612565	4.80499629	2484657	127908	4.6606	2484657	2482474	99.34	2468231	233358	3E+06	1637	genus	Listeria
LM-7	TURKEY	Listeria monocytogenes	SD	High	KrakenUniq	FBP-specific	42.197	2790760	418881395	25.485	1395380	1372144	1.66520948	1371906	238	0.0171	1371906	1371486	99.15	1360275	160868	3E+06	1637	genus	Listeria
LM-7	TURKEY	Listeria monocytogenes	NSD	Low	KrakenUniq	FBP-specific	7.988	7563752	1.127E+09	23.945	3781876	3609682	4.55313712	3238107	371575	9.8252	3238107	3232046	99.06	3207679	389064	3E+06	1637	genus	Listeria
LM-7	TURKEY	Listeria monocytogenes	SD	Low	KrakenUniq	FBP-specific	9.954	3616622	542466638	25.4	1808311	1773582	1.92052142	1770340	3242	0.1793	1770340	1769651	98.93	1751412	252246	3E+06	1637	genus	Listeria
LM-7	TURKEY	Listeria monocytogenes	NSD	Medium	KrakenUniq	FBP-specific	16.833	3755130	562559669	23.79	1877565	1788119	4.76393627	1638701	149418	7.9581	1638701	1634476	98.72	1617764	227285	3E+06	1637	genus	Listeria
LM-7	TURKEY	Listeria monocytogenes	SD	Medium	KrakenUniq	FBP-specific	23.805	3604948	541048755	25.65	1802474	1774299	1.56312934	1774120	179	0.0099	1774120	1773609	98.91	1754792	250577	3E+06	1637	genus	Listeria
LM-7	TURKEY	Listeria monocytogenes	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	1439872	NA	NA	NA	NA	NA	1439872	1432062	99.08	1426609	128715	3E+06	1637	genus	Listeria
LM-8	TURKEY	Listeria monocytogenes	NSD	High	KrakenUniq	FBP-specific	18.765	8310900	1.247E+09	23.97	4155450	3967568	4.52133945	3713138	254430	6.1228	3713138	3704007	98.57	3659870	552053	3E+06	1637	genus	Listeria
LM-8	TURKEY	Listeria monocytogenes	SD	High	KrakenUniq	FBP-specific	31.831	3222556	483741157	25.705	1611278	1586724	1.52388353	1586617	107	0.0066	1586617	1586121	98.96	1570148	216794	3E+06	1637	genus	Listeria
LM-8	TURKEY	Listeria monocytogenes	NSD	Low	KrakenUniq	FBP-specific	7.156	5553364	830785271	24.015	2776682	2654024	4.41743059	2264331	389693	14.0345	2264331	2258563	99.13	2244704	241755	3E+06	1637	genus	Listeria
LM-8	TURKEY	Listeria monocytogenes	SD	Low	KrakenUniq	FBP-specific	9.626	4688542	702917427	25.63	2344271	2306720	1.60181993	2306287	433	0.0185	2306287	2305026	99.09	2285357	304435	3E+06	1637	genus	Listeria
LM-8	TURKEY	Listeria monocytogenes	NSD	Medium	KrakenUniq	FBP-specific	17.126	2959422	443754557	23.965	1479711	1414789	4.38747837	1303043	111746	7.5519	1303043	1299869	98.63	1285246	193077	3E+06	1637	genus	Listeria
LM-8	TURKEY	Listeria monocytogenes	SD	Medium	KrakenUniq	FBP-specific	28.44	8303588	1.247E+09	25.47	4151794	4080190	1.72465204	4080032	158	0.0038	4080032	4078722	98.83	4032494	614067	3E+06	1637	genus	Listeria
LM-8	TURKEY	Listeria monocytogenes	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	1512364	NA	NA	NA	NA	NA	1512364	1504389	99.2	1500328	116430	3E+06	1637	genus	Listeria
LW-1	TURKEY	Listeria welshimeri	NSD	High	KrakenUniq	FBP-specific	17.807	3938408	589847266	23.99	1969204	1885958	4.2273934	1739435	146523	7.4407	1739435	1696051	97.49	1695826	175285	2E+06	1637	genus	Listeria
LW-1	TURKEY	Listeria welshimeri	SD	High	KrakenUniq	FBP-specific	37.556	4005030	601103043	25.72	2002515	1971447	1.55144905	1971376	71	0.0035	1971376	1923157	97.55	1922993	179544	2E+06	1637	genus	Listeria
LW-1	TURKEY	Listeria welshimeri	NSD	Low	KrakenUniq	FBP-specific	10.016	4113496	615672711	23.96	2056748	1963171	4.54975525	1687604	275567	13.3982	1687604	1643259	97.35	1642961	171368	2E+06	1637	genus	Listeria
LW-1	TURKEY	Listeria welshimeri	SD	Low	KrakenUniq	FBP-specific	9.562	5964526	894188373	25.695	2982263	2935534	1.56689735	2934636	898	0.0301	2934636	2872468	97.87	2872151	229887	2E+06	1637	genus	Listeria
LW-1	TURKEY	Listeria welshimeri	NSD	Medium	KrakenUniq	FBP-specific	12.763	5354084	802385175	23.685	2677042	2535734	5.27851263	2295582	240152	8.9708	2295582	2237039	97.43	2236691	236689	2E+06	1637	genus	Listeria
LW-1	TURKEY	Listeria welshimeri	SD	Medium	KrakenUniq	FBP-specific	19.733	23282244	3.491E+09	25.885	11641122	1.1E+07	1.47936771	11468202	705	0.0061	11468202	11230937	97.92	11229817	839212	3E+06	1637	genus	Listeria
LW-1	TURKEY	Listeria welshimeri	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	885911	NA	NA	NA	NA	NA	885911	862212	97.25	861563	37551	2E+06	1637	genus	Listeria
NC	TURKEY	Negative_control	NSD	NC	KrakenUniq	FBP-specific	1.068	16515342	2.419E+09	23.905	8257671	7829216	5.18856951	39533	7789683	94.3327	39533	4571	0.9106	360	49	46184	1637	genus	Listeria
NC	TURKEY	Negative_control	SD	NC	KrakenUniq	FBP-specific	0.326	1436562	208472466	26.195	718281	701204	2.37748179	273723	427481	59.5145	273723	40490	5.695	15588	1485	2E+06	1637	genus	Listeria

Appendix 4.2. Sequencing and taxonomic classification metrics for VTEC short-read quasi-metagenomic datasets generated in Chapter 4. Performance is compared across two laboratory workflows: no saponin-based eukaryotic DNA depletion (NSD) and saponin depletion (SD). Samples were generated from two food matrices (ground beef and lettuce) at three spike levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). Taxonomic results are reported for two reference databases: the foodborne pathogen (FBP)-specific database and *Escherichia coli* (EC)-specific database. Sample identifiers include VTEC serotypes (EC-1, 2, 5, 6 = O157:H7; EC-3 = O26:H11; EC-4 = O103:H25). Negative control (NC) samples are included to establish baseline taxonomic background. PE indicates paired-end sequencing.

Strain	Food	Pathogen	Wet Lab	Spike Level	Classifier	Database	DNA Concentration	Total Reads R1 R2	Total Bases	Avg Base Quality	Raw Reads PE	fastp	Percent fastp	fastp HostRemoved	Eukaryotic Reads	Percent ER	Number of Reads Available for Classification (PE)	Classified reads	Percent classified	reads clade	taxReads	kmers	taxID	rank	taxName
EC-1	GROUND BEEF	Escherichia coli O157:H7	NSD	High	KrakenUniq	FBP-specific	25.644	4093274	614792481	26.825	2046637	1999689	2.29390947	1992613	7076	0.3457	1992613	1992227	59.28	1181291	0	1E+06	561	genus	Escherichia
EC-1	GROUND BEEF	Escherichia coli O157:H7	NSD	High	KrakenUniq	EC-specific	25.644	4093274	614792481	26.825	2046637	1999689	2.29390947	1992613	7076	0.3457	1992613	1992191	96.65	1925880	0	6E+06	561	genus	Escherichia
EC-1	GROUND BEEF	Escherichia coli O157:H7	SD	High	KrakenUniq	FBP-specific	46.186	2487350	372074600	27.325	1243675	1219439	1.94874063	1219251	188	0.0151	1219251	1219080	58.54	713772	0	1E+06	561	genus	Escherichia
EC-1	GROUND BEEF	Escherichia coli O157:H7	SD	High	KrakenUniq	EC-specific	46.186	2487350	372074600	27.325	1243675	1219439	1.94874063	1219251	188	0.0151	1219251	1219071	97.11	1184051	0	5E+06	561	genus	Escherichia
EC-1	GROUND BEEF	Escherichia coli O157:H7	NSD	Low	KrakenUniq	FBP-specific	30.019	4108958	617230874	26.795	2054479	2008802	2.22328873	2003289	5513	0.2683	2003289	2002886	58.74	1176723	0	1E+06	561	genus	Escherichia
EC-1	GROUND BEEF	Escherichia coli O157:H7	NSD	Low	KrakenUniq	EC-specific	30.019	4108958	617230874	26.795	2054479	2008802	2.22328873	2003289	5513	0.2683	2003289	2002845	96.56	1934312	0	6E+06	561	genus	Escherichia
EC-1	GROUND BEEF	Escherichia coli O157:H7	SD	Low	KrakenUniq	FBP-specific	59.833	1105410	165827590	27.245	552705	541574	2.01391339	541519	55	0.01	541519	541463	58.5	316791	0	916824	561	genus	Escherichia
EC-1	GROUND BEEF	Escherichia coli O157:H7	SD	Low	KrakenUniq	EC-specific	59.833	1105410	165827590	27.245	552705	541574	2.01391339	541519	55	0.01	541519	541455	96.84	524407	0	5E+06	561	genus	Escherichia
EC-1	GROUND BEEF	Escherichia coli O157:H7	NSD	Medium	KrakenUniq	FBP-specific	26.425	3465458	520419422	26.92	1732729	1693865	2.24293585	1689561	4304	0.2484	1689561	1689275	58.82	993721	0	1E+06	561	genus	Escherichia
EC-1	GROUND BEEF	Escherichia coli O157:H7	NSD	Medium	KrakenUniq	EC-specific	26.425	3465458	520419422	26.92	1732729	1693865	2.24293585	1689561	4304	0.2484	1689561	1689249	96.49	1630252	0	5E+06	561	genus	Escherichia
EC-1	GROUND BEEF	Escherichia coli O157:H7	SD	Medium	KrakenUniq	FBP-specific	54.102	1138412	170581827	27.28	569206	557657	2.02896667	557577	80	0.0141	557577	557501	58.48	326074	0	919968	561	genus	Escherichia
EC-1	GROUND BEEF	Escherichia coli O157:H7	SD	Medium	KrakenUniq	EC-specific	54.102	1138412	170581827	27.28	569206	557657	2.02896667	557577	80	0.0141	557577	557496	96.95	540561	0	5E+06	561	genus	Escherichia
EC-1	GROUND BEEF	Escherichia coli O157:H7	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	929401	NA	NA	NA	NA	NA	929401	928954	55.54	516152	0	2E+06	561	genus	Escherichia
EC-1	GROUND BEEF	Escherichia coli O157:H7	SI	SI	KrakenUniq	EC-specific	NA	NA	NA	NA	929401	NA	NA	NA	NA	NA	929,401	928,947	97.19	903256	0	6E+06	561	genus	Escherichia
EC-2	GROUND BEEF	Escherichia coli O157:H7	NSD	High	KrakenUniq	FBP-specific	17.456	5561498	834254966	26.92	2780749	2720343	2.17229243	2709322	11021	0.3963	2709322	2708644	58.43	1583048	0	1E+06	561	genus	Escherichia
EC-2	GROUND BEEF	Escherichia coli O157:H7	NSD	High	KrakenUniq	EC-specific	17.456	5561498	834254966	26.92	2780749	2720343	2.17229243	2709322	11021	0.3963	2709322	2708579	96.99	2627832	0	6E+06	561	genus	Escherichia
EC-2	GROUND BEEF	Escherichia coli O157:H7	SD	High	KrakenUniq	FBP-specific	32.391	12610408	1.883E+09	27.665	6305204	6201495	1.64481593	6189401	12094	0.1918	6189401	6188061	56.5	3496771	0	1E+06	561	genus	Escherichia
EC-2	GROUND BEEF	Escherichia coli O157:H7	SD	High	KrakenUniq	EC-specific	32.391	12610408	1.883E+09	27.665	6305204	6201495	1.64481593	6189401	12094	0.1918	6189401	6187989	97.51	6035555	0	6E+06	561	genus	Escherichia
EC-2	GROUND BEEF	Escherichia coli O157:H7	NSD	Low	KrakenUniq	FBP-specific	22.836	5239754	787110140	26.61	2619877	2553148	2.54702797	2545274	7874	0.3005	2545274	2544653	59.11	1504411	0	1E+06	561	genus	Escherichia
EC-2	GROUND BEEF	Escherichia coli O157:H7	NSD	Low	KrakenUniq	EC-specific	22.836	5239754	787110140	26.61	2619877	2553148	2.54702797	2545274	7874	0.3005	2545274	2544596	96.67	2460578	0	6E+06	561	genus	Escherichia
EC-2	GROUND BEEF	Escherichia coli O157:H7	SD	Low	KrakenUniq	FBP-specific	53.725	3019834	450531020	27.11	1509917	1477699	2.13375967	1477330	369	0.0244	1477330	1477092	57.19	844882	0	1E+06	561	genus	Escherichia
EC-2	GROUND BEEF	Escherichia coli O157:H7	SD	Low	KrakenUniq	EC-specific	53.725	3019834	450531020	27.11	1509917	1477699	2.13375967	1477330	369	0.0244	1477330	1477078	96.6	1427064	0	5E+06	561	genus	Escherichia
EC-2	GROUND BEEF	Escherichia coli O157:H7	NSD	Medium	KrakenUniq	FBP-specific	20.471	3421562	513452435	27	1710781	1674559	2.1172786	1668059	6500	0.3799	1668059	1667667	58.63	977947	0	1E+06	561	genus	Escherichia
EC-2	GROUND BEEF	Escherichia coli O157:H7	NSD	Medium	KrakenUniq	EC-specific	20.471	3421562	513452435	27	1710781	1674559	2.1172786	1668059	6500	0.3799	1668059	1667602	96.88	1616023	0	5E+06	561	genus	Escherichia
EC-2	GROUND BEEF	Escherichia coli O157:H7	SD	Medium	KrakenUniq	FBP-specific	36.284	2874950	429366512	27.39	1437475	1410739	1.859928	1410227	512	0.0356	1410227	1409943	58.14	819960	0	1E+06	561	genus	Escherichia
EC-2	GROUND BEEF	Escherichia coli O157:H7	SD	Medium	KrakenUniq	EC-specific	36.284	2874950	429366512	27.39	1437475	1410739	1.859928	1410227	512	0.0356	1410227	1409925	97.14	1369853	0	5E+06	561	genus	Escherichia
EC-2	GROUND BEEF	Escherichia coli O157:H7	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	880997	NA	NA	NA	NA	NA	880997	879940	53.06	467465	0	1E+06	561	genus	Escherichia
EC-2	GROUND BEEF	Escherichia coli O157:H7	SI	SI	KrakenUniq	EC-specific	NA	NA	NA	NA	880997	NA	NA	NA	NA	NA	880997	879921	97.3	857239	0	6E+06	561	genus	Escherichia
EC-3	GROUND BEEF	Escherichia O26:H11	NSD	High	KrakenUniq	FBP-specific	23.843	3169826	475485259	26.955	1584913	1553298	1.99474672	1548777	4521	0.2853	1548777	1548473	32.17	498195	0	714894	561	genus	Escherichia
EC-3	GROUND BEEF	Escherichia O26:H11	NSD	High	KrakenUniq	EC-specific	23.843	3169826	475485259	26.955	1584913	1553298	1.99474672	1548777	4521	0.2853	1548777	1548381	97.63	1512052	0	6E+06	561	genus	Escherichia
EC-3	GROUND BEEF	Escherichia O26:H11	SD	High	KrakenUniq	FBP-specific	44.129	1519974	228282039	27.12	759987	745549	1.89976934	745416	133	0.0175	745416	745352	33.17	247244	0	568003	561	genus	Escherichia
EC-3	GROUND BEEF	Escherichia O26:H11	SD	High	KrakenUniq	EC-specific	44.129	1519974	228282039	27.12	759987	745549	1.89976934	745416	133	0.0175	745416	745320	97.86	729488	0	5E+06	561	genus	Escherichia
EC-3	GROUND BEEF	Escherichia O26:H11	NSD	Low	KrakenUniq	FBP-specific	13.076	6674774	1.001E+09	26.865	3337387	3266109	2.13574272	3255882	10227	0.3064	3255882	3254210	29.14	948708	0	952769	561	genus	Escherichia
EC-3	GROUND BEEF	Escherichia O26:H11	NSD	Low	KrakenUniq	EC-specific	13.076	6674774	1.001E+09	26.865	3337387	3266109	2.13574272	3255882	10227	0.3064	3255882	3253908	97.47	3173490	0	6E+06	561	genus	Escherichia
EC-3	GROUND BEEF	Escherichia O26:H11	SD	Low	KrakenUniq	FBP-specific	25.293	1566036	234355764	27.37	783018	769807	1.68718982	767690	2117	0.2704	767690	767200	30.8	236478	0	581780	561	genus	Escherichia
EC-3	GROUND BEEF	Escherichia O26:H11	SD	Low	KrakenUniq	EC-specific	25.293	1566036	234355764	27.37	783018	769807	1.68718982	767690	2117	0.2704	767690	767108	97.56	748936	0	5E+06	561	genus	Escherichia
EC-3	GROUND BEEF	Escherichia O26:H11	NSD	Medium	KrakenUniq	FBP-specific	12.147	5714026	854271765	26.995	2857013	2795576	2.15039274	2784135	11441	0.4005	2784135	2764540	29.52	821854	0	936346	561	genus	Escherichia
EC-3	GROUND BEEF	Escherichia O26:H11	NSD	Medium	KrakenUniq	EC-specific	12.147	5714026	854271765	26.995	2857013	2795576	2.15039274	2784135	11441	0.4005	2784135	2763073	96.76	2694052	0	6E+06	561	genus	Escherichia
EC-3	GROUND BEEF	Escherichia O26:H11	SD	Medium	KrakenUniq	FBP-specific	24.311	3169606	473984912	27.3	1584803	1555497	1.84918883	1551250	4247	0.268	1551250	1539491	30.12	467261	0	696321	561	genus	Escherichia
EC-3	GROUND BEEF	Escherichia O26:H11	SD	Medium	KrakenUniq	EC-specific	24.311	3169606	473984912	27.3	1584803	1555497	1.84918883	1551250	4247	0.268	1551250	1538616	96.81	1501816	0	6E+06	561	genus	Escherichia
EC-3	GROUND BEEF	Escherichia O26:H11	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	1100888	NA	NA	NA	NA	NA	1100888	1100663	38.11	419585	0	2E+06	561	genus	Escherichia
EC-3	GROUND BEEF	Escherichia O26:H11	SI	SI	KrakenUniq	EC-specific	NA	NA	NA	NA	1100888	NA	NA	NA	NA	NA	1100888	1100060	97.7	1075605	0	7E+06	561	genus	Escherichia
NC	GROUND BEEF	Negative_control	NSD	NC	KrakenUniq	FBP-specific	26.312	8756098	1.313E+09	25.475	4378049	4181872	4.4809229	1532453	2649419	60.516	1532453	293779	4.318	66173	0	186059	561	genus	Escherichia
NC	GROUND BEEF	Negative_control	NSD	NC	KrakenUniq	EC-specific	26.312	8756098	1.313E+09	25.475	4378049	4181872	4.4809229	1532453	2649419	60.516	1532453	198219	11.26	172508	0	352182	561	genus	Escherichia
NC	GROUND BEEF	Negative_control	SD	NC	KrakenUniq	FBP-specific	15.785	4543722	675432765	27.57	2271861	2235489	1.60097823	2209715	25774	1.1345	2209715	341722	4.082	90210	0	177845	561	genus	Escherichia
NC	GROUND BEEF	Negative_control	SD	NC	KrakenUniq	EC-specific	15.785	4543722	675432765	27.57	2271861	2235489	1.60097823	2209715	25774	1.1345	2209715	225942	8.697	192168	0	338492	561	genus	Escherichia

Appendix 4.2. Sequencing and taxonomic classification metrics for VTEC short-read quasi-metagenomic datasets generated in Chapter 4. Performance is compared across two laboratory workflows: no saponin-based eukaryotic DNA depletion (NSD) and saponin depletion (SD). Samples were generated from two food matrices (ground beef and lettuce) at three spike levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). Taxonomic results are reported for two reference databases: the foodborne pathogen (FBP)-specific database and *Escherichia coli* (EC)-specific database. Sample identifiers include *E. coli* serotypes (EC-1, 2, 5, 6 = O157:H7; EC-3 = O26:H11; EC-4 = O103:H25). Negative control (NC) samples were included to establish baseline taxonomic background. PE indicates paired-end sequencing. ER indicates eukaryotic reads. Single-isolate (SI) reference genomes were previously sequenced and were not evaluated using the preprocessing workflow developed in this study; these entries are indicated as NA in the table (Continued).

Strain	Food	Pathogen	Wet Lab	Spike Level	Classifier	Database	DNA Concentration	Total Reads R1 R2	Total Bases	Avg Base Quality	Raw Reads PE	fastp	Percent fastp	fastp HostRemoved	Eukaryotic Reads	Percent ER	Number of Reads Available for Classification (PE)	Classified reads	Percent classified	reads clade	taxReads	kmers	taxID	rank	taxName
EC-4	LETTUCE	Escherichia coli O103:H25	NSD	High	KrakenUniq	FBP-specific	5.417	4286356	639561769	25.32	2143178	2092594	2.36023326	2092484	110	0.0051	2092484	2092003	29.59	619160	0	838929	561	genus	Escherichia
EC-4	LETTUCE	Escherichia coli O103:H25	NSD	High	KrakenUniq	EC-specific	5.417	4286356	639561769	25.32	2143178	2092594	2.36023326	2092484	110	0.0051	2092484	2079429	96.97	2029015	0	6E+06	561	genus	Escherichia
EC-4	LETTUCE	Escherichia coli O103:H25	SD	High	KrakenUniq	FBP-specific	13.799	2986958	447183414	23.315	1493479	1390352	6.90515233	1390258	94	0.0063	1390258	1390088	30.92	429919	0	749542	561	genus	Escherichia
EC-4	LETTUCE	Escherichia coli O103:H25	SD	High	KrakenUniq	EC-specific	13.799	2986958	447183414	23.315	1493479	1390352	6.90515233	1390258	94	0.0063	1390258	1389360	97.53	1355983	0	6E+06	561	genus	Escherichia
EC-4	LETTUCE	Escherichia coli O103:H25	NSD	Low	KrakenUniq	FBP-specific	8.266	2116136	316453721	25.23	1058068	1032396	2.42630908	1032325	71	0.0067	1032325	1032115	30.45	314295	0	629870	561	genus	Escherichia
EC-4	LETTUCE	Escherichia coli O103:H25	NSD	Low	KrakenUniq	EC-specific	8.266	2116136	316453721	25.23	1058068	1032396	2.42630908	1032325	71	0.0067	1032325	1030597	97.39	1005389	0	5E+06	561	genus	Escherichia
EC-4	LETTUCE	Escherichia coli O103:H25	SD	Low	KrakenUniq	FBP-specific	13.299	4819526	719565663	23.325	2409763	2248528	6.69090695	2248404	124	0.0051	2248404	2248109	31.03	697658	0	918633	561	genus	Escherichia
EC-4	LETTUCE	Escherichia coli O103:H25	SD	Low	KrakenUniq	EC-specific	13.299	4819526	719565663	23.325	2409763	2248528	6.69090695	2248404	124	0.0051	2248404	2246961	97.54	2193193	0	6E+06	561	genus	Escherichia
EC-4	LETTUCE	Escherichia coli O103:H25	NSD	Medium	KrakenUniq	FBP-specific	6.483	2528386	376665299	25.245	1264193	1233239	2.44851854	1233161	78	0.0062	1233161	1232887	30.11	371351	0	669180	561	genus	Escherichia
EC-4	LETTUCE	Escherichia coli O103:H25	NSD	Medium	KrakenUniq	EC-specific	6.483	2528386	376665299	25.245	1264193	1233239	2.44851854	1233161	78	0.0062	1233161	1231691	97.42	1201395	0	5E+06	561	genus	Escherichia
EC-4	LETTUCE	Escherichia coli O103:H25	SD	Medium	KrakenUniq	FBP-specific	58.315	6265548	934073950	23.485	3132774	2923352	6.68487417	2923158	194	0.0062	2923158	2922708	29.84	872137	0	1E+06	561	genus	Escherichia
EC-4	LETTUCE	Escherichia coli O103:H25	SD	Medium	KrakenUniq	EC-specific	58.315	6265548	934073950	23.485	3132774	2923352	6.68487417	2923158	194	0.0062	2923158	2921660	97.6	2852927	0	6E+06	561	genus	Escherichia
EC-4	LETTUCE	Escherichia coli O103:H25	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	1159936	NA	NA	NA	NA	NA	1159936	1158319	32.7	379257	0	1E+06	561	genus	Escherichia
EC-4	LETTUCE	Escherichia coli O103:H25	SI	SI	KrakenUniq	EC-specific	NA	NA	NA	NA	1159936	NA	NA	NA	NA	NA	1,159,936	1156796	97.49	1130845	0	6E+06	561	genus	Escherichia
EC-5	LETTUCE	Escherichia coli O157:H7	NSD	High	KrakenUniq	FBP-specific	1.915	10408688	1.555E+09	25.425	5204344	5070308	2.57546388	5070124	184	0.0035	5070124	5069041	58.02	2941541	0	2E+06	561	genus	Escherichia
EC-5	LETTUCE	Escherichia coli O157:H7	NSD	High	KrakenUniq	EC-specific	1.915	10408688	1.555E+09	25.425	5204344	5070308	2.57546388	5070124	184	0.0035	5070124	5063086	97.87	4961990	0	6E+06	561	genus	Escherichia
EC-5	LETTUCE	Escherichia coli O157:H7	SD	High	KrakenUniq	FBP-specific	4.947	8809958	1.318E+09	23.365	4404979	4099064	6.94475502	4098828	236	0.0054	4098828	4097978	58.42	2394681	0	2E+06	561	genus	Escherichia
EC-5	LETTUCE	Escherichia coli O157:H7	SD	High	KrakenUniq	EC-specific	4.947	8809958	1.318E+09	23.365	4404979	4099064	6.94475502	4098828	236	0.0054	4098828	4093868	97.77	4007439	0	6E+06	561	genus	Escherichia
EC-5	LETTUCE	Escherichia coli O157:H7	NSD	Low	KrakenUniq	FBP-specific	1.422	12201640	1.821E+09	25.865	6100820	5970079	2.14300701	5969846	233	0.0038	5969846	5968649	58.4	3486616	0	2E+06	561	genus	Escherichia
EC-5	LETTUCE	Escherichia coli O157:H7	NSD	Low	KrakenUniq	EC-specific	1.422	12201640	1.821E+09	25.865	6100820	5970079	2.14300701	5969846	233	0.0038	5969846	5960989	97.91	5845256	0	7E+06	561	genus	Escherichia
EC-5	LETTUCE	Escherichia coli O157:H7	SD	Low	KrakenUniq	FBP-specific	3.085	18021014	2.699E+09	23.095	9010507	8284165	8.06105583	8283675	490	0.0054	8283675	8282218	59.52	4930707	0	2E+06	561	genus	Escherichia
EC-5	LETTUCE	Escherichia coli O157:H7	SD	Low	KrakenUniq	EC-specific	3.085	18021014	2.699E+09	23.095	9010507	8284165	8.06105583	8283675	490	0.0054	8283675	8260901	97.66	8090251	0	7E+06	561	genus	Escherichia
EC-5	LETTUCE	Escherichia coli O157:H7	NSD	Medium	KrakenUniq	FBP-specific	1.637	11015074	1.648E+09	25.575	5507537	5382790	2.26502337	5382599	191	0.0035	5382599	5381537	58.4	3143455	0	2E+06	561	genus	Escherichia
EC-5	LETTUCE	Escherichia coli O157:H7	NSD	Medium	KrakenUniq	EC-specific	1.637	11015074	1.648E+09	25.575	5507537	5382790	2.26502337	5382599	191	0.0035	5382599	5376450	97.89	5269081	0	6E+06	561	genus	Escherichia
EC-5	LETTUCE	Escherichia coli O157:H7	SD	Medium	KrakenUniq	FBP-specific	3.723	12458298	1.841E+09	23.75	6229149	5839630	6.25316556	5839293	337	0.0054	5839293	5837510	55.95	3266960	0	2E+06	561	genus	Escherichia
EC-5	LETTUCE	Escherichia coli O157:H7	SD	Medium	KrakenUniq	EC-specific	3.723	12458298	1.841E+09	23.75	6229149	5839630	6.25316556	5839293	337	0.0054	5839293	5829471	97.84	5713319	0	7E+06	561	genus	Escherichia
EC-5	LETTUCE	Escherichia coli O157:H7	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	895071	NA	NA	NA	NA	NA	895071	894340	58.13	520274	0	1E+06	561	genus	Escherichia
EC-5	LETTUCE	Escherichia coli O157:H7	SI	SI	KrakenUniq	EC-specific	NA	NA	NA	NA	895071	NA	NA	NA	NA	NA	895071	893490	97.95	876739	0	6E+06	561	genus	Escherichia
EC-6	LETTUCE	Escherichia coli O157:H7	NSD	High	KrakenUniq	FBP-specific	1.326	12205262	1.818E+09	25.825	6102631	5972240	2.13663582	5971982	258	0.0042	5971982	5970680	57.19	3415456	0	2E+06	561	genus	Escherichia
EC-6	LETTUCE	Escherichia coli O157:H7	NSD	High	KrakenUniq	EC-specific	1.326	12205262	1.818E+09	25.825	6102631	5972240	2.13663582	5971982	258	0.0042	5971982	5956486	97.71	5835144	0	7E+06	561	genus	Escherichia
EC-6	LETTUCE	Escherichia coli O157:H7	SD	High	KrakenUniq	FBP-specific	6.72	14727304	2.182E+09	23.51	7363652	6885048	6.49954669	6884626	422	0.0057	6884626	6882532	56.68	3901927	0	2E+06	561	genus	Escherichia
EC-6	LETTUCE	Escherichia coli O157:H7	SD	High	KrakenUniq	EC-specific	6.72	14727304	2.182E+09	23.51	7363652	6885048	6.49954669	6884626	422	0.0057	6884626	6872829	97.79	6732353	0	7E+06	561	genus	Escherichia
EC-6	LETTUCE	Escherichia coli O157:H7	NSD	Low	KrakenUniq	FBP-specific	1.749	16534908	2.472E+09	25.67	8267454	8089711	2.14991217	8089381	330	0.004	8089381	8087799	58.37	4722099	0	2E+06	561	genus	Escherichia
EC-6	LETTUCE	Escherichia coli O157:H7	NSD	Low	KrakenUniq	EC-specific	1.749	16534908	2.472E+09	25.67	8267454	8089711	2.14991217	8089381	330	0.004	8089381	8056037	97.54	7890074	0	7E+06	561	genus	Escherichia
EC-6	LETTUCE	Escherichia coli O157:H7	SD	Low	KrakenUniq	FBP-specific	11.673	22572014	3.339E+09	23.56	11286007	1.1E+07	6.93496823	10502681	645	0.0057	10502681	10499126	56.95	5981421	0	3E+06	561	genus	Escherichia
EC-6	LETTUCE	Escherichia coli O157:H7	SD	Low	KrakenUniq	EC-specific	11.673	22572014	3.339E+09	23.56	11286007	1.1E+07	6.93496823	10502681	645	0.0057	10502681	10479676	97.75	10265872	0	8E+06	561	genus	Escherichia
EC-6	LETTUCE	Escherichia coli O157:H7	NSD	Medium	KrakenUniq	FBP-specific	0.906	12427132	1.853E+09	25.775	6213566	6079598	2.1560566	6079380	218	0.0035	6079380	6077780	58.88	3579560	0	2E+06	561	genus	Escherichia
EC-6	LETTUCE	Escherichia coli O157:H7	NSD	Medium	KrakenUniq	EC-specific	0.906	12427132	1.853E+09	25.775	6213566	6079598	2.1560566	6079380	218	0.0035	6079380	6065763	97.78	5944621	0	7E+06	561	genus	Escherichia
EC-6	LETTUCE	Escherichia coli O157:H7	SD	Medium	KrakenUniq	FBP-specific	1.333	37382186	5.458E+09	23.825	18691093	1.8E+07	6.36950445	17499358	1205	0.0064	17499358	17493114	53.71	9398032	0	3E+06	561	genus	Escherichia
EC-6	LETTUCE	Escherichia coli O157:H7	SD	Medium	KrakenUniq	EC-specific	1.333	37382186	5.458E+09	23.825	18691093	1.8E+07	6.36950445	17499358	1205	0.0064	17499358	17438963	97.59	17078143	0	9E+06	561	genus	Escherichia
EC-6	LETTUCE	Escherichia coli O157:H7	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	1141246	NA	NA	NA	NA	NA	1141246	1140626	59.59	680056	0	2E+06	561	genus	Escherichia
EC-6	LETTUCE	Escherichia coli O157:H7	SI	SI	KrakenUniq	EC-specific	NA	NA	NA	NA	1141246	NA	NA	NA	NA	NA	1141246	1139692	97.79	1116023	0	6E+06	561	genus	Escherichia
NC	LETTUCE	Negative_control	NSD	NC	KrakenUniq	FBP-specific	0.392	11726554	1.735E+09	26.13	5863277	5734695	2.19300572	134384	5600311	95.515	134384	90716	1.72	2312	0	127076	561	genus	Escherichia
NC	LETTUCE	Negative_control	NSD	NC	KrakenUniq	EC-specific	0.392	11726554	1.735E+09	26.13	5863277	5734695	2.19300572	134384	5600311	95.515	134384	32270	20.24	27193	0	1E+06	561	genus	Escherichia
NC	LETTUCE	Negative_control	SD	NC	KrakenUniq	FBP-specific	0.334	248094	34928382	23.495	124047	115066	7.23999774	112207	2859	2.3048	112207	110037	1.804	2024	0	104197	561	genus	Escherichia
NC	LETTUCE	Negative_control	SD	NC	KrakenUniq	EC-specific	0.334	248094	34928382	23.495	124047	115066	7.23999774	112207	2859	2.3048	112207	32918	26.47	29698	0	1E+06	561	genus	Escherichia

Appendix 4.3. Sequencing and taxonomic classification metrics for *Shigella* spp. short-read quasi-metagenomic datasets generated in Chapter 4. Performance is compared across two laboratory workflows: no saponin-based eukaryotic DNA depletion (NSD) and saponin depletion (SD). Samples were generated from a single food matrix (carrot) at three spike levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). Taxonomic results are reported for two reference databases: the foodborne pathogen (FBP)-specific database and *Shigella* (SH)-specific database. Sample identifiers include *Shigella* species (SF-1 = *S. flexneri*; SS-1, 2 = *S. sonnei*). Negative control (NC) samples were included to establish baseline taxonomic background. PE indicates paired-end sequencing. ER indicates eukaryotic reads. Single-isolate (SI) reference genomes were previously sequenced and were not evaluated using the preprocessing workflow developed in this study; these entries are indicated as NA in the table.

Strain	Food	Pathogen	Wet Lab	Spike Level	Classifier	Database	DNA Concentration	Total Reads R1 R2	Total Bases	Avg Base Quality	Raw Reads PE	fastp	Percent fastp	fastp HostRemoved	Eukaryotic Reads	Percent ER	Number of Reads Available for Classification (PE)	Classified reads	Percent classified	reads clade	taxReads	kmers	taxID	rank	taxName
NC	CARROT	Negative_control	NSD	NC	KrakenUniq	FBP-specific	0.328	10647764	1.572E+09	25.54	5323882	5130825	3.62624491	4816160	314665	5.9104	4816160	264889	0.049	2360	110	38229	620	genus	Shigella
NC	CARROT	Negative_control	NSD	NC	KrakenUniq	SH-specific	0.328	10647764	1.572E+09	25.54	5323882	5130825	3.62624491	4816160	314665	5.9104	4816160	122486	1.413	68045	32600	1E+06	620	genus	Shigella
NC	CARROT	Negative_control	SD	NC	KrakenUniq	FBP-specific	0.317	814	85316	22.995	407	349	14.2506143	349	0	0	349	114	3.725	13	3	308	620	genus	Shigella
NC	CARROT	Negative_control	SD	NC	KrakenUniq	SH-specific	0.317	814	85316	22.995	407	349	14.2506143	349	0	0	349	103	14.61	51	29	5482	620	genus	Shigella
SF-1	CARROT	Shigella flexneri	NSD	High	KrakenUniq	FBP-specific	3.274	4983122	738733065	25.215	2491561	2409930	3.27629948	2409782	148	0.0059	2409782	2407719	41.37	996912	68491	383406	620	genus	Shigella
SF-1	CARROT	Shigella flexneri	NSD	High	KrakenUniq	SH-specific	3.274	4983122	738733065	25.215	2491561	2409930	3.27629948	2409782	148	0.0059	2409782	2407403	96.38	2322545	905331	5E+06	620	genus	Shigella
SF-1	CARROT	Shigella flexneri	SD	High	KrakenUniq	FBP-specific	13.102	7074814	1.052E+09	27.4	3537407	3477201	1.70198114	3477007	194	0.0055	3477007	3475410	43.6	1515950	107196	376276	620	genus	Shigella
SF-1	CARROT	Shigella flexneri	SD	High	KrakenUniq	SH-specific	13.102	7074814	1.052E+09	27.4	3537407	3477201	1.70198114	3477007	194	0.0055	3477007	3475132	96.33	3349482	1233279	5E+06	620	genus	Shigella
SF-1	CARROT	Shigella flexneri	NSD	Low	KrakenUniq	FBP-specific	5.465	5178524	769803798	24.97	2589262	2498507	3.50505279	2498344	163	0.0063	2498344	2496465	41.47	1035946	66246	382226	620	genus	Shigella
SF-1	CARROT	Shigella flexneri	NSD	Low	KrakenUniq	SH-specific	5.465	5178524	769803798	24.97	2589262	2498507	3.50505279	2498344	163	0.0063	2498344	2496260	96.52	2411292	934258	5E+06	620	genus	Shigella
SF-1	CARROT	Shigella flexneri	SD	Low	KrakenUniq	FBP-specific	9.726	4343946	649996499	27.305	2171973	2133245	1.78307926	2133155	90	0.0041	2133155	2132156	43.56	929125	61549	356633	620	genus	Shigella
SF-1	CARROT	Shigella flexneri	SD	Low	KrakenUniq	SH-specific	9.726	4343946	649996499	27.305	2171973	2133245	1.78307926	2133155	90	0.0041	2133155	2131998	96.39	2056180	751744	5E+06	620	genus	Shigella
SF-1	CARROT	Shigella flexneri	NSD	Medium	KrakenUniq	FBP-specific	2.651	26181722	3.902E+09	25.52	13090861	1.3E+07	3.03779866	12692484	703	0.0054	12692484	12681391	40.9	5191264	340370	489691	620	genus	Shigella
SF-1	CARROT	Shigella flexneri	NSD	Medium	KrakenUniq	SH-specific	2.651	26181722	3.902E+09	25.52	13090861	1.3E+07	3.03779866	12692484	703	0.0054	12692484	12680258	96.61	12262203	4851489	5E+06	620	genus	Shigella
SF-1	CARROT	Shigella flexneri	SD	Medium	KrakenUniq	FBP-specific	8.641	4433744	658347286	27.115	2216872	2173420	1.96005904	2173305	115	0.0052	2173305	2172158	42.66	927075	62491	366825	620	genus	Shigella
SF-1	CARROT	Shigella flexneri	SD	Medium	KrakenUniq	SH-specific	8.641	4433744	658347286	27.115	2216872	2173420	1.96005904	2173305	115	0.0052	2173305	2171965	96.37	2094380	789102	5E+06	620	genus	Shigella
SF-1	CARROT	Shigella flexneri	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	683576	NA	NA	NA	NA	NA	683576	683357	48.49	331451	21285	388713	620	genus	Shigella
SF-1	CARROT	Shigella flexneri	SI	SI	KrakenUniq	SH-specific	NA	NA	NA	NA	683576	NA	NA	NA	NA	NA	683576	683199	96.34	658580	216014	5E+06	620	genus	Shigella
SS-1	CARROT	Shigella sonnei	NSD	High	KrakenUniq	FBP-specific	5.201	5060728	752081888	25.075	2530364	2443648	3.42701682	2442968	680	0.0269	2442968	2432907	25.03	611439	27253	224525	620	genus	Shigella
SS-1	CARROT	Shigella sonnei	NSD	High	KrakenUniq	SH-specific	5.201	5060728	752081888	25.075	2530364	2443648	3.42701682	2442968	680	0.0269	2442968	2425011	92.81	2267428	1355148	4E+06	620	genus	Shigella
SS-1	CARROT	Shigella sonnei	SD	High	KrakenUniq	FBP-specific	9.887	8166532	1.216E+09	27.235	4083266	4006901	1.8701941	4006665	236	0.0058	4006665	4005275	25.91	1038187	45889	223244	620	genus	Shigella
SS-1	CARROT	Shigella sonnei	SD	High	KrakenUniq	SH-specific	9.887	8166532	1.216E+09	27.235	4083266	4006901	1.8701941	4006665	236	0.0058	4006665	3980663	92.6	3710080	2120983	4E+06	620	genus	Shigella
SS-1	CARROT	Shigella sonnei	NSD	Low	KrakenUniq	FBP-specific	3.987	7270848	1.081E+09	24.98	3635424	3508987	3.47791619	3507511	1476	0.0406	3507511	3488411	24.95	875012	39071	236782	620	genus	Shigella
SS-1	CARROT	Shigella sonnei	NSD	Low	KrakenUniq	SH-specific	3.987	7270848	1.081E+09	24.98	3635424	3508987	3.47791619	3507511	1476	0.0406	3507511	3474545	92.57	3246763	1935496	4E+06	620	genus	Shigella
SS-1	CARROT	Shigella sonnei	SD	Low	KrakenUniq	FBP-specific	10.596	4152746	617072691	27.33	2076373	2037407	1.87663777	2037282	125	0.006	2037282	2036254	25.16	512541	22103	200906	620	genus	Shigella
SS-1	CARROT	Shigella sonnei	SD	Low	KrakenUniq	SH-specific	10.596	4152746	617072691	27.33	2076373	2037407	1.87663777	2037282	125	0.006	2037282	2016106	91.79	1870057	1044174	4E+06	620	genus	Shigella
SS-1	CARROT	Shigella sonnei	NSD	Medium	KrakenUniq	FBP-specific	4.187	19554818	2.913E+09	25.09	9777409	9449878	3.34987521	9446707	3171	0.0324	9446707	9402783	25.15	2375729	105990	307050	620	genus	Shigella
SS-1	CARROT	Shigella sonnei	NSD	Medium	KrakenUniq	SH-specific	4.187	19554818	2.913E+09	25.09	9777409	9449878	3.34987521	9446707	3171	0.0324	9446707	9374845	92.84	8770094	5232596	5E+06	620	genus	Shigella
SS-1	CARROT	Shigella sonnei	SD	Medium	KrakenUniq	FBP-specific	5.034	7984996	1.187E+09	27.455	3992498	3924407	1.70547362	3924163	244	0.0061	3924163	3922730	26.15	1026112	46217	219845	620	genus	Shigella
SS-1	CARROT	Shigella sonnei	SD	Medium	KrakenUniq	SH-specific	5.034	7984996	1.187E+09	27.455	3992498	3924407	1.70547362	3924163	244	0.0061	3924163	3911525	93.26	3659573	2157679	4E+06	620	genus	Shigella
SS-1	CARROT	Shigella sonnei	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	1047493	NA	NA	NA	NA	NA	1047493	1046787	26.68	279503	11044	255653	620	genus	Shigella
SS-1	CARROT	Shigella sonnei	SI	SI	KrakenUniq	SH-specific	NA	NA	NA	NA	1047493	NA	NA	NA	NA	NA	1,047,493	1,034,512	92.8	972107	557395	5E+06	620	genus	Shigella
SS-2	CARROT	Shigella sonnei	NSD	High	KrakenUniq	FBP-specific	4.837	4248978	635875875	24.95	2124489	2045542	3.71604654	2045283	259	0.0122	2045283	2042862	24.92	509699	22897	216029	620	genus	Shigella
SS-2	CARROT	Shigella sonnei	NSD	High	KrakenUniq	SH-specific	4.837	4248978	635875875	24.95	2124489	2045542	3.71604654	2045283	259	0.0122	2045283	2033057	92.94	1900919	1097679	4E+06	620	genus	Shigella
SS-2	CARROT	Shigella sonnei	SD	High	KrakenUniq	FBP-specific	5.309	10652750	1.595E+09	27.38	5326375	5229323	1.82210227	5229036	287	0.0054	5229036	5227089	25.66	1342023	60461	228921	620	genus	Shigella
SS-2	CARROT	Shigella sonnei	SD	High	KrakenUniq	SH-specific	5.309	10652750	1.595E+09	27.38	5326375	5229323	1.82210227	5229036	287	0.0054	5229036	5210846	93.25	4876055	2814228	4E+06	620	genus	Shigella
SS-2	CARROT	Shigella sonnei	NSD	Low	KrakenUniq	FBP-specific	5.215	4289146	639547613	24.8	2144573	2060166	3.93584177	2059884	282	0.0131	2059884	2056340	24.7	508889	22948	215951	620	genus	Shigella
SS-2	CARROT	Shigella sonnei	NSD	Low	KrakenUniq	SH-specific	5.215	4289146	639547613	24.8	2144573	2060166	3.93584177	2059884	282	0.0131	2059884	2046914	92.9	1913619	1118404	4E+06	620	genus	Shigella
SS-2	CARROT	Shigella sonnei	SD	Low	KrakenUniq	FBP-specific	12.401	4021884	597375722	27.32	2010942	1972851	1.8941869	1972716	135	0.0067	1972716	1972025	25.06	494354	21549	200305	620	genus	Shigella
SS-2	CARROT	Shigella sonnei	SD	Low	KrakenUniq	SH-specific	12.401	4021884	597375722	27.32	2010942	1972851	1.8941869	1972716	135	0.0067	1972716	1960203	92.59	1826502	1045623	4E+06	620	genus	Shigella
SS-2	CARROT	Shigella sonnei	NSD	Medium	KrakenUniq	FBP-specific	7.287	7916244	1.185E+09	24.925	3958122	3819994	3.48973579	3819508	486	0.0123	3819508	3814735	25.25	964366	43300	240402	620	genus	Shigella
SS-2	CARROT	Shigella sonnei	NSD	Medium	KrakenUniq	SH-specific	7.287	7916244	1.185E+09	24.925	3958122	3819994	3.48973579	3819508	486	0.0123	3819508	3799841	93.09	3555479	2064110	5E+06	620	genus	Shigella
SS-2	CARROT	Shigella sonnei	SD	Medium	KrakenUniq	FBP-specific	6.347	3991104	598016481	27.175	1995552	1956968	1.9335001	1956850	118	0.0059	1956850	1956144	25.91	507099	22672	201951	620	genus	Shigella
SS-2	CARROT	Shigella sonnei	SD	Medium	KrakenUniq	SH-specific	6.347	3991104	598016481	27.175	1995552	1956968	1.9335001	1956850	118	0.0059	1956850	1948523	93.11	1822047	1037768	4E+06	620	genus	Shigella
SS-2	CARROT	Shigella sonnei	SI	SI	KrakenUniq	FBP-specific	NA	NA	NA	NA	1428870	NA	NA	NA	NA	NA	1428870	1428409	27.29	389963	14600	263413	620	genus	Shigella
SS-2	CARROT	Shigella sonnei	SI	SI	KrakenUniq	SH-specific	NA	NA	NA	NA	1428870	NA	NA	NA	NA	NA	1428870	1408446	92.27	1318364	696758	5E+06	620	genus	Shigella

Appendix 4.4. PulseNet Canada Guidelines for Assessing the Quality of Sequence Data in BioNumerics.

Title: PulseNet Canada Guidelines for Assessing the Quality of Sequence Data in BioNumerics **Chart # DED-C-350A-4**
Date of Issue: 2020-11-05

Organism	Quality Metric	Value**
Listeria monocytogenes	AvgQuality	≥30
	N50•	~ 50,000
	NrContigs	~ 200
	Length	~ 2,700,000 to 3,200,000 base pairs
	NrConsensus	~ 2500
	Core Percent	≥95%^
Escherichia coli	AvgQuality	≥30
	N50•	~ 100,000
	NrContigs	~ 500
	Length	~ 4,500,000 to 6,000,000 base pairs
	NrConsensus	~ 3800ǂ
	Core Percent	≥90%
Shigella spp.	AvgQuality	≥30
	N50•	~ 18,000
	NrContigs	~ 500
	Length	~4,300,000 to 5,000,000 base pairs
	NrConsensus	~ 3300ǂ
	Core Percent	≥90%
Salmonella enterica	AvgQuality	≥30
	N50•	~ 100,000
	NrContigs	~ 200
	Length	~ 4,400,000 to 6,000,000 base pairs*
	NrConsensus	~ 3800ǂ
	Core Percent	≥90%
Vibrio cholerae	AvgQuality	≥30
	N50•	~ 100,000
	NrContigs	~ 150
	Length	~3,800,000 to 4,300,000 base pairs
	NrConsensus	~ 2500
	Core Percent	Core Scheme not yet available
Campylobacter coli/jejuni	AvgQuality	≥30
	N50•	~ 100,000
	NrContigs	~ 150
	Length	~1,400,000 to 2,000,000 base pairs
	NrConsensus	~ 1100
	Core Percent	Not yet validated

Sequences must pass the QC metrics highlighted in red prior to submission to the PulseNET Canada National Databases. Seqeunces that do not pass the other QC metrics may be submitted to the National Databases, but may be rejected upon further QC review. ***AvgQuality*** =Average quality of the sequence read set using the quality scores obtained from the raw data; ***N50***=Length of the median contig (in terms of sequence length); ***NrContigs***=Number of contigs in the SPAdes assembled sequence; ***Length***=Length of the SPAdes assembled genome; ***NrConsensus***=Number of alleles that are similar between both allele identification algorithms; ***Core Percent***=Percentage of loci found belonging to the subset of core loci.

**Values listed are anticipated to become progressively more refined as sequencing continues and experience with the data increases. Metrics were validated on isolates sequenced with the Illumina MiSeq v3 600 cycle kit.

•Impacted by read length; a shorter read length will result in a smaller N50, an N50 greater than the value listed is acceptable.

^ This value can be as low as 90% for some lineage III / IV Listeria. Isolates with a core percent of 90-94% will be accepted into the national database but will not be assigned nomenclature.

ǂNrConsensus has been observed to be as low as 3600 alleles for *E. coli* O157:H7 and even lower for some non-O157 serotypes; this value has also been observed to be as low as 3500 for some rare *Salmonella* serovars; this value has also been observed to be as low as 3000 for Shigella dysenteriae.

a This value is typically between 4.4 to 5.6 Mbp for the majority of Salmonella species. Some subspecies IIIb and Paratyphi B var. Java may genome sizes closer to or slightly larger than 6Mbp.

Appendix 5.1. Sequencing and taxonomic classification metrics for *Listeria* spp. long-read quasi-metagenomic datasets generated in Chapter 5. Performance is compared across two laboratory workflows: no saponin-based eukaryotic DNA depletion (NSD) and saponin depletion (SD). Samples were generated from four food matrices (ham, roast beef, salad kit, turkey) at three spike levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). Taxonomic results are reported for the foodborne pathogen (FBP)-specific database. Sample identifiers include *Listeria* species (LI = *L. innocua*; LM-1–8 = *L. monocytogenes*; LS = *L. seeligeri*; LW = *L. welshimeri*). Negative control (NC) samples were included to establish baseline taxonomic background. Single-isolate (SI) reference genomes were previously sequenced and were not evaluated using the preprocessing workflow developed in this study; these entries are indicated as NA in the table (Continued). FL90 = filtlong filtering retaining best 90% of reads; pcABI = porcechop ABI; ER = Eukaryotic reads

Strain	Food	Pathogen	Wet Lab	Spike	Classifier	Database	DNA Concentration	Total Reads	Total Bases	Mean Read Length	Mean Base Quality	Total Reads (FL90)	Percent Filtered (FL90)	Total Reads (FL90+pcABI)	Percent Filtered (FL90+pcABI)	Total Reads (FL90+pcABI+ER)	Eukaryotic Reads	Percent Eukaryotic	Number of Reads Available for Classification	Percent Classified	reads clade	taxReads	kmers	taxID	rank	taxName
LI-1	HAM	Listeria innocua	NSD	Low	KrakenUniq	FBP-Specific	6.718	68162	409304669	6004.9	13.4	33181	48.67961621	33248	48.77791145	32671	577	1.76609225	32671	99.66	32560	13	4E+06	1637	genus	Listeria
LI-1	HAM	Listeria innocua	NSD	Medium	KrakenUniq	FBP-Specific	17.228	85319	851311738	9978	13.7	45054	52.8065261	45087	52.84520447	44774	313	0.69906642	44774	99.68	44630	5	4E+06	1637	genus	Listeria
LI-1	HAM	Listeria innocua	NSD	High	KrakenUniq	FBP-Specific	32.792	65066	650507576	9997.7	13.5	34583	53.15064704	34607	53.18753266	34408	199	0.57835387	34408	99.52	34242	5	3E+06	1637	genus	Listeria
LI-1	HAM	Listeria innocua	SD	Low	KrakenUniq	FBP-Specific	19.079	112185	946449526	8436.5	13.2	64092	57.13063244	64092	57.13063244	64022	70	0.10933742	64022	99.76	63869	14	4E+06	1637	genus	Listeria
LI-1	HAM	Listeria innocua	SD	Medium	KrakenUniq	FBP-Specific	22.107	94416	656205024	6950.1	13.4	40451	42.843374	40618	43.0202508	40484	134	0.33099496	40484	99.46	40265	11	4E+06	1637	genus	Listeria
LI-1	HAM	Listeria innocua	SD	High	KrakenUniq	FBP-Specific	71.297	135715	826428343	6089.4	13.3	52415	38.62137568	52641	38.78790112	52407	234	0.44650524	52407	99.44	52112	22	5E+06	1637	genus	Listeria
LI-1	HAM	Listeria innocua	SI	SI	KrakenUniq	FBP-Specific	32.2	98395	1611147253	13535.1	13.4	NA	NA	NA	NA	NA	NA	NA	98644	99.84	98484	16	5E+06	1637	genus	Listeria
LM-1	HAM	Listeria monocytogenes	NSD	Low	KrakenUniq	FBP-Specific	7.217	51170	422475607	8256.3	13.1	26803	52.38030096	26857	52.48583154	26593	264	0.99274245	26593	99.83	26549	78	4E+06	1637	genus	Listeria
LM-1	HAM	Listeria monocytogenes	NSD	Medium	KrakenUniq	FBP-Specific	20.13	164625	1593141606	9677.4	13.3	90639	55.05785877	90749	55.1246773	90333	416	0.4605183	90333	99.96	90296	156	5E+06	1637	genus	Listeria
LM-1	HAM	Listeria monocytogenes	NSD	High	KrakenUniq	FBP-Specific	37.853	68882	782794874	11364.3	13.7	35993	52.25312854	36119	52.43605006	35975	144	0.40027797	35975	99.91	35944	29	4E+06	1637	genus	Listeria
LM-1	HAM	Listeria monocytogenes	SD	Low	KrakenUniq	FBP-Specific	16.245	156822	1217244741	7762	13.1	75606	48.2113479	75723	48.28595478	75551	172	0.22766079	75551	99.97	75526	150	4E+06	1637	genus	Listeria
LM-1	HAM	Listeria monocytogenes	SD	Medium	KrakenUniq	FBP-Specific	60.379	201494	1908751452	9473	13	103107	51.17125076	103315	51.27447964	103062	253	0.2454833	103062	99.93	102986	139	5E+06	1637	genus	Listeria
LM-1	HAM	Listeria monocytogenes	SD	High	KrakenUniq	FBP-Specific	72.31	129614	1000542161	7719.4	13.9	58675	45.26902958	58785	45.35389696	58457	328	0.56109619	58457	99.85	58369	70	4E+06	1637	genus	Listeria
LM-1	HAM	Listeria monocytogenes	SI	SI	KrakenUniq	FBP-Specific	20	52535	1073934032	16915	13.6	NA	NA	NA	NA	NA	NA	NA	52673	99.99	52669	47	5E+06	1637	genus	Listeria
LM-2	HAM	Listeria monocytogenes	NSD	Low	KrakenUniq	FBP-Specific	2.847	74401	491403506	6604.8	13.4	40999	55.10544213	41063	55.19146248	40579	484	1.19273516	40579	99.96	40561	0	4E+06	1637	genus	Listeria
LM-2	HAM	Listeria monocytogenes	NSD	Medium	KrakenUniq	FBP-Specific	9.354	158509	1788912193	11285.9	13.2	89081	56.19933253	89248	56.30468932	88963	285	0.3203579	88963	99.93	88899	0	4E+06	1637	genus	Listeria
LM-2	HAM	Listeria monocytogenes	NSD	High	KrakenUniq	FBP-Specific	68.949	210923	2238945802	10615	13.6	113584	53.85093138	113831	53.96803573	113557	274	0.24128852	113557	99.96	113507	0	5E+06	1637	genus	Listeria
LM-2	HAM	Listeria monocytogenes	SD	Low	KrakenUniq	FBP-Specific	18.08	76625	615804195	8036.6	13.5	39175	51.12561175	39175	51.12561175	39069	106	0.27131485	39069	99.97	39058	1	4E+06	1637	genus	Listeria
LM-2	HAM	Listeria monocytogenes	SD	Medium	KrakenUniq	FBP-Specific	32.008	100032	715605147	7153.8	12.9	40943	40.92990243	41035	41.021873	40922	113	0.27613509	40922	99.86	40865	0	4E+06	1637	genus	Listeria
LM-2	HAM	Listeria monocytogenes	SD	High	KrakenUniq	FBP-Specific	36.666	222122	1910649825	8601.8	13.1	109921	49.48676853	111267	50.09274183	110958	309	0.27848375	110958	99.95	110897	5	5E+06	1637	genus	Listeria
LM-2	HAM	Listeria monocytogenes	SI	SI	KrakenUniq	FBP-Specific	15.6	83624	1344552803	13441.1	13.5	NA	NA	NA	NA	NA	NA	NA	83799	100	83797	0	5E+06	1637	genus	Listeria
NC	HAM	NC	NSD	NC	KrakenUniq	FBP-Specific	0.76	4434	5266739	1187.8	12.9	1931	43.54984213	1925	43.41452413	749	1176	157.009346	749	10.81	81	0	322267	1637	genus	Listeria
NC	HAM	NC	SD	NC	KrakenUniq	FBP-Specific	0.049	2494	2547995	1021.6	12.2	224	8.981555734	224	8.981555734	221	3	1.35746606	221	80.54	178	0	960194	1637	genus	Listeria
LM-3	ROAST_BEEF	Listeria monocytogenes	NSD	Low	KrakenUniq	FBP-Specific	3.964	76221	268999267	3529.2	12.9	20524	26.92696239	20578	26.997809	20193	385	1.9066013	20193	99.16	20024	63	4E+06	1637	genus	Listeria
LM-3	ROAST_BEEF	Listeria monocytogenes	NSD	Medium	KrakenUniq	FBP-Specific	9.864	93205	365376588	3920.1	13.2	22308	23.93433829	22393	24.02553511	22222	171	0.7695077	22222	99.65	22144	27	5E+06	1637	genus	Listeria
LM-3	ROAST_BEEF	Listeria monocytogenes	NSD	High	KrakenUniq	FBP-Specific	15.764	162701	910105105	5593.7	13.1	59791	36.74900585	59894	36.81231216	59571	323	0.54221014	59571	99.93	59529	88	4E+06	1637	genus	Listeria
LM-3	ROAST_BEEF	Listeria monocytogenes	SD	Low	KrakenUniq	FBP-Specific	0	9893	23952652	2421.2	12.9	4034	40.77630648	4068	41.11998383	4058	10	0.24642681	4058	98.45	3995	38	3E+06	1637	genus	Listeria
LM-3	ROAST_BEEF	Listeria monocytogenes	SD	Medium	KrakenUniq	FBP-Specific	8.66	76391	582653459	7627.3	13.7	38784	50.77037871	39195	51.3084002	39008	187	0.47938884	39008	99.87	38957	41	4E+06	1637	genus	Listeria
LM-3	ROAST_BEEF	Listeria monocytogenes	SD	High	KrakenUniq	FBP-Specific	25.122	91341	549368007	6014.5	13.6	41687	45.63886973	41896	45.86768264	41690	206	0.49412329	41690	99.94	41663	72	4E+06	1637	genus	Listeria
LM-3	ROAST_BEEF	Listeria monocytogenes	SI	SI	KrakenUniq	FBP-Specific	16.2	66237	1183415982	14970.8	13.6	NA	NA	NA	NA	NA	NA	NA	66387	100	66385	48	5E+06	1637	genus	Listeria
LM-4	ROAST_BEEF	Listeria monocytogenes	NSD	Low	KrakenUniq	FBP-Specific	3.962	66134	198731891	3005	12.7	16581	25.07182387	16581	25.07182387	15318	1263	8.24520172	15318	98.91	15151	47	3E+06	1637	genus	Listeria
LM-4	ROAST_BEEF	Listeria monocytogenes	NSD	Medium	KrakenUniq	FBP-Specific	7.452	112347	531135951	4727.6	13.2	33517	29.8334624	34008	30.27050121	33855	153	0.45192734	33855	99.81	33792	30	4E+06	1637	genus	Listeria
LM-4	ROAST_BEEF	Listeria monocytogenes	NSD	High	KrakenUniq	FBP-Specific	20.998	94013	719922423	7657.7	13.7	39854	42.39200962	39885	42.42498378	39677	208	0.52423318	39677	99.98	39670	28	4E+06	1637	genus	Listeria
LM-4	ROAST_BEEF	Listeria monocytogenes	SD	Low	KrakenUniq	FBP-Specific	2.5	53563	109896141	2051.7	13.8	24876	46.44250695	24901	46.48918096	24857	44	0.17701251	24857	98.65	24521	247	3E+06	1637	genus	Listeria
LM-4	ROAST_BEEF	Listeria monocytogenes	SD	Medium	KrakenUniq	FBP-Specific	70	161500	1272468154	7879.1	13.9	89610	55.48606811	89680	55.52941176	89344	336	0.3760745	89344	99.89	89242	85	4E+06	1637	genus	Listeria
LM-4	ROAST_BEEF	Listeria monocytogenes	SD	High	KrakenUniq	FBP-Specific	12.8	116316	894486144	7690.1	13.3	58045	49.90285085	58045	49.90285085	57731	314	0.54390189	57731	99.95	57700	52	4E+06	1637	genus	Listeria
LM-4	ROAST_BEEF	Listeria monocytogenes	SI	SI	KrakenUniq	FBP-Specific	18.6	32049	633726654	16268.6	13.7	NA	NA	NA	NA	NA	NA	NA	32123	100	32122	15	4E+06	1637	genus	Listeria
NC	ROAST_BEEF	NC	NSD	NC	KrakenUniq	FBP-Specific	1.907	37616	16054166	426.8	12.7	25140	66.83326244	25140	66.83326244	1560	23580	1511.53846	1560	9.936	155	3	444635	1637	genus	Listeria
NC	ROAST_BEEF	NC	SD	NC	KrakenUniq	FBP-Specific	0	4128	1480242	358.6	12.9	2391	57.92151163	1935	46.875	450	1485	330	450	23.56	106	1	352634	1637	genus	Listeria

Appendix 5.1. Sequencing and taxonomic classification metrics for *Listeria* spp. long-read quasi-metagenomic datasets generated in Chapter 5. Performance is compared across two laboratory workflows: no saponin-based eukaryotic DNA depletion (NSD) and saponin depletion (SD). Samples were generated from four food matrices (ham, roast beef, salad kit, turkey) at three spike levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). Taxonomic results are reported for the foodborne pathogen (FBP)-specific database. Sample identifiers include *Listeria* species (LI = *L. innocua*; LM-1–8 = *L. monocytogenes*; LS = *L. seeligeri*; LW = *L. welshimeri*). Negative control (NC) samples were included to establish baseline taxonomic background. Single-isolate (SI) reference genomes were previously sequenced and were not evaluated using the preprocessing workflow developed in this study; these entries are indicated as NA in the table (Continued). FL90 = filtlong filtering retaining best 90% of reads; pcABI = porcechop ABI; ER = Eukaryotic reads (Continued).

Strain	Food	Pathogen	Wet Lab	Spike	Classifier	Database	DNA Concentration	Total Reads	Total Bases	Mean Read Length	Mean Base Quality	Total Reads (FL90)	Percent Filtered (FL90)	Total Reads (FL90+pcABI)	Percent Filtered (FL90+pcABI)	Total Reads (FL90+pcABI+ER)	Eukaryotic Reads	Percent Eukaryotic	Number of Reads Available for Classification	Percent Classified	reads clade	taxReads	kmers	taxID	rank	taxName
LM-5	SALAD_KIT	Listeria monocytogenes	NSD	Low	KrakenUniq	FBP-Specific	13.17	147649	935138323	6333.5	13.4	74360	50.36268447	74982	50.78395384	71865	3117	4.3372991	71865	82.15	59038	1	4E+06	1637	genus	Listeria
LM-5	SALAD_KIT	Listeria monocytogenes	NSD	Medium	KrakenUniq	FBP-Specific	23.073	185890	1535854020	8262.2	14	90813	48.85308516	91009	48.95852386	87353	3656	4.18531705	87353	97.95	85564	0	4E+06	1637	genus	Listeria
LM-5	SALAD_KIT	Listeria monocytogenes	NSD	High	KrakenUniq	FBP-Specific	22.665	144447	1290882473	8936.7	12.9	66058	45.73165244	66190	45.82303544	63327	2863	4.52097841	63327	98.95	62665	0	4E+06	1637	genus	Listeria
LM-5	SALAD_KIT	Listeria monocytogenes	SD	Low	KrakenUniq	FBP-Specific	11.3	117809	892783612	7578.2	13.4	61621	52.305851	62094	52.70734834	59883	2211	3.69219979	59883	79.19	47419	2	4E+06	1637	genus	Listeria
LM-5	SALAD_KIT	Listeria monocytogenes	SD	Medium	KrakenUniq	FBP-Specific	20.6	249756	1717900943	6878.3	13.4	126726	50.73992216	127915	51.2159868	122811	5104	4.15597951	122811	97.74	120035	2	5E+06	1637	genus	Listeria
LM-5	SALAD_KIT	Listeria monocytogenes	SD	High	KrakenUniq	FBP-Specific	22.4	238358	1357914154	5697	13.5	124724	52.32633266	125717	52.7429329	121193	4524	3.73288886	121193	98.85	119796	1	4E+06	1637	genus	Listeria
LM-5	SALAD_KIT	Listeria monocytogenes	SI	SI	KrakenUniq	FBP-Specific	29.2	73063	1482126069	16986	13.8	NA	NA	NA	NA	NA	NA	NA	73237	100	73234	0	5E+06	1637	genus	Listeria
LM-6	SALAD_KIT	Listeria monocytogenes	NSD	Low	KrakenUniq	FBP-Specific	0.112	6179	5723908	926.3	12.2	2188	35.41026056	2181	35.29697362	2102	79	3.7583254	2102	65.98	1387	6	1E+06	1637	genus	Listeria
LM-6	SALAD_KIT	Listeria monocytogenes	NSD	Medium	KrakenUniq	FBP-Specific	16.324	142824	832526547	5829	13.4	57242	40.07869826	57487	40.25023806	55279	2208	3.99428354	55279	97.78	54052	1	5E+06	1637	genus	Listeria
LM-6	SALAD_KIT	Listeria monocytogenes	NSD	High	KrakenUniq	FBP-Specific	21.522	327950	2185646981	6664.6	13.2	160089	48.81506327	161748	49.32093307	154812	6936	4.48027285	154812	99.78	154478	1	5E+06	1637	genus	Listeria
LM-6	SALAD_KIT	Listeria monocytogenes	SD	Low	KrakenUniq	FBP-Specific	7	148788	872067088	5861.1	13.3	73891	49.6619351	73976	49.71906337	71919	2057	2.86016213	71919	82.28	59176	0	4E+06	1637	genus	Listeria
LM-6	SALAD_KIT	Listeria monocytogenes	SD	Medium	KrakenUniq	FBP-Specific	29.6	87481	625814173	7153.7	13.5	40711	46.53696231	41175	47.0673632	39299	1876	4.77365836	39299	98.36	38653	0	4E+06	1637	genus	Listeria
LM-6	SALAD_KIT	Listeria monocytogenes	SD	High	KrakenUniq	FBP-Specific	31.125	181500	1209169794	6662.1	13.7	94073	51.83085399	94253	51.93002755	91276	2977	3.26153644	91276	99.78	91077	1	4E+06	1637	genus	Listeria
LM-6	SALAD_KIT	Listeria monocytogenes	SI	SI	KrakenUniq	FBP-Specific	40	68162	1019477094	12263.9	13.6	NA	NA	NA	NA	NA	NA	NA	68311	99.99	68302	0	4E+06	1637	genus	Listeria
LS-1	SALAD_KIT	Listeria seeligeri	NSD	Low	KrakenUniq	FBP-Specific	6.134	152126	649994600	4272.7	13.7	71770	47.17799719	72403	47.59409963	69869	2534	3.62678727	69869	92.65	64734	59	3E+06	1637	genus	Listeria
LS-1	SALAD_KIT	Listeria seeligeri	NSD	Medium	KrakenUniq	FBP-Specific	20.849	211603	1502132082	7098.8	13.2	112003	52.93072404	113084	53.44158637	108298	4786	4.41928752	108298	97.42	105505	60	3E+06	1637	genus	Listeria
LS-1	SALAD_KIT	Listeria seeligeri	NSD	High	KrakenUniq	FBP-Specific	12.343	195237	1164148412	5962.7	13	92932	47.5995841	93114	47.69280413	89760	3354	3.73663102	89760	98.67	88570	48	3E+06	1637	genus	Listeria
LS-1	SALAD_KIT	Listeria seeligeri	SD	Low	KrakenUniq	FBP-Specific	18.726	135685	870266351	6413.9	13.5	62562	46.10826547	62708	46.21586763	60595	2113	3.48708639	60595	93.99	56952	18	3E+06	1637	genus	Listeria
LS-1	SALAD_KIT	Listeria seeligeri	SD	Medium	KrakenUniq	FBP-Specific	28	265879	2261066695	8504.1	14.1	134618	50.63130221	134895	50.73548494	129740	5155	3.97333128	129740	98.09	127261	26	3E+06	1637	genus	Listeria
LS-1	SALAD_KIT	Listeria seeligeri	SD	High	KrakenUniq	FBP-Specific	35	118670	960193908	8091.3	13.7	56670	47.75427657	57338	48.3171821	53857	3481	6.46341237	53857	99.08	53364	15	2E+06	1637	genus	Listeria
LS-1	SALAD_KIT	Listeria seeligeri	SI	SI	KrakenUniq	FBP-Specific	77.4	81943	1608451005	15991.5	13.7	NA	NA	NA	NA	NA	NA	NA	82143	99.71	81901	22	4E+06	1637	genus	Listeria
NC	SALAD_KIT	NC	NSD	NC	KrakenUniq	FBP-Specific	1.578	96275	290539801	3017.8	12.6	54443	56.54946767	54566	56.67722669	53411	1155	2.16247589	53411	0.4437	237	10	1E+06	1637	genus	Listeria
NC	SALAD_KIT	NC	SD	NC	KrakenUniq	FBP-Specific	2.74	61188	210259690	3436.3	13.8	35331	57.74171406	35385	57.82996666	35049	336	0.95865788	35049	0.291	102	3	419743	1637	genus	Listeria

Appendix 5.1. Sequencing and taxonomic classification metrics for *Listeria* spp. long-read quasi-metagenomic datasets generated in Chapter 5. Performance is compared across two laboratory workflows: no saponin-based eukaryotic DNA depletion (NSD) and saponin depletion (SD). Samples were generated from four food matrices (ham, roast beef, salad kit, turkey) at three spike levels: low (10^1 CFU/mL), medium (10^2 CFU/mL), and high (10^3 CFU/mL). Taxonomic results are reported for the foodborne pathogen (FBP)-specific database. Sample identifiers include *Listeria* species (LI = *L. innocua*; LM-1–8 = *L. monocytogenes*; LS = *L. seeligeri*; LW = *L. welshimeri*). Negative control (NC) samples were included to establish baseline taxonomic background. Single-isolate (SI) reference genomes were previously sequenced and were not evaluated using the preprocessing workflow developed in this study; these entries are indicated as NA in the table (Continued). FL90 = filtlong filtering retaining best 90% of reads; pcABI = porehop A; ER = Eukaryotic reads (Continued).

Strain	Food	Pathogen	Wet Lab	Spike	Classifier	Database	DNA Concentration	Total Reads	Total Bases	Mean Read Length	Mean Base Quality	Total Reads (FL90)	Percent Filtered (FL90)	Total Reads (FL90+pcABI)	Percent Filtered (FL90+pcABI)	Total Reads (FL90+pcABI+ER)	Eukaryotic Reads	Percent Eukaryotic	Number of Reads Available for Classification	Percent Classified	reads clade	taxReads	kmers	taxID	rank	taxName
LM-7	TURKEY	Listeria monocytogenes	NSD	Low	KrakenUniq	FBP-Specific	5.3	18832	181115442	9617.4	13.6	10409	55.27293968	10429	55.37914189	10313	116	1.12479395	10313	99.97	10310	15	3E+06	1637	genus	Listeria
LM-7	TURKEY	Listeria monocytogenes	NSD	Medium	KrakenUniq	FBP-Specific	23.883	50100	544934153	10876.9	13.8	29576	59.03393214	29643	59.16766467	29438	205	0.69637883	29438	99.97	29428	25	4E+06	1637	genus	Listeria
LM-7	TURKEY	Listeria monocytogenes	NSD	High	KrakenUniq	FBP-Specific	21.4	69658	616247353	8846.8	14.1	39525	56.74150851	39525	56.74150851	39198	327	0.83422624	39198	99.89	39156	50	4E+06	1637	genus	Listeria
LM-7	TURKEY	Listeria monocytogenes	SD	Low	KrakenUniq	FBP-Specific	15.845	27095	277124972	10227.9	13.6	14144	52.20151319	14144	52.20151319	14112	32	0.22675737	14112	99.81	14085	11	4E+06	1637	genus	Listeria
LM-7	TURKEY	Listeria monocytogenes	SD	Medium	KrakenUniq	FBP-Specific	64.685	56182	700374486	12466.2	13.3	33875	60.29511231	33875	60.29511231	33795	80	0.23672141	33795	99.98	33787	35	4E+06	1637	genus	Listeria
LM-7	TURKEY	Listeria monocytogenes	SD	High	KrakenUniq	FBP-Specific	51.12	51718	602986122	11659.1	12.9	27885	53.9173982	27885	53.9173982	27752	133	0.47924474	27752	99.91	27728	16	4E+06	1637	genus	Listeria
LM-7	TURKEY	Listeria monocytogenes	SI	SI	KrakenUniq	FBP-Specific	28.4	54823	704641142	12853	13	NA	NA	NA	NA	NA	NA	NA	30280	100	30280	9	4E+06	1637	genus	Listeria
LM-8	TURKEY	Listeria monocytogenes	NSD	Low	KrakenUniq	FBP-Specific	7.156	33876	324754729	9586.6	13.9	19604	57.86987838	19604	57.86987838	19275	329	1.70687419	19275	99.98	19272	28	3E+06	1637	genus	Listeria
LM-8	TURKEY	Listeria monocytogenes	NSD	Medium	KrakenUniq	FBP-Specific	9.72	46527	578640527	12436.7	13.4	28254	60.72603005	28290	60.80340447	28101	189	0.67257393	28101	99.92	28078	40	4E+06	1637	genus	Listeria
LM-8	TURKEY	Listeria monocytogenes	NSD	High	KrakenUniq	FBP-Specific	32.6	38641	511622855	13240.4	13.4	21796	56.4064077	22017	56.97833907	21864	153	0.69978046	21864	99.94	21850	12	4E+06	1637	genus	Listeria
LM-8	TURKEY	Listeria monocytogenes	SD	Low	KrakenUniq	FBP-Specific	46.263	48480	481860193	9939.4	13.5	27436	56.59240924	27436	56.59240924	27378	58	0.21184893	27378	99.96	27366	39	4E+06	1637	genus	Listeria
LM-8	TURKEY	Listeria monocytogenes	SD	Medium	KrakenUniq	FBP-Specific	52.979	28097	350640627	12479.6	13.5	17974	63.97124248	18108	64.44816173	18046	62	0.34356644	18046	99.94	18036	35	4E+06	1637	genus	Listeria
LM-8	TURKEY	Listeria monocytogenes	SD	High	KrakenUniq	FBP-Specific	40.721	29622	513001175	17318.2	13.5	18448	62.27803659	18448	62.27803659	18345	103	0.56146089	18345	99.97	18339	13	4E+06	1637	genus	Listeria
LM-8	TURKEY	Listeria monocytogenes	SI	SI	KrakenUniq	FBP-Specific	24	88556	1150773139	12994.9	13	NA	NA	NA	NA	NA	NA	NA	48279	100	48279	21	4E+06	1637	genus	Listeria
LW-1	TURKEY	Listeria welshimeri	NSD	Low	KrakenUniq	FBP-Specific	5.04	43093	394539961	9155.5	13.5	25622	59.45745249	25622	59.45745249	25205	417	1.65443364	25205	99.85	25168	8	3E+06	1637	genus	Listeria
LW-1	TURKEY	Listeria welshimeri	NSD	Medium	KrakenUniq	FBP-Specific	6.82	42070	478123775	11365	13.4	24091	57.26408367	24091	57.26408367	23884	207	0.866689	23884	99.91	23863	5	3E+06	1637	genus	Listeria
LW-1	TURKEY	Listeria welshimeri	NSD	High	KrakenUniq	FBP-Specific	18.7	75075	840601336	11196.8	13.5	46255	61.61172161	46312	61.68764569	45875	437	0.95258856	45875	99.88	45819	17	3E+06	1637	genus	Listeria
LW-1	TURKEY	Listeria welshimeri	SD	Low	KrakenUniq	FBP-Specific	16.346	27576	391944888	14213.3	13.5	15549	56.38598782	15562	56.43313026	15489	73	0.47130221	15489	99.97	15484	1	3E+06	1637	genus	Listeria
LW-1	TURKEY	Listeria welshimeri	SD	Medium	KrakenUniq	FBP-Specific	39.837	37415	441548815	11801.4	13	18774	50.1777362	18802	50.2525725	18749	53	0.28268174	18749	99.77	18706	3	3E+06	1637	genus	Listeria
LW-1	TURKEY	Listeria welshimeri	SD	High	KrakenUniq	FBP-Specific	49.551	71277	929508147	13040.8	13.2	42126	59.10181405	42179	59.17617184	42018	161	0.38316912	42018	99.88	41969	7	3E+06	1637	genus	Listeria
LW-1	TURKEY	Listeria welshimeri	SI	SI	KrakenUniq	FBP-Specific	40.4	44604	833896627	15556.3	13.6	NA	NA	NA	NA	NA	NA	NA	44776	99.91	44734	13	4E+06	1637	genus	Listeria
NC	TURKEY	NC	NSD	NC	KrakenUniq	FBP-Specific	0	5349	4494715	840.3	13	1787	33.40811367	1787	33.40811367	120	1667	1389.16667	120	50	60	0	253417	1637	genus	Listeria
NC	TURKEY	NC	SD	NC	KrakenUniq	FBP-Specific	0.03	7970	2568388	322.3	12.5	3744	46.9761606	1376	17.26474279	1339	37	2.76325616	1339	11.65	156	1	671327	1637	genus	Listeria