Hunting down what makes good bacteria bad company

A study investigating the genetic and phenotypic differences between *Limosilactobacillus fermentum* AGR1485 and AGR1487 to elucidate the molecular basis underlying *L. fermentum* AGR1487's barrier-disruptive phenotype.

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Massey University,

Manawatū,

New Zealand

Marc Alex Bailie

2025

Abstract

Background: Bacteria play crucial roles in human society, from food production to natural symbiosis. Among these, probiotic species such as *Limosilactobacillus fermentum* are widely recognised for their health benefits, although not all strains exhibit equivalent properties. Notably, strain AGR1487 of *L. fermentum* has been identified as an outlier, exhibiting a barrier-disruptive phenotype in an *in vitro* model of the intestinal epithelial barrier. Disruption of this barrier is clinically relevant, as it can increase intestinal permeability, facilitate translocation of luminal antigens or pathogens, and contribute to gastrointestinal disorders including inflammatory bowel disease and irritable bowel syndrome. The *in vitro* model employs Caco-2 cells, derived from human colorectal adenocarcinoma, which differentiate into enterocyte-like cells, and measures transepithelial electrical resistance (TEER) to assess barrier function. In contrast, strain AGR1485, isolated from the same human population, does not display a barrier-disruptive phenotype. This study investigates the genetic and phenotypic differences between AGR1485 and AGR1487 to elucidate the molecular basis underlying the barrier-disruptive phenotype.

Methods: A variety of methods, including a multi-omics approach were employed to analyse the genetic and phenotypic differences between AGR1485 and AGR1487. The barrier-disruptive phenotype was confirmed through TEER assays. Growth characteristics were assessed using spectrophotometry, and bacterial morphology was examined via electron microscopy. Genomic sequencing was performed using hybrid assembly approaches, combining Illumina short-read and PacBio long-read sequencing. Comparative genomic and transcriptomic analyses of both strains were conducted to identify non-identical genes and differential gene expression patterns.

Results: TEER assays demonstrated that AGR1487 significantly disrupts intestinal barrier integrity (-20.11 % TEER per hour), whereas AGR1485 had minimal effects. Electron microscopy revealed AGR1487's tendency to autoaggregate, a feature absent in AGR1485. Both genomes were sequenced, and the assemblies were of exceptionally high quality, with BUSCO

completeness scores of 99.5 % (98.8 % single-copy, 0.7 % duplicated, 0.5 % fragmented, 0 % missing), confirming their suitability for detailed comparative and functional analyses. Comparative genomic analysis showed that AGR1487 possesses 62 non-identical genes within regions of high genomic variability, suggesting horizontal gene transfer events and diversification or loss of these genes in other strains. Transcriptomic profiling indicated that AGR1487 expresses these non-identical genes before TEER assays, with 25 candidate genes potentially implicated in its barrier-disruptive properties. Among them, a phage-associated truncated metallopeptidase and mucin-binding protein were identified as strong candidates.

Conclusion: The findings of this PhD thesis highlight strain-specific variations within *L. fermentum*, emphasising the necessity of evaluating probiotics at the strain level. The identification of genetic elements responsible for the barrier-disruptive phenotype in AGR1487 underscores the complexity of bacterial-host interactions and the need for careful screening in probiotic applications. Future research will focus on the functional validation of candidate genes through knockout and complementation studies to confirm their roles in gut barrier integrity disruption.

Acknowledgements

A wholehearted thank you to my supervisors for their endless patience and guidance.

Prof. Nicole Roy, thank you for encouraging and guiding me. Your editorial superpowers and work ethic are an inspiration that will continue to guide me beyond this PhD.

Dr. Eric Altermann, your direction and friendship fuelled my progress throughout this journey. Your thought-provoking questions and insightful feedback often led to hours of internal debate and further research. Moreover, thank you for always being direct, for your willingness to teach, and for engaging in discussions that encouraged me to explore new methods and strategies.

Dr. Jon Palmer, thank you for always knowing the next step. You seemed to have an uncanny ability to say exactly what I needed to hear, whether it was to push through writer's block or move forward to the next phase.

A heartfelt thank you to Dr. Wayne Young for knowing where to look and how to find the best solutions for coding issues, scripting challenges, and workflow optimisations.

Lastly, thank you, Prof. Warren McNabb, for your well-timed encouragement and for those moments of casual conversation in your office. Those discussions helped diffuse stress and provided much-needed moments of respite. Furthermore, thank you for throwing me into the deep end, pushing me to do uncomfortable things like public speaking, while also providing the support that turned those challenges into successes.

To my wife and my life, thank you, Kym. You have been my rock, steady and gracious throughout this journey. Without you and the girls, this entire project would have been impossible. You have been my joy and sunshine, yet you also allowed me to disappear into my work for days without complaint. I am endlessly grateful.

I extend my gratitude to the Tertiary Education Commission for funding this PhD project through the Riddet Institute Centre of Research Excellence, Massey University, and AgResearch. I would also like to thank the Riddet Institute for being a welcoming and inclusive home

throughout my studies. The student colloquia, social events, and presentation opportunities allowed me to flourish. Additionally, I appreciate Massey University and AgResearch for providing both the facilities and an on-site home for my research.

Dr. Jane Mullaney and Dr. Alicia Barnett, thank you for your mentorship, constant companionship, and stimulating debates about everything. Dr. Jane, I will miss the hard questions, our thorough conversations, and your strength and conviction. Dr. Alicia, thank you for sharing your lab space, your time, and your knowledge.

To the many friends I made at the Riddet Institute, thank you. The laughter, the hairbrained schemes, and the outrageously silly ideas we convinced ourselves were true will never be forgotten. These friendships are an unexpected yet invaluable reward from my time at the Institute.

Lastly, I would like to thank the Manawatū Imaging Centre and Massey University's Sequencing Service for their expertise, access to facilities, and assistance with the imaging, microscopy, and sequencing work that contributed to this thesis.

Contents

List of Figures

List of Tables

Abbreviations

AGR1485 / AGR1487 – Strain designations for *Limosilactobacillus fermentum*

ABySS – Assembly By Short Sequences

ANI – Average Nucleotide Identity

AT – Adenine-Thymine

ATCC – American Type Culture Collection

BAM – Binary Alignment Map (file format)

Bandage – Bioinformatics Assembly Graph Toolkit

BEDTools – Suite of utilities for genome analysis

BLAST – Basic Local Alignment Search Tool

bp – Base Pairs

BUSCO – Benchmarking Universal Single-Copy Orthologs

BWA – Burrows-Wheeler Aligner

Canu – Long-read genome assembler

Cas9 – CRISPR-associated protein 9

CECT – Colección Española de Cultivos Tipo (Spanish Type Culture Collection)

CheckM – Tool for genome quality assessment

CI – Confidence Interval

CLR – C-type Lectin Receptor

CO_2 – Carbon Dioxide

Circos – Visualisation tool for genomic data

CRISPR – Clustered Regularly Interspaced Short Palindromic Repeats

Cas – CRISPR-associated systems

DMEM – Dulbecco's Modified Eagle's Media

DNA – Deoxyribonucleic Acid

DSM – Deutsche Sammlung von Mikroorganismen und Zellkulturen

DUF – Domain of Unknown Function

EdgeR – R package for differential gene expression analysis

EPS – Exopolysaccharides

EVOM – Epithelial Volt/Ohm Meter

FBS – Foetal Bovine Serum

FASTQ – File format for sequence data

FDR – False Discovery Rate

GC – Guanine-Cytosine (content)

GC Content – Guanine-Cytosine percentage in a genome

GC-skew – Guanine-Cytosine skew

GIT – Gastrointestinal Tract

GlimmerHMM – Gene prediction tool

GRAS – Generally Recognised As Safe

GUI – Graphical User Interface

HGT – Horizontal Gene Transfer

hisat2 – Alignment program for mapping RNA-seq reads

HMM – Hidden Markov Model

HTB-37 – The ATCC designation for the Caco-2 cell line

ICEs – Integrative and Conjugative Elements

IFN-γ – Interferon Gamma

IgA – Immunoglobulin A

IL-1β – Interleukin-1 beta

IL-4 – Interleukin-4

IL-6 – Interleukin-6

IL-8 – Interleukin-8

IL-10 – Interleukin-10

Illumina NovaSeq 6000 – Sequencing platform

in vitro – (Latin: “in glass”; experiments conducted outside living organisms)

in vivo – (Latin: “in life”; experiments conducted within living organisms)

ISO – International Organisation for Standardization

L. fermentum – *Limosilactobacillus fermentum*

LAB – Lactic Acid Bacteria

LCBs – Locally Collinear Blocks

LTA – Lipoteichoic Acid

LPXTG – Cell Wall Anchor Domain Motif

LTS – Long-Term Support (Ubuntu Linux)

M199 – Media 199

MaSuRCA – Maryland Super-Read Celera Assembler

Mb – Megabases (a measure of genome size)

ME-3 – Specific strain of L. fermentum

MEGAHIT – Metagenomic assembler

MEM – Minimal Essential Media

Metassembler – Hybrid genome assembly tool

MAMPs – Microbe-Associated Molecular Patterns

MAPK – Mitogen-Activated Protein Kinase

MCP-1 – Monocyte Chemoattractant Protein-1

mmseq2 – Sequence alignment and clustering tool

MMP-1 / MMP-3 – Matrix Metalloproteinases 1 and 3

MucBP – Mucin Binding Protein

MRS – de Man, Rogosa, and Sharpe (growth medium)

MTCC – Microbial Type Culture Collection

MurNAc – N-acetylmuramic acid

N – Ambiguous base in genome assemblies

NCBI – National Center for Biotechnology Information

NCBI RefSeq – NCBI Reference Sequence Database

NEAA – Non-Essential Amino Acids

NF-κB – Nuclear Factor Kappa-Light-Chain-Enhancer of Activated B Cells

NGS – Next-Generation Sequencing

N.J. – Neighbour-Joining

NLRs – NOD-Like Receptors

NOD – Nucleotide-binding Oligomerisation Domain

NZ – New Zealand

OLC – Overlapping Layout Consensus

OD600 – Optical Density at 600 nm

ORFs – Open Reading Frames

PAdj – Adjusted P-value

PacBio – Pacific Biosciences (sequencing technology)

PAMPs – Pathogen-Associated Molecular Patterns

PCR – Polymerase Chain Reaction

PCC – Probiotic Candidate Culture

PFGE – Pulse-Field Gel Electrophoresis

PG – Peptidoglycan

PGAP – Prokaryotic Genome Annotation Pipeline

Pilon – Genome assembly polisher

PRRs – Pattern Recognition Receptors

qCML – Quantile-Conditional Maximum Likelihood (statistical method)

QUAST – Quality Assessment Tool for Genome Assemblies

Racon – Consensus sequence polisher

Ray – Genome assembler

RBH – Reciprocal Best-Hit

REAPR – Read Error and Assembly Parameter Refinement

RNA – Ribonucleic Acid

RNA-seq – RNA sequencing

RIN – RNA Integrity Number

ROARY – A rapid large-scale prokaryote pan-genome analysis tool

RPM – Revolutions Per Minute

SAM – S-Adenosyl Methionine

SAMtools – Sequence Alignment/Map tools

SDS – Sodium Dodecyl Sulfate

SNPs – Single Nucleotide Polymorphisms

SOAPdenovo – Short Oligonucleotide Analysis Package

SPAdes – St. Petersburg genome assembler

SMRT – Single-Molecule Real-Time (sequencing)

Spp. – Abbreviation for “species” (plural)

TEER – Transepithelial Electrical Resistance

TGF-β – Transforming Growth Factor Beta

TLRs – Toll-Like Receptors

TNF-α – Tumour Necrosis Factor Alpha

TnpB – Transposase B

Trimmomatic – Tool for trimming Illumina sequence reads

UGENE – Bioinformatics software

UPGMA – Unweighted Pair Group Method with Arithmetic Mean

USA – United States of America

Velvet – Short-read genome assembler

WGS – Whole-Genome Sequencing

Note to the reader

The data and scripts to support this thesis can be found in the digital appendix under the following link:

https://drive.google.com/drive/folders/1g454dGCx3WajDMRYtdMnhaQ52_FMLgzd?usp=sharing

Thesis specific definitions

In this study, "unique" or "non-identical" refers to AGR1487 genes classified as cloud genes by Roary pan-genome analysis at a 95% sequence identity threshold. Cloud genes are present in ≤15% of strains analysed (≤6 of 40 L. fermentum genomes), representing the rare portion of the pan-genome. A gene is considered "unique" to AGR1487 if it was identified as a cloud gene by Roary and subsequently assigned to AGR1487 through BLASTP filtering against a custom AGR1487 database. This approach identifies genes that are either: (1) absent or below 95% identity in other strains, or (2) present in only a small minority of strains. These genes are "unique" because they represent rare genetic content within the *L. fermentum* pan-genome that is present in AGR1487. Subsequent BLASTP validation against full-length Prokka-annotated proteins confirmed that most cloud genes have near-identical homologs (96-100% identity) in at least one other strain but remain rare across the species. Only three genes (RS02170, RS05945, RS09860) had no detectable homolog in any other strain.

Limosilactobacillus fermentum AGR1485: A companion strain isolated from the same human population and during the same sampling experiment that yielded AGR1487. AGR1485 served as a negative control and as a canonical representative of the broader, expected behaviour of the L. fermentum species.

Chapter 1: General Introduction

1.1. Introduction

Bacteria are integral to human society, playing crucial roles in food production and natural symbiosis. Over the years, humans have selectively consumed specific bacteria for their health benefits, particularly those recognised as probiotics. Probiotics are defined as "live microorganisms which when administered in adequate amounts confer a health benefit on the host" (Bernardeau, Vernoux, Henri-Dubernet, & Gueguen, 2008; Bron, van Baarlen, & Kleerebezem, 2011; Hill et al., 2014; Lebeer, Vanderleyden, & De Keersmaecker, 2008; Marco, Pavan, & Kleerebezem, 2006). It is also general knowledge that some species of bacteria are pathogenic, and the distinction between beneficial and harmful bacteria is not always straightforward.

Small genetic variations, such as differences in gene composition or the presence or absence of specific genes, can differentiate bacterial strains, even within the same species. Recent research has shown that these minor genetic differences can significantly impact the effects (positive or negative) a bacterium has on a human host (Zeevi et al., 2019). As such, evaluating bacteria at the strain level is crucial when assessing their safety and potential benefits.

The phylum Firmicutes contains more than 200 species of bacteria that are widely employed in biotechnology and fermented food preservation (Oren & Garrity, 2021; Salvetti, Harris, Felis, & Toole, 2018). These bacteria are under intense study due to their considerable economic impact and involvement in food fermentation and spoilage, biotechnology, and therapeutics (Chaillou et al., 2005; Cotter, Ross, & Hill, 2013; Ganzle & Ripari, 2016; Pot et al., 2013; Stefanovic, Fitzgerald, & McAuliffe, 2017).

Advancements in genetic analyses have recently led to the reclassification of Firmicutes as Bacillota and the establishment of new genera, including *Limosilactobacillus* (Ksiezarek, Grosso,

Ribeiro, & Peixe, 2022; Pallen, 2023). Among the members of Bacillota is the species *Limosilactobacillus fermentum*, a species classified as 'generally recognised as safe' (GRAS). This species is widely utilised as a therapeutic agent, an acidifying starter culture, and a probiotic (Fuochi, Li Volti, & M Furneri, 2017; Verce, Vuyst, & Weckx, 2018; Zheng et al., 2020). The popularity of *L. fermentum* has been steadily increasing in recent years, primarily due to its commercial significance and potential in food production and therapeutic applications (Naghmouchi et al., 2020; Zhao, Zhong, Liu, Wang, & Gao, 2021). However, it is important to note that not all strains of *L. fermentum* strains are equally beneficial for human consumption, with strain-level variability playing a critical role in their effectiveness and safety (Anderson et al., 2016; Sengupta et al., 2015).

L. fermentum is an irregular, non-sporing, Gram-positive, rod-shaped bacterium (McClung, 1987). It is a gas-producing, heterofermentative, air-tolerant, obligate anaerobe that produces equimolar amounts of lactic acid, carbon dioxide, and ethanol from the metabolism of hexoses (Adegoke, 2004). This metabolic profile arises because *L. fermentum* lacks the majority of the citric acid cycle and quinone/ubiquinone biosynthesis pathways, rendering it incapable of respiration (Illeghems, De Vuyst, & Weckx, 2015). Despite its preference for anaerobic conditions, the species is oxygen-tolerant, making it commercially adaptable. *L. fermentum* has been isolated from diverse environments, including spontaneously fermenting plant-based food systems and the human gastrointestinal tract (Duar et al., 2017; Ganzle & Ripari, 2016; Hammes & Hertel, 2006; Mundt & Hammer, 1968).

Interestingly, the properties of *L. fermentum* strains cannot always be predicted based on their source of isolation (Verce, De Vuyst, & Weckx, 2020). For example, the strains *L. fermentum* AGR1485 and AGR1487 were both isolated from the oral cavity of different healthy humans but displayed contrasting phenotypes (Sengupta et al., 2015). Strain AGR1487 exhibited greater bile acid tolerance and a broader pH range permissiveness than AGR1485. Additionally, AGR1487's genome is approximately 200 kbp smaller, a feature often associated with host adaptation to the

gastrointestinal tract (Sengupta et al., 2015). However, AGR1487 also disrupts gastrointestinal barrier integrity in model systems by uniformly degrading the tight junctions between Caco-2 cells, and it shows greater permissiveness to mannitol passage across epithelial barriers, indicating that small molecules can bypass the cells via paracellular pathways rather than passing through them (Rachel C Anderson et al., 2013). While AGR1487 demonstrates several host-adaptive traits, it also induces heightened inflammatory responses in germ-free mice, a phenotype associated with gastrointestinal disease (Anderson et al., 2016; Sengupta et al., 2015; Vancamelbeke & Vermeire, 2017). Notably, Sengupta's work demonstrated that AGR1487's barrier-disruptive phenotype is retained even when using heat-killed cells or purified cell wall extracts, suggesting that this effect is mediated by structural components rather than active metabolism (Sengupta et al., 2015). The genetic factors underlying this barrier disruption remain unknown.

These findings underscore the importance of strain-level diversity in shaping human health impacts. Identifying the genes or genetic elements responsible for these variations is vital, as it could improve probiotic development and lead to pre-screening systems for pathogenic markers in GRAS species.

The concept of the "bacterial theatre of activity" provides a framework for understanding bacterial interactions with their environment. This theatre encompasses microbial structures, macromolecules, metabolites, mobile genetic elements, and environmental conditions (Berg et al., 2020). Despite significant advancements, the human gastrointestinal tract microbiome remains poorly characterised, with many species and functional traits still unknown (Qin et al., 2010). Despite much progress over the last ten years, many species and functional diversity of the bacteria found in the human gastrointestinal tract (GIT) remain uncharacterised (Almeida et al., 2021). Overcoming these gaps requires complex analyses such as multi-omics and time-series studies rather than simple species-phenotype correlations (Cani, 2018).

This research employs a variety of methodologies and tools that include a multi-omics approach to identify gene candidates responsible for *L. fermentum* AGR1487's ability to disrupt the intestinal epithelial barrier compared to AGR1485. These insights could redefine how we view and utilise bacteria in probiotics, therapeutics, and food production.

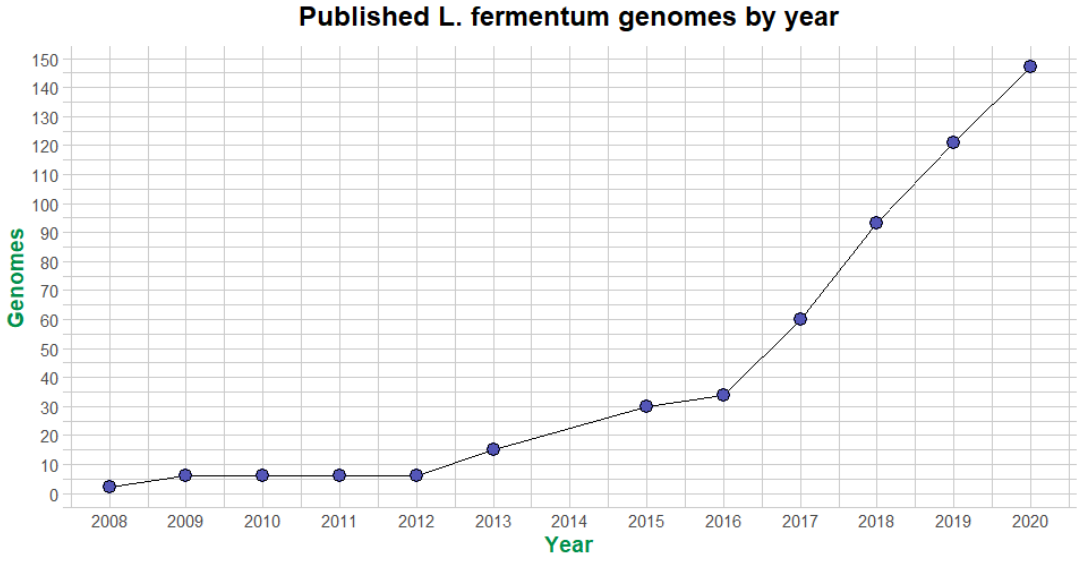

Figure 1.1-1: The number of *L. fermentum* genomes published by year. The accumulative number of genomes available to download from NCBI by year. Data was accessed on 9-12-2021 at 11:14 am.

1.2. References

Adegoke, G. (2004). *Understanding food microbiology* (2nd ed.). Alleluia Ventures.

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., . . . Finn, R. D. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*, 39(1), 105–114. https://doi.org/10.1038/s41587-020-0603-3

Anderson, R. C., Ulluwishewa, D., Young, W., Ryan, L. J., Henderson, G., Meijerink, M., . . . Roy, N. C. (2016). Human oral isolate *Lactobacillus fermentum* AGR1487 induces a pro-inflammatory response in germ-free rat colons. Scientific Reports, 6, 20318. https://doi.org/10.1038/srep20318

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Verges, M. C., Charles, T., . . . Schloter, M. (2020). Microbiome definition re-visited: Old concepts and new challenges. *Microbiome, 8*(1), 103. https://doi.org/10.1186/s40168-020-00875-0

Bernardeau, M., Vernoux, J. P., Henri-Dubernet, S., & Gueguen, M. (2008). Safety assessment of dairy microorganisms: The *Lactobacillus* genus. *International Journal of Food Microbiology, 126*(3), 278–285. https://doi.org/10.1016/j.ijfoodmicro.2007.08.015

Bron, P. A., van Baarlen, P., & Kleerebezem, M. (2011). Emerging molecular insights into the interaction between probiotics and the host intestinal mucosa. *Nature Reviews Microbiology, 10*(1), 66–78. https://doi.org/10.1038/nrmicro2690

Cani, P. D. (2018). Human gut microbiome: Hopes, threats, and promises. *Gut, 67*(9), 1716–1725. https://doi.org/10.1136/gutjnl-2018-316723

Chaillou, S., Champomier-Verges, M. C., Cornet, M., Crutz-Le Coq, A. M., Dudez, A. M., Martin, V., . . . Zagorec, M. (2005). The complete genome sequence of the meat-borne lactic acid bacterium *Lactobacillus sakei* 23K. *Nature Biotechnology, 23*(12), 1527–1533. https://doi.org/10.1038/nbt1160

Cotter, P. D., Ross, R. P., & Hill, C. (2013). Bacteriocins – a viable alternative to antibiotics? *Nature Reviews Microbiology, 11*(2), 95–105. https://doi.org/10.1038/nrmicro2937

Duar, R. M., Lin, X. B., Zheng, J., Martino, M. E., Grenier, T., Perez-Munoz, M. E., . . . Walter, J. (2017). Lifestyles in transition: Evolution and natural history of the genus *Lactobacillus*. *FEMS Microbiology Reviews, 41*(Supp_1), S27–S48. https://doi.org/10.1093/femsre/fux030

Fuochi, V., Li Volti, G., & Furneri, P. M. (2017). Probiotic properties of *Lactobacillus fermentum* strains isolated from human oral samples and description of their antibacterial activity. *Current Pharmaceutical Biotechnology, 18*(2), 138–149.

Gänzle, M., & Ripari, V. (2016). Composition and function of sourdough microbiota: From ecological theory to bread quality. *International Journal of Food Microbiology, 239*, 19–25. https://doi.org/10.1016/j.ijfoodmicro.2016.05.004

Hammes, W. P., & Hertel, C. (2006). The genera *Lactobacillus* and *Carnobacterium*. In M. Dworkin, S. Falkow, E. Rosenberg, K. H. Schleifer, & E. Stackebrandt (Eds.), *The prokaryotes* (pp. 320–403). Springer US.

Hill, C., Guarner, F., Reid, G., Gibson, G. R., Merenstein, D. J., Pot, B., . . . Sanders, M. E. (2014). Expert consensus document: The International Scientific Association for Probiotics and Prebiotics consensus statement on the scope and appropriate use of the term probiotic. *Nature Reviews Gastroenterology & Hepatology, 11*(8), 506–514. https://doi.org/10.1038/nrgastro.2014.66

Illeghems, K., De Vuyst, L., & Weckx, S. (2015). Comparative genome analysis of the candidate functional starter culture strains *Lactobacillus fermentum* 222 and *Lactobacillus plantarum* 80 for controlled cocoa bean fermentation processes. *BMC Genomics, 16*, 766. https://doi.org/10.1186/s12864-015-1927-0

Ksiezarek, M., Grosso, F., Ribeiro, T. G., & Peixe, L. (2022). Genomic diversity of genus *Limosilactobacillus*. *Microbial Genomics, 8*(7). https://doi.org/10.1099/mgen.0.000847

Lebeer, S., Vanderleyden, J., & De Keersmaecker, S. C. (2008). Genes and molecules of lactobacilli supporting probiotic action. *Microbiology and Molecular Biology Reviews, 72*(4), 728–764. https://doi.org/10.1128/MMBR.00017-08

Marco, M. L., Pavan, S., & Kleerebezem, M. (2006). Towards understanding molecular modes of probiotic action. *Current Opinion in Biotechnology, 17*(2), 204–210. https://doi.org/10.1016/j.copbio.2006.02.005

McClung, L. S. (1987). Review of *Bergey's Manual of Systematic Bacteriology, Volume 2* (P. H. A. Sneath, N. S. Mair, & M. E. Sharpe, Eds.). *International Journal of Systematic Bacteriology, 37*(1), 91. https://doi.org/10.1099/00207713-37-1-91

Mundt, J. O., & Hammer, J. L. (1968). Lactobacilli on plants. *Applied and Environmental Microbiology, 16*(9), 1326–1330.

Naghmouchi, K., Belguesmia, Y., Bendali, F., Spano, G., Seal, B. S., & Drider, D. (2020). *Lactobacillus fermentum*: A bacterial species with potential for food preservation and biomedical applications. *Critical Reviews in Food Science and Nutrition, 60*(20), 3387–3399. https://doi.org/10.1080/10408398.2019.1688250

Oren, A., & Garrity, G. M. (2021). Valid publication of the names of forty-two phyla of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology, 71*(10), 005056.

Pallen, M. J. (2023). Request for an opinion on the standing and retention of *Firmicutes* as a phylum name. *International Journal of Systematic and Evolutionary Microbiology, 73*(7), 005933.

Pot, B., Felis, G., De Bruyne, K., Tsakalidou, E., Papadimitriou, K., Leisner, J., & Vandamme, P. (2013). Lactic acid bacteria: Biodiversity and taxonomy. In *Lactic Acid Bacteria* (pp. 249–353).

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., . . . MetaHIT Consortium. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature, 464*(7285), 59–65. https://doi.org/10.1038/nature08821

Salvetti, E., Harris, H. M. B., Felis, G. E., & Toole, P. W. (2018). Comparative genomics of the genus *Lactobacillus* reveals robust phylogroups that provide the basis for reclassification. *Applied and Environmental Microbiology, 84*(17), e00993-00918. https://doi.org/10.1128/AEM.00993-18

Sengupta, R., Anderson, R. C., Altermann, E., McNabb, W. C., Ganesh, S., Armstrong, K. M., . . . Roy, N. C. (2015). *Lactobacillus fermentum* AGR1487 cell surface structures and supernatant increase paracellular permeability through different pathways. *MicrobiologyOpen, 4*(4), 541–552. https://doi.org/10.1002/mbo3.260. Epub 2015 May 6.

Stefanovic, E., Fitzgerald, G., & McAuliffe, O. (2017). Advances in the genomics and metabolomics of dairy lactobacilli: A review. *Food Microbiology, 61*, 33–49. https://doi.org/10.1016/j.fm.2016.08.009

Vancamelbeke, M., & Vermeire, S. (2017). The intestinal barrier: A fundamental role in health and disease. *Expert Review of Gastroenterology & Hepatology, 11*(9), 821–834. https://doi.org/10.1080/17474124.2017.1343143

Verce, M., De Vuyst, L., & Weckx, S. (2020). Comparative genomics of *Lactobacillus fermentum* suggests a free-living lifestyle of this lactic acid bacterial species. *Food Microbiology, 89*, 103448. https://doi.org/10.1016/j.fm.2020.103448

Verce, M., Vuyst, L. D., & Weckx, S. (2018). Complete and annotated genome sequence of the sourdough lactic acid bacterium *Lactobacillus fermentum* IMDO 130101. *Genome Announcements, 6*(19), e00256-00218. https://doi.org/10.1128/genomeA.00256-18

Zeevi, D., Korem, T., Godneva, A., Bar, N., Kurilshikov, A., Lotan-Pompan, M., . . . Segal, E. (2019). Structural variation in the gut microbiome associates with host health. *Nature, 572*(7771), 327–332. https://doi.org/10.1038/s41586-019-1065-y

Zhao, X., Zhong, X., Liu, X., Wang, X., & Gao, X. (2021). Therapeutic and improving function of lactobacilli in the prevention and treatment of cardiovascular-related diseases: A novel perspective from gut microbiota. *Frontiers in Nutrition, 8*, 693412

Zheng, J., Wittouck, S., Salvetti, E., Franz, C. M. A. P., Harris, H. M. B., Mattarelli, P., . . . Lebeer, S. (2020). A taxonomic note on the genus *Lactobacillus*: Description of 23 novel genera, emended description of the genus *Lactobacillus* Beijerinck 1901, and union of *Lactobacillaceae* and *Leuconostocaceae*. *International Journal of Systematic and Evolutionary Microbiology, 70*(4), 2782–2858. https://doi.org/10.1099/ijsem.0.004107

Chapter 2: Literature review

2.1. Introduction

Microorganisms have long been recognised for their critical roles in human biology, impacting numerous aspects of development, health and disease (Rackaityte & Lynch, 2020; Young, 2017). In recent decades, a growing body of research has shifted our understanding of the human body from being a collection of human cells to including a complex ecosystem inhabited by vast communities of microorganisms. These communities of microorganisms are collectively known as the microbiome (Proctor et al., 2019). The microbiome, comprised of fungi, bacteria, archaea, and viruses, inhabits various niches within the human body, such as the skin, oral cavity, and, most notably, the gastrointestinal tract (GIT) (Dekaboruah, Suryavanshi, Chettri, & Verma, 2020; Proctor et al., 2019). The human microbiome plays a central role in many physiological processes, from digestion and immune regulation to metabolic and neurological functions (Bull & Plummer, 2014; Gwak & Chang, 2021; Oliphant & Allen-Vercoe, 2019).

The significance of the gut microbiome has become increasingly clear, with studies highlighting its role in maintaining health, preventing disease, and even influencing the host's gene expression (Nichols & Davenport, 2021). Among the diverse microbial populations within the human microbiome, bacteria represent the most abundant and functionally versatile group (Walter, 2008). Within this group, the *Limosilactobacillus* (previously part of the *Lactobacillus*) genus stands out due to its widespread presence in the GIT, as well as its long history of association with food fermentation and probiotic use (Ibrahim, 2016; Owusu-Kwarteng, Tano-Debrah, Akabanda, & Jespersen, 2015; Salvetti, Harris, Felis, & Toole, 2018). *Limosilactobacillus* species are considered beneficial to human health, promoting intestinal homeostasis, enhancing immune responses, and acting as a defence against pathogenic organisms (Lacerda et al., 2022; Paulino do Nascimento, Lacerda, Ferreira, de Souza, & de Brito Alves, 2022).

One specific species within the *Limosilactobacillus* genus, *Limosilactobacillus fermentum*, has garnered attention for its unique properties and its potential health benefits (Lacerda et al., 2022; Paulino do Nascimento et al., 2022). *L. fermentum* is commonly found in fermented foods and is known for its probiotic capabilities (Lacerda et al., 2022; Naghmouchi et al., 2020a; Paulino do Nascimento et al., 2022). Although less abundant in the human GIT compared to other species, *L. fermentum* plays an important role in modulating the intestinal environment, strengthening the epithelial barrier, and influencing immune responses (De Gregorio et al., 2023; Peng et al., 2020). Understanding *L. fermentum*'s interactions with the host, its potential probiotic benefits and its role in food systems provides valuable insights into both human health and microbial ecology. This species is particularly interesting for its strain-level diversity and its significant role in human health and food science (Y. Zhao et al., 2022).

2.2. Taxonomy and Genomics of *L. fermentum*

The taxonomy and genomic structure of *L. fermentum* have undergone significant revisions due to advancements in whole-genome sequencing (WGS) technologies (Belova et al., 2016). These innovations have provided deeper insights into the evolutionary history, genetic diversity, and functional capabilities of *Lactobacillus* species (Mendes-Soares, Suzuki, Hickey, & Forney, 2014). The most notable shift in recent years is the reclassification of many species within the *Lactobacillus* genus, including *L. fermentum*, into a newly defined genus, *Limosilactobacillus*, based on phylogenomic analysis (J. Zheng et al., 2020). This reclassification reflects a more nuanced understanding of the relationships between species and their ecological and evolutionary adaptations (Inglin, Meile, & Stevens, 2018).

Historically, the genus *Lactobacillus* encompassed a large and diverse group of lactic acid bacteria (LAB) renowned for their role in fermentation and probiotic applications. However, the sheer diversity within the genus revealed limitations in the traditional classification system, which lacked precision. Advances in WGS enabled scientists to reassess *Lactobacillus* taxonomy by

examining phylogenetic relationships and conducting detailed genomic comparisons, ultimately prompting a significant revision (J. Zheng et al., 2020).

In 2020, Zheng et al. conducted a comprehensive reclassification study using genomic data from 283 species previously categorised under *Lactobacillus*. Employing a polyphasic approach that integrated genomic, phenotypic, and metabolic data, the study divided the genus into 25 distinct genera, one of which was *Limosilactobacillus. L. fermentum* was reclassified into this new genus, alongside 30 other species, based on shared genomic and ecological characteristics (J. Zheng et al., 2020).

The genus *Limosilactobacillus* primarily comprises species that are either host-associated, commonly residing in the GIT of animals, or found in fermented foods (Racines et al., 2023; Y. Zhao et al., 2022). These species share certain metabolic traits, such as the production of lactic acid through heterofermentation (Sugahara, Kato, Nagayama, Sashihara, & Nagatomi, 2022). This reclassification provides a more refined understanding of the evolutionary pathways and ecological niches occupied by these bacteria (Ksiezarek, Grosso, Ribeiro, & Peixe, 2022).

This taxonomic shift and the inclusion of modern datasets and approaches have significant implications for understanding the evolutionary history of the *Lactobacillus* genus (Ksiezarek et al., 2022; J. Zheng et al., 2020). By organising species into more distinct and accurate genera, researchers can better trace the adaptive traits and evolutionary pressures that have shaped these microbes (Ksiezarek et al., 2022). For *L. fermentum*, its inclusion in *Limosilactobacillus* highlights its capacity to inhabit diverse environments, ranging from the human intestine to plant-based fermented foods (Y. Zhao et al., 2022). Additionally, it underscores its evolutionary relationship with other host-adapted species, such as *Limosilactobacillus reuteri* (F. Li et al., 2023).

2.3. Genome Structure and Function in *L. fermentum*

The genome of *L. fermentum* has been extensively studied, revealing a compact yet highly versatile genetic structure. With a typical genome size ranging between 1.8 and 2.2 megabases

(Mb), it is relatively small compared to free-living bacterial species but aligns with the expected range for host-adapted microbes (K. Brandt, Nethery, O'Flaherty, & Barrangou, 2020; Duar et al., 2017). This reduced genome size is attributed to evolutionary processes such as genome reduction, where genes unnecessary for survival in specific niches are lost over time (Makarova et al., 2006). However, *L. fermentum* has a higher Guanine and Cytosine (GC) content (52.50%) than other members of the *Lactobacillus* genus, which is considered a low-GC genus (Katelyn Brandt & Barrangou, 2018). This higher GC content means that *L. fermentum* has undergone less genomic drift than the other species and does not have a specific niche adaptation like many LAB species (K. Brandt et al., 2020; Duar et al., 2017).

This genomic versatility is reflected in *L. fermentum*'s metabolic capabilities, particularly its facultative heterofermentative nature. It is capable of producing lactic acid through multiple fermentation pathways depending on substrate availability (Adegoke, 2004). Its genome encodes enzymes involved in the fermentation of carbohydrates like glucose, maltose, and sucrose to produce equimolar amounts of lactate, carbon dioxide and ethanol as the primary metabolic byproducts, demonstrating its metabolic flexibility (Adegoke, 2004; O'Donnell, O'Toole, & Ross, 2013).

A notable feature of the *L. fermentum* genome is its ability to adapt and thrive in diverse environments, such as the human GIT and fermented foods. Genes conferring bile salt tolerance and acid resistance enable it to survive the harsh conditions of the GIT. Additionally, the genome encodes adhesion factors critical for attaching to the mucosal surfaces of the GIT, a key trait for probiotic functionality (Bhattacharya et al., 2023; Lehri, Seddon, & Karlyshev, 2015b).

When compared to closely related species such as *L. reuteri*, *L. fermentum* exhibits a broader carbohydrate fermentation profile, enabling it to utilise a wider range of substrates. In contrast, *L. reuteri* has a more specialised genome tailored to its strong adaptation to the GIT of its host (Morita et al., 2008).

The genomic diversity of *L. fermentum* is further evident at the strain level. Strains isolated from fermented foods often harbour genes associated with carbohydrate metabolism that are absent in strains from the human GIT. This strain-specific variation highlights the remarkable ecological versatility of this species and its ability to adapt to distinct environmental niches (Y. Zhao et al., 2022).

2.4. Ecological Niche and Distribution of *L. fermentum*

L. fermentum has a small genome but is an adaptable and versatile species of LAB that thrives across a wide range of ecological niches. It is commonly found in diverse environments, including plant surfaces, fermented food systems, and the GIT of various animal hosts (Pswarayi & Gänzle, 2019; Y. Zhao et al., 2022). The ability of *L. fermentum* to inhabit such diverse ecosystems stems from its genetic adaptability, which allows it to modify its metabolism and survive under different environmental conditions (Y. Zhao et al., 2022). This section will discuss the ecological distribution of *L. fermentum*, focusing on its environmental origins, role in food systems, and presence in animal hosts.

2.4.1. Environmental Origins

The hypothesis that *L. fermentum* originates from plant matter or free-living environments is supported by its frequent isolation from the surfaces of plants, particularly in association with plant fermentation systems (Figure 2.7-2) (Pswarayi & Gänzle, 2019; Sánchez, Palop, & Ballesteros, 2000). *L. fermentum* has been detected on the surfaces of leaves, roots, and other plant tissues, suggesting that it may have evolved as part of the plant microbiota, specifically within the rhizosphere (root-associated microbial communities) and phyllosphere (leaf-associated microbial communities) (Al-Yami, Al-Mousa, Al-Otaibi, & Khalifa, 2022; Ekundayo, 2014; A. O. Yu, Leveau, & Marco, 2020). Despite being less abundant in these environments compared to other microbial species, *L. fermentum* plays a significant role in the fermentation of plant materials (A. O. Yu et al., 2020).

Research has shown that *L. fermentum* is present in a variety of plant-based environments, though often in low abundance (A. O. Yu et al., 2020). For instance, studies have isolated *L. fermentum* from the surfaces of wheat, maize, barley, and rice plants, where it contributes to the initial stages of spontaneous fermentation (Agati, Guyot, Morlon-Guyot, Talamond, & Hounhouigan, 1998; Khlestkin et al., 2022). Its presence in these environments suggests that it may utilise the sugars and other fermentable substrates available in plant tissues for growth and survival (Sudeepa & Bhavini, 2020). While *L. fermentum* is not a dominant species in these ecosystems, its ability to perform lactic acid fermentation under anaerobic conditions or low-oxygen environments gives it a unique ecological niche in breaking down plant materials (X. Yang et al., 2024; M. Zhao et al., 2023).

The role of *L. fermentum* as a fermentative species in plant matter is further supported by its genome, which contains genes for carbohydrate metabolism and lactic acid fermentation (Hossain, 2022). These metabolic capabilities allow it to thrive in plant-associated environments where fermentable carbohydrates are abundant. However, due to its relatively low abundance in the rhizosphere and phyllosphere, *L. fermentum* is considered an opportunistic coloniser rather than a primary component of plant microbiota (Thapa & Prasanna, 2018).

Agati et al. (1998) documented the role of *L. fermentum* in early fermentation stages but highlighted its replacement by other species as fermentation progressed. Comparative genomic studies reveal that *L. fermentum* has genetic adaptations that allow it to exploit specific, short-term ecological niches (Buron-Moles, Chailyan, Dolejs, Forster, & Mikš, 2019). For example, carbohydrate metabolism and stress tolerance genes are well-represented, supporting its capacity to inhabit transient environments like fermenting plant matter (Buron-Moles et al., 2019). Its transient nature in these ecosystems suggests that *L. fermentum* may have evolved to take advantage of short-term fermentative niches, such as those that occur during the decomposition of plant matter or during spontaneous fermentation processes making it ideal for food systems.

2.4.2. Food Systems

One of the most well-established ecological niches for *L. fermentum* is in human-made fermentation systems, where it plays a critical role in the production of fermented foods (Figure 2.7-2). *L. fermentum* is a key species involved in the fermentation of plant-based foods, such as cereals, as well as in dairy and sourdough systems (Angelescu, Zamfir, Stancu, & Grosu-Tudor, 2019; Dan et al., 2015; Naghmouchi et al., 2020). Its ability to thrive in these environments is attributed to its metabolic flexibility and genomic adaptations that enhance its fermentative capacity (K. Brandt et al., 2020; O'Donnell et al., 2013).

In cereal-based fermentation, such as sourdough and traditional fermented beverages like Mahewu (a fermented maize drink common in Zimbabwe), *L. fermentum* is often found as one of the dominant species (Corsetti & Settanni, 2007; Pswarayi & Gänzle, 2019). The ability of *L. fermentum* to spontaneously ferment cereals is linked to its ability to metabolise a wide range of sugars, including glucose, maltose, and raffinose, producing lactic acid as the primary byproduct (Y. Wang et al., 2021). The production of lactic acid lowers the pH of the fermentation environment, creating an acidic condition that favours the growth of LAB while inhibiting the growth of spoilage organisms and pathogens (W. Wu & Li, 2018).

L. fermentum is also commonly found in dairy fermentation systems, particularly in traditional fermented milk products like kefir and yoghurt (Ann Catherine Archer & Halami, 2015; Bao et al., 2010). In these systems, *L. fermentum* contributes to the acidification and flavour development of the final product (Reginensi, Olivera, Bermúdez, & González, 2016). Studies have shown that certain strains of *L. fermentum* isolated from fermented dairy products can produce exopolysaccharides (EPS), which improve the texture and viscosity of fermented milk products (Ale et al., 2016).

Sourdough fermentation is another environment where *L. fermentum* thrives (Corsetti & Settanni, 2007). In this system, *L. fermentum* often coexists with other lactic acid bacteria and yeasts, contributing to the development of the dough's flavour and texture (Corsetti & Settanni,

2007). The ability of *L. fermentum* to ferment complex carbohydrates in sourdough is facilitated by its genome, which encodes enzymes capable of breaking down polysaccharides and oligosaccharides commonly found in flour (Ganzle & Follador, 2012).

The success of *L. fermentum* in human-made food systems can be attributed to its genetic adaptations for energy metabolism. The genome of *L. fermentum* contains genes for many carbohydrate utilisation pathways, allowing it to metabolise both simple and complex sugars. Additionally, *L. fermentum* has developed mechanisms to tolerate the acidic conditions that arise during fermentation, as well as the osmotic stress associated with high sugar concentrations in food systems (Gaucher et al., 2019; Stefanello et al., 2019). This ability to tolerate acidic conditions may also contribute to *L. fermentum*'s survival and adaptability in animal hosts, such as the GIT.

2.4.3. Animal Hosts

Lactobacillus fermentum is also a member of the gut microbiota in various animal hosts, including humans (Duar et al., 2017). Its presence in the GIT of mammals, birds, and other vertebrates suggests that it plays a role in the microbial ecology of the GIT. However, it is generally considered a transient member of the gut microbiota rather than a permanent resident (Chou, Ho, Chen, Wu, & Pan, 2020; Duar et al., 2017; Jang, Kim, Han, Lee, & Ko, 2019).

In humans, *L. fermentum* has been isolated from faecal, oral, and vaginal samples, indicating its broad distribution across different regions of the body (Figure 2.7-2) (Mann et al., 2021; Yan Zhao et al., 2021). Studies have shown that *L. fermentum* is part of the normal gut microbiota. However, it is often in lower abundance than other dominant gut bacteria from the *Bacteroides* genus and other Firmicutes spp. (Mariat et al., 2009). In the GIT, *L. fermentum* produces lactic acid and other antimicrobial compounds that help regulate the composition of the gut microbiota and protect against pathogenic bacteria, conferring intestinal health benefits (Ann Catherine Archer & Halami, 2015).

The role of *L. fermentum* in the oral cavity and vaginal microbiota is similarly important. In the oral cavity, *L. fermentum* is part of the diverse microbial community that colonises the mucosal surfaces, helping to maintain oral health by preventing the overgrowth of harmful microbes (Ahirwar, Gupta, & Snehi, 2019). In the vaginal microbiome, *L. fermentum* is one of the species associated with maintaining a healthy microbial balance, which is crucial for preventing infections such as bacterial vaginosis (Borges, Silva, & Teixeira, 2014).

Although *L. fermentum* is commonly found in the GIT, its presence is generally transient. Unlike other probiotic species, such as *L. reuteri*, which can stably colonise the human intestine, *L. fermentum* tends to pass through the digestive system without establishing a long-term niche (Walter, 2008). This transient nature is likely due to its limited ability to adhere to the intestinal epithelium and its lower competitive advantage compared to other gut-resident microbes (Gharbi et al., 2019). However, despite its transient presence, *L. fermentum* can benefit intestinal health by maintaining an acidic environment unfavourable to pathogens (Pessione, 2012).

In addition to humans, *L. fermentum* has been isolated from the GIT of various other animals, including pigs, chickens, and rodents. In these animals, *L. fermentum* plays a similar role in contributing to intestinal health and modulating the composition of the gut microbiota (Hautefort et al., 1999). In poultry, for example, *L. fermentum* has been identified as a potential probiotic that can enhance growth performance and improve intestinal health by reducing the colonisation of harmful pathogens (Jha, Das, Oak, & Mishra, 2020). Additionally, studies have shown that *L. fermentum* can colonise the GIT of rodents, suggesting its role in modulating the gut microbiota composition and promoting intestinal health (Walter, 2008).

Overall, *L. fermentum* demonstrates remarkable adaptability to several ecological niches, from plant-based environments and food systems to animal hosts (Figure 2.7-2). Its ability to survive in diverse ecosystems is supported by its versatile genome, which enables it to metabolise various substrates and tolerate different environmental stresses. Whether in the rhizosphere,

fermented foods, or the gut microbiota, *L. fermentum* remains a species of significant interest due to its ecological diversity and potential applications in both health and food industries.

2.5. Comparative Genomic: *L. fermentum* vs other LAB Species

The comparative genomic of *L. fermentum* and other closely related LAB species, such as *L. reuteri*, *L. plantarum*, and *L. rhamnosus*, provides valuable insights into the genetic adaptations that enable these species to thrive in different ecological niches (Drissi et al., 2014; Pasolli et al., 2020; Salvetti et al., 2018). While sharing fundamental features like lactic acid fermentation, these LAB species differ significantly in their genomic content, reflecting their adaptation to specific environments, including fermented foods and host-associated habitats like the GIT (Pasolli et al., 2020).

2.5.1. Similarities and Differences in Genomic Content

The genomic content of *L. fermentum* shares several core similarities with other LAB species, particularly in genes related to lactic acid fermentation and basic metabolic processes (Illeghems, De Vuyst, & Weckx, 2015; Ullah et al., 2024). However, there are also significant differences in terms of gene content, particularly those associated with environmental adaptability, host interactions, and niche specialisation (Lehri, Seddon, & Karlyshev, 2017b).

2.5.2. Fermentation Pathways

Like other LAB species, *L. fermentum* primarily relies on the heterofermentative production of lactic acid from carbohydrates (Illeghems et al., 2015). This ability is encoded by a core set of genes, including those responsible for glycolysis, the pentose phosphate pathway, and the production of lactic acid via lactate dehydrogenase (Illeghems et al., 2015). Comparative genomic studies have shown that the genes involved in carbohydrate metabolism are highly conserved across *L. fermentum*, *L. reuteri*, *L. plantarum*, and *L. rhamnosus*, reflecting the central role of lactic acid fermentation in all of these species (Salvetti et al., 2018).

However, *L. fermentum* differs from species like *L. plantarum* in its metabolic versatility. While *L. plantarum* can utilise a wider range of carbon sources, *L. fermentum* is more specialised,

relying on simpler sugars like glucose, maltose, and sucrose (Illeghems et al., 2015). This metabolic specialisation is reflected in its genome, which lacks some of the complex carbohydrate metabolism genes found in *L. plantarum* (Illeghems et al., 2015). On the other hand, *L. reuteri* and *L. rhamnosus* exhibit a similar range of carbohydrate metabolism genes as *L. fermentum* but with slight variations depending on their specific niches, such as intestine or fermented food systems (O'Donnell et al., 2013).

2.5.3. Host Adaptation Genes

One of the key differences between *L. fermentum* and other LAB species lies in the genes associated with host adaptation. Species like *L. reuteri* and *L. rhamnosus* have undergone extensive genetic adaptations to thrive in the GIT of humans and other animals (Chervinets et al., 2018; Frese et al., 2013). This is evidenced by the presence of genes that confer resistance to bile acids, the ability to adhere to the intestinal epithelium, and mechanisms for modulating host immune responses (Chervinets et al., 2018).

In *L. fermentum*, several genes involved in host interaction and adaptation are also present, including genes encoding surface proteins such as mucin-binding proteins and aggregation-promoting factors (Lehri et al., 2015b). These genes help *L. fermentum* adhere to the intestinal lining, albeit transiently, and allow it to modulate the intestinal environment by producing antimicrobial compounds. However, compared to *L. reuteri*, which is known for its long-term colonisation of the GIT, *L. fermentum* has fewer genes directly involved in sustained host-microbe interactions, which correlates with its more transient presence in the GIT (Duar et al., 2017).

2.5.4. Stress Response Genes

Another area of divergence in the genomic content of *L. fermentum* and other LAB species is their ability to respond to environmental stressors. *L. fermentum* possesses genes that confer tolerance to acidic conditions and bile salts, which are critical for its survival in the GIT and fermented food systems (Lehri et al., 2017b). This is similar to other LAB species like *L. rhamnosus*, which also harbours genes for acid and bile resistance, ensuring its survival in both

the GIT and dairy fermentation processes (De Angelis & Gobbetti, 2011). In contrast, *L. plantarum*, which is more versatile and often found in diverse environments, *L. fermentum* has an expanded set of genes for stress responses, including osmotic stress and temperature fluctuations, reflecting its broader ecological niche than *L. fermentum* (De Angelis & Gobbetti, 2011).

2.6. Genome Plasticity of *L. fermentum*

Genome reduction is a common evolutionary strategy observed in host-adapted microorganisms, including *Lactobacillus* species (Duar et al., 2017). This process involves the loss of genes that are unnecessary for survival in stable, nutrient-rich environments such as the GIT or human-made food systems (K. Brandt et al., 2020; Duar et al., 2017). For *L. fermentum*, genome reduction has eliminated pathways that are redundant or not required in these specific niches, enabling greater specialisation (K. Brandt et al., 2020; Lehri, Seddon, & Karlyshev, 2015a).

One notable adaptation resulting from genome reduction in *L. fermentum* is the loss of genes related to the citric acid cycle and respiratory pathways, reflecting its preference for environments where anaerobic fermentation is advantageous (Lehri et al., 2017b). This metabolic streamlining is balanced by the retention of genes essential for survival in acidic and bile-rich environments, such as those in the human GIT (Bernard, Chinnaiyan, Almeda, Catala-Valentin, & Andl, 2023; Ruiz, Margolles, & Sánchez, 2013).

Horizontal gene transfer (HGT) drives genetic diversity in *L. fermentum* and other *Lactobacillus* species. HGT allows bacteria to acquire genetic material from different organisms, enabling rapid adaptation to new environments (van Reenen & Dicks, 2011). Through HGT, *L. fermentum* strains have acquired genes that enhance their fitness in diverse environments, including those related to antibiotic resistance, bile salt hydrolase activity and proton pumps, and carbohydrate metabolism (Anisimova & Yarullina, 2020; Chen, Yu, Tian, Zhao, & Zhai, 2022; Sathyanarayanan Jayashree, Pooja, Pushpanathan, Rajendhran, & Gunasekaran, 2014).

Key vectors for HGT in *L. fermentum* include mobile genetic elements such as plasmids, transposons, and integrative and conjugative elements (ICEs). These elements often carry genes related to antibiotic resistance, carbohydrate metabolism, and stress tolerance, contributing to strain-specific differences and enhancing survival under various conditions (van Reenen & Dicks, 2011).

Plasmids, autonomously replicating DNA molecules, are instrumental in transferring genes that confer advantageous traits (Ojha, Shah, & Mishra, 2021). For example, in fermented food systems, *L. fermentum* has been shown to acquire plasmids that enhance its ability to metabolise specific carbohydrates found in cereal-based fermentations, providing a competitive advantage (Akamine, Mansoldo, & Vermelho, 2023; Alkay & Dertli, 2024; Y. Wang et al., 2021). Additionally, plasmid-encoded antibiotic resistance genes have been identified in strains isolated from fermented foods, likely facilitated by the use of antibiotics in agriculture and food production (S. Jayashree et al., 2013; Ojha et al., 2021).

ICEs are another class of mobile genetic elements that reside within the bacterial chromosome but can excise themselves and transfer to recipient cells via conjugation. ICEs often carry genes beneficial for adaptation, including those involved in metabolism and stress response (Gonçalves, de Assis, & Santana, 2022). Their ability to mobilise and integrate into different genomic contexts makes them significant contributors to genetic diversity in *L. fermentum*.

Bacteriophages, viruses that infect bacteria, also play a role in the genomic diversity of *L. fermentum*. Phage-mediated HGT can transfer genetic material between bacterial strains, introducing new traits (van Reenen & Dicks, 2011). *L. fermentum* strains have developed bacteriophage resistance genes, which protect them from phage attacks in fermented food systems, a common issue in food fermentation processes (Lv et al., 2023; Pei et al., 2021). The acquisition of these resistance genes through HGT enhances the long-term survival and stability of *L. fermentum* populations in these environments. However, these resistance factors, such as CRISPR-Cas systems, are less frequent in *L. fermentum* than in other LAB species (Pei et al.,

2021). The novel genetic potential HGT provides is further shaped by environmental selective pressures.

Selective pressure has further shaped the genome of *L. fermentum*, preserving genes that enhance its probiotic functions. These include genes involved in adhesion to mucosal surfaces, immune modulation, and pathogen inhibition (Colautti, Orecchia, Comi, & Iacumin, 2022; Valeria Garcia-Castillo et al., 2019; Lehri et al., 2015a). These traits contribute to its value as a probiotic, particularly for modulating gut microbiota and promoting GIT health.

The genome of *L. fermentum* reflects a balance between minimalism, achieved through genome reduction, and adaptability, facilitated by HGT (K. Brandt et al., 2020; Lehri et al., 2015a). This dual strategy enables *L. fermentum* to thrive in various environments, from the human GIT to fermented foods. Continued research into its genomic structure will provide further insights into its evolutionary history and potential applications in health and industry.

2.7. Phenotypic Diversity of *L. fermentum*

The phenotypic diversity of *L. fermentum* spans a wide range of traits, from probiotic effects and metabolic adaptations to potential pathogenic behaviour. These phenotypic traits vary considerably depending on the strain's environmental origin, genetic background, and the specific functions they perform. For instance, the average genome size of *L. fermentum* strains ranges between 1.8 Mb to 2.2 Mb indicating the vast potential for genetic difference between strains. Yet, the average GC content is typically narrow, ranging between 51.5% to 52.5% (Ncbi.nlm.nih.gov, 2021). This section explores the functional diversity across various strains of *L. fermentum*, their environmental adaptations, and the unique phenotypic traits observed in specific strains, such as AGR1487's barrier-disruptive properties.

L. fermentum has been isolated from diverse environments, including fermented foods, human and animal microbial communities, and plant materials. The phenotypic traits of these strains often reflect their ecological niches. For example, strains isolated from fermented foods exhibit enhanced fermentative capabilities, contributing to improved food quality and

preservation. In contrast, strains from the human microbial community are more likely to display probiotic properties, such as supporting GIT health and modulating the immune system.

Given the transient nature of *L. fermentum* within various environments, it is likely that these phenotypes overlap across strains isolated from different sources, indicating an adaptive versatility of the species. Understanding these strains' characteristics and potential applications requires examining their origins and their most notable features, like the production of exopolysaccharides (EPS).

EPS production is a key characteristic of several *L. fermentum* strains, significantly contributing to food quality and potential health benefits. For example, the strain YL-11, isolated from fermented milk, is renowned for its EPS production, enhancing fermented foods' texture and viscosity (Wei, Li, Li, Huang, & Li, 2019; Yildiz & Karatas, 2018). Furthermore, YL-11 has shown antitumor potential, underscoring its importance in food systems and its potential therapeutic applications (Wei et al., 2019). Similarly, strain D12, sourced from freshly smoked cheese, and strain MC1, derived from mother's milk, produce EPS that enhances probiotic properties and supports GIT health (Čuljak et al., 2024). Strain Lf2, isolated from hard cheese, is used in yoghurt production due to its EPS production (Ale et al., 2016).

Some EPS-producing *L. fermentum* strains were isolated from spontaneously fermented plant materials. For instance, strains 139, 263, and 296 were isolated from the spontaneous fermentation of *Almagro* eggplants (Guérin, Silva, Garcia, & Remize, 2020). However, only one has been described as isolated from an animal host and that was human breast milk. Given the connection between food systems and animal hosts, these characteristics are likely to be found in *L. fermentum* isolated from animal hosts despite the lack of reported cases, much like the reports of the near-ubiquitous probiotic claims made for the species.

Many *L. fermentum* strains exhibit notable probiotic properties, particularly those isolated from the human GIT. The strain UCO-979C modulates the innate immune response by reducing pro-inflammatory cytokines, including Interleukin-8 (IL-8), Tumour Necrosis Factor Alpha

(TNF-α), Interleukin-1 beta (IL-1β), Interleukin-6 (IL-6), and Monocyte Chemoattractant Protein-1 (MCP-1) while enhancing the production of the regulatory cytokine, transforming growth factor-beta (TGF-β), in gastric epithelial cells. These actions demonstrate its anti-inflammatory properties, which may be beneficial in managing gastric inflammation (Valeria Garcia-Castillo et al., 2019). Another strain, XJC60, has demonstrated the ability to reduce reactive oxygen species by up to 70% in damaged skin while enhancing mitochondrial function (H. Chen et al., 2022). Additionally, nicotinamide produced by XJC60 has been shown to downregulate transcript levels of matrix metalloproteinases (MMP-1 and MMP-3) and pro-inflammatory cytokines (IL-1β, IL-6, and IL-8) in skin cells, thereby suppressing collagen degradation and reducing inflammation in both *in vitro* and *in vivo* models (H. Chen et al., 2022).

Strains 12-1 and X6L1 stimulate the secretion of immunoglobulin A (IgA) in the intestine, enhancing mucosal immunity. Increased IgA levels are crucial in maintaining intestinal immune regulation and protecting against pathogens (Mei et al., 2022). *L. fermentum* CS57 was isolated from vaginal swabs; this strain produces a bacteriocin-like substance that can modulate immune responses. It potentially enhances mucosal immunity and inhibits the growth of pathogenic bacteria, offering protection and promoting microbial balance (La Fata, Weber, & Mohajeri, 2018). Strain MTCC 5898, isolated from human faeces, has enhanced immune responses by stimulating cytokine production and boosting immunoglobulin levels (R. Sharma et al., 2014). Strain 47-7, isolated from the faeces of human infants, has been identified as a promising probiotic for infants, aiding in intestinal development and supporting a healthy gut microbiota (Konyanee et al., 2019; Yotpanya et al., 2016).

The strains Ogi E1 and Mw2 were isolated from Ogi and mawè fermented maize doughs, respectively. These strains exhibit amylolytic activity, breaking down starches into simpler sugars, which can influence the flavour profile of the fermented product (Agati et al., 1998). *L. fermentum* FUA3589 contributes to the fermentation process of mahewu (a fermented cereal) by producing lactic acid, which lowers the pH and enhances the beverage's flavour profile. Its metabolic activities help in developing the characteristic taste of mahewu (Pswarayi & Gänzle,

2019). Similarly, strain FUA3589, derived from mahewu, is used as a starter culture in traditional fermentation, improving flavour profiles and fostering beneficial microbial communities (Pswarayi & Gänzle, 2019).

Some strains, such as DSM 20052, ME-3 and CECT, have been incorporated into commercial dietary supplements due to their extensive probiotic potential and positive effects on intestinal health (K. Brandt et al., 2020; Pereira & Gibson, 2002; Pereira, McCartney, & Gibson, 2003). These strains are widely regarded as safe and effective for improving GIT function and modulating the gut microbiota (Ale, Rojas, Reinheimer, & Binetti, 2020).

The phenotypic diversity of *L. fermentum* is closely linked to the genomic adaptations of each strain. For example, many strains with larger genomes, such as CRL1446 and LF2, have been shown to possess a broader range of functional properties, including the ability to prevent metabolic syndromes and promote cheese fermentation (Russo et al., 2020). Conversely, most strains with smaller genomes, such as ATCC 14931 and Lf1, tend to have more specialised functions, including specific enzymatic activities and probiotic traits often associated with host specialisation (Harris, Ale, Reinheimer, Binetti, & O'Toole, 2018). Still, the trend remains unclear due to the transient nature of the species. However, the health benefits of *L. fermentum* are well established.

While *L. fermentum* is generally recognised for its beneficial traits and is widely considered safe, evidence suggests its potential pathogenicity in specific contexts (Chery, Dvoskin, Morato, & Fahoum, 2013). Notably, *L. fermentum* has been implicated in cholecystitis, empyema, and lung abscesses. For instance, Chery et al. (2013) reported a case of cholecystitis caused by *L. fermentum*, highlighting its ability to act as an opportunistic pathogen. Similarly, Stadler et al. (2018) documented an instance of a lung abscess associated with this species. Additionally, Salminen et al. (2006) discussed the potential pathogenicity of *L. fermentum* in various clinical scenarios. Despite these findings, detailed strain-level information was not recorded or discussed,

limiting further studies necessary to elucidate the underlying mechanisms contributing to such pathogenic behaviour.

A rare and potentially unique phenotype in *L. fermentum* is the barrier-disruptive capability observed in strain AGR1487 (Anderson et al., 2013). AGR1487 was isolated from the human oral cavity of a healthy human. AGR1487 exhibits characteristics of host adaptation, including bile acid resistance and a reduced genome size compared to other strains (Anderson et al., 2013; Bailie, Altermann, Young, Roy, & McNabb, 2021). However, AGR1487 has been shown to disrupt tight junctions between epithelial cells, leading to a loss of intestinal barrier integrity *in vitro* (Anderson et al., 2013). This disruption triggers a pro-inflammatory response in germ-free mice, highlighting the strain's potential to activate the immune system (Anderson et al., 2016). At the time of writing, this phenotype has not been discovered in any other strain of *L. fermentum*.

In contrast, many strains of *L. fermentum* have the opposite effect, such as L930BB and CVM-347, demonstrating protective effects on the intestinal epithelial barrier by enhancing its integrity (Paveljšek et al., 2018). These strains promote reorganisation of the actin cytoskeleton, decrease apoptosis, prevent pathogen invasion, and contribute to GIT health (Gou, Zhang, Ren, Li, & Zhang, 2022; Paveljšek et al., 2018). Since most *L. fermentum* strains are probiotic, show no adverse effects on the host and support intestinal epithelial integrity, AGR1487's phenotype is likely rare. It may stem from specific genetic adaptations absent in other strains. Furthermore, an analysis of the literature and metadata of 80 strains downloaded from NCBI's database in 2022 revealed that only AGR1487 was recorded to have adverse effects on human health or a barrier-disruptive phenotype (Figure 2.7-2).

Considering the genomic plasticity of *L. fermentum*, comprehensive genomic analyses are critical for identifying potential virulence factors and evaluating safety, especially when considering its application in probiotics and fermented foods. Understanding the conditions under which *L. fermentum* transitions from a benign to a pathogenic state is essential to mitigate potential risks associated with its use.

L. fermentum demonstrates remarkable phenotypic diversity, shaped by its adaptation to different ecological niches, such as food systems, the human GIT, and industrial processes (Figure 2.7-2). The rare barrier-disruptive phenotype of AGR1487 contrasts with the typical beneficial effects observed in most strains, reinforcing the need for strain-level characterisation when evaluating *L. fermentum*'s probiotic potential and safety. Research into the genetic and phenotypic diversities of *L. fermentum* continues to better understand the functional capabilities of this species and its applications in both health and industry.

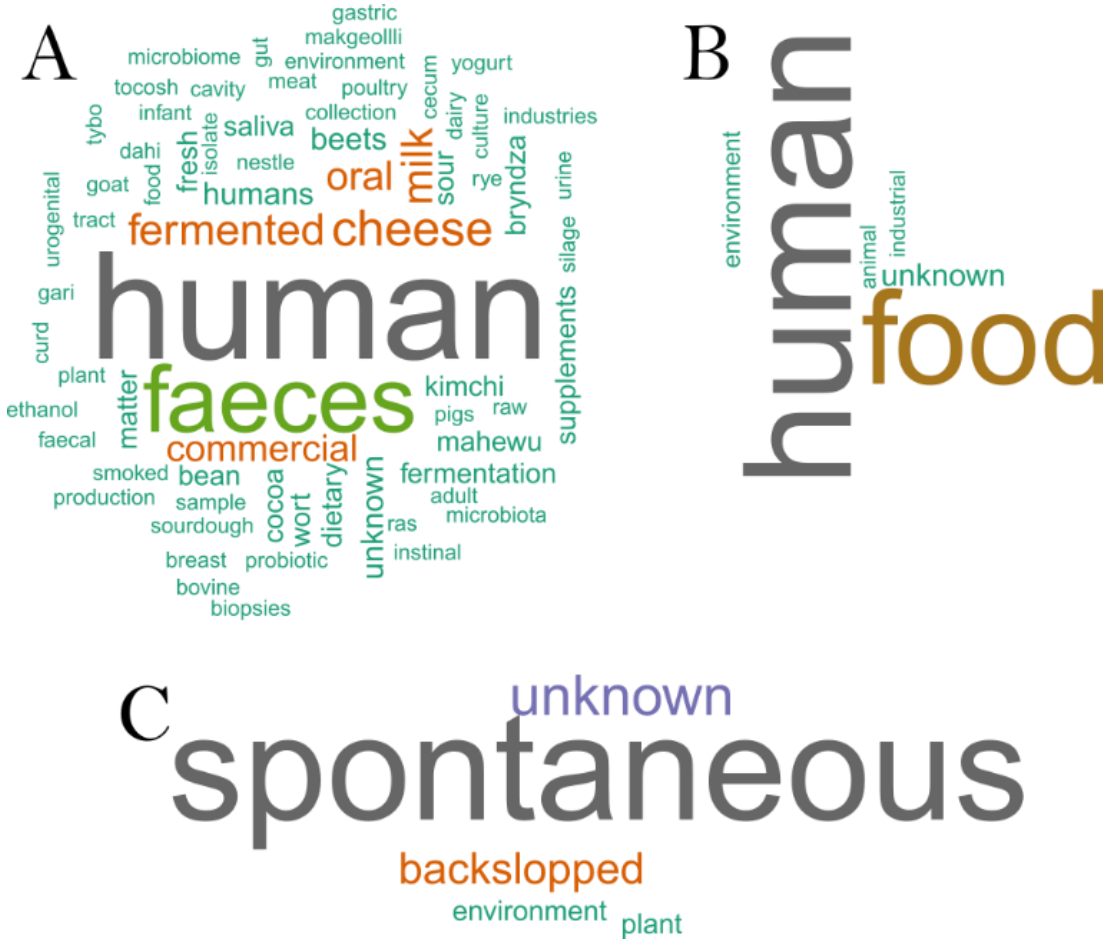

Figure 2.7-2: Word clouds produced from NCBI metadata for 80 strains of *Limosilactobacillus fermentum*. (A) Unmodified keywords extracted *L. fermentum's* source metadata. (B) The source metadata was modified to highlight the source more directly. For example, all food-system related sources were simplified to "food", from leaf matter or environmental samples to "environment", etc. However, industrial or trademarked strains that did not have source metadata were renamed to industrial, and for strains that did not have source metadata were named unknown. (C) How food systems were inoculated extracted from publications associated with each strain where possible.

Table 2.7-1: Strains of *L. fermentum* and their properties

STRAIN	SIZE (MB)	GC%	SOURCE	PHENOTYPE OR USE	REFERENCE
NCC2970	1.95	52.2	Nestle Culture Collection	Acidifying starter culture	(Barretto et al., 2016)
LMT2-75	2.33	50.5	Kimchi	Alcohol and acetaldehyde-decomposing activity	None
3872	2.33	50.6	Milk, humans, poultry, and pigs	Probiotic - bacteriocin-producing bacteria for combatting *Campylobacter jejuni* infections	(Lehri et al., 2015a; Lehri, Seddon, & Karlyshev, 2017a)
2760 AND 2759	2.27	51.4	human faeces matter and dairy	Anti-inflammatory gene, adhesion and immune stimulatory properties	(A. C. Archer, Muthukumar, & Halami, 2021; Palani Kumar, Halami, & Serva Peddha, 2021)
FTDC 8312	2.24	51.0	Human faeces	Probiotic combats hypercholesterolemia	(Lye et al., 2017)
USM 8633	2.24	51.0	Fermented meat sausage	Probiotic, brain health promoting factor	(Ewe, Wan-Abdullah, Alias, & Liong, 2012)
AGR1485	2.23	51.2	Human oral isolate	Probiotic, suspected to have anti-inflammatory behaviour	(Bailie, Altermann, Young, Roy, & McNabb, 2020)
SNUV175	2.28	51.1	Human vagina	Probiotic, bacterial vaginosis and pathogenic viral infections	(S. Lee, You, Kwon, & Ko, 2017; Jang, S. (2018).
SRCM103285	2.15	51.3	Food	Starter culture - acetate fermentation and alcohol tolerance	(Kang et al., 2020)
SRCM 103290	2.12	51.4	Environmental	Unknown effect	None
HFD1	2.10	51.8	Human faeces	Anti-bacterial peptide producing	(Ozhegov et al., 2020)
IFO 3956	2.10	51.5	Egyptian Ras cheese	Starter culture - proteolytic activity	(El-Ghaish et al., 2010)
SK152	2.09	51.9	Kimchi	Antimicrobial properties	(Yoo et al., 2017)
IMDO 130101	2.10	51.5	Rye Sourdough	Acidifying starter culture	(Verce, Vuyst, & Weckx, 2018)
F-6	2.06	51.7	Raw milk	Starter culture and immunosuppression	(Z. Sun, Zhang, Bilige, & Zhang, 2015)
LAC FRN-92	2.06	51.8	Human oral cavity	Unknown effect	None
LDTM 7301	2.05	51.7	Makgeolli	Immunosuppression	(Ann, Choi, & Yoon, 2023)
CBA7106	2.04	51.7	Human faeces	Unknown effect	(D.-Y. Park et al., 2022)
MTCC 25067	2.01	51.2	Indian fermented milk, dahl	Starter culture - thickener	(Aryantini, Prajapati, Urashima, & Fukuda, 2017)
AGR1487	1.94	52.2	Human oral isolate	Disrupts barrier integrity and causes inflammation	(Bailie et al., 2021)
YL-11	1.91	51.9	Fermented milk	Exopolysaccharide Antitumoral Activity	(Wei et al., 2019)
B1 28	1.91	52.3	Fermented beets	Unknown effect	None

DSM 20052	1.89	52.5	Fermented beets	Gastrointestinal function and modulating the gut microbiota	(K. Brandt et al., 2020)
MTCC 5898	2.10	52.1	Human faeces	Enhances systemic immune response and antioxidant capacity in ageing mice	(Gupta, Kaur, Kapila, & Kapila, 2021)
VRI-003	1.95	52.0	Human faeces	Disease resistance, long-term effect on the gut microbiota in humans.	(Cox, Pyne, Saunders, & Fricker, 2010)
47-7	2.10	52.2	Human faeces	Probiotic (effect unspecified)	(Konyanee et al., 2019)
LFQI6	2.10	52.1	Human faeces	Probiotic. Creates a strong biofilm. Upregulating the expression of the essential skin barrier homeostatic protein, filaggrin, in human skin.	(Subhadra, Krier, Hofstee, Monsul, & Berkes, 2015)
ATCC 14931	1.90	52.6	Fermented beets	Produces serine dehydratase L-serine deaminase, L-serine dehydratase	None
DSM 20055	1.90	52.4	Human saliva	Biotechnology potential	(dos Santos et al., 2021)
39	1.83	51.6	Unknown	Unknown	(Soloveva et al., 2021)
CRL1446	2.15	51.4	Oral, Cheese	Probiotic - metabolic syndromes - fat	(Russo et al., 2020)
D12	2.02	52	Cheese	Probiotic- exopolysaccharides	(Čuljak et al., 2024)
222	1.95	52.1	Cocoa bean fermentation	Starter culture - flavour changer	(Illeghems et al., 2015)
AF15-40LB	1.97	52	Human faeces	Gut microbiota analysis	None
MGYG-HGUT-00166	1.97	52	Human faeces	Unknown effect	None
FUA3588	2.02	51.7	Mahewu	Spontaneously inoculate from raw ingredients. Only present during summer but not in winter from the same sites. Flavour starter culture.	(Pswarayi & Gänzle, 2019)
SHI-2	1.94	52.1	Saliva humans	Unknown effect - found during a metabolic study of the human oral microbiome to understand pathogenic biofilm formation. Small peptide molecules are secreted in the oral cavity.	None
S6	1.91	52.3	Sour wort	Spontaneous starter culture	None
S13	1.92	52.3	Sour wort	Spontaneous starter culture	None
L930BB	1.98	52.1	Unknown - store shelf	Probiotic-epithelial barrier protection - RNA expression shows the increase of the cytoskeleton-signalling pathways.	(Paveljšek et al., 2018)
103	2.05	51.8	Human faeces	Unknown effect	None
L13	1.95	52.6	Human faeces	Inhibition of biofilm formation - various responses to medication and antibiotics	None
279	1.99	52.0	Human faeces	Unknown effect	(Tomaro-Duchesneau et al., 2015)
28-3-CHN	2.03	52.0	Human vagina	Unknown effect	None

MD IIE-4657	1.91	52.3	Silage	Fermentation culture	None
DS19_7	2.02	51.6	Commercial dietary supplements	Broadly considered beneficial	None
311	2.04	51.8	Human faeces	Unknown effect	None
317	1.92	51.5	Nono African fermented milk	Bacteriocin and for the potential production of aroma compounds.	None
CECT 9269	2.08	51.7	Tocosh, fermented potatoes	High biotechnological	None
S30	2.16	51.2	Human	Unknown effect	None
L18	2.11	52.0	Human	GABA-producing probiotic	(S. Kaur et al., 2023)
CVM-347	2.09	52.5	Human oral and intestine	Probiotic	None
DS13_7	1.99	51.8	Humans - isolated from Commercial dietary supplements	Probiotic	None
AF16-22LB	1.99	51.9	Human faeces	Unknown effect	None
90 TC-4	1.82	51.9	Unknown commercial	Probiotic - unknown characteristics.	(Tochilina, Belova, Soloveva, Ivanova, & Zhirnov, 2019)
LFU21	2.13	51.4	Human faeces	Unknown effect	(K. Brandt et al., 2020)
BIO6529	1.96	51.7	Human faeces	Unknown effect	None
NBRC 3959	1.93	52.1	Plant material	Commercial Biotechnological	None
BFE 6620	1.98	52.1	Gari - fermented ginger	Starter culture for African vegetable fermentation	(Wafula et al., 2017)
KMB_612	1.92	52.2	Bryndza cheese - sheep milk cheese	Unknown effect	None
FAM 19471	2.04	51.5	Cheese	Unknown effect	None
RI-508	1.91	52.2	Cocoa bean fermentation	Unknown effect	None
KMB_613	2.00	52.1	bryndza cheese	Unknown	None
UMB0187	2.00	52.1	Catheter - urine - human	Female Urinary Microbiome	None
FUA3589	2.08	51.5	Mahewu	Starter culture for African fermented home brew porridge - none alcoholic	(Gaur & Gänzle, 2024)
LF2	2.05	51.7	Regional Tybo cheese	Cheese - nonstarter culture??	(Ale et al., 2016; Harris et al., 2018)
LF1	1.82	52.5	Human intestine	Probiotic	(Chauhan et al., 2014)
NCDC 400	1.90	51.6	Indian Isolate of Dairy Origin - curd	Starter culture for milk fermentation	(Azmal Ali, Kumar, Mohanty, & Behare, 2018)
NB-22	2.01	51.8	Human vaginal microbiota	Probiotic	(Chaplin et al., 2015)
779_LFER	1.94	52.1	Humans	Hospital infections	None

I2	2.08	51.6	Brazilian ethanol production industries	Brazilian ethanol production industries	(Chaplin et al., 2015)
FTDC8312	1.97	51.8	Human faeces	Hypercholesterolemia	(Lye et al., 2017)
CECT 5716	2.10	51.5	breast milk - human	Infection control	(Jimenez et al., 2010)
DR9	2.36	50.4429	Fresh milk - bovine	Reduced ageing - ageing bone and muscle	(Hor et al., 2019)
MTCC 8711	2.57	49.6491	Yoghurt	Probiotic – effect unknown	(Sathyanarayanan Jayashree et al., 2014; S. Jayashree et al., 2013)
HFB3	2.04	51.8	Human faeces	Probiotic – effect unknown	(Kumari, Swarnkar, Kumar, Singh, & Gupta, 2015)
AF11-4-H	1.94	52.2	Human faeces	Unknown effect	None
UCO-979C	2.01	51.9	Human gastric biopsies	Probiotic strain - antibacterial against *Helicobacter pylori* - immunomodulation	(Valeria Garcia-Castillo et al., 2019; V. Garcia-Castillo et al., 2020)
CIM:MAG 1415	1.76	49.7	Human faeces	Intestinal cancer metagenomics	None
39	1.83	51.6	Federal State Scientific-Industrial Company MICROGEN	Unknown effect	(Soloveva et al., 2021)

The strain list, genome size, GC content and isolation were downloaded from NCBI (Ncbi.nlm.nih.gov, 2021). When a strain did not have a primary literature source, none was listed for reference. In situations when no reference was found, the data was pulled from NCBI where possible.

2.8. Molecular Interactions with the Host

Interactions between *L. fermentum* and its host are critical in determining its probiotic potential, immune modulation capabilities, and effects on intestinal health (Valeria Garcia-Castillo et al., 2019; C. Ren et al., 2016). These interactions are primarily mediated by microbe-associated molecular patterns (MAMPs), which are recognised by host pattern recognition receptors (PRRs) (C. Ren et al., 2016). Such interactions influence immune responses, intestinal barrier integrity, and competition with pathogens. This section examines these mechanisms, focusing on the distinct behaviours of *L. fermentum* where possible.

MAMPs enable the host's immune system to distinguish between commensal and pathogenic bacteria (Figure 2.8-3) (Thoda & Touraki, 2023). In *L. fermentum*, key MAMPs include peptidoglycan (PG), lipoteichoic acids (LTA), and exopolysaccharides (EPS) (Abdalla et al., 2021; Chapot-Chartier & Kulakauskas, 2014; Teame et al., 2020). These molecules interact with PRRs such as Toll-like receptors (TLRs) and Nucleotide-binding oligomerisation domain- (NOD) like receptors (NLRs), triggering immune responses that either support homeostasis or lead to inflammation (W. Wang, Li, Han, Li, & Kong, 2022). PG fragments are recognised by NOD2 receptors, modulating intestinal immune balance. LTAs bind to TLR2, promoting immune signalling that prevents excessive inflammation, while EPS strengthens the mucus layer and inhibits pathogen adherence (Figure 2.8-3) (Oliveira-Nascimento, Massari, & Wetzler, 2012; Strober & Watanabe, 2011). The recognition of these MAMPs determines *L. fermentum*'s probiotic potential by supporting commensalism or signalling potential threats (Thoda & Touraki, 2023; W. Wang et al., 2022).

Many *L. fermentum* strains exhibit immune-modulating properties (Valeria Garcia-Castillo et al., 2019; Jang et al., 2019; M. R. Park et al., 2020). MTCC 5898 and VRI-003 stimulate anti-inflammatory cytokines like IL-10, mitigating intestinal inflammation (Cox et al., 2010; Gupta et al., 2021). MTCC 5898 enhances mucosal immunity by increasing IgA production (H. Kaur, Gupta,

Kapila, & Kapila, 2022). Clinical trials with VRI-003 and CECT5716 have demonstrated reduced inflammatory markers following long-term consumption (four months to a year), suggesting *L. fermentum* enhances immune tolerance and reduces inflammation linked to irritable bowel syndrome and inflammatory bowel disease (Cox et al., 2010; Rodríguez-Sojo, Ruiz-Malagón, Rodríguez-Cabezas, Gálvez, & Rodríguez-Nogales, 2021). Strain-specific modifications of surface molecules like EPS and LTAs influence whether *L. fermentum* promotes immune tolerance or disrupts intestinal homeostasis (Teame et al., 2020).

Molecular interactions between *Lactobacillus* species and the host maintain intestinal homeostasis by regulating mucosal integrity and immune responses (Yan & Polk, 2020). The intestinal epithelial barrier separates host immunity from a diverse microbial community (Okumura & Takeda, 2017). *Lactobacillus* species contribute to mucosal health by promoting the production of mucin, antimicrobial peptides, and secretory IgA, which protect the intestinal epithelial barrier and maintain microbial diversity (Ohland & MacNaughton, 2010; J. Sun et al., 2019; Zheng, Liwinski, & Elinav, 2020). Mucin forms a physical barrier that minimises microbial translocation, allowing beneficial bacteria to adhere to the mucosal layer (Fang et al., 2021; Paone & Cani, 2020). Dysbiosis, an imbalance resulting from infection, poor diet, or antibiotic use, can lead to epithelial barrier breakdown, chronic inflammation, diseases such as irritable bowel syndrome and inflammatory bowel disease, and metabolic disorders (Di Vincenzo, Del Gaudio, Petito, Lopetuso, & Scaldaferri, 2024; Santana, Rosas, Ribeiro, Marinho, & de Souza, 2022).

Direct microbial interactions with host cells involve MAMP recognition by PRRs, including TLRs, C-type lectin receptors (CLRs), and NLRs (Figure 2.8-3) (Kawai & Akira, 2009; Thoda & Touraki, 2023). PRRs help distinguish between commensal and pathogenic bacteria, modulating immune responses (Tanoue, Umesaki, & Honda, 2010). MAMPs such as PG, LTA, and EPS are integral to *L. fermentum*'s interactions with the host (Abdalla et al., 2021; R. Sengupta et al., 2013). PG, a major component of the Gram-positive bacterial cell wall, is recognised by TLR2 and NOD

receptors, triggering immune responses ranging from inflammation to immune tolerance (Dubé & Behr, 2023; Le, Kulatheepan, & Jeyaseelan, 2023).

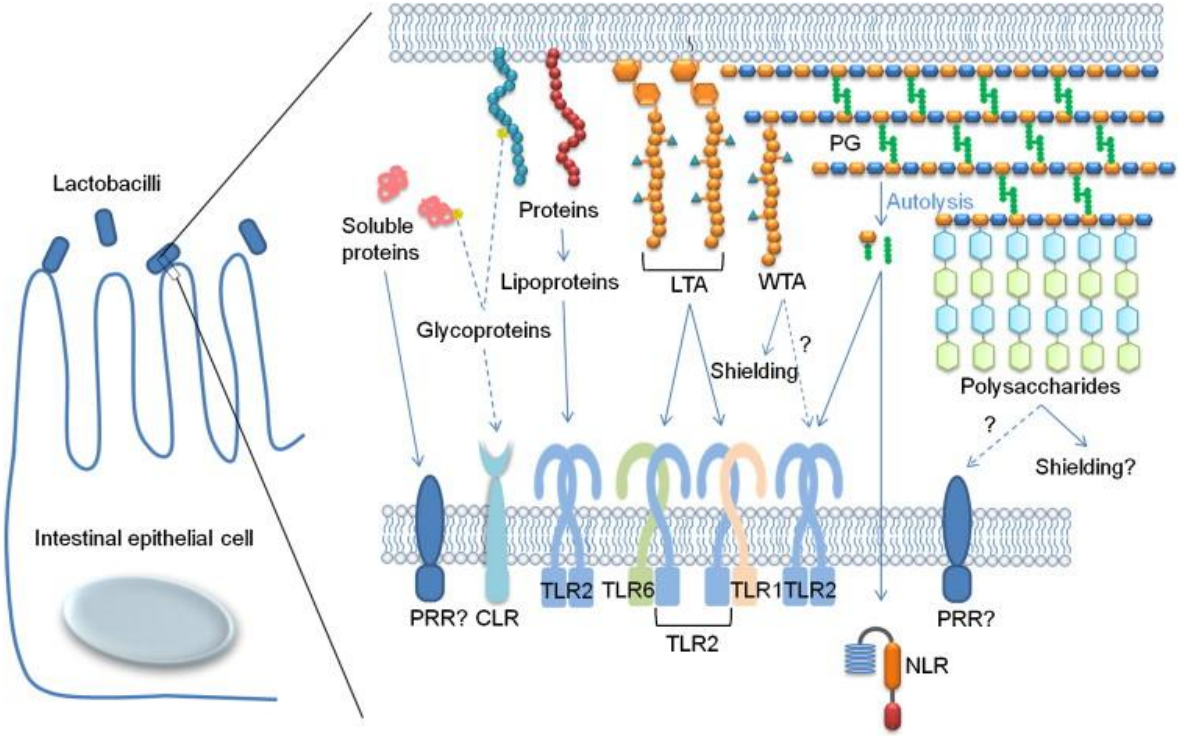

Figure 2.8-3: A representation of the surface interface between human gastrointestinal epithelial cells and the surface of *Lactobacillus* indicating MAMPs and their respective PRRs. The Lactobacillus cell envelope contains a variety of macromolecules like peptidoglycan (PG), teichoic acids, polysaccharides and lipoproteins or surface (glyco)proteins. These macromolecules or fragments of them are specifically recognised by the host's Toll-Like Receptors (TLRs), Nucleotide Oligomerization Domain (NOD)-Like Receptors (NLRs), and C-type Lectin Receptors (CLRs). Fragments of the PG from autolysis or enzymatic digestion and lipoproteins are detected via TLR2. Lipoteichoic acids are thought to signal through heterodimers of TLR2/6 and TLR2/1. While wall teichoic acid (WTA) may interact via TLR2, they are predominantly viewed as protective elements that shield surface molecules from interacting. This shielding effect is also a function of the *Lactobacillus* cell surface polysaccharides. Glycoproteins are detected through C-type Lectin Receptors (CLRs). However, other direct interactions like mucin binding, cell adhesion and degradation are not represented here. Adapted with permission from Elsevier and reproduced from I. C. Lee, Tomita, Kleerebezem, and Bron (2013).

2.9. Peptidoglycan Structure in *L. fermentum*

Peptidoglycan (PG) in Gram-positive bacteria like *L. fermentum* consists of glycan chains of alternating sugars, N-acetylglucosamine (GlcNAc) and N-acetylmuramic acid (MurNAc), cross-linked by short peptides (Rajagopal & Walker, 2017; Vollmer, Blanot, & De Pedro, 2008). These glycan strands form the backbone of PG, with peptide bridges connecting the MurNAc residues of adjacent strands (Vollmer et al., 2008). The primary peptide chain typically consists of L-alanine, D-glutamic acid, meso-diaminopimelic acid (mDAP) or L-lysine, and D-alanine. In *Lactobacillus* species, these peptide bridges undergo modifications that alter immune recognition and resistance to environmental stress (R. Sengupta et al., 2013; Torrens & Cava, 2024).

Cross-linking of these peptide chains gives PG a three-dimensional mesh-like structure, giving the bacterial cell wall strength and flexibility (Vollmer et al., 2008). The degree of cross-linking varies between species, influencing cell wall rigidity. In Gram-positive bacteria like *L. fermentum*, the PG layer is significantly thicker than in Gram-negative bacteria, often reaching up to 40 layers (Silhavy, Kahne, & Walker, 2010; Vollmer et al., 2008). This thickness provides structural support, allowing survival in diverse environments, including the GIT (Silhavy et al., 2010).

PG structure varies among *L. fermentum* strains, with unique adaptations influencing host interactions (Chapot-Chartier & Kulakauskas, 2014). For instance, a study comparing two *L. fermentum* strains, AGR1485 and AGR1487, found notable differences in their effects on intestinal epithelial barrier function. AGR1487 decreased transepithelial electrical resistance (TEER) across Caco-2 cell layers, indicating a reduction in barrier integrity, whereas AGR1485 did not exhibit this effect (Ranjita Sengupta et al., 2015). These findings suggest that variations in cell surface structures, including PG composition and surface proteins, may lead to strain-specific interactions with host cells, resulting in differential activation of metabolic pathways and immune responses (Ranjita Sengupta et al., 2015). Additionally, PG provides mechanical strength and shape, preventing osmotic

lysis and contributing to *L. fermentum*'s rod-like morphology (Chapot-Chartier & Kulakauskas, 2014; Vollmer et al., 2008). It is a dynamic structure undergoing constant remodelling during bacterial growth and division, ensuring structural integrity through the activity of PG-synthesising and degrading enzymes (Alvarez et al., 2024; Garde, Chodisetti, & Reddy, 2021; Weaver, Taguchi, & Dörr, 2023).

PG also serves as an anchor for other cell wall components, such as LTAs and surface proteins, which mediate interactions with host tissues and gut microbes (Fischetti, 2019; Rajagopal & Walker, 2017). PG fragments act as MAMPs, recognised by host PRRs, particularly TLRs and NODs (Bersch et al., 2021; D. Li & Wu, 2021). This recognition initiates immune responses, with PG playing a key role in immune modulation by promoting tolerance or triggering inflammation, depending on its structure and modifications (Bastos, Wheeler, & Boneca, 2020; Bersch et al., 2021).

2.10. Peptidoglycan Modifications and Host Interactions

Variations in PG structure profoundly affect bacterial physiology and host immune responses. Modifications like amidation, *O*-acetylation, and *N*-deacetylation enable bacteria to evade immune defences and persist within the intestinal environment (Davis & Weiser, 2011).

Amidation converts carboxylic acid groups in peptide chains into amide groups, enhancing resistance to antimicrobial peptides and antibiotics. This modification alters the cell wall structure, thereby influencing bacterial interactions with host cells and modulating immune recognition (Bishop, Boyle, & Finlay, 2007).

O-acetylation involves adding acetyl groups to MurNAc residues via *O*-acetyltransferase (*OatA*). This change increases bacterial resistance to lysozyme and reduces susceptibility to bacterial autolysins, which could allow *L. fermentum* strains to thrive in lysozyme-rich environments, such as the GIT (Bishop et al., 2007).

Similarly, N-deacetylation removes acetyl groups from GlcNAc residues through the action of peptidoglycan deacetylase (*PgdA*). This process further enhances lysozyme resistance by preventing PG degradation, bolstering the bacteria's defences against host immune mechanisms (Davis & Weiser, 2011).

Together, these modifications enable *L. fermentum* to fine-tune its interactions with the host, promoting symbiosis while avoiding excessive immune activation. By altering its PG structure, the bacterium can also better adapt to environmental stresses, including osmotic pressure, pH fluctuations, and temperature changes, which are critical for survival in the GIT (R. Sengupta et al., 2013; Yadav, Espaillat, & Cava, 2018).

2.11. Peptidoglycan and Immune Response

The interaction between PG and the host immune system determines whether a bacterial strain behaves as a probiotic or pathogen, as was seen for *L. rhamnosus* (Huang et al., 2020). PG is recognised as a MAMP by PRRs on immune cells such as TLR2, NOD1, and NOD2, initiating immune responses that can be pro-inflammatory or homeostatic (Trindade & Chen, 2020). During pro-inflammatory responses, PG fragments can stimulate cytokine production, including TNF-α and IL-6, which are crucial for infection control but can lead to tissue damage if overproduced (Chiu et al., 2009; J. E. Wang et al., 2000). However, during tolerance and immune homeostasis, non-pathogenic *Lactobacillus* strains, PG recognition promotes immune tolerance, contributing to intestinal homeostasis and preventing unnecessary inflammation (Mazziotta, Tognon, Martini, Torreggiani, & Rotondo, 2023; H. J. Wu & Wu, 2012). *L. fermentum* PG plays a pivotal role in host-microbe interactions by influencing immune responses and balancing beneficial probiotic effects with potential immunogenicity (Kim, Lee, Kang, Kang, & Cho, 2023; Mazziotta et al., 2023). Understanding these modifications and their effects on host immunity is critical for assessing *L. fermentum* probiotic potential and its role in intestinal health.

2.12. Exopolysaccharides in *Lactobacillus* Strains

Exopolysaccharides (EPS) produced by *Lactobacillus* strains play critical roles in bacterial survival, host-microbe interactions, immune modulation, and environmental adaptation (Dertli, Mayer, & Narbad, 2015; Nowak et al., 2020). These complex carbohydrate polymers are secreted by bacteria into their external environment, forming a protective matrix that encapsulates the bacterial cells (Dertli et al., 2015; Nowak et al., 2020). EPS is essential in biofilm formation, enhancing bacterial adhesion to surfaces and protecting against hostile environmental factors, including the host immune system. The production of EPS is a highly strain-specific trait, with significant variation across different species and strains of *Lactobacillus*, reflecting their diverse ecological niches and functions (Arsköld et al., 2007; Dertli et al., 2015).

2.12.1. Structure and Composition of Exopolysaccharides

EPS in *Lactobacillus* species typically consists of repeating units of monosaccharides such as glucose, galactose, fructose, mannose, and rhamnose. Glycosidic bonds link these units, forming long, branched or linear polymer chains. EPS can be classified into capsular or free based on their attachment to the bacterial cell (Netrusov et al., 2023).

Capsular polysaccharides (CPS) are covalently attached to the bacterial cell surface, forming a tightly bound capsule around the cell. Free or loosely bound EPS (slime) are secreted into the surrounding environment, forming a slimy matrix around the bacterial colonies. The specific composition and structure of EPS vary widely among *Lactobacillus* strains, and these variations are thought to be linked to the bacteria's specific functions and environmental adaptations. In *L. fermentum*, for instance, distinct EPS-producing strains have been associated with differing capabilities in biofilm formation and immune modulation (Netrusov et al., 2023; Sørensen et al., 2022).

2.12.2. Role of EPS in Biofilm Formation

One of the most important functions of EPS in *Lactobacillus* strains is its contribution to biofilm formation (Dertli et al., 2015; Salas-Jara, Ilabaca, Vega, & García, 2016). Biofilms are structured communities of bacteria embedded within a self-produced EPS matrix, which attaches to surfaces such as the GIT, teeth, or medical devices (Khatoon, McTiernan, Suuronen, Mah, & Alarcon, 2018). When within a biofilm, bacteria are better protected from environmental stressors such as antimicrobial agents, immune responses, and fluctuations in pH or temperature (Khatoon et al., 2018).

EPS is the primary scaffold of the biofilm, providing structural integrity and facilitating bacterial adhesion to surfaces (Limoli, Jones, & Wozniak, 2015). This adhesion is crucial for bacterial colonisation and persistence in host environments (Limoli et al., 2015). Studies have shown that EPS produced by *L. fermentum* strains enhances their ability to form biofilms in both GIT and oral environments, contributing to their probiotic effect (Abdalla et al., 2021). The production of EPS-rich biofilms also allows these bacteria to withstand environmental challenges, such as oxidative stress and desiccation, further enhancing their survival (Abdalla et al., 2021; Limoli et al., 2015).

2.12.3. Protective Role of EPS against Host Immune Factors

EPS is critical in protecting *Lactobacillus* strains from the host immune system (Ciszek-Lenda, Nowak, Sróttek, Gamian, & Marcinkiewicz, 2011; Nowak et al., 2020; Tarannum et al., 2024). The EPS matrix can shield against host immune factors such as antimicrobial peptides, complement proteins, and phagocytic cells (Ciszek-Lenda et al., 2011; Nowak et al., 2020; Tarannum et al., 2024). By encapsulating bacterial cells, EPS prevents direct contact between bacterial surface molecules and host immune receptors, reducing immune detection and response (Ciszek-Lenda et al., 2011; Nowak et al., 2020; Tarannum et al., 2024).

This protective role of EPS is critical in the GIT, where *Lactobacillus* strains must withstand constant exposure to bile salts, digestive enzymes, and immune molecules (Lebeer, Claes, Verhoeven,

Vanderleyden, & De Keersmaecker, 2011). EPS production allows these bacteria to evade immune recognition by masking PAMPs on their surface, such as LTAs or PG, that would otherwise be recognised by host PRRs, such as TLRs (Gunn, Bakaletz, & Wozniak, 2016; Moser et al., 2021; Peters et al., 2022). By reducing immune recognition, EPS helps maintain a commensal relationship between *Lactobacillus* strains and their host, supporting the stability of the host gut microbiota (Jurášková, Ribeiro, & Silva, 2022; Laiño, Villena, Kanmani, & Kitazawa, 2016).

In addition to its role as a physical barrier, EPS can actively interact with host immune cells to modulate immune responses. For example, *L. fermentum* strains produce EPS that reduces the secretion of pro-inflammatory cytokines like TNF-α and IL-6 by immune cells (V. Garcia-Castillo et al., 2020; Su et al., 2022). This immune-modulatory effect helps prevent excessive inflammation in the intestine, promoting health and preventing inflammatory diseases such as inflammatory bowel disease (Su et al., 2022; M. Sun, Wu, Liu, & Cong, 2017).

2.12.4. Immune Modulation and Probiotic Potential of EPS

The probiotic effects of *Lactobacillus* strains are closely linked to their ability to modulate the host immune system, and EPS production plays a central role in this process (Laiño et al., 2016). Studies have demonstrated that the EPS produced by *L. fermentum* and other *Lactobacillus* species can promote immune homeostasis by influencing the balance between pro-inflammatory and anti-inflammatory responses (Laiño et al., 2016; Xiu et al., 2018).

One mechanism through which EPS exerts its immune-modulatory effects is interacting with dendritic cells and macrophages, key players in the host immune system (Tang et al., 2015). EPS has been shown to stimulate the production of anti-inflammatory cytokines, such as IL-10, while reducing the production of pro-inflammatory cytokines, like TNF-α (Ciszek-Lenda et al., 2011). This modulation of immune responses helps to maintain a healthy intestinal environment, reducing the risk of inflammatory conditions and promoting overall intestinal health (Hou et al., 2022).

In addition to its effects on cytokine production, EPS can influence the differentiation of regulatory T cells (Tregs), which play a crucial role in controlling immune responses and maintaining tolerance to commensal bacteria (Nowak et al., 2020; S. Ren et al., 2021). The ability of *Lactobacillus* strains to enhance Treg differentiation is an essential aspect of their probiotic potential, as it helps prevent inappropriate immune responses to harmless or beneficial microbes (Nowak et al., 2020; S. Ren et al., 2021).

2.12.5. EPS and Microbial Adaptation

EPS production is key in adapting *Lactobacillus* strains to different environments, particularly within the host (Nguyen et al., 2020). The ability to produce EPS allows *Lactobacillus* strains to colonise a variety of niches, including the GIT, oral cavity, and skin, by providing protection against environmental stressors and enhancing biofilm formation (Dertli et al., 2015; Giordani et al., 2023; Nguyen et al., 2020).

In the GIT, EPS helps *Lactobacillus* strains persist in bile acids, digestive enzymes, and fluctuating pH levels (Khalil, Abd Manap, Mustafa, Alhelli, & Shokryazdan, 2018). In the oral cavity, EPS contributes to biofilm formation on dental surfaces, protecting salivary enzymes and immune molecules (Cugini, Shanmugam, Landge, & Ramasubbu, 2019). On the skin, *Lactobacillus* strains are exposed to desiccation and antimicrobial peptides. While specific studies on EPS production by *Lactobacillus* strains on the skin are limited, it is plausible that EPS provides a protective barrier, aiding in moisture retention and offering resistance against antimicrobial substances, thereby supporting bacterial survival in the skin's challenging environment.

2.12.6. Therapeutic Potential of EPS

Given its role in biofilm formation, immune modulation, and environmental adaptation, EPS produced by *Lactobacillus* strains holds significant potential for therapeutic applications. EPS-producing strains of *Lactobacillus* strains have been explored for use in the treatment of

gastrointestinal disorders, including inflammatory bowel disease and irritable bowel syndrome, due to their ability to enhance intestinal barrier integrity and reduce inflammation (Compare et al., 2017; Noda, Danshiitsoodol, Kanno, Uchida, & Sugiyama, 2021).

In addition, EPS-producing probiotics have shown promise in oral health applications, where they can help prevent dental caries and periodontal disease by promoting biofilm formation and protecting against pathogenic bacteria (Zhang, Ding, & Guo, 2022). EPS has also been investigated for its potential in skin health, where it may enhance barrier function and protect against microbial infections (Morifuji, Kitade, Fukasawa, Yamaji, & Ichihashi, 2017).

The therapeutic potential of EPS extends beyond its direct effects on the host, as EPS can also influence the composition and function of the host microbiota. By promoting biofilm formation and immune modulation, EPS-producing *Lactobacillus* strains can help to maintain a healthy gut microbiota composition, which is essential for overall health and well-being (Bengoa, Dardis, Gagliarini, Garrote, & Abraham, 2020).

EPS produced by *Lactobacillus* strains plays a central role in bacterial survival, biofilm formation, and host-microbe interactions (Netrusov et al., 2023; Singh, Datta, Narayanan, & Rajnish, 2021; Werning et al., 2022). The strain-specific nature of EPS production reflects the diverse ecological niches occupied by *Lactobacillus* strains and their adaptive capabilities (Nguyen et al., 2020). EPS provides a protective shield against host immune factors and promotes immune modulation, making EPS-producing *Lactobacillus* strains promising candidates for therapeutic applications (Kwon et al., 2020; Nowak et al., 2020). Understanding the mechanisms underlying EPS production and function is essential for harnessing the full potential of *Lactobacillus* strains as probiotics and for developing new strategies to promote health and prevent disease.

2.13. TEER Assays in Gastrointestinal Research

Trans-epithelial/Endothelial Electrical Resistance (TEER) assays have become an essential non-invasive method for quantifying the integrity of epithelial barriers, particularly within GIT studies (Srinivasan et al., 2015). The epithelial barrier serves as a critical defence, regulating the passage of nutrients while blocking harmful substances such as toxins, pathogens, and microbes (Srinivasan et al., 2015). Disruptions in this barrier, observed in conditions like inflammatory bowel disease or so-called "leaky gut syndrome", can lead to increased permeability and inflammation (Srinivasan et al., 2015). This section reviews the TEER assay, the applications of the assay in gastrointestinal research, and the TEER assay in *L. fermentum* research.

2.13.1. Overview of the TEER Assays

TEER assays provide a quantitative measure of the integrity of epithelial or endothelial monolayers by assessing the electrical resistance across them (Srinivasan et al., 2015). In these assays, cell lines such as Caco-2 (derived from colorectal adenocarcinoma cells) are cultured on permeable membranes (Kus, Ibragimow, & Piotrowska-Kempisty, 2023). Electrodes positioned above and below the cell monolayer apply a small electrical current, and the resulting resistance reflects the formation and maintenance of tight junctions (Figure 3.2.1) (Srinivasan et al., 2015). Higher resistance values indicate robust tight junction integrity, while lower values suggest cell lysis, mechanical disruption, or separation of cells, allowing current to pass through gaps in the monolayer (Srinivasan et al., 2015).

TEER is widely accepted as a quantitative technique for monitoring tight junction dynamics in both endothelial and epithelial cell culture models (Figure 2.13.1) (Srinivasan et al., 2015). In a typical TEER setup, Caco-2 cells are grown into a confluent monolayer on a semi-permeable membrane housed within an electrically insulating cup suspended in growth media. The design of the TEER cup

ensures that the applied current travels across the membrane rather than around it (Srinivasan et al., 2015).

In co-culture experiments designed to simulate the intestinal lumen interface, bacteria are added to the apical chamber (above the cell monolayer), while the basal chamber contains separate media (Janssen et al., 2024; Srinivasan et al., 2015). This configuration permits direct interaction between the bacteria and the apical surface of differentiated Caco-2 cells while preventing the mixing of the two media compartments (Verhoeckx et al., 2015).

Despite its widespread use, the reproducibility of TEER assays can be affected by the choice of cell culture media (Kus et al., 2023; M. Sharma et al., 2025). Much of the validation of TEER assays has been carried out using media such as Minimal Essential Media (MEM) or Dulbecco's Modified Eagle's Media (DMEM), both of which contain at least 10% foetal bovine serum (FBS) (Kus et al., 2023). Although FBS supplies essential growth factors, its presence can influence Caco-2 cell morphology, proliferation, differentiation, and ultimately, barrier integrity (Sambuy et al., 2005; Xiu-Wen, Ru-Feng, Ming, Wei, & Xiu-Wei, 2013).

In studies involving *L. fermentum*, notably when testing strain AGR1487, the apical media typically excludes FBS to avoid pH alterations resulting from lactic acid production. Media 199 (M199) is often used instead of MEM or DMEM. However, the literature did not discuss the rationale for this media substitution. The choice of which media to use during research may profoundly impact the results. This is an important consideration given the wide range of applications the TEER assay has in intestinal research.

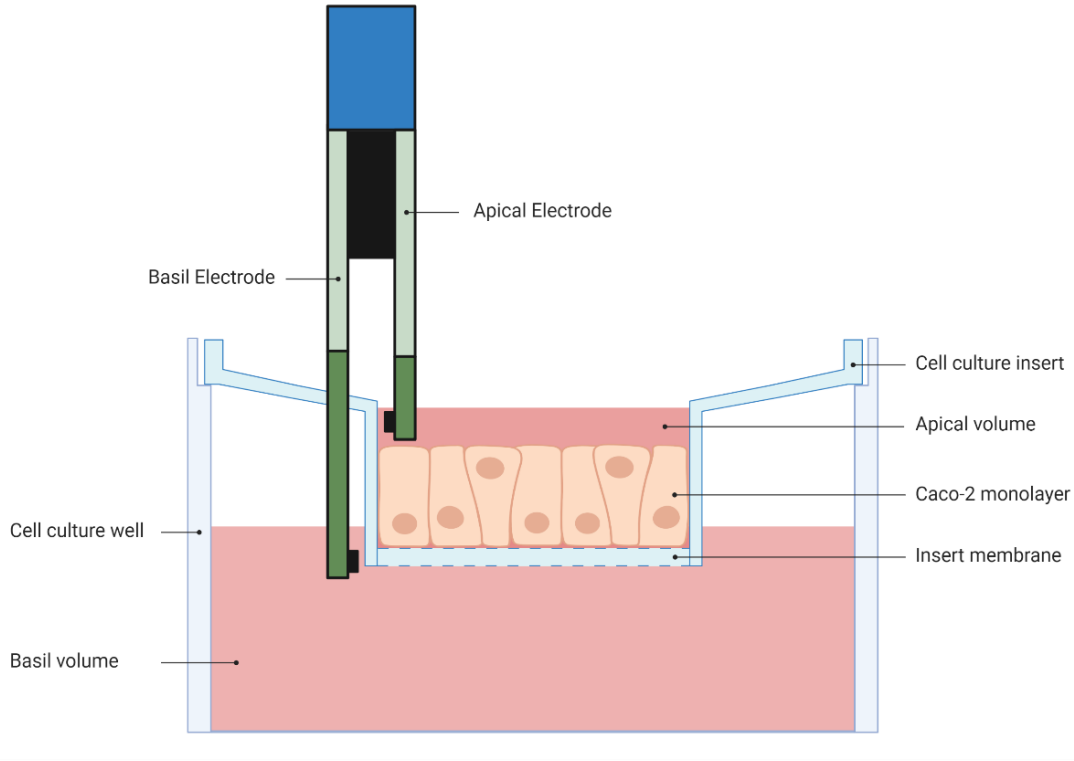

Figure 2.13-4: Diagrammatical layout of a typical TEER experiment. A cell (e.g., Caco-2) monolayer is grown on a semi-permeable membrane attached to a cell culture insert suspended in growth media. Media is applied to the top of the cells (Apical), and a volume is provided beneath the cells (Basil). The TEER electrodes are placed so that one electrode is above the monolayer and the other is below the monolayer.

2.14. Applications of TEER Assays in Gastrointestinal Research

TEER assays offer critical insights into the functioning and health of the intestinal barrier. Researchers use these assays to study a wide variety of factors that impact the epithelial barrier, including the effects of pathogens, toxins, and probiotic treatments (Boll, Winther, Knudsen, Copani, & Cappellozza, 2024; F. Yang et al., 2015; Yuan, van der Mei, Busscher, & Peterson, 2020).

2.14.1. Inflammatory Responses

Inflammation in the GIT, triggered by diseases such as Crohn's disease or ulcerative colitis, can lead to a breakdown in tight junctions, which are equivalent to reducing TEER values (Lechuga, Braga-Neto, Naydenov, Rieder, & Ivanov, 2023). In research, TEER assays are used to investigate how inflammation compromises epithelial integrity and to test the effectiveness of anti-inflammatory treatments in restoring barrier function (Lechuga et al., 2023; Metz et al., 2020).

2.14.2. Probiotic Efficacy

Probiotics, particularly strains of *Lactobacillus*, including *L. fermentum*, are known for their ability to modulate immune responses and strengthen or weaken epithelial barriers. TEER assays allow researchers to evaluate the ability of probiotics to enhance barrier function by promoting tight junction formation, as is the case for *L. reuteri* I5007 (F. Yang et al., 2015). Increased TEER values after treatment with probiotics indicate an improvement in barrier integrity. However, a reduction in TEER indicates a disruption of the barrier function, as shown for *L. fermentum* AGR1487 (Anderson et al., 2016).

2.14.3. Toxin or Pathogen Exposure

TEER assays are instrumental in understanding how pathogens or toxins affect the intestinal epithelial barrier. A decrease in TEER upon exposure to bacterial toxins or pathogens, such as *Clostridium difficile* or *Salmonella enterica*, indicates barrier disruption (Feltis et al., 1999; Lépine et al., 2018). These assays can also assess the protective effects of treatments, such as probiotics or anti-microbial agents, in preventing pathogen-induced barrier damage (Han et al., 2023).

2.15. TEER findings with *L. fermentum*

Several recent investigations have applied TEER measurements to assess the impact of different *L. fermentum* strains on epithelial models, revealing promising roles for these bacteria in enhancing

and protecting barrier integrity. However, TEER studies involving *L. fermentum* are rare at the time of writing.

2.15.1. TEER Barrier disruption and immune activation by AGR1487

Studies in FBS-free M199 media have demonstrated that *L. fermentum* AGR1487 compromises the barrier integrity of Caco-2 monolayers. Immunofluorescent staining of tight junction proteins reveals that AGR1487 uniformly degrades these junctions, a disruptive effect observed whether the bacteria are alive or dead (Anderson et al., 2013; Ranjita Sengupta et al., 2015). Notably, both whole cells and bacterial cell debris induce barrier disruption, whereas the bacterial supernatant alone does not, suggesting that the active component is associated with the bacterial cell surface rather than being a secreted factor (Ranjita Sengupta et al., 2015).

Further supporting this conclusion, experiments mixing AGR1487 with the non-disruptive strain AGR1485 showed that the presence of AGR1485 does not mitigate the barrier-disruptive effects of AGR1487, reinforcing the idea that AGR1487's action is both intrinsic and remains potent when the strain is dead (Anderson et al., 2013). Although the mechanism remains unknown, one hypothesis links this barrier disruption to an immune response similar to that observed in germ-free rat models.

In vivo, the presence of AGR1487 in the intestines of germ-free rats causes colon-specific inflammation, a response not seen with AGR1485 (Anderson et al., 2016). AGR1487 appears to activate this immune response via TLRs and the proinflammatory regulator NF-κB (Anderson et al., 2016). The strain also demonstrates enhanced capabilities in stimulating dendritic cell maturation, cytokine secretion, and the recruitment of macrophages and lymphocytes in the colon (Anderson et al., 2016). Despite inoculation at similar concentrations (as measured by OD_{600}), AGR1487 achieves higher colonisation, up to 1,000-fold greater colony-forming units, in the caecum, colon, and faeces compared to AGR1485. However, inflammation was restricted to the colon (Anderson et al., 2016).

The TEER assay data and the inflammatory response in germ-free rats might initially seem linked through a common TLR-mediated mechanism (Anderson et al., 2016). However, the Caco-2 TEER assay does not have a mucosal layer, and supernatants from AGR1487 did not produce the barrier-disruptive phenotype, suggesting a distinct mechanism is possible. In TEER assays, the disruptive effect requires direct contact, which is impossible with a mucosal layer. The inflammatory response could be triggered by MAMPs engaging TLRs, but these MAMPs would have to overcome the mucosal layer in germ-free rats (Anderson et al., 2016; Anderson et al., 2013).

It is possible that these MAMPs are not present in the bacterial supernatant if they are associated with larger cell surface structures that become pelleted during centrifugation. Alternatively, variations in the integrity of the intestinal mucus layer might allow contact-dependent effects in some circumstances but not others, explaining why inflammation is localised to the colon. These findings collectively suggest that AGR1487 may employ two distinct modes of action: one that directly disrupts epithelial tight junctions and another that incites TLR-mediated inflammatory responses. Unlike *L. fermentum* AGR1487, TEER disruption was not seen for *L. fermentum* PCC and other strains (Srimahaeak, Bianchi, Chlumsky, Larsen, & Jespersen, 2021).

2.15.2. TEER Enhancement by *L. fermentum* PCC

In an *in vitro* study using Caco-2 cell monolayers, exposure to *L. fermentum* PCC increased TEER by approximately 33% compared to untreated controls (Srimahaeak, Bianchi, Chlumsky, Larsen, & Jespersen, 2021). The study further noted that the addition of pectin amplified this effect, suggesting that dietary fibres may work synergistically with the probiotic to improve tight junction function and stabilise the epithelial barrier (Srimahaeak, Bianchi, Chlumsky, Larsen, & Jespersen, 2021).

2.15.3. Barrier Stabilisation with *L. fermentum* 53

Another investigation focused on the metabolites secreted by *L. fermentum* strain 53 in HT-29 cell monolayers. HT-29 is the human colorectal adenocarcinoma cell line, meaning it is derived from a human colon cancer (adenocarcinoma) tumour, like Caco-2 cells. The results showed that treatment with these bioactive compounds increased TEER and induced the reorganisation of actin filaments, key components in cell-to-cell adhesion. This dual effect indicates that the secreted metabolites reinforce the barrier, helping maintain cell polarity and tight junction integrity (Moskova-Doumanova et al., 2024).

2.15.4. Protective Effects of *L. fermentum* ME-3

The strain *L. fermentum* ME-3 has been evaluated in Caco-2 cell models as a protective agent against barrier-disrupting challenges. When Caco-2 monolayers were exposed to either the chemotherapeutic agent Irinotecan or inflammatory stimuli (such as lipopolysaccharide), preincubation with ME-3 effectively prevented the typical drop in TEER. This preservation of electrical resistance indicates that ME-3 can both prevent and repair tight junction disruption, thereby safeguarding epithelial barrier function under adverse conditions (Moskova-Doumanova et al., 2024)

2.15.5. Consortium Approaches Involving *L. fermentum* 3872

Further extending the concept of barrier protection, a consortium that included *L. fermentum* strain 3872, combined with *Ligilactobacillus salivarius* 7247, was tested across human, porcine, and chicken enterocyte models. In these systems, the consortium increased TEER and reduced paracellular permeability. This multi-strain approach suggests that complementary interactions among probiotic strains can lead to enhanced stabilisation of the intestinal barrier, thereby reducing the risk of pathogen translocation and inflammation (Abramov et al., 2024).

2.16. Conclusion

Limosilactobacillus fermentum has remarkable diversity and multifaceted roles, a compact, versatile genome, varied phenotypic traits, and strain-specific interactions with host systems. Advances in whole-genome sequencing and comparative genomics have unveiled the genetic foundations of its probiotic functions and the adaptations that enable different strains to thrive in niches such as the GIT, fermented foods, and plant environments.

Moreover, the strain-specific differences, evident in studies ranging from metabolic profiling and immune modulation to TEER assays of epithelial barrier function, are critical for achieving beneficial probiotic effects and, in rare cases, may lead to the detection of adverse outcomes. These insights are pivotal for the rational selection and safe application of *L. fermentum* strains in health-promoting products and food fermentations. Future research should continue to elucidate the molecular mechanisms behind these strain-dependent effects, ensuring that only the most effective and safe strains are developed into targeted probiotic interventions and industrial applications.

TEER assays have long served as an essential tool for evaluating the integrity of epithelial barriers, particularly within the GIT. They provide valuable insights into how probiotics, pathogens, and inflammation affect epithelial cell tight junctions and overall barrier function. Although TEER assays are limited, co-culture systems and molecular profiling are expanding their capabilities, making them even more potent for understanding epithelial health.

TEER assays will remain at the forefront of GIT studies as research progresses, especially in developing treatments for conditions like inflammatory bowel disease and irritable bowel syndrome. By integrating TEER measurements with advanced technologies, a more comprehensive understanding of epithelial barrier function in health and disease and more effective strategies might help target and restore barrier integrity to improve overall human health.

The strain-specific differences observed in *L. fermentum* and other *Lactobacillus* species play a pivotal role in their interactions with the host and their potential as probiotics. This review encompassed published literature on 89 unique *L. fermentum* strains, of which only one exhibited a reported barrier-disruptive phenotype. This finding suggests that AGR1487's barrier-disruptive behaviour may be rare or unique to that strain. While some strains promote intestinal health and favourably modulate immune responses, others may disrupt intestinal barrier integrity and contribute to inflammation. Therefore, understanding the genetic, metabolic, and functional diversity among *L. fermentum* strains is essential for developing safe and effective probiotic products. Future research should continue to explore the mechanisms underlying these strain-specific effects, focusing on identifying those that offer the greatest potential for enhancing human health.

2.17. Aims of this thesis

Many *L. fermentum* strains are recognised for their probiotic benefits, including enhancing intestinal barrier integrity. However, strain AGR1487 uniquely disrupts this barrier. Mechanistic studies on *L. fermentum* phenotypes are rare or absent. Although TEER assays are commonly used to assess barrier function in cell culture models, the molecular basis for AGR1487's disruptive effects remains largely unexplored. Unlike beneficial strains that stabilise tight junctions and promote intestinal mucosal health, AGR1487 appears to alter epithelial junction proteins and impair cell-cell adhesion in Caco-2 cells. This discrepancy highlights the need to investigate the specific cellular pathways and molecular events activated by AGR1487.

Current evidence supports the thesis hypothesis that *L. fermentum AGR1487 induces a barrier-disruptive phenotype in TEER assays through a genetically mediated cell surface structure*. However, research has yet to fully characterise the cell surface structures of AGR1487 compared to its non-disruptive counterparts. Although bacterial surface properties are recognised as critical mediators of host interactions, limited data exist on how AGR1487's cell envelope differs from strains like

AGR1485. Advanced imaging techniques, such as high-resolution electron microscopy (discussed in Chapter 3), are needed to reveal these structural differences, potentially identifying targets that could be responsible for AGR1487's barrier-disruptive phenotype.

Moreover, AGR1487 appears to harbour unique genetic elements and variations in genome size that may underlie its negative impact on barrier integrity. Complete whole-genome sequencing is essential to identify the genetically mediated cell surface structures involved. Chapter 4 focuses on generating reference-quality genome assemblies that accurately reflect the true genomes, ensuring that no critical genetic information, possibly responsible for the barrier-disruptive phenotype, is overlooked.

Although comparative genomic analyses for *L. fermentum* exist, they typically compare different species rather than differences between strains. Existing pan-genome studies often rely on draft assemblies that may miss core gene similarities and structural differences.

Chapters 5 and 6 aim to integrate multi-omics data, comparative genomics in Chapter 5 and transcriptomic profiles during TEER assays in Chapter 6 to identify candidate genes responsible for AGR1487's disruptive phenotype. This comprehensive approach is crucial for pinpointing the genetic determinants underlying AGR1487's barrier-disruptive phenotype. Once identified, functional studies such as gene knockouts and complementation experiments will be essential to confirm the link between these genetic factors and the observed barrier disruption, thereby clarifying the molecular interplay between bacterial surface structures and host epithelial responses.

Addressing these gaps will validate the hypothesis that AGR1487 disrupts the intestinal epithelial barrier via a genetically mediated cell surface structure and inform the screening and selection of probiotic strains for clinical applications and food production.

A summary and discussion of the important results presented in this thesis is given in Chapter 7.

2.18. References

Abdalla, A. K., Ayyash, M. M., Olaimat, A. N., Osaili, T. M., Al-Nabulsi, A. A., Shah, N. P., & Holley, R. (2021). Exopolysaccharides as antimicrobial agents: Mechanism and spectrum of activity. *Frontiers in Microbiology, 12*, 664395. https://doi.org/10.3389/fmicb.2021.664395

Abramov, V. M., Kosarev, I. V., Machulin, A. V., Deryusheva, E. I., Priputnevich, T. V., Panin, A. N., . . . Karlyshev, A. V. (2024). Anti-Salmonella defence and intestinal homeostatic maintenance *in vitro* of a consortium containing *Limosilactobacillus fermentum* 3872 and *Ligilactobacillus salivarius* 7247 strains in human, porcine, and chicken enterocytes. *Antibiotics, 13*(1), 30. Retrieved from https://www.mdpi.com/2079-6382/13/1/30

Adegoke, G. O. (2004). *Understanding food microbiology* (2nd ed.). Alleluia Ventures.

Agati, V., Guyot, J. P., Morlon-Guyot, J., Talamond, P., & Hounhouigan, D. J. (1998). Isolation and characterization of new amylolytic strains of *Lactobacillus fermentum* from fermented maize doughs (*mawè* and *ogi*) from Benin. *Journal of Applied Microbiology, 85*(3), 512–520. https://doi.org/10.1046/j.1365-2672.1998.853527.x

Ahirwar, S. S., Gupta, M., & Snehi, S. K. (2019). Dental caries and *Lactobacillus*: Role and ecology in the oral cavity. *International Journal of Pharmaceutical Sciences and Research, 11*, 4818–4829.

Akamine, I. T., Mansoldo, F. R. P., & Vermelho, A. B. (2023). Probiotics in the sourdough bread fermentation: Current status. *Fermentation, 9*(2), 90. Retrieved from https://www.mdpi.com/2311-5637/9/2/90

Al-Yami, A. M., Al-Mousa, A. T., Al-Otaibi, S. A., & Khalifa, A. Y. (2022). *Lactobacillus* species as probiotics: Isolation sources and health benefits. *Journal of Pure and Applied Microbiology, 16*, 2270–2291.

Ale, E. C., Perezlindo, M. J., Burns, P., Tabacman, E., Reinheimer, J. A., & Binetti, A. G. (2016). Exopolysaccharide from *Lactobacillus fermentum* Lf2 and its functional characterization as a yogurt additive. *Journal of Dairy Research, 83*(4), 487–492. https://doi.org/10.1017/S0022029916000571

Ale, E. C., Rojas, M. F., Reinheimer, J. A., & Binetti, A. G. (2020). *Lactobacillus fermentum*: Could EPS production ability be responsible for functional properties? *Food Microbiology, 90*, 103465. https://doi.org/10.1016/j.fm.2020.103465

Alkay, Z., & Dertli, E. (2024). Sourdough lactic acid bacteria as functional microorganisms for technological purposes. In E. B. Ceresino, G. Juodeikiene, S. M. Schwenninger, & J. M. Ferreira da Rocha (Eds.), *Sourdough microbiota and starter cultures for industry* (pp. 309–341). Springer International Publishing.

Alvarez, L., Hernandez, S. B., Torrens, G., Weaver, A. I., Dörr, T., & Cava, F. (2024). Control of bacterial cell wall autolysins by peptidoglycan crosslinking mode. *Nature Communications, 15*(1), 7937. https://doi.org/10.1038/s41467-024-52325-2

Anderson, R. C., Ulluwishewa, D., Young, W., Ryan, L. J., Henderson, G., Meijerink, M., . . . Roy, N. C. (2016). Human oral isolate *Lactobacillus fermentum* AGR1487 induces a pro-inflammatory response in germ-free rat colons. *Scientific Reports, 6*, 20318. https://doi.org/10.1038/srep20318

Anderson, R. C., Young, W., Clerens, S., Cookson, A. L., McCann, M. J., Armstrong, K. M., & Roy, N. C. (2013). Human oral isolate *Lactobacillus fermentum* AGR1487 reduces intestinal barrier integrity by increasing the turnover of microtubules in Caco-2 cells. *PLOS ONE, 8*(11), e78774. https://doi.org/10.1371/journal.pone.0078774

Angelescu, I.-R., Zamfir, M., Stancu, M.-M., & Grosu-Tudor, S.-S. (2019). Identification and probiotic properties of lactobacilli isolated from two different fermented beverages. *Annals of Microbiology, 69*(13), 1557–1565. https://doi.org/10.1007/s13213-019-01540-0

Anisimova, E. A., & Yarullina, D. R. (2020). Antibiotic resistance and the mobility of its genetic determinants in *Lactobacillus fermentum*. *Molecular Genetics, Microbiology and Virology, 35*(4), 202–209. https://doi.org/10.3103/S0891416820040035

Ann, S., Choi, Y., & Yoon, Y. (2023). Comparative genomic analysis and physiological properties of *Limosilactobacillus fermentum* SMFM2017-NK2 with ability to inflammatory bowel disease. *Microorganisms, 11*(3), 547.

Archer, A. C., & Halami, P. M. (2015). Probiotic attributes of *Lactobacillus fermentum* isolated from human feces and dairy products. *Applied Microbiology and Biotechnology, 99*(19), 8113–8123. https://doi.org/10.1007/s00253-015-6679-x

Archer, A. C., Muthukumar, S. P., & Halami, P. M. (2021). *Lactobacillus fermentum* MCC2759 and MCC2760 alleviate inflammation and intestinal function in high-fat diet-fed and streptozotocin-induced diabetic rats. *Probiotics and Antimicrobial Proteins, 13*(4), 1068–1080. https://doi.org/10.1007/s12602-021-09744-0

Arsköld, E., Svensson, M., Grage, H., Roos, S., Rådström, P., & van Niel, E. W. (2007). Environmental influences on exopolysaccharide formation in *Lactobacillus reuteri* ATCC 55730. *International Journal of Food Microbiology, 116*(1), 159–167. https://doi.org/10.1016/j.ijfoodmicro.2006.12.010

Aryantini, N. P., Prajapati, J. B., Urashima, T., & Fukuda, K. (2017). Complete genome sequence of *Lactobacillus fermentum* MTCC 25067 (formerly TDS030603), a viscous exopolysaccharide-producing strain isolated from Indian fermented milk. *Genome Announcements, 5*(13). https://doi.org/10.1128/genomeA.00091-17

Azmal Ali, S., Kumar, S., Mohanty, A. K., & Behare, P. (2018). Draft genome sequence of *Lactobacillus fermentum* NCDC 400, isolated from a traditional Indian dairy product. *Genome Announcements, 6*(2). https://doi.org/10.1128/genomeA.01492-17

Bailie, M. A., Altermann, E., Young, W., Roy, N. C., & McNabb, W. C. (2020). Complete genome sequence of *Lactobacillus fermentum* strain AGR1485, a human oral isolate. *Microbiology Resource Announcements, 9*(36), e00841-00820. https://doi.org/10.1128/MRA.00841-20

Bailie, M. A., Altermann, E., Young, W., Roy, N. C., & McNabb, W. C. (2021). Complete annotated genome sequence of *Limosilactobacillus fermentum* AGR1487. *Microbiology Resource Announcements, 10*(1). https://doi.org/10.1128/mra.01056-20

Bao, Y., Zhang, Y., Zhang, Y., Liu, Y., Wang, S., Dong, X., . . . Zhang, H. (2010). Screening of potential probiotic properties of *Lactobacillus fermentum* isolated from traditional dairy products. *Food Control, 21*(5), 695–701. https://doi.org/10.1016/j.foodcont.2009.10.010

Barretto, C., Ngom-Bru, C., Genevaz, A., Fournier, C., Moine, D., Kassam, M., . . . Duboux, S. (2016). Genome sequence of *Lactobacillus fermentum* strain NCC2970 (CNCM I-5068). *Genome Announcements, 4*(6). https://doi.org/10.1128/genomeA.01254-16

Bastos, P. A. D., Wheeler, R., & Boneca, I. G. (2020). Uptake, recognition and responses to peptidoglycan in the mammalian host. *FEMS Microbiology Reviews, 45*(1). https://doi.org/10.1093/femsre/fuaa044

Belova, I. V., Tochilina, A. G., Solovyeva, I. V., Efimov, E. I., Gorlova, I. S., Ivanova, T. P., & Zhirnov, V. A. (2016). *Lactobacillus fermentum* 90 TC-4 taxonomic status confirmation using whole genome sequencing and MALDI TOF mass spectrum. *Russian Journal of Genetics, 52*(9), 907–913. https://doi.org/10.1134/S1022795416090039

Bengoa, A. A., Dardis, C., Gagliarini, N., Garrote, G. L., & Abraham, A. G. (2020). Exopolysaccharides from *Lactobacillus paracasei* isolated from kefir as potential bioactive compounds for microbiota modulation. *Frontiers in Microbiology, 11*, 583254. https://doi.org/10.3389/fmicb.2020.583254

Bernard, J. N., Chinnaiyan, V., Almeda, J., Catala-Valentin, A., & Andl, C. D. (2023). *Lactobacillus* sp. facilitate the repair of DNA damage caused by bile-induced reactive oxygen species in experimental models of gastroesophageal reflux disease. *Antioxidants, 12*(7), 1314. Retrieved from https://www.mdpi.com/2076-3921/12/7/1314

Bersch, K. L., DeMeester, K. E., Zagani, R., Chen, S., Wodzanowski, K. A., Liu, S., . . . Grimes, C. L. (2021). Bacterial peptidoglycan fragments differentially regulate innate immune signaling. *ACS Central Science, 7*(4), 688–696. https://doi.org/10.1021/acscentsci.1c00200

Bhattacharya, R., Mukhopadhyay, B. C., Mishra, M., Arukha, A. P., Mandal, S., & Biswas, S. R. (2023). Draft genome sequence of probiotic bacterium *Lactobacillus fermentum* S2, isolated from raw cow milk. *Microbiology Resource Announcements, 12*(11), e00499-00423. https://doi.org/10.1128/MRA.00499-23

Bishop, J. L., Boyle, E. C., & Finlay, B. B. (2007). Deception point: Peptidoglycan modification as a means of immune evasion. *Proceedings of the National Academy of Sciences, 104*(3), 691–692. https://doi.org/10.1073/pnas.0611133104

Boll, E. J., Winther, K. D., Knudsen, T. T. M., Copani, G., & Cappellozza, B. I. (2024). *Ligilactobacillus animalis* 506 protects the intestinal barrier from the damaging effects of enteric pathogens and deoxynivalenol. *Animals (Basel), 14*(2). https://doi.org/10.3390/ani14020269

Borges, S., Silva, J., & Teixeira, P. (2014). The role of lactobacilli and probiotics in maintaining vaginal health. *Archives of Gynecology and Obstetrics, 289*(3), 479–489. https://doi.org/10.1007/s00404-013-3064-9

Brandt, K., & Barrangou, R. (2018). Using glycolysis enzyme sequences to inform *Lactobacillus* phylogeny. *Microbial Genomics, 4*(6). https://doi.org/10.1099/mgen.0.000187

Brandt, K., Nethery, M. A., O'Flaherty, S., & Barrangou, R. (2020). Genomic characterization of *Lactobacillus fermentum* DSM 20052. *BMC Genomics, 21*(1), 328. https://doi.org/10.1186/s12864-020-6740-8

Bull, M. J., & Plummer, N. T. (2014). Part 1: The human gut microbiome in health and disease. *Integrative Medicine (Encinitas), 13*(6), 17–22.

Buron-Moles, G., Chailyan, A., Dolejs, I., Forster, J., & Mikš, M. H. (2019). Uncovering carbohydrate metabolism through a genotype-phenotype association study of 56 lactic acid bacteria genomes. *Applied Microbiology and Biotechnology, 103*(7), 3135–3152. https://doi.org/10.1007/s00253-019-09701-6

Chaplin, A. V., Shkoporov, A. N., Efimov, B. A., Pikina, A. P., Borisova, O. Y., Gladko, I. A., . . . Kafarskaia, L. I. (2015). Draft genome sequence of *Lactobacillus fermentum* NB-22. *Genome Announcements, 3*(4). https://doi.org/10.1128/genomeA.00896-15

Chapot-Chartier, M. P., & Kulakauskas, S. (2014). Cell wall structure and function in lactic acid bacteria. *Microbial Cell Factories, 13*(Suppl 1), S9. https://doi.org/10.1186/1475-2859-13-S1-S9

Chauhan, R., Sudhakaran Vasanthakumari, A., Panwar, H., Mallapa, R. H., Duary, R. K., Batish, V. K., & Grover, S. (2014). Amelioration of colitis in mouse model by exploring antioxidative potentials of an indigenous probiotic strain of *Lactobacillus fermentum* Lf1. *BioMed Research International, 2014*(1), 206732. https://doi.org/10.1155/2014/206732

Chen, C., Yu, L., Tian, F., Zhao, J., & Zhai, Q. (2022). Identification of novel bile salt-tolerant genes in *Lactobacillus* using comparative genomics and its application in the rapid screening of tolerant strains. *Microorganisms, 10*(12), 2371. Retrieved from https://www.mdpi.com/2076-2607/10/12/2371

Chen, H., Li, Y., Xie, X., Chen, M., Xue, L., Wang, J., . . . Zhao, H. (2022). Exploration of the molecular mechanisms underlying the anti-photoaging effect of *Limosilactobacillus fermentum* XJC60. *Frontiers in Cellular and Infection Microbiology, 12*, 838060. https://doi.org/10.3389/fcimb.2022.838060

Chervinets, Y., Chervinets, V., Shenderov, B., Belyaeva, E., Troshin, A., Lebedev, S., & Danilenko, V. (2018). Adaptation and probiotic potential of lactobacilli, isolated from the oral cavity and intestines of healthy people. *Probiotics and Antimicrobial Proteins, 10*(1), 22–33. https://doi.org/10.1007/s12602-017-9348-9

Chery, J., Dvoskin, D., Morato, F. P., & Fahoum, B. (2013). *Lactobacillus fermentum*, a pathogen in documented cholecystitis. *International Journal of Surgery Case Reports, 4*(8), 662–664. https://doi.org/10.1016/j.ijscr.2013.04.034

Chiu, Y.-C., Lin, C.-Y., Chen, C.-P., Huang, K.-C., Tong, K.-M., Tzeng, C.-Y., . . . Tang, C.-H. (2009). Peptidoglycan enhances IL-6 production in human synovial fibroblasts via TLR2 receptor, focal adhesion kinase, Akt, and AP-1-dependent pathway. *The Journal of Immunology, 183*(4), 2785–2792. https://doi.org/10.4049/jimmunol.0802826

Chou, Y. C., Ho, P. Y., Chen, W. J., Wu, S. H., & Pan, M. H. (2020). *Lactobacillus fermentum* V3 ameliorates colitis-associated tumorigenesis by modulating the gut microbiome. *American Journal of Cancer Research, 10*(4), 1170–1181.

Ciszek-Lenda, M., Nowak, B., Sróttek, M., Gamian, A., & Marcinkiewicz, J. (2011). Immunoregulatory potential of exopolysaccharide from *Lactobacillus rhamnosus* KL37: Effects on the production of inflammatory mediators by mouse macrophages. *International Journal of Experimental Pathology, 92*(6), 382–391. https://doi.org/10.1111/j.1365-2613.2011.00788.x

Colautti, A., Orecchia, E., Comi, G., & Iacumin, L. (2022). Lactobacilli, a weapon to counteract pathogens through the inhibition of their virulence factors. *Journal of Bacteriology, 204*(11), e00272-00222. https://doi.org/10.1128/jb.00272-22

Compare, D., Rocco, A., Coccoli, P., Angrisani, D., Sgamato, C., Iovine, B., . . . Nardone, G. (2017). *Lactobacillus casei* DG and its postbiotic reduce the inflammatory mucosal response: An ex-vivo organ culture model of post-infectious irritable bowel syndrome. *BMC Gastroenterology, 17*(1), 53. https://doi.org/10.1186/s12876-017-0605-x

Corsetti, A., & Settanni, L. (2007). Lactobacilli in sourdough fermentation. *Food Research International, 40*(5), 539–558.

Cox, A. J., Pyne, D. B., Saunders, P. U., & Fricker, P. A. (2010). Oral administration of the probiotic *Lactobacillus fermentum* VRI-003 and mucosal immunity in endurance athletes. *British Journal of Sports Medicine, 44*(4), 222–226. https://doi.org/10.1136/bjsm.2007.044628

Cugini, C., Shanmugam, M., Landge, N., & Ramasubbu, N. (2019). The role of exopolysaccharides in oral biofilms. *Journal of Dental Research, 98*(7), 739–745. https://doi.org/10.1177/0022034519845001

Čuljak, N., Bellich, B., Pedroni, A., Butorac, K., Pavunc, A. L., Novak, J., . . . Kos, B. (2024). *Limosilactobacillus fermentum* strains MC1 and D12: Functional properties and exopolysaccharides characterization. *International Journal of Biological Macromolecules, 273*, 133215. https://doi.org/10.1016/j.ijbiomac.2024.133215

Dan, T., Liu, W., Song, Y., Xu, H., Menghe, B., Zhang, H., & Sun, Z. (2015). The evolution and population structure of *Lactobacillus fermentum* from different naturally fermented products as determined by multilocus sequence typing (MLST). *BMC Microbiology, 15*, 107. https://doi.org/10.1186/s12866-015-0447-z

Davis, K. M., & Weiser, J. N. (2011). Modifications to the peptidoglycan backbone help bacteria to establish infection. *Infection and Immunity, 79*(2), 562–570. https://doi.org/10.1128/iai.00651-10

De Angelis, M., & Gobbetti, M. (2011). Stress responses of lactobacilli. In E. Tsakalidou & K. Papadimitriou (Eds.), *Stress Responses of Lactic Acid Bacteria* (pp. 219–249). Springer US.

De Gregorio, A., Serafino, A., Krasnowska, E. K., Superti, F., Di Fazio, M. R., Fuggetta, M. P., . . . Fiorentini, C. (2023). Protective effect of *Limosilactobacillus fermentum* ME-3 against the increase in paracellular permeability induced by chemotherapy or inflammatory conditions in Caco-2 cell models. *International Journal of Molecular Sciences, 24*(7), 6225. Retrieved from https://www.mdpi.com/1422-0067/24/7/6225

Dekaboruah, E., Suryavanshi, M. V., Chettri, D., & Verma, A. K. (2020). Human microbiome: An academic update on human body site-specific surveillance and its possible role. *Archives of Microbiology, 202*(8), 2147–2167. https://doi.org/10.1007/s00203-020-01931-x

Dertli, E., Mayer, M. J., & Narbad, A. (2015). Impact of the exopolysaccharide layer on biofilms, adhesion, and resistance to stress in *Lactobacillus johnsonii* FI9785. *BMC Microbiology, 15*(1), 8. https://doi.org/10.1186/s12866-015-0347-2

Di Vincenzo, F., Del Gaudio, A., Petito, V., Lopetuso, L. R., & Scaldaferri, F. (2024). Gut microbiota, intestinal permeability, and systemic inflammation: A narrative review. *Internal and Emergency Medicine, 19*(2), 275–293. https://doi.org/10.1007/s11739-023-03374-w

dos Santos, C. I., Campos, C. D. L., Nunes-Neto, W. R., do Carmo, M. S., Nogueira, F. A. B., Ferreira, R. M., . . . Monteiro-Neto, V. (2021). Genomic analysis of *Limosilactobacillus fermentum* ATCC 23271, a potential probiotic strain with anti-*Candida* activity. *Journal of Fungi, 7*(10), 794. Retrieved from https://www.mdpi.com/2309-608X/7/10/794

Drissi, F., Merhej, V., Angelakis, E., El Kaoutari, A., Carrière, F., Henrissat, B., & Raoult, D. (2014). Comparative genomics analysis of *Lactobacillus* species associated with weight gain or weight protection. *Nutrition & Diabetes, 4*(2), e109. https://doi.org/10.1038/nutd.2014.2

Duar, R. M., Lin, X. B., Zheng, J., Martino, M. E., Grenier, T., Perez-Munoz, M. E., . . . Walter, J. (2017). Lifestyles in transition: Evolution and natural history of the genus *Lactobacillus*. *FEMS Microbiology Reviews, 41*(Supp_1), S27–S48. https://doi.org/10.1093/femsre/fux030

Dubé, J. Y., & Behr, M. A. (2023). A nod to the bond between NOD2 and *Mycobacteria*. *PLoS Pathogens, 19*(6), e1011389. https://doi.org/10.1371/journal.ppat.1011389

Ekundayo, F. (2014). Isolation and identification of lactic acid bacteria from rhizosphere soils of three fruit trees, fish, and ogi. *International Journal of Current Microbiology and Applied Sciences,* 3(11), 991-998.

El-Ghaish, S., Dalgalarrondo, M., Choiset, Y., Sitohy, M., Ivanova, I., Haertlé, T., & Chobert, J.-M. (2010). Characterization of a new isolate of *Lactobacillus fermentum* IFO 3956 from Egyptian Ras cheese with proteolytic activity. *European Food Research and Technology, 230*(4), 635–643. https://doi.org/10.1007/s00217-009-1206-x

Ewe, J. A., Wan-Abdullah, W. N., Alias, A. K., & Liong, M. T. (2012). Effects of ultrasound on growth, bioconversion of isoflavones, and probiotic properties of parent and subsequent passages of *Lactobacillus fermentum* BT 8633 in biotin-supplemented soymilk. *Ultrasonics Sonochemistry, 19*(4), 890–900. https://doi.org/10.1016/j.ultsonch.2012.01.003

Fang, J., Wang, H., Zhou, Y., Zhang, H., Zhou, H., & Zhang, X. (2021). Slimy partners: The mucus barrier and gut microbiome in ulcerative colitis. *Experimental & Molecular Medicine, 53*(5), 772–787. https://doi.org/10.1038/s12276-021-00617-8

Feltis, B. A., Kim, A. S., Kinneberg, K. M., Lyerly, D. L., Wilkins, T. D., Erlandsen, S. L., & Wells, C. L. (1999). *Clostridium difficile* toxins may augment bacterial penetration of intestinal epithelium. *Archives of Surgery, 134*(11), 1235–1242. https://doi.org/10.1001/archsurg.134.11.1235

Fischetti, V. A. (2019). Surface proteins on Gram-positive bacteria. *Microbiology Spectrum, 7*(4), 10.1128/microbiolspec.gpp3-0012-2018. https://doi.org/10.1128/microbiolspec.gpp3-0012-2018

Frese, S. A., Mackenzie, D. A., Peterson, D. A., Schmaltz, R., Fangman, T., Zhou, Y., . . . Walter, J. (2013). Molecular characterization of host-specific biofilm formation in a vertebrate gut symbiont. *PLoS Genetics, 9*(12), e1004057. https://doi.org/10.1371/journal.pgen.1004057

Ganzle, M. G., & Follador, R. (2012). Metabolism of oligosaccharides and starch in lactobacilli: A review. *Frontiers in Microbiology, 3*, 340. https://doi.org/10.3389/fmicb.2012.00340

Garcia-Castillo, V., Komatsu, R., Clua, P., Indo, Y., Takagi, M., Salva, S., . . . Villena, J. (2019). Evaluation of the immunomodulatory activities of the probiotic strain *Lactobacillus fermentum* UCO-979C. *Frontiers in Immunology, 10.* https://doi.org/10.3389/fimmu.2019.01376

Garcia-Castillo, V., Marcial, G., Albarracín, L., Tomokiyo, M., Clua, P., Takahashi, H., . . . Villena, J. (2020). The exopolysaccharide of *Lactobacillus fermentum* UCO-979C is partially involved in its immunomodulatory effect and its ability to improve the resistance against *Helicobacter pylori* infection. *Microorganisms, 8*(4). https://doi.org/10.3390/microorganisms8040479

Garde, S., Chodisetti, P. K., & Reddy, M. (2021). Peptidoglycan: Structure, synthesis, and regulation. *EcoSal Plus, 9*(2). https://doi.org/10.1128/ecosalplus.ESP-0010-2020

Gaucher, F., Bonnassie, S., Rabah, H., Marchand, P., Blanc, P., Jeantet, R., & Jan, G. (2019). Review: Adaptation of beneficial *Propionibacteria, Lactobacilli,* and *Bifidobacteria* improves tolerance toward technological and digestive stresses. *Frontiers in Microbiology, 10.* https://doi.org/10.3389/fmicb.2019.00841

Gaur, G., & Gänzle, M. (2024). Biochemical characterization of HcrF from *Limosilactobacillus fermentum*, a NADH-dependent 2-ene reductase with activity on hydroxycinnamic acids. *Letters in Applied Microbiology, 77*(12). https://doi.org/10.1093/lambio/ovae109

Gharbi, Y., Fhoula, I., Ruas-Madiedo, P., Afef, N., Boudabous, A., Gueimonde, M., & Ouzari, H.-I. (2019). In-vitro characterization of potentially probiotic *Lactobacillus* strains isolated from human microbiota: Interaction with pathogenic bacteria and the enteric cell line HT29. *Annals of Microbiology, 69*(1), 61–72. https://doi.org/10.1007/s13213-018-1396-1

Giordani, B., Naldi, M., Croatti, V., Parolin, C., Erdoğan, Ü., Bartolini, M., & Vitali, B. (2023). Exopolysaccharides from vaginal lactobacilli modulate microbial biofilms. *Microbial Cell Factories, 22*(1), 45. https://doi.org/10.1186/s12934-023-02053-x

Gonçalves, O. S., de Assis, J. C. S., & Santana, M. F. (2022). Breaking the ICE: An easy workflow for identifying and analyzing integrative and conjugative elements in bacterial genomes. *Functional & Integrative Genomics, 22*(6), 1139–1145. https://doi.org/10.1007/s10142-022-00903-2

Gou, H.-Z., Zhang, Y.-L., Ren, L.-F., Li, Z.-J., & Zhang, L. (2022). How do intestinal probiotics restore the intestinal barrier? *Frontiers in Microbiology, 13.* https://doi.org/10.3389/fmicb.2022.929346

Grover, M., Vanuytsel, T., & Chang, L. (2024). Intestinal permeability in disorders of gut-brain interaction: From bench to bedside. *Gastroenterology.* https://doi.org/10.1053/j.gastro.2024.08.033

Guérin, M., Silva, C. R.-D., Garcia, C., & Remize, F. (2020). Lactic acid bacterial production of exopolysaccharides from fruit and vegetables and associated benefits. *Fermentation, 6*(4), 115. Retrieved from https://www.mdpi.com/2311-5637/6/4/115

Gunn, J. S., Bakaletz, L. O., & Wozniak, D. J. (2016). What's on the outside matters: The role of the extracellular polymeric substance of gram-negative biofilms in evading host immunity and as a target for therapeutic intervention. *Journal of Biological Chemistry, 291*(24), 12538–12546. https://doi.org/10.1074/jbc.R115.707547

Gupta, T., Kaur, H., Kapila, S., & Kapila, R. (2021). *Lactobacillus fermentum* (MTCC-5898) alleviates *Escherichia coli*-induced inflammatory responses in intestinal epithelial cells by modulating immune genes and NF-κB signaling. *Journal of Applied Microbiology, 131*(6), 3008–3017. https://doi.org/10.1111/jam.15153

Gwak, M. G., & Chang, S. Y. (2021). Gut-brain connection: Microbiome, gut barrier, and environmental sensors. *Immune Network, 21*(3), e20. https://doi.org/10.4110/in.2021.21.e20

Han, H., You, Y., Cha, S., Kim, T. R., Sohn, M., & Park, J. (2023). Multi-species probiotic strain mixture enhances intestinal barrier function by regulating inflammation and tight junctions in lipopolysaccharide-stimulated Caco-2 cells. *Microorganisms, 11*(3). https://doi.org/10.3390/microorganisms11030656

Harris, H. M. B., Ale, E. C., Reinheimer, J. A., Binetti, A. G., & O'Toole, P. W. (2018). Draft genome sequence of *Lactobacillus fermentum* Lf2, an exopolysaccharide-producing strain isolated from Argentine cheese. *Microbiology Resource Announcements, 7*(18). https://doi.org/10.1128/mra.01072-18

Hautefort, I., Roels, A., Ladiré, M., Raibaud, P., Ducluzeau, R., & Fons, M. (1999). Selection of *Lactobacillus fermentum* strains able to durably colonize the digestive tract of mice harboring a complex human flora. *FEMS Microbiology Ecology, 29*(1), 23–31.

Hor, Y. Y., Ooi, C. H., Khoo, B. Y., Choi, S. B., Seeni, A., Shamsuddin, S., . . . Liong, M. T. (2019). *Lactobacillus* strains alleviated aging symptoms and aging-induced metabolic disorders in aged rats. *Journal of Medicinal Food, 22*(1), 1–13. https://doi.org/10.1089/jmf.2018.4229

Hossain, T. J. (2022). Functional genomics of the lactic acid bacterium *Limosilactobacillus fermentum* LAB-1: Metabolic, probiotic, and biotechnological perspectives. *Heliyon, 8*(11).

Hou, K., Wu, Z.-X., Chen, X.-Y., Wang, J.-Q., Zhang, D., Xiao, C., . . . Chen, Z.-S. (2022). Microbiota in health and diseases. *Signal Transduction and Targeted Therapy, 7*(1), 135. https://doi.org/10.1038/s41392-022-00974-4

Huang, J., Li, J., Li, Q., Li, L., Zhu, N., Xiong, X., & Li, G. (2020). Peptidoglycan derived from *Lactobacillus rhamnosus* MLGA up-regulates the expression of chicken β-defensin 9 without triggering an inflammatory response. *Innate Immunity, 26*(8), 733–745. https://doi.org/10.1177/1753425920949917

Ibrahim, S. A. (2016). *Lactic acid bacteria: Lactobacillus spp.: Other species*. In *Reference Module in Food Science*. Elsevier.

Illeghems, K., De Vuyst, L., & Weckx, S. (2015). Comparative genome analysis of the candidate functional starter culture strains *Lactobacillus fermentum* 222 and *Lactobacillus plantarum* 80 for controlled cocoa bean fermentation processes. *BMC Genomics, 16*, 766. https://doi.org/10.1186/s12864-015-1927-0

Inglin, R. C., Meile, L., & Stevens, M. J. A. (2018). Clustering of pan- and core-genome of *Lactobacillus* provides novel evolutionary insights for differentiation. *BMC Genomics, 19*(1), 284. https://doi.org/10.1186/s12864-018-4601-5

Jang, Y. J., Kim, W. K., Han, D. H., Lee, K., & Ko, G. (2019). *Lactobacillus fermentum* species ameliorate dextran sulfate sodium-induced colitis by regulating the immune response and altering gut microbiota. *Gut Microbes, 10*(6), 696–711. https://doi.org/10.1080/19490976.2019.1589281

Janssen, A. W. F., van der Lugt, B., Duivenvoorde, L. P. M., Vos, A. P., Bastiaan-Net, S., Tomassen, M. M. M., . . . van der Zande, M. (2024). Comparison of iPSC-derived human intestinal epithelial cells with Caco-2 cells and human in vivo data after exposure to *Lactiplantibacillus plantarum* WCFS1. *Scientific Reports, 14*(1), 26464. https://doi.org/10.1038/s41598-024-74802-w

Jayashree, S., Pooja, S., Pushpanathan, M., Rajendhran, J., & Gunasekaran, P. (2014). Identification and characterization of bile salt hydrolase genes from the genome of *Lactobacillus fermentum* MTCC 8711. *Applied Biochemistry and Biotechnology, 174*(2), 855–866. https://doi.org/10.1007/s12010-014-1118-5

Jayashree, S., Pooja, S., Pushpanathan, M., Vishnu, U., Sankarasubramanian, J., Rajendhran, J., & Gunasekaran, P. (2013). Genome sequence of *Lactobacillus fermentum* strain MTCC 8711, a probiotic bacterium isolated from yogurt. *Genome Announcements, 1*(5). https://doi.org/10.1128/genomeA.00770-13

Jha, R., Das, R., Oak, S., & Mishra, P. (2020). Probiotics (direct-fed microbials) in poultry nutrition and their effects on nutrient utilization, growth and laying performance, and gut health: A systematic review. *Animals, 10*(10), 1863. Retrieved from https://www.mdpi.com/2076-2615/10/10/1863

Jimenez, E., Langa, S., Martin, V., Arroyo, R., Martin, R., Fernandez, L., & Rodriguez, J. M. (2010). Complete genome sequence of *Lactobacillus fermentum* CECT 5716, a probiotic strain isolated from human milk. *Journal of Bacteriology, 192*(18), 4800. https://doi.org/10.1128/JB.00702-10

Jurášková, D., Ribeiro, S. C., & Silva, C. C. G. (2022). Exopolysaccharides produced by lactic acid bacteria: From biosynthesis to health-promoting properties. *Foods, 11*(2). https://doi.org/10.3390/foods11020156

Kang, W., Jeong, S.-Y., Cha, I.-S., Cho, M. S., Jeong, D.-Y., Kim, B.-Y., & Lee, J. H. (2020). Complete genome sequence of *Lactobacillus fermentum* SRCM103285 isolated from Korean rice wine, makgeolli. *미생물학회지, 56*(3), 337–339.

Kaur, H., Gupta, T., Kapila, S., & Kapila, R. (2022). *Lactobacillus fermentum* (MTCC-5898)-based fermented whey renders prophylactic action against colitis by strengthening the gut barrier function and maintaining immune homeostasis. *Microbial Pathogenesis, 173*, 105887.

Kaur, S., Sharma, P., Mayer, M. J., Neuert, S., Narbad, A., & Kaur, S. (2023). Beneficial effects of GABA-producing potential probiotic *Limosilactobacillus fermentum* L18 of human origin on intestinal permeability and human gut microbiota. *Microbial Cell Factories, 22*(1), 256. https://doi.org/10.1186/s12934-023-02264-2

Kawai, T., & Akira, S. (2009). The roles of TLRs, RLRs, and NLRs in pathogen recognition. *International Immunology, 21*(4), 317–337. https://doi.org/10.1093/intimm/dxp017

Khalil, E. S., Abd Manap, M. Y., Mustafa, S., Alhelli, A. M., & Shokryazdan, P. (2018). Probiotic properties of exopolysaccharide-producing *Lactobacillus* strains isolated from tempoyak. *Molecules, 23*(2). https://doi.org/10.3390/molecules23020398

Khatoon, Z., McTiernan, C. D., Suuronen, E. J., Mah, T. F., & Alarcon, E. I. (2018). Bacterial biofilm formation on implantable devices and approaches to its treatment and prevention. *Heliyon, 4*(12), e01067. https://doi.org/10.1016/j.heliyon.2018.e01067

Khlestkin, V. K., Lockachuk, M. N., Savkina, O. A., Kuznetsova, L. I., Pavlovskaya, E. N., & Parakhina, O. I. (2022). Taxonomic structure of bacterial communities in sourdoughs of spontaneous fermentation. *Vavilovskii Zhurnal Genet Selektsii, 26*(4), 385–393. https://doi.org/10.18699/vjgb-22-47

Kim, S., Lee, H. H., Kang, C. H., Kang, H., & Cho, H. (2023). Immune-enhancing effects of *Limosilactobacillus fermentum* in BALB/c mice immunosuppressed by cyclophosphamide. *Nutrients, 15*(4). https://doi.org/10.3390/nu15041038

Konyanee, A., Yotpanya, P., Panya, M., Engchanil, C., Suebwongsa, N., Namwat, W., . . . Lulitanond, V. (2019). Genome sequence of *Lactobacillus fermentum* 47-7, a good in vitro

probiotic strain isolated from a healthy Thai infant. *Microbiology Resource Announcements, 8*(39), e01014–19. https://doi.org/10.1128/mra.01014-19

Ksiezarek, M., Grosso, F., Ribeiro, T. G., & Peixe, L. (2022). Genomic diversity of genus *Limosilactobacillus*. *Microbial Genomics, 8*(7). https://doi.org/10.1099/mgen.0.000847

Kumari, M., Swarnkar, M. K., Kumar, S., Singh, A. K., & Gupta, M. (2015). Genome sequence of a potential probiotic strain, *Lactobacillus fermentum* HFB3, isolated from a human gut. *Genome Announcements, 3*(6). https://doi.org/10.1128/genomeA.01296-15

Kus, M., Ibragimow, I., & Piotrowska-Kempisty, H. (2023). Caco-2 cell line standardization with pharmaceutical requirements and in vitro model suitability for permeability assays. *Pharmaceutics, 15*(11). https://doi.org/10.3390/pharmaceutics15112523

Kwon, M., Lee, J., Park, S., Kwon, O.-H., Seo, J., & Roh, S. (2020). Exopolysaccharide isolated from *Lactobacillus plantarum* L-14 has anti-inflammatory effects via the toll-like receptor 4 pathway in LPS-induced RAW 264.7 cells. *International Journal of Molecular Sciences, 21*(23), 9283. Retrieved from https://www.mdpi.com/1422-0067/21/23/9283

La Fata, G., Weber, P., & Mohajeri, M. H. (2018). Probiotics and the gut immune system: Indirect regulation. *Probiotics and Antimicrobial Proteins, 10*(1), 11–21. https://doi.org/10.1007/s12602-017-9322-6

Lacerda, D. C., Trindade da Costa, P. C., Pontes, P. B., Carneiro Dos Santos, L. A., Cruz Neto, J. P. R., Silva Luis, C. C., . . . de Brito Alves, J. L. (2022). Potential role of *Limosilactobacillus fermentum* as a probiotic with anti-diabetic properties: A review. *World Journal of Diabetes, 13*(9), 717–728. https://doi.org/10.4239/wjd.v13.i9.717

Laiño, J., Villena, J., Kanmani, P., & Kitazawa, H. (2016). Immunoregulatory effects triggered by lactic acid bacteria exopolysaccharides: New insights into molecular interactions with host cells. *Microorganisms, 4*(3). https://doi.org/10.3390/microorganisms4030027

Le, J., Kulatheepan, Y., & Jeyaseelan, S. (2023). Role of toll-like receptors and nod-like receptors in acute lung infection. *Frontiers in Immunology, 14*, 1249098. https://doi.org/10.3389/fimmu.2023.1249098

Lebeer, S., Claes, I. J., Verhoeven, T. L., Vanderleyden, J., & De Keersmaecker, S. C. (2011). Exopolysaccharides of *Lactobacillus rhamnosus* GG form a protective shield against innate immune factors in the intestine. *Microbial Biotechnology, 4*(3), 368–374. https://doi.org/10.1111/j.1751-7915.2010.00199.x

Lechuga, S., Braga-Neto, M. B., Naydenov, N. G., Rieder, F., & Ivanov, A. I. (2023). Understanding disruption of the gut barrier during inflammation: Should we abandon traditional epithelial cell lines and switch to intestinal organoids? *Frontiers in Immunology, 14*, 1108289. https://doi.org/10.3389/fimmu.2023.1108289

Lee, I. C., Tomita, S., Kleerebezem, M., & Bron, P. A. (2013). The quest for probiotic effector molecules—Unraveling strain specificity at the molecular level. *Pharmacological Research, 69*(1), 61–74. https://doi.org/10.1016/j.phrs.2012.09.010

Lee, S., You, H. J., Kwon, B., & Ko, G. (2017). Complete genome sequence of the plasmid-bearing *Lactobacillus fermentum* strain SNUV175, a probiotic for women's health isolated from the vagina of a healthy South Korean woman. *Genome Announcements, 5*(12), e00045–17. https://doi.org/10.1128/genomeA.00045-17

Lehri, B., Seddon, A. M., & Karlyshev, A. V. (2015a). *Lactobacillus fermentum* 3872 genome sequencing reveals plasmid and chromosomal genes potentially involved in a probiotic activity. *FEMS Microbiology Letters, 362*(11). https://doi.org/10.1093/femsle/fnv068

Lehri, B., Seddon, A. M., & Karlyshev, A. V. (2015b). *Lactobacillus fermentum* 3872 genome sequencing reveals plasmid and chromosomal genes potentially involved in a probiotic activity. *FEMS Microbiology Letters, 362*(11). https://doi.org/10.1093/femsle/fnv068

Lehri, B., Seddon, A. M., & Karlyshev, A. V. (2017a). *Lactobacillus fermentum* 3872 as a potential tool for combatting *Campylobacter jejuni* infections. *Virulence, 8*(8), 1753–1760. https://doi.org/10.1080/21505594.2017.1362533

Lehri, B., Seddon, A. M., & Karlyshev, A. V. (2017b). Potential probiotic-associated traits revealed from completed high-quality genome sequence of *Lactobacillus fermentum* 3872. *Standards in Genomic Science, 12*, 19. https://doi.org/10.1186/s40793-017-0228-4

Lépine, A. F. P., de Wit, N., Oosterink, E., Wichers, H., Mes, J., & de Vos, P. (2018). *Lactobacillus acidophilus* attenuates *Salmonella*-induced stress of epithelial cells by modulating tight-junction genes and cytokine responses. *Frontiers in Microbiology, 9*. https://doi.org/10.3389/fmicb.2018.01439

Li, D., & Wu, M. (2021). Pattern recognition receptors in health and diseases. *Signal Transduction and Targeted Therapy, 6*(1), 291. https://doi.org/10.1038/s41392-021-00687-0

Li, F., Li, X., Cheng, C. C., Bujdoš, D., Tollenaar, S., Simpson, D. J., . . . Zheng, J. (2023). A phylogenomic analysis of *Limosilactobacillus reuteri* reveals ancient and stable evolutionary relationships with rodents and birds and zoonotic transmission to humans. *BMC Biology, 21*(1), 53. https://doi.org/10.1186/s12915-023-01541-1

Limoli, D. H., Jones, C. J., & Wozniak, D. J. (2015). Bacterial extracellular polysaccharides in biofilm formation and function. *Microbiology Spectrum, 3*(3), 10.1128/microbiolspec.mb-0011-2014. https://doi.org/10.1128/microbiolspec.mb-0011-2014

Lv, R., Gao, X., Zhang, C., Lian, W., Quan, X., Guo, S., & Chen, X. (2023). Characteristics and whole-genome analysis of *Limosilactobacillus fermentum* phage LFP02. *Foods, 12*(14), 2716. Retrieved from https://www.mdpi.com/2304-8158/12/14/2716

Lye, H. S., Kato, T., Low, W. Y., Taylor, T. D., Prakash, T., Lew, L. C., . . . Liong, M. T. (2017). *Lactobacillus fermentum* FTDC 8312 combats hypercholesterolemia via alteration of gut microbiota. *Journal of Biotechnology, 262*, 75–83. https://doi.org/10.1016/j.jbiotec.2017.09.007

Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., . . . Polouchine, N. (2006). Comparative genomics of the lactic acid bacteria. *Proceedings of the National Academy of Sciences, 103*(42), 15611–15616.

Mann, S., Park, M. S., Johnston, T. V., Ji, G. E., Hwang, K. T., & Ku, S. (2021). Isolation, characterization, and biosafety evaluation of *Lactobacillus fermentum* OK with potential oral probiotic properties. *Probiotics and Antimicrobial Proteins, 13*(5), 1363–1386. https://doi.org/10.1007/s12602-021-09761-z

Mariat, D., Firmesse, O., Levenez, F., Guimarães, V. D., Sokol, H., Doré, J., . . . Furet, J. P. (2009). The *Firmicutes/Bacteroidetes* ratio of the human microbiota changes with age. *BMC Microbiology, 9*(1), 123. https://doi.org/10.1186/1471-2180-9-123

Mazziotta, C., Tognon, M., Martini, F., Torreggiani, E., & Rotondo, J. C. (2023). Probiotics mechanism of action on immune cells and beneficial effects on human health. *Cells, 12*(1). https://doi.org/10.3390/cells12010184

Mei, L., Chen, Y., Wang, J., Lu, J., Zhao, J., Zhang, H., . . . Chen, W. (2022). *Lactobacillus fermentum* stimulates intestinal secretion of immunoglobulin A in an individual-specific manner. *Foods, 11*(9), 1229. Retrieved from https://www.mdpi.com/2304-8158/11/9/1229

Mendes-Soares, H., Suzuki, H., Hickey, R. J., & Forney, L. J. (2014). Comparative functional genomics of *Lactobacillus* spp. reveals possible mechanisms for specialization of vaginal lactobacilli to their environment. *Journal of Bacteriology, 196*(7), 1458–1470. https://doi.org/10.1128/jb.01439-13

Metz, J. K., Wiegand, B., Schnur, S., Knoth, K., Schneider-Daum, N., Groß, H., . . . Hittinger, M. (2020). Modulating the barrier function of human alveolar epithelial (hAELVi) cell monolayers as a model of inflammation. *Alternatives to Laboratory Animals, 48*(5–6), 252–267. https://doi.org/10.1177/0261192920983015

Morifuji, M., Kitade, M., Fukasawa, T., Yamaji, T., & Ichihashi, M. (2017). Exopolysaccharides isolated from milk fermented with lactic acid bacteria prevent ultraviolet-induced skin damage in hairless mice. *International Journal of Molecular Sciences, 18*(1), 146. Retrieved from https://www.mdpi.com/1422-0067/18/1/146

Morita, H., Toh, H., Fukuda, S., Horikawa, H., Oshima, K., Suzuki, T., . . . Hattori, M. (2008). Comparative genome analysis of *Lactobacillus reuteri* and *Lactobacillus fermentum* reveal a genomic island for reuterin and cobalamin production. *DNA Research, 15*(3), 151–161. https://doi.org/10.1093/dnares/dsn009

Moser, C., Jensen, P., Thomsen, K., Kolpen, M., Rybtke, M., Lauland, A. S., . . . Tolker-Nielsen, T. (2021). Immune responses to *Pseudomonas aeruginosa* biofilm infections. *Frontiers in Immunology, 12*, 625597. https://doi.org/10.3389/fimmu.2021.625597

Moskova-Doumanova, V., Vaseva, A., Veleva, R., Mladenova, K., Melniska, D., Doumanov, J., . . . Danova, S. (2024). In vitro effects of postmetabolites from *Limosilactobacillus fermentum* 53 on the survival and proliferation of HT-29 cells. *Microorganisms, 12*(7). https://doi.org/10.3390/microorganisms12071365

Naghmouchi, K., Belguesmia, Y., Bendali, F., Spano, G., Seal, B. S., & Drider, D. (2020a). *Lactobacillus fermentum*: A bacterial species with potential for food preservation and biomedical

applications. *Critical Reviews in Food Science and Nutrition, 60*(20), 3387–3399. https://doi.org/10.1080/10408398.2019.1688250

Ncbi.nlm.nih.gov. (2021). *Limosilactobacillus fermentum* (ID 711) - Genome - NCBI. Retrieved from https://www.ncbi.nlm.nih.gov/genome/?term=Lactobacillus+fermentum%5BOrganism%5D

Netrusov, A. I., Liyaskina, E. V., Kurgaeva, I. V., Liyaskina, A. U., Yang, G., & Revin, V. V. (2023). Exopolysaccharides producing bacteria: A review. *Microorganisms, 11*(6), 1541. Retrieved from https://www.mdpi.com/2076-2607/11/6/1541

Nguyen, P. T., Nguyen, T. T., Bui, D. C., Hong, P. T., Hoang, Q. K., & Nguyen, H. T. (2020). Exopolysaccharide production by lactic acid bacteria: The manipulation of environmental stresses for industrial applications. *AIMS Microbiology, 6*(4), 451–469. https://doi.org/10.3934/microbiol.2020027

Nichols, R. G., & Davenport, E. R. (2021). The relationship between the gut microbiome and host gene expression: A review. *Human Genetics, 140*(5), 747–760. https://doi.org/10.1007/s00439-020-02237-0

Noda, M., Danshiitsoodol, N., Kanno, K., Uchida, T., & Sugiyama, M. (2021). The exopolysaccharide produced by *Lactobacillus paracasei* IJH-SONE68 prevents and ameliorates inflammatory responses in DSS-induced ulcerative colitis. *Microorganisms, 9*(11), 2243. Retrieved from https://www.mdpi.com/2076-2607/9/11/2243

Nowak, B., Śróttek, M., Ciszek-Lenda, M., Skałkowska, A., Gamian, A., Górska, S., & Marcinkiewicz, J. (2020). Exopolysaccharide from *Lactobacillus rhamnosus* KL37 inhibits T cell-dependent immune response in mice. *Archivum Immunologiae et Therapiae Experimentalis, 68*(3), 17. https://doi.org/10.1007/s00005-020-00581-7

O'Donnell, M. M., O'Toole, P. W., & Ross, R. P. (2013). Catabolic flexibility of mammalian-associated lactobacilli. *Microbial Cell Factories, 12*(1), 48. https://doi.org/10.1186/1475-2859-12-48

Ohland, C. L., & MacNaughton, W. K. (2010). Probiotic bacteria and intestinal epithelial barrier function. *American Journal of Physiology-Gastrointestinal and Liver Physiology, 298*(6), G807–G819. https://doi.org/10.1152/ajpgi.00243.2009

Ojha, A. K., Shah, N. P., & Mishra, V. (2021). Conjugal transfer of antibiotic resistances in *Lactobacillus* spp. *Current Microbiology, 78*(8), 2839–2849. https://doi.org/10.1007/s00284-021-02554-1

Okumura, R., & Takeda, K. (2017). Roles of intestinal epithelial cells in the maintenance of gut homeostasis. *Experimental & Molecular Medicine, 49*(5), e338. https://doi.org/10.1038/emm.2017.20

Oliphant, K., & Allen-Vercoe, E. (2019). Macronutrient metabolism by the human gut microbiome: Major fermentation by-products and their impact on host health. *Microbiome, 7*(1), 91. https://doi.org/10.1186/s40168-019-0704-8

Oliveira-Nascimento, L., Massari, P., & Wetzler, L. M. (2012). The role of TLR2 in infection and immunity. *Frontiers in Immunology, 3*, 79. https://doi.org/10.3389/fimmu.2012.00079

Owusu-Kwarteng, J., Tano-Debrah, K., Akabanda, F., & Jespersen, L. (2015). Technological properties and probiotic potential of *Lactobacillus fermentum* strains isolated from West African fermented millet dough. *BMC Microbiology, 15*(1), 261. https://doi.org/10.1186/s12866-015-0602-6

Ozhegov, G., Pavlova, A., Zhuravleva, D., Gogoleva, N., Shagimardanova, E., Markelova, M., . . . Kayumov, A. (2020). Whole genome sequence data of *Lactobacillus fermentum* HFD1, the producer of antibacterial peptides. *Data in Brief, 32*, 106105.

Palani Kumar, M. K., Halami, P. M., & Serva Peddha, M. (2021). Effect of *Lactobacillus fermentum* MCC2760-based probiotic curd on hypercholesterolemic C57BL6 mice. *ACS Omega, 6*(11), 7701–7710. https://doi.org/10.1021/acsomega.1c00045

Paone, P., & Cani, P. D. (2020). Mucus barrier, mucins and gut microbiota: The expected slimy partners? *Gut, 69*(12), 2232–2243. https://doi.org/10.1136/gutjnl-2020-322260

Park, D.-Y., Hwang, J., Kim, Y., Kim, Y.-Y., Kim, H.-S., & Hwang, I. (2022). Whole-genome sequence of *Limosilactobacillus fermentum* strain DM075, isolated from the human oral cavity. *Microbiology Resource Announcements, 11*(11), e00819-00822. https://doi.org/10.1128/mra.00819-22

Park, M. R., Shin, M., Mun, D., Jeong, S.-Y., Jeong, D.-Y., Song, M., . . . Oh, S. (2020). Probiotic *Lactobacillus fermentum* strain JDFM216 improves cognitive behavior and modulates immune response with gut microbiota. *Scientific Reports, 10*(1), 21701. https://doi.org/10.1038/s41598-020-77587-w

Pasolli, E., De Filippis, F., Mauriello, I. E., Cumbo, F., Walsh, A. M., Leech, J., . . . Ercolini, D. (2020). Large-scale genome-wide analysis links lactic acid bacteria from food with the gut microbiome. *Nature Communications, 11*(1), 2610. https://doi.org/10.1038/s41467-020-16438-8

Paulino do Nascimento, L. C., Lacerda, D. C., Ferreira, D. J. S., de Souza, E. L., & de Brito Alves, J. L. (2022). *Limosilactobacillus fermentum*, current evidence on the antioxidant properties and opportunities to be exploited as a probiotic microorganism. *Probiotics and Antimicrobial Proteins, 14*(5), 960–979. https://doi.org/10.1007/s12602-022-09943-3

Paveljšek, D., Juvan, P., Košir, R., Rozman, D., Hacin, B., Ivičak-Kocjan, K., & Rogelj, I. (2018). *Lactobacillus fermentum* L930BB and *Bifidobacterium animalis* subsp. *animalis* IM386 initiate signalling pathways involved in intestinal epithelial barrier protection. *Beneficial Microbes, 9*(3), 515–525. https://doi.org/10.3920/bm2017.0107

Pei, Z., Sadiq, F. A., Han, X., Zhao, J., Zhang, H., Ross, R. P., . . . Chen, W. (2021). Comprehensive scanning of prophages in *Lactobacillus*: Distribution, diversity, antibiotic resistance genes, and linkages with CRISPR-Cas systems. *mSystems, 6*(3), e01211-01220. https://doi.org/10.1128/msystems.01211-20

Peng, Z., Wei, B., Huang, T., Liu, Z., Guan, Q., Xie, M., . . . Xiong, T. (2020). Screening, safety evaluation, and mechanism of two *Lactobacillus fermentum* strains in reducing the translocation of *Staphylococcus aureus* in the Caco-2 monolayer model. *Frontiers in Microbiology, 11*. https://doi.org/10.3389/fmicb.2020.566473

Pereira, D. I., & Gibson, G. R. (2002). Cholesterol assimilation by lactic acid bacteria and *Bifidobacteria* isolated from the human gut. *Applied and Environmental Microbiology, 68*(9), 4689–4693.

Pereira, D. I., McCartney, A. L., & Gibson, G. R. (2003). An in vitro study of the probiotic potential of a bile-salt-hydrolyzing *Lactobacillus fermentum* strain, and determination of its cholesterol-lowering properties. *Applied and Environmental Microbiology, 69*(8), 4743–4752.

Pessione, E. (2012). Lactic acid bacteria contribution to gut microbiota complexity: Lights and shadows. *Frontiers in Cellular and Infection Microbiology, 2*. https://doi.org/10.3389/fcimb.2012.00086

Peters, D. T., Reifs, A., Alonso-Caballero, A., Madkour, A., Waller, H., Kenny, B., . . . Lakey, J. H. (2022). Unraveling the molecular determinants of the anti-phagocytic protein cloak of plague bacteria. *PLOS Pathogens, 18*(3), e1010447. https://doi.org/10.1371/journal.ppat.1010447

Proctor, L. M., Creasy, H. H., Fettweis, J. M., Lloyd-Price, J., Mahurkar, A., Zhou, W., . . . The Integrative HMP Research Network Consortium. (2019). The integrative human microbiome project. *Nature, 569*(7758), 641–648. https://doi.org/10.1038/s41586-019-1238-8

Pswarayi, F., & Gänzle, M. G. (2019). Composition and origin of the fermentation microbiota of Mahewu, a Zimbabwean fermented cereal beverage. *Applied and Environmental Microbiology, 85*(11), e03130-03118. https://doi.org/10.1128/AEM.03130-18

Racines, M. P., Solis, M. N., Šefcová, M. A., Herich, R., Larrea-Álvarez, M., & Revajová, V. (2023). An overview of the use and applications of *Limosilactobacillus fermentum* in broiler chickens. *Microorganisms, 11*(8). https://doi.org/10.3390/microorganisms11081944

Rackaityte, E., & Lynch, S. V. (2020). The human microbiome in the 21st century. *Nature Communications, 11*(1), 5256. https://doi.org/10.1038/s41467-020-18983-8

Rajagopal, M., & Walker, S. (2017). Envelope structures of Gram-positive bacteria. *Current Topics in Microbiology and Immunology, 404*, 1–44. https://doi.org/10.1007/82_2015_5021

Ren, C., Zhang, Q., de Haan, B. J., Zhang, H., Faas, M. M., & de Vos, P. (2016). Identification of TLR2/TLR6 signalling lactic acid bacteria for supporting immune regulation. *Scientific Reports, 6*(1), 34561. https://doi.org/10.1038/srep34561

Ren, S., Zhang, X., Guan, H., Wu, L., Yu, M., Hou, D., . . . Fang, X. (2021). *Lactobacillus acidipiscis* induced regulatory gamma delta T cells and attenuated experimental autoimmune encephalomyelitis. *Frontiers in Immunology, 12*. https://doi.org/10.3389/fimmu.2021.623451

Rodríguez-Sojo, M. J., Ruiz-Malagón, A. J., Rodríguez-Cabezas, M. E., Gálvez, J., & Rodríguez-Nogales, A. (2021). *Limosilactobacillus fermentum* CECT5716: Mechanisms and therapeutic insights. *Nutrients, 13*(3). https://doi.org/10.3390/nu13031016

Ruiz, L., Margolles, A., & Sánchez, B. (2013). Bile resistance mechanisms in *Lactobacillus* and *Bifidobacterium*. *Frontiers in Microbiology, 4*, 396. https://doi.org/10.3389/fmicb.2013.00396

Russo, M., Marquez, A., Herrera, H., Abeijon-Mukdsi, C., Saavedra, L., Hebert, E., . . . Medina, R. (2020). Oral administration of *Lactobacillus fermentum* CRL1446 improves biomarkers of metabolic syndrome in mice fed a high-fat diet supplemented with wheat bran. *Food & Function, 11*(5), 3879–3894.

Salas-Jara, M. J., Ilabaca, A., Vega, M., & García, A. (2016). Biofilm-forming *Lactobacillus*: New challenges for the development of probiotics. *Microorganisms, 4*(3). https://doi.org/10.3390/microorganisms4030035

Salminen, M. K., Rautelin, H., Tynkkynen, S., Poussa, T., Saxelin, M., Valtonen, V., & Järvinen, A. (2006). *Lactobacillus* bacteremia, species identification, and antimicrobial susceptibility of 85 blood isolates. *Clinical Infectious Diseases, 42*(5), e35–e44. https://doi.org/10.1086/500214

Salvetti, E., Harris, H. M. B., Felis, G. E., & O'Toole, P. W. (2018). Comparative genomics of the genus *Lactobacillus* reveals robust phylogroups that provide the basis for reclassification. *Applied and Environmental Microbiology, 84*(17), e00993-00918. https://doi.org/10.1128/AEM.00993-18

Sambuy, Y., De Angelis, I., Ranaldi, G., Scarino, M. L., Stammati, A., & Zucco, F. (2005). The Caco-2 cell line as a model of the intestinal barrier: Influence of cell and culture-related factors on Caco-2 cell functional characteristics. *Cell Biology and Toxicology, 21*(1), 1–26. https://doi.org/10.1007/s10565-005-0085-6

Sánchez, I., Palop, L., & Ballesteros, C. (2000). Biochemical characterization of lactic acid bacteria isolated from spontaneous fermentation of *Almagro* eggplants. *International Journal of Food Microbiology, 59*(1–2), 9–17. https://doi.org/10.1016/s0168-1605(00)00256-7

Santana, P. T., Rosas, S. L. B., Ribeiro, B. E., Marinho, Y., & de Souza, H. S. P. (2022). Dysbiosis in inflammatory bowel disease: Pathogenic role and potential therapeutic targets. *International Journal of Molecular Sciences, 23*(7). https://doi.org/10.3390/ijms23073464

Sengupta, R., Altermann, E., Anderson, R. C., McNabb, W. C., Moughan, P. J., & Roy, N. C. (2013). The role of cell surface architecture of lactobacilli in host-microbe interactions in the gastrointestinal tract. *Mediators of Inflammation, 2013*, 237921. https://doi.org/10.1155/2013/237921

Sengupta, R., Anderson, R. C., Altermann, E., McNabb, W. C., Ganesh, S., Armstrong, K. M., . . . Roy, N. C. (2015). *Lactobacillus fermentum* AGR1487 cell surface structures and supernatant increase paracellular permeability through different pathways. *MicrobiologyOpen, 4*(4), 541–552. https://doi.org/10.1002/mbo3.260

Sharma, M., Huber, E., Arnesdotter, E., Behrsing, H. P., Bettmann, A., Brandwein, D., . . . Clippinger, A. J. (2025). Minimum information for reporting on the TEER (trans-epithelial/endothelial electrical resistance) assay (MIRTA). *Archives of Toxicology, 99*(1), 57–66. https://doi.org/10.1007/s00204-024-03879-z

Sharma, R., Kapila, R., Kapasiya, M., Saliganti, V., Dass, G., & Kapila, S. (2014). Dietary supplementation of milk fermented with probiotic *Lactobacillus fermentum* enhances systemic immune response and antioxidant capacity in aging mice. *Nutrition Research, 34*(11), 968–981. https://doi.org/10.1016/j.nutres.2014.09.006

Silhavy, T. J., Kahne, D., & Walker, S. (2010). The bacterial cell envelope. *Cold Spring Harbor Perspectives in Biology, 2*(5), a000414. https://doi.org/10.1101/cshperspect.a000414

Singh, S., Datta, S., Narayanan, K. B., & Rajnish, K. N. (2021). Bacterial exopolysaccharides in biofilms: Role in antimicrobial resistance and treatments. *Journal of Genetic Engineering and Biotechnology, 19*(1), 140. https://doi.org/10.1186/s43141-021-00242-y

Soloveva, I. V., Novikova, N. A., Tochilina, A. G., Belova, I. V., Kashnikov, A. Y., Sashina, T. A., . . . Molodtsova, S. B. (2021). The probiotic strain *Lactobacillus fermentum* 39: Biochemical properties, genomic features, and antiviral activity. *Microbiology, 90*(2), 219–225. https://doi.org/10.1134/S0026261721020132

Sørensen, H. M., Rochfort, K. D., Maye, S., MacLeod, G., Brabazon, D., Loscher, C., & Freeland, B. (2022). Exopolysaccharides of lactic acid bacteria: Production, purification, and health benefits towards functional food. *Nutrients, 14*(14). https://doi.org/10.3390/nu14142938

Srimahaeak, T., Bianchi, F., Chlumsky, O., Larsen, N., & Jespersen, L. (2021). In-vitro study of *Limosilactobacillus fermentum* PCC adhesion to and integrity of the Caco-2 cell monolayers as affected by pectins. *Journal of Functional Foods, 79*, 104395.

Srinivasan, B., Kolli, A. R., Esch, M. B., Abaci, H. E., Shuler, M. L., & Hickman, J. J. (2015). TEER measurement techniques for in vitro barrier model systems. *Journal of Laboratory Automation, 20*(2), 107–126. https://doi.org/10.1177/2211068214561025

Stadler, C., Sanivarapu, R., Mashaal, H., & Iqbal, J. (2018). A rare pathogen: *Lactobacillus fermentum* in empyema. *CHEST, 154*(4), 207A. https://doi.org/10.1016/j.chest.2018.08.181

Stefanello, R. F., Nabeshima, E. H., Iamanaka, B. T., Ludwig, A., Fries, L. L. M., Bernardi, A. O., & Copetti, M. V. (2019). Survival and stability of *Lactobacillus fermentum* and *Wickerhamomyces anomalus* strains upon lyophilisation with different cryoprotectant agents. *Food Research International, 115*, 90–94. https://doi.org/10.1016/j.foodres.2018.07.044

Strober, W., & Watanabe, T. (2011). NOD2, an intracellular innate immune sensor involved in host defense and Crohn's disease. *Mucosal Immunology, 4*(5), 484–495. https://doi.org/10.1038/mi.2011.29

Su, L., Ma, F., An, Z., Ji, X., Zhang, P., Yue, Q., . . . Zhao, L. (2022). The metabolites of *Lactobacillus fermentum* F-B9-1 relieved dextran sulfate sodium-induced experimental ulcerative colitis in mice. *Frontiers in Microbiology, 13*, 865925. https://doi.org/10.3389/fmicb.2022.865925

Subhadra, B., Krier, J., Hofstee, K., Monsul, N., & Berkes, E. (2015). Draft whole-genome sequence of *Lactobacillus fermentum* LfQi6, derived from the human microbiome. *Genome Announcements, 3*(3). https://doi.org/10.1128/genomeA.00423-15

Sudeepa, E., & Bhavini, K. (2020). Review on *Lactobacillus fermentum*.

Sugahara, H., Kato, S., Nagayama, K., Sashihara, K., & Nagatomi, Y. (2022). Heterofermentative lactic acid bacteria such as *Limosilactobacillus* as a strong inhibitor of aldehyde compounds in plant-based milk alternatives. *Frontiers in Sustainable Food Systems, 6*. https://doi.org/10.3389/fsufs.2022.965986

Sun, J., Qi, C., Zhu, H., Zhou, Q., Xiao, H., Le, G., . . . Yu, R. (2019). IgA-targeted *Lactobacillus jensenii* modulated gut barrier and microbiota in high-fat diet-fed mice. *Frontiers in Microbiology, 10*. https://doi.org/10.3389/fmicb.2019.01179

Sun, M., Wu, W., Liu, Z., & Cong, Y. (2017). Microbiota metabolite short-chain fatty acids, GPCR, and inflammatory bowel diseases. *Journal of Gastroenterology, 52*, 1–8.

Sun, Z., Zhang, W., Bilige, M., & Zhang, H. (2015). Complete genome sequence of the probiotic *Lactobacillus fermentum* F-6 isolated from raw milk. *Journal of Biotechnology, 194*, 110–111. https://doi.org/10.1016/j.jbiotec.2014.12.010

Tang, Y., Dong, W., Wan, K., Zhang, L., Li, C., Zhang, L., & Liu, N. (2015). Exopolysaccharide produced by *Lactobacillus plantarum* induces maturation of dendritic cells in BALB/c mice. *PLOS ONE, 10*(11), e0143743. https://doi.org/10.1371/journal.pone.0143743

Tanoue, T., Umesaki, Y., & Honda, K. (2010). Immune responses to gut microbiota-commensals and pathogens. *Gut Microbes, 1*(4), 224–233. https://doi.org/10.4161/gmic.1.4.12613

Tarannum, N., Ali, F., Khan, M. S., Alhumaidan, O. S., Zawad, A. N. M. S., & Hossain, T. J. (2024). Bioactive exopolysaccharide from *Limosilactobacillus fermentum* LAB-1: Antioxidant, anti-inflammatory, antibacterial, and antibiofilm properties. *Bioactive Carbohydrates and Dietary Fibre, 31*, 100409. https://doi.org/10.1016/j.bcdf.2024.100409

Teame, T., Wang, A., Xie, M., Zhang, Z., Yang, Y., Ding, Q., . . . Zhou, Z. (2020). Paraprobiotics and postbiotics of probiotic *Lactobacilli*, their positive effects on the host, and action mechanisms: A review. *Frontiers in Nutrition, 7*. https://doi.org/10.3389/fnut.2020.570344

Thapa, S., & Prasanna, R. (2018). Prospecting the characteristics and significance of the phyllosphere microbiome. *Annals of Microbiology, 68*(5), 229–245. https://doi.org/10.1007/s13213-018-1331-5

Thoda, C., & Touraki, M. (2023). Immunomodulatory properties of probiotics and their derived bioactive compounds. *Applied Sciences, 13*(8), 4726. Retrieved from https://www.mdpi.com/2076-3417/13/8/4726

Tochilina, A. G., Belova, I. V., Soloveva, I. V., Ivanova, T. P., & Zhirnov, V. A. (2019). Bioinformatic analysis of the genome of the *Lactobacillus fermentum* 90 TC-4 production strain.

Molecular Genetics, Microbiology and Virology, 34(3), 176–181. https://doi.org/10.3103/S0891416819030078

Tomaro-Duchesneau, C., Saha, S., Malhotra, M., Jones, M., Rodes, L., & Prakash, S. (2015). *Lactobacillus fermentum* NCIMB 5221 and NCIMB 2797 as cholesterol-lowering probiotic biotherapeutics: *In vitro* analysis. *Beneficial Microbes, 6*(6), 861–869.

Torrens, G., & Cava, F. (2024). Mechanisms conferring bacterial cell wall variability and adaptivity. *Biochemical Society Transactions, 52*(5), 1981–1993. https://doi.org/10.1042/bst20230027

Trindade, B. C., & Chen, G. Y. (2020). NOD1 and NOD2 in inflammatory and infectious diseases. *Immunological Reviews, 297*(1), 139–161. https://doi.org/10.1111/imr.12902

Ullah, M., Rizwan, M., Han, J., Raza, A., Chen, Y., Yan, M., . . . Chen, H. (2024). Comparative probio-genomics analysis of *Limosilactobacillus fermentum* 3872. *Probiotics and Antimicrobial Proteins*. https://doi.org/10.1007/s12602-024-10286-4

van Reenen, C. A., & Dicks, L. M. T. (2011). Horizontal gene transfer amongst probiotic lactic acid bacteria and other intestinal microbiota: What are the possibilities? A review. *Archives of Microbiology, 193*(3), 157–168. https://doi.org/10.1007/s00203-010-0668-3

Verce, M., Vuyst, L. D., & Weckx, S. (2018). Complete and annotated genome sequence of the sourdough lactic acid bacterium *Lactobacillus fermentum* IMDO 130101. *Genome Announcements, 6*(19), e00256-00218. https://doi.org/10.1128/genomeA.00256-18

Verhoeckx, K., Cotter, P., López-Expósito, I., Kleiveland, C., Lea, T., Mackie, A., . . . Wichers, H. (Eds.). (2015). *The impact of food bioactives on health: In vitro and ex vivo models.* Springer. https://doi.org/10.1007/978-3-319-16104-4.

Vollmer, W., Blanot, D., & De Pedro, M. A. (2008). Peptidoglycan structure and architecture. *FEMS Microbiology Reviews, 32*(2), 149–167. https://doi.org/10.1111/j.1574-6976.2007.00094.x

Wafula, E. N., Brinks, E., Becker, B., Huch, M., Trierweiler, B., Mathara, J. M., . . . Franz, C. (2017). Draft genome sequence of *Lactobacillus fermentum* BFE 6620, a potential starter culture for African vegetable foods, isolated from fermented cassava. *Genome Announcements, 5*(33). https://doi.org/10.1128/genomeA.00801-17

Walter, J. (2008). Ecological role of *Lactobacilli* in the gastrointestinal tract: Implications for fundamental and biomedical research. *Applied and Environmental Microbiology, 74*(16), 4985–4996. https://doi.org/10.1128/AEM.00753-08

Wang, J. E., Jørgensen, P. F., Almlöf, M., Thiemermann, C., Foster, S. J., Aasen, A. O., & Solberg, R. (2000a). Peptidoglycan and lipoteichoic acid from *Staphylococcus aureus* induce tumor necrosis factor alpha, interleukin 6 (IL-6), and IL-10 production in both T cells and monocytes in a human whole blood model. *Infection and Immunity, 68*(7), 3965–3970. https://doi.org/10.1128/iai.68.7.3965-3970.2000

Wang, W., Li, Y., Han, G., Li, A., & Kong, X. (2022). *Lactobacillus fermentum* CECT5716 alleviates the inflammatory response in asthma by regulating TLR2/TLR4 expression. *Frontiers in Nutrition, 9.* https://doi.org/10.3389/fnut.2022.931427

Wang, Y., Wu, J., Lv, M., Shao, Z., Hungwe, M., Wang, J., . . . Geng, W. (2021). Metabolism characteristics of lactic acid bacteria and the expanding applications in the food industry. *Frontiers in Bioengineering and Biotechnology, 9.* https://doi.org/10.3389/fbioe.2021.612285

Weaver, A., Taguchi, A., & Dörr, T. (2023). Masters of misdirection: Peptidoglycan glycosidases in bacterial growth. *Journal of Bacteriology, 205*(3), e00428-00422. https://doi.org/10.1128/jb.00428-22

Wei, Y., Li, F., Li, L., Huang, L., & Li, Q. (2019). Genetic and biochemical characterization of an exopolysaccharide with *in vitro* antitumoral activity produced by *Lactobacillus fermentum* YL-11. *Frontiers in Microbiology, 10.* https://doi.org/10.3389/fmicb.2019.02898

Werning, M. L., Hernández-Alcántara, A. M., Ruiz, M. J., Soto, L. P., Dueñas, M. T., López, P., & Frizzo, L. S. (2022). Biological functions of exopolysaccharides from lactic acid bacteria and their potential benefits for humans and farmed animals. *Foods, 11*(9), 1284. Retrieved from https://www.mdpi.com/2304-8158/11/9/1284

Wu, H. J., & Wu, E. (2012). The role of gut microbiota in immune homeostasis and autoimmunity. *Gut Microbes, 3*(1), 4–14. https://doi.org/10.4161/gmic.19320

Wu, W., & Li, H. (2018). Metabolites of lactic acid bacteria. In W. Chen & A. Narbad (Eds.), *Lactic Acid Bacteria in Foodborne Hazards Reduction: Physiology to Practice* (pp. 87–113). Springer Singapore.

Xiu-Wen, W., Ru-Feng, W., Ming, Y., Wei, X., & Xiu-Wei, Y. (2013). Dulbecco's modified eagle's medium and minimum essential medium–which one is more preferred for establishment of Caco-2 cell monolayer model used in evaluation of drug absorption? *Die Pharmazie - An International Journal of Pharmaceutical Sciences, 68(10),* 805-810.

Xiu, L., Zhang, H., Hu, Z., Liang, Y., Guo, S., Yang, M., . . . Wang, X. (2018). Immunostimulatory activity of exopolysaccharides from probiotic *Lactobacillus casei* WXD030 strain as a novel adjuvant *in vitro* and *in vivo. Food and Agricultural Immunology, 29(1),* 1086-1105. https://doi.org/10.1080/09540105.2018.1513994

Yadav, A. K., Espaillat, A., & Cava, F. (2018). Bacterial strategies to preserve cell wall integrity against environmental threats. *Frontiers in Microbiology, 9.* https://doi.org/10.3389/fmicb.2018.02064

Yan, F., & Polk, D. B. (2020). Probiotics and probiotic-derived functional factors—mechanistic insights into applications for intestinal homeostasis. *Frontiers in Immunology, 11.* https://doi.org/10.3389/fimmu.2020.01428

Yang, F., Wang, A., Zeng, X., Hou, C., Liu, H., & Qiao, S. (2015). *Lactobacillus reuteri* I5007 modulates tight junction protein expression in IPEC-J2 cells with LPS stimulation and in newborn

piglets under normal conditions. *BMC Microbiology, 15(1),* 32. https://doi.org/10.1186/s12866-015-0372-1

Yang, X., Hong, J., Wang, L., Cai, C., Mo, H., Wang, J., . . . Liao, Z. (2024). Effect of lactic acid bacteria fermentation on plant-based products. *Fermentation, 10(1),* 48. Retrieved from https://www.mdpi.com/2311-5637/10/1/48

Yildiz, H., & Karatas, N. (2018). Microbial exopolysaccharides: Resources and bioactive properties. *Process Biochemistry, 72,* 41-46.

Yoo, D., Bagon, B. B., Valeriano, V. D. V., Oh, J. K., Kim, H., Cho, S., & Kang, D.-K. (2017). Complete genome analysis of *Lactobacillus fermentum* SK152 from kimchi reveals genes associated with its antimicrobial activity. *FEMS Microbiology Letters, 364(18).*

Yotpanya, P., Panya, M., Engchanil, C., Suebwongsa, N., Namwat, W., Thaw, H., & Lulitanond, V. (2016). Probiotic characterization of lactic acid bacteria isolated from infants' feces and its application for the expression of green fluorescent protein. *Malaysian Journal of Microbiology,* 76-84.

Young, V. B. (2017). The role of the microbiome in human health and disease: An introduction for clinicians. *BMJ, 356,* j831. https://doi.org/10.1136/bmj.j831

Yu, A. O., Leveau, J. H. J., & Marco, M. L. (2020). Abundance, diversity and plant-specific adaptations of plant-associated lactic acid bacteria. *Environmental Microbiology Reports, 12(1),* 16-29. https://doi.org/10.1111/1758-2229.12794

Yu, S., Sun, Y., Shao, X., Zhou, Y., Yu, Y., Kuai, X., & Zhou, C. (2022). Leaky gut in IBD: Intestinal barrier-gut microbiota interaction. *J Microbiol Biotechnol, 32(7),* 825-834. https://doi.org/10.4014/jmb.2203.03022

Yuan, L., van der Mei, H. C., Busscher, H. J., & Peterson, B. W. (2020). Two-stage interpretation of changes in TEER of intestinal epithelial layers protected by adhering *Bifidobacteria* during *E. coli* challenges. *Frontiers in Microbiology, 11.* https://doi.org/10.3389/fmicb.2020.599555

Zhang, Y., Ding, Y., & Guo, Q. (2022). Probiotic species in the management of periodontal diseases: An overview. *Frontiers in Cellular and Infection Microbiology, 12,* 806463. https://doi.org/10.3389/fcimb.2022.806463

Zhao, M., Bao, J., Wang, Z., Du, S., Gao, C., Nan, D., . . . Ge, G. (2023). Evaluation of the fermentation performance and functional properties of bacterial communities of amaranth silage supplemented with *Limosilactobacillus fermentum* and *Latilactobacillus graminis. Chemical and Biological Technologies in Agriculture, 10(1),* 103. https://doi.org/10.1186/s40538-023-00476-7

Zhao, Y., Yu, L., Tian, F., Zhao, J., Zhang, H., Chen, W., . . . Zhai, Q. (2022). Environment-related genes analysis of *Limosilactobacillus fermentum* isolated from food and human gut: Genetic diversity and adaptation evolution. *Foods, 11(19).* https://doi.org/10.3390/foods11193135

Zhao, Y., Yu, L., Tian, F., Zhao, J., Zhang, H., Chen, W., & Zhai, Q. (2021). An optimized culture medium to isolate *Lactobacillus fermentum* strains from the human intestinal tract. *Food & Function, 12(15),* 6740-6754.

Zheng, D., Liwinski, T., & Elinav, E. (2020). Interaction between microbiota and immunity in health and disease. *Cell Research, 30(6),* 492-506. https://doi.org/10.1038/s41422-020-0332-7

Zheng, J., Wittouck, S., Salvetti, E., Franz, C. M. A. P., Harris, H. M. B., Mattarelli, P., . . . Lebeer, S. (2020). A taxonomic note on the genus *Lactobacillus*: Description of 23 novel genera, emended description of the genus *Lactobacillus Beijerinck* 1901, and union of *Lactobacillaceae* and *Leuconostocaceae*. *International Journal of Systematic and Evolutionary Microbiology, 70*(4), 2782–2858. https://doi.org/10.1099/ijsem.0.004107

Jang, S. (2018). *Inhibitory effect of Lactobacillus fermentum SNUV175 and Lactobacillus crispatus SNUV220 isolated from healthy Korean women on vulvovaginal candidiasis* (Master's thesis). **Seoul National University Graduate School.** https://hdl.handle.net/10371/143927

Chapter 3: TEER and Growth Phenotyping for AGR1485 and AGR1487

3.1. Abstract

Background: Probiotic bacteria, including *Limosilactobacillus fermentum*, are generally considered safe for consumption. However, strain AGR1487 of *L. fermentum* has been identified as an exception, as it disrupts gut barrier integrity in Caco-2 cells and induces colon inflammation in germ-free mice, unlike strain AGR1485, which was isolated from a healthy human concurrently.

Aim: This study aimed to confirm the barrier-disruptive phenotype of AGR1487 and to examine any additional phenotypic differences between strains AGR1485 and AGR1487, which had not been previously reported.

Methods: The barrier-disruptive phenotype was confirmed using transepithelial electrical resistance (TEER) assays on Caco-2 cell monolayers. Bacterial growth was monitored by measuring the culture turbidity at OD_{600} with a spectrophotometer. Cell morphology and autoaggregation properties were examined by electron microscopy.

Results: In TEER assays, co-culture with AGR1487 resulted in a significant decline in barrier integrity, with a drop of –20.11 %TEER per hour, compared to AGR1485, which exhibited minimal effects (2.90 %TEER per hour) similar to the control (5.24 %TEER per hour). Electron microscopy revealed that AGR1487 cells tend to autoaggregate, a feature not observed in AGR1485. It was also noted that the media format (M199 without foetal bovine serum) used for TEER assays did not support robust bacterial growth, as confirmed by OD_{600} measurements.

Conclusion: These results confirm that *L. fermentum* AGR1487 disrupts barrier integrity in Caco-2 cells and suggest that the gene candidate(s) are likely expressed before the TEER assay in support of past findings that dead AGR1487 cells also create the phenotype. These phenotypes underscore the need for further investigation into their respective mechanisms of action. Sequencing both strains will enable examining their gene content, which is necessary to identify the gene candidate(s) responsible for the barrier-disruptive phenotype.

3.2. Introduction

Limosilactobacillus fermentum (formerly *Lactobacillus fermentum*) is a Gram-positive, heterofermentative bacterium characterised by short, long, thin rods arranged in pairs or chains (McClung, 1987). Commonly found in cereal and vegetable fermentations, this species has long been used for food preservation. It produces bacteriocins, lactic acid, and exopolysaccharides, which help it compete with pathogenic microorganisms (Alfano et al., 2020; Di Cagno et al., 2008; Sanchez, Palop, & Ballesteros, 2000).

Due to its phenotypic diversity, *L. fermentum* is extensively studied as a probiotic. For example, Jo et al. (2021) evaluated 25 strains. They identified *L. fermentum* MG7011 for its superior gastrointestinal tract (GIT) tolerance, adhesion to Caco-2 and HT-29 cells, antioxidative and anti-inflammatory activities, and ability to protect the epithelial barrier. Other strains have shown benefits, such as alleviating hyperuricemia in rats (Wu et al., 2021), preventing lactational mastitis in breastfeeding women (Hurtado et al., 2017), and reducing cholesterol in both rats and humans (Kullisaar, Zilmer, Salum, Rehema, & Zilmer, 2016; Pan, Zeng, & Yan, 2011; Tomaro-Duchesneau et al., 2015). Despite these promising attributes, few studies have explored potentially detrimental phenotypes within this species.

One notable exception is strain AGR1487, which exhibits increased resistance to bile acids and low pH, along with distinct carbon utilisation patterns compared to strain AGR1485 (Sengupta et al., 2015). More importantly, AGR1487 compromises intestinal barrier integrity, as demonstrated by a reduction in transepithelial electrical resistance (TEER) in Cancer coli-2 cells (Caco-2) derived from colorectal adenocarcinoma cells, a phenomenon not observed with AGR1485 (Sengupta et al., 2015).

TEER is a widely accepted quantitative method for assessing tight junction integrity in epithelial and endothelial cell models (Srinivasan et al., 2015). Caco-2 cells are cultured on semi-permeable membranes in these assays to form a monolayer that mimics the intestinal barrier. Bacteria

are added to the apical compartment (above the monolayer), and any disruption of tight junctions allows electrical current to pass more readily, signalling compromised barrier function.

In experiments using FBS-free M199 media, *L. fermentum* AGR1487 was shown to uniformly degrade tight junctions, as confirmed by tight junction staining (R. C. Anderson et al., 2013). Notably, this disruptive effect occurred with live cells, dead cells, and cell debris, whereas the bacterial supernatant had no effect. Moreover, mixing AGR1487 with AGR1485 in different ratios did not mitigate AGR1487's barrier-disruptive phenotype. Despite these observations, the exact mechanism by which AGR1487 impairs barrier integrity remains unclear (Sengupta et al., 2015).

In vivo, the presence of AGR1487 in germ-free rats induces colon inflammation, a response not observed with AGR1485 (Rachel C. Anderson et al., 2016). AGR1487 appeared to activate the immune response through Toll-like receptors (TLRs) and the proinflammatory regulator Nuclear Factor Kappa-Light-Chain-Enhancer of Activated B Cells (NF-κB), which led to enhanced dendritic cell maturation, cytokine secretion, and increased infiltration of macrophages and lymphocytes in the colon. Although both strains were administered at the same inoculation concentration, AGR1487 was recovered significantly throughout the GIT, up to 1,000-fold higher in the caecum, colon, and faeces. Yet, inflammation was detected only in the colon (Rachel C. Anderson et al., 2016).

These findings underscore the complex and dual nature of *L. fermentum* strains, highlighting the need to understand their beneficial and potentially detrimental effects on host health and further explore their phenotypic differences of AGR1485 and AGR1487.

3.3. Aims and Hypotheses

Because media composition can affect Caco-2 cell physiology, the TEER methodology from previous experimental designs established by Sengupta et al. (2015) was used to ensure reproducibility. Although animal testing is beyond the scope of this PhD program, further phenotypic characterisation of *L. fermentum* AGR1487 relative to AGR1485 may guide future research.

Aim 1: Reproduce the disruptive barrier phenotype in TEER assays using a Caco-2 monolayer grown in M199.

Aim 2: Expand the phenotype characterisation of both *L. fermentum* AGR1485 and AGR1487.

Hypothesis: *L. fermentum* AGR1487 reproducibly reduces TEER in Caco-2 cell monolayers, but *L. fermentum* AGR1485 does not.

3.4. Materials and methods

3.4.1. Cell culture media and broths

Numerous cell culture media were needed to maintain the bacterial strains and Caco-2 cell line. The recipes for all media formats are listed below (Table 3.4-2).

Table 3.4-2: Cell growth media and bacterial growth broth.

GROWTH MEDIA	COMPONENTS AND RECIPE
MRS MEDIA	52.2 g of lactobacilli de Man-Rogosa-Sharpe (MRS) broth (Merck Ltd., Auckland, NZ) was added to 1 L of Milli-Q water and autoclaved for 15 minutes at 121 ºC before storage at room temperature.
OMNILOG 60% MRS MEDIA	31.32 g of lactobacilli de Man-Rogosa-Sharpe (MRS) broth (Merck Ltd., Auckland, NZ) was added to 1 L of Milli-Q water and filter sterilised with 0.22 µm filter. The media was made fresh with 1% v/v with redox dye mix G (Biolog, Hayward, USA) and used immediately.
MEM	0.5 mL of NEAA was added to 45 mL of MEM. The solution was mixed, filtered, and sterilised before storage at 4 ºC.
MEM WITH FBS	0.5 mL of NEAA and 4.5 mL of FBS were added to 45 mL of MEM. The solution was mixed, filtered, and sterilised before storage at 4 ºC
M199	0.5 mL of NEAA was added to 45 mL of M199. The solution was mixed, filtered, and sterilised before storage at 4 ºC
M199 WITH FBS	0.5 mL of NEAA and 4.5 mL of FBS were added to 45 mL of M199. The solution was mixed, filtered, and sterilised before storage at 4 ºC

MEM = minimal essential media, M199 = medium 199, FBS = foetal bovine serum, NEAA = non-essential amino acids. All human cell culture media were prepared to 50 mL. All reagents were sourced from Life Technologies, NZ, except MRS broth.

3.4.2. Culture of Caco-2 cells

Caco-2 cells (HTB-37 American Type Culture Collection (ATCC) were kindly provided at passage 16 by Dr Alicia Barnett (AgResearch Ltd, Grasslands, Palmerston North, NZ) and used at passage 28-33 (Briske-Anderson, Finley, & Newman, 1997). All cells, cell culture media, and reagents were handled without antibiotics and aseptically in various class II biological safety cabinets (laminar flow hood) provided by AgResearch and The Riddet Institute

3.4.3. Maintenance of Caco-2 cells

The Caco-2 cells were grown in 75 cm^2 rectangular canted-neck cell culture flasks (Corning, Lindfield, New South Wales, Australia) in 15 mL of minimal essential media (MEM) containing foetal bovine serum at 37 ºC in a 5% CO_2 atmosphere (Table 3.4-2). The cell culture volumes for various vessel sizes were calculated using Equation 3.4-1.

The cell culture media was first replaced 24 hours after seeding the cells and then every 2 to 3 days. The cells were passaged at around 80% confluence once a week after inspection by microscope. The passage was diluted 1:5 into a new culture flask using pre-warmed media and reagents to reduce stress on the cells.

$$\text{Volume}_{\text{minimum}} = \text{SA}\ (\text{cm}^2) \times 0.2\ \text{mL}$$

$$\text{Volume}_{\text{maximum}} = \text{SA}\ (\text{cm}^2) \times 0.5\ \text{mL}$$

$$\text{SA} = \text{surface area of the culture vessle}\ (\text{cm}^2)$$

Equation 3.4-1: The equations for determining the minimum and maximum amount of media required for cell culture.

3.4.4. Harvesting Caco-2 cells

The Caco-2 cells were dissociated from the flask (75 cm^2 flask) by incubating the cells at 37 ºC for 12 to 15 minutes in the presence of 4.0 mL of pre-warmed (37 ºC) TrypLE Express (Life Technologies, Penrose, Auckland, NZ). The flask was gently agitated to aid the detachment of any loosely adhered cells and reduce clumping. The flask was inspected using an inverted light microscope to ensure the cells were dislodged. The cell suspension was removed and added to 8.0 mL of fresh cell culture media to inactivate the TrypLE Express in a 15 mL Nunc tube. The cell suspension was centrifuged at 240 x g, the supernatant was removed, and the cells were gently resuspended in 5 mL of fresh cell culture media.

3.4.5. Caco-2 cell counting

Ten µL of 0.4% Trypan Blue solution (Life Technologies, Auckland, NZ) was mixed with 10 µL cell suspension acquired during cell harvesting (Section 3.4.4). Ten µL of the cell suspension with dye was gently pipetted into a cell counting chamber (ThermoFisher Scientific, for Applied Biosystems by AsureQuality, NZ) and allowed to fill by capillary action. The cells were counted in brightfield mode using the Countess Automated Cell Counter (Applied Biosystems, ThermoFisher Scientific, AsureQuality, NZ).

3.4.6. Culture of bacterial strains

L. fermentum AGR1485 and AGR1487 were acquired from the AgResearch culture collection (AgResearch, Grasslands, Palmerston North, NZ). The strains were previously isolated from the oral cavity of healthy volunteers (R. C. Anderson, Cookson, McNabb, Kelly, & Roy, 2010).

Bacteria were revived from frozen stocks (Section 3.4.7) by streaking onto MRS (Table 3.4-2) infused agar (Bacteriological agar, Sigma-Aldrich, Merck Ltd., Auckland, NZ) plates that were incubated for 24 hours at 37 ºC aerobically. Single colonies were picked and used to inoculate 5 mL of sterile MRS media in a 15 mL Nunc tube and grown overnight at 37 ºC without shaking.

3.4.7. Long-term storage of bacterial strains

Bacteria were stored in glycerol stocks made from 15% glycerol and MRS media at -80 ºC. Bacteria were revived from -80 ºC stocks by isolation streaking onto MRS agar plates and incubated aerobically overnight at 37 °C. A single colony was selected from an overnight plate to inoculate 5 mL of sterile MRS broth. The inoculated MRS broth was grown overnight at 37 ºC without shaking. The overnight culture was gently resuspended by pipetting, and 800 µL was then added to 200 µL of 75% glycerol in a 1.5 mL Eppendorf tube and mixed thoroughly by pipette. The Eppendorf tubes were transferred to a storage container and stored at -80 ºC.

3.4.8. TEER assay co-cultured with bacterial strains

The Caco-2 cells were seeded into inserts at 6.3 x 10^4 cells/cm^2. The TEER assays were conducted using Corning 6 insert PET Transwells with a pore size of 0.4 μm and a growth area of 4.67 cm^2 (Corning, Lindfield, New South Wales, Australia). The TEER resistance measurements were carried out by inserting the seeded Transwell insert into an EndOhm 24G Snapwell culture cup (World Precision Instruments, Sarasota, Florida, USA), connected to an EVOM 3 TEER meter with automatic data logging (World Precision Instruments, Sarasota, Florida, USA). Once the monolayers achieved 300–500 Ω·cm² or higher for at least a minute without significant fluctuations. Once adequate resistance was achieved, the apical medium was replaced with respective test-condition cell culture media and equilibrated for 24 hours before co-culture experiments.

An overnight culture of either AGR1485 or AGR1487 was pelleted at 4,000 x g for 15 minutes and washed with 0.1 M phosphate buffer (pH 7.2). The cells were resuspended and centrifuged twice in phosphate buffer before being resuspended in cell culture media to OD_{600} = 0.9 (Rachel C. Anderson et al., 2016; R. C. Anderson et al., 2013). Before cell suspensions were applied, each Transwell was tested for an initial TEER reading, recorded as the background resistance. The apical volume of the Transwells was then replaced with the bacterial suspensions, and the 0-time point measurement was taken.

Three independent biological replicates were performed, each using new bacterial colonies, with two technical replicates per condition on each plate (three plates in total, 6 × 3 wells measured).

TEER measurements were recorded every 2 hours for 8 hours. A final measurement was taken the following morning at the 24-hour mark. The TEER values were calculated following the formulas shown in Equation 3.4-2 and Equation 3.4-3.

$$\text{TEER} = (\text{raw resistance} - \text{background resistance}) \times \text{membrane area}$$

Equation 3.4-2: The equations for calculating the TEER from raw resistance values and removing background resistance.

$$\text{TEER}_{\text{Change}} = \frac{(\text{TEER}_{\text{current}} - \text{TEER}_{\text{initial}})}{\text{TEER}_{\text{initial}}} \times 100$$

Equation 3.4-3: The equation for calculating the change in TEER from background-corrected raw TEER values. TEERintial is the time point 0 before the treatment condition is added to the apical volume of the Transwell.

3.4.9. pH measurements

The pH measurements were made aerobically using a thermal probe with a PL-700ALS desktop pH meter (EZDO, NanKang District, Taipei City, Taiwan). All bacterial growth was completed aseptically in aerobic conditions. Frozen bacterial stocks were revived on MRS agar plates by streaking roughly 10 μL of stock with a sterile plastic disposable loop. The streaked MRS agar plates were incubated overnight at 37 ºC. Individual colonies for each replicate were picked and inoculated into 10 mL MRS broth and incubated overnight at 37 ºC without shaking in 15 mL Nunc tubes. The overnight cultures were centrifuged at 4,000 x g, and the MRS media supernatant was removed. The pellets were gently washed with 1 mL 0.1 M phosphate buffer, pH 7.2, before resuspending in 10 mL of the same phosphate buffer. The cells were pelleted and resuspended once more using the same procedure.

The bacterial suspensions were diluted into MEM with or without foetal bovine serum (FBS) to an OD_{600} = 0.9, replicating the methodology for the TEER assays. Four replicates of each strain (AGR1485 and AGR1487) against each condition (MEM with or without FBS) alongside a no-treatment control with no added bacteria of the same media composition. Due to pH being temperature tolerant and reducing the chance of cross-contamination, each condition was accompanied by a 15 mL-Nunc tube containing autoclaved MilliQ water. The MilliQ water was used for the thermal probe

while pH measurements were taken. The pH measurements were taken every hour for the first 3 hours and the following day, 20 hours after inoculating the MEM cell culture media.

The experiment was performed four times, each using a different bacterial colony to represent a biological replicate.

3.4.10. Bacterial growth assays

Growth experiments were conducted in various clean biosafety cabinets provided by Riddet Institute and AgResearch (Te Ohu Rangahau Kai, Massey, Palmerston North). Single colonies were picked from MRS recovery plates for overnight cultures (Section 3.4.6). Each growth experiment was conducted as eight biological replicates and repeated 2 to 3 times on separate titre plates over 24 hours in aerobic conditions (technical replicates). Overnight cultures were pelleted at 4,000 x g, the supernatant removed and resuspended in 10 mL of 0.1 M phosphate buffer, pH 7.2, twice before use. Cell growth media M199, MEM, M199 with FBS, MEM with FBS, FBS and MRS bacterial broth were inoculated with the washed bacterial strains to an OD_{600} = 0.2 without altering the atmosphere.

Two hundred µL of each condition (Table 3.4-2) were loaded into untreated, sterile, flatbottomed 96 well-titre plates (ThermoFisher Scientific, Auckland, NZ). The titre plates were sealed with microplate sealing tape (ThermoFisher Scientific, Auckland, NZ) before being relocated to the Multiskan Microplate Spectrophotometer (ThermoFisher Scientific, for Applied Biosystems by AsureQuality, NZ). Growth measurements (OD_{600}) were taken every 30 minutes for 24 hours with light shaking in a Multiskan (ThermoFisher Scientific, for Applied Biosystems by AsureQuality, NZ).

Each growth experiment comprised eight biological replicates, each derived from a separate bacterial colony, and was repeated on three separate titre plates, with each well on a plate representing a technical replicate. The order of the conditions across each plate was randomised to account for potential plate position effects.

3.4.11. Bacterial colony imaging

Approximately 10 µL of an overnight culture of both strains were independently streaked onto MRS agar without antibiotics using sterile plastic loops. The streaked plates were incubated overnight at 37 ºC aerobically in a temperature-controlled room. The colonies were examined using a Leica MZ12 (Leica microsystems, Bio-Strategy, NZ) stereomicroscope. Light-normalised, calibrated images of the colonies were taken with an attached DFC320 (Leica microsystems, Bio-Strategy, NZ) camera. Both microscope and camera were calibrated, maintained, and provided by the Manawatu Microscopy and Imaging Centre, Palmerston North.

3.4.12. Electron microscopy

Overnight cultures of both strains were pelleted in 15 mL Nunc tubes at 4,000 x g for 15 minutes. The pellets were washed in 0.1 M phosphate buffer (pH 7.2) and resuspended in 10 mL of 0.1 M phosphate buffer. The resuspension was centrifuged once more, and the supernatant was discarded. The pellets were resuspended in 10 mL of 0.1 M phosphate buffer. An aliquot of the cell suspension was diluted in phosphate buffer to an OD_{600} = 0.1 and fixed 1:1 with a fixing solution (3 % glutaraldehyde and 2% formaldehyde in 0.1 M phosphate buffer, pH 7.2) provided by the Manawatu Microscopy and Imaging Centre (Palmerston North, NZ).

Raoul Solomon prepared the fixed samples for transmission electron microscopy (TEM) (without any dye) and scanning electron microscopy (SEM) by sputter-coating with gold at the Manawatu Microscopy and Imaging Centre (Palmerston North, NZ).

Prepared TEM samples were viewed with the FEI Tecnai G2 Spirit BioTWIN electron microscope (Czech Republic). The SEM coupons with the fixed sample were mounted onto aluminium stubs and imaged with an FEI Quanta 200 (FEI Co., Hillsboro, OR, USA) electron microscope.

3.4.13. Statistics

The confidence intervals, standard deviations, and P-values were calculated in R, with replicate structures clearly defined in the analysis. Each growth experiment comprised eight biological replicates derived from separate bacterial colonies and was repeated on three independent titre plates, with each well representing a technical replicate. The order of conditions was randomised across plates to minimise positional effects. These statistical measures were then integrated into the graphing scripts provided in the digital appendix.

3.5. Results

3.5.1. Bacterial Caco-2 co-culture TEER assays

To assess the reliability and reproducibility of these findings, experiments were repeated under the same conditions and further experiments were performed using MEM media and foetal bovine serum (FBS). A 95% confidence interval (95% CI) represented the possible differences between the strains AGR1485, AGR1487, and a control (no bacteria added). The assay measured the percentage of transepithelial/transendothelial electrical resistance (%TEER), determined using Equation 3.4-2 and Equation 3.4-3. Each experiment concluded at the 24-hour mark.

L. fermentum AGR1487 caused a decrease in TEER across a Caco-2 monolayer approximately 3 hours post-initiation. A complete loss of resistance was detected after 7 hours of contact with AGR1487 and did not improve until the end of the 24-hour assay (Figure 3.5.1-1 A). AGR1487 elicited a maximum drop of -−427.31 % TEER between 3 to 24 hours, averaging −20.11 % TEER/h (95 % CI: −20.58 to −19.64) (Figure 3.5.1-1 A). Neither strain AGR1485 nor the control decreased TEER (Figure 3.5.1-1 A). Strain AGR1485 and the control presented positive slopes of 2.90 % TEER/h (95 % CI: 2.73–3.07) and 5.24 % TEER/h (95 % CI: 4.59–5.89), respectively. The overlapping 95% confidence intervals for the control and strain AGR1485 throughout the assay

(Figure 3.5.1-1 A) indicated no significant differences between these treatments in M199 without FBS in the apical volume.

Incorporating FBS into the apical M199 volume in the TEER assay compromised the assay's reliability, yielding inconclusive outcomes (Figure 3.5.1-1 B). While FBS did not alter the effect of AGR1487 on TEER in M199 media, it increased the variance for AGR1485 readings and the control's behaviour, which showed an increase in the integrity of the time of the assay that did not occur without FBS. Conversely, the 95% CI for AGR1485 indicated no significant difference among either strain or the control due to a wide variance when FBS was included with M199 (Figure 3.5.1-1 B).

Substituting M199 media with MEM yielded inconclusive results, as differences between conditions were not statistically significant because of the media and not the bacterial strains. Without FBS, all conditions, including the control, appeared to lose barrier integrity during the TEER assay. Even with FBS, the conditions could not be statistically differentiated due to wide variance. The root cause of this increased variance with MEM media remains unknown but may be related to insufficient nutrients since all conditions lost TEER over time without FBS. With FBS, the control behaved similarly to M199 with FBS, indicating that the observed effect is likely attributable to the media rather than the bacteria (Appendix: Figure 3.8.1-1 A and B).

Table 3.5-3: Media trials during TEER assays

MEDIA	OUTCOME	COMMENTS
M199	Reproducible	Reliable, reproducible TEER results.
M199 WITH FBS	Inconclusive	FBS generated high variation in TEER results when AGR1485 was co-cultured.
MEM	Inconclusive	All conditions lost TEER. Possibly insufficient nutrients.
MEM WITH FBS	Inconclusive	MEM media was incapable of generating enough resolution between conditions.

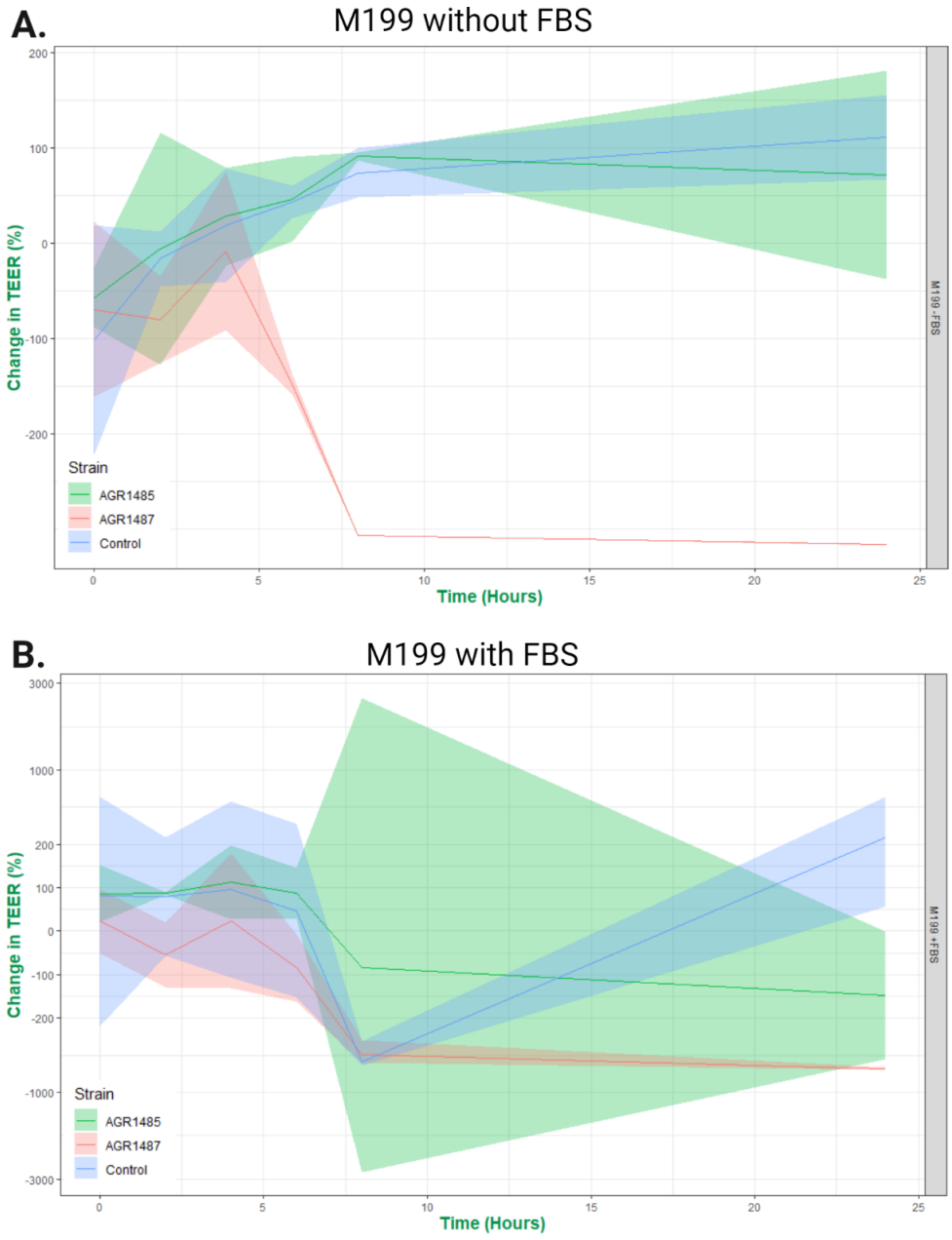

Figure 3.5.1-1: Caco-2-AGR1485/87 M199 TEER assays. The change in electrical resistance (% change) of Caco-2 cells grown in (A) M199 with or (B) without FBS along with either AGR1487 (Red), AGR1485 (Green) or without any bacteria (Control in Blue) in the apical volume over 24 hours. The shading represents the 95% confidence interval (CI) of the data shaded over the trend line. The control comprises Caco-2 cells treated the same as the other conditions but without bacteria. For each condition, n = 3.

3.5.2. Bacterial growth in different cell culture media

Both bacterial strains were expected to grow in the cell culture media, with FBS providing an environment conducive to lactic acid production. One hypothesis was that AGR1487's barrier-disruptive phenotype might be related to pH changes, as lactic acid, a metabolic product of *L. fermentum*, could compromise the barrier integrity of Caco-2 cells. Each media format contained phenol red as a pH indicator; healthy Caco-2 cells require a pH near seven and phenol red changes colour from yellow at pH 6.6 to bright pink or fuchsia at pH 8.1. It was anticipated that both strains would acidify the media, turning the dye yellow over 24 hours in the presence of FBS (Sengupta et al., 2015) and that AGR1487 might do so even in the absence of FBS. However, after 24 hours of incubation at 37 °C, neither strain induced a colour change in the phenol red indicator (Figure 3.5.2-1).

Each growth condition included a control without bacteria (Blank). In every media format, whether broth or FBS, the samples were inoculated with either AGR1485 or AGR1487 at an initial concentration of OD_{600} = 0.2. Although an increase in OD_{600} was detectable in bacteria-inoculated samples compared to the blank, no further increase in OD_{600} was observed for either strain in any media format or in FBS alone over 18 hours (Figure 3.5.2-2 and Figure 3.5.2-3).

In contrast, both AGR1485 and AGR1487 exhibited robust growth in MRS broth, exceeding the linear range defined by Beer's Law during the assay (Figure 3.5.2-3 and Figure 3.5.2-2). The strains displayed statistically similar lag phases (95% confidence interval) from inoculation until approximately 3 hours, after which they entered the log phase. AGR1485 transitioned into the stationary phase between 10 and 12 hours, whereas AGR1487 did not reach a stationary phase within the 18-hour experimental period. Notably, neither the standard cell culture media nor FBS alone supported the growth of either strain (Figure 3.5.2-3).

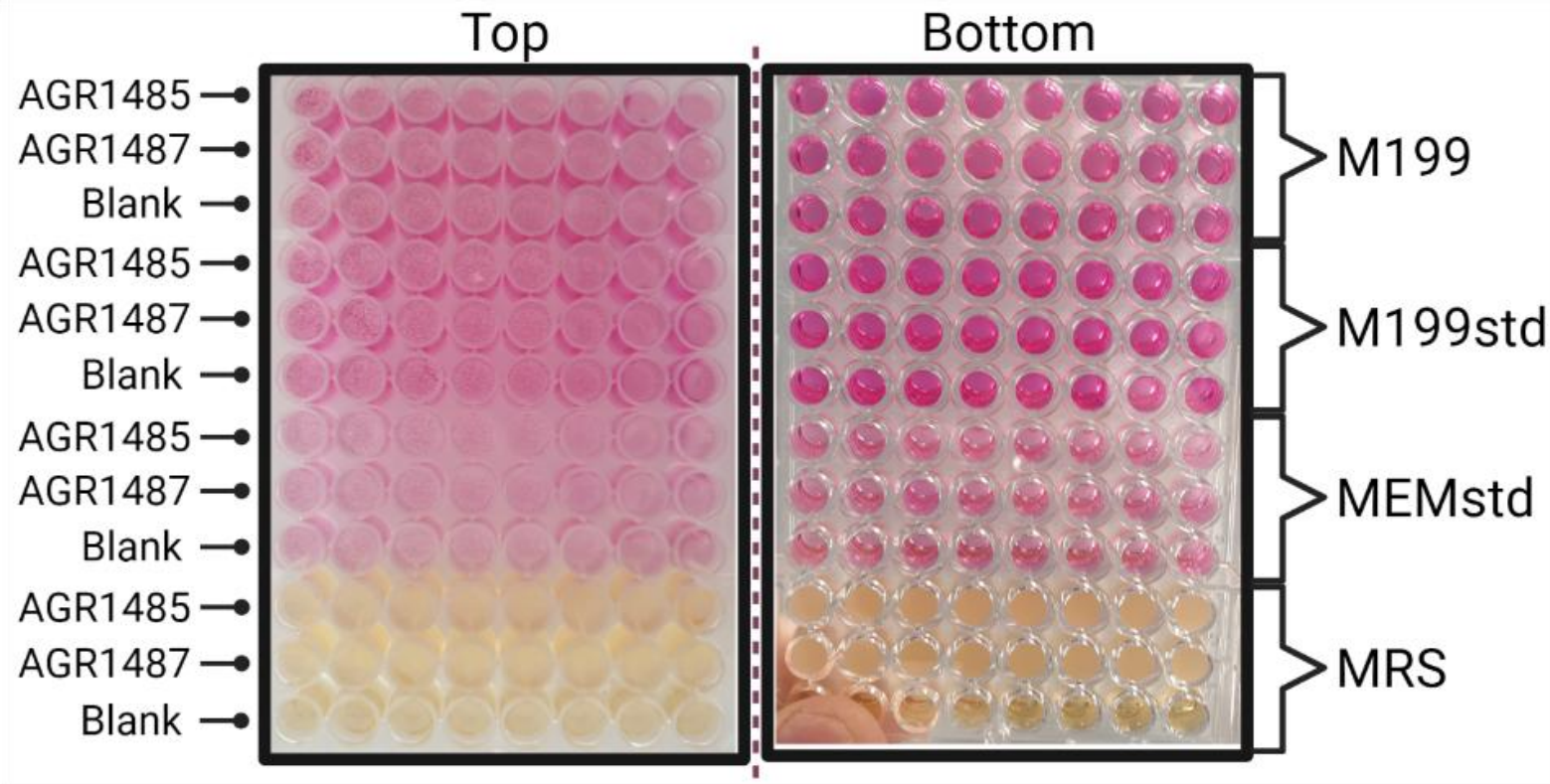

Figure 3.5.2-1: 96-well growth assay. One of three replicates from which data was collected for Figure 3.5.2-2 after 24 hours incubated at 37 ºC. Top: Photo taken from the top of the plate through the plate sealing tape. Bottom: The same plate was photographed from the bottom through the base of the plate. Phenol red was not included in MRS.

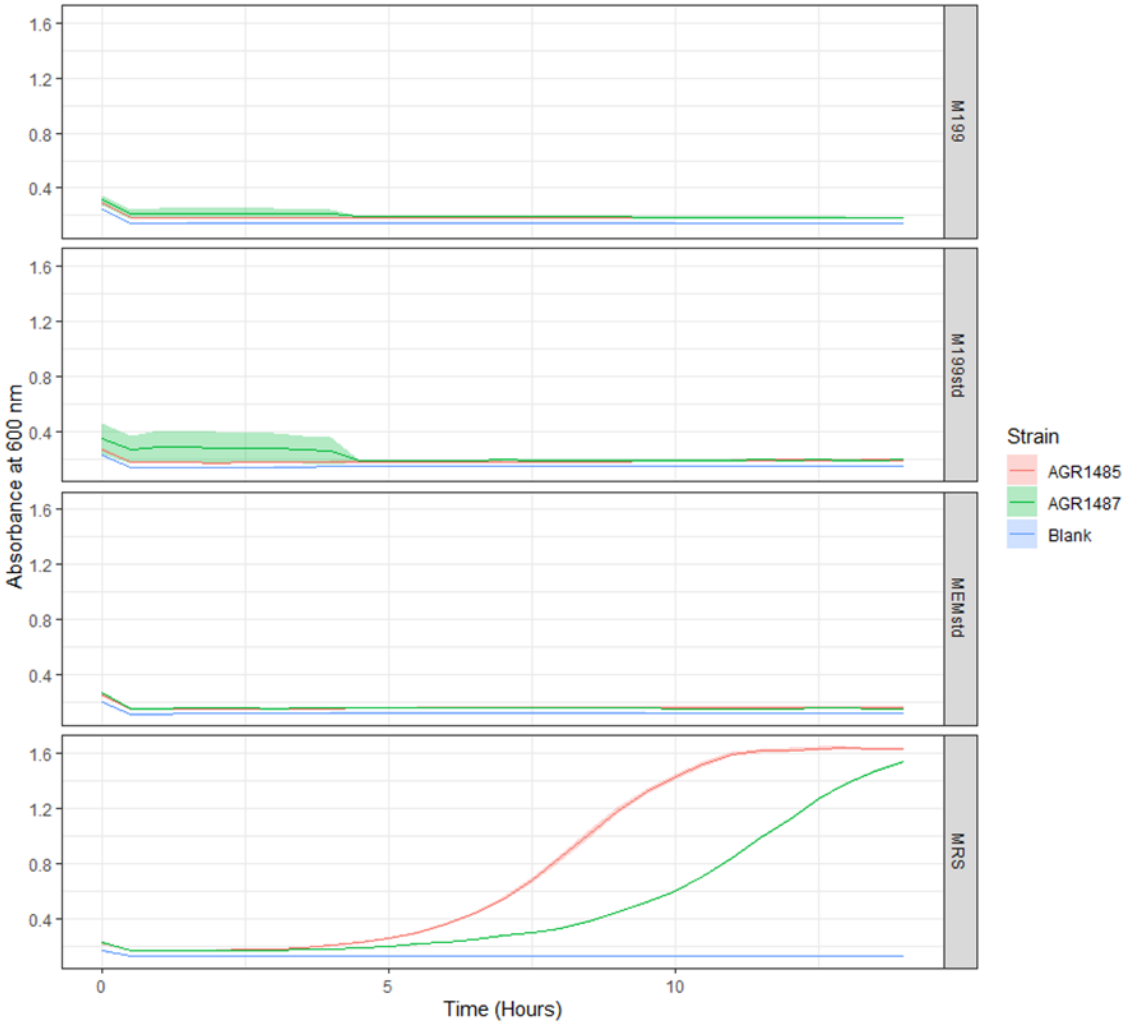

Figure 3.5.2-2: Bacterial growth in M199 media. The growth of *L. fermentum* AGR1485 (Red) and AGR1487 (Green) in M199, M199 with FBS (M199std), MEM with FBS (MEMstd) and MRS media, along with a control with no bacteria. The samples' optical density (OD600) was recorded in a 96-well titre plate reader over 24 hours. Shading corresponding to the sample colouring shows the sample's 95% confidential. N = 16 for each strain in each environmental condition. Sample sizes (N): M199 =32, M199std = 16, MEMstd = 24, MRS = 24.

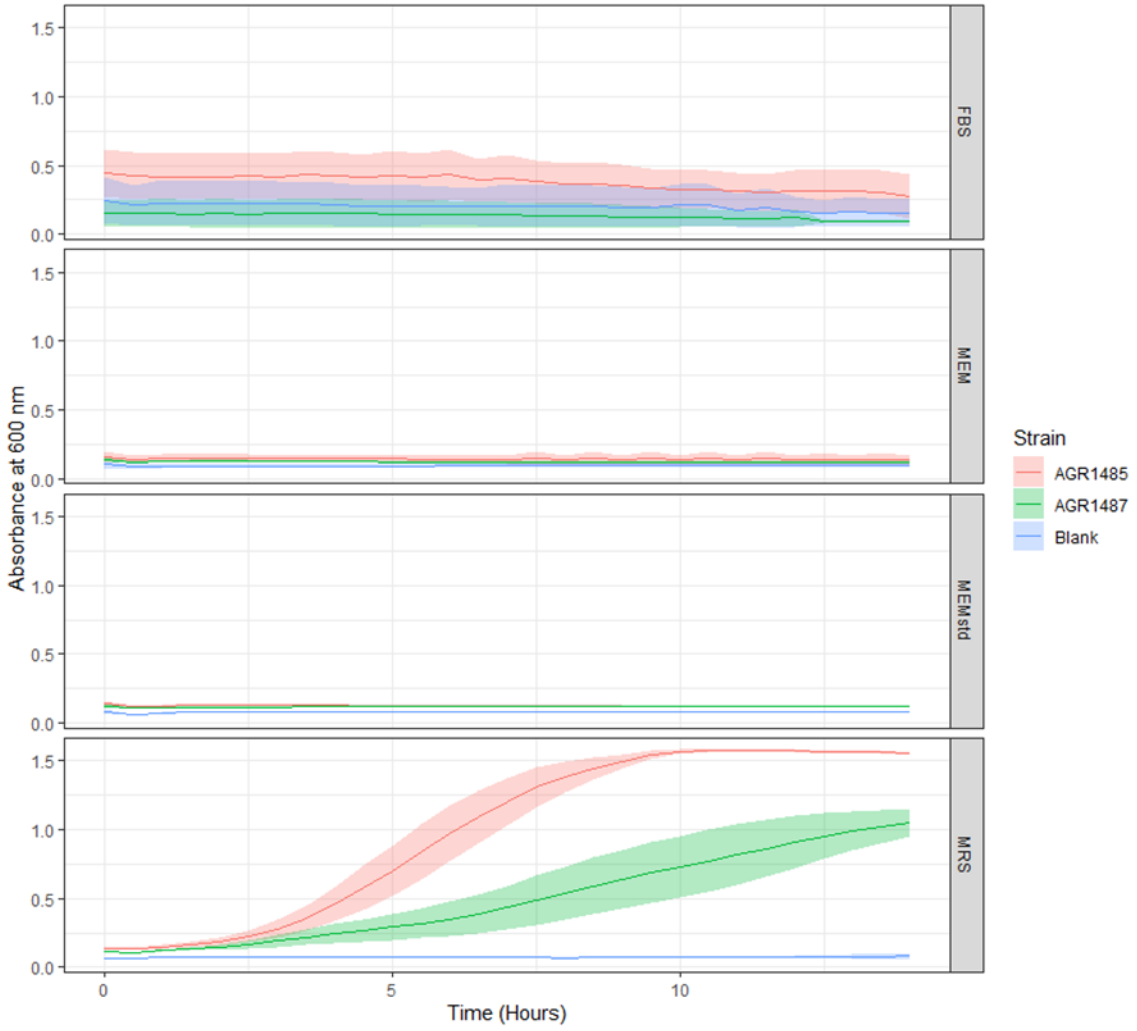

Figure 3.5.2-3: Bacterial growth in MEM media. The growth of *L. fermentum* AGR1485 (Red) and AGR1487 (Green) in FBS, MEM, MEM with FBS (MEMstd) and MRS media, and a control with no bacteria. The samples' optical density (OD600) was recorded in a 96-well titre plate reader over 24 hours. Shading corresponding to the sample colouring shows the sample's 95% confidence interval. N = 16 for each strain in each environmental condition.

3.5.3. Agar plate imaging

Light microscopy was used to compare the colony morphology of AGR1485 and AGR1487 grown on MRS agar, in order to determine differences in surface-associated growth traits such as extracellular polymeric matrix production, texture, and colour. This comparison aimed to further characterise phenotypic differences between the strains on this medium (Section 3.4.11). Both strains displayed round, smooth, uniform, and glossy colonies (Figure 3.5 8). While AGR1487 colonies appeared white and nearly transparent, isolated AGR1485 colonies were distinctly yellow and denser (Figure 3.5.3-1). At a 6.3x magnification, the AGR1485 colonies' boundaries were visible (Figure 3.5.3-2). However, the same magnification and depth of field did not delineate the edges of AGR1487 colonies (Figure 3.5.3-2). On MRS at 37 ºC, AGR1485 formed colonies with an average radius of approximately 0.5 mm. In comparison, AGR1487 produced consistently smaller colonies, with a typical radius of 0.25 mm (Figure 3.5.3-2).

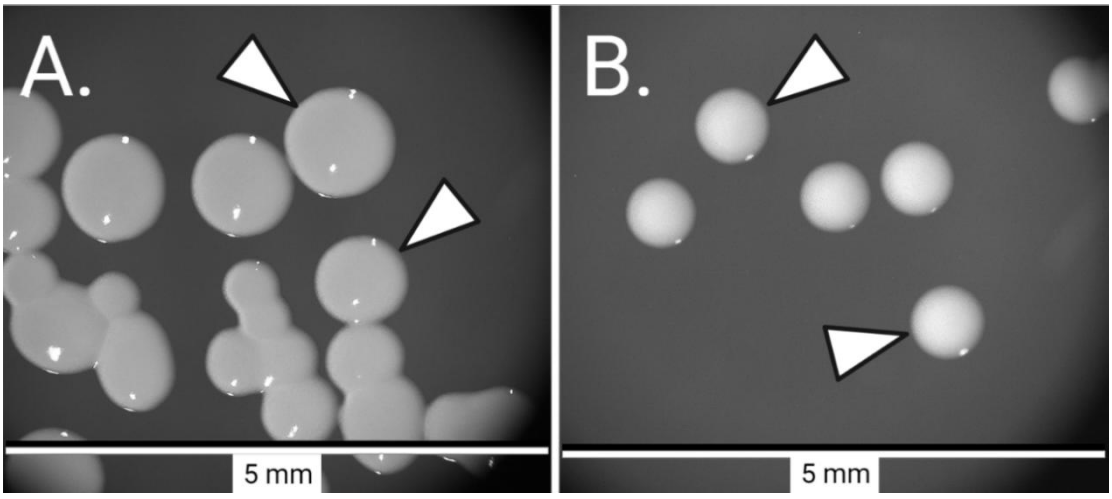

Figure 3.5.3-1: MRS agar plate colony imaging 2.5x. Colour and light normalised imaging of *L. fermentum* (A) AGR1485 and (B) AGR1487 colonies (indicated by white arrows) grown overnight at 37°C on MRS agar plates.

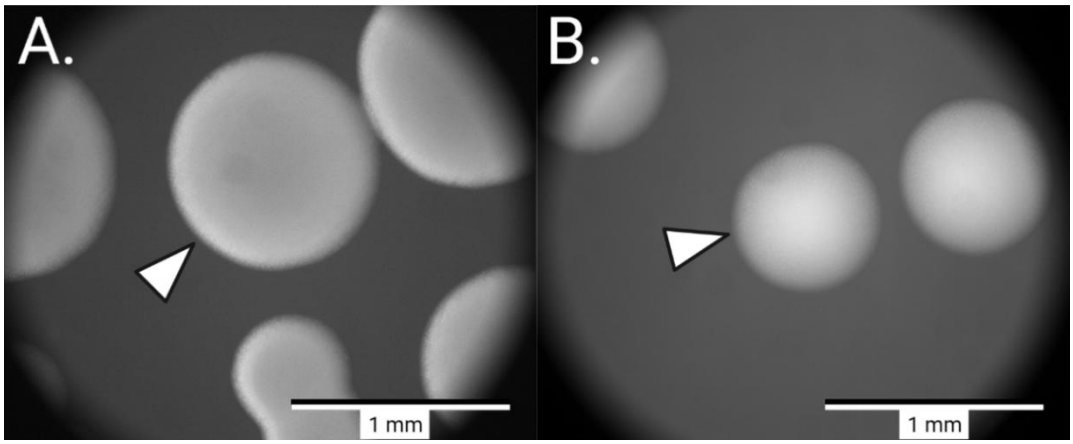

Figure 3.5.3-2: MRS agar plate colony imaging 6.3x. Colour and light normalised imaging of *L. fermentum* (A) AGR1485 and (B) AGR1487 colonies (indicated by white arrows) grown overnight at 37°C on MRS agar plates imaged at 6.3x magnification.

3.5.4. Electron microscopy

Scanning electron microscopy (SEM) and transmission electron microscopy (TEM) were employed to analyse the cell morphology of both *L. fermentum* strains.

The AGR1485 strain's cells were a mix of long and short rods, isolated or paired (Figure 3.5.4-1 A). These rod-shaped cells were typically free and ranged from 1 to 5 µm in length (Figure 3.5.4-2 A). High magnification TEM revealed no discernible external structures on the cell's edges (Figure 3.5.4-4 A).

The AGR1487 strain exhibited a distinctively different cell morphology compared to AGR1485. AGR1487 autoaggregates into tight clusters persisting even after vigorous pipetting (Figure 3.5.4-1 B1). Although vortexing did not disaggregate AGR1487, manual pipetting did break down major aggregates to yield a visually consistent suspension. These clusters appeared cocci-like (Figure 3.5.4-1 B1 and B2). Yet, dislodged mature cells revealed a short rod shape (Figure 3.5.4-1 B, 3). Most cells from this strain were around 1 µm long (Figure 3.5.4-2 B), but some elongated cells

reached 40 µm (Figure 3.5.4-1 B, 2). High magnification TEM showed no visible external structures on their edges, similar to AGR1485 (Figure 3.5.4-4 B).

Both strains' morphology seemed to comprise long and cocci-like cells (Figure 3.5.4-2 and Figure 3.5.4-3). However, AGR1485 did not autoaggregate, and the cells were, on average, larger than AGR1487. Additionally, high-magnification cross-sections of cell walls from both strains unveiled neither internal nor external structures (data not shown).

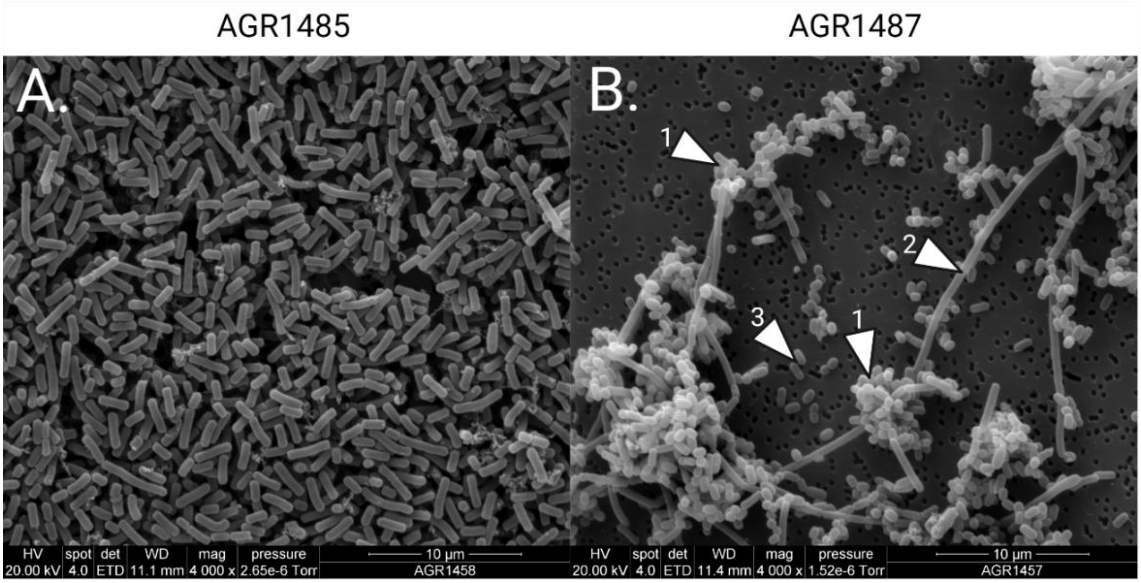

Figure 3.5.4-1: SEM of both strains at 4,000x magnification. Gold-coated scanning electron microscopy of OD_{600} normalised bacterial overnight cultures for *L. fermentum* AGR1485 and *L. fermentum* AGR1487 (A) *L. fermentum* AGR1485; the cells are separated, rod-shaped and near-uniform in length and shape. (B) *L. fermentum* AGR1487, in contrast to AGR1485, features (1) tight clusters. The cells (2 and 3) have diverse shapes and sizes. The aggregated AGR1487 cells (1) are connected via long cells (2). A single, dislodged AGR1487 cell (3).

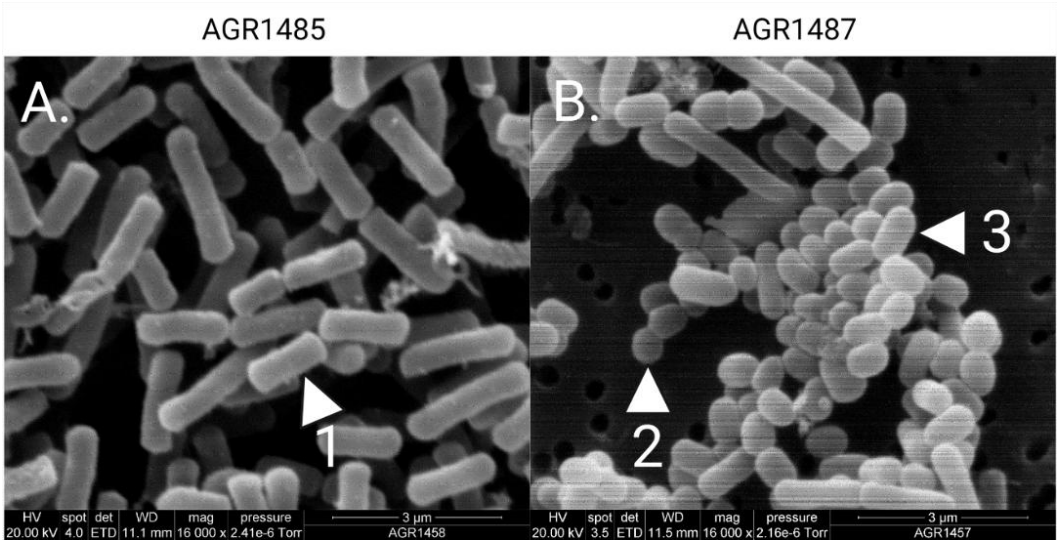

Figure 3.5.4-2: SEM of both strains at 16,000x magnification. Gold-coated scanning electron microscopy of OD_{600}-normalised bacterial cultures for *L. fermentum* AGR1485 and *L. fermentum* AGR1487. (A) *L. fermentum* AGR1485 cells are rod-shaped, uniform, and unconnected. (B) *L. fermentum* AGR1487 cells are tightly connected and near coccoidal in shape within the tight bundles (2 and 3).

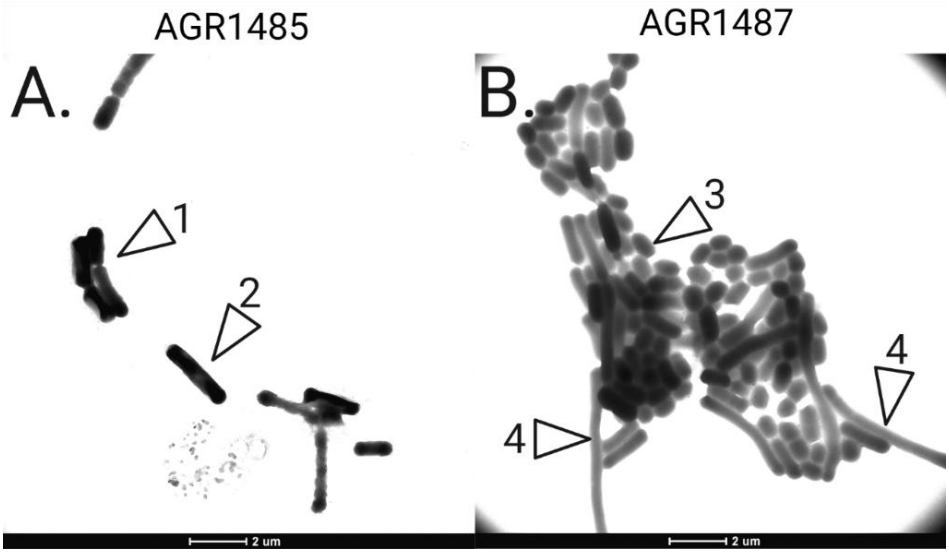

Figure 3.5.4-3: Unstrained TEM of both strains at 8,200x. Unstained transmission electron microscopy of OD_{600} normalised bacterial cultures for (A) *L. fermentum* AGR1485 and (B) *L. fermentum* AGR1487 fixed with formalin. (1) *L. fermentum* AGR1485 cells undergoing binary fission. (2) A mature *L. fermentum* AGR1485 cell. (3) Mature *L. fermentum* AGR1487 cell. (4) Long *L. fermentum* AGR1487 cells linking bundle structures.

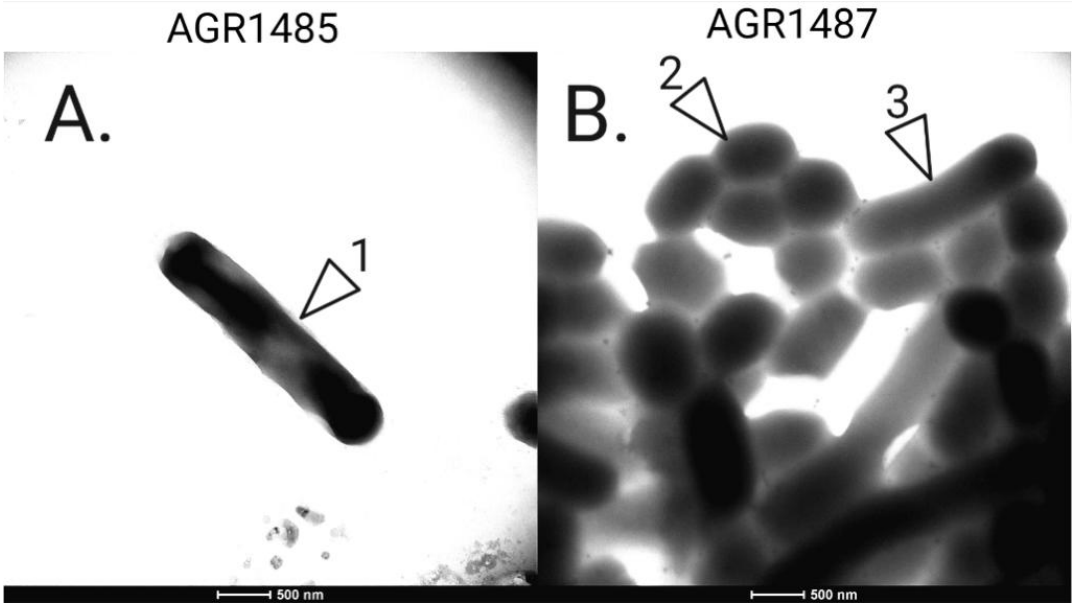

Figure 3.5.4-4: Unstained TEM of both strains at 20,500x. Unstained transmission electron microscopy of OD_{600}-normalised bacterial cultures for (A) *L. fermentum* AGR1485 and (B) *L. fermentum* AGR1487 fixed with formalin. (1) Mature *L. fermentum* AGR1485 cell. The cell is rod-shaped with smooth edges. (2) *L. fermentum* AGR1487 cells appear nearly coccoidal in shape. The cells are tightly connected to their membranes or by an extracellular matrix.

3.6. Discussion

The growth medium composition for Caco-2 cells affects their physiology, monolayer integrity, and TEER assay consistency (Sambuy et al., 2005; Xiu-Wen, Ru-Feng, Ming, Wei, & Xiu-Wei, 2013). As these authors suggest, TEER outcomes obtained with one medium (e.g., M199) may not be directly comparable to those from another (e.g., MEM), particularly when FBS concentrations differ. Therefore, maintaining identical conditions across assays is crucial.

Previous studies investigating the barrier-disruptive TEER phenotype modified the medium, most notably the removal of FBS, which was believed to reduce lactic acid levels produced by the bacteria during the TEER assays (R. C. Anderson et al., 2013; Sengupta et al., 2015). However, aside from noting that FBS can stimulate bacterial lactic acid production (resulting in a pH shift detrimental to Caco-2 cells), the rationale for deviating from standard protocols remains underexplored. Lactic acid is believed to impair cell growth, weaken the monolayer, and reduce barrier integrity, as evidenced by decreased TEER. Consequently, these assays excluded FBS from the apical chamber (Sengupta et al., 2015). Such modifications suggest that results from these altered TEER assays may not align with those obtained under standard conditions, emphasising the need to understand how different culture conditions impact outcomes.

Given the TEER assay modifications, it was essential to conduct controlled experiments to determine how media composition affects TEER measurements. For example, experiments using M199 media without FBS in the apical volume replicated previously published TEER results (Figure 3.5.1-1 A) (Sengupta et al., 2015).

During the first three hours, there was no difference between the treatment and control groups based on their 95% CI. However, after three hours, conditions diverged, and AGR1487 caused a drop in TEER that reached a steady state or complete loss of resistance by hour seven of the assay that persisted until the assay's end. This barrier disruption was marked by an average %TEER reduction

of −20.11 % TEER/h (95 % CI: −20.58 to −19.64) from three to 24 hours, culminating in a maximum drop of −427.31 %TEER (Figure 3.5.1-1 A). This steady state indicates a complete loss of the monolayer's electrical insulation in the presence of AGR1487, whereas AGR1485 remained indistinguishable from the control (95% CI).

Both AGR1485 and the control exhibited positive slopes (2.90 and 5.24, respectively), indicating that electrical resistance increased over 24 hours. Since the control showed a similar trend, this rise is inherent to the assay rather than specific to AGR1485 (Figure 3.5.1-1).

Adding FBS to the apical volume increased variability within all treatment groups (AGR1485, AGR1487, and the control) (Figure 3.5.1-1 B). With FBS present, AGR1485's variation overlapped with that of the other groups, obscuring treatment differences based on the 95% CI. This discrepancy was most evident for AGR1487, which consistently reduced TEER as seen in FBS-free experiments, while the control showed a pronounced TEER increase absent in the AGR1485 treatment group. These results raise concerns about the assay's reliability when FBS is included in the apical volume, especially since no changes in the pH indicator were observed. This finding suggests that another factor may be influencing the outcomes with this specific configuration of the TEER assay and warrants further investigation (Figure 3.5.1-1 B).

Switching from M199 to MEM was unsuccessful, as all conditions—including the control—lost TEER, indicating that MEM without FBS cannot support the Caco-2 monolayer (Appendix: Figure 3.8.1-1 A). Adding FBS increased variability, rendering treatment performances indistinguishable in MEM-based TEER assays (Appendix: Figure 3.8.1-1 B). AGR1487 consistently exhibited a barrier-disruptive phenotype regardless of media type or FBS presence (Figure 3.5.1-1 A and B, Figure 3.8.1-1 A and B). However, the addition of FBS or a change in media obscured differences between strains and the control. The most reliable TEER measurements were obtained using M199 without FBS in the apical volume.

Sengupta et al. (2015) reported that FBS alters media pH via lactic acid production, thereby harming Caco-2 monolayers. In contrast, pH studies found no evidence that FBS induced enough lactic acid production to shift the media's pH enough to cause damage to the monolayer. All TEER assays incorporated phenol red (pH 6.8 [yellow] to 8.2 [red]) as a pH indicator, which remained pink (around pH 7) throughout the 24 hours, indicating a stable, healthy pH for Caco-2 cells (Figure 3.5 2). Furthermore, direct pH measurements of MEM, conducted with and without FBS at cell concentrations equivalent to the TEER assays but in larger volumes (15 mL versus 1 mL), revealed no deviations from the physiological range (pH 7 to 7.9) over 20 hours at 37ºC in a 5% CO2 environment (Appendix: 3.8.2).

Additionally, when bacteria were inoculated into M199 and MEM (with or without FBS) or FBS alone at an initial OD_{600} of 0.2 and incubated overnight, neither a colour change in the pH indicator (Figure 3.5.2-1) nor an increase in optical density (Figure 3.5.2-2and Figure 3.5.2-3) was observed. While the initial OD_{600} confirmed bacterial presence, no further growth occurred. In contrast, both strains thrived in MRS, a nutrient-rich microbial broth, suggesting that the cell culture media either did not support bacterial growth or contained undisclosed inhibitors. These differences were further reflected in distinct colony morphologies and growth on solid MRS agar, hinting at variations in nutrient utilisation or extracellular matrix production between the strains.

AGR1487 produced less extracellular matrix on solid MRS agar compared to AGR1485. AGR1485's colonies appeared glossy and creamy with a soft texture upon manipulation, indicating an abundant extracellular matrix (Figure 3.5.3-1 A). In contrast, AGR1487 colonies were opaque, uniform, and nearly rigid, often lifting as a cohesive sheet when picked (Figure 3.5.3-1 B; sheets not shown). Over 24 hours, AGR1485 colonies expanded to an average diameter of 1 mm (Figure 3.5.3-2 A) while AGR1487 colonies reached only one-third of that size (Figure 3.5.3-2 B). Although the impact of these differences on colony size was not directly tested, the observations prompted further microscopic analysis.

SEM and TEM analyses were used to examine the cellular morphology underlying the colony and rigidity differences between AGR1485 and AGR1487. SEM revealed that AGR1485 cells were small, elongated rods, either isolated or paired and loosely dispersed (Figure 3.5.4-1). In contrast, AGR1487 formed tightly clustered bundles (Figure 3.5.4-1, B1) with elongated cells budding into near-cocci forms, likely contributing to colony cohesion (Figure 3.5.4-1 B2). When these clusters were dislodged, the cells appeared more rounded and smaller than AGR1485 (Figure 3.5 7, B3).

Magnified inspection showed that AGR1487 cells were roughly one-third to one-half the size of AGR1485 cells (Figure 3.5.4-2). While AGR1485 typically arranged end-to-end or singly (Figure 3.5.4-2 A1), AGR1487 exhibited both end-to-end and lateral attachments (Figure 3.5.4-2 B3). These morphological differences suggest that the bundle structures unique to AGR1487 cannot be solely attributed to the fixing solution.

TEM confirmed AGR1485's standard morphology (Figure 3.5.4-3, A1 and A2) and further supported AGR1487's autoaggregation (Figure 3.5.4-3, B3 and B4). The AGR1487 bundles comprised a mixture of elongated cells and predominantly short, near-cocci cells. At higher magnifications, both strains exhibited smooth cell surfaces, indicating that AGR1487's autoaggregation was not due to cell wall projections but may result from incomplete cell division or the presence of an extracellular matrix (Figure 3.5.4-4, B2, B3 and B4).

The genetic basis for the morphological differences between AGR1485 and AGR1487 remains unknown. AGR1487's traits may reduce the surface area available to lytic enzymes, conferring resistance to host-derived lytic activity. This resistance could decrease the release of Microbe-Associated Molecular Patterns (MAMPs), modulating TLR responses and potentially leading to heightened intestinal inflammation. Although speculative, this hypothesis warrants further investigation. Evidence indicates that constant MAMP production is essential for maintaining intestinal homeostasis (Rakoff-Nahoum, Paglino, Eslami-Varzaneh, Edberg, & Medzhitov, 2004);

however, this balance may require a diverse gut microbiome, which was absent in the germ-free rats used with AGR1487.

Interestingly, AGR1487's behaviour in rodent intestines appears to differ from its performance in TEER assays. In TEER assays, AGR1487 consistently exhibits a barrier-disruptive phenotype regardless of bacterial growth (this study) or cell viability (e.g., UV-sterilised or lysed cells) (Sengupta et al., 2015). suggesting that an antagonist is already present on the bacterial surface. *In vivo*, however, AGR1487 produced 1,000-fold more colony-forming units in rat faeces and was more prevalent throughout the GIT than AGR1485 despite similar inoculation levels (Rachel C. Anderson et al., 2016). Whether non-viable AGR1487 cells can trigger the inflammation observed in the colon of germ-free rats remains untested.

Nevertheless, AGR1487 appears more host-adapted than AGR1485 (Sengupta et al., 2015) yet induced colon inflammation in germ-free rats, indicating direct interaction with host tissues (Rachel C. Anderson et al., 2016). The presumed mechanism involves TLR2-mediated activation of p38 MAPK signalling, which disrupts tight junctions, a notion supported by observed gene expression changes (Rachel C. Anderson et al., 2016). Given that TLR receptors require MAMPs for activation, this hypothesis is plausible, especially considering the sterile mucus lining in the animal intestine. However, these findings contrast with TEER assays, where AGR1487's barrier-disruptive response depended on direct contact with Caco-2 cells (R. C. Anderson et al., 2013), with only cell debris, not the supernatant, eliciting the effect (Sengupta et al., 2015). This discrepancy suggests that the MAMPs required for TLR activation may be large enough to pellet during differential centrifugation. Therefore, the colon inflammation observed in germ-free rats and the TEER barrier-disruptive phenotype may result from distinct underlying mechanisms.

3.7. Conclusion

This study demonstrates that AGR1487 consistently exhibits a barrier-disruptive phenotype in TEER assays, unlike AGR1485. This effect persists in both MEM and M199 media and is independent of FBS in the apical volume. Importantly, neither strain's impact on Caco-2 monolayer integrity correlates with pH changes in the culture medium, alleviating concerns that FBS-induced acidity affects cell physiology.

While FBS appears to influence Caco-2 responses to AGR1485, the resolution of FBS-containing assays was insufficient to assess its impact on AGR1487 interactions. This limitation calls for further investigation using more precise methods. Additionally, AGR1485 and AGR1487 differ in cell morphology and tolerances to pH, bile, and salt, suggesting complex environmental adaptations. The relationship between AGR1487's morphological traits and its potential to modulate host immune responses, particularly via TLR2-mediated inflammation, requires further study. Unravelling the genetic and molecular basis of these properties may deepen our understanding of host-microbe dynamics and intestinal inflammation.

The following chapters will explore the genetic factors underlying AGR1487's barrier-disruptive phenotype in TEER assays, focusing on results obtained using M199 without foetal bovine serum in the apical volume under aerobic conditions, which produced the most pronounced effects. These findings will help elucidate the molecular mechanisms governing AGR1487's interactions with the gastrointestinal tract model.

3.8. Appendix

3.8.1. TEER assays with MEM

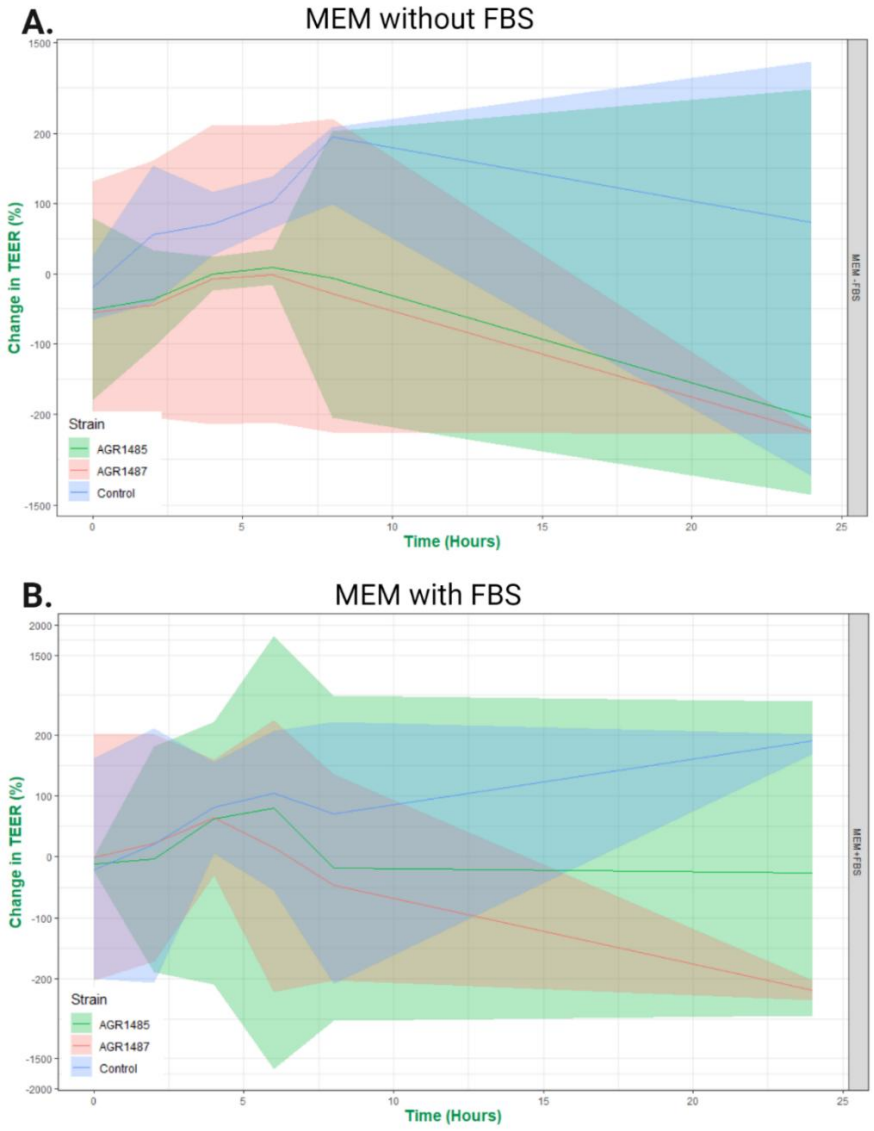

Figure 3.8.1-1: Caco-2-AGR1485/87 MEM TEER assays. The change in electrical resistance (% change) of Caco-2 cells grown in (A) M199 with or (B) without FBS along with either AGR1487 (Red), AGR1485 (Green) or without any bacteria (Control in Blue) in the apical volume over 24 hours. The shading represents the 95% confidence interval (CI) of the data shaded over the trend line. For each condition, n = 3

3.8.2. pH measurements

During the TEER assays, the addition of FBS in the presence of AGR1485 resulted in a loss of electrical resistance during TEER assays (Figure 3.8.1-1). However, no change in the colour of the indicator dye Phenol Red was observed. Phenol Red changes to yellow below pH 6.2 and shifts to bright fuchsia above pH 8.2. The apical volume of the TEER assays was recreated in 10 mL volumes in 15 mL falcon tubes. AGR1485, AGR1487 or no bacteria (blank) were added to MEM culture media with or without FBS and were tested directly with a pH meter. The sample populations are represented as box plots and grouped by replicate (A) or time (C). This data (B replicates, or D time) was plotted with a 95% CI to compare the samples statistically. None of the samples shifted the media pH below pH 6.2 or over pH 8.2 over the 20-hour time course (Figure 3.8.2-1, Figure 3.8.2-2, Figure 3.8.2-3, Figure 3.8.2-4 A and B). Thus, the indicator dye did not shift colour during the TEER assays because the pH remained around pH 7.

When measuring the pH of MEM media with FBS containing AGR1487, the solution remained around pH 7.3 (Figure 3.8.2-2 C and D). AGR1487 cultured in MEM, without FBS, shifted the pH to around 7.9. However, the pH dropped to 7.7 over 20 hours (Figure 3.8.2-1, C and D). AGR1485 cells with MEM media with FBS increased the pH over the same 20-hour period from pH 7.4 to 7.7 (Figure 3.8.2-3C and D). AGR1485, in the presence of MEM without FBS, had a higher starting pH (pH 7.8) (Figure 3.8.2-4 C and D) than the same test condition with FBS (pH 7.4).

Regardless of the presence of FBS or bacteria, the pH stayed within the red pH (above pH 6.3) range of the indicator dye. The variation between test conditions could also be due to random sampling, which does not appear to result from biological activity.

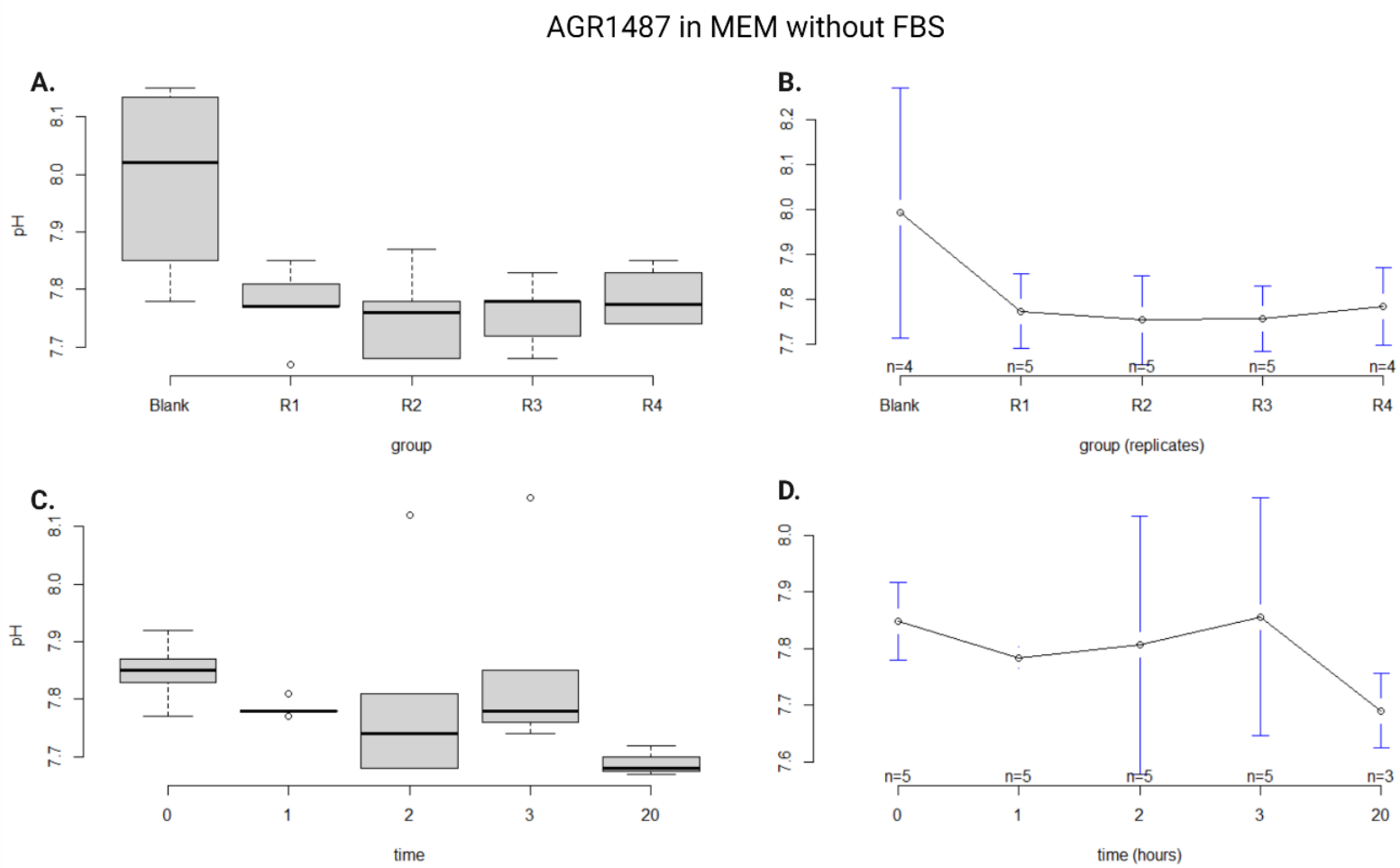

Figure 3.8.2-1: pH measurements of AGR1487 in MEM without FBS. A. Box plots of the data grouped by replicate and averaged across the time points. B. The same data as A represent as a dot plot with a 95% confidence interval. C. Box plot of the replicates grouped by time. D. The average per time point was plotted with a 95% confidence interval. The pH of the samples did not exceed the upper or lower limit of the Phenol Red indicator dye.

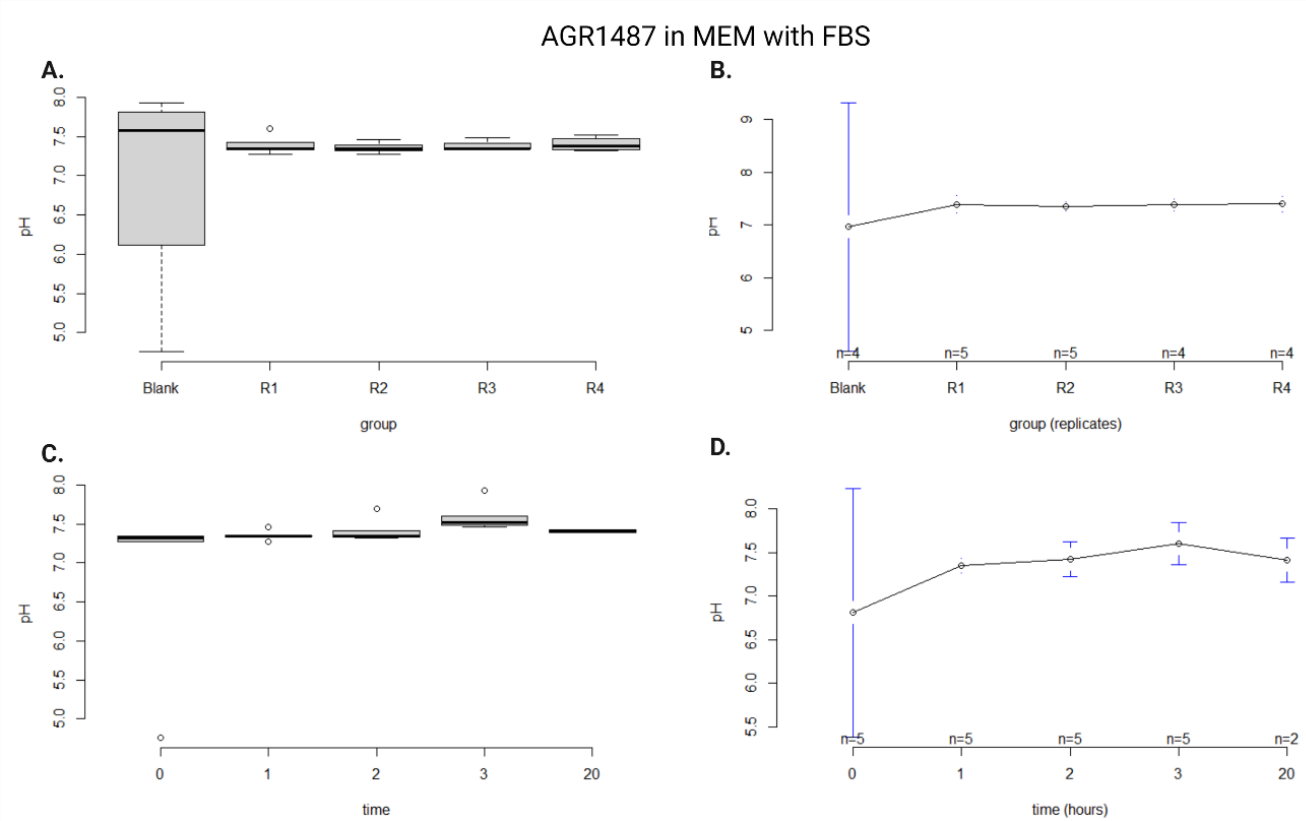

Figure 3.8.2-2: pH measurements of AGR1487 in MEM in the presence of FBS. A. Box plots of the data grouped by replicate and averaged across the time points. B. The same data as A represent as a dot plot with a 95% confidence interval. C. Box plot of the replicates grouped by time. D. The average per time point was plotted with a 95% confidence interval.

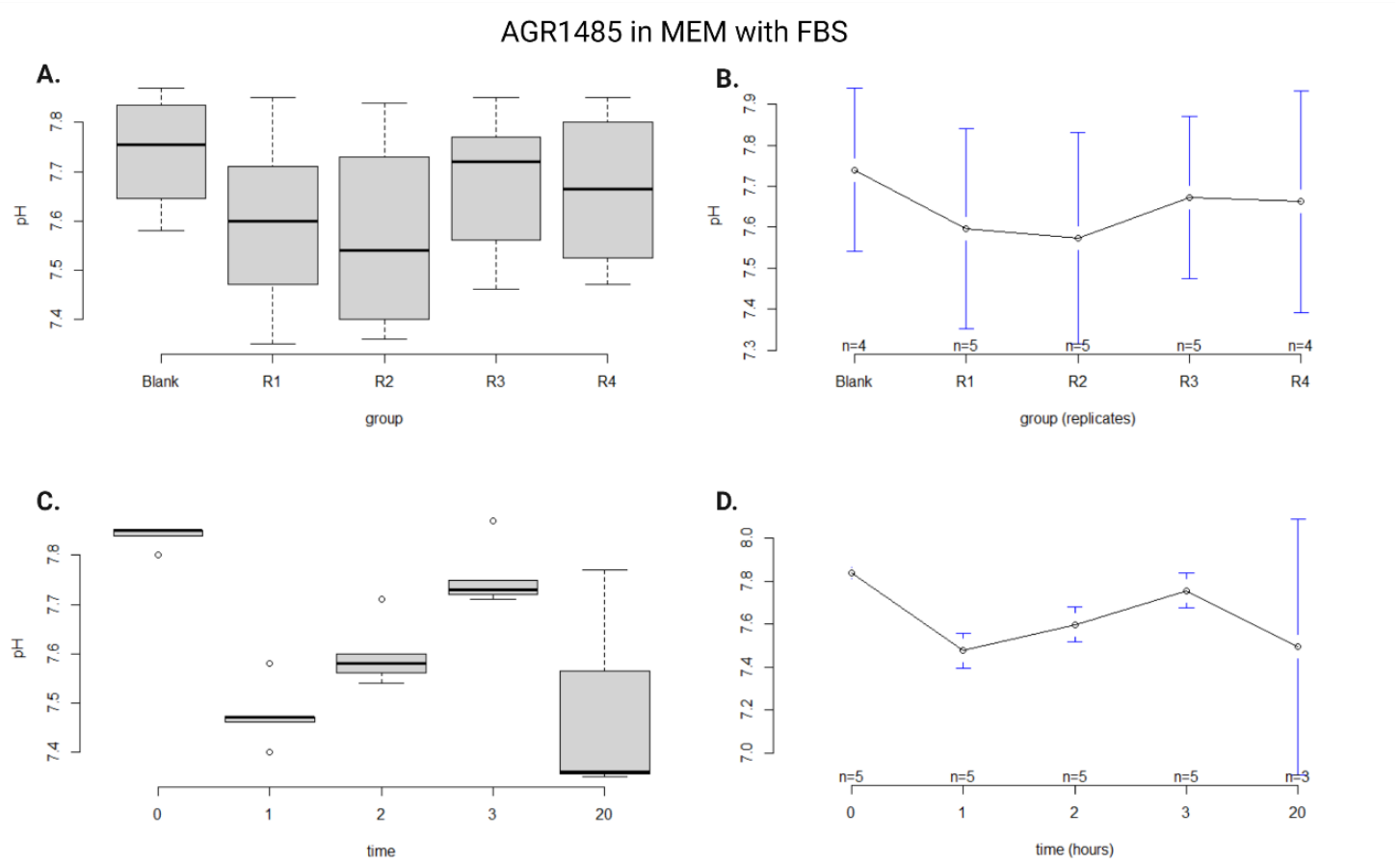

Figure 3.8.2-3: pH measurements of AGR1485 in MEM in the presence of FBS. A. Box plots of the data grouped by replicate and averaged across the time points. B. The same data as A represent as a dot plot with a 95% confidence interval. C. Box plot of the replicates grouped by time. D. The average per time point was plotted with a 95% confidence interval

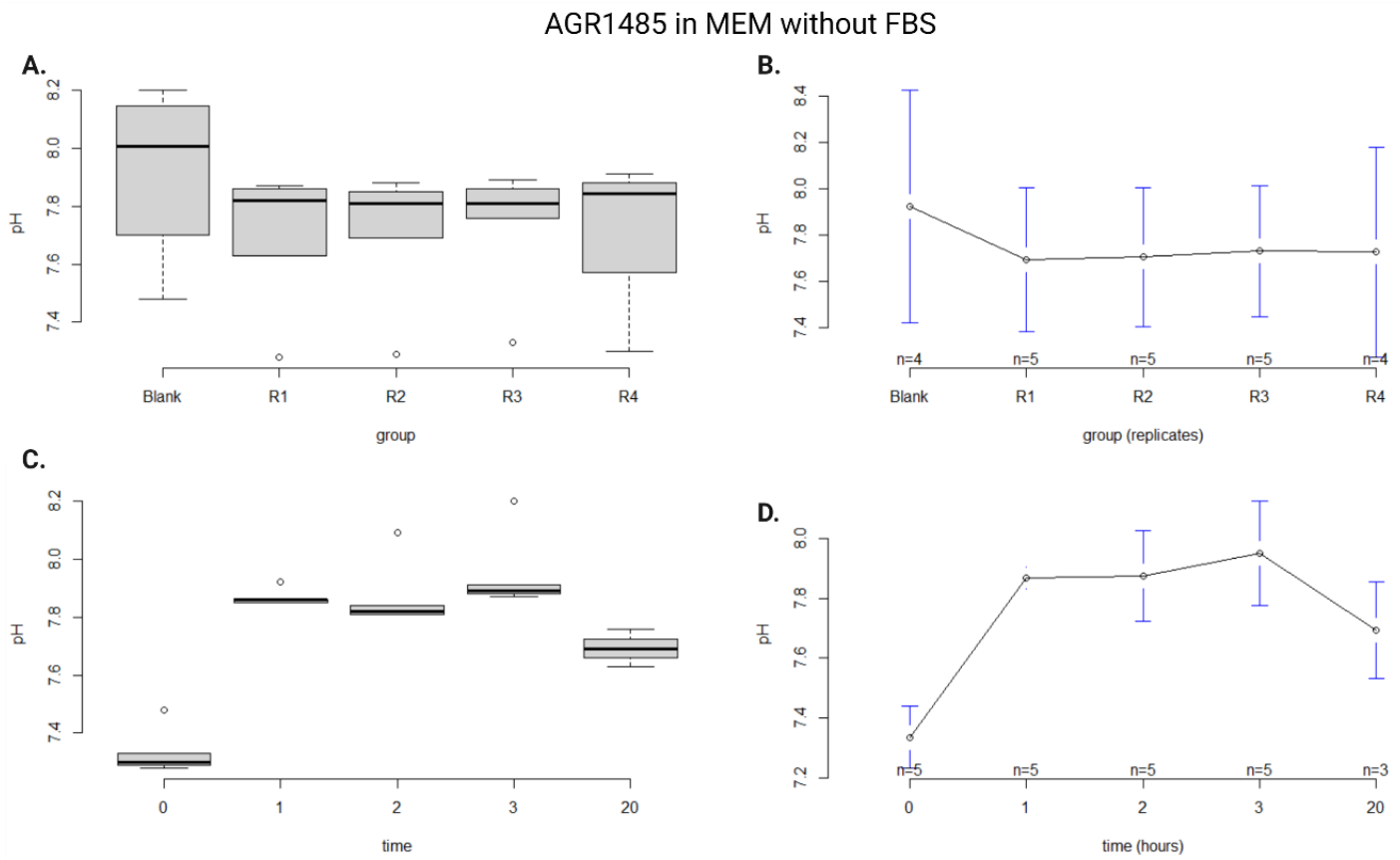

Figure 3.8.2-4: pH measurements of AGR1485 without FBS. A. Box plots of the data grouped by replicate and averaged across the time points. B. The same data as A represent as a dot plot with a 95% confidence interval.. C. Box plot of the replicates grouped by time. D. The average per time point was plotted with a 95% confidence interval. The pH of the samples did not exceed the upper or lower limit of the Phenol Red indicator dye, nor did the pH change significantly over time

3.8.3. Enzyme-assisted chemical lysis

AGR1487 is highly resistant to host-derived lytic enzymes, complicating nucleic acid extractions. The overnight cultures were diluted to an OD_{600} = 0.9 and incubated with 100 units of Mutanolysin (not shown), 8 – 16 mg of Lysozyme (Figure 3.8.3-1 A), or 16 mg of Lysozyme, and 300 units of Mutanolysin (not shown). Regardless of the combination, there was little to no improvement in AGR1487 lysis. However, it profoundly affected *L. fermentum* AGR1485 (Figure 3.8.3-1 B). The cell morphology of AGR1487 (seen above) could be a reason for this resistance to lytic enzymes by reducing the surface area or preventing the enzymes from making contact with the substrate.

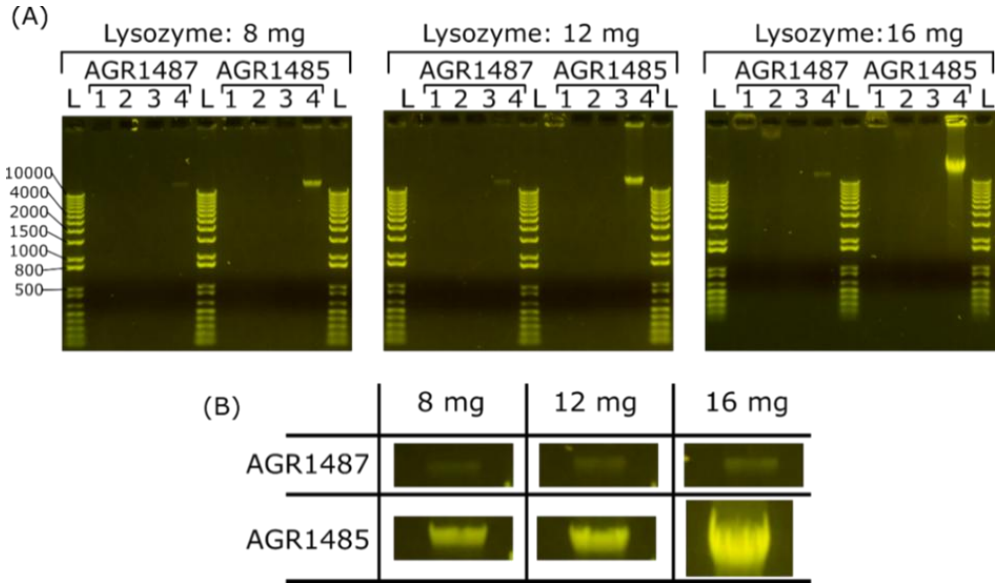

Figure 3.8.3-1: Whole-genome purification. (A) Diagnostic aliquots taken of Lysozyme modified purification steps of 15mL (~OD_{600}: 12) static phase, OD normalised bacterial cultures run on 1% agarose gels after precipitation in isopropanol. (1) Aliquot 1: cell lysis step sampled after B2 was added and incubated; (2) Aliquot 2: Vortexed samples before column loading; (3) Aliquot 3: Runoff after two washes; (4) Aliquot 4: Elution sample; (L) Ladders: 1kb plus. (B) Enlarged results from lane 4 of each condition arranged according to strain and lysozyme concentration.

3.9.References

Alfano, A., Perillo, F., Fusco, A., Savio, V., Corsaro, M. M., Donnarumma, G., . . . Cimini, D. (2020). *Lactobacillus brevis* CD2: Fermentation strategies and extracellular metabolites characterisation. *Probiotics and Antimicrobial Proteins, 12*(4), 1542–1554.

Anderson, R. C., Cookson, A. L., McNabb, W. C., Kelly, W. J., & Roy, N. C. (2010). *Lactobacillus plantarum* DSM 2648 is a potential probiotic that enhances intestinal barrier function. *FEMS Microbiology Letters, 309*(2), 184–192. https://doi.org/10.1111/j.1574-6968.2010.02038.x

Anderson, R. C., Ulluwishewa, D., Young, W., Ryan, L. J., Henderson, G., Meijerink, M., . . . Roy, N. C. (2016). Human oral isolate *Lactobacillus fermentum* AGR1487 induces a pro-inflammatory response in germ-free rat colons. *Scientific Reports, 6*, 20318. https://doi.org/10.1038/srep20318

Anderson, R. C., Young, W., Clerens, S., Cookson, A. L., McCann, M. J., Armstrong, K. M., & Roy, N. C. (2013). Human oral isolate *Lactobacillus fermentum* AGR1487 reduces intestinal barrier integrity by increasing the turnover of microtubules in Caco-2 cells. *PLOS ONE, 8*(11), e78774. https://doi.org/10.1371/journal.pone.0078774

Briske-Anderson, M. J., Finley, J. W., & Newman, S. M. (1997). The influence of culture time and passage number on the morphological and physiological development of Caco-2 cells. *Proceedings of the Society for Experimental Biology and Medicine, 214*(3), 248–257.

Di Cagno, R., Surico, R. F., Siragusa, S., De Angelis, M., Paradiso, A., Minervini, F., . . . Gobbetti, M. (2008). Selection and use of autochthonous mixed starter for lactic acid fermentation of carrots, French beans, or marrows. *International Journal of Food Microbiology, 127*(3), 220–228.

Hurtado, J. A., Maldonado-Lobón, J. A., Díaz-Ropero, M. P., Flores-Rojas, K., Uberos, J., Leante, J. L., . . . Olivares, M. (2017). Oral administration to nursing women of *Lactobacillus fermentum* CECT5716 prevents lactational mastitis development: A randomized controlled trial. *Breastfeeding Medicine, 12*(4), 202–209.

Jo, Y. M., Kim, G. Y., Kim, S. A., Cheon, S. W., Kang, C. H., & Han, N. S. (2021). *Limosilactobacillus fermentum* MG7011: An amylase and phytase producing starter for the preparation of rice-based probiotic beverages. *Frontiers in Microbiology, 12*, 745952. https://doi.org/10.3389/fmicb.2021.745952

Kullisaar, T., Zilmer, K., Salum, T., Rehema, A., & Zilmer, M. (2016). The use of probiotic *L. fermentum* ME-3 containing Reg'Activ Cholesterol supplement for 4 weeks has a positive influence on blood lipoprotein profiles and inflammatory cytokines: An open-label preliminary study. *Nutrition Journal, 15*(1), 1–6.

McClung, L. S. (1987). *Bergey's Manual of Systematic Bacteriology, Volume 2.* Edited by Peter H. A. Sneath, Nicholas S. Mair, & M. Elisabeth Sharpe. The Williams & Wilkins Co., Baltimore, 1986, pp. 962–1599. ISBN 07893-3. $65.50. *International Journal of Systematic Bacteriology, 37*(1), 91–91. https://doi.org/10.1099/00207713-37-1-91

Pan, D. D., Zeng, X. Q., & Yan, Y. T. (2011). Characterisation of *Lactobacillus fermentum* SM-7 isolated from koumiss, a potential probiotic bacterium with cholesterol-lowering effects. *Journal of the Science of Food and Agriculture, 91*(3), 512–518.

Rakoff-Nahoum, S., Paglino, J., Eslami-Varzaneh, F., Edberg, S., & Medzhitov, R. (2004). Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis. *Cell, 118*(2), 229–241.

Sambuy, Y., De Angelis, I., Ranaldi, G., Scarino, M. L., Stammati, A., & Zucco, F. (2005). The Caco-2 cell line as a model of the intestinal barrier: Influence of cell and culture-related factors on Caco-2 cell functional characteristics. *Cell Biology and Toxicology, 21*(1), 1–26. https://doi.org/10.1007/s10565-005-0085-6

Sanchez, I., Palop, L., & Ballesteros, C. (2000). Biochemical characterisation of lactic acid bacteria isolated from spontaneous fermentation of 'Almagro' eggplants. *International Journal of Food Microbiology, 59*(1–2), 9–17.

Sengupta, R., Anderson, R. C., Altermann, E., McNabb, W. C., Ganesh, S., Armstrong, K. M., . . . Roy, N. C. (2015). *Lactobacillus fermentum* AGR1487 cell surface structures and supernatant increase paracellular permeability through different pathways. *MicrobiologyOpen, 4*(4), 541–552. https://doi.org/10.1002/mbo3.260

Srinivasan, B., Kolli, A. R., Esch, M. B., Abaci, H. E., Shuler, M. L., & Hickman, J. J. (2015). TEER measurement techniques for *in vitro* barrier model systems. *Journal of Laboratory Automation, 20*(2), 107–126. https://doi.org/10.1177/2211068214561025

Tomaro-Duchesneau, C., Saha, S., Malhotra, M., Jones, M., Rodes, L., & Prakash, S. (2015). *Lactobacillus fermentum* NCIMB 5221 and NCIMB 2797 as cholesterol-lowering probiotic biotherapeutics: *In vitro* analysis. *Beneficial Microbes, 6*(6), 861–869.

Wu, Y., Ye, Z., Feng, P., Li, R., Chen, X., Tian, X., . . . Li, X. (2021). *Limosilactobacillus fermentum* JL-3 isolated from "Jiangshui" ameliorates hyperuricemia by degrading uric acid. *Gut Microbes, 13*(1), 1897211.

Xiu-Wen, W., Ru-Feng, W., Ming, Y., Wei, X., & Xiu-Wei, Y. (2013). Dulbecco's modified eagle's medium and minimum essential medium—Which one is more preferred for establishment of Caco-2 cell monolayer model used in evaluation of drug absorption? *Die Pharmazie – An International Journal of Pharmaceutical Sciences, 68*(10), 805–810.

Chapter 4: DNA sequencing and genome assembly for AGR1485 and AGR1487

The following chapter contains material from the published manuscripts listed below:

Book Chapter:

Altermann, E., Bailie, M., Fraser, Karl., E., Young, W. (2021). NexGen Sequencing Data: Bioinformatic Tools for Visualisation and Analysis. Elsevier, Comprehensive Foodomics, pages 47 - 90. https://doi.org/10.1016/B978-0-12-816395-5.00001-0

Publications:

Bailie, M., Altermann, E., Young, W., Roy, N., & McNabb, W. (2020). Complete Genome Sequence of *Lactobacillus fermentum* Strain AGR1485, a Human Oral Isolate. Microbiology Resource Announcements, 9(36). doi: 10.1128/mra.00841-20

Bailie, M., Altermann, E., Young, W., Roy, N., & McNabb, W. (2020). Complete Genome Sequence of *Lactobacillus fermentum* Strain AGR1485, a Human Oral Isolate. Microbiology Resource Announcements, 9(36). doi: 10.1128/mra.00841-20

4.1. Abstract

Background: *Limosilactobacillus fermentum* is widely recognised as a probiotic species found in food products and holds Generally Recognised as Safe (GRAS) status with the U.S. Food and Drug Administration. However, strains AGR1487 and AGR1485 exhibit divergent phenotypes. While AGR1487 disrupts barrier integrity in Caco-2 transepithelial electrical resistance (TEER) assays and increases colon inflammation in germ-free mice, AGR1485 shows neither effect.

Aim: This study aimed to sequence and assemble the complete genomes of both strains to elucidate the genetic basis for their differing phenotypes.

Methods: Initial *de novo* assemblies using 454 Pyrosequencing and Illumina short-read data produced incomplete, error-prone draft genomes. These limitations were overcome by a hybrid assembly that combined Illumina short-reads with PacBio long-read sequencing. The resulting assemblies were validated by comparing *in silico* restriction digest fragmentation patterns with published pulse-field gel electrophoresis data.

Results: The hybrid assembly yielded reference-quality genomes of approximately 2.2 Mbp for AGR1485 and 1.9 Mbp for AGR1487. Both genomes exhibit similar GC contents (51.15% for AGR1485 and 52.17% for AGR1487) and lack plasmids. The agreement between the *in silico* and published fragmentation patterns confirms that the assemblies accurately represent the true genomic structures.

Conclusion: The generation of high-quality genome assemblies for both strains provides a robust foundation for identifying the genetic determinants underlying their divergent phenotypes. These findings pave the way for future studies to elucidate the molecular mechanisms responsible for the barrier-disruptive properties observed in AGR1487.

4.2. Introduction

Limosilactobacillus fermentum (*L. fermentum*) is a food-grade, lactic acid-producing microorganism used as an acidifying starter culture in food production (Akabanda, Owusu-Kwarteng, Tano-Debrah, Parkouda, & Jespersen, 2014). However, recent literature indicates that at least one strain of *L. fermentum* may exhibit pathogenic behaviour (Rachel C. Anderson et al., 2016). In particular, *L. fermentum* AGR1487 induces colon inflammation in orally inoculated germ-free mice and reduces epithelial integrity in between human epithelial colorectal adenocarcinoma (Caco-2) cells, as measured by transepithelial electrical resistance (TEER) assays (Rachel C. Anderson et al., 2016; Sengupta et al., 2015). A decline in TEER reflects disrupted tight junctions between epithelial cells. Although the genes, molecules, and mechanisms underlying this phenotype remain unknown, the leaky intestinal phenotype is absent in *L. fermentum* AGR1485 (Sengupta et al., 2015).

Since the advent of omics and recent advances in computing, the top-down method has been the dominant approach to studying biological systems (Bruggeman & Westerhoff, 2007). This strategy examines system-wide characteristics, phenotypes, and behaviours from a "bird's-eye view," capturing non-specific patterns that inform predictive technologies (Bruggeman & Westerhoff, 2007). In this context, the bacterial genome is the starting point for top-down methods such as genomics and transcriptomics, and genome assemblies serve as computational reconstructions of these genomes. A major limitation of this reconstruction is the potential to overlook genetic elements responsible for the barrier-disruptive phenotype. (Bruggeman & Westerhoff, 2007; Tanay, Sharan, Kupiec, & Shamir, 2004).

One of the primary challenges in genome sequencing is that current technologies produce reads of limited length. Due to constraints, such as the time required for polymerase denaturation, depletion of fluorescent indicators, or detector saturation (Amarasinghe et al., 2020), a bacterial genome cannot be sequenced as a single, continuous fragment. Instead, the genome is fragmented, and these shorter

sequences, or reads, are generated. The reads' length, quantity, and error rate depend on the sequencing technology used (Altermann, Bailie, Fraser, & Young, 2021).

Short-read technologies (e.g., Illumina) yield highly accurate, high-coverage data that are excellent for detecting small-scale variations (Altermann et al., 2021). However, their limited read length hampers the ability to resolve repetitive sequences, complex genomic regions, and structural variations, often resulting in fragmented or incomplete assemblies (Altermann et al., 2021). In contrast, long-read technologies (e.g., PacBio and Oxford Nanopore) produce reads that can span these challenging regions, leading to more contiguous and comprehensive genome assemblies (Altermann et al., 2021). Yet, long-reads typically come with higher error rates and lower throughput, which can compromise base-level accuracy (Altermann et al., 2021).

By combining short-read and long-read data, researchers can capitalise on both technologies' strengths while overcoming their individual limitations (Altermann et al., 2021). Short-reads contribute high accuracy and depth, while long-reads provide the contiguity necessary to resolve complex genomic architectures (Altermann et al., 2021). This integrated approach facilitates the construction of reference-quality genome assemblies that are both precise and complete, offering superior results compared to using either technology alone (Altermann et al., 2021).

4.3. Aims and Hypotheses

L. fermentum AGR1487 exhibits a barrier-disruptive phenotype (Sengupta et al., 2015). Because the cell surface is predominantly shaped by gene products encoded within its genome or plasmids and their associated chemical reactions, sequencing and assembling a complete genome is essential to investigate this unique phenotype further.

Chapter 4 outlines two primary objectives. The first was to evaluate and employ various sequencing technologies to capture comprehensive genomic information for AGR1485 and AGR1487. The second was to integrate these data to reconstruct an accurate, reference-quality digital copy of each strain's genome (genome assembly).

Aim 1: Generate complete genome assembly for both strains (AGR1485 and AGR1487) without missing information and validate the two assemblies using published laboratory-based methods.

Hypothesis: Different sequencing technologies have unique strengths and weaknesses; leveraging multiple platforms will yield complete *L. fermentum* genomes without gaps.

4.4. Materials and methods

4.4.1. Bacterial cell culture and broths

L. fermentum strains AGR1485 and AGR1487 were obtained from the AgResearch culture collection (Grasslands, Palmerston North, NZ). These strains were originally isolated by R. C. Anderson et al. (2010). Cells were revived from frozen stocks by streaking onto sterile agar plates infused with de Man-Rogosa-Sharpe (MRS) broth (using bacteriological agar from Sigma-Aldrich, Merck Ltd., Auckland, NZ) and incubated aerobically at 37 °C for 24 hours. Single colonies were then used to inoculate 5 mL of sterile MRS broth (Merck Ltd., Auckland, NZ) in 15 mL Nunc tubes, and the cultures were grown overnight at 37 °C with shaking at 100 RPM.

4.4.2. Cell lysis and DNA purification

The bacterial cells were lysed by grinding under liquid nitrogen and digested using the Qiagen bacterial lysis method (QIAGEN, 2015, June). High-molecular-weight DNA was then purified with Qiagen Genomic-tip/100G columns (Bio-Strategy, Auckland, New Zealand), and plasmids were isolated using the GeneJET Plasmid Miniprep Kit (ThermoFisher, Auckland, New Zealand) according to the manufacturer's instructions.

4.4.3. 16S rRNA gene sequencing

PCR amplification of the 16S rRNA gene was performed using the universal primers 27F (positions 8–27; sequence: AGAGTTTGATCMTGGCTCAG) and 1492R (positions 1492–1513; sequence: TACGGYTACCTTGTTACGACTT) following the protocol of Weisburg, Barns, Pelletier, and Lane (1991). The resulting ~1,400 bp amplicons were submitted to the Massey University Genome Sequencing Service for Sanger sequencing.

4.4.4. Whole-genome sequencing

Whole-genome sequencing was performed using Illumina short-read and PacBio single-molecule real-time (SMRT) sequencing to enable a hybrid assembly. Illumina library preparation and sequencing were conducted by BGI (Beijing Genomic Institute, Shenzhen, China, 2014). Template DNA was sheared into 500-bp fragments using a TruSeq library preparation kit and sequenced on a HiSeq 2000 genome analyser, yielding 2,738,965 and 2,523,872 100-bp paired-end reads for strains AGR1485 and AGR1487, respectively. In parallel, PacBio SMRTbell libraries were prepared from sheared DNA, with magnetic beads used to purify hairpin dimers under defined size selection conditions. The adapters were removed using PacBio's MagBead kit. SMRT sequencing was carried out on a PacBio Sequel platform, generating the reads described in Table 4.4-4. The original PacBio sequencing report is available in the digital appendix for this thesis.

Table 4.4-4: PacBio Subread dataset parameters

	AGR1485	**AGR1487**
NUMBER OF SUBREADS	249,954	344,060
AVERAGE READ LENGTH	8,790 bp	8,498 bp
N_{50}	10,057 bp	9,837 bp

4.4.5. Software and parameters

Custom Python scripts were developed for data processing, analysis, and automation, streamlining the handling of complex datasets. The research environment was built on various versions of the Linux Ubuntu operating system, ensuring compatibility and stability, while essential dependencies were managed via package managers like Conda (2022) (Grüning et al., 2018) to guarantee seamless integration of all software tools. Although not detailed in **Error! Reference s ource not found.** these components were critical for maintaining reproducibility and operational efficiency.

Illumina reads was evaluated for length, quality, and adapter content using FastQC (v0.11.9) (Andrews, 2010), and primers and low-quality read ends were removed with Trimmomatic (v0.39) (Bolger, Lohse, & Usadel, 2014). Assemblers were run with default settings unless specified otherwise, and assembly graphs were inspected using Bandage (v0.8.1) (Ryan R Wick, Schultz, Zobel, & Holt, 2015) before final assembly and polishing. Assemblies were iteratively refined with Pilon (v1.22) (Walker et al., 2014) until no further improvements were observed. Genome completeness and contamination were assessed with CheckM (v.1.0.18) (Parks, Imelfort, Skennerton, Hugenholtz, & Tyson, 2015), and assembly statistics, including GC content and sequencing errors, were calculated using QUAST (v4.6.3) (Gurevich, Saveliev, Vyahhi, & Tesler, 2013).

For manual k-mer optimised short read assemblies, SPAdes (v3.11.1) (A. Bankevich et al., 2012) was run manually (designated SPAdes_3.11.1). The Unicycler pipeline provided automatically optimised SPAdes assemblies (yielding Uni-SPAdes assemblies) (Ryan R. Wick, Judd, Gorrie, & Holt, 2017), and the Ray assembler by S. Boisvert, Laviolette, and Corbeil (2010) was employed as a benchmark control using a de Bruijn graph assembly followed by overlapping layout consensus (OLC) for contig extension. Various long-read assemblers were trialled in parallel (Table 4.4-5) and recombined with short-read assemblies using Metassebler to generate hybrid assemblies for both strains (Wences & Schatz, 2015). Unicycler (v0.4.7) was also run in hybrid mode using the Pacbio and Illumina datasets (Ryan R. Wick et al., 2017). This methodological framework provides valuable insights into the trade-offs and efficiencies of different assembly strategies.

Table 4.4-5: List of software used during genome assembly

SOFTWARE PACKAGE	USE	LICENSE	REFERENCE
NEWBLER 3.0	Assembler	Proprietary	(Margulies et al., 2005)
BANDAGE	Assembly graph visualisation	Free	(Ryan R Wick et al., 2015)
UNICYCLER	Assembler	Free	(Ryan R. Wick et al., 2017)
ABYSS	Assembler	Free	(Simpson et al., 2009)
CELERA	Assembler	Free	(Myers et al., 2000)
C.A.B.O.G.	Assembler	Free	(Miller et al., 2008)
SOAPDENOVO	Assembler	Free	(R. Li et al., 2010)
KIKI	Assembler	Free	(GitHub, 2021)
MEGAHIT	Assembler	Free	(D. Li, Liu, Luo, Sadakane, & Lam, 2015)
VELVET	Assembler	Free	(Zerbino & Birney, 2008)
SPADES	Assembler	Free	(Anton Bankevich et al., 2012)
MASURCA	Assembler	Free	(Zimin et al., 2013)
RAY	Assembler	Free	(Sébastien Boisvert, Raymond, Godzaridis, Laviolette, & Corbeil, 2012)
FASTQC V.1.0.4	Quality checker	Free	(Andrews, 2010)
CHECKM	Quality checker	Free	(Parks et al., 2015)
Q.U.A.S.T.	Quality checker	Free	(Gurevich et al., 2013; Mikheenko, Valin, Prjibelski, Saveliev, & Gurevich, 2016)
R.E.A.P.R.	Quality checker	Free	(Hunt et al., 2013)
GENEIOUS	Bioinformatics Software for Sequence Data Analysis	Proprietary	(Geneious, 2021)
UGENE	Bioinformatics Software for Sequence Data Analysis	Free	(Golosova et al., 2014; Okonechnikov, Fursov, Golosova, & team, 2012; Rose, Tiunov, Sukhomlinov, Golosova, & Prosperi, 2018)

ARTEMIS	Bioinformatics Software	Free	(McQuillan, Parkhill, Berriman, Harris, & Carver, 2011)
ARTEMIS COMPARISON TOOL	Genome visulisation	Free	(Carver et al., 2008)
GAP5	Genome finishing	Free	(Whitwham & Bonfield, 2010)
METASSEMBLER	Hybrid assembler	Free	(Wences & Schatz, 2015)
SEALER	Gap closer and finishing tool	Free	(Paulino et al., 2015)
GAPCLOSER	Gap closer and finishing tool	Free	(Luo et al., 2012)
GAPBLASTER	Gap closer and finishing tool	Free	(de Sá et al., 2016)
GAPFILLER	Gap closer and finishing tool	Free	(Nadalin, Vezzi, & Policriti, 2012)
SAMTOOLS	Data formatting	Free	(H. Li, 2011; H. Li et al., 2009)

4.4.6. Data handling

All sequencing data were converted to FASTQ format for quality control. Roche 454-Pyrosequencing data were screened for adapters and converted to FASTQ using custom Python scripts(Weisburg et al.). In contrast, Illumina data were provided as FASTQ files with standardised headers featuring paired-end reads free of adapters. Both datasets were assessed with FastQC (Andrews, 2010) before and after trimming with Trimmomatic (Bolger et al., 2014). Trimmomatic was configured to search for all known adapter sequences to ensure complete removal. FastQC was used iteratively to determine optimal trimming positions, eliminating low-quality sequence information from the read ends.

4.5. Results

4.5.1. Roche 454-Pyrosequencing assembler benchmarks

Genome assembly benchmarking was conducted using several assemblers on 454 pyrosequencing data. Notably, only the Newbler assembler could process the 454-pyrosequencing data in paired-end mode due to unique header information not understandable by modern assemblers. The other assemblers were limited to single-end mode without the header information. Benchmarking with CheckM assessed completion and contamination for the short-read assemblers. Despite optimisation, Velvet failed to produce any assemblies, whereas Megahit showed improvements with further iterations (Figure 4.5-1). QUAST statistics indicated that most assemblers yielded fragmented assemblies, with Newbler producing the most contiguous results (Figure 4.5-2). However, the Newbler assemblies contained 12,217 Ns for the AGR1485 assembly and 7,289 Ns for the AGR1487 assembly.

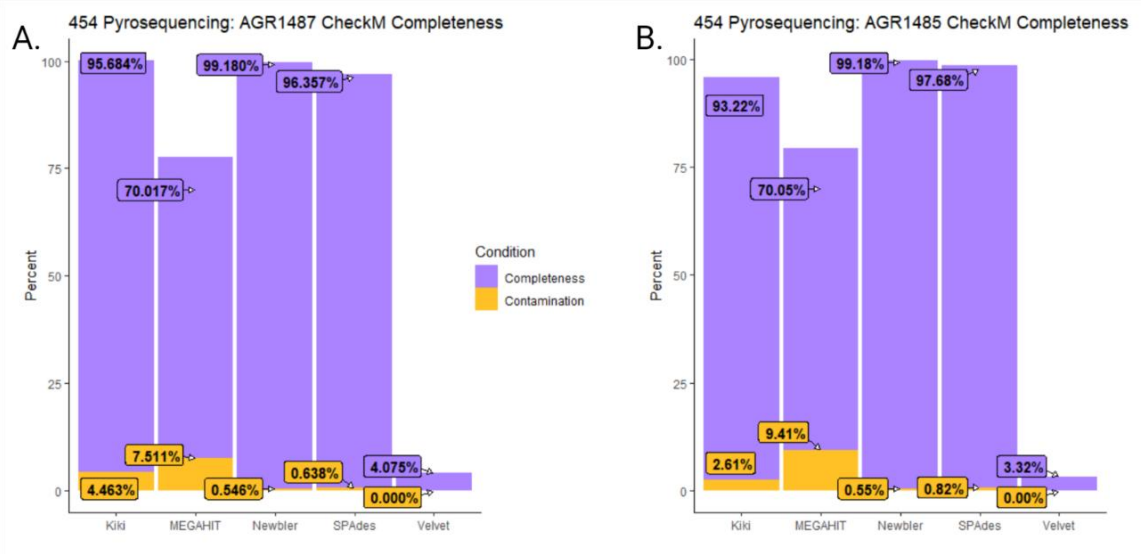

Figure 4.5-1: Roche 454-Pyrosequencing CheckM results for different assembly results. CheckM completeness (Purple) and contamination (Gold bar) scores were estimated using gene markers inferred automatically from a reference tree based on detected gene content. The assembler results were split by strain. A. AGR1487 and B. AGR1485. All assemblers were run using default parameters.

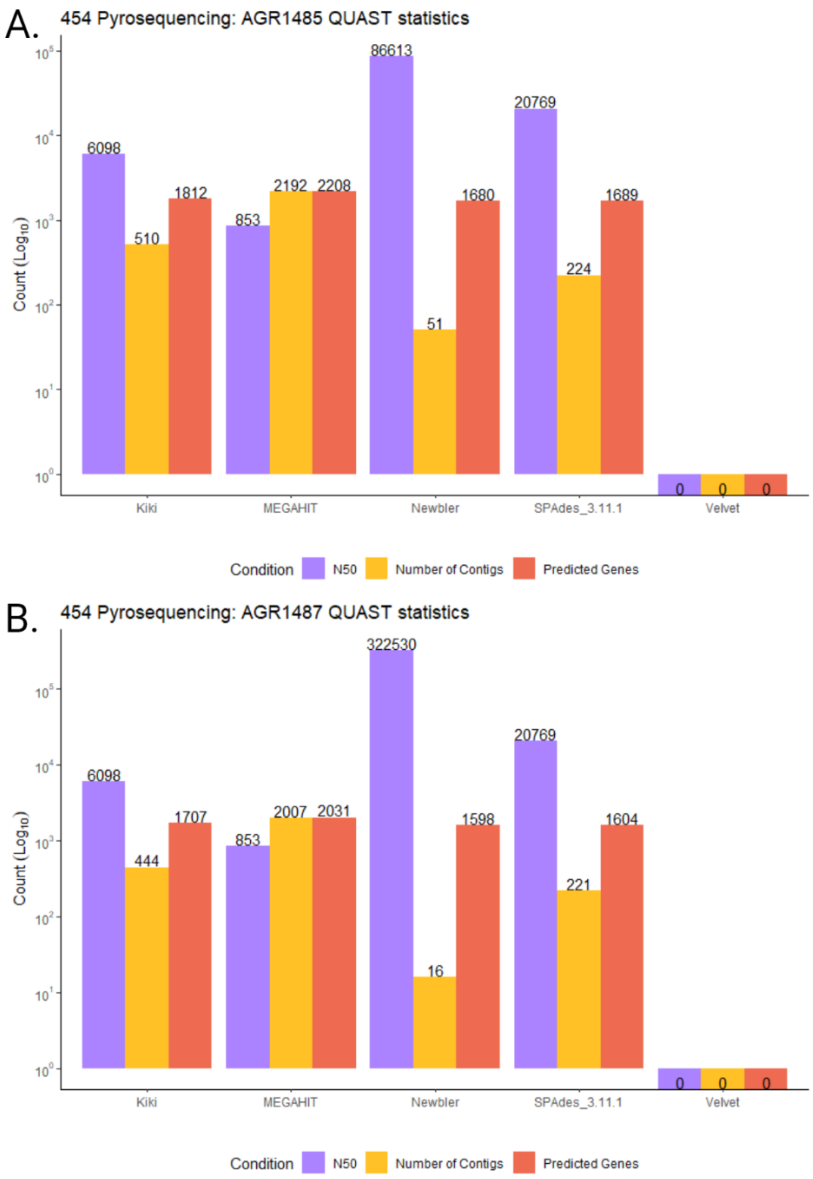

Figure 4.5-2: Roche 454-Pyrosequencing QUAST assembly quality statistics. The assembler performance results were split by strain A. AGR1487 and B. AGR1485. The QUAST statistics were calculated directly from the assemblies. N_{50} is defined as the sequence length of the shortest contig at 50% of the total assembly length. (Purple). The number of contigs is a measure of genome fragmentation (Gold rod). The number of unique genes is a prediction based on open reading frame discovery (Coral). Only Newbler produced detectible errors in the assembly or "Ns" (AGR1485: 12217, AGR1487: 7289). QUAST was run using default parameters.

4.5.2. Illumina assembler benchmarks

The assembler selection for Illumina data was narrowed due to the abundance of excellent tools available. Although Velvet might have produced better assemblies, it was excluded in favour of modern algorithms offering more convenient k-mer optimisation. The chosen assemblers yielded near-complete assemblies with minimal contamination. Notably, Uni-SPAdes performed approximately 1% worse than the manually optimised SPAdes_3.11.1 in terms of completeness but exhibited lower contamination (Figure 4.5-3).

QUAST analysis revealed that none of the assemblies contained ambiguous bases (Ns) or miscalled bases (Figure 4.5-4). As expected, Ray produced the longest and fewest contigs, owing to its use of OLC. In contrast, the other assemblers, which did not employ OLC, generated contigs of similar lengths with comparable *N50* values (Figure 4.5-4). Interestingly, the number of predicted genes varied among these non-OLC assemblers, particularly between the SPAdes assemblies using different k-mer optimisation strategies (Figure 4.5-4).

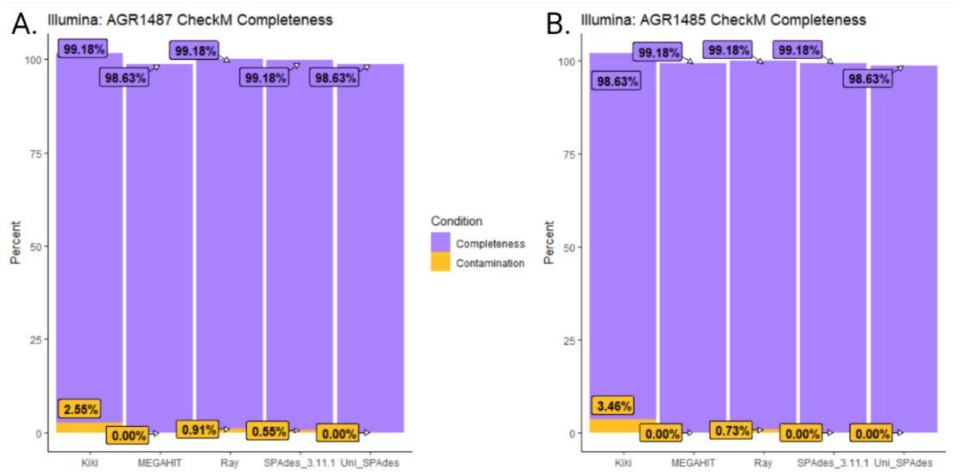

Figure 4.5-3: Illumina CheckM Completeness results. CheckM completeness (Purple) and contamination (Gold bar) scores were estimated using gene markers inferred automatically from a reference tree based on detected gene content. The results were split by strain. A. AGR1487 and B. AGR1485. "Uni_SPAdes" was Unicycler optimised but used the same version (version 3.11.1) as "SPAdes_3.11.1", which was manually K-mer optimised. All assemblers were run using default parameters.

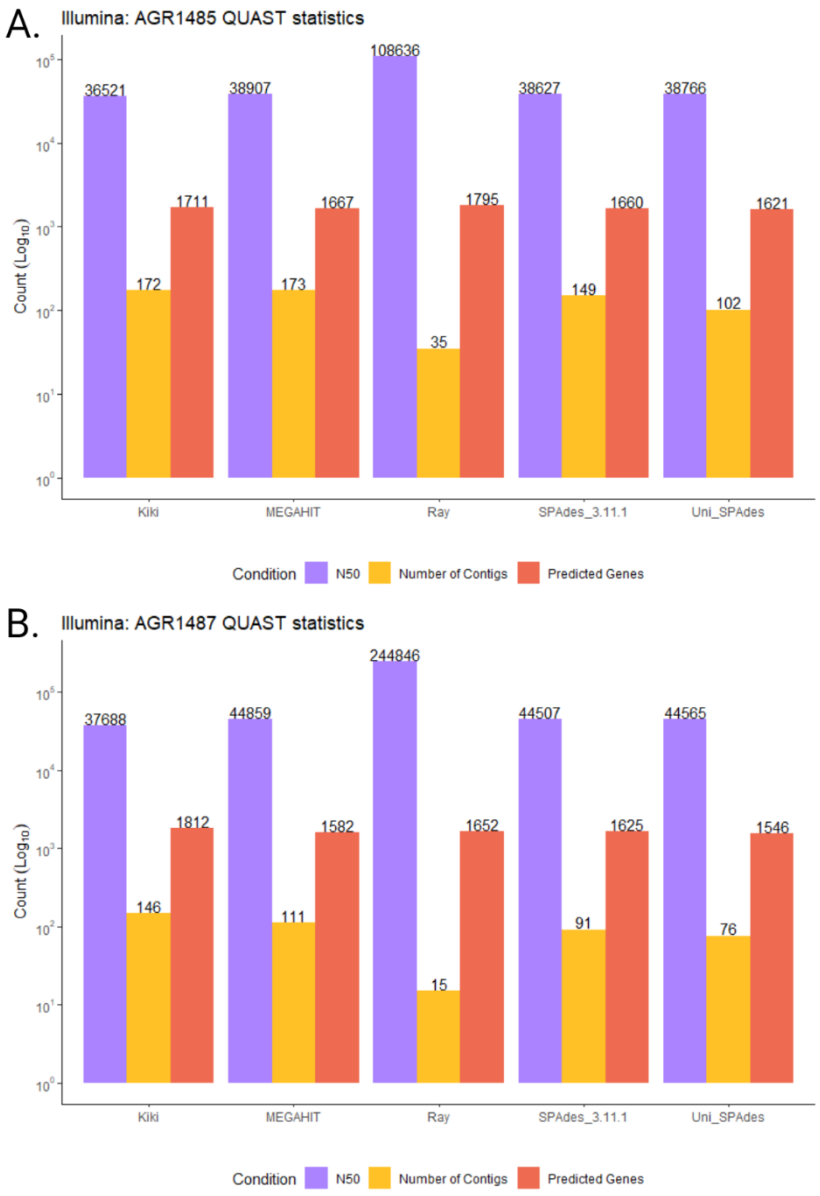

Figure 4.5-4: Illumina QUAST assembly quality statistics. The assembly statics were split by strain, AGR1485 (A.) and AGR1487 (B.). The QUAST statistics were calculated directly from the assemblies. N_{50} is defined as the sequence length of the shortest contig at 50% of the total assembly length. (Purple). The number of contigs is a measure of genome fragmentation (Gold rod). The number of unique genes is a prediction based on open reading frame discovery (Coral). "Uni_SPAdes" was Unicycler optimised but used the same version (version 3.11.1) as "SPAdes_3.11.1", which was manually K-mer optimised. No "Ns" were detected in any of these assemblies QUAST was run using default parameters

4.5.3. 16S rRNA gene sequencing

Strain and data fidelity were confirmed by 16S rRNA gene sequencing. NCBI BLAST's default threshold (E-value of 10) was applied, yielding alignments with a probability well over 99% that the matches were not due to random chance. These results confirmed that the amplicons from both strains belonged to *L. fermentum*. The four amplicon sequences (forward and reverse for both strains) were then aligned with the best Illumina (Ray) and Pyrosequencing (Newbler) assemblies, showing concordance with the 16S rRNA genes in the assemblies and no differences between strains or assembly methods.

Geneious was used to align the amplicons to the contigs, revealing four miscalled bases and two nucleotide polymorphisms (SNPs) (Figure 4.5-6). These discrepancies are unlikely to have resulted from suboptimal PCR conditions (Figure 4.5-5) or sample concentration differences, as submission concentrations were normalised. Instead, they appear to be sequencing errors stemming from low resolution at the beginning and ends of each amplicon that were not fully corrected during trimming.

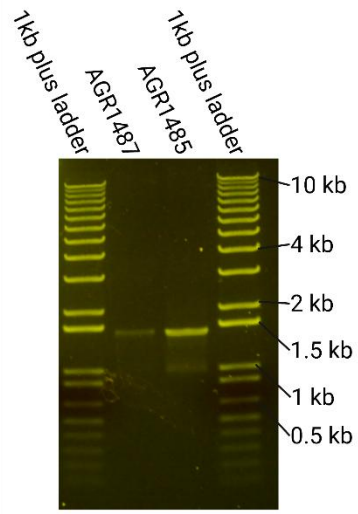

Figure 4.5-5: PCR 16S rRNA gene amplicons. Agarose gel electrophoresis of 16S rRNA gene amplicons from AGR1485 and AGR1487 amplified from the genomic DNA extracted from glycerol stocks and Sanger sequenced.

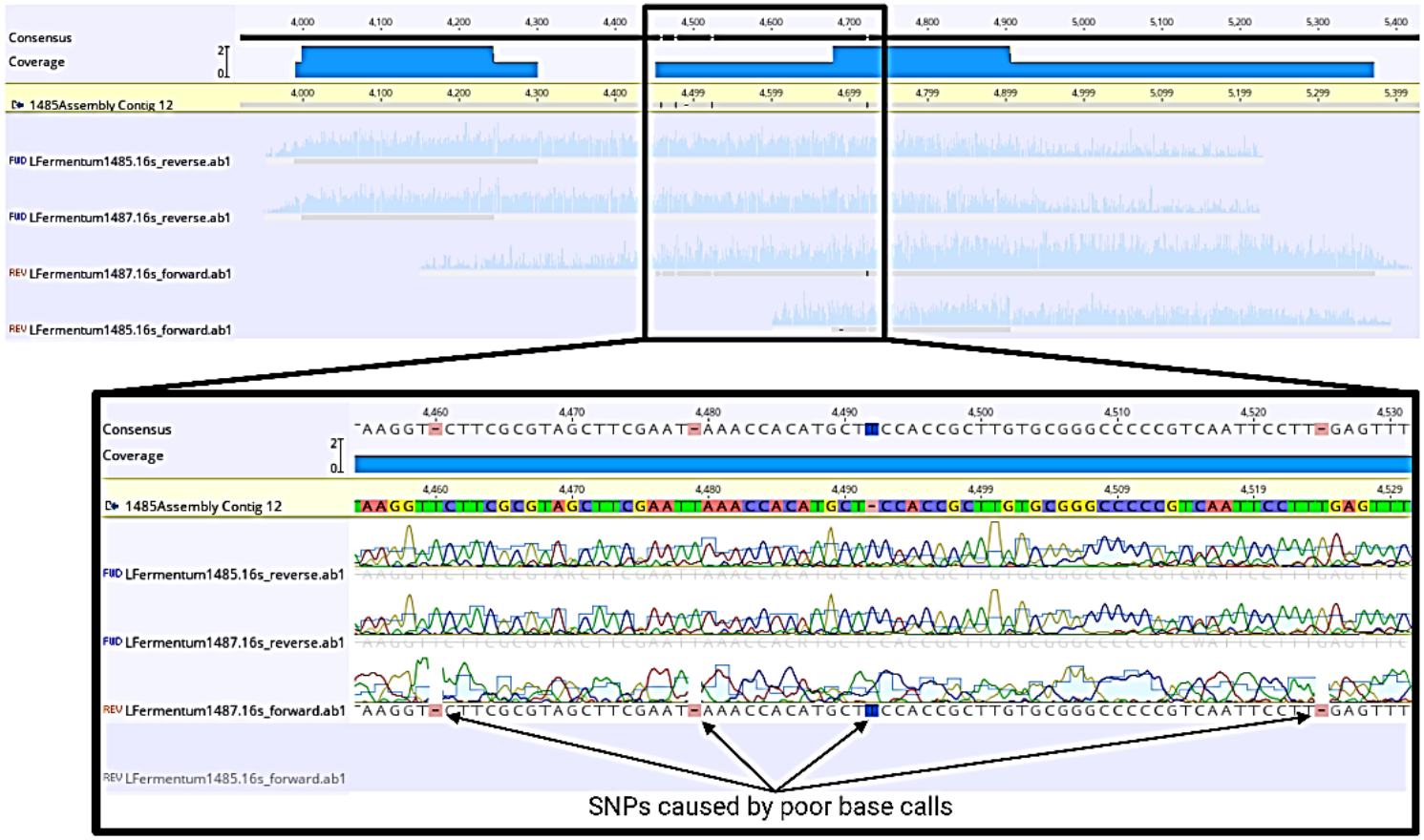

Figure 4.5-6: 16S rRNA amplicon alignments. 16S rRNA gene multiple sequence alignments (MSA) from AGR1485 and AGR1487 aligned against the corresponding gene in contig 12 from the AGR1485 454-Pyrosequencing assembly. The enlarged region shows three miscalled bases and one SNP. A total of six nucleotide differences between the two strain's 16S rRNA genes were found. These differences appeared to be ambiguous base calls caused by low-resolution regions at the distal end of the reads. These errors were not representative of actual changes in the sequence but were a limitation of the sequencing chemistry. MSA for the amplicons was done in Geneious.

4.5.4. Data set similarity

Genome assemblies generated from Illumina and Pyrosequencing data were aligned using Nucmer. Alignments between AGR1485 and AGR1487 revealed significant differences, consistent with their distinct genome sizes and expected structural variations. In contrast, alignments of Pyrosequencing and Illumina datasets for the same strain demonstrated strong concordance.

4.5.5. Novel sequence detection by global alignment

Whole-genome alignments using Nucmer were performed on the Newbler (454-Pyrosequencing) and Ray (Illumina) assemblies to evaluate the efficacy of short-read hybrid assembly, focusing on detecting novel sequences. The Ray assembly was chosen as the reference because it contained fewer errors and was constructed using de Bruijn graph methods combined with contig extension via OLC. The Newbler assemblies for AGR1485 and AGR1487 were globally aligned to their corresponding Ray assemblies (Figure 4.5-7 and Figure 4.5-8).

The Nucmer plots were partitioned into panels using R, each representing a single contig. Alignments were filtered to retain only those with a minimum match length of 10 kbp, and the contigs were then graphed by sequence identity and annotated for comparison (Figure 4.5-9 A and B). This analysis revealed novel genomic regions in each dataset that did not appear to be contaminated but represented sequences absent from the reference (Figure 4.5-9 C and D). Thus, the two datasets captured different genomic regions, suggesting their combined use in a hybrid assembly may facilitate better genome closure or the discovery of new genes.

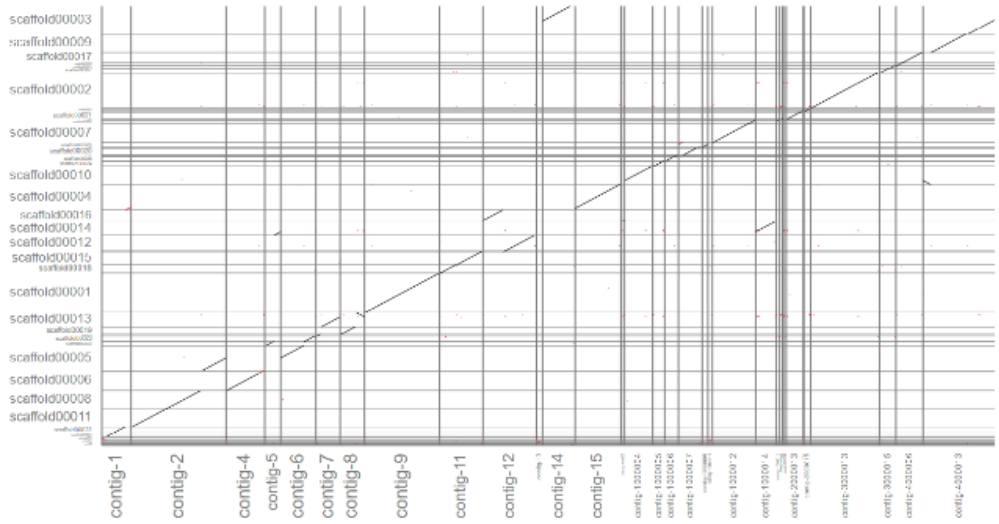

Figure 4.5-7: Nucmer global genome alignment for AGR1485. This figure displays the default MUMmerplot output comparing the draft reference genome (Ray assembly) with the Newbler assembly derived from 454 pyrosequencing data. Note that both assemblies are in draft form, so the contig order is not established. Nucmer was run using the following options: -maxmatch -l 100 -c 500..

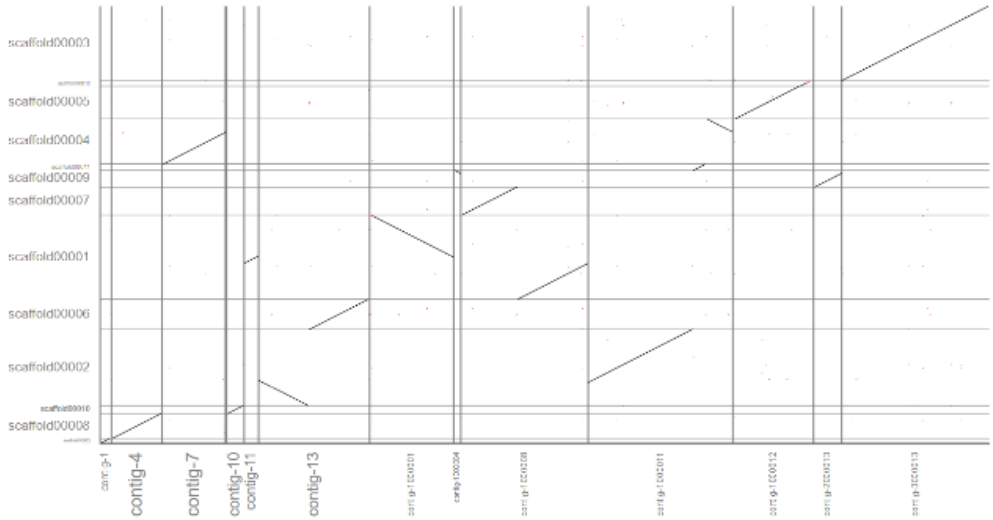

Figure 4.5-8: Nucmer global genome alignment for AGR1487. This figure displays the default MUMmerplot output comparing the draft reference genome (Ray assembly) with the Newbler assembly derived from 454 pyrosequencing data. As both assemblies are in draft form, the contig order is not established. The alignment was generated using Nucmer with the options -maxmatch -l 100 -c 500.

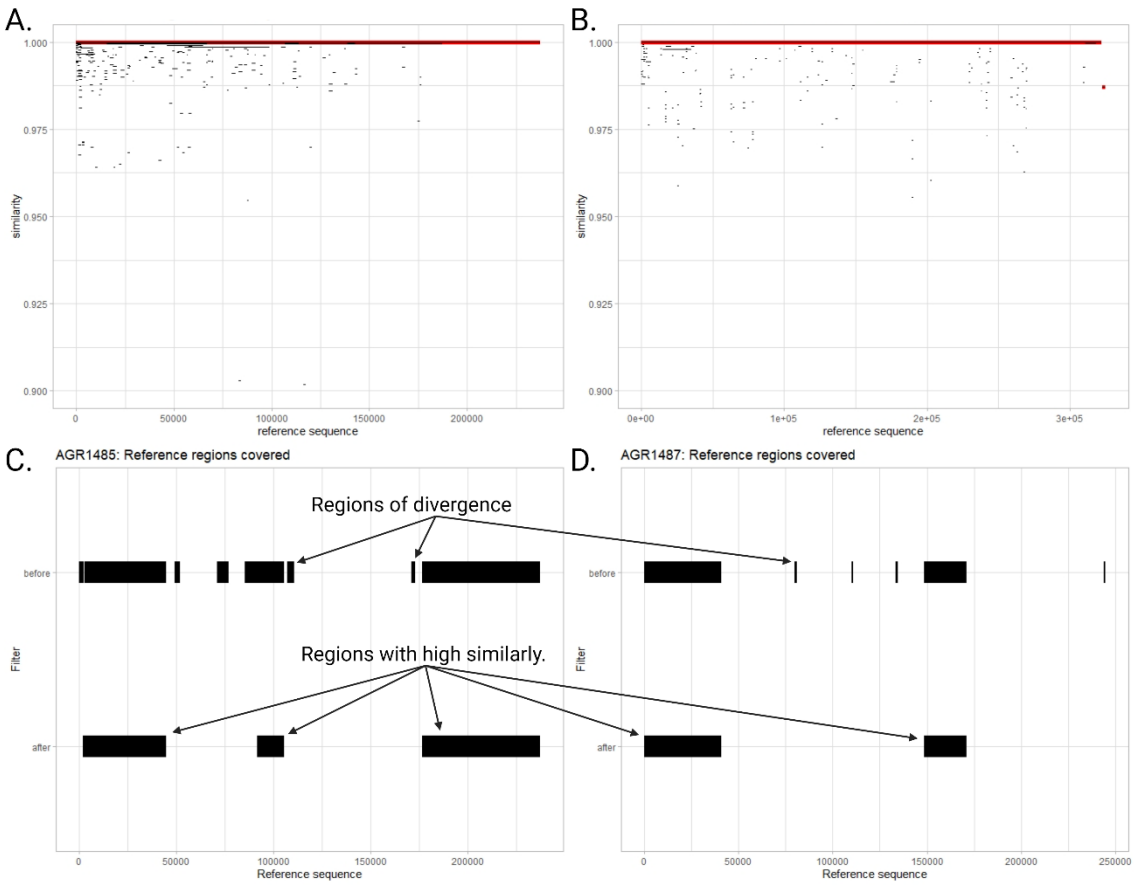

Figure 4.5-9: Filtered Nucmer alignments reveal novel sequence information. (A) AGR1485: Newbler assembly aligned to the Ray reference. (B) AGR1487: Newbler assembly aligned similarly. Maximal unique matches (MUMs) computed with the –maxmatch parameter are shown in red. (C) AGR1485: Alignment filtered to exclude identical regions; remaining imperfect alignments (black bars) are classified as 'high similarity' or 'divergent' if novel sequences are present. (D) AGR1487: Comparison as in (C). "Before" filtering shows all regions with <97% identity, while "After" filtering removes contigs containing stretches (>1000 bp) absent from the reference. The most diverse regions are confined to discrete contigs, indicating novel sequences rather than contamination or plasmids. Percent similarity was calculated using the suffix tree algorithm (Delcher et al., 1999), with updates to the –maxmatch parameter as described by Kurtz et al. (2004).

4.5.6. Short-read hybrid assembly and finishing

Combining two sequencing technologies offers a streamlined path to genome closure by leveraging their complementary strengths. Since each technology captures different genomic regions, one can fill the gaps left by the other. Metassembler employs an OLC approach to connect similar contig regions and extend them iteratively, recombining two assemblies into a single hybrid that unifies common regions while retaining novel sequences. However, this approach has limitations. Scaffolding the assembly is not achievable. The resulting assembly remains draft quality, and errors in the original datasets will persist.

The recombination of Newbler and Illumina assemblies using Metassembler was successful. Attempts to close remaining gaps with GapBlaster or TSG-GapCloser yielded no further improvements, and scaffolding with a downloaded reference would bias the *de novo* assembly by losing novel rearrangements and mis-scaffolding insertions and thus was abandoned. Complete genome closure was not achievable.

For AGR1485, the Metassembler results were less favourable than for AGR1487, as expected, given AGR1485's larger, more complex genome (Figure 4.5-10). The recombined scaffold for AGR1485 contained 23 contigs (down from 35) and 22 fewer open reading frames (ORFs) than the Newbler assembly (Figure 4.5-10 A & B). ORF duplications persisted. Ambiguous base calls (Ns) were also carried over from the Newbler assembly. They were not fully corrected, likely due to low read coverage, a limitation unlikely to be resolved with further optimisation. However, additional iterations might improve gene duplication issues and overall structure.

In contrast, the AGR1487 Metassembler assembly reduced the contig count from 16 (Newbler assembly) to 8 (Figure 4.5-10 D), removed the Ns, and eliminated homologous ORF duplications detected by CheckM. These improvements yielded 1,634 predicted ORFs, which were more accurate despite being fewer than in the individual assemblies.

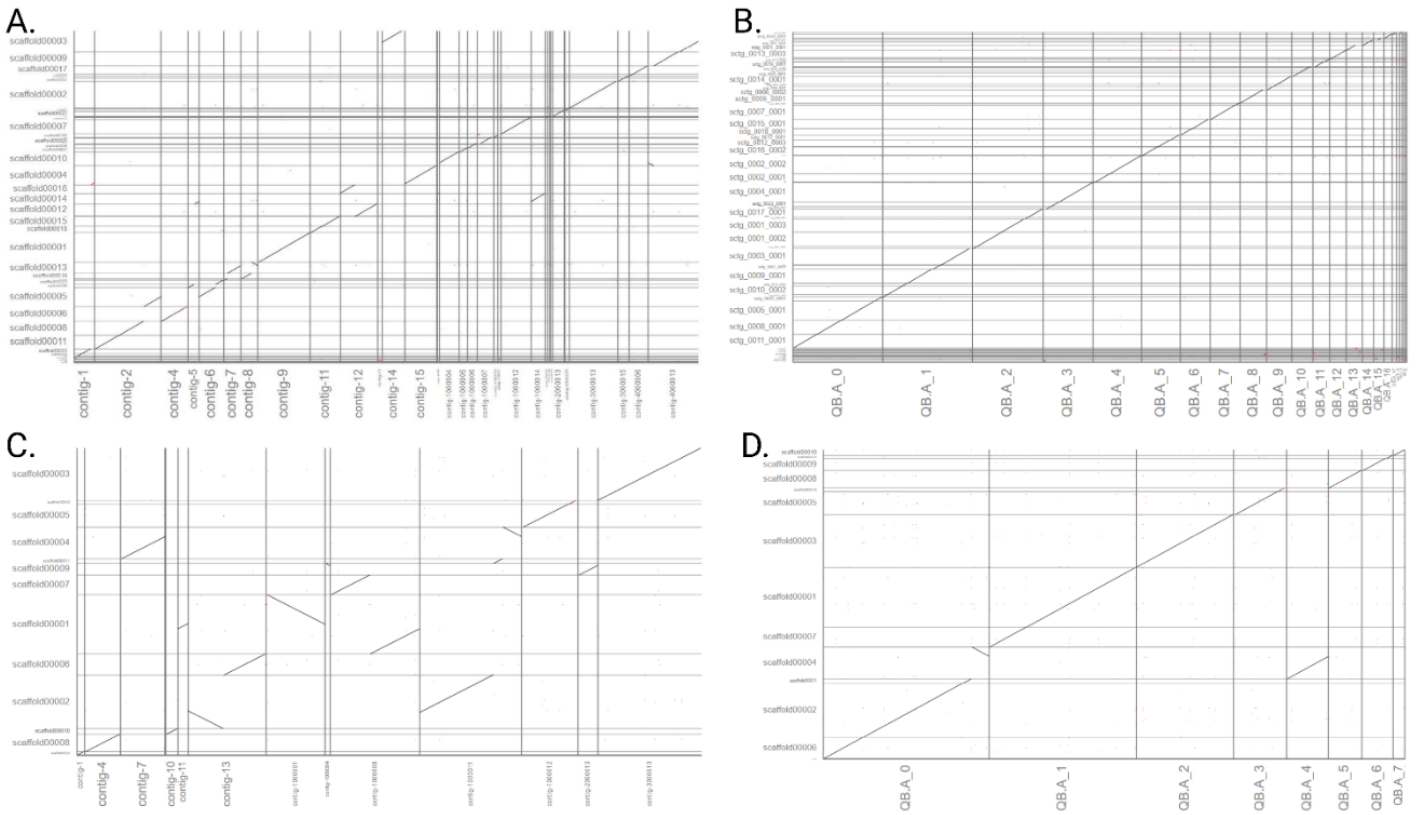

Figure 4.5-10: Short-read hybrid assembly with Metassembler. Nucmer alignments are shown before and after combining datasets, with Newbler on the Y-axis and the Illumina Ray or hybrid assembly on the X-axis. (A) AGR1485: Newbler vs Ray assembly before hybridisation. (B) AGR1485: Metassembler output comparing Newbler with the Metassembler scaffold from the recombination of Ray, AByss, and Megahit Illumina assemblies. (C) AGR1487: Metassembler output comparing Newbler 3.0 with the corresponding Metassembler scaffold. (D) AGR1487: Metassembler output comparing Newbler 3.0 with the corresponding Metassembler scaffold. Note that the reduction in contig count simplifies the scaffold appearance; this improved scaffolding is an artefact of including the reference in the hybrid assembly and does not reflect the true genome structure.

4.5.7. Long-read hybrid assembly

The Unicycler hybrid assembly approach (Figure 4.5-11) independently assembles long- and short-read datasets before recombining them to close gaps, much like the Metassembler approach, rather than using short-reads solely for error correction of long-reads. The resulting assembly graphs, visualised with Bandage (Figure 4.5-12), show that the automatically k-mer optimised de novo SPAdes graphs for AGR1485 and AGR1487 were incomplete (Figure 4.5-12 A). Miniasm assemblies using only long-read PacBio reads successfully were contiguous, and the reads could be successfully mapped back to the assemblies (Figure 4.5-12 B). The Miniasm assemblies contained whole genome duplications and underestimated genome sizes that were not corrected by Racon polishing (Figure 4.5-12 B). In contrast, the Unicycler hybrid assembly corrected internal SNPs and provided accurate genome sizes by integrating short-read data that prevented genome duplication (Figure 4.5-12 C). The PacBio long-read data scaffolded the short-read SPAdes assemblies by bridging gaps, leaving complete genome assemblies by sampling repetitive elements that prevented the SPAdes assembly from completing contigs due to frayed ropes in the graph (Figure 4.5-12 C).

The Unicycler approach outperformed single-dataset and short-read hybrid assemblies by producing the fewest contigs (Table 4.5-6), and, despite yielding fewer predicted genes (Table 4.5-7), it eliminated detectable gene duplications and errors. Final assemblies assessed by CheckM showed 99.18% completeness and 0.546% contamination for both genomes. QUAST confirmed that AGR1485 and AGR1487 were each assembled into a single contig spanning 2,226,862 bp and 1,939,032 bp, respectively, with G+C contents of 51.15% and 52.17%, and no missing or ambiguous nucleotides. Average read coverage was 989.2× for AGR1485 and 1,510.51× for AGR1487, the latter benefiting from an extra PacBio lane. The chromosomes were predicted to be Circular by Unicycler and validated by *in silico* digestion (Figure 4.5-13) with UGENE and pulse-field gel electrophoresis (Table 4.5-8). No plasmids were detected during assembly or by commercial plasmid preparations. Overall, the final assemblies are closed, circular, reference-quality genomes.

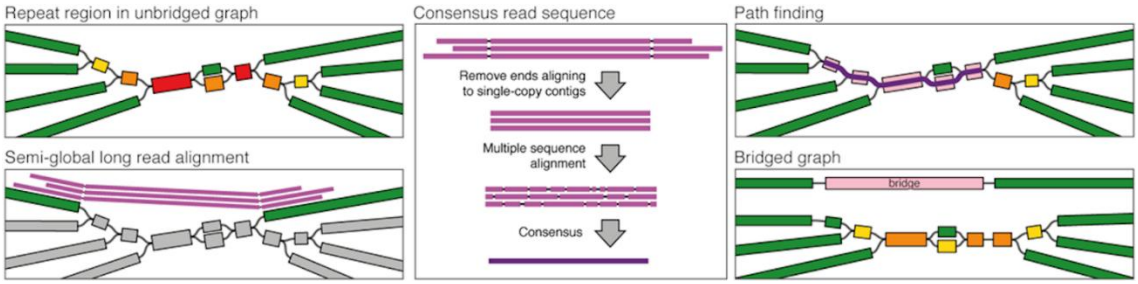

Figure 4.5-11: Unicycler hybrid assembly. Regions with multiple possible paths from a short-read assembly are bridged with long-reads by semi-global long-read alignment. MSA finds a consensus concerning the long-reads and is used for gap bridging and pathfinding. This process simplifies the path that the assembler must follow to join contigs. The consensus read from the long-read data acts like a "bridge" or gap sequence. The contigs are ordered by resolving the gaps, allowing the assembler to join the contigs from the short-read data set into a larger super contig. Modified with permission from https://github.com/rrwick/Unicycler under the Creative Commons Attribution 4.0 International (CC BY) license.

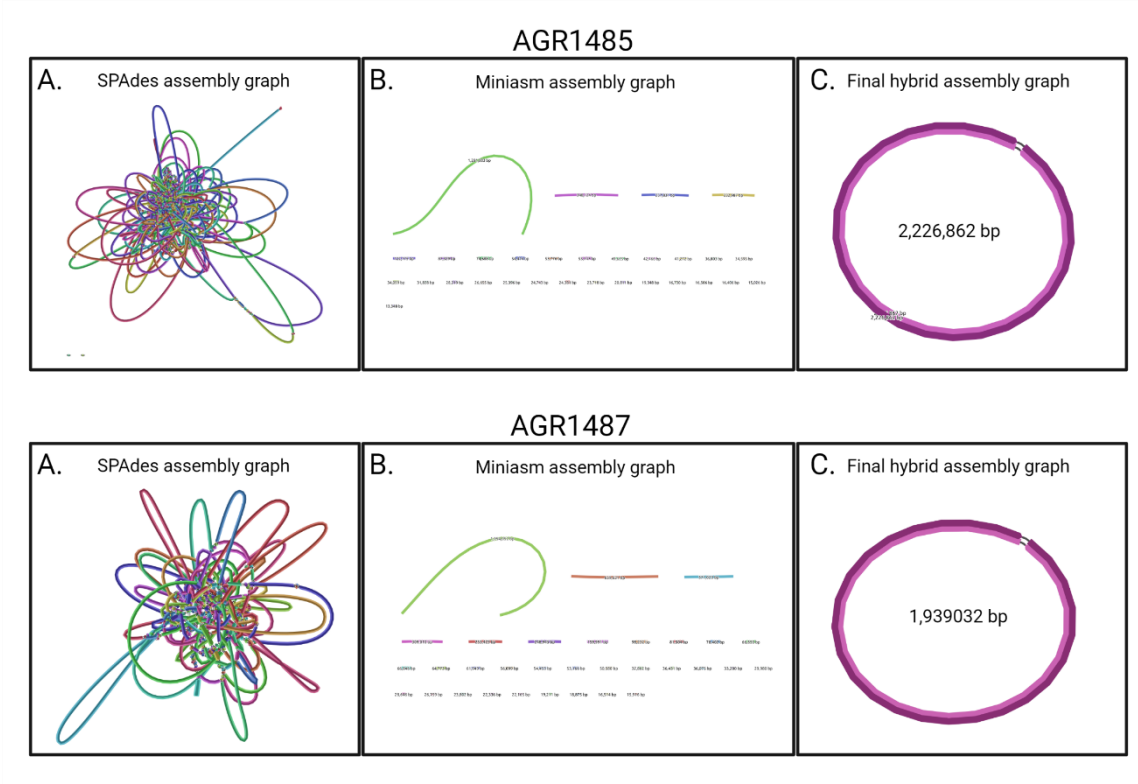

Figure 4.5-12: Hybrid assembly graphs for each assembler. A. The best-optimised SPAdes assembly using Illumina short-reads. B. The miniasm PacBio assemblies use only subreads, including polishing with Racon. C. Final assembly path of the combined data sets. In both cases, the Eulerian walks are resolved into a single path.

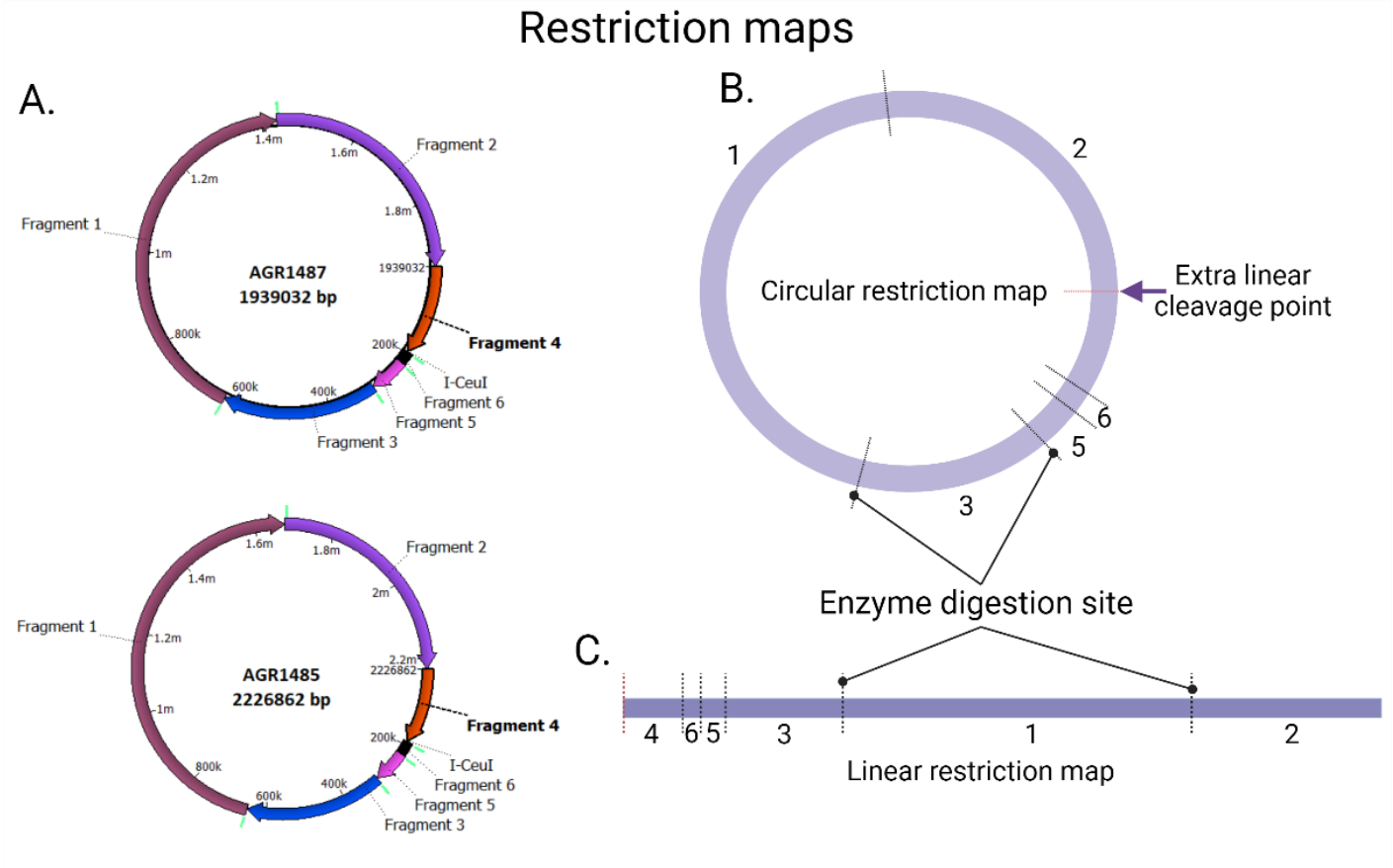

Figure 4.5-13: Restriction maps. A. Circular genome atlases of both genomes were annotated to show the I-CeuI restriction sites and fragments numbered according to size. B. A diagram of the restriction sites and predicted number of fragments for a circular genome topology. C. A diagram of the same genome atlas as in A. but in a linear configuration. The linear structure has the same restrictions sites as B. The marked edges of the fasta sequence are in orange. If the genomes are circular, fragments 2 and 4 are expected to be shown as a single fragment during gel electrophoresis due to the loss of the "extra linear cleavage point".

Table 4.5-6: The number of contigs per assembly method

ASSEMBLER	AGR1485	AGR1487
SPADES	140	110
METASSEMBLER	23	8
UNICYCLER	1	1

Table 4.5-7: The number of unique genes predicted with each method

ASSEMBLER	AGR1485	AGR1487
SPADES	1,654	1,577
METASSEMBLER	1,773	1,634
UNICYCLER	1,742	1,587

Table 4.5-8: Digital restriction digest results for a linear and circular topology

FRAGMENT NUMBER	**AGR1485 LINEAR (BP)**	**AGR1485 CIRCULAR (BP)**	**AGR1487 LINEAR (BP)**	**AGR1485 CIRCULAR (BP)**
1	1,030,370	1,030,370	802,398	802,398
2	551,219	736,861	511,150	701,934
3	551,219	341,542	329,243	329,243
4	185,642	83,273	190,784	77,822
5	83,273	34,796	77,822	27,615
6	34,796	None	27,615	None

4.6. Discussion

A major risk of using incomplete (draft) reference sequences is that target genes may be absent due to missing regions in the genome assembly. In an attempt to mitigate this risk, a short-read assembly strategy was implemented that utilised two short-read datasets per strain, one 454-pyrosequencing dataset and one Illumina dataset.

Several assemblers were benchmarked to determine the best performer. Most de Bruijn graph assemblers failed to assemble the Roche 454-pyrosequencing data due to a file format incompatibility and the non-standardised 2009 headers. Although fastq could be generated from the .sff format using Python scripts, this process removed paired-end information, forcing single-end assembly and increasing contig counts. In contrast, Roche's Newbler retained paired-end information and produced longer contigs through its overlapping layout consensus (OLC) and greedy graphing methods. However, Newbler's raw files were incompatible with standard quality-control tools like Trimmomatic, and its internal read corrections could not be validated. Initially, the assemblies were evaluated using CheckM for completeness and contamination and QUAST for contig statistics.

CheckM, which assesses gene completeness and contamination based on reference tree markers, showed that most short-read assemblers achieved over 90% completeness, except for Velvet and MEGAHIT. Further k-mer optimisation did not improve Velvet's performance, and although MEGAHIT (an updated version of SOAPdenovo designed for metagenomic assembly) provided better scaffolding, its CheckM scores remained unchanged.

QUAST metrics indicated that while Newbler assemblies ranked highest in performance, their quality was poor, often containing spurious rearrangements. The best Newbler assemblies had 12,217 miscalled bases in AGR1485 and 7,289 in AGR1487, likely due to insufficient preprocessing and low coverage. In contrast, the de Bruijn graph assemblers (MEGAHIT, Velvet, Kiki, and SPAdes) operated in single-end mode, hindering contig extension, but produced consistent results, passed read

alignment quality control, and contained no miscalled bases. However, these assemblies were highly fragmented and would require additional sequencing to achieve reference-quality reconstruction.

Illumina data produced in 2014 was more standardised and compatible with modern assemblers. Assembly pipelines were constrained by automated k-mer optimisation and computational demands, affecting efficiency. For comparative purposes, manually and automatically optimised SPAde assemblies were created. This approach improved assembly accuracy by approximately 1% (based on CheckM homolog scanning) but introduced 0.55% more contamination in the AGR1487 assembly compared to the automated system. QUAST results suggested that although manual k-mer optimisation improved unique gene discovery rates, it also increased the apparent contamination rate in CheckM. Parameter adjustments can alter nucleotide sequences and ultimately affect ORF positions and structures. Although QUAST uses GeneMarkS, GeneMark-ES, MetaGeneMark, and GlimmerHMM for gene prediction, the method for determining gene uniqueness remains unclear, particularly for truncated or structurally erroneous genes.

Assembling the Illumina data was more straightforward than assembling the pyrosequencing data. Across both strains, Illumina assemblers achieved approximately 99% completeness with no ambiguous bases. Unicycler produced the best assemblies with additional polishing, though it required longer runtimes (~45 minutes and 40 GB RAM) compared to MEGAHIT, which completed a draft assembly in under a minute using roughly 9 GB RAM. Independently run SPAdes took about 9 minutes per strain, while RAY generated the longest contigs by performing a second round on the initial de Bruijn graphed contigs using OLC. However, the accuracy of these extended contigs was not evaluated.

Comparing the Newbler (454-pyrosequencing data) and RAY (Illumina data) assemblies revealed nearly identical results within each strain but significant structural differences between strains, which aligned with published PFGE data. Furthermore, 16S rRNA sequencing confirmed that

both strains were *L. fermentum*, matching the draft assemblies except for six SNPs, likely arising from low-resolution regions at the ends of the Sanger sequences.

Alignments between the Newbler and RAY assemblies suggested that merging the datasets could reduce contig counts before genome finishing. Since each sequencing chemistry captured different genomic regions, their combination could yield a more complete assembly. However, limitations such as lack of scaffolding, low coverage, and poor representation of repetitive elements, typically addressed by PCR, remained.

Global alignments with Nucmer showed disordered contigs due to the draft nature of the assemblies; however, because the goal was *de novo* assembly, the contig order did not impact downstream analysis. Quality scoring corrected ambiguous nucleotides (Ns) in the Newbler assembly using RAY data, even though order and orientation differed. To determine if merging datasets could add new information, the Nucmer alignments were analysed in more detail.

Nucmer plots processed in R revealed long regions of similarity (~10 kbp) across most contigs, though several regions contained structural differences and SNPs, potentially arising from sequencing errors, inadequate trimming, or difficulties in assembling repetitive regions. Given the high overall alignment similarity, the absence of a secondary chromosome, and the lack of detected plasmids in electrophoresis or commercial preparations, these differences were likely genomic.

To simplify the noisy alignment graphs, only contigs with at least 10 kbp of correctly aligned sequence were retained. Newbler contigs spanned the entire RAY draft genome but were skewed toward the beginning, with an estimated genome size of ~2 Mbp. Perfectly aligned contigs were removed, and regions with less than 97% agreement were marked. Filtering out sections containing novel sequences indicated that these regions likely represented new or divergent DNA within partially matching contigs.

These findings, including the distribution of novel regions, support the potential of hybrid assembly to capture more genomic information, although scaffolding remains a challenge.

Metassembler combined the Illumina and pyrosequencing datasets, reducing contig counts, but digital gap closure was ineffective. Although genome walking could theoretically resolve low-coverage regions, SNPs, and ambiguous bases, it would require extensive PCR reactions and sequencing cycles, taking months or even a year to optimise.

Moreover, genome walking may not accurately capture true genome structure, as repetitive elements lost during short-read assembly cannot be recovered. Given the success of hybrid assembly, the approach was repeated with 10 kb PacBio long-read sequencing to generate closed genomes. These long-reads increase coverage, bridge gaps, and span repetitive regions, thereby preserving genomic structure.

Several long-read hybrid assembly tools were Benchmarked. MaSuRCA and ABySS failed due to resource limitations, while Canu successfully assembled the data after correcting PacBio errors with Illumina reads. However, Canu produced two to four mega-contigs per genome that essentially repeated 2 Mbp regions. Despite subsampling, genome duplication persisted, necessitating additional rounds of polishing and PCR for contig connection and orientation. Polishing with long-reads, however, exceeded our computational resources.

Unicycler provided a more efficient hybrid assembly approach using an OLC-based method. It automatically corrected PacBio errors with high-accuracy Illumina reads and independently assembled datasets before merging them into a simplified assembly graph. Long-reads resolved gaps and repetitive regions that short-read assemblers typically struggle to handle.

Assembly paths were visualised in Bandage. The Illumina short read SPAdes assembly appeared initially as a tangled ball of yarn due to the number of decisions left as possible paths due to frayed ropes in the assembly graph. Frayed ropes are a computational abstraction for regions where

repetitive elements cause multiple, equally plausible paths for the computer to assemble the data. Short-reads cannot resolve these regions, and the contig extension must be terminated, leading to bubbles representing these frayed ropes. When the long-read data was assembled independently, several contigs could span the entire genome. However, the long-read assemblies could not generate an assembly that resembled the genome sizes obtained by pulse field gel electrophoresis (Sengupta et al., 2015). In this regard, the genome assemblies were two or three times the published size.

However, the Unicycler assembly approach successfully created a high-accuracy genome assembly from the PacBio and Illumina datasets. The PacBio contigs were laced into the short-read Illumina assembly. The short-read assembly limited the genome reconstruction's size, preventing genome duplication. The long-reads resolved the contiguity issues caused by repetitive sequences, resulting in a single Eulerian walk that resolved the genomes. These new hybrid paths only contained a single, contiguous solution to the genome reconstructions. No further scaffolding was needed since the genomes were already in a single contigs per assembly.

Racon long read polishing refined the assemblies, followed by Pilon polishing with short-reads until no further improvements were detected. Unicycler reoriented the genomes to begin with dnaA or repA, and all assemblies passed quality control (CheckM, BUSCO, and QUAST). REAPR read alignment tests further confirmed structural accuracy.

Hybrid assembly using two short-read datasets reduced contig counts but left gaps that long-reads subsequently resolved, producing a single, scaffolded, reference-quality assembly. Although QUAST predicted varying numbers of unique genes, Unicycler's error-free results likely provide the most accurate genome representation.

Because the assembly pipeline was entirely *de novo*, the digital reconstructions might not fully reflect the true genome structures. However, validation using *in silico* restriction digestion and

published PFGE results confirmed their accuracy and a comparison was published in each genome announcement (Sengupta et al., 2015, Bailie et al., 2020, Bailie et al., 2021).

The final assemblies, as assessed with CheckM (99.18% completeness, 0.546% contamination), BUSCO (C:99.5%[S:98.8%,D:0.7%],F:0.5%,M:0.0%,n:402), comprised single contigs of 2,226,862 bp (AGR1485) and 1,939,032 bp (AGR1487), with GC contents of 51.15% and 52.17%, respectively, and no ambiguous nucleotides. Average read coverages were 989.2× for AGR1485 and 1,510.51× for AGR1487. Unicycler predicted circular chromosomes, confirmed via *in silico* digestion with UGENE, and both BWA and Bowtie successfully mapped all Illumina reads.

No plasmids were detected during laboratory DNA extractions using plasmid kits or analysing the long- and short-read datasets for plasmid sequences and structures. Furthermore, all the data aligned back to the genome assemblies with no plasmids. Given that all the data could be realigned with the genome assemblies, those assemblies matched the PFGE results for genome sizes. It is unlikely that plasmids were present in strains or datasets generated from them. The lack of plasmids suggests that the barrier-disruptive phenotype of AGR1487 is likely encoded within its single, closed genomic sequence.

4.7. Conclusion

The genomes of *L. fermentum* AGR1485 and AGR1487 were successfully sequenced using multiple technologies. Although short-read hybrid assemblies initially produced high completeness, they exhibited significant limitations that could not be resolved within the timeframe of this PhD program. Consequently, an alternative strategy that combined PacBio SMRT long-read sequencing with Illumina short-reads was successful. This hybrid approach yielded closed, reference-quality genome assemblies for both strains. The final assemblies were circular and plasmid-free, and they passed all quality control tests, including validation against published PFGE results. In subsequent chapters, these high-quality genome assemblies will serve as the foundation for genetic analyses and as alignment templates for mRNA sequencing.

4.8. References

Akabanda, F., Owusu-Kwarteng, J., Tano-Debrah, K., Parkouda, C., & Jespersen, L. (2014). The use of lactic acid bacteria starter culture in the production of nunu, a spontaneously fermented milk product in Ghana. *International Journal of Food Science, 2014*, 721067. https://doi.org/10.1155/2014/721067

Altermann, E., Bailie, M. A., Fraser, K., & Young, W. (2021). NexGen sequencing data: Bioinformatic tools for visualization and analysis. In A. Cifuentes (Ed.), *Comprehensive Foodomics* (pp. 47–90). Oxford: Elsevier.

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology, 21*(1), 30. https://doi.org/10.1186/s13059-020-1935-5

Anderson, R. C., Cookson, A. L., McNabb, W. C., Kelly, W. J., & Roy, N. C. (2010). *Lactobacillus plantarum* DSM 2648 is a potential probiotic that enhances intestinal barrier function. *FEMS Microbiology Letters, 309*(2), 184–192. https://doi.org/10.1111/j.1574-6968.2010.02038.x

Anderson, R. C., Ulluwishewa, D., Young, W., Ryan, L. J., Henderson, G., Meijerink, M., . . . Roy, N. C. (2016). Human oral isolate *Lactobacillus fermentum* AGR1487 induces a pro-inflammatory response in germ-free rat colons. *Scientific Reports, 6*, 20318. https://doi.org/10.1038/srep20318

Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., . . . Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology, 19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Boisvert, S., Laviolette, F., & Corbeil, J. (2010). Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of Computational Biology, 17*(11), 1519–1533. https://doi.org/10.1089/cmb.2009.0238

Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F., & Corbeil, J. (2012). Ray Meta: Scalable *de novo* metagenome assembly and profiling. *Genome Biology, 13*(12), R122. https://doi.org/10.1186/gb-2012-13-12-r122

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114–2120.

Bruggeman, F. J., & Westerhoff, H. V. (2007). The nature of systems biology. *Trends in Microbiology, 15*(1), 45–50. https://doi.org/10.1016/j.tim.2006.11.003

Carver, T., Berriman, M., Tivey, A., Patel, C., Böhme, U., Barrell, B. G., . . . Rajandream, M.-A. (2008). Artemis and ACT: Viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics, 24*(23), 2672–2676. https://doi.org/10.1093/bioinformatics/btn529

de Sá, P. H. C. G., Miranda, F., Veras, A., de Melo, D. M., Soares, S., Pinheiro, K., . . . Ramos, R. T. J. (2016). GapBlaster—A graphical gap filler for prokaryote genomes. *PLOS ONE, 11*(5), e0155327. https://doi.org/10.1371/journal.pone.0155327

Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O., & Salzberg, S. L. (1999). Alignment of whole genomes. *Nucleic Acids Research, 27*(11), 2369–2376.

Geneious. (2021). *About | Geneious*. Retrieved from https://www.geneious.com/about/

GitHub. (2021). *GitHub - GeneAssembly/kiki: De novo metagenomic assembly.* Retrieved from https://github.com/GeneAssembly/kiki

Golosova, O., Henderson, R., Vaskin, Y., Gabrielian, A., Grekhov, G., Nagarajan, V., . . . Huyen, Y. (2014). Unipro UGENE NGS pipelines and components for variant calling, RNA-seq and ChIP-seq data analyses. *PeerJ, 2,* e644. https://doi.org/10.7717/peerj.644

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., . . . The Bioconda Team. (2018). Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nature Methods, 15(7),* 475-476. https://doi.org/10.1038/s41592-018-0046-7

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

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., & Otto, T. D. (2013). REAPR: A universal tool for genome assembly evaluation. *Genome Biology, 14(5),* R47. https://doi.org/10.1186/gb-2013-14-5-r47

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology, 5(2),* 1-9.

Li, D., Liu, C. M., Luo, R., Sadakane, K., & Lam, T. W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics, 31(10),* 1674-1676. https://doi.org/10.1093/bioinformatics/btv033

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics, 27(21),* 2987-2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25(16),* 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., . . . Kristiansen, K. (2010). *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Research, 20*(2), 265–272.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., . . . Liu, Y. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read *de novo* assembler. *Gigascience, 1*(1), 18.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., . . . Chen, Z. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature, 437*(7057), 376.

McQuillan, J. A., Parkhill, J., Berriman, M., Harris, S. R., & Carver, T. (2011). Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics, 28*(4), 464–469. https://doi.org/10.1093/bioinformatics/btr703

Mikheenko, A., Valin, G., Prjibelski, A., Saveliev, V., & Gurevich, A. (2016). Icarus: Visualizer for *de novo* assembly evaluation. *Bioinformatics, 32*(21), 3321–3323. https://doi.org/10.1093/bioinformatics/btw379

Miller, J. R., Delcher, A. L., Koren, S., Venter, E., Walenz, B. P., Brownley, A., . . . Sutton, G. (2008). Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics, 24*(24), 2818–2824.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., . . . Remington, K. A. (2000). A whole-genome assembly of *Drosophila*. *Science, 287*(5461), 2196–2204.

Nadalin, F., Vezzi, F., & Policriti, A. (2012). GapFiller: A *de novo* assembly approach to fill the gap within paired reads. *BMC Bioinformatics, 13*(Suppl 14), S8. https://doi.org/10.1186/1471-2105-13-s14-s8

Okonechnikov, K., Fursov, M., Golosova, O., & team, t. U. (2012). Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics, 28*(8), 1166–1167. https://doi.org/10.1093/bioinformatics/bts091

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research, 25*(7), 1043–1055.

Paulino, D., Warren, R. L., Vandervalk, B. P., Raymond, A., Jackman, S. D., & Birol, I. (2015). Sealer: A scalable gap-closing application for finishing draft genomes. *BMC Bioinformatics, 16*(1), 230. https://doi.org/10.1186/s12859-015-0663-4

QIAGEN. (2015, June). *QIAGEN Genomic DNA Handbook.* Retrieved from https://www.qiagen.com/us/resources/resourcedetail?id=d2b85b26-16dd-4259-a3a7-a08cbd2a08a3&lang=en

Rose, R., Tiunov, A., Sukhomlinov, D., Golosova, O., & Prosperi, M. (2018). Flexible design of multiple metagenomics classification pipelines with UGENE. https://doi.org/10.1093/bioinformatics/bty901

Sengupta, R., Anderson, R. C., Altermann, E., McNabb, W. C., Ganesh, S., Armstrong, K. M., . . . Roy, N. C. (2015). *Lactobacillus fermentum* AGR1487 cell surface structures and supernatant increase paracellular permeability through different pathways. *MicrobiologyOpen, 4(4),* 541-552. https://doi.org/10.1002/mbo3.260

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research, 19*(6), 1117–1123. https://doi.org/10.1101/gr.089532.108

Tanay, A., Sharan, R., Kupiec, M., & Shamir, R. (2004). Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proceedings of the National Academy of Sciences of the United States of America, 101*(9), 2981–2986. https://doi.org/10.1073/pnas.0308661100

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., . . . Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE, 9*(11), e112963. https://doi.org/10.1371/journal.pone.0112963

Weisburg, W. G., Barns, S. M., Pelletier, D. A., & Lane, D. J. (1991). 16S ribosomal DNA amplification for phylogenetic study. *Journal of Bacteriology, 173*(2), 697–703.

Wences, A. H., & Schatz, M. C. (2015). Metassembler: Merging and optimizing *de novo* genome assemblies. *Genome Biology, 16*, 207. https://doi.org/10.1186/s13059-015-0764-4

Whitwham, A., & Bonfield, J. K. (2010). Gap5—Editing the billion fragment sequence assembly. *Bioinformatics, 26*(14), 1699–1703. https://doi.org/10.1093/bioinformatics/btq268

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology, 13*(6), e1005595.

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: Interactive visualization of *de novo* genome assemblies. *Bioinformatics, 31*(20), 3350–3352.

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for *de novo* short read assembly using *de Bruijn* graphs. *Genome Research, 18*(5), 821–829. https://doi.org/10.1101/gr.074492.107

Zimin, A. V., Puiu, D., Marçais, G., Yorke, J. A., Roberts, M., & Salzberg, S. L. (2013). The MaSuRCA genome assembler. *Bioinformatics, 29*(21), 2669–2677. https://doi.org/10.1093/bioinformatics/btt476

Chapter 5: Genetic Analysis of AGR1487

5.1. Abstract

Background: *Limosilactobacillus fermentum* strain AGR1487 exhibits a unique phenotype by disrupting tight junctions in transepithelial electrical resistance (TEER) assays, an effect not observed in other strains, raising concerns about the potential for genes that may compromise gastrointestinal health.

Aim: This study aimed to identify candidate genes responsible for AGR1487's barrier-disruptive phenotype using comparative genomics.

Methods: Genomic similarity among 40 *L. fermentum* strains was assessed using Average Nucleotide Identity (ANI) and Mash analyses. Synteny analyses were performed on eight closed genomes, including AGR1487 and AGR1485, to identify regions of high sequence variation with elevated AT content indicative of horizontal gene transfer. Pan-genome analysis was conducted to delineate core versus accessory genomic elements, and custom BLAST scripts identified genes unique to AGR1487 and their genomic locations. A gene was considered "unique" or "non-identical" if it lacked a 100% identical match in any of the other 39 strains, either due to complete absence or sequence divergence.

Results: The genomes share an average nucleotide similarity of 96.9%, with approximately 3% divergence localised to dispersed high-variation regions. Pan-genome analysis revealed that each strain possesses about 67 unique genes on average, with AGR1487 containing 62. A focused examination of AGR1487's unique gene set identified 14 candidate targets, 12 hypothetical proteins and two enzymes within domesticated phage regions that may underlie its disruptive phenotype.

Conclusion: The discovery of 14 candidate genes in AGR1487's high-variation regions provide a targeted basis for future studies on the genetic determinants of its barrier-disruptive properties, offering new insights into strain-specific genomic variations that may impact gastrointestinal health.

5.2. Introduction

The rapid progress in genomics and biotechnology has illuminated the versatile potential of microorganisms across domains like medicine, agriculture, and industry. Lactic acid bacteria (LAB), particularly *Limosilactobacillus fermentum*, have garnered interest due to their beneficial impact on human and animal health, food preservation, and biomedical applications (Naghmouchi et al., 2020). This species, previously recognised as *Lactobacillus fermentum*, is noteworthy for its role in various food fermentation processes (Zheng et al., 2020). Conversely, a specific strain of *L. fermentum*, AGR1487, demonstrates a "barrier-disruptive phenotype", meaning it may be able to weaken the gastrointestinal tract (GIT)'s epithelial barrier (Chapter 3) (Sengupta et al., 2015). This phenomenon could cause contents to be translocated across the intestinal epithelium, negatively impacting the host and food safety. Comprehending the genetic underpinnings of this characteristic is paramount. Today, genome sequencing and genetic analysis are vital instruments for probing the genetic attributes of LAB and other microorganisms, enabling the scrutiny of potentially deleterious traits at the strain level (Sun et al., 2015).

Next-generation sequencing (NGS) technologies have catalysed genomic research by providing rapid, high-throughput, and economical whole-genome sequencing (Mardis, McPherson, Martienssen, Wilson, & McCombie, 2002) and by extension, long-read sequencing facilitated the transition to whole complete bacterial genome assemblies, like those created for *L. fermentum* AGR1485 and AGR1487 (Bailie, Altermann, Young, Roy, & McNabb, 2020, 2021). These technologies have been harnessed to sequence and analyse the genomes of various LAB strains, allowing an in-depth understanding of their metabolic functions, stress responses, and other properties (Douillard et al., 2013). The comparative genomics of *Lactobacillus rhamnosus* via pan-genomics and genome synteny analysis elucidated genes related to carbohydrate utilisation, bile salt hydrolase activity, and host adhesion (Ceapa et al., 2016). Although many of these genes were described, such

techniques may be powerful tools for detecting genetic foundations of potentially harmful traits by contrasting genomes of different strains of *L. fermentum*.

The *L. fermentum* AGR1487 genome (1,939,032-bp) is 287,830-pb smaller than *L. fermentum* AGR1485's genome (2,226,862-bp), indicating a genetic difference between the strain's genomes (Chapter 4). Examining the genetics of *L. fermentum* AGR1487 promises insight into the genes underpinning its barrier-disruptive phenotype, ascertained through transepithelial electrical resistance assays (TEER).

5.3. Aims and Hypotheses

The intestinal epithelial barrier-disruptive characteristic of AGR1487 is not ubiquitous within this bacterial species (Chapter 2). Therefore, genetic constituents underlying this phenotype in AGR1487 could be attributed to genetic divergences unique to this strain.

Hypothesis 1: The strains of *L. fermentum* are closely related, justifying the exclusion of conserved regions of the genomes from further detailed genetic analysis.

Hypothesis 2: The ability of AGR1487 to compromise the barrier integrity of Caco-2 monolayers during TEER assays is due to one or more non-identical genes unique to AGR1487.

Aim 1: Conduct a comparative genomic analysis of various *L. fermentum* strains to identify and characterise conserved and variable regions that may be linked to the barrier-disruptive phenotype.

Rationale: Since the barrier-disruptive phenotype is presumed rare among *L. fermentum* strains, most of their genomes are potentially conserved. Ignoring these conserved regions allows a more efficient narrowing of the search space, highlighting the limited variable segments that could harbour the genes responsible for the barrier-disruptive phenotype.

Aim 2: Systematically identify genes unique to the AGR1487 and refine this selection by leveraging insights from previous research.

Rationale: Including genetic data from a broad spectrum of *L. fermentum* strains in the analysis enhances the likelihood of isolating genes exclusive to the AGR1487's genome. These regions of uniqueness may be essential for the barrier-disruptive properties of AGR1487. Subsequent refinement of gene annotations, informed by metadata from earlier studies, will enable a more precise identification of likely gene candidates that play a pivotal role in this phenotype.

5.4. Materials and Methods

All software applications were run using their default settings unless otherwise stated. The applications were regularly updated before running the analyses. All software packages were run locally on workstations or servers within Conda (v22.11.1) environments on the Linux Ubuntu 20.04.5 LTS (Focal Fossa) operating system. NCBI RefSeq GenBank files were used for all analyses. All scripts associated with the workflows described here are provided in the digital appendix under the same chapter name. In this chapter, various scripts were utilised and written in multiple programming languages, including Bash (Bourne Again SHell), Perl, Python, and R. These diverse languages facilitate the efficient execution of different analytical tasks in these bioinformatics pipelines. Statistics were handled by the packages and pipelines used for these analyses and stored in the digital appendix.

5.4.1. Linearisation and Start Sites

For correct alignment, the circular genomes of the *L. fermentum* strains were linearised for comparison. Since linearisation could occur at any locus within a genome assembly, a fixed reference site was established for all strains to avoid alignment artefacts. The *dna*A gene (accession number: UniRef90_B2GEU8) was selected as a reference point because the gene is conserved and necessary for chromosome replication. The software Circulator (1.5.5) was employed to synchronise the start of all linearised genome assemblies (n = -1) upstream of the dnaA gene locus.

circlator fixstart \

--genes_fa Starting_gene_sequence.fasta \

--orig_fa original_genome_assembly.fasta output_prefix

5.4.2. Genomic Similarity

Average Nucleotide Identity (ANI) was the primary method for assessing genomic similarity between bacterial strains. ANI, a quantitative measure, calculates the mean nucleotide identity of coding regions shared between two genomic sequences. This analysis utilised 34 NCBI RefSeq genomic sequences in FASTA format and included *Lactiplantibacillus plantarum* WCFS1 as an external control and outlier. The Skani package, part of the Biopython software suite, performed the ANI. Skani (0.2.1) (Shaw & Yu, 2023) calculates the ANI value by aligning orthologous gene pairs to assess nucleotide identity as a percentage. The Skani pipeline and the resulting output were run, processed and visualised into an ANI clustermap by the ANIclustermap pipeline (1.2.0) (Shimoyama, 2022a) with the following command: *ANIclustermap -i Anicluster/data -o Anicluster/all_genomes3 --fig_width 20 --fig_height 15 --annotation*.

In addition to ANI, the genomic similarity was also calculated by Mash (V2.3 on GCC-11.3.0) for the same dataset (Ondov et al., 2016). Mash, based on the MinHash algorithm, estimates genomic sequence similarity through the Mash score (Ondov et al., 2016). This score is derived from the Jaccard index, comparing k-mers from genomic sequences. Mash simplifies this process by converting k-mers into bit strings using hash functions and then applying MinHash to create a condensed 'sketch' represented as a triangular matrix. This method facilitates rapid similarity estimation across large datasets. The Mash score is particularly useful for identifying closely related species, analysing microbial diversity, and monitoring the spread of genetic variants. Command: *mash triangle 25_lf_fs.fa > 25_lf_fs.fa_mash*

5.4.3. Whole Genome Alignments

The dataset was downloaded from NCBI's RefSeq database. All the genomes were annotated with the Prokaryotic Genome Annotation Pipeline (PGAP) to identify and annotated protein conding

ORFs. Once the ORFs were annotated they were referred to as genes throughout the text, except where unannotated sequence was translated and then used for a comparison. The translated ORF were compared with MMseq2.

Gene comparisons were performed using Mmseqs2's reciprocal best-hit (RBH) approach. Mmseqs2 is a fast and sensitive software suite for protein sequence analysis, capable of identifying orthologs between species based on the highest-scoring protein sequence alignments. The RBH approach involves protein sequence extraction, sequence search, reciprocal best-hit identification, and clustering of orthologous pairs into groups. The Mmseqs2 alignments (June 2021) were visualised with Moshi pgv-pmauve (version 0.3.2) (Shimoyama, 2022-06-20). This tool provides interactive visualisation of whole-genome alignments and comparisons, facilitating the identification of conserved regions and potential rearrangements. Brown lines represented features in the same orientation, while green lines represented features reversed between adjacent genomes. The MMseq2 calculated the identity match between two protein sequences to be a number between 0 - 100%. The resulting identity match score between linked proteins was graphically indicated by the opacity of the lines, from transparent (0% match) to solid (100%).

Locally Collinear Blocks (LCBs) blocks were computed with ProgressiveMauve (2.4.0), a multiple genome alignment tool designed explicitly for detecting homologous regions between multiple input nucleotide sequences (Darling, Mau, & Perna, 2010). ProgressiveMauve identifies LCBs as conserved segments across input genomes while accounting for potential rearrangements and inversions. The generated alignments were analysed for genomic rearrangements, synteny, and structural variations via the included Windows 10 general user interface (GUI) and presented as screenshots.

5.4.4. PAN-Genomics

To ensure a comprehensive and accurate pan-genome representation, the pan-genome analysis was performed using three software packages: PANX, Bacterial Pan-genome Analysis (BPGA), and Roary.

PANX (version 1.6.0) (Ding, Baumdicker, & Neher, 2017) is an HTML-based pan-genome analysis tool that was run locally. The software suite employs a pipeline that includes gene annotation, orthologous gene clustering, and pan-genome profiling. PANX's default output was visualised using the interactive HTML package pan-genome-visualisation provided for PANX. The non-identical genes were identified from the HTML output and combined using a custom script.

BPGA (version 1.3) (Chaudhari, Gupta, & Dutta, 2016) is a versatile and efficient software for pan-genome analysis that utilises a scalable and memory-efficient algorithm. The BPGA pipeline includes gene annotation with Prokka, orthologous gene clustering, pan-genome profiling, and additional comparative genomic analyses. The BPGA clustering was always processed with USEARCH (version 11) (Edgar, 2010). The system was queried for the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and Neighbour-Joining (N.J.). No difference between the inferred trees was detected for the 33 closed RefSeq *L. fermentum* genomes.

Roary (version 3.11.2) (Page et al., 2015) is a rapid large-scale prokaryote pan-genome analysis tool that uses a graph-based approach for orthologous gene clustering. The downstream data was analysed with Python and Perl packages provided by the Sanger Institute. Further analysis was conducted using Pagoo following the published protocol (Ferrés & Iraola, 2021). Pagoo is a computational tool developed for systematically organising and visualising pangenome data. It consolidates pangenome analysis outputs, enabling interactive exploration and interpretation of the pangenome structure and gene content. Identifying non-identical genes per strain is not a feature of

the pangenome software packages used for this work; therefore, it was necessary to write custom scripts and BLAST databases to find the non-identical genes.

5.4.5. Non-Identical Gene Identification

A unique gene is defined as an AGR1487 genes for which no 100% identical, full-length BLASTP match exists in any other *L. fermentum* strain (xxii - Thesis specific definitions). Standard pan-genome analysis tools (PANX, BPGA, Roary) identify orthologous gene clusters across strains but do not directly output strain-specific gene lists. To identify genes present only in AGR1487, a custom BLAST database was constructed containing all AGR1487 protein sequences with their accession numbers. The pan-genome outputs were queried against this database to identify AGR1487-specific entries. For Roary analysis, the internal BLASTP results (_blast_results file) and clustered protein sequences (_clustered file) were used to extract percent identity, percent coverage, and bit scores for each AGR1487 gene. This approach enabled filtering of the pan-genome 'cloud' gene category to identify candidates that were non-identical to proteins in other *L. fermentum* strains at the specified identity thresholds (100% and 95%).

Nucleotide and protein databases were constructed with the following commands:

esearch -db protein -query PRJNA596816 | efetch -format fasta > ref/AGR1487_proteins.fa

makeblastdb -in ref/AGR1487_proteins.fa -dbtype prot -out ref/AGR1487_prots -parse_seqids

PANX results were combined into a single file for ease of use and translated into protein sequences to employ the BLAST pipeline. The individual text files for each of the positive, unique hits in nucleic acid were translated with a standard loop:

for file in ref/panx/alignments/.fa; do transeq -sequence $file -outseq ref/panx/translation/$(basename $file .fa).fasta; done as follows: cat ref/panx/translation/*.fasta > ref/panx/panx_combined.txt*; *done*

The resulting translated outputs were combined into a single file as follows:

cat ref/panx/translation/.fasta > ref/panx/panx_combined.txt*

The sequences were systematically searched against the custom database for the associated protein accession numbers.

blastp -db ref/AGR1487_prots -query ref/panx/panx_combined.txt -outfmt '6 qacc sacc pident qlen length' | sort -k 3rn > ref/panx/PANX_AGR1487_UNIQUE.txt

Protein accession numbers were extracted from the file with the following pipeline:

cut -f2 ref/panx/PANX_AGR1487_UNIQUE.txt | grep . > ref/panx/IDX_panx.txt

Additional detail, annotation and information for the final protein accession numbers were requested and automatically formatted with the following pipeline:

cat ref/panx/IDX_panx.txt | epost -db protein -format acc | efetch -db protein -format fasta > ref/panx/panx_output_final.fasta

BPGA and Roary followed a process similar to PANX without translating and combining the outputs. Instead, the accession numbers were extracted from the results and saved as a single-column list of accession numbers. These accession number lists were searched against the NCBI database to retrieve annotation details using the command:

blastp -db ref/AGR1487_prots -query ref/BPGA_default/REPSEQ_UNIQUE.txt -outfmt '6 qacc sacc pident qlen length' | sort -k 3rn > ref/BPGA_default/REPSEQ_AGR1487_UNIQUE.txt

A complete script, workflow and terminal transcript are provided in the digital appendix titled “Blast_for_unique_genes”, complete with annotation and commentary. The resulting files were compared with a custom Python script called “Search_merge_annotate.py” in the digital appendix. The script was written to find accession numbers in all three outputs or combinations. PANX and

BPGA analyses were set with a unique tolerance of 100%, while Roary was set to a default 95% threshold for uniqueness. Two comparisons were conducted: firstly, only PANX and BPGA, the genes with 100% uniqueness, and then all three datasets to find the genes within the PANX and BPGA datasets that did not undergo reductive evolution based on the hits from Roary.

Table 5.4-9: BLAST and sorting options

OPTION	DESCRIPTION
QACC	Query accession, which is an identifier for the sequence being compared.
SACC	Subject accession is an identifier for the sequence the query is being compared to.
PIDENT	Percent identity is the percentage of identical amino acids or nucleotides between the query and subject sequences.
QLEN	Query length, which is the length of the query sequence.
SORT	Sorting lines of text in a file.
-K 3	Specifies that the sorting should be done based on the third column of the file.
-R	Specifies that the sorting should be done in reverse order.
-N	Sorting should be done numerically rather than alphabetically.

The BLASTP statistics presented in Table 5.5-13 and Table 5.5-14 were derived from comparative analysis against 39 non-AGR1487 *L. fermentum* strains using the custom script "blast_against_all_strains.py" that is available in the digital appendix. Each column is defined as follows:

- **ID**: The RefSeq locus tag identifier for each gene in the AGR1487 genome (e.g., RS00335).
- **Annotation**: Functional annotation assigned by Prokka during genome annotation.
- **Start/End**: Genomic coordinates (base pairs) of the gene on the AGR1487 chromosome.
- **nt Length**: Nucleotide length of the gene, calculated as (End - Start + 1).

- **aa Length**: Amino acid length of the translated protein, calculated as nucleotide length divided by 3.
- **Best Hit Strain**: The strain containing the closest homolog to the AGR1487 protein, or "No hit" if no significant similarity was found.
- **Best Hit Percent ID**: Percentage sequence identity from BLASTP alignment against the closest homolog in other strains. "No hit" indicates genes unique to AGR1487.
- **Best Hit Percent Coverage**: Percentage of the AGR1487 protein sequence aligned in the best BLASTP hit, indicating whether the match represents a full-length or partial protein alignment.
- **Best Hit E-value**: BLASTP E-value representing the statistical significance of the alignment; lower values indicate more significant matches.
- **Best Hit Bit score**: BLASTP bit score representing the quality of the alignment; higher scores indicate better matches.

Genes showing "No hit" across all columns represent proteins that are absent from the other 39 L. fermentum strains, indicating AGR1487-specific gene content. Genes with low percent identity (<50%) represent strain-specific variants with distant homologs, while higher identity values indicate conserved proteins with sequence variation.

5.4.6. Gene ranking

Two custom pipelines (*map.sh and annotate.py*) were developed to identify, score, and annotate AGR1487 candidate genes based on sequence homology and specificity across *L. fermentum* strains using the MMseqs2 bitscores are provided in the deletion priority score (DPS) folder in the digital appendix. The bitscore reported by MMseqs2 are a normalised measure of alignment strength that

integrates both match quality and alignment length. It is derived from the raw alignment score and adjusted using statistical constants that allow comparisons across databases and sequence lengths. In practical terms, a higher bitscore indicates a longer, more identical, and more confident alignment between two sequences. Unlike percentage identity alone, the bitscore accounts for alignment coverage, substitution matrix weights, and gap penalties, making it a robust indicator of overall sequence similarity. The ranking also integrates transcriptional activity measured during AGR1487 cultivation in MRS media, prioritizing genes that are actively expressed and down-weighting those with minimal or absent expression.

The bitscores were used because they provide a unified measure of sequence relatedness, enabling fair comparisons across genes of different lengths and between self and non-self-alignments. This allows the ranking algorithm to emphasize genes that are strongly conserved within AGR1487 but weakly matched in other *L. fermentum* strains, highlighting potential strain-specific loci.

Table 5.4-10: terms, names and variables used for the ranking output

TERM	DESCRIPTION	CONTRIBUTION TO RANKING
RSID	Gene identifier from the AGR1487 genome (e.g., RS03620).	Key identifier used for joining annotations.
QID	Query ID of the AGR1487 protein sequence in the reference genome.	Maps the sequence in the assembly.
CP	Contig or chromosome accession (e.g., NZ_CP047585.1).	Context for genomic origin.
BEST_SELF_BITSCORE	Highest BLAST bitscore when aligned against AGR1487 (self).	Reflects intra-strain sequence strength.
MEAN_SELF_BITSCORE	Mean bitscore of all self-hits for that RSid.	Captures within-strain conservation.
N_SELF_HITS	Number of self-hits.	Indicates paralogy or redundancy.
BEST_NONSELF_BITSCORE	Highest bitscore against any non-AGR1487 strain.	Measures potential cross-strain similarity.
SELF_VS_NONSELF_RATIO	Ratio of self vs non-self bitscore.	Higher ratio = higher strain specificity.
NEW RANK	Composite score integrating the above metrics.	Used to order genes by AGR1487-specificity.
ORDER	Rank order (1 = highest specificity).	Final priority measure for candidate selection.

5.4-4: Ranking formula

$$rank\_score = (\ (best\_self\_bitscore) \times (1 + self\_vs\_nonself\_ratio) + 0.1 \times mean\_self\_bitscore \ + 0.01 \times n\_self\_hits) \times f(Exp)$$

$$where\ f(Exp) =$$

$$\begin{cases} 1 + 0.2 \times Exp, & if\ Exp > 0 \\ e\text{^}(Exp\ /\ 2), & if\ Exp \leq 0 \end{cases}$$

To validate the cloud gene identification and refine the ranking, full-length protein sequences were extracted from Prokka-annotated AGR1487 proteins and compared against Prokka-annotated proteins from 38 other *L. fermentum* strains using BLASTP. This validation addressed potential discrepancies arising from the use of Roary pan-genome representative sequences, which may not capture complete open reading frame (ORF) boundaries.

To incorporate these validation results into the ranking, a penalty/bonus factor g(pident) was applied to the original ranking scores:

5.4-5: Prokka data modification to the ranking formula

$$new_{rank_{score}} = rank_{score} \times g(pident)$$

$$g(pident) =$$

$$\begin{cases} 2.0 & if\ no\ homolog\ detected \\ 1.0 - 0.5 \times (pident/100)^2 & if\ pident \geq 95\% \\ 1.0 - 0.2 \times (pident/100) & if\ pident < 95\% \end{cases}$$

This modification rewards genes that are truly absent from other *L. fermentum* strains (2× bonus) while penalising genes with near-identical homologs in other strains. The quadratic penalty for genes with ≥95% identity ensures that those approaching 100% identity receive substantially harsher penalties (factor of 0.5 at 100% identity) than genes with moderate similarity (factor of 0.9 at 50% identity). Genes with lower identity values receive milder linear penalties, reflecting that sequence divergence may indicate functional differentiation even when homologs are present.

5.4.7. Genome Atlas

A genome atlas was constructed with pyCirclize (1.2.0) (Shimoyama, 2022b). pyCirclize is a package for producing Circos Plots and Chord Diagrams in Python. The AGR1487 genome atlas was created with a customised configuration file designed to align the genome sequencing in its circular form with the mmseq2 protein alignments (5.4.3) and blast results indicating the location of unique genes (5.4.5). The entire file, subsequent runs and configuration files are provided in the digital appendix titled “genome_atlas_AGR1487”.

5.5. Results

5.5.1. Similarity of the Genome Assemblies

An ANI was calculated in an all-versus-all experimental design to determine the degree of difference between the genomes in the dataset. The ANI analysis estimates genome similarity between 0% (no similarity) to 100% (identical). A strain of *Lactiplantibacillus plantarum* (WCFS1) was added as a control. WCFS1 had no similarity in ANI with any *L. fermentum* strains. The strains of *L. fermentum* had a range of similarity from 96.9% to 99.9.% (Figure 5.5-1). The apparent lack of similarity between *L. plantarum* WCFS1 and the *L. fermentum* genomes may result from nucleotide homology below 70%, a threshold of accuracy for the Skani software package (Shaw & Yu, 2023).

However, all *L. fermentum* strains exhibited an ANI above 96.9%. In addition to the ANI, the mash distances were also calculated as an alternative method to the ANI in an all-versus-all experimental design to probe the genome similarity further without whole genome alignments.

The mash distance, also called the Mash score, measures the dissimilarity between a pair of genomes based on the MinHash signatures (Ondov et al., 2016). The score is calculated as one minus the fraction of matching hashes between the two signatures (Ondov et al., 2016). A Mash distance of 0 implies identical signatures (and thus, highly similar genomes), while a distance of 1 implies completely different signatures. Along with the Mash distance, Mash can also provide a p-value and confidence interval (CI), offering a statistical measure of the likelihood that the observed similarity or dissimilarity is not due to chance. The p-value range detected between two signatures was $3.05e^{-04}$ - $4.43e^{-08}$. The test indicated that in addition to the *L. plantarum* WCFS1 genome, ten other extra-chromosomal contigs also scored a dissimilarity score of 1 compared to all genomes, indicating that they were unique. These ten contigs did belong to *L. fermentum* strains but also deviated from the patterns of similarity already observed. The contigs were not detected by chance, nor were they false positives. The genomes these ten sequences were associated with also had the largest deviation of the *L. fermentum* genomes, around Mash distances of 0.24. In contrast, the other genomes ranged between 0.012 and 0.00019 in support of the ANI values. These contigs were not assembled into the complete genomes associated with the sequences. This would only be tolerated during genome assembly for metagenomes, plasmids or viruses.

A closer inspection revealed that the ten divergent strains contained plasmids that were detected as extra-chromosomal contigs. All ten plasmids' accession numbers corresponded to 10 of the 11 genomes with lower Mash scores (Table 5.5-11). Removal of the plasmid sequences during a subsequent experiment brought all the genomes into the same 0.012 – 0.00019 score range, indicating that plasmids produce the largest detectable source of variation using this method. Both Mash and ANI show that the genomes have a 96.9 – 99.9% overall similarity over the entire dataset. Given that

the genomes have roughly a 3.1% ANI and Mash distance difference, each genome could theoretically hold 66.6 unique genes, which is 3.1% of the average number of detected OPFs found for *L. fermentum*. This value will be tested during the pan-genome analysis.

Table 5.5-11: Plasmids found by Mash

ACCESSION NUMBER	NCBI ANNOTATION
CP011537.1	*Limosilactobacillus fermentum* 3872
CP047585.1	*Limosilactobacillus fermentum* strain AGR1487 chromosome, complete genome
CP033371.1	DR9 host of 3375.1
CP033375.1	*Limosilactobacillus fermentum* strain DR9 plasmid unnamed4, complete sequence
CP076447.1	*Limosilactobacillus fermentum* strain L1 plasmid unnamed1, complete sequence
CP034100.1	*Limosilactobacillus fermentum* strain LMT2-75 plasmid p1, complete sequence
AP017974.1	*Lactobacillus fermentum* plasmid pLF25067 DNA, complete genome, strain: MTCC 25067
CP019032.1	*Limosilactobacillus fermentum* strain SNUV175 plasmid pSNU175-3, complete sequence
CP035055.1	*Limosilactobacillus fermentum* strain SRCM 103290 chromosome, complete genome (plasmid inserted)
AP024321.1	*Limosilactobacillus fermentum* ike38 plasmid pike38 DNA, complete sequenc

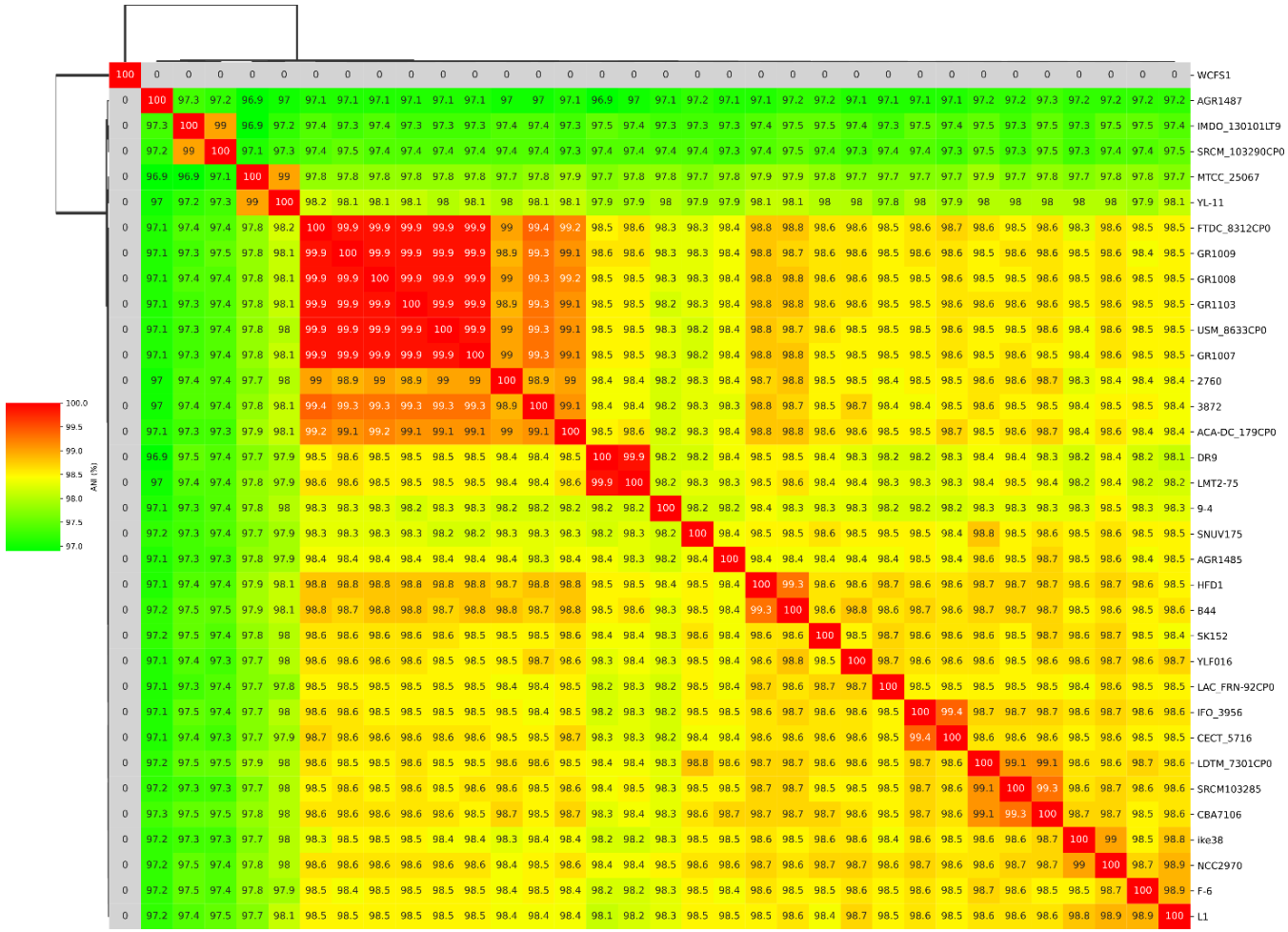

Figure 5.5-1: ANI heat map. The ANI results were coloured from red (100% similarity) to green (97%). Genomes with little or no similarity were coloured grey if their similarity fell below 50%. The dataset contains 30 closed genomes of *L. fermentum* and the *L. plantarum* WCFS1 genome (control) from the NCBI RefSeq library.

5.5.2. Whole Genome Alignments

Please note that ORF in this chapter referred to annotation translated open reading frames (ORF) so that the protein sequences can be compared. MMseq2 is a software package that employs the RBH approach to compare protein sequences and identify orthologous regions (Steinegger & Söding, 2017). Individual p-scores for each translated ORF alignment were stored in the digital appendix. The results were filtered ($p < 0.05$) before further analysis. On average, *L. fermentum* genomes have a total of 2,147 ORFs, of which Mmseq2 found homology for an average of 1,622 of the available ORFs. The average identity match for the homologous ORFs was greater than 95% for all the strains of *L fermentum*. The Mmseq2 results implied that, on average, 525 of the ORFs were non-homologous because the application reported partial homology between ORFs. For instance, the alignment between AGR1485 and *L. plantarum* WCFS1 found 1297 homologous ORF with an average similarity of 55.80%. The number of homologous ORF and non-homologous reading frames changed depending on the genome pairing and was sensitive to which genome was used as the reference. An example of some of the alignments is provided in Table 5.5-12 and visually in Figure 5.5-2.

The ORF order between *L. fermentum* genomes was highly conserved and limited to defined regions, separated by gaps with unconserved content. The unconserved content ranged from unique genes to degraded or missing loci. The gaps had few or no alignments with other genomes, indicating areas of high variation. The inversion observed in AGR1485 when the AGR1487 genome was used as a reference is no exception and fell within one of these high variation zones. The locus was not inverted in strains AGR1487, B44, L1 and 2760. Still, its location was inconsistent between the genomes. Palindromic repeats border the inversion, and the gene content between the repeats contains transposons and various phage-related genes. Despite this inversion, the genomes had an average similarity of 98.37% (Table 5.5-12).

LCBs were calculated using ProgressiveMauve to enhance the visualisation of indels, inversions, and translocation events for the AGR1487, B44, L1 and 2760 strains. A clear inversion was identified to start at position 1,550,000 nt into the AGR1485 genome, which translocated and

inverted at position 750,000 nt in the AGR1487 genome (Figure 5.5-3). While other translocation events were observed, no other large inversions were found.

These results show that the structure, gene order and identity of the features within genomes from strains of *L. fermentum* are conserved. Unfortunately, TEER assay testing is limited, and the barrier-disruptive phenotype has not been described for any other strain tested here. Given that AGR1485 does not produce the barrier-disruptive phenotype and shares these regions of similarity with AGR1487 and the other strains, these regions of similarity are unlikely to be the source of the barrier-disruptive phenotype. With the understanding of what was shared between strains, finding the unique gene content was the next step. However, an all-versus-all analysis was impractical for the 40 NCBI RefSeq genomic sequences available as a dataset at the time of writing. Pan-genome analyses could achieve this goal.

Table 5.5-12: MMseq2 alignment

ALIGNMENT	NUMBER OF ALIGNED FEATURES	AVERAGE MATCH AS A PERCENTAGE
AGR1487 VS B44	1,568	97.88%
B44 VS L1	1,632	98.77%
L1 VS 2760	1,680	97.95%
2760 VS AGR1485	1,934	98.87%
AGR1485 VS WCFS1	1,297	55.80%

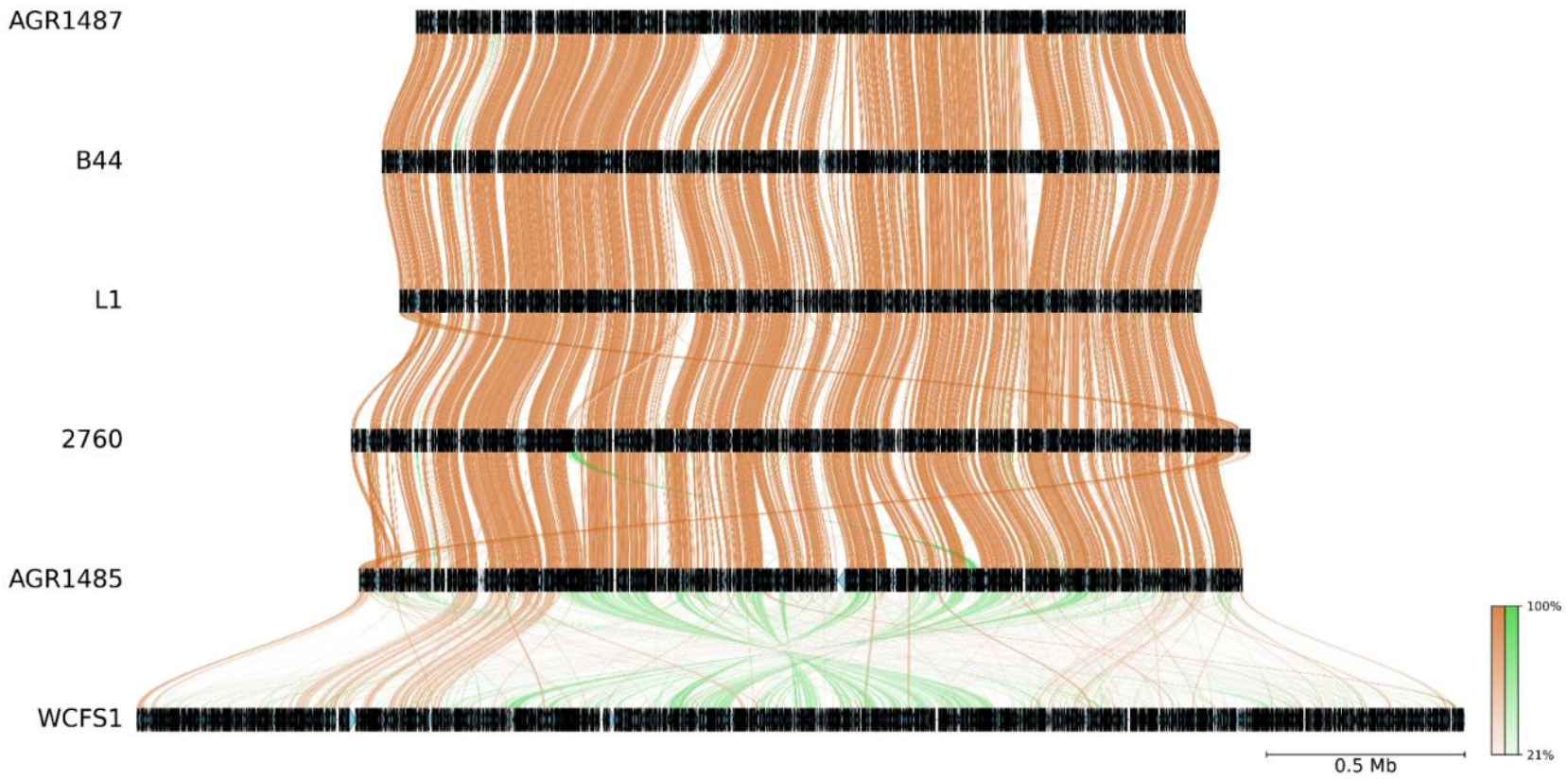

Figure 5.5-2: Mmseqs2 reciprocal best-hit. Closed *L. fermentum* AGR1487, 2760, L1, AGR1485, B44 and *L. plantarum* WCFS1 directly aligned against one another. Black arrows indicate genomic features provided by feature files downloaded from NCBI's RefSeq database. Brown lines show homologous matches in the same orientation. Matches in the reverse orientation are shown as green lines. The opacity of the linking lines is directly related to the identity score for the feature and its similarity between genomes.

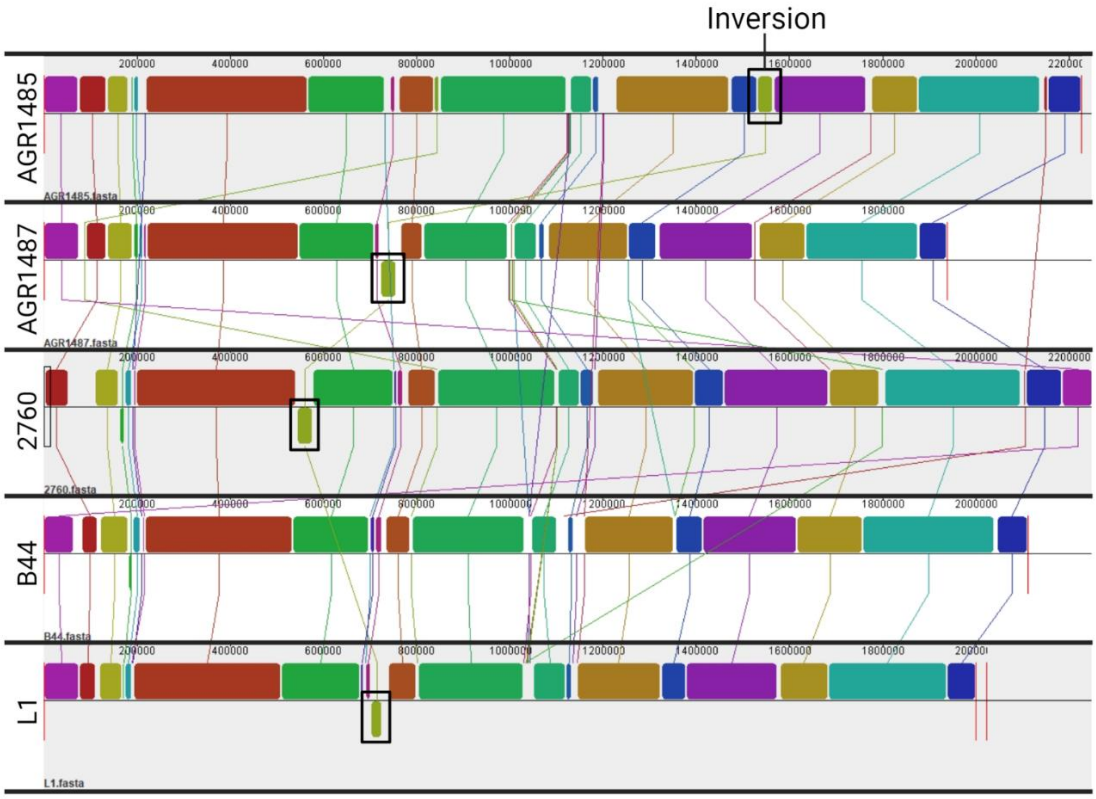

Figure 5.5-3: ProgressiveMauve LCB map. Closed *L. fermentum* AGR1487, 2760, L1, AGR1485, B44 and *L. plantarum* WCFS1 aligned with ProgressiveMauve. Locally, Collinear Blocks are coloured independently according to their identity, which is determined as a homologous region. A connecting line with the same colour as the LCB was used to connect the LCBs between genomes. All LCBs in the forward or similar direction are shown above the genome line, and any LCBs shown below the line are inversions relative to the reference sequence. AGR1485 was held as the reference sequence for this experiment.

5.5.3. PAN-Genomics

A gene-level analysis of a single strain can be complex and lacks contextual information derived from the species. A pan-genome analysis was conducted to establish this species-level context. The pan-genome was explored using PANX, BPGA, and Roary. The Roary pan-genome software was utilised to derive graphical results because of the accessibility and the availability of numerous community-developed pipelines tailored specifically for analysing Roary's output data. The pan-genome was constructed using a dataset comprised of 40 closed and PGAP-annotated *L. fermentum* genomes from the National Center for Biotechnology Information's (NCBI) RefSeq database. The data was procured from the NCBI database on September 10, 2022.

Pan-genome analyses were conducted with 1,000 iterations. Core genes, representing the conserved elements across the species, reached a stable convergence with a minimum inclusion of 17 genomes (Figure 5.5-4). In this context, convergence refers to the point at which calculated estimates stabilise and no longer significantly alter with additional genomes. Only fully closed genomes were included to maintain the robustness of the analysis.

The pan-genome structure was compartmentalised into core, soft-core, accessory, and unique (or cloud) genes (Figure 5.5-4 A). Core genes, present in all strains, typically encompass genes essential for survival. The soft-core category is used to describe genes that would have been core genes but are missing from one or two genes. Accessory (or shell) genes, found in six or more strains, often endow strains with additional capabilities like antibiotic resistance or virulence factors, enhancing their competitive advantage in certain environmental conditions. Unique (or cloud) genes, exclusively present in a single strain or two, illustrate individual strains' unique adaptations and capabilities. Among the analysed genes, 4,887 were classified as cloud genes across more than six strains. Furthermore, 1,490 shell genes were found to be common to many strains. The pan-genome encompassed 1,101 core genes in all 40 strains (Figure 5.5-4 C) and 161 soft core genes in 38 of the 40 strains.

The same dataset generated the rarefaction curve with 1,000 iterations (Figure 5.5-4 B). This graph plots the discovery rate of new genes (Y-axis) against the number of genomes sampled (X-axis) (Figure 5.5-4 B). As genome sampling increases, the discovery rate of new genes per genome tends to decrease. However, the data showed no evidence of convergence in discovering new genes, suggesting an 'open' pan-genome and implying ongoing species adaptation.

Although most genes reside in the cloud category, they do not belong to all strains by definition. By plotting the number of new genes, each genome contributed to the pan-genome. This analysis found that, on average, each genome contributed 67 unique genes. This value was comparable to the theoretical value of 66.557 calculated earlier (Figure 5.5-4 D). The unique barrier disruptive phenotype of AGR1487 may fall within these 67 unique genes. Unfortunately, the pan-genome software packages only captured a count of the genes and did not record which strains contributed each new gene to the pan-genome or the gene accession numbers. Manual scripts were written to data farm those gene identities.

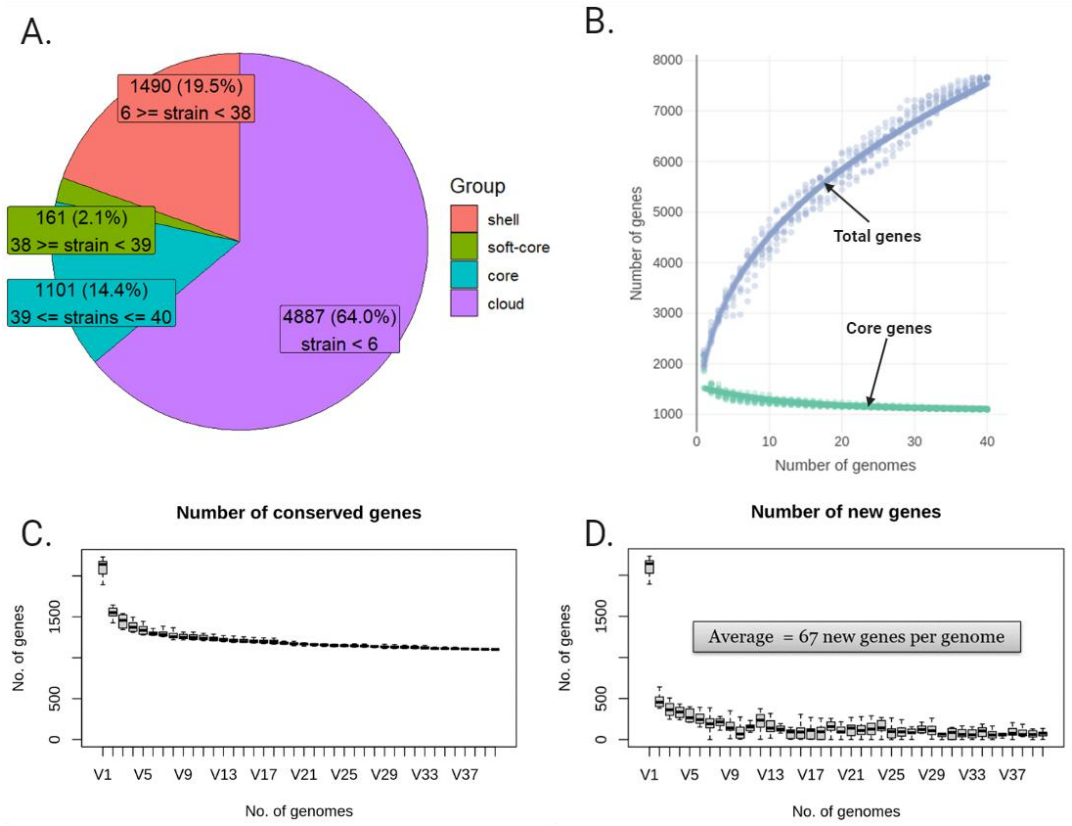

Figure 5.5-4: Roray Pan-genome analysis. A. Pan-genome open reading frames displayed as a pie chart. B. Refractory plot of total genes and core genes as genomes are randomly added with each iteration. C. After all interactions were completed, the number of conserved genes or core genes was extracted from the refractory plot dataset. D. The average number of unique genes detected as genomes are randomly added to the analysis per iteration. The pan-genome analyses were done with 1,000 interactions. The data reach convergence with the inclusion of 17 genomes.

5.5.4. AGR1487's Non-Identical or Unique Genes

Identifying non-identical genes involved comparing computed cloud or "unique" genes from PANX, BPGA, and Roary using a custom script. In this context, a gene is considered "unique" or "non-identical" if the BLASTP protein alignment does not produce a 100% identical match in any other *L. fermentum* strain within the pan-genome dataset; that is, the aligned region varies by one or more amino acids, or homologs in other strains are absent or truncated (Thesis specific definitions). The comparison revealed that AGR1487 contained 62 unique genes (Table 5.5-13), which agreed with the pan-genome data estimate of 67 genes. To test if these unique genes were unique or if homologous fragments were present in other strains, the analysis was repeated with a 95% similarity threshold to overcome the innate limitations of using a 100% identity cutoff threshold. At 100%, each alignment is sensitive to SNPs and minor assembly errors. However, this error can be alleviated by lowering the cutoff threshold at the cost of a change of higher false positive discoveries. The results from each attempt were stored in lists with names corresponding to the identity cutoff threshold used during the analysis. The two lists (95% and 100%) were compared, leaving 21 genes agreed between lists, and the lists generated more relaxed identity matches (95% identity match) (Table 5.5-14).

A 100% identity match in BLAST searches and amino acid alignments indicates perfect alignment with identical sequences, suggesting potential evolutionary conservation. A drawback of this method is that at 100%, the system is sensitive to errors such as SNPs, assembly errors or miss-called bases. This could be corrected in the 95% list, which was more tolerant of these minor errors. Notably, several genes exhibited less than 100% query coverage in BLASTP alignments, indicating they may represent gene fragments rather than full-length proteins. For example, three IS5 family transposases (RS02170, RS05375, and RS05945) showed approximately 48% coverage, suggesting these are remnants of mobile genetic elements that have undergone partial deletion or decay. Similarly, RS03620, encoding a putative ImmA/IrrE family metallo-endopeptidase, was identified as only 75 amino acids based on alignment to Roary pan-genome

representative sequences, substantially shorter than typical full-length members of this protein family.

To validate these findings and resolve discrepancies arising from the use of Roary representative sequences, full-length Prokka-annotated AGR1487 proteins were compared against Prokka-annotated proteins from 39 other *L. fermentum* strains using BLASTP. This validation revealed that RS03620 encodes a full-length protein of 249 amino acids with 98.8% identity to a homolog in strain YLF016, indicating that the original 75 amino acid length reflected alignment to a truncated Roary representative sequence rather than the true ORF boundary. The Prokka validation demonstrated that most cloud genes (59 of 62) had detectable homologs in at least one other strain, with identity values ranging from 42.7% to 100%. Six genes showed 100% identity to proteins in other strains (RS00910, RS01680, RS02660, RS03615, RS04890, RS06245), indicating that their classification as cloud genes reflected rarity within the pan-genome (presence in ≤15% of strains) rather than absolute sequence uniqueness. Only one gene, RS09860 (putative holin-like toxin), had no detectable homolog in any other strain and represents a truly unique AGR1487-specific gene.

Three genes were excluded from the final ranking analysis. RS02170 and RS05945 are identical IS5 family transposases (123 amino acids each) that were masked during MMseqs2 analysis due to their repetitive nature; their absence of Prokka BLAST hits reflects technical masking of low-complexity sequences rather than biological uniqueness.

The updated ranking (Table 5.5-15) incorporates a penalty/bonus factor based on Prokka validation results to down-weight genes with near-identical homologs in other strains and reward truly unique genes. Additional columns from the Prokka validation include: percent identity, query coverage, and full-length protein size in amino acids. These values were derived from BLASTP searches of full-length Prokka-annotated AGR1487 proteins against a combined database of Prokka-annotated proteins from 38 non-AGR1487 *L. fermentum* strains.

The location of each gene relative to the genome was not provided by the pan-genome software. The gene accession numbers were matched to the genes annotated on the genome. The genome gene model and gene annotation calculated by PGAP, Prokka, and Glimmer3 were mapped onto the AGR1487 genome atlas, revealing their location. The unique genes were found within regions of high variation. The genome was arbitrarily divided into five regions to aid the discussion (Figure 5.5-5).

Table 5.5-13: 62 AGR1487 specific cloud genes identified by the Roary pan-genome analysis with 100% sequence identity threshold.

ID	ANNOTATION	START	END	NT LENGTH	AA LENGTH	BEST HIT STRAIN	BEST HIT % ID	BEST HIT % COVERAGE	BEST HIT E VALUE	BEST HIT BIT SCORE
RS00155	ABC transporter permease	33309	34472	1164	388	YL-11	82.3	99.7	0.00E+00	513
RS00325	aspartate ammonia-lyase	78029	79408	1380	460	SRCM103285	96.4	100	6.91E-16	67.4
RS00335	diguanylate cyclase	79768	79980	213	71	No hit	No hit	No hit	No hit	No hit
RS00340	GNAT family N-acetyltransferase	80164	80748	585	195	No hit	No hit	No hit	No hit	No hit
RS00350	aminotransferase class I/II-fold pyridoxal phosphate-dependent enzyme	82665	83840	1176	392	No hit	No hit	No hit	No hit	No hit
RS00355	hypothetical protein	84100	84333	234	78	SRCM103285	99.3	100	0.00E+00	833
RS00580	DUF4422 domain-containing protein	133628	134314	687	229	YL-11	32.5	53.5	1.06E-15	76.3
RS00585	capsular polysaccharide synthesis family protein	134503	135306	804	268	SRCM103285	36.5	61.9	2.14E-28	112
RS00590	glycosyltransferase	135358	135633	276	92	No hit	No hit	No hit	No hit	No hit
RS00595	EpsG family protein	136344	137534	1191	397	MTCC 25067	95.9	100	4.54E-113	318
RS00600	serine acetyltransferase	137657	138151	495	165	L1	91.4	99.6	0.00E+00	872
RS00910	hypothetical protein	198689	199249	561	187	PMC101	96.3	100	1.58E-106	301
RS00925	hypothetical protein-flavodoxin	201038	201547	510	170	No hit	No hit	No hit	No hit	No hit
RS00930	hypothetical protein	201534	201890	357	119	PCC	92.7	87.4	6.62E-148	416
RS01080	hypothetical protein	231127	231687	561	187	B44	48.9	96.9	1.87E-93	278
RS01085	aldo/keto reductase	231926	232783	858	286	PCC	97.4	100	0.00E+00	603
RS01090	aldo/keto reductase	232797	233660	864	288	PCC	98.9	100	2.35E-128	358
RS01665	beta-galactosidase	351969	353981	2013	671	NCC2970	100	100	1.37E-68	200
RS01670	glycoside-pentoside-hexuronide (GPH):cation symporter	353974	355419	1446	482	YL-11	99.5	100	5.13E-149	411
RS01675	LacI family DNA-binding transcriptional regulator	355552	356574	1023	341	ike38	100	100	3.13E-57	170
RS01680	SDR family oxidoreductase	356885	357679	795	265	YL-11	98.6	100	2.25E-156	431
RS02170	IS5 family transposase	453952	454727	776	258	No hit	No hit	No hit	No hit	No hit
RS02660	hypothetical protein	548139	548309	171	57	VHProbi O48	99.7	100	0.00E+00	663
RS03575	Arm DNA-binding domain-containing protein	718176	718397	222	74	AGR1485	98.7	100	5.11E-50	160

RS03615	hypothetical protein (metal-sulfur cluster assembly factor)	723640	724083	444	148	No hit	No hit	No hit	No hit	No hit
RS03620	ImmA/IrrE family metallo-endopeptidase	724110	724334	225	75	SNUV175	99.2	100	7.24E-83	238
RS03625	hypothetical protein	724306	724605	300	100	JNU 532	97.9	100	0.00E+00	1052
RS03635	transposase	725263	725586	324	108	YLF016	100	100	6.12E-148	408
RS03640	helix-turn-helix transcriptional regulator	725742	725957	216	72	M4	94.6	100	9.22E-119	332
RS03645	phage antirepressor	725990	726778	789	263	M4	98	100	4.22E-109	308
RS03655	hypothetical protein	727859	728179	321	107	2760	97.1	100	0.00E+00	1175
RS03660	Inositolphosphorylceramide synthase	728323	728490	168	56	AGR1485	92.6	100	1.15E-82	238
RS03695	hypothetical protein (lactate utilisation protein C)	731789	732790	1002	334	AGR1485	98.7	100	0.00E+00	1099
RS03700	hypothetical protein (LutB/LldF family L-lactate oxidation iron-sulfur protein)	732832	733278	447	149	AGR1485	61.6	80.6	0.00E+00	723
RS03795	phage tail	751294	753579	2286	762	3872	98.7	100	1.02E-51	156
RS03820	hypothetical protein	754886	755383	498	166	AGR1485	65.8	100	2.16E-57	174
RS03860	IS5 family transposase	761687	762007	321	107	IMDO 130101	99	100	0.00E+00	579
RS03875	MucBP domain-containing protein	762847	765993	3147	1049	DM075	97.6	90.3	0.00E+00	548
RS03905	TIGR02328 family protein	769281	769643	363	121	IMDO 130101	99	100	0.00E+00	783
RS03925	site-specific integrase	772259	773065	807	269	SNUV175	100	100	3.77E-51	154
RS04875	Rgg/GadR/MutR family transcriptional regulator	969916	970833	918	306	YL-11	99.6	100	0.00E+00	521
RS04890	phage scaffolding protein	973696	973944	249	83	YL-11	99.3	100	8.49E-110	308
RS05065	hypothetical protein	1004935	1005072	138	46	PCC	92.2	74.7	0.00E+00	649
RS05330	type II toxin-antitoxin system RelE/ParE family toxin	1061383	1061667	285	95	SRCM 103290	98.6	100	4.95E-99	280
RS05370	methyltransferase domain-containing protein	1068394	1069104	711	237	YLF016	98.9	100	1.73E-61	181
RS05375	transposase	1069264	1069599	336	112	SRCM103285	100	100	0.00E+00	588
RS05405	hypothetical protein	1076705	1076884	180	60	MTCC 25067	100	100	2.24E-111	317
RS05410	SEC10/PgrA surface exclusion domain-containing protein	1076986	1077975	990	330	SCB0035	99.4	100	0.00E+00	685

RS05415	SEC10/PgrA surface exclusion domain-containing protein	1078030	1078374	345	115	M4	100	100	7.04E-82	234
RS05420	LPXTG cell wall anchor domain-containing protein	1079146	1079724	579	193	ike38	99.7	100	0.00E+00	804
RS05425	hypothetical protein	1079715	1080389	675	225	ike38	100	100	5.43E-128	357
RS05540	TetR family transcriptional regulator C-terminal domain-containing protein	1101609	1101818	210	70	AGR1485	99.6	100	0.00E+00	553
RS05945	IS5 family transposase	1181855	1182630	776	258	No hit	No hit	No hit	No hit	No hit
RS06245	hypothetical protein	1248542	1248784	243	81	JNU 532	100	100	0.00E+00	516
RS07340	IS200/IS605 family element transposase accessory protein TnpB	1446759	1447004	246	82	No hit	No hit	No hit	No hit	No hit
RS08060	DUF421 domain-containing protein	1572916	1573539	624	208	SRCM 103290	98.5	100	6.60E-43	132
RS08065	DUF3290 family protein	1573552	1573995	444	148	YLF016	100	100	8.42E-162	445
RS08410	hypothetical protein	1685595	1685798	204	68	ike38	98	100	1.88E-110	310
RS08610	glycerophosphodiester phosphodiesterase	1686096	1687742	1647	549	No hit	No hit	No hit	No hit	No hit
RS08770	HigA family addiction module antidote protein	1714184	1714540	357	119	PMC101	99.2	100	0.00E+00	537
RS09860	putative holin-like toxin	1500217	1500309	93	31	No hit	No hit	No hit	No hit	No hit
RS09865	Subtilisin-like protein hydrolase activity (serine)	1500407	1500604	198	66	MTCC 25067	97.1	100	5.59E-73	215

BLASTP STATISTICS SHOW PERCENT IDENTITY, PERCENT QUERY COVERAGE, AND BIT SCORES FOR THE BEST MATCH TO PROTEINS ACROSS 39 NON-AGR1487 *L. FERMENTUM STRAINS*. GENE COORDINATES AND LENGTHS ARE DERIVED FROM BLAST ALIGNMENTS AGAINST ROARY PAN-GENOME REPRESENTATIVE SEQUENCES AND MAY NOT REFLECT FULL OPEN READING FRAME BOUNDARIES. GENES WITH "NO HIT" HAD NO SIGNIFICANT MATCH IN OTHER STRAINS AT THE SEARCH THRESHOLD. VALIDATION USING FULL-LENGTH PROKKA-ANNOTATED PROTEINS (SEE SECTION 5.5.5) CONFIRMED GENE IDENTIFICATION BUT REVEALED DIFFERENT SEQUENCE STATISTICS DUE TO ORF BOUNDARY DIFFERENCES.

Table 5.5-14: 21 AGR1487 specific cloud genes identified by the Roary pan-genome analysis with 95% sequence identity threshold.

ID	ANNOTATION	START	END	NT LENGTH	AA LENGTH	BEST HIT STRAIN	BEST HIT % ID	BEST HIT % COVERAGE	BEST HIT E VALUE	BEST HIT BIT SCORE
RS00335	diguanylate cyclase	79768	79980	213	71	No hit	No hit	No hit	No hit	No hit
RS00340	GNAT family N-acetyltransferase	80164	80748	585	195	No hit	No hit	No hit	No hit	No hit
RS00350	aminotransferase class I/II-fold pyridoxal phosphate-dependent enzyme	82665	83840	1176	392	No hit	No hit	No hit	No hit	No hit
RS00355	hypothetical protein	84100	84333	234	78	SRCM103285	99.3	100	0.00E+00	833
RS00580	DUF4422 domain-containing protein	133628	134314	687	229	YL-11	32.5	53.5	1.06E-15	76.3
RS00585	capsular polysaccharide synthesis family protein	134503	135306	804	268	SRCM103285	36.5	61.9	2.14E-28	112
RS00590	glycosyltransferase	135358	135633	276	92	No hit	No hit	No hit	No hit	No hit
RS00595	EpsG family protein	136344	137534	1191	397	MTCC 25067	95.9	100	4.54E-113	318
RS00910	hypothetical protein	198689	199249	561	187	PMC101	96.3	100	1.58E-106	301
RS00930	hypothetical protein	201534	201890	357	119	PCC	92.7	87.4	6.62E-148	416
RS03615	hypothetical protein (metal-sulfur cluster assembly factor)	723640	724083	444	148	No hit	No hit	No hit	No hit	No hit
RS03695	hypothetical protein (lactate utilisation protein C)	731789	732790	1002	334	AGR1485	98.7	100	0.00E+00	1099
RS03700	hypothetical protein (LutB/LldF family L-lactate oxidation iron-sulfur protein)	732832	733278	447	149	AGR1485	61.6	80.6	0.00E+00	723
RS03795	phage tail	751294	753579	2286	762	3872	98.7	100	1.02E-51	156
RS05370	methyltransferase domain-containing protein	1068394	1069104	711	237	YLF016	98.9	100	1.73E-61	181

RS05415	SEC10/PgrA surface exclusion domain-containing protein	1078030	1078374	345	115	M4	100	100	7.04E-82	234
RS07340	IS200/IS605 family element transposase accessory protein TnpB	1446759	1447004	246	82	No hit	No hit	No hit	No hit	No hit
RS08060	DUF421 domain-containing protein	1572916	1573539	624	208	SRCM 103290	98.5	100	6.60E-43	132
RS08065	DUF3290 family protein	1573552	1573995	444	148	YLF016	100	100	8.42E-162	445
RS08410	hypothetical protein	1685595	1685798	204	68	ike38	98	100	1.88E-110	310
RS08610	glycerophosphodiester phosphodiesterase	1686096	1687742	1647	549	No hit	No hit	No hit	No hit	No hit

THIS SUBSET OF 21 GENES FROM TABLE 5.5-13 SHOWS BLASTP RESULTS AGAINST 39 OTHER *L. FERMENTUM* STRAINS. GENE BOUNDARIES ARE DERIVED FROM ALIGNMENTS TO ROARY REPRESENTATIVE SEQUENCES RATHER THAN COMPLETE ORF ANNOTATIONS. GENES WITH "NO HIT" WERE NOT DETECTED IN OTHER STRAINS, WHILE THOSE WITH LOW SIMILARITY SCORES REPRESENT REGIONS WHERE THE ROARY QUERY SEQUENCE DIVERGED FROM AGR1487. SUBSEQUENT VALIDATION USING FULL-LENGTH PROKKA ANNOTATIONS DEMONSTRATED THAT MOST GENES IN THIS SUBSET HAVE NEAR-IDENTICAL HOMOLOGS (96-100% IDENTITY) WHEN COMPLETE ORF SEQUENCES ARE COMPARED.

5.5.5. Candidate gene ranking

The gene ranking analysis compared protein sequences from *Limosilactobacillus fermentum* AGR1487 against both its own genome and 39 other *L. fermentum* genomes to test and rank the unique gene candidates. Each gene was assessed using quantitative parameters: best_self_bitscore, which is the highest BLAST alignment score for a gene matching itself; self_vs_nonself_ratio, representing how much more strongly a gene aligns within AGR1487 than with other strains; mean_self_bitscore, the average strength of all intra-strain matches; and n_self_hits, the number of homologous hits identified within the same genome. The bitscore is a normalised measure derived from BLAST that indicates the statistical strength of an alignment and the length of an alignment, with higher values corresponding to greater similarity and lower probabilities of random matches.

To further refine the ranking, RNA expression data from AGR1487 cultured in MRS media were incorporated into the scoring model. This adjustment rewarded genes that were transcriptionally active under growth conditions while penalising those with low or no expression, ensuring that the final ranking prioritised genes that are not only sequence-specific but also biologically expressed.

Several genes were among the highest-ranked based on strong self-conservation and limited similarity to other *L. fermentum* strains. The top five genes identified were RS08610 (glycerophosphodiester phosphodiesterase), RS03875 (MucBP domain-containing protein), RS01665 (beta-galactosidase), RS00585 (capsular polysaccharide synthesis protein), and RS03795 (S74 domain protein with hydrolase activity similar to *Lactobacillus* phage JNU P5 tail fibre). These genes demonstrate high within-strain similarity and divergence from related *L. fermentum* genomes.

Table 5.5-15: Candidate rankings

RSID	NEW RANK SCORE	BEST SELF BITSCORE	MEAN SELF BITSCORE	N SELF HITS	BEST NONSELF BITSCORE	SELF VS NONSELF RATIO	PROKKA PIDENT	PROKKA COVERAGE	PENALTY FACTOR	PROKKA AA LENGTH	ANNOTATION
RS08610	5752.1	1036.0	569.0	2.0	104.0	10.0	99.6	100.0	0.5	224.0	glycerophosphodiester phosphodiesterase
RS03875	4838.6	1918.0	1918.0	1.0	516.0	3.7	97.6	100.0	0.5	375.0	MucBP domain-containing protein
RS03795	3052.8	1425.0	1425.0	1.0	485.0	2.9	96.9	100.0	0.5	655.0	S74 domain; Lactobacillus phage JNU P5 tail fibre; hydrolase activity
RS01665	3021.9	1425.0	1425.0	1.0	461.0	3.1	99.4	100.0	0.5	315.0	beta-galactosidase
RS00585	3001.0	625.0	625.0	1.0	74.0	8.4	99.7	99.0	0.5	395.0	capsular polysaccharide synthesis protein
RS00325	2985.1	913.0	417.3	3.0	361.0	2.5	42.7	97.0	0.9	391.0	aspartate ammonia-lyase
RS00590	2268.8	657.0	349.5	2.0	117.0	5.6	98.2	89.0	0.5	62.0	glycosyltransferase
RS05415	1756.5	595.0	356.5	2.0	130.0	4.6	97.6	100.0	0.5	492.0	SEC10/PgrA surface exclusion domain-containing protein
RS00350	1479.1	807.0	258.2	9.0	321.0	2.5	98.3	100.0	0.5	298.0	aminotransferase class I/II-fold pyridoxal phosphate-dependent enzyme
RS05410	1427.0	595.0	332.5	2.0	166.0	3.6	98.3	100.0	0.5	229.0	SEC10/PgrA surface exclusion domain-containing protein

RS07340	1330.9	837.0	837.0	1.0			98.7	100.0	0.5	234.0	IS200/IS605 family accessory protein TnpB-related protein
RS01670	1243.3	945.0	627.5	2.0	612.0	1.5	99.6	100.0	0.5	243.0	glycoside-pentoside-hexuronide (GPH):cation symporter
RS01675	1175.8	687.0	414.0	2.0	295.0	2.3	99.5	100.0	0.5	412.0	LacI family DNA-binding transcriptional regulator
RS00595	1174.2	753.0	753.0	1.0			99.7	100.0	0.5	346.0	EpsG family protein
RS01085	1071.1	591.0	224.9	8.0	233.0	2.5	99.3	100.0	0.5	273.0	aldo/keto reductase
RS04875	1068.8	613.0	436.5	2.0	267.0	2.3	98.2	100.0	0.5	221.0	Rgg/GadR/MutR family transcriptional regulator
RS03695	1068.3	668.0	668.0	1.0			98.4	100.0	0.5	183.0	hypothetical protein
RS05370	970.9	493.0	240.7	3.0	172.0	2.9	99.7	100.0	0.5	360.0	class I SAM-dependent methyltransferase
RS00155	912.5	773.0	773.0	1.0	641.0	1.2	98.8	100.0	0.5	325.0	ABC transporter permease
RS00580	894.7	558.0	352.0	2.0	265.0	2.1	99.4	100.0	0.5	181.0	DUF4422 Family-glycosyltransferase
RS03925	887.7	531.0	379.5	2.0	261.0	2.0	96.1	100.0	0.5	388.0	site-specific integrase
RS00925	850.6	359.0	169.7	3.0	98.0	3.7	99.7	100.0	0.5	346.0	hypothetical protein-flavodoxin
RS05405	770.3	709.0	709.0	1.0	702.0	1.0	98.5	100.0	0.5	66.0	hypothetical protein
RS01080	748.7	376.0	175.3	3.0	134.0	2.8	98.3	100.0	0.5	242.0	hypothetical protein

RS08770	713.1	664.0	223.1	8.0	653.0	1.0	97.6	100.0	0.5	248.0	HigA family addiction module antidote protein
RS01090	673.4	589.0	230.9	8.0	488.0	1.2	99.1	100.0	0.5	451.0	aldo/keto reductase
RS00340	648.9	417.0	417.0	1.0			99.8	100.0	0.5	466.0	GNAT family N-acetyltransferase
RS03620	648.3	603.0	603.0	1.0	603.0	1.0	98.8	100.0	0.5	249.0	ImmA/IrrE family metallo-endopeptidase
RS03645	647.8	536.0	536.0	1.0	419.0	1.3	99.2	100.0	0.5	121.0	phage antirepressor
RS08060	639.0	409.0	409.0	1.0			99.6	100.0	0.5	269.0	DUF421 domain-containing protein
RS00910	615.4	397.0	397.0	1.0			100.0	100.0	0.5	215.0	hypothetical protein
RS05425	595.6	376.0	376.0	1.0			98.9	100.0	0.5	443.0	hypothetical protein
RS01680	562.6	529.0	180.2	6.0	484.0	1.1	100.0	100.0	0.5	148.0	SDR family oxidoreductase
RS03820	545.9	357.0	357.0	1.0	195.0	1.8	97.8	100.0	0.5	183.0	hypothetical protein
RS05420	523.9	334.0	334.0	1.0			99.4	100.0	0.5	505.0	LPXTG cell wall anchor domain-containing protein
RS08065	465.2	292.0	292.0	1.0			98.6	100.0	0.5	295.0	DUF3290 family protein
RS03625	463.2	440.0	282.0	2.0	440.0	1.0	99.0	100.0	0.5	195.0	hypothetical protein
RS03615	458.8	296.0	296.0	1.0			100.0	100.0	0.5	120.0	hypothetical protein
RS03700	441.3	277.0	277.0	1.0			98.6	100.0	0.5	357.0	hypothetical protein
RS03905	414.0	265.0	265.0	1.0			99.6	100.0	0.5	246.0	TIGR02328 family protein
RS00930	402.2	252.0	252.0	1.0			98.5	100.0	0.5	841.0	hypothetical protein

RS08410	395.4	297.0	297.0	1.0	194.0	1.5	99.4	100.0	0.5	171.0	hypothetical protein alpha/beta hydrolase
RS05375	370.2	344.0	253.3	13.0	344.0	1.0	98.1	100.0	0.5	466.0	transposase
RS00600	357.6	332.0	332.0	1.0	324.0	1.0	99.3	100.0	0.5	553.0	serine acetyltransferase
RS03635	356.2	227.0	227.0	1.0	119.0	1.9	97.8	100.0	0.5	321.0	transposase
RS03860	332.0	229.0	109.0	3.0	126.0	1.8	99.4	100.0	0.5	162.0	IS5 family transposase
RS03575	322.2	143.0	143.0	1.0	46.0	3.1	96.4	100.0	0.5	1028.0	Arm DNA-binding domain-containing protein
RS09860	311.5	64.0	64.0	1.0	48.0	1.3	No hit	-	2.0		putative holin-like toxin
RS03655	280.4	236.0	236.0	1.0	187.0	1.3	99.7	100.0	0.5	386.0	hypothetical protein
RS00355	263.9	165.0	165.0	1.0			98.4	100.0	0.5	314.0	hypothetical protein
RS00335	221.6	141.0	141.0	1.0			99.3	100.0	0.5	406.0	diguanylate cyclase
RS03640	219.3	139.0	139.0	1.0			99.1	100.0	0.5	215.0	helix-turn-helix transcriptional regulator
RS05540	212.2	199.0	199.0	1.0	196.0	1.0	99.6	100.0	0.5	271.0	TetR family transcriptional regulator C-terminal domain-containing protein
RS04890	195.5	155.0	155.0	1.0	109.0	1.4	100.0	100.0	0.5	148.0	phage scaffolding protein
RS02660	173.6	112.0	112.0	1.0			100.0	100.0	0.5	247.0	hypothetical protein
RS05330	169.3	156.0	156.0	1.0	156.0	1.0	98.3	100.0	0.5	235.0	type II toxin-antitoxin system RelE/ParE family toxin
RS06245	154.3	144.0	144.0	1.0	138.0	1.0	100.0	100.0	0.5	123.0	hypothetical protein

RS03660	128.6	107.0	107.0	1.0	87.0	1.2	98.4	74.0	0.5	250.0	Inositolphosphorylceramide synthase
RS05065	123.3	91.0	91.0	1.0	59.0	1.5	98.7	100.0	0.5	386.0	hypothetical protein
RS09865	82.3	64.0	64.0	1.0	48.0	1.3	97.1	65.0	0.5	267.0	Subtilisin-like protein hydrolase (serine)

CLOUD GENES WERE RANKED USING A COMPOSITE SCORE INTEGRATING MMSEQS2 ALIGNMENT METRICS AND A PENALTY FACTOR DERIVED FROM FULL-LENGTH PROKKA PROTEIN BLAST VALIDATION AGAINST 39 *L. FERMENTUM* STRAINS.

COLUMN DESCRIPTIONS: RSID, AGR1487 GENE IDENTIFIER; NEW RANK SCORE, COMPOSITE RANKING SCORE AFTER APPLYING THE PROKKA-BASED PENALTY FACTOR; BEST SELF BITSCORE, HIGHEST MMSEQS2 BITSCORE WHEN ALIGNED AGAINST AGR1487 PROTEINS (SELF-ALIGNMENT); MEAN SELF BITSCORE, AVERAGE BITSCORE ACROSS ALL SELF-HITS; N SELF HITS, NUMBER OF SELF-HITS INDICATING POTENTIAL PARALOGS; BEST NONSELF BITSCORE, HIGHEST BITSCORE AGAINST ANY NON-AGR1487 STRAIN; SELF VS NONSELF RATIO, RATIO OF SELF TO NON-SELF BITSCORE WHERE HIGHER VALUES INDICATE GREATER STRAIN SPECIFICITY; PROKKA PIDENT, PERCENT IDENTITY OF FULL-LENGTH PROKKA-ANNOTATED PROTEIN TO BEST HIT IN OTHER STRAINS; PROKKA COVERAGE, PERCENT QUERY COVERAGE OF THE PROKKA BLAST ALIGNMENT; PENALTY FACTOR, MULTIPLIER APPLIED TO ORIGINAL RANKING SCORE BASED ON PROKKA IDENTITY (VALUES <1.0 PENALISE GENES WITH NEAR-IDENTICAL HOMOLOGS; 2.0 REWARDS TRULY UNIQUE GENES); PROKKA AA LENGTH, FULL-LENGTH PROTEIN SIZE FROM PROKKA ANNOTATION; ANNOTATION, FUNCTIONAL ANNOTATION.

TWO IS5 FAMILY TRANSPOSASES (RS02170, RS05945) WERE EXCLUDED AS THESE IDENTICAL SEQUENCES WERE MASKED DURING MMSEQS2 ANALYSIS DUE TO THEIR REPETITIVE NATURE, AND THEIR ABSENCE OF PROKKA BLAST HITS REFLECTS TECHNICAL MASKING RATHER THAN BIOLOGICAL UNIQUENESS.

5.5.6. Genome atlas with all results

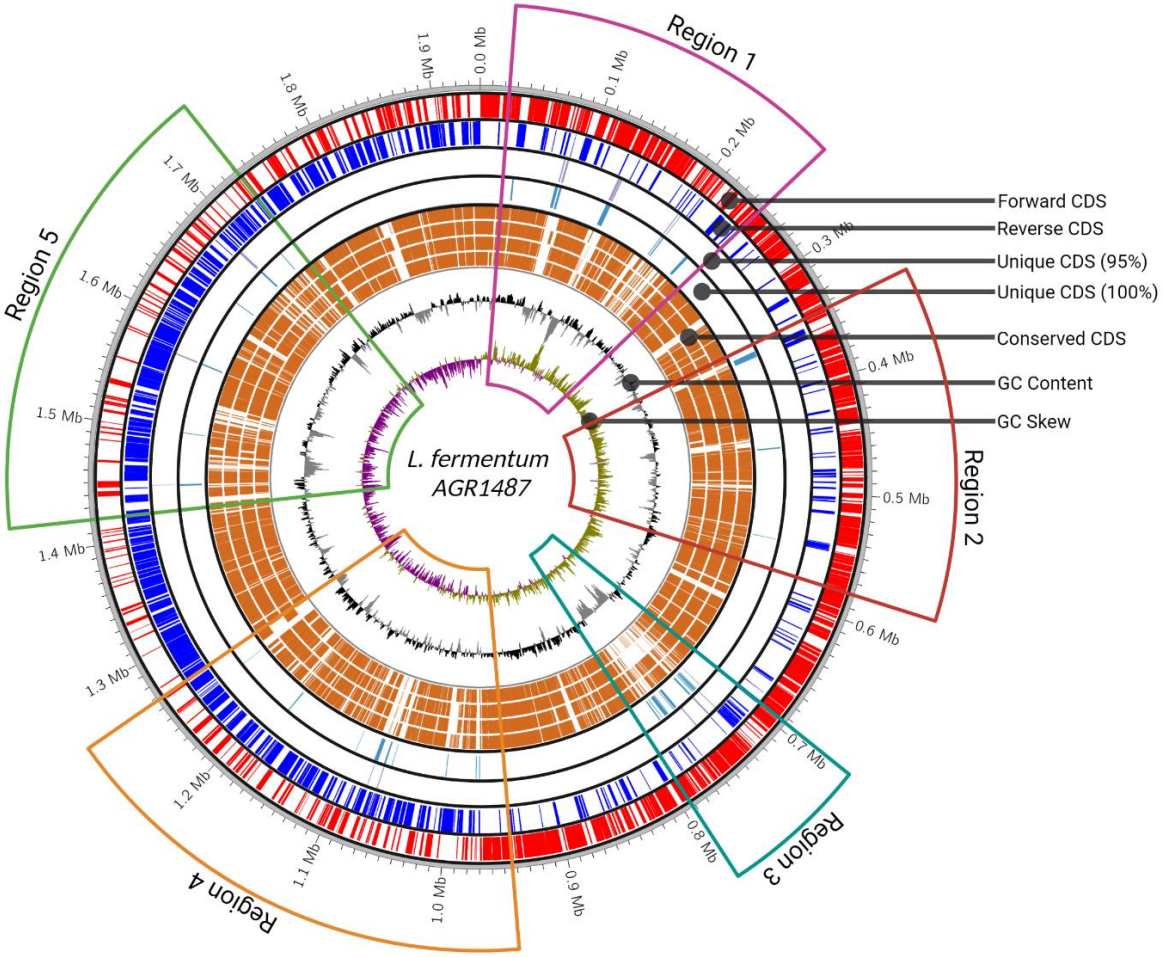

Figure 5.5-5: Genome atlas of AGR1487. The genome atlas of AGR1487 over layed with additional detail. Outter ring to inner ring Forward CDS, Reverse CDS, Unque genes from the 95% list, Unique genes from the 100% list, Conserved CDS alignments calculated with Mmseq2 reciprocal best hits approach, GC content, GC skew.

5.6. Discussion

The bioinformatics analyses were designed to identify strain-specific, unique gene (see Thesis specific definitions) candidates by establishing a genetic commonality baseline. In genomics, conserved genes serve as markers of genetic stability, offering a comparative framework against which deviations are measured. With an adequate baseline, isolating and identifying the genetic variations that may have given rise to the barrier-disruptive phenotype in AGR1487 became possible. This approach facilitated the identification of potential candidate genes and enhanced our understanding of genetic diversity and evolution within this microbial species.

An initial assessment of genomic similarity within *L. fermentum* strains was conducted using ANI, which calculates the genetic nucleotide identity between strains. The ANI results indicated a high degree of similarity among the *L. fermentum* strains, ranging from 96.9% to 99.9%, with a notable absence of similarity to the *L. plantarum* control. The *L. plantarum* strain was added to test where the ANI could detect species' differences, which was successful. However, ANI did not find any homology between the strains because it cannot accurately compare genomes when the datasets fall below roughly 82% similarity; this is a known system flaw (Shaw & Yu, 2023R). In addition, the ANI's inability to capture structural genomic nuances due to its aggregation of sequence data into single numerical values highlighted the need for a more detailed analysis. Some strains were known to contain plasmids before ANI, and those could not be distinguished from genomic differences between the strains. A Mash analysis was done to find support for the ANI results while retaining some of the contig information within the files.

Mash distances were calculated for a more granular comparison. Mash was more computationally expensive than ANI, but Mash distance is a probabilistic estimate and, as a result, also provides a p-value for each comparison. In addition, the Mash comparison is done by sequence in the file rather than by file, meaning that each contig id within a FASTA file remains in the analysis. Since each genome should be complete, only one sequence should be in each

FASTA file. Therefore, any deviation from this rule would immediately indicate a contaminated assembly, plasmid or phage sequence associated with the bacterial strain.

The Mash supported the ANI results in that the genomes *L. fermentum* genomes had a high degree of similarity but also outlined the outliers in the dataset that contained additional sequences beyond their core genome sequence. Ten outlier sequences were in the *L. fermentum* dataset, and none had sequence homology to any other *L. fermentum* strains. However, these sequences did lower the total score of the *L. fermentum* strain they were associated with. A closer examination revealed that the unique sequences were plasmids. These plasmids could be ignored because neither AGR1485 nor AGR1487 contained plasmids, and the Mash already established that these sequences were unique and limited to the strains carrying them. However, how much the plasmids and structures like these contributed to the genetic variation between strains remained unanswered. Removing the ten plasmids from the analysis improved the Mash scores for the harbouring strains from roughly 0.24 to align with the other strains to fall within the same range of 0.012-0.00019. The improved scores without plasmids indicate that much of the genetic diversity of *L. fermentum* was linked to extra-chromosomal structures. These plasmids provided nearly 20% more dissimilarity than the genomes directly (roughly 3.1%). Since neither AGR1485 nor AGR1487 contains plasmids, identifying AGR1487's barrier-disruptive phenotype would benefit from removing plasmids from the analysis, but a plasmid or phage could have integrated itself into the genome. To prevent the loss of information, the strains and plasmids were left intact for downstream experiments. These experiments further support that the gene candidates for the barrier disruptive phenotype expressed by AGR1487 in a TEER assay are limited to its genome. The ANI and the Mash distances support the hypothesis that the genomes are closely related, with an average difference of around 3.1% between the genomes. Within this 3.1%, the gene candidates responsible for the barrier disruptive phenotype will likely be found.

Despite the close genetic relationships revealed by the ANI and Mash distances, these methods do not provide insight into the structural variations within the genomes. Is all the variation limited to a single region within the genome or spread throughout each strain? Are all

the genes present in various decay states? A comparative and synteny analysis of the coding sequences (CDS) regions was conducted using Mmseqs2 to answer these questions. Mmseqs2 is a fast and sensitive software suite for protein sequence analysis, capable of identifying orthologs between species based on the highest-scoring protein sequence alignments. The Mmseq2 alignments highlighted a high degree of conservation in identity match, and CDS order yet revealed significant genomic variations. For instance, *L. fermentum* AGR1487 has a genome size of 1,939,032-bp, which is 287,830-bps smaller than the strain AGR1485 (2,226,862-bp). The 287,830 bps of DNA missing from AGR1487 was not located in a single deletion but was spread through the genome. There was also a notable inversion and sequence translocation. The inverted locus was also present in half the strains in the dataset. The inverted locus did not produce the same alignment score between strains, nor was it the same size, underscoring the complex nature of genetic evolution within *L. fermentum*.

To better visualise these genomic rearrangements, LCBs for the genome were calculated by ProgressiveMauve, enabling the identification and illustration of homologous regions and rearrangements such as inversions and translocations. This analysis provided a clearer picture of the genomic landscape. The most prominent genetic event distinguishing the AGR1485 and AGR1487 is a single inversion that also translocates from a nucleotide position of approximately 1,550,000 in the AGR1485 genome to around 750,000 in the AGR1487 genome. These distinct inversion and translocation are indicators of horizontal gene transfer (HGT). To further support that this was a HGT event, the LCBs for this region were compared to the other strains. The translocation position was inconsistent between genomes, and the LCB was absent from the B44 genome, a sign that this LCB is not a universal feature of this species of bacteria in support of previous BLAST analyses. The LCB could have eroded in B44, a naturally occurring event commonly known as reductive evolution, where unconserved genes are lost gradually over time. Given the varying MMseq2 scores and the change in locus alignment sizes between strains, this locus showed signs of reductive evolution.

Nevertheless, the order and orientation of all other LCBs remain consistent across genomes, except for their sizes, which did vary, indicating gene loss and mutations. A genome-wide association study could not be carried out because the gene markers could not be confirmed in other strains. This situation could be addressed in a traditional genome-wide association study by exploring inheritance patterns. However, this was impossible because bacteria typically undergo binary fusion as a primary replication method. A process that will ensure that all offspring will contain the marker, providing no deductive advantage. Thus, no common markers for the phenotype would be forthcoming. However, comparative genomics techniques standard to pan-genome analyses could be used to find unique genes for each strain. There were a few limitations to a pan-genome approach. Most pan-genome software packages did not report strain-level information, which was unsuitable for work to identify unique genes within a single strain.

Multiple analytical packages revealed that each strain contained an average of 2,147 ORFs. Since the genomes differed by around 3.1%, each genome was theoretically expected to hold approximately 66.557 genes. A pan-genome analysis was used to explore the genes further.

The pan-genome encompassed 1,101 core genes present in all 40 strains, along with the soft-core genes (38 or more strains), which could be excluded from further analyses. It is unlikely that the phenotype is common enough to be found in six or more strains, given that there were no other recorded examples of inflammation or barrier disruption from this well-studied food-borne microbe at the time of writing. However, it remains a possibility; therefore, the 1,490 shell genes could also be tentatively ignored from the initial experiments but could be explored as part of future work.

Roary found that each genome contains an average of 67 non-identical or “unique” genes (Figure 5.5-4 D), comparable to the theoretical value of 66.557 calculated earlier. These unique genes, defined as those absent or lacking 100% identical matches in other strains (xxii), represent the best opportunity to find the gene(s) that gave rise to the barrier-disruptive phenotype of AGR1487. A custom BLAST filtering process was implemented to find the unique genes specific

to AGR1487 within the remaining 4,887 genes representing the cloud genes. The resulting outputs contained the 62 genes unique to AGR1487.

The pan-genome analyses were run at 100% and 95% alignment match cutoff thresholds. A drawback of the 100% identity match was that the outputs are sensitive to assembly errors, as any assembly error resulting in changes as small as a single SNP could be detected as a unique gene, thereby increasing the rate of false discoveries. Alternatively, reducing the alignment cutoff threshold reduces the false discovery rate but increases the number of gene candidates. It is also a point of consideration that the gene responsible for any unique phenotype could exist as an inactive mutant in other strains.

The data was tested by comparing the 95% identity match list with the 100% identity match list for agreement to produce a more lenient but limited list of genes. This approach found that AGR1487 has 21 unique genes (regardless of the cutoff value) that were dispersed throughout the genome. Despite this, none of the unique genes were excluded from downstream analyses. The 41 genes that did not agree between lists represent those that could have undergone reductive evolution or were detected as unique due to an assembly error. The genes were checked by alignment using the PANX HTML. The alignments showed that many of the genes selected were longer in AGR1487 than other strains. This indicated that the mutations may have arisen in different strains and potentially not AGR1487. For instance, the inversion previously identified during the genome alignments contains many of AGR1487's unique genes. This locus had fewer genes between the palindromic repeats in the other 16 strains of *L. fermentum*, where it can be found, and many of those genes were also less complete. The sequence alignments for the remaining genes also had decreasing alignment matches, many with apparent deletions in the genes, indicating reductive evolution at this locus. This pattern of gene fragmentation was also evident in the BLASTP coverage statistics (Table 5.5-13 and Table 5.5-14), where several genes, particularly transposases (RS02170, RS05375, RS05945), exhibited only ~48% query coverage, consistent with partial gene loss or pseudogenisation in other strains. These fragmented transposases likely represent decaying mobile genetic elements, a hallmark of ongoing genome

reduction in bacterial lineages. Due to the reductive evolution observed in the locus above, the phenotype may still reside within the 41 genes that did not agree between lists when changing the cutoff threshold. Still, the list of 21 genes that agreed between the lists represents the unique genes in AGR1487. However, their locations relative to all the other data (such as the MMseqs2 alignments) could provide further information regarding which parts of the genome could be responsible for the barrier disruptive-phenotype.

The gene candidates were further tested to quantify their sequence-level uniqueness and evaluate their specificity across all 40 *L. fermentum* strains. Using a custom MMseqs2- and BLAST-based ranking framework, each gene was assigned a composite score reflecting both its conservation within AGR1487 and its divergence from homologous sequences in other strains. The ranking was further weighted by whether the genes were expressed by AGR1487 when exposed to MRS media. To validate these results and address potential discrepancies from the use of Roary pan-genome representative sequences, full-length Prokka-annotated proteins were compared against 39 other *L. fermentum* (AGR1487 excluded) strains using BLASTP. This validation revealed that most cloud genes had near-identical homologs (96-100% identity) in at least one other strain when complete ORF sequences were compared, indicating that their classification as cloud genes reflects rarity within the pan-genome rather than absolute sequence uniqueness. The original ranking scores were subsequently modified by a penalty/bonus factor based on Prokka BLAST identity: genes with high identity to other strains were penalised, while the single truly unique gene (RS09860) received a bonus. This quantitative approach complements the qualitative alignment results by identifying which genes are truly AGR1487-specific rather than artifacts of assembly variation, truncation, or pan-genome rarity. Genes with the highest rank scores represent the most distinct loci in the AGR1487 genome and may therefore underpin the barrier-disruptive phenotype observed in this strain.

To visualise the regions where other AGR1487 unique genes were found relative to other *L. fermentum* strains, the Mmseq2 results were aligned to the AGR1487 genome atlas from Chapter 4, along with the unique genes found during the pan-genome analysis. The AGR1487

genome atlas was arbitrarily divided into five regions, 1 to 5, to aid the discussion below. Because the phenotype is associated with the cell surface, cytosolic proteins such as transcriptional regulators could be excluded from the analysis. Each numbered area was searched for signal peptides (a protein transportation sequence) and motifs (transmembrane domains) that would indicate whether or not a gene would be sent to the cell surface.

Region 1 stood out as it contained the highest number of unique genes, totalling 18 of the 62 genes. These genes cover various functions, including ABC transporters, serine acetyltransferase, reductases, and hypothetical proteins. Of the 18 genes, four are hypothetical proteins, and none were predicted to have cell surface signalling peptides, transmembrane helices, cell surface domains, or folds. Notably, one of the hypothetical proteins exhibits a significant similarity (greater than 60%) to a known flavodoxin sequence, which was manually added to the gene's annotation. The remaining genes with signal peptides identified in Region 1 were exclusively linked to sugar transporters and synthases, which are less likely to act on tight junctions but are known to act on small molecules in the supernatant. Neither feature was characteristic of this barrier-disruptive phenotype.

Region 2 comprised a beta-galactosidase, a Glycoside-pentoside-hexuronide symporter, a ribosomal protein S18-alanine N-acetyltransferase, and a couple of transcriptional regulators. Only the beta-galactosidase and the Glycoside-pentoside-hexuronide symporter were predicted to have membrane-spanning protein structures and signal peptides. Given that the phenotype did not respond to the supernatant, it is unlikely that these transporters and their substrates are involved in the mechanism.

Region 3 exhibited a more compact cluster of unique genes than other regions, as all the unique genes in this region are part of a domesticated prophage previously identified as an inversion. Seventeen genes in this region are exclusive to AGR1487 and located within palindromic repeats. Interestingly, not all phage genes are unique, and many can be found in other strains. However, AGR1487 had the most complete phage (with the most genes) among all the

strains analysed (Figure 5.8-1). The phage harboured phage tail genes and three hypothetical proteins. NCBI BLAST and the GAMOLA2 annotations from Chapter 4 predicted the hypothetical proteins were a metal-sulphur cluster assembly factor, lactate utilisation protein C, and a LutB/LldF family L-lactate oxidation iron-sulphur protein. None of these genes contained signal peptides.

The phage also contained the gene RS03620, an ImmA/IrrE family metallo-endopeptidase with a signal peptide. Roary pan-genome analysis initially identified RS03620 as a truncated 75 amino acid fragment (225 nucleotides); however, full-length Prokka annotation revealed it encodes a complete 249 amino acid protein, consistent with typical ImmA/IrrE family proteins that span 200–300 amino acids**.** RS03620-1 is a phage anti-repressor targeted towards inactivating ImmR in *Bacillus subtilis* through proteolysis (Bose, Auchtung, Lee, & Grossman, 2008). This metallo-endopeptidase is also present in *L. fermentum* strains 2760 and L1 but is incomplete (99.26% similar to AGR1487's). Homologous sequences from the gene were found in 13 other strains, with BLAST alignment scores ranging from 93.38 to 30.74 for all other strains. This gene has a signal peptide sequence and appears to be the most promising candidate from this region. Whether this truncated metallopeptidase retains enzymatic activity or represents a pseudogene requires functional validation; however, the presence of a signal peptide suggests it may still be secreted and could potentially interact with host cell surface proteins. However, it remains unknown if it can target tight junctions, but it is known to cleave proteins.

Region 4 ranged from 969,916-bp to 973,944-bp and comprised five highly varied areas encompassing phage proteins. The region from 1,004,935-pb to 1,080,389-bp contains a series of antitoxin genes, transposases, and a gene with an LPXTG cell wall anchor domain (Siegel, Reardon, & Ton-That, 2017). While some of these genes have signal peptides and transmembrane domains, they act on small molecules or target the cell's DNA. Thus, they are less likely to cause the barrier-disruptive phenotype than the genes in Region 3.

Region 5 encompassed eight unique genes, including RS07340, an IS200/IS605 family accessory protein TnpB-related protein; RS08060, a DUF421 domain-containing protein; RS08065, a DUF3290 family protein; RS08410, a hypothetical protein with a predicted alpha/beta hydrolase; RS08610, a glycerophosphodiester phosphodiesterase; RS08770, a HigA family addiction module antidote protein; RS09860, a putative holin-like toxin; and RS09865, a subtilisin-like protein with hydrolase activity targeted at serine; and CP047585.1, an IS5 family transposase. Transmembrane helices, domains, or signalling peptides were not detected within this region due to incomplete annotations or annotations that did not span the entire reading frame; these genes cannot be ruled out. Some of these genes have predicted functions that could give rise to the barrier-disruptive phenotype.

Overall, the unique genes were all observed to align with AT-rich regions of the genome in high variation zones, which often had genes associated with phage, palindromic repeats, and transposases. This observation indicates potential occurrences of horizontal gene transfer events. The presence of fragmented transposases and genes initially identified as fragments by Roary (such as RS03620, later confirmed as full-length) within these variable regions is consistent with the dynamic nature of these loci, where mobile genetic elements facilitate gene acquisition, duplication, and subsequent decay through pseudogenisation. These events were not localised, suggesting they did not happen simultaneously but gradually over time.

5.7. Conclusion

The genomes of *L. fermentum* were highly conserved in gene identity and gene order. The genomes had a similarity of approximately 97%. The 3.1% difference between the genomes amounted to around 67 unique genes per genome. AGR1487 contained 62 unique genes that were not localised into a single region but dispersed throughout the genome. All the unique genes were limited to areas with a high AT content surrounded by palindromic repeats. Many of these also

contained fragments of phages or transposons, suggesting that the primary source of divergence between genomes of *L. fermentum* arose from horizontal gene transfer events.

The gene candidates contributing to this barrier-disruptive phenotype most likely fall within these regions of high variation. Two of these regions contained genes that have the potential to cause the barrier-disruptive phenotype and are part of domesticated phages. Two of these candidates are enzymes, a metalloendopeptidase and a hydrolase, which are rare in other *L. fermentum* strains and may represent the primary focus for future research.

Integrating genomic data with other omics approaches like transcriptomics during the TEER assay may offer a more comprehensive perspective on the molecular mechanisms behind this disruptive phenotype and facilitate the discovery of biomarkers. The next chapter focuses on which high-variation regions are turned on during the barrier-disruptive phenotype, and the identified unique genes were tested for expression during the transcriptomics assays carried out in Chapter 6.

5.8. Appendix

5.8.1. SyntTax

Prokaryotic Synteny & Taxonomy Explorer (Figure 5.8-1) is a web application for bacterial comparative genomics https://archaea.i2bc.paris-saclay.fr/SyntTax/. The site links NCBI's database to BLAST search engines, allowing the user to define the species or strains for analysis and then provide a protein sequence to BLAST against all the strains. The output is a local scan of the region where the gene is found and local annotations of all the other genes around the locus (Figure 5.8-1). Each positive alignment is assigned a colour that represents and is shown per genome. The results are for the metallopeptidase within the suspected domesticated phage found in variable region 3 of the AGR1487 genome. The phage is found in several strains, but its most complete form is only in AGR1487. The collapsed phage indicates that this region has undergone reductive evolution in other strains. The metallopeptidase in the centre of this page module is a good example. The gene in question is found in five strains but with reduced identity alignments with AGR1487, where it is at its longest.

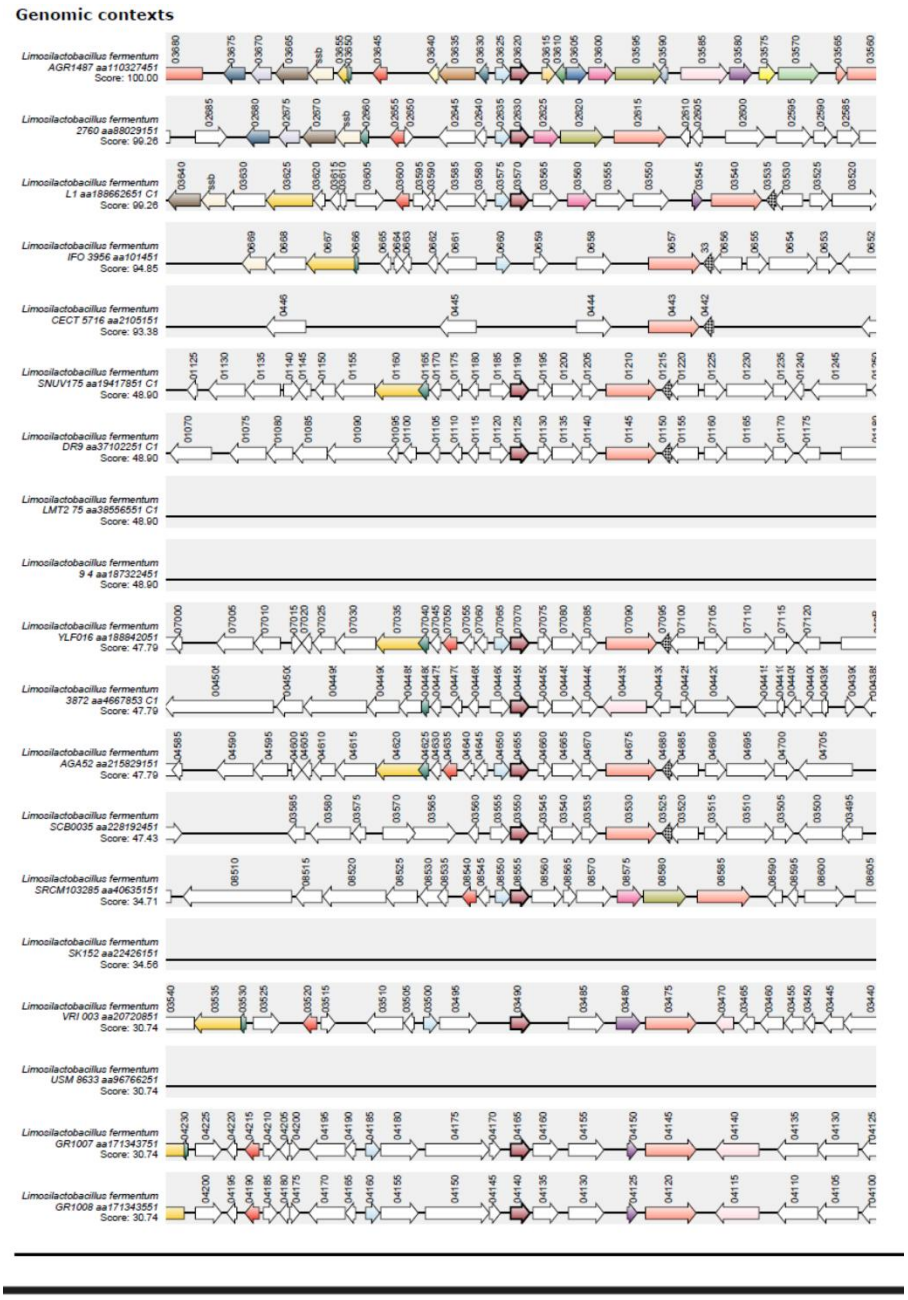

Figure 5.8-1: Syntax map. The domesticated phage was found in AGR1487's 3rd region and contained the closest clustering of unique genes out of all the regions. The centred object in red with a black border is the aligned metallopeptidase protein sequence extracted from AGR1487 and aligned against all *L. fermentum* records present on NCBI at the time of writing. Successful alignments are coloured the same between genomes. The left-hand side of the image contains the record name and the score match of the alignment. The phage is present in several other strains but has lost genes.

5.9. References

Bailie, M. A., Altermann, E., Young, W., Roy, N. C., & McNabb, W. C. (2020). Complete genome sequence of *Lactobacillus fermentum* strain AGR1485, a human oral isolate. *Microbiology Resource Announcements, 9*(36), e00841-00820. https://doi.org/10.1128/MRA.00841-20

Bailie, M. A., Altermann, E., Young, W., Roy, N. C., & McNabb, W. C. (2021). Complete annotated genome sequence of *Limosilactobacillus fermentum* AGR1487. *Microbiology Resource Announcements, 10*(1). https://doi.org/10.1128/mra.01056-20

Bose, B., Auchtung, J. M., Lee, C. A., & Grossman, A. D. (2008). A conserved anti-repressor controls horizontal gene transfer by proteolysis. *Molecular Microbiology, 70*(3), 570–582. https://doi.org/10.1111/j.1365-2958.2008.06414.x

Ceapa, C., Davids, M., Ritari, J., Lambert, J., Wels, M., Douillard, F. P., . . . Kleerebezem, M. (2016). The variable regions of *Lactobacillus rhamnosus* genomes reveal the dynamic evolution of metabolic and host-adaptation repertoires. *Genome Biology and Evolution, 8*(6), 1889–1905. https://doi.org/10.1093/gbe/evw123

Chaudhari, N. M., Gupta, V. K., & Dutta, C. (2016). BPGA—an ultra-fast pan-genome analysis pipeline. *Scientific Reports, 6*(1), 24373. https://doi.org/10.1038/srep24373

Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: Multiple genome alignment with gene gain, loss, and rearrangement. *PLoS ONE, 5*(6), e11147. https://doi.org/10.1371/journal.pone.0011147

Ding, W., Baumdicker, F., & Neher, R. A. (2017). panX: Pan-genome analysis and exploration. *Nucleic Acids Research, 46*(1), e5. https://doi.org/10.1093/nar/gkx977

Douillard, F. P., Ribbera, A., Jarvinen, H. M., Kant, R., Pietila, T. E., Randazzo, C., . . . de Vos, W. M. (2013). Comparative genomic and functional analysis of *Lactobacillus casei* and *Lactobacillus rhamnosus* strains marketed as probiotics. *Applied and Environmental Microbiology, 79*(6), 1923–1933. https://doi.org/10.1128/AEM.03467-12

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics, 26*(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Ferrés, I., & Iraola, G. (2021). An object-oriented framework for evolutionary pangenome analysis. *Cell Reports Methods, 1*(5), 100085. https://doi.org/10.1016/j.crmeth.2021.100085

Mardis, E., McPherson, J., Martienssen, R., Wilson, R. K., & McCombie, W. R. (2002). What is finished, and why does it matter. *Genome Research, 12*(5), 669–671.

Naghmouchi, K., Belguesmia, Y., Bendali, F., Spano, G., Seal, B. S., & Drider, D. (2020). *Lactobacillus fermentum*: A bacterial species with potential for food preservation and biomedical

applications. *Critical Reviews in Food Science and Nutrition, 60*(20), 3387–3399. https://doi.org/10.1080/10408398.2019.1688250

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology, 17*(1), 132. https://doi.org/10.1186/s13059-016-0997-x

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., . . . Parkhill, J. (2015). Roary: Rapid large-scale prokaryote pan-genome analysis. *Bioinformatics, 31*(22), 3691–3693. https://doi.org/10.1093/bioinformatics/btv421

Sengupta, R., Anderson, R. C., Altermann, E., McNabb, W. C., Ganesh, S., Armstrong, K. M., . . . Roy, N. C. (2015). *Lactobacillus fermentum* AGR1487 cell surface structures and supernatant increase paracellular permeability through different pathways. *MicrobiologyOpen, 4*(4), 541–552. https://doi.org/10.1002/mbo3.260

Shaw, J., & Yu, Y. W. (2023). Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nature Methods, 20*(11), 1661–1665. https://doi.org/10.1038/s41592-023-02018-3

Shimoyama, Y. (2022a). *ANIclustermap: A tool for drawing ANI clustermap between all-vs-all microbial genomes.* Retrieved from https://github.com/moshi4/ANIclustermap

Shimoyama, Y. (2022b). *pyCirclize: Circular visualization in Python.* Retrieved from https://github.com/moshi4/pyCirclize

Shimoyama, Y. (2022-06-20). *pyGenomeViz: A genome visualization python package for comparative genomics.* Retrieved from https://github.com/moshi4/pyGenomeViz

Siegel, S. D., Reardon, M. E., & Ton-That, H. (2017). Anchoring of LPXTG-Like Proteins to the Gram-Positive Cell Wall Envelope. *Current Topics in Microbiology and Immunology, 404,* 159-175. https://doi.org/10.1007/82_2016_8

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology, 35(11),* 1026-1028. https://doi.org/10.1038/nbt.3988

Sun, Z., Harris, H. M. B., McCann, A., Guo, C., Argimón, S., Zhang, W., . . . O'Toole, P. W. (2015). Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. *Nature Communications, 6(1),* 8322. https://doi.org/10.1038/ncomms9322

Zheng, J., Wittouck, S., Salvetti, E., Franz, C. M. A. P., Harris, H. M. B., Mattarelli, P., . . . Lebeer, S. (2020). A taxonomic note on the genus *Lactobacillus*: Description of 23 novel genera, emended description of the genus *Lactobacillus* Beijerinck 1901, and union of *Lactobacillaceae* and *Leuconostocaceae*. *International Journal of Systematic and Evolutionary Microbiology, 70(4),* 2782-2858. https://doi.org/10.1099/ijsem.0.004107

Transcriptomics of AGR1485 and AGR1487 before and during the TEER assay

6.1. Abstract

Background: In previous chapters, *Limosilactobacillus fermentum* AGR1487 has been shown to cause a drop in transepithelial electrical resistance (TEER) assays, that was not expressed by AGR1485 (Chapter 3). Prior genomic analyses revealed that AGR1487 harbours 62 unique genes, possibly contributing to its barrier-disruptive properties (Chapters 4 and 5).

Aim: This study aimed to identify candidate genes associated with AGR1487's barrier-disruptive phenotype by investigating its transcriptomic response of AGR1485 and AGR1487 during the TEER assays and comparing the results to the discoveries of previous chapters.

Methods: Transcriptomic profiling was performed on AGR1487 and AGR1485 cultures maintained in MRS media and during TEER assays in M199 media. Differential gene expression analysis compared the transcriptional profiles between these two conditions per strain. Additionally, a Classifications Of Genes (COG) analysis was conducted to evaluate functional gene expression patterns in AGR1487 and AGR1485 across both media types. The transcripts recorded in each condition were aligned to their respective strain's genome and evaluated.

Results: AGR1487 expressed 875 genes in MRS, which decreased to 97 genes during the TEER assay. Notably, none of the 97 gene expressed by AGR1487 during the TEER assay were identified as unique to the strain, suggesting that the barrier-disruptive phenotype may be primed before the TEER assay (M199 media). In addition, AGR1487 and AGR1485 exhibited downregulation of genes involved in ribosomal structure, biogenesis, cell cycle control, cell division, chromosome partitioning, and vesicle transport in M199, a finding consistent with the limited bacterial growth observed in this medium. However, all these COGs were observed in MRS. From these results, 25 candidate genes were identified as potential mediators of AGR1487's barrier-disruptive effect.

Conclusion: The transcriptomic profiling of AGR1485 and AGR1487 supported that neither strain can grow in M199 media that is required to support the growth of Caco-2 cells. Furthermore,

the gene candidate(s) responsible for the barrier disruptive phenotype of AGR1487 are likely expressed in MRS media (the bacterial growth broth) before the TEER assay. The 62 gene candidates identified in Chapter 5 were successfully reduced to 25 candidate genes that may underpin AGR1487's unique barrier-disruptive properties. Future studies employing gene knockout, overexpression, and protein localisation approaches will be essential to validate these candidates and further elucidate the mechanisms driving this phenotype.

6.2. Introduction

Identifying and characterising genes responsible for specific phenotypes in bacteria is fundamental to understanding their physiological roles and potential impacts on host health (Han et al., 2017; Kurilshikov, Wijmenga, Fu, & Zhernakova, 2017). One such phenotype is the barrier-disruptive activity exhibited by *Limosilactobacillus fermentum* AGR1487, which has the potential to negatively affect the host (Anderson et al., 2013). This Chapter focuses on the transcriptomics of *L. fermentum* AGR1487 to uncover the gene candidate(s) that may be responsible for this phenotype during the transepithelial electrical resistance (TEER) assay of human intestinal epithelial cellular (Caco-2) monolayers in M199 cell culture media.

M199 is a complex medium formulated for human cell culture and an integral component of maintaining Caco-2 monolayers required for TEER assay. However, before the TEER assay, the bacterial strains (AGR1487 and AGR1485) were grown and maintained in MRS (de Man, Rogosa, and Sharpe) broth. MRS is a nutrient-rich medium designed for the optimal growth of lactic acid bacteria and is incompatible with TEER assays.

The first question is in which medium (MRS or M199) are the gene(s) responsible for the barrier-disruptive phenotype expressed. Previous studies have indicated that the barrier-disruptive phenotype is retained in AGR1487 cultured in MRS even after the cells are killed (Sengupta et al., 2015). This observation suggests that the responsible factors are pre-expressed into gene products and remain active after killing the bacteria. This finding indicates that the genes responsible for the barrier-disruptive phenotype were likely to be expressed when AGR1487 was cultured in MRS medium.

In Chapter 5, the genetic content of AGR1487 was explored, and it was determined that AGR1487 harbours 62 non-identical or “unique” genes and the regions in the genome where these unique genes could be identified. This Chapter expands on these findings by providing further

evidence to determine in which media format the gene candidate(s) responsible for the phenotype were expressed. Finally, the transcriptomics data was explored to determine which of the 62 unique genes were expressed and could be involved in the barrier-disruptive phenotype.

6.3. Aims and Hypotheses

Aim 1: Determine the gene transcripts of both strains during the TEER assay of Caco-2 monolayers and compare them to the gene transcripts of both strains grown in MRS.

Aim 2: Determine whether the gene candidate(s) predicted to be responsible for the barrier disruptive phenotype of Caco-2 monolayers were expressed during the TEER assay or before the assay in MRS broth.

Hypothesis 1: The cell growth media M199 is insufficient to maintain the growth of AGR1487 and AGR1485 during the TEER assay. Therefore, the genes responsible for the barrier-disruptive phenotype of *L. fermentum* AGR1487 are expressed before the TEER assay.

Hypothesis 2: Transcriptomic analysis, which identifies genes that are expressed and those that are not, is hypothesised to simplify the identification of unique genes and pinpoint potential candidate genes responsible for the barrier-disruptive phenotype observed in TEER assays.

6.4. Materials and methods

6.4.1. Caco-2 cells cell cultures

All aspects of Caco-2 cell culture are described in Chapter 3, sections 3.4.1 – 3.4.5.

6.4.2. Bacterial cell cultures

All aspects of bacterial cell culture and maintenance are described in Chapter 3, sections 3.4.6 – 3.4.8.

6.4.3. TEER assay co-cultured with bacterial strains

Except for the changes listed below, the TEER assays were implemented as described in Chapter 3, Section 3.4.8. The experiments were conducted using Corning 6-insert PET Transwells with a pore size of 0.4 μm and a growth area of 4.67 cm^2 (Corning, Lindfield, New South Wales, Australia). Each plate contained two wells of each strain (AGR1485 and AGR1487) and two no-treatment controls containing only media. These two wells per condition represented technical replicates on the same plate. A separate, independently grown overnight culture was used for each plate, providing three biological replicates per condition across the experiment. From each plate, one well per condition was selected for RNA extraction, while the other was used for TEER measurement continuity. The assay was run for 8 hours before the selected wells were removed and stored in RNAlater (#AM7020, ThermoFisher Scientific, Auckland, New Zealand).

6.4.4. Bacterial RNA isolation

The 1.5 mL apical volume was removed from the Corning 6 insert PET Transwells and added to pre-filled 15 mL Nunc tubes containing 5 mL of RNAlater (#AM7020, ThermoFisher Scientific,

Auckland, New Zealand). The mixture was mixed by inversion and snap-frozen in liquid nitrogen for storage at -80 ºC for RNA extraction.

The samples were allowed to thaw before a brief centrifugation. The supernatant was removed, and the pellet was refrozen in liquid nitrogen. The cells were lysed by grinding under liquid nitrogen, followed by the TRIzol™ Max™ Bacterial RNA Isolation Kit using the manufacturer's protocol with Max bacterial enhancement reagent (#16096020, ThermoFisher Scientific, Auckland, New Zealand) ("Thermofisher document connect," 2020).

6.4.5. mRNA quality control

The mRNA extracted, using the method in Section 6.4.4, was quantified on an Agilent Bioanalyser 2100 using RNA 6000 Nano Kit (SKU: 5067-1511, Agilent, New Zealand) following the manufacturer's protocol. The mRNA was run alongside the Agilent RNA 6000 Nano Ladder provided with the kit. The samples had an RNA integrity number (RIN) between 8.1 – 9.7. A 25µL aliquot of ~2000 $ng.\mu L^{-1}$ mRNA per sample was applied to GenTegra RNA tubes (SKU: ISC09, Custom Science, Auckland, New Zealand) in a biosafety hood for approximately 18 hours following the manufacturer's protocol.

6.4.6. Transcript sequencing

The mRNA was sent to Custom Science LTD (Newmarket, Auckland, New Zealand) for sequencing. The service provided quality control, stranded mRNA library preparation, library quality control and sequencing. Sequencing was ordered for 150 pair-end reads delivered at ≥ 6Gbs of data per sample from an Illumina NovaSeq 6000. All the samples were sequenced on the same lane and supplied as raw reads.

6.4.7.Software and packages

All software applications were run using their default settings unless otherwise stated. The applications were regularly updated before running the analyses. All software packages were run locally on workstations or servers within Conda (v22.11.1) environments on the Linux Ubuntu 20.04.5 LTS (Focal Fossa) operating system. NCBI RefSeq GenBank files were used for all analyses. All scripts associated with the workflows described here are provided in the digital appendix under the same chapter name. In this Chapter, various scripts were utilised and written in multiple programming languages, including Bash (Bourne Again SHell), Perl (v5.32.1), Python (v3.8), and R (v4.1.0). Statistics were handled by the packages and pipelines used for these analyses and are stored in the digital appendix. A complete workflow for the differential expression that accesses the nested scripts stored in the digital appendix is also provided. Most of the differential expression analyses and scripts were provided by the course "RNA-seq by example" presented by Biostars (Albert, 2020). A step-by-step script for all operations is provided in the digital appendix.

6.4.8.Read quality control

Default parameters were applied for all software packages unless otherwise specified. Illumina short-read quality control was analysed using FastQC (v0.11.9) (Andrews, 2010) before and after trimming with Trimmomatic (v0.39) (Bolger, Lohse, & Usadel, 2014). FastQC (v0.11.9) was used to compare the trimmed and raw reads before alignment.

6.4.9. Read statistics

The read statistics were called with seqkit (v2.3.0) and stored with the following command (Shen, Le, Li, & Hu, 2016). seqkit stats refs/*.fna > stats_fasta_files.txt

6.4.10. Read indexing and alignment-based RNA-Sequencing

A sample ID file was created to store the sample names using "parellel" (Conda installed 2021). A variable was assigned for the reference genome before the genome was indexed with hisat2 and Samtools. The fastq files were aligned to the reference with hisat2 (v2.2.1) (Zhang, Park, Bennett, Thornton, & Kim, 2021) and sorted with Samtools (v1.12) (Li et al., 2009). The alignment files (.BAM) were converted to bed files for manual inspection with bedtools (v2.31.0) (Quinlan, 2014) and bedGraphToBigWig (v302.1). These files were, in turn, used to assess the quality of the features and to determine if any were missing.

6.4.11. Feature counting RNA-Sequencing

Reference-based transcriptomics read counts were conducted using the NCBI RefSeq GenBank files and genomes recorded in the bam format created in section 6.4.10. Bio fetch was used to download the files via the terminal interface (e.g. bio fetch NZ_CP047585.1 > AGR1487_gen.gb). The annotation details were extracted from the GenBank file and stored in a gene feature file with the ".gtf" extension. The .gtf file was used as a feature reference file for Featurecount to find features from the alignments (.bam) above. The default type "–t" parameter was changed to "gene" from "exon" to match the annotation categorisation. The file type "-F" was changed to the .gtf from the default "gff". Changing the file type did not impact any of the counts. Read pairs were counted with setting "-p" on the same strand "--countReadPairs -S 1".

6.4.12. Classification-based transcriptomics

Kallisto was chosen for downstream differential gene expression analyses due to inconsistencies with the stability of the Salmon install at the time of writing. Kallisto was run with

default parameters, and the output was normalised for differential expression analysis using the "combine.py" Python script provided by Biostars. The script can be downloaded using curl.

"curl http://data.biostarhandbook.com/books/rnaseq/code/combine.py > ~/bin/combine"

6.4.13. Differential gene expression analysis

The differential gene expression was processed using R-based methods via the terminal. The differential statistics were calculated with EdgeR using the Kallisto counts created above. The distribution estimate was calculated using the qCML method recommended by EdgeR and BioStars. The typical dispersion and tagwise dispersions were calculated in a single run using default parameters. Testing the differentially expressed genes was accomplished with a 3x3 experimental design comparing M199 to MRS with an exactTest. A heatmap of the results.csv was plotted with the Rscript "create_heatmap.r."

The volcano plots were created in R using the custom script Volcano_plot.R saved in the scripts folder in the digital appendix. The scripts sort and label the data based on adjusted p-value (PAdj) equal to 0.01 and a log fold change of 0.58. Genes below these thresholds are labelled "no." However, genes with a PAdj greater than 0.01 were sorted by log2 fold change. fold change > 0.58 = UP, and fold change < 0.58 = DOWN.

6.4.14. Classification of genes analysis

The genes from the Volcano plot were sorted into individual files for the Classification of Genes (COG) analysis. The analysis was completed using the default parameters of the COGclassifier (v1.0.5) provided by Moshi4.

6.4.15. Genome atlas

A genome atlas was constructed with Circos (v0.69-9). The AGR1487 genome atlas was created with a customised configuration file designed to align the genome sequencing in its circular form with the mmseq2 protein alignments and blast results indicating the location of unique genes from Chapter 5. The atlas was also augmented with the log fold change data extracted from the Kallisto EdgeR results. The entire file, subsequent runs and configuration files are provided in the digital appendix titled “genome_atlas_AGR1487”.

6.5. Results

6.5.1. Differential gene expression of AGR1485 and AGR1487

Heatmaps were used to present a high-level overview of the expression data through the normalised matrix of gene counts. The matrix was filtered at a 5% false discovery rate and grouped by inter-replicate and inter-sample. Switching AGR1485 from MRS to M199 caused the downregulation of the expression of 575 genes in M199 and an up-regulation of 251 genes (Figure 6.5-1). However, AGR1485 had a lower inter-replicate consistency in M199 than in MRS (Figure 6.5-1). By contrast, AGR1487 downregulated the expression of 875 genes and upregulated the expression of 93 genes in M199 (Figure 6.5-2). Unlike AGR1485, the inter-replicate consistency was more consistent in M199 but less consistent in MRS than ARG1485 (Figure 6.5-1 and Figure 6.5-2). Volcano plots were prepared to compare the significance and magnitude of all the genes while allowing the annotation of the most upregulated genes in each media type per strain.

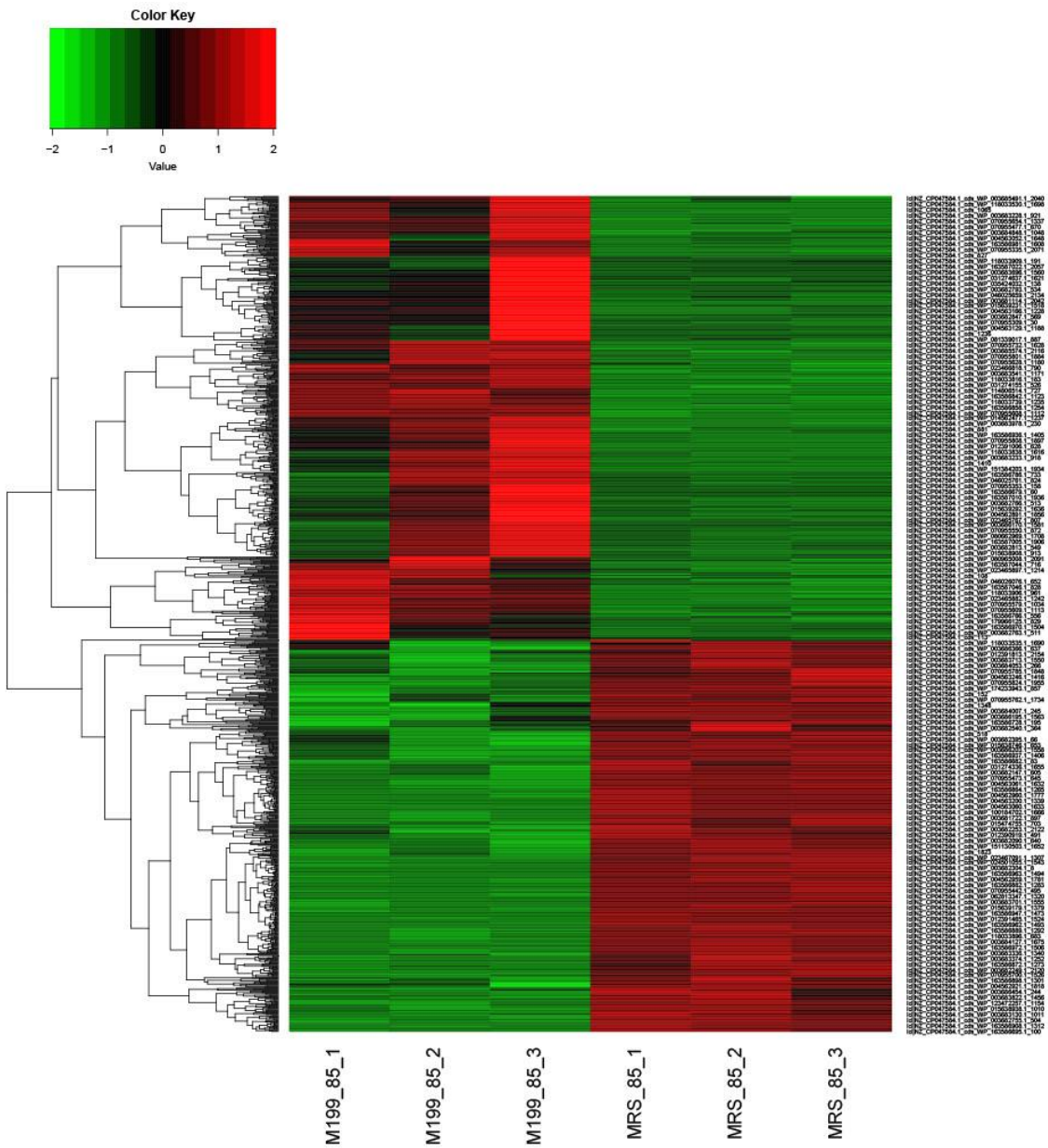

Figure 6.5-1: Gene expression heatmap for AGR1485. Samples are grouped by growth medium: M199 (left) and MRS (right). Negative values (decreases) are represented in Green, while positive values (increases) are shown in red. The normalised matrix false discovery rate was filtered at 5% before plotting.

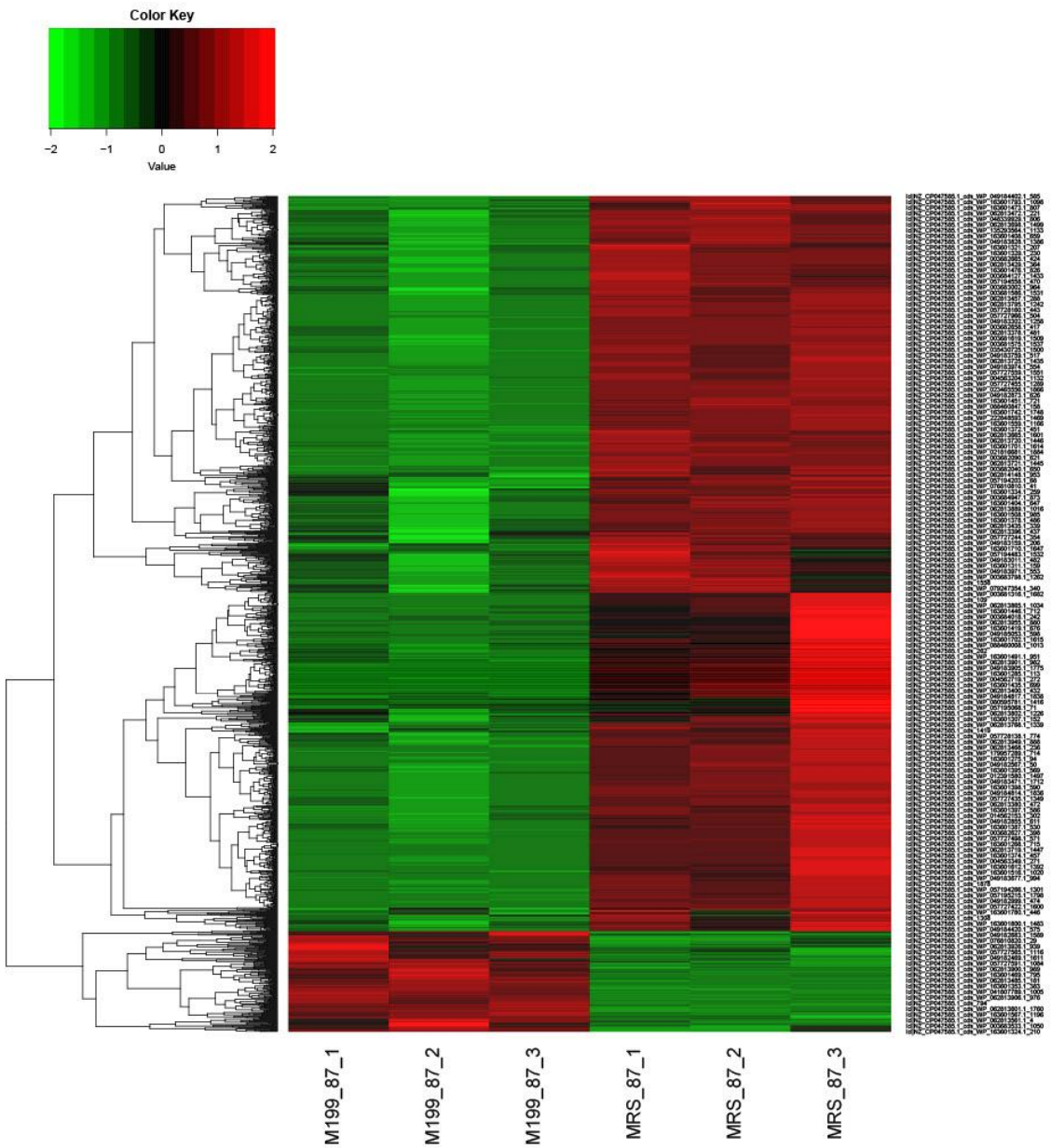

Figure 6.5-2: Gene expression heatmap for AGR1487. Samples are grouped by growth medium: M199 (left) and MRS (right). Negative values (decreases) are represented in Green, while positive values (increases) are shown in red. The normalised matrix false discovery rate was filtered at 5% before plotting.

6.5.2. Volcano plots

Volcano plots were used to display the transcriptional difference between the two strains grown in MRS and M199 (Figures 6.5 3 and 6.5 4) using a log2 fold change. The genes were first filtered to show only those with a log2 fold change greater than 0.58. The vertical (black) lines show a log2 fold change cut-off of 0.8, while the red horizontal line shows a p-value cut-off of 5%. Green shows the most upregulated genes in M199 vs the most upregulated genes in MRS in Red.

The most upregulated MRS genes in both strains were predominantly from genes associated with domesticated phages (Figure 6.5-3 and Figure 6.5-4). For instance, in AGR1485, the most upregulated gene in MRS was a phage-capsid protein (Figure 6.5-3). In AGR1487, a single-stranded binding protein and phage anti-repressor protein were the two most upregulated genes in MRS (Figure 6.5-4). Again, these genes and all the hypothetical proteins trace back to region three. This region is bordered by palindromic repeats and contains transposases discussed in Chapter 5.

Both strains expressed fewer genes and fewer transcripts per gene when switched to M199 media from MRS (Figure 6.5-3 and Figure 6.5-4). The expression pattern of both strains in M199 media did not correlate to any unique genes found for the strains in Chapter 5; for instance, the result, SPFH/Band protein, was common to both strains and many of the 40 *L. fermentum* strains used in the pan-genome dataset (Chapter 5). However, when the strains were grown in MRS media, the expression patterns for both strains included many, but not all, their unique genes. For example, phage anti-repressor and single-stranded binding proteins unique to AGR1487 were highly in MRS but were absent in M199 cell culture media. A COG analysis was conducted for the transcripts of both strains to further explore the function of the expressed genes in each media type.

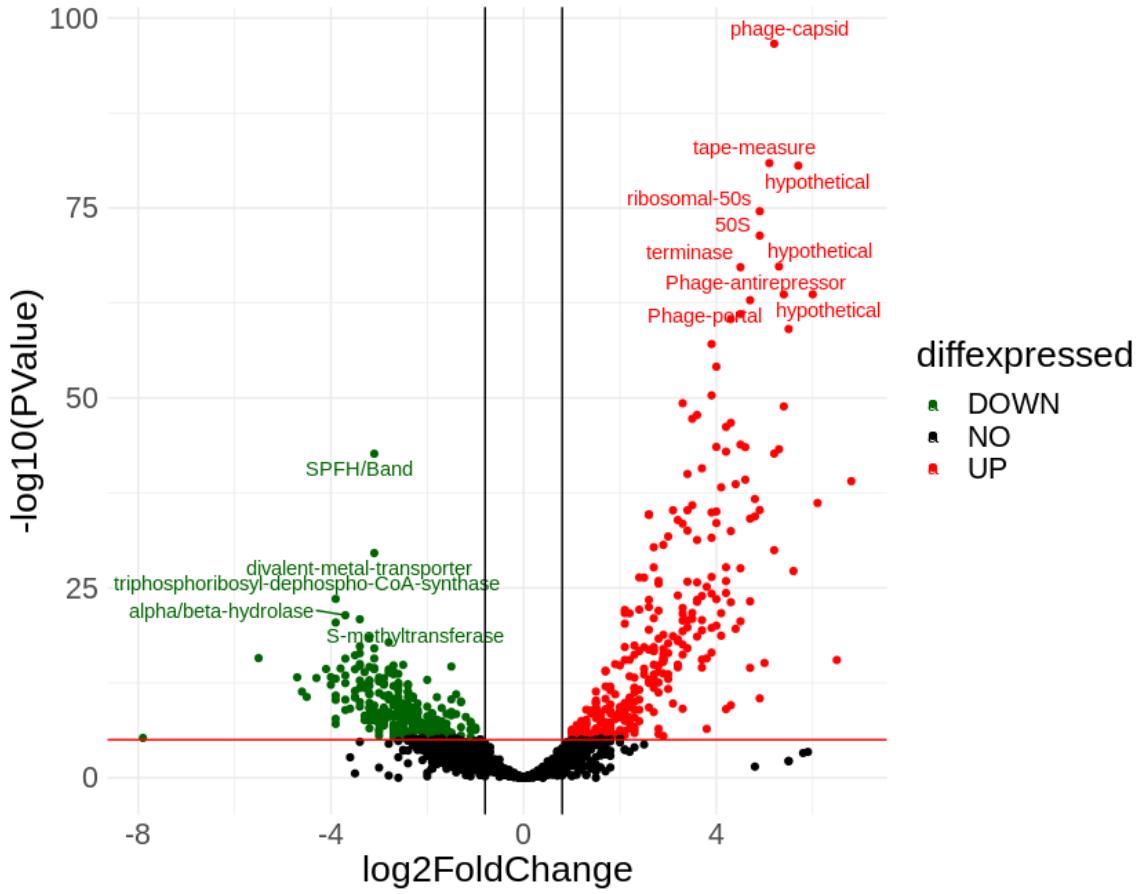

Figure 6.5-3: Volcano plot AGR1485 gene expression. The genes most upregulated in MRS media were coloured RED, while the most down-regulated genes were coloured Green. Vertical black lines show a log2 fold change of 0.8, while the horizontal red indicates the 5% p-value cut-off. The top 10 genes in each category were annotated.

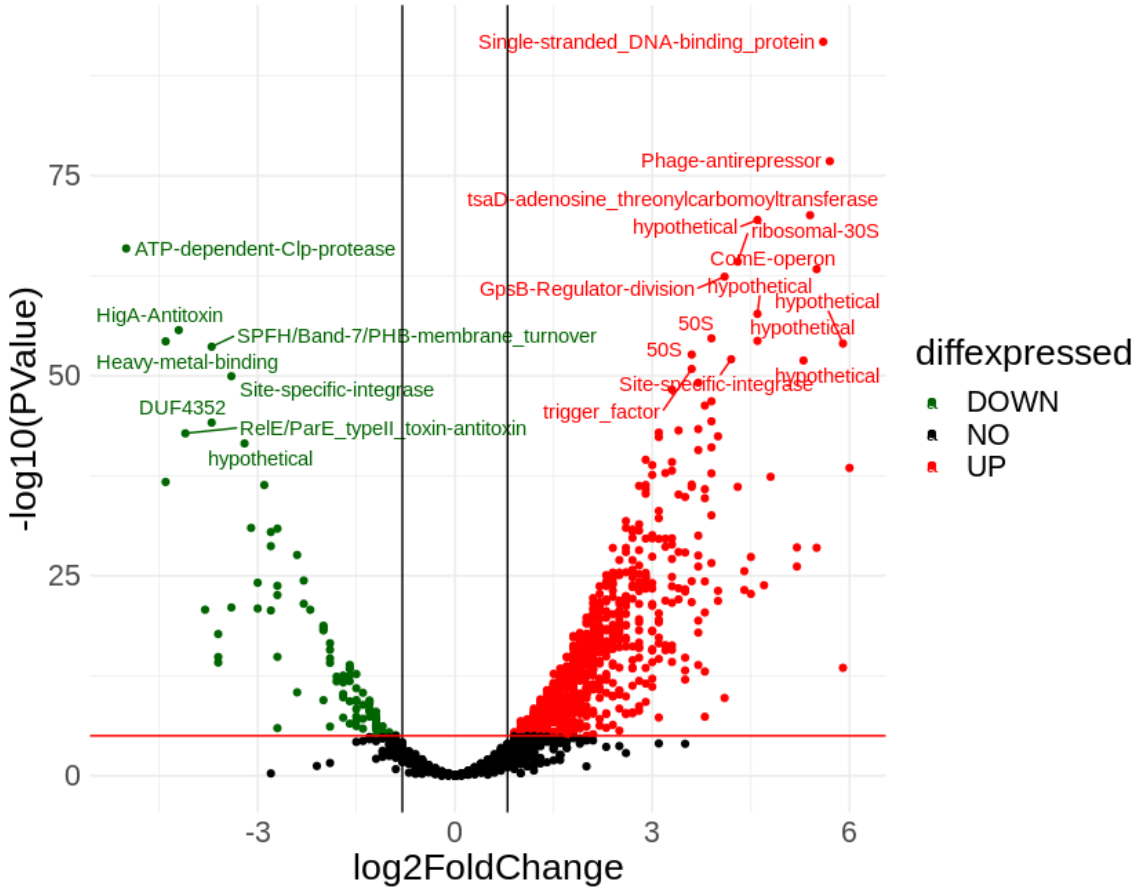

Figure 6.5-4: Volcano plot AGR1487 gene expression. The genes most upregulated in MRS media were coloured RED, while the most down-regulated genes were coloured Green. Vertical black lines show a log2 fold change of 0.8, while the horizontal red indicates the 5% p-value cut-off. The top 10 genes in each category were annotated.

6.5.3. COG functional analysis

The COG functional analysis provides a systematic framework for categorising genes based on their predicted functions and evolutionary relationships. This classification scheme was used to study the high-level function of overexpressed genes in both strains within each media type.

AGR1485 expressed fewer genes in M199 than in MRS, as outlined in section 6.5.1. Most notably, the category most expressed changed from **J**, translation, ribosomal structure and biogenesis in MRS to **E** amino acid transport and metabolism in M199 (Figure 6.5-5). All genes associated with categories **D** and **U**, cell cycle control, cell division, chromosome partitioning, intercellular trafficking, secretion, and vesicle transport, are missing from AGR1485 transcripts when exposed to M199 media during the TEER assay but were present in MRS media (Figure 6.5-5). Additionally, there was a reduction in transcripts associated with the categories **M** and **P** when AGR1485 was transitioned to M199. M-class genes refer to the cell wall, membrane, and envelope biogenesis genes, and P-class genes are related to inorganic ion transport and metabolism (Figure 6.5-5).

N and **Q**-class transcripts were exclusively detected in M199 media and were not expressed in MRS. **N**-class genes control cell motility and were not unique to AGR1485, yet they were expressed only when the cells were in M199 media (Figure 6.5-5). Furthermore, **Q**-class genes are related to secondary metabolite biosynthesis, transport, and catabolism. **Q** is the only category containing catabolism genes, and these genes were inactive when AGR1485 was in MRS but became active when the cells were in M199 (Figure 6.5-5). When AGR1485 was in M199 cell culture media during the TEER assay, there was an absence of gene expression relating to the categories cell wall, membrane, envelope biogenesis, cell cycle control, cell division, and chromosome partitioning (Figure 6.5-5). The cells shifted expression to secondary metabolite biosynthesis, transport, catabolism, and cell mobility genes.

AGR1487 displayed similar transcriptome behaviour in MRS and M199 as AGR1485 but expressed far fewer genes (97 in total) in M199 than AGR1485. AGR1487 saw a sharp shift from 875 genes in MRS to 93 genes in M199. The 93 genes are representative of most of the COG categories. However, the same **J** (translation, ribosomal structure and biogenesis) to **E** (amino acid transport and metabolism) shifts were present in AGR1487 as in AGR1485 (Figure 6.5-6 and Figure 6.5-5). Genes associated with **L** (replication, recombination and repair) and **M** (cell wall/membrane/envelope biogenesis) processes dropped from 46 and 61 genes to 2 and 1, respectively. **D** (Cell cycle control, cell division, chromosome partitioning) and **U** (Intracellular trafficking, secretion, and vesicular transport)-class genes were present in MRS, but these functions were lost in M199 (Figure 6.5-6). However, the **Q**-class (Secondary metabolites biosynthesis, transport, and catabolism) genes were expressed in MRS for AGR1487 but not in M199 media, contrary to AGR1485's expression behaviour, which may have to do with AGR1487's unusual cell morphology (Figure 6.5-6). In addition, AGR1487 also had a complete loss of C-class genes, which were associated with energy production and conversion, along with a sharp reduction in **F**, **H**, **I**, and **P**-class genes (Figure 6.5-6). **F**, **H**, **I**, and **P**-class genes were associated with nucleotide transport and metabolism, coenzyme transport and metabolism, lipid transport and metabolism, and inorganic ion transport and metabolism (Figure 6.5-6).

When AGR1487 was moved to M199 during the TEER assay, the gene expression for genes related to energy production and conversion and reduced expression genes associated with various metabolic processes were no longer detected. AGR1487, like AGR1485, lost the cell wall, membrane, envelope biogenesis, cell cycle control, cell division, and chromosome partitioning gene expression. These changes indicate that both strains did not proliferate and produce new cell walls, and had less translation, ribosomal structure creation, and biogenesis in M199 cell culture media. To visualise the relative location of the unique genes relative to other genes and gene transcripts, the data was plotted on a single AGR1487 genome atlas for examination.

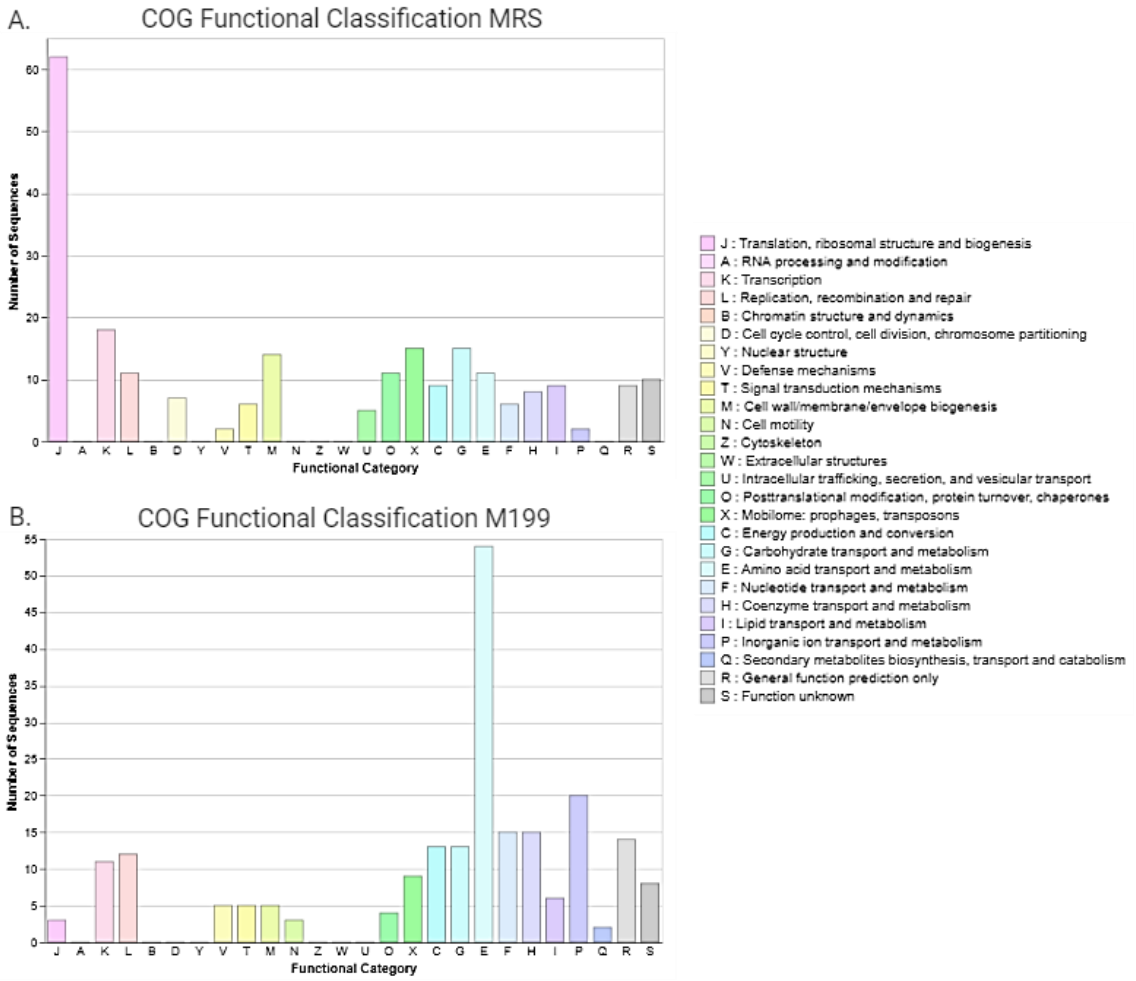

Figure 6.5-5: Functional classification of AGR1485's genes. A. Functional classifications of genes were expressed while the AGR1485 was in MRS before the TEER assay. B. Functional classifications of all the genes when AGR1485 was cultured in M199 during the TEER assay. The gene lists were extracted from the filtered results from the Volcano plots, excluding any genes below the 5% p-value cut-off line. The data was further filtered to remove genes with a log fold change equal to or less than 0.58 fold.

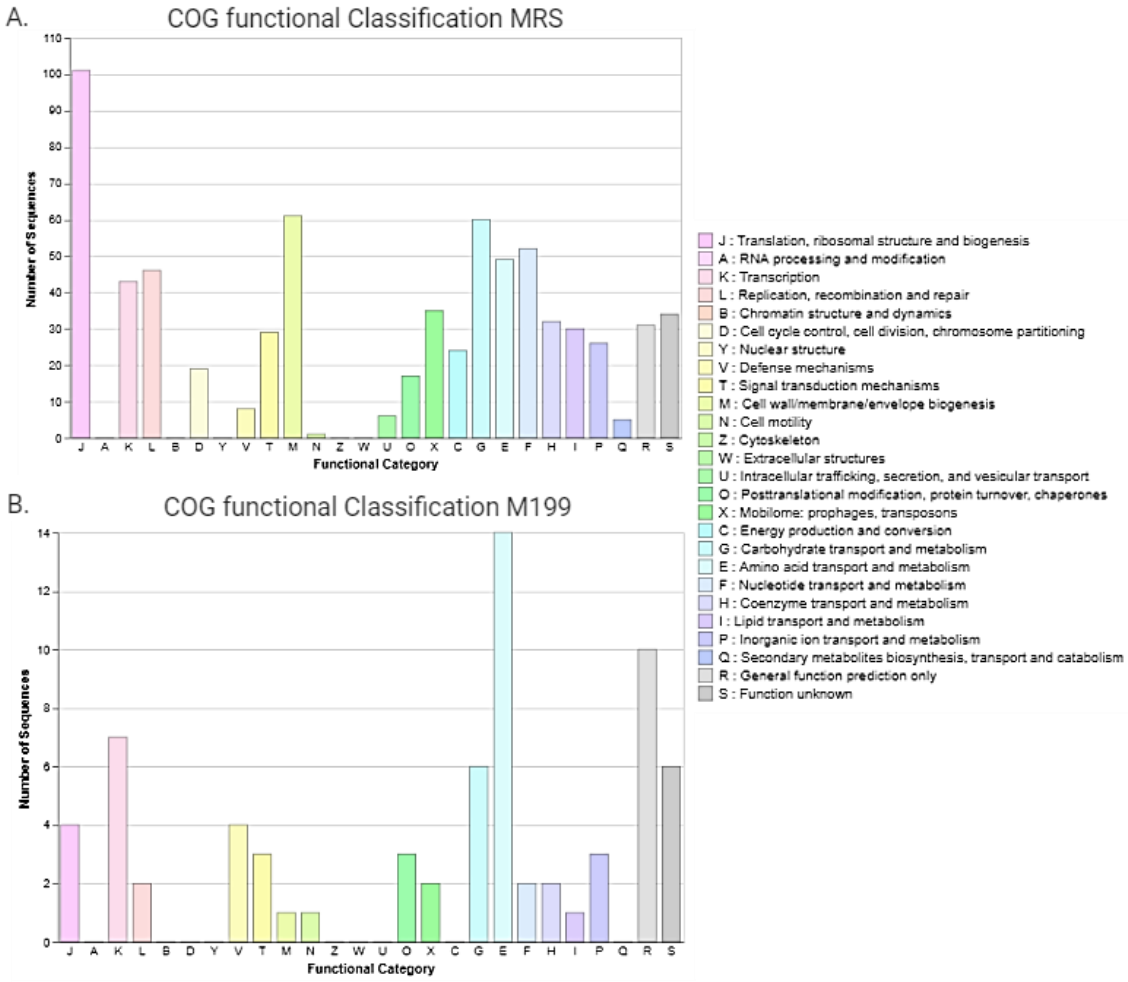

Figure 6.5-6: Functional classification of AGR1487's genes. A. Functional classifications of genes were expressed while the AGR1487 was in MRS before the TEER assay. B. Functional classifications of all the genes when AGR1487 was cultured in M199 during the TEER assay. The gene lists were extracted from the filtered results from the Volcano plots, excluding any genes below the 5% p-value cut-off line. The data was further filtered to remove genes with a log fold change equal to or less than 0.58 fold.

6.5.4. AGR1487 genome atlas integrated with expression data

The transcripts detected for AGR1487 in MRS media did not include all the unique genes (Chapter 5). The list of 62 unique genes from Chapter 5 was refined by aligning the transcriptomics reads to the genome to determine which of these 64 genes were expressed. Furthermore, the resulting genes were cross-referenced with the statistical data generated during the EdgeR differential expression analysis to investigate the statistical likelihood that the gene expression levels were not false positives.

Of the 62 unique genes, 27 were not expressed or were detected as false positives by EdgeR. All the results are provided in the digital appendix. The remaining 37 genes were filtered by function and described per genomic region using the same atlas described in Chapter 5, augmented with the transcriptomics data (Figure 6.5-7). Within each region, only genes that were unique to AGR1487 are described.

In Region 1 of the atlas (Figure 6.5-7), 17 unique genes were present, of which 8 were false positives during the statistical analysis, one gene was downregulated, and two genes had known cytosolic functions (Table 6.5-16). The remaining six genes from Region 1 might play roles in various phenotypes unique to AGR1487, which are explored in the discussion.

Region 2 contained enzymes with known functions, such as beta-galactosidase, a cytosolic protein responsible for breaking down lactose. In addition, this cluster of unique genes included a symporter, a DNA-binding protein, a DNA transposase, and one oxidoreductase. The gene products are predicted to be predominantly cytosolic and not linked to the cell wall or cell wall functions. It is unlikely that these genes were involved in the barrier-disruptive phenotype of AGR1487.

Region 2 contains a single hypothetical protein with an unknown function. Still, the measured gene expression in MRS was low, and the associated gene did not contain a detectable surface localisation signal sequence within or near the open reading frame.

Region 3 contained the highest number of potential targets and all of the most highly expressed genes. The region included 17 unique genes, of which only one was downregulated. Furthermore, six genes had known cytosolic functions and were targeted towards the DNA. These genes included three transposases and three DNA-binding proteins or transcriptional regulators. The remaining eleven proteins included seven hypothetical proteins, a single-cell surface-targeted truncated metallopeptidase, and a surface-targeted S74 domain phage tail chaperon with hydrolase activity.

Region 4 contained 14 genes; eight were downregulated or detected as false positives. This region appeared to contain the remains of a domesticated phage in other strains that include phage capsid and tail protein genes. In AGR1487, only the phage scaffolding protein and palindromic repeats bordered the region remained. Four genes had known functions: a Type II toxin and antitoxin system, a DNA binding Class 1 SAM-dependent methyltransferase, and a couple of transposases. None of these functions have a known interaction with tight junctions and are not targeted towards the cell surface. It is, therefore, unlikely that they are involved in the phenotype. Two hypothetical proteins are in the cluster; one was downregulated, and the other did not have a signal peptide in its sequence or surrounding area. Region 4 was unlikely to be involved in AGR1487's barrier-disruptive phenotype.

Region 5 contained eight unique genes. Four were found to be false positives, and one was downregulated. The remaining three genes were annotated to be two DUF-containing proteins that are membrane-targeted and a single HigA toxin/antitoxin system. While none of these genes had known functions that could result in the barrier-disruptive phenotype, the DUF proteins are largely uncharacterised. The HigA system is known to be involved in biofilm formation when the harbouring strains are under stress.

Table 6.5-16: Unque genes of AGR1487 with MRS expression data

	ID	Annotation details	Start	End	Exp	FDR	
Region 1	RS00155	ABC transporter permease	33,309	34,472	-1.1	TRUE	
	RS00325	Aspartate ammonia-lyase	78,029	79,408	1.9	TRUE	**N**
	RS00335	Diguanylate cyclase	79,768	79,980	1.1	FALSE	
	RS00340	GNAT family N-acetyltransferase	80,164	80,748	1.8	TRUE	**Y**
	RS00350	Aminotransferase_class_I/II-fold_pyridoxal_phosphate-dependent_enzyme	82,665	83,840	1	FALSE	
	RS00355	Hypothetical protein	84,100	84,333	0.5	FALSE	
	RS00580	DUF4422_Family-glycosyltransferase	133,622	134,413	3.7	TRUE	**Y**
	RS00585	Capsular polysaccharide synthesis protein	134,410	135,309	3.6	TRUE	**Y**
	RS00590	Glycosyltransferase	135,334	136,320	3.5	TRUE	**Y**
	RS00595	EpsG family protein	136,344	137,534	3.6	TRUE	**Y**
	RS00600	Serine acetyltransferase	137,657	138,151	3.3	TRUE	**N**
	RS00910	Hypothetical protein	198,689	199,249	0.4	FALSE	
	RS00925	Hypothetical protein-flavodoxin	201,038	201,547	0.3	FALSE	
	RS00930	Hypothetical protein	201,534	201,890	0.3	FALSE	
	RS01080	Hypothetical protein	231,127	231,687	0.1	FALSE	
	RS01085	Aldo/keto reductase	231,926	232,783	1.2	FALSE	
	RS01090	Aldo/keto reductase	232,797	233,660	0.9	TRUE	**N**
Region 2	RS01665	Beta-galactosidase	351,969	353,981	1	TRUE	**N**
	RS01670	Glycoside-pentoside-hexuronide (GPH): cation symporter	353,974	355,419	1.3	TRUE	**N**
	RS01675	LacI family DNA-binding transcriptional regulator	355,552	356,574	1.3	TRUE	**N**
	RS01680	SDR family oxidoreductase	356,885	357,679	2	TRUE	**N**
	RS02170	IS5 family transposase	453,952	454,727	1.6	TRUE	**N**
	RS02660	Hypothetical protein	548,139	548,309	0.6	TRUE	**Y**
Region 3	RS03575	Arm DNA-binding_domain-containing protein	718,176	718,397	3.3	TRUE	**N**
	RS03615	Hypothetical protein	723,640	724,083	3.9	TRUE	**Y**
	RS03620	ImmA/IrrE family metalloendopeptidase	724,110	724,334	4.6	TRUE	**Y**
	RS03625	Hypothetical protein	724,306	724,605	2.8	TRUE	**Y**
	RS03635	Transposase	725,263	725,586	3.8	TRUE	**N**
	RS03640	Helix-turn-helix transcriptional regulator	725,742	725,957	6	TRUE	**N**
	RS03645	Phage antirepressor	725,990	726,778	5.7	TRUE	**N**
	RS03655	Hypothetical protein	727,859	728,179	4.6	TRUE	**Y**
	RS03660	Inositolphosphorylceramide synthase	728,323	728,490	4.4	TRUE	**Y**
	RS03695	Hypothetical protein	731,789	732,790	2.8	TRUE	**Y**
	RS03700	Hypothetical protein	732,832	733,278	4.4	TRUE	**Y**
	RS03795	S74 domain- Lactobacillus phage JNU_P5 - virus tail fibre, function hydrolase activity.	751,294	753,579	3.3	TRUE	**Y**
	RS03820	Hypothetical protein	754,886	755,383	3.8	TRUE	**Y**
	RS03860	IS5_family_transposase	761,687	762,007	0.9	TRUE	**N**
	RS03875	MucBP domain-containing protein	762,847	765,993	2.1	TRUE	**Y**
	RS03905	TIGR02328 Hypothetical protein	769,281	769,643	2.7	TRUE	**Y**
	RS03925	Site-specific integrase	772,259	773,065	-1.9	FALSE	

Region 4	RS04875	Rgg/GadR/MutR family transcriptional regulator	969,916	970,833	0.3	FALSE	
	RS04890	Phage scaffolding protein	973,696	973,944	0.5	FALSE	
	RS05065	Hypothetical protein	1,004,935	1,005,072	1	FALSE	
	RS05330	Type II toxin-antitoxin system RelE/ParE family toxin	1,061,383	1,061,667	-1.7	TRUE	
	RS05370	Class I SAM-dependent methyltransferase	1,068,394	1,069,104	2	TRUE	**N**
	RS05375	Transposase	1,069,264	1,069,599	2.3	TRUE	**N**
	RS05405	Hypothetical protein	1,076,705	1,076,884	-0.2	TRUE	
	RS05410	SEC10/PgrA surface exclusion domain-containing protein	1,076,986	1,077,975	0.8	FALSE	
	RS05415	SEC10/PgrA surface exclusion domain-containing protein	1,078,018	1,079,124	0	FALSE	
	RS05420	LPXTG cell wall anchor domain-containing protein	1,079,146	1,079,724	0.8	FALSE	
	RS05425	Hypothetical protein	1,079,715	1,080,389	0	FALSE	
	RS05540	TetR_family_transcriptional_regulator_C-terminal_domain-containing_protein	1,101,609	1,101,818	0.4	FALSE	
	RS05945	IS5_family_transposase	1,181,855	1,182,630	1.6	TRUE	**N**
	RS06245	Hypothetical_protein	1,248,542	1,248,784	1.6	TRUE	**Y**
Region 5	RS07340	IS200/IS605_family_accessory_protein_TnpB-related_protein	1,445,913	1,447,250	0.1	FALSE	
	RS08060	DUF421_domain-containing_protein	1,572,904	1,573,542	1.4	TRUE	**Y**
	RS08065	DUF3290_family_protein	1,573,552	1,573,995	5.2	TRUE	**Y**
	RS08410	Hypothetical protein alpha/beta hydrolase	1,645,492	1,645,905	0.9	FALSE	
	RS08610	Glycerophosphodiester phosphodiesterase	1,686,096	1,687,742	0.6	FALSE	
	RS08770	HigA family addiction module antidote protein	1,714,184	1,714,540	1	TRUE	**Y**
	RS09860	Putative holin-like toxin	1,500,217	1,500,309	0.6	FALSE	
	RS09865	Subtilisin-like protein hydrolase activity targeted at serine	1,500,407	1,500,604	0.6	FALSE	
	CP047585	IS5 family transposase	1,181,855	1,182,630	-4.2	TRUE	

The 62 unique genes augmented with MRS expression data. The hits were colour-coded **Green** for positive hits or colour-coded **Red** for false positives or if the gene was downregulated. Gene candidates that meet all requirements are indicated with a **Y** mark or marked with **N** if they were ruled out.

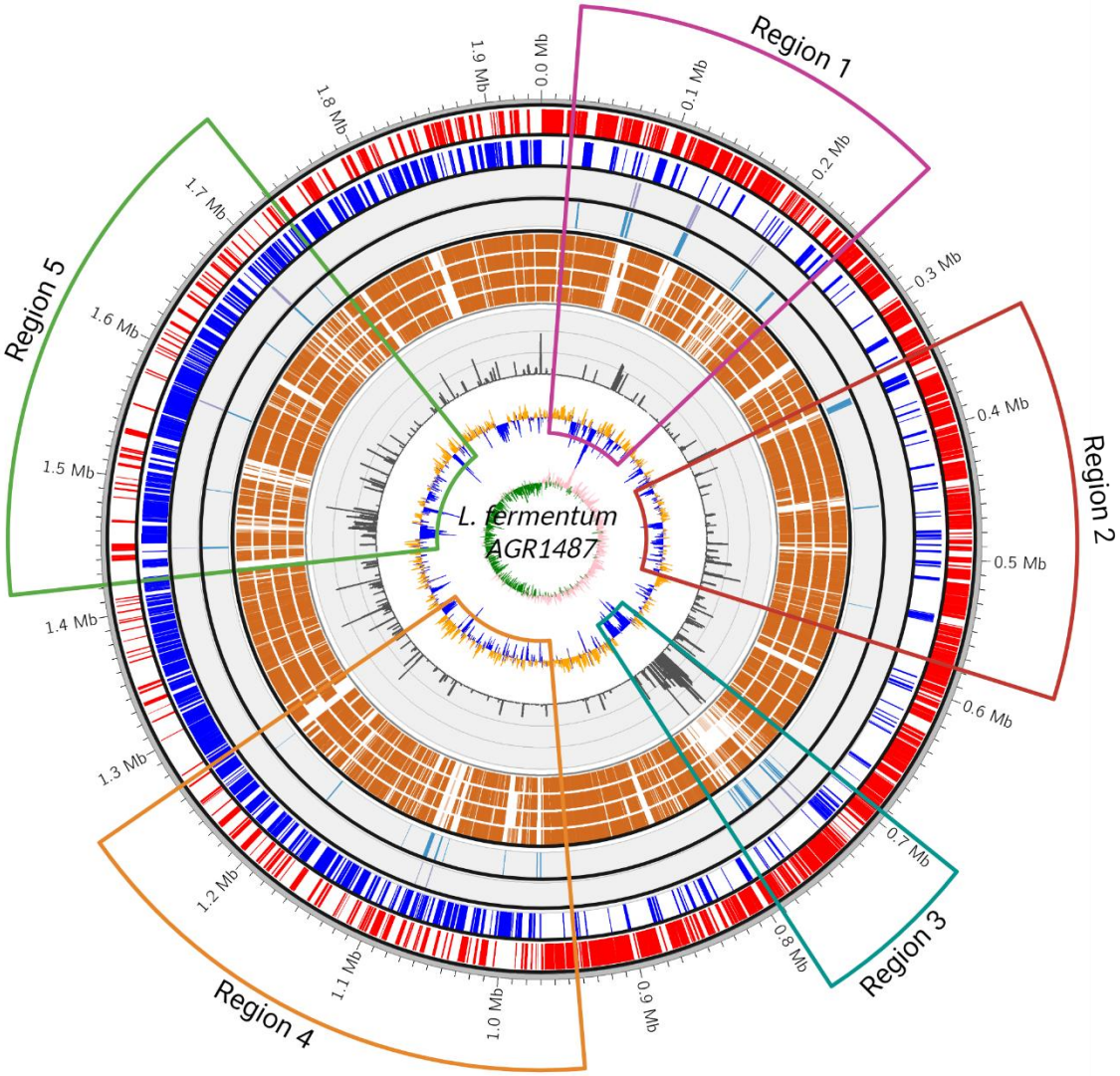

Figure 6.5-7: Genome Atlas of AGR1487 with gene expression data. The genome sequenced in Chapter 4 is overlayed with the genetic analysis data from Chapter 5 and transcriptomic data from Chapter 6. The regions defined in Chapter 5 are displayed on the atlas. The rings go from the innermost to the outermost. 1st ring: GC-skew. 2nd ring: GC-content. 3rd ring: MRS transcriptomic data filtered for data less or equal to 1.4-fold upregulation. 4th ring, MMSeq2 reciprocal best hits alignment of *L. fermentum* strains, including AGR1485. 5th ring: unique genes with alignment stringency set to 100%. 6th ring: unique genes that agree between 95% stringency and 100% datasets. 7th ring (blue): reverse genes. 8th ring (red): forward genes. 9th ring (grey): genome coordinates with ticks at every 100mbs.

6.6. Discussion

The main research question was, which of the two media formats (MRS and M199 during a TEER assay) did AGR1487 express the gene(s) responsible for the barrier-disruptive phenotype? The MRS growth broth was explicitly designed for bacteria, while M199 was designed for human cell culture. Therefore, it was expected that the two media types would impact the bacteria cells differently. mRNA was extracted while both strains were exposed to MRS and M199. The M199 samples were taken during the TEER assays while the phenotype was being monitored and successfully reproduced, as in Chapter 3. The differential gene expression in the two environments was visualised with heat maps (Figure 6.5-1 and Figure 6.5-2). AGR1485 and AGR1487 showed a switch in active genes within each media type. In MRS, more genes were expressed than in M199 (TEER assay). AGR1485 switched on nearly half its genes from MRS to M199, while AGR1487 turned off almost its entire genome, upregulating only 97 genes. These 97 genes did not align with any unique genes predicted for AGR1487.

Given that these 97 genes are not unique and assuming the unique genes produce the phenotype, these data support past findings that the barrier-disruptive phenotype is expressed before the TEER assay. Past findings showed that the barrier-disruptive phenotype occurred even when AGR1487 cells were killed (Sengupta et al., 2015). Therefore, the barrier-disruptive phenotype was likely expressed before the AGR1487 cells were washed and resuspended in M199.

The Volcano plots showed that many genes predicted as unique for AGR1487 were expressed in MRS media before the TEER assay. In MRS media, both strains' genes with the highest expression level were associated with phages (Figure 6.5-3 and Figure 6.5-4). In AGR1487, the genes that showed the highest upregulation in MRS predominantly originated from Region 3, a previously described domesticated phage (Chapter 5). Again, the Volcano plots displayed that the number and magnitude of genes expressed dropped when the strains were moved from MRS to M199. It is not

surprising that there was a marked switch in gene expression. The bacterial cells did not increase their optical density in M199 as in MRS media (Chapter 3). However, the functional behaviour of both strains in each media format remained unexplored. The genes expressed in each media format were examined by functional COG analysis to understand better the behaviour of both strains in the two media formats.

The COG analysis showed that while both strains expressed fewer genes in M199 than in MRS, the genes expressed by the strains covered various categories. Both strains shifted most genes from translation, ribosomal structure, and biogenesis in MRS to amino acid transport and metabolism in M199 (Figure 6.5-5 and Figure 6.5-6). The lack of ribosomal structure and biogenesis in M199 is vital because the genes involved in AGR1487's disruptive phenotype are unlikely to be produced and mobilised to the cell surface in the absence of expression from these genes unless latent ribosomes and mRNAs were present from the MRS exposure that persisted through the TEER assay period. Again, it is most likely that the causative genes were expressed and mobilised to the cell wall before the TEER assay started. The cells also lacked the genes for cell growth and reproduction when exposed to M199.

Neither strain had detectable expression of genes associated with cell cycle control, cell division, chromosome partitioning, intercellular trafficking, secretion, and vesicle transport transcripts when exposed to M199 media during the TEER assay. Still, these were present in MRS media (Figure 6.5-5 and Figure 6.5-6). These genes are specifically required for cell growth and division. The complete lack of transcripts from these genes indicates that the cells could not grow and divide in M199 cell culture media. Additionally, there was a reduction in transcripts associated with cell wall, membrane, and envelope biogenesis genes. Without the production of new cell walls and cell membranes, the source of the barrier-disruptive phenotype is unlikely to be created, given that the cell surface is believed to generate the barrier-disruptive phenotype (Anderson et al., 2013; Sengupta et al., 2015). However, there is still the possibility that the causative gene products were

made before washing the bacterial cell and were still mobilised to the cell surface in M199. But, unlike AGR1485, in AGR1487, intracellular trafficking, secretion, and vesicular transport transcripts were not detected in M199, making mobilisation to the cell surface unlikely in M199 as well (Figure 6.5-5 and Figure 6.5-6). It is clear from these results that the gene(s) that give rise to the barrier-disruptive phenotype are likely expressed before M199 media resuspension. No unique genes in AGR1487 were expressed in M199 during the TEER assay. Therefore, the MRS genes can be exclusively considered for gene candidates. In addition, not all of AGR1487's unique genes are expressed in MRS either. The entire list from Chapter 5 of unique genes was replicated and considered in light of transcript counting and differential expression statistics for false positives while maintaining the same five arbitrary regions to aid the discussion.

In **Region 1** of the atlas (Figure 6.5-7), many true hits (transcripts aligned with unique genes) had known functions. For instance, the aspartate ammonia-lyase starting at position 78,029 has no signal peptide, is involved in the reversible deamination of alanine and aspartate amino acids, and has had a known function for almost 100 years (Viola, 2000). The region also contains a serine acetyltransferase known to be involved in cysteine synthase and is cytosolically located (Takahashi, 2010). At position 232,797, the region also contains an aldo/keto reductase with low expression levels. This gene is cytosolically targeted and unlikely to contribute to the barrier-disruptive phenotype because it is NADPH-dependent (Penning, 2015). Given its predicted cellular location, it could not contribute directly to the barrier-disruptive phenotype.

This region does have a cluster of genes that could contribute to the unique autoaggregation of AGR1487. A cluster of unique genes in the region's centre is known to be involved in extracellular matrix production. The first of these were the GNAT family N-acetyltransferase genes. These genes produce proteins that have been documented to conduct a wide range of functions and assist in modifying the peptidoglycan and resistance to lytic enzymes and antibiotics (Burckhardt & Escalante-Semerena, 2020; Favrot, Blanchard, & Vergnolle, 2016). The GNAT genes could be the source of

the autoaggregation and lytic enzyme resistance of AGR1487, as shown in Chapter 3. In addition, the region also contained two DUF4422 family genes, one of which was accompanied by a glycosyltransferase. Glycosyltransferases catalyse the transfer of sugars to acceptor molecules and are involved in the biosynthesis of complex glycoconjugates and carbohydrates (Zhu, Wu, Zhang, & Wu, 2013). When targeted towards the cell surface, their functions include increased adhesion, signalling and cell wall biosynthesis (Drickamer & Taylor, 1998). Improved adhesion has already been shown for AGR1487 (Anderson et al., 2013).

Given that the glycosyltransferase is attached to a DUF transmembrane domain, it is likely that this transferase will function on the cell surface of the strain. In addition to the DUF4422 genes, AGR1487 also contains unique DUF421 and DUF-containing genes. These genes produce transmembrane proteins and contain signal peptides, targeting the genes towards the cell surface. DUF421 was found to be involved in extreme wet heat-resistant spore formation in wild-type *Bacillus cereus* (Yu et al., 2023). While its role in *L. fermentum* is speculative, the mechanism of action of extreme wet heat-resistant spore formation was through surface stabilisation and altering the permeability of the cells. These glucosyltransferases and DUF proteins would make good targets for a deletion analysis as they can contribute to several unique characteristics credited to AGR1487.

The region also contained an acetyltransferase with a signal peptide. Again, these genes have a known function of modifying peptidoglycan that makes the cell walls of the bacteria convey an increased resistance to lytic enzymes, a phenotype of AGR1487 not expressed by AGR1485 (Sychantha, Brott, Jones, & Clarke, 2018). In addition to the aforementioned cell surface modification genes, the region contains a capsular polysaccharide synthesis gene and an EpsG gene. These genes produce proteins known to produce the exopolysaccharide (EPS) components of the extracellular matrix and enhance EPS during biofilm formation in *L. fermentum* strain TDS030603 (Shi, Aryantini, Uchida, Urashima, & Fukuda, 2014). While systems like these are not unique to *L. fermentum*, this specific EpsG gene is unique to AGR1487. **Region 1** was characterised by genes that modify the cell

surface and extracellular environment of AGR1487 and provides a robust series of candidates for both the barrier disruptive phenotype and unique colony morphology of AGR1487.

Region 2 of the genome atlas (Figure 6.5-7) contains a single gene candidate. The gene was a hypothetical protein with an unknown function and had a low expression level. The remaining genes included beta-galactosidase, a well-documented enzyme for digesting lactose (Juers, Matthews, & Huber, 2012). The beta-galactosidase gene expression was accompanied by its associated *Lac*I transcriptional regulator (Juers et al., 2012). The other unique genes were an SDR oxidoreductase, an IS5 transposase, and a Glycoside-pentoside-hexuronide (GPH) cation symporter. These last three genes are either targeted towards the DNA or act on small molecules (Guan & Kaback, 2004; Humayun, Zhang, Butcher, Moshayedi, & Saier, 2017; Kavanagh, Jörnvall, Persson, & Oppermann, 2008). These candidates do not fit the inclusion criteria. Therefore, **Region 2**, except for the lone hypothetical protein, can be disregarded.

Region 3 of the atlas contains a genes from a domesticated phage. The highest upregulated genes from AGR1487 were traced back to this phage. The region contained seven hypothetical proteins with unknown functions, many of which have predicted signal peptides, meaning they are targeted towards the cell surface. Several genes, such as the DNA-binding Arm protein, transcriptional regulator, phage anti-repressor proteins and transposase genes, could be ruled out due to their DNA-binding nature. The remaining four genes, a truncated ImmA/IrrE metalloendopeptidase, Inositolphosphorylceramide synthase, S74 domain JNU P5 phage tail fibre protein with hydrolase activity and a Mucin Binding Protein (MucBP) domain-containing gene, all contain signal peptides and should be considered. Metalloendopeptidase gene expression in *Enterococcus faecalis* is stimulated by collagen (Hirakawa, Sugimoto, Tsuji, Inamoto, & Maeda, 2022). However, the effect the metalloendopeptidase has on tight junctions remains unknown, but there is some interaction between this protein and the host. The MucBP domain is documented to

increase the retention time of bacteria (Chatterjee et al., 2018). This region contained 11 possible gene candidates for the barrier-disruptive phenotype.

Region 4 can be disregarded except for a single hypothetical protein. The other genes in this region were not expressed. The remaining genes are a couple of DNA-targeted transposases and SAM-dependent methyltransferases. SAM-dependant methyltransferases have been well-studied and understood (Struck, Thompson, Wong, & Micklefield, 2012) and were found to be cytosolic. In addition, no surface signalling peptides were found in or around the gene. However, the region contains a hypothetical protein showing gene expression when AGR1487's barrier disruptive phenotype occurs. No internal structures of the gene were identified, such as domains or peptide motifs, leading to a complete lack of annotation. Due to this lack of annotation, this hypothetical protein cannot be ruled out and remains a gene candidate.

Of the nine genes found to be unique to **Region 5**, five were found not to be expressed and could be excluded from further analysis. The remaining genes included an antidote-toxin or toxin-antitoxin (TA) system. Despite this system being unique, homologues of these systems are present in nearly all bacteria and archaeal species (Wang, Yao, Sun, & Wood, 2021). Of these toxin-antitoxin systems, most are proteinaceous, and they are targeted towards DNA and cytosolic targets. However, these TA systems are also known to be involved with biofilms and persister cell formation by creating a dormancy state for the bacterial cell (Singh, Yadav, Ghosh, & Rathore, 2021). These genes could be involved in mechanisms but not directly with the degradation of the tight junctions between human intestinal epithelial cells. The remaining genes were another transposase and several DUF proteins. As described before, transposons are DNA-targeted and cannot be involved in mechanisms. The DUF protein domains, on the other hand, are surface-targeted and could have some involvement; however, the domains they were linked with could not be identified at the time of writing and, therefore, must be kept as tentative gene targets for later testing.

Of the 62 genes identified as unique for AGR1487, only 22 were expressed, cell surface targeted, or had an unknown function (Table 6.5-16 marked with ü). These 22 genes represent the best possible candidates for the barrier-disruptive phenotype. **Regions 2 and 4** can be ignored for the two hypothetical proteins, while **Regions 1, 3 and 5** have the most promising gene candidates.

In summary, AGR1487 shows a distinct differential gene expression profile when exposed to MRS compared to M199. The identified regions and genes of interest, particularly those associated with the cell surface and extracellular matrix modifications, provide a robust foundation for further investigation into the barrier-disruptive phenotype of Caco2-2 monolayers. The unique gene candidates, especially those with signal peptides and involvement in peptidoglycan modification, extracellular matrix production, and cell adhesion, merit detailed functional studies to confirm their roles. Future research should focus on the characterisation of these gene products and their mechanisms of action. Additionally, further studies should address the potential persistence of gene products through the TEER assay period and explore the functional behaviour of the strains in different media formats. Detailed functional validation through gene knockout and overexpression experiments, as well as protein localisation studies, will be crucial in confirming the involvement of these genes in the barrier-disruptive phenotype of AGR1487.

6.7. Conclusion

The differential gene expression analysis of AGR1487 revealed more upregulated genes in MRS than in M199, with a notable upregulation of only 97 genes in M199. However, these 97 genes in M199 did not align with the unique genes predicted for AGR1487. This suggests that the barrier-disruptive phenotype of Caco2 monolayers was likely expressed before the TEER assay, consistent with past findings.

Functional analysis through COG categorisation indicated a shift in gene expression from translation, ribosomal structure, and biogenesis in MRS to amino acid transport and metabolism in M199. The lack of ribosomal structure and biogenesis genes in M199 implies that the barrier-disruptive phenotype genes are not expressed in this medium unless latent ribosomes and mRNAs persist from the MRS exposure. The absence of cell cycle control and cell wall biogenesis genes in M199 further supports that these genes are expressed and mobilised before the cells are resuspended in M199.

The genomic analysis identified five regions of interest within the AGR1487 genome. In the MRS medium, genes from these five regions were expressed. In **Region 1**, several candidates for the barrier-disruptive phenotype were found, including genes involved in extracellular matrix production and peptidoglycan modification. **Region 3** contained potential gene candidates, including hypothetical proteins with signal peptides and truncated metallopeptides unique to AGR1487. **Region 5** contained several unique genes, including toxin-antitoxin systems and DUF proteins requiring further functional analysis.

Overall, the results suggest that the barrier-disruptive phenotype of AGR1487 is pre-expressed in MRS and carried over to the TEER assay of Caco-2 monolayers in M199. The identified candidate genes, particularly those involved in cell surface modifications and extracellular matrix production, provide a foundation for future research. Functional validation of these candidates through gene knockout, overexpression, and protein localisation studies will be crucial to confirm their roles and elucidate the mechanisms underlying the barrier-disruptive phenotype of AGR1487.

6.8. References

Albert, I. (2020, March 5). *The Biostar Handbook: 2nd Edition.* Retrieved from https://www.biostarhandbook.com/

Anderson, R. C., Young, W., Clerens, S., Cookson, A. L., McCann, M. J., Armstrong, K. M., & Roy, N. C. (2013). Human oral isolate *Lactobacillus fermentum* AGR1487 reduces intestinal barrier integrity by increasing the turnover of microtubules in Caco-2 cells. *PLoS ONE, 8*(11), e78774. https://doi.org/10.1371/journal.pone.0078774

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. In: *Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.*

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

Burckhardt, R. M., & Escalante-Semerena, J. C. (2020). Small-molecule acetylation by GCN5-related N-acetyltransferases in bacteria. *Microbiology and Molecular Biology Reviews, 84*(2). https://doi.org/10.1128/mmbr.00090-19

Chatterjee, M., Pushkaran, A. C., Vasudevan, A. K., Menon, K. K. N., Biswas, R., & Mohan, C. G. (2018). Understanding the adhesion mechanism of a mucin binding domain from *Lactobacillus fermentum* and its role in enteropathogen exclusion. *International Journal of Biological Macromolecules, 110*, 598-607. https://doi.org/10.1016/j.ijbiomac.2017.10.107

Drickamer, K., & Taylor, M. E. (1998). Evolving views of protein glycosylation. *Trends in Biochemical Sciences, 23*(9), 321-324. https://doi.org/10.1016/S0968-0004(98)01218-8

Favrot, L., Blanchard, J. S., & Vergnolle, O. (2016). Bacterial GCN5-related N-acetyltransferases: From resistance to regulation. *Biochemistry, 55*(7), 989-1002. https://doi.org/10.1021/acs.biochem.5b01269

Guan, L., & Kaback, H. R. (2004). Glucose/sugar transport in bacteria. In W. J. Lennarz & M. D. Lane (Eds.), *Encyclopedia of Biological Chemistry* (pp. 204-207). New York: Elsevier.

Han, B., Sivaramakrishnan, P., Lin, C.-C. J., Neve, I. A., He, J., Tay, L. W. R., . . . Wang, J. (2017). Microbial genetic composition tunes host longevity. *Cell, 169*(7), 1249-1262.e1213. https://doi.org/10.1016/j.cell.2017.05.031

Hirakawa, Y., Sugimoto, S., Tsuji, N., Inamoto, T., & Maeda, H. (2022). Analysis of gene expression profiles of *Enterococcus faecalis* induced by type I collagen. *World Journal of Advanced Research and Reviews, 13*(1), 129-139. https://doi.org/10.30574/wjarr.2022.13.1.0223

Humayun, M. Z., Zhang, Z., Butcher, A. M., Moshayedi, A., & Saier, M. H., Jr. (2017). Hopping into a hot seat: Role of DNA structural features on IS5-mediated gene activation and inactivation under stress. *PLoS ONE, 12*(6), e0180156. https://doi.org/10.1371/journal.pone.0180156

Juers, D. H., Matthews, B. W., & Huber, R. E. (2012). LacZ β-galactosidase: Structure and function of an enzyme of historical and molecular biological importance. *Protein Science, 21*(12), 1792-1807. https://doi.org/10.1002/pro.2165

Kavanagh, K. L., Jörnvall, H., Persson, B., & Oppermann, U. (2008). Medium- and short-chain dehydrogenase/reductase gene and protein families: The SDR superfamily: Functional and structural diversity within a family of metabolic and regulatory enzymes. *Cellular and Molecular Life Sciences, 65*(24), 3895-3906. https://doi.org/10.1007/s00018-008-8588-y

Kurilshikov, A., Wijmenga, C., Fu, J., & Zhernakova, A. (2017). Host genetics and gut microbiome: Challenges and perspectives. *Trends in Immunology, 38*(9), 633-647. https://doi.org/10.1016/j.it.2017.06.003

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Penning, T. M. (2015). The aldo-keto reductases (AKRs): Overview. *Chemico-Biological Interactions, 234*, 236-246. https://doi.org/10.1016/j.cbi.2014.09.024

QIAGEN. (2015, June). *QIAGEN Genomic DNA Handbook.* Retrieved from https://www.qiagen.com/us/resources/resourcedetail?id=d2b85b26-16dd-4259-a3a7-a08cbd2a08a3&lang=en

Quinlan, A. R. (2014). BEDTools: The Swiss-army tool for genome feature analysis. *Current Protocols in Bioinformatics, 47*(1), 11.12.1-11.12.34. https://doi.org/10.1002/0471250953.bi1112s47

Sengupta, R., Anderson, R. C., Altermann, E., McNabb, W. C., Ganesh, S., Armstrong, K. M., . . . Roy, N. C. (2015). *Lactobacillus fermentum* AGR1487 cell surface structures and supernatant increase paracellular permeability through different pathways. *MicrobiologyOpen, 4*(4), 541-552. https://doi.org/10.1002/mbo3.260

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE, 11*(10), e0163962. https://doi.org/10.1371/journal.pone.0163962

Shi, T., Aryantini, N. P., Uchida, K., Urashima, T., & Fukuda, K. (2014). Enhancement of exopolysaccharide production of *Lactobacillus fermentum* TDS030603 by modifying culture conditions. *Bioscience of Microbiota, Food and Health, 33*(2), 85-90. https://doi.org/10.12938/bmfh.33.85

Singh, G., Yadav, M., Ghosh, C., & Rathore, J. S. (2021). Bacterial toxin-antitoxin modules: Classification, functions, and association with persistence. *Current Research in Microbial Sciences, 2*, 100047. https://doi.org/10.1016/j.crmicr.2021.100047

Struck, A. W., Thompson, M. L., Wong, L. S., & Micklefield, J. (2012). S-adenosyl-methionine-dependent methyltransferases: Highly versatile enzymes in biocatalysis, biosynthesis, and other

biotechnological applications. *Chembiochem, 13*(18), 2642-2655. https://doi.org/10.1002/cbic.201200556

Sychantha, D., Brott, A. S., Jones, C. S., & Clarke, A. J. (2018). Mechanistic pathways for peptidoglycan O-acetylation and de-O-acetylation. *Frontiers in Microbiology, 9*, 2332. https://doi.org/10.3389/fmicb.2018.02332

Takahashi, H. (2010). Chapter 4 - Regulation of sulfate transport and assimilation in plants. In K. W. Jeon (Ed.), *International Review of Cell and Molecular Biology* (Vol. 281, pp. 129-159). Academic Press.

TRIzol™ Max™ Bacterial RNA Isolation Kit. (2020). Retrieved from https://www.thermofisher.com/document-connect/document-connect.html?url=https%3A%2F%2Fassets.thermofisher.com%2FTFS-Assets%2FLSG%2Fmanuals%2Ftrizolmax_man.pdf

Viola, R. E. (2000). L-aspartase: New tricks from an old enzyme. *Advances in Enzymology and Related Areas of Molecular Biology, 74*, 295-341. https://doi.org/10.1002/9780470123201.ch7

Wang, X., Yao, J., Sun, Y.-C., & Wood, T. K. (2021). Type VII toxin/antitoxin classification system for antitoxins that enzymatically neutralise toxins. *Trends in Microbiology, 29*(5), 388-393. https://doi.org/10.1016/j.tim.2020.12.001

Yu, B., Kanaan, J., Shames, H., Wicander, J., Aryal, M., Li, Y., . . . Setlow, P. (2023). Identification and characterisation of new proteins crucial for bacterial spore resistance and germination. *Frontiers in Microbiology, 14*. https://doi.org/10.3389/fmicb.2023.1161604

Zhang, Y., Park, C., Bennett, C., Thornton, M., & Kim, D. (2021). Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N. *Genome Research, 31*(7), 1290-1295. https://doi.org/10.1101/gr.275647.120

Zhu, F., Wu, R., Zhang, H., & Wu, H. (2013). Structural and biochemical analysis of a bacterial glycosyltransferase. *Methods in Molecular Biology, 1022*, 29-39. https://doi.org/10.1007/978-1-62703-465-4_3

General Discussion

7.1. Discussion

7.1.1. Historical and Taxonomic Background

Limosilactobacillus fermentum is a well-studied species; it was first described as *Lactobacillus fermentum* in Bergey's Manual of Systematic Bacteriology in 1901, according to a publication in 1938 (Pederson, 1938). In the 1987 edition of Bergey's Manual of Systematic Bacteriology, *L. fermentum* is described as an irregular, nonsporing, Gram-positive rod (McClung, 1987). The species was further described in 2004 as a gas-producing heterofermentative, air-tolerant, obligate anaerobic bacterium that produces equimolar amounts of lactic acid, carbon dioxide, and ethanol from the metabolism of hexoses (Adegoke, 2004). Furthermore, the species is commonly described as abundant in milk products, sewage, manure, fermenting vegetables and the mouth and faeces of humans (Zhao et al., 2019). *L. fermentum* has been studied for over 120 years, and due to this research and its connection with food systems, animal hosts and its impact on humans, it has been considered safe to consume (Chapter 2).

7.1.2. Contradictory Evidence: AGR1487's Barrier-Disruptive Phenotype

With regards to safety, *L. fermentum* has been labelled as "Qualified Presumption of Safety" (QPS) by the European Food Safety Authority (EFSA) and listed as a "generally recognised as safe" (GRAS) by the US Food and Drug Administration (FDA) in the United States of America (Leuschner et al., 2010). *L. fermentum* is believed to be safe to consume because "It has no ill-effects, being consumed in a variety of fermented foods by people of different regions and ages" (Zhao et al., 2019). Furthermore, *L. fermentum* was found to have little or no ill effect on healthy elderly humans, and it was even postulated that it could be the source of longevity in these people (Park et al., 2015). In general, these claims are substantiated by a connection with foods like cereal- and vegetable-based fermented foods (Di Cagno et al., 2008; Sanchez, Palop, & Ballesteros, 2000) and, traditionally, *L.*

fermentum has been used for food preservation. Along with food preservation, *L. fermentum* also has mechanisms for preventing or eradicating infections by other pathogenic microorganisms (Alfano et al., 2020). All these features of *L. fermentum* are discussed in detail in Chapter 2, along with *L. fermentum*'s connection with humans, health and its probiotic status.

Given the historical context, the "probiotic" symbiont *L. fermentum* is believed to be safe. However, this is contradicted by the barrier-disruptive phenotype of *L. fermentum* AGR1487, the subject of this PhD dissertation, which can negatively impact intestinal health. The overarching hypothesis of this PhD dissertation was that *L. fermentum* AGR1487 decreased intestinal barrier integrity in TEER assays via a genetically mediated cell surface structure

One of the knowledge gaps is that few publications explore the negative impact of this species but instead focus on its probiotic potential. Is this species as safe as the consensus describes? To date, the literature leans towards the side that claims the species is safe to consume. Another consideration is whether AGR1487's barrier-disruptive phenotype is unique. This phenotype could be ignored if widespread throughout the species, given the evidence surrounding the safe consumption of *L. fermentum*. If unique, how did this strain gain this phenotype, or how did the other strains lose it? At the very least, the barrier disruptive phenotype is not universal, as the phenotype is not expressed by the strain AGR1485. Unfortunately, no other strains have been tested for this phenotype via a transepithelial electrical resistance assay (TEER) and based on the literature review (Chapter 2), the phenotype remains undiscovered in other strains due to this limitation.

7.1.3. Phenotypic Evidence from TEER and Animal Models

In addition to the literature review (Chapter 2), the claim that AGR1487's barrier-disruptive phenotype is rare or potentially unique is partially due to an emerging field of wound healing in animal models. Wound healing is used here because it also highlights an earlier claim that unfavourable phenotypes and some aspects of the species' behaviour highlight phenotypic diversity

when interacting with epithelial surfaces, especially in wound healing, where AGR1487's phenotype does not fit well within the consensus. AGR1487 produces the barrier-disruptive phenotype by interacting with the intestinal epithelial cells (Caco-2 cells) in TEER assays, which was also visualised as diffused staining of the tight junctions between the cells (Rachel C Anderson et al., 2013). Anderson et al. (2013) demonstrate that AGR1487 caused an increase in the turnover of microtubules in Caco-2 cells during the TEER assay, disrupting the monolayer. Both strains also caused inflammation in germ-free rats (Rachel C. Anderson et al., 2016). AGR1487 elicited a stronger inflammatory response than the AGR1485 strain in mono-colonised germ-free rats (Rachel C. Anderson et al., 2016). Sengupta et al. (2015) showed that AGR1487 produces the barrier-disruptive phenotype even when killed, and the bacterial supernatant is not involved in the mechanism disrupting the Caco-2 monolayer in TEER assays.

7.1.4. Comparisons to Wound Healing and Other Strain Effects

In animal models, Golkar et al. (2021) showed that some *L. fermentum* supernatants, when mixed with "Cold Creams" (Cold creams are emulsions of water and fats, typically including mineral oil and beeswax, designed to cleanse and moisturise the skin), increase wound closure speeds by promoting collagenesis and reducing inflammation, ultimately improving the integrity of the skin epithelial layer. It should be noted that AGR1487 supernatants have not been tested in cold creams, but an increase in barrier integrity in response to supernatants was not observed in TEER assays either (Sengupta et al., 2015). Golkar et al. (2021) further postulated that these effects could be caused by cell fractions, cell lysates, short-chain fatty acids (SCFAs), extracellular polysaccharides (EPS), teichoic acid, or proteins (Golkar et al., 2021). These components of a cell lysate's supernatant are considered components of the bacterial cell surface. Similar results were reported for Probiotic Lysate-Loaded Nanogels (Ashoori et al., 2020).

It has also been stated that the "presence of lactobacilli attracted neutrophils and macrophages to the wound site by increasing the concentration of cytokines and chemokines (TNF-α, IFN-γ, IL-4,

IL-6, TGF-β, and matrix metalloproteinases) while preventing infection through pH, bacteriocin production, or the production of undescribed biosurfactants" (Bekiaridou, Karlafti, Oikonomou, Ioannidis, & Papavramidis, 2021). Evidence shows that the inflammation response improves wound healing (Bekiaridou et al., 2021). Unfortunately, these authors did not disclose the strains of *L. fermentum* used in their studies. Still, their strains exhibit the opposite effects than those discovered with AGR1487. During wound healing, the tight junctions are expected to reform and would not be able to do so without microtubules (Glotfelty, Zahs, Iancu, Shen, & Hecht, 2014). The cell surface components of AGR1487 may be responsible for the loss of microtubules, and these findings suggested that more research is required, but this could also be explained by limitations in testing.

7.1.5. Assay Limitations and Methodological Considerations

Some of those limitations were linked to the TEER assay, which prevented parallel analysis with other species and compounds that could have provided the context to unravel this barrier-disruptive phenotype. The limitation is that the specific TEER assays used to study AGR1485 and AGR1487 cannot be directly compared with other published TEER assays with a different design. Given that AGR1485 and AGR1487 are the first *L. fermentum* to be tested using the TEER assay, optimisations were introduced to accommodate the requirements for this species. However, these optimisations lead to a deviation from the more common methodologies, especially those that use foetal bovine serum (FBS). For example, experiments on *L. fermentum* TSF331 also applied the bacterium to the upper compartment but used DMEM as a growth media instead of M199 (Lin et al., 2024). The composition of the growth medium for Caco-2 cells significantly impacts their physiology and biology, affecting monolayer integrity and TEER assay results (Srinivasan & Kolli, 2019). Results obtained by TEER assays using different media, such as M199 and MEM, are not directly comparable, with varying levels of FBS.

However, modifications to the TEER assay were necessary to improve its reliability when co-culturing the CaCo-2 cells with *L. fermentum*. The impact these changes had on the Caco-2 cell

physiology had yet to be explored in detail; other than that, the barrier integrity was maintained for the duration of the assay (Chapter 3). AGR1487 disrupted the integrity of Caco-2 monolayers, resulting in a drop in the electrical resistance of the monolayer applied to the circuit during the TEER assay (Chapter 3). This phenotype was not triggered by the supernatant (Sengupta et al., 2015) and appeared independent of changes to the Caco-2 cell culture media (Chapter 3). Due to the limited number of strains of *L. fermentum* exposed to the modified TEER assays, AGR1487 could only be compared to AGR1485, a strain exposed to the same conditions as AGR1487.

7.1.6. Morphological and Phenotypic Differences Between AGR1485 and AGR1487

Phenotypic diversity in the species is usual, as demonstrated by the comparison during the literature review (Chapter 2). Given that AGR1485 does not produce the barrier-disruptive phenotype, the genetic elements shared between AGR1487 and AGR1485 are unlikely to be the origin of AGR1487's barrier-disruptive phenotype. Understanding the genetic differences between the strains was carried out to expand on the differences between AGR1485 and AGR1487 (Chapter 4).

Phenotypes of these two strains, such as pH and bile salt tolerance differences, were confirmed by OmniLog assays (not shown). In addition, AGR1485 and AGR1487 grew differently on the same solid and liquid growth media. The AGR1487 colonies grew more uniform over the same period as AGR1485 but were much smaller (Chapter 3). AGR1487 colonies were also rigid; they remained intact when picked from the plate and did not resuspend well into liquid culture. AGR1485 grew larger colonies than AGR1487, had a more yellow tone than the white colouring of AGR1487 colonies, and had a paste-like texture (Chapter 3). A similar observation was made when cultures were growing overnight. The resulting pellet of AGR1487 cells was more challenging to resuspend than AGR1485 (Chapter 3).

Given these morphological differences, the strains were compared via electron microscopy to determine if there were any visible structural differences between the strains. Scanning and

transmission electron micrographs revealed that AGR1487 autoaggregated into clusters while AGR1485 did not. AGR1485 had a general morphology that was more in line with what was described by Bergey's Manual of Systematic Bacteriology. The autoaggregation behaviour of AGR1487 explains the colony rigidity on solid media and the resistance to resuspension in liquid culture. However, the electron micrographs did not suggest a mechanism for this autoaggregation behaviour. It was postulated that autoaggregation could be the source of AGR1487's resistance to lytic enzymes (Chapter 3). Following the reasoning that reducing access to surface area could restrict enzyme access to the cell walls and protect the cells towards the cluster's centre, autoaggregation may be a mechanism by which AGR1487 resists lysis by host-derived lytic enzymes like mutanolysin and lysozyme used in Chapter 3. Resistance to lytic enzymes is also a known mechanism for increasing inflammation by reducing detected Microbe-Associated Molecular Patterns (MAMPs) in mammalian intestines (Wiertsema, van Bergenhenegouwen, Garssen, & Knippels, 2021). Future studies could focus on testing these phenotypes to causally link the autoaggregation phenotype to lytic enzyme resistance and animal intestine inflammation.

7.1.7. Metabolic Considerations: pH and Lactic Acid Debate

Sengupta et al. (2015) stated that FBS-induced *L. fermentum* produced excess lactic acid to a concentration capable of harming the Caco-2 cells, which was not reproducible (Chapter 3). The Phenol Red indicator dye used in M199 cell culture media showed that the pH levels remained stable and physiologically normal during the TEER assays. Additionally, overnight pH assays in MEM with FBS and FBS alone did not reveal further changes in pH (Chapter 3). Furthermore, overnight cultures of both strains in all combinations of growth media used for the TEER assays and in FBS alone did not induce a colour change in the indicator dye overnight (Chapter 3). However, if AGR1487 was poorly resuspended during the washing steps, the sample could shift the dye to yellow when resuspended in the cell culture media. This phenomenon was likely due to the non-resuspension of the bacterial cells, which led to the carry-over of spent MRS. Unfortunately, pH titrations were not

carried out to find how much MRS media from the overnight culture of *L. fermentum* it would take to shift the Phenol Red dye to yellow. In cases where incomplete resuspension of AGR1487 was observable, the media could take an hour or more to start showing changes in pH, which could have been misinterpreted as lactic acid production.

These observations provided a dilemma. Was autoaggregation, and thus poor resuspension and MRS carry-over the source of Phenol Red's colour change in the presence of AGR1487, or did AGR1487 exhibit some metabolism in M199 media that AGR1485 did not? To test these hypotheses, the bacteria were inoculated into all combinations of cell cultures again, and the OD_{600} was monitored over 48 hours for any biomass increases (Chapter 3). In addition, both strains were also exposed to pure FBS to test if they could use FBS as a substrate. The experiment also required a positive control. Given that MRS was exclusively used to grow and maintain the bacterial cultures before the TEER assays, it was included as a positive control. Surprisingly, no increase in bacterial biomass was observed in any of the cell culture media formats nor for FBS alone over 48 hours (Chapter 3). This observation suggests that cell culture media may inhibit bacterial growth or does not have the required nutrients to support the growth of *L. fermentum*. In addition, FBS did not provide any additional nutrients to the bacteria or rescue bacteria growth. Similarly, the colour of the Phenol Red in the cell culture media did not change during the experiment (Chapter 3). These results do not support the hypothesis that the strains are actively metabolising during the TEER assay, but could lactic acid still be the driving force behind the phenotype?

As stated before, the pH did not change during the assays. If lactic acid was a product of *L. fermentum* metabolism, there was no evidence that sufficient metabolism took place to shift the pH throughout the TEER assays or supplemental experiments (Chapter 3). It is unlikely that the bacteria were producing lactic acid during the TEER assay based on the transcriptomic results (Chapter 6, discussed later). Furthermore, no increase in biomass in cell culture media supports these findings. Lactic acid is soluble in water, and AGR1487's supernatant did not cause the barrier-disruptive

phenotype discussed earlier. It also stands to reason that human cells have some tolerance toward lactate. Given that the Caco-2 cell line is of human origin, it also has some lactic acid tolerance. Taking all this evidence together, it is therefore unlikely that the barrier disruptive phenotype observed during the TEER assay for AGR1487 is driven by pH or the pH of the supernatant. Therefore, the source of acid that changed the colour of Phenol red in past experiments (Sengupta et al., 2015) is likely due to autoaggregation and poor resuspension, leading to carry-over of spent MRS media.

7.1.8. Genomic Analyses: Assembly, Closure, and Annotation

Given the morphological and phenotypic differences between AGR1487 and AGR1485, confirming that these were members of the same species, was performed in Chapter 4. 16S rRNA gene sequencing confirmed that the two strains were from the same species. Still, this method was unsuitable for unravelling the potential genetic basis of AGR1487's barrier-disruptive phenotype. Several datasets from previous work, along with the physical strains, were available: a 454-Pyrosequencing dataset and an Illumina short-read dataset per strain. Unfortunately, there was no extraction information, and only the most basic library preparation and sequencing information existed. The four datasets were assembled into draft genomes and compared in Chapter 4.

The comparison confirmed that the two datasets were from *L. fermentum*. Furthermore, the assemblies were fragmented, which was expected for genome assemblies created from a short next-generation sequencing system. The Illumina datasets featured better coverage, accuracy, and completion scores, but Pyrosequencing comprised fewer contigs and harboured genes absent in the Illumina assembly. Could combining both data types into a short-read hybrid assembly increase completion and reduce contiguity? The closure of both genomes was undertaken in Chapter 4 to circumvent the problem that draft assemblies need more genetic or genomic information due to their incompleteness and assembly gaps of varying lengths. The genes responsible for AGR1487's barrier disruptive phenotype may reside within one of these missing regions, obfuscating results.

Alignments of the two sequencing technologies revealed that the Illumina and Pyrosequencing genome assemblies map sufficiently to show that both cover the same strains with minimal differences. However, both assemblies suffered from low coverage, and as palindromic repeats are common in both strains, these elements in the genomes prevented the genomes' closure. The assembled datasets were combined into a "hybrid" assembly to improve coverage by merging the best Illumina assembly with the best Pyrosequencing assembly. However, post-assembly statistics and alignment tests revealed that the hybrid assembly was lower quality than the genome assembly that relied on Illumina alone. Instead, regions of the hybrid-assembled genome did not align with the individual assemblies, revealing an unexpected aspect of the data and methodology.

Both datasets comprised short-read sequencing methods, meaning that the fragments assembled by computer programs were all less than 500 base pairs in length. These short-reads could not resolve genome regions that contained palindromic repeats and other repetitive elements. However, the two data types confirmed that a single Illumina or Pyrosequencing dataset alone did not sample the entire genome. However, the Pyrosequencing dataset had to be excluded due to the increased error rate caused by poor base calls and low coverage, which complicated the discovery of genes during downstream analysis.

One of the most significant drawbacks of short-read sequencing technologies like Illumina (a short-read technology) is that computers cannot successfully resolve repetitive elements. These repetitive elements give rise to spurious alternative genome structures that computers cannot solve, called "frayed ropes". The sequence must be terminated at each frayed rope because the actual path cannot be resolved without a guide, usually a reference genome or long-read. With *de novo* assemblies, there are no pre-assembled genomes to guide the new assembly. However, short-read sequencing graphs built with the "De Bruijn graph" are good at estimating the genome size. A solution to the absence of a guide was found, leveraging high molecular weight DNA extractions and long-read sequencing provided by Pacific Biosciences (Pacbio). Genome size estimates using long-read

sequencing often lead to larger than natural estimations, and thus, the genome assemblers create gene duplications. In return, the longer 10kb reads were long enough to sequence through all the frayed ropes, solving the weakness of short-read sequencing.

A hybrid assembly of Pacbio and Illumina data was created for both strains. The data analysis of these hybrid assemblies showed that the paths generated by the assembly algorithms had only one path to completion. No possible alternative genome structures could be found with these data. It remains true that an algorithm can lead to artefacts, and the assemblies had to be validated to a biological sample. In past publications, pulse field gel electrophoresis was performed on both strains. The fragmentation pattern from these results was compared to the digital fragmentation patterns computationally calculated using the same enzyme restriction sites. The *in-silico* analysis revealed that the genome assemblies were correctly structured, were of the right size, and had a circular structure. In addition, the data and additional laboratory work confirmed that neither strain contained any plasmids. The genome assemblies were thus complete, without missing information, and allowed for the search for gene candidate(s) that give rise to the barrier disruptive phenotype.

The accuracy of the PacBio data and Illumina short-reads improved base-call accuracy. The datasets had high coverage and accuracy base-calling, removing ambiguity and improving the open reading frame detection accuracy in their final closed assembly. The genome was annotated in three stages to assess and reduce the chances that genomic details were missed. The first round of annotation was performed by running multiple annotation packages and comparing the discovered open reading frames using the GAMOLA2 pipeline (Altermann, Lu, & McCulloch, 2017). The second was a manual curation of the open reading frames and their annotations created by GAMOLA2 (Chapter 4). Lastly, transcriptomic reads were aligned to annotated genome assemblies to verify that no regions containing read alignments were missing a gene model or annotation. Therefore, the resulting genome assemblies were confirmed complete, robustly annotated, and appeared to need no more information.

7.1.9. Comparative Genomics and Pan-Genome Insight

The first genome comparison was carried out between the two strains AGR1485 (~2,225 kb) and AGR1487 (~1,930 kb), and despite displaying large differences in their genome sizes of approximately 295 kb, the genome alignments revealed that shared genetic information was nearly identical (Chapter 5). However, a two-way comparison between the two strains revealed many points of divergence, preventing narrowing potential regions that could harbour the genes mediating the barrier-disruptive phenotype. More *L. fermentum* genomes were needed for further comparison. The closed genome assemblies for both strains were used to find other strains of equal completeness deposited in public databases. As stated before, incomplete genomes pose the risk of missing information. In addition, the genomic regions that give rise to the barrier-disruptive phenotype are rare or unique. How many genomes would a comparison need to verify genomic regions common between strains?

Finding the minimum number of genomes can be stated in another way. How many genomes does it take for the data to reach convergence? Convergence meant the results would stay the same if more samples were added. A parameter directly tested by the refractory plots customarily done during pan-genome analyses to find core genes. Core genes are a classification of genes universally found in all species members and are commonly believed to be housekeeping genes or to protect or allow survival in a given environment. However, the analysis can only be applied to closed genomes that do not have missing information. Early attempts to use pan-genomics to simplify the search for common elements were eroded because the core genes were randomly missing from various *L. fermentum* draft assemblies due to incomplete sampling and, therefore, would not be considered core genes (Chapter 5). In this situation, only around 400 genes were found to be in common, and these results needed to be more consistent when adding new genomes to the dataset. To correct this, 40 closed genomes of similar quality to AGR1485 and AGR1487 were downloaded from NCBI and

reannotated with GAMOLA2 to create a comparable dataset (Chapter 5). The experiment revealed that complete convergence of 1,101 core genes occurs with a minimum of 17 genomes (Chapter 5).

While the core genes reached convergence, the total number of genes did not, meaning that the pan-genome was open, indicating that the species was actively evolving by including new genes into their genomes. Each genome provided an average increase of 67 new or "unique" genes per genome added to the refractory plot (Chapter 5). Furthermore, the pan-genomic analysis gave some structure to the gene content dataset of the 40 strains. For instance, 1,101 genes were present in all strains, representing 14.4% of the pan-genome. An additional 21.6% of the pan-genome genes were found in more than six but less than thirty-eight strains. Most of the pan-genome (around 64%) was in the cloud or unique gene list, meaning they were present in less than six strains (Chapter 5). The pan-genome left a question: How similar was each genome of *L. fermentum* to the others? Because if the phenotype was unique, common or conserved regions could be ignored from further analysis.

Hashing methods could provide the answer. Hashing is a computational process typically employed to determine if two data file copies are identical. In genomics, hashmaps index k-mers (short nucleic acid sequences) into a database that can serve as a "fingerprint" for that genome and enable a fast lookup of the sequences. A complementary method calculates a genome assembly's average nucleotide identity (ANI) and compares it to the ANI from another genome.

Both methods provided a similarity score but did not offer specific gene or strain level information (Chapter 5). Hashing is also sensitive to small changes and can be stringent, especially with genomes that have homologues or single nucleotide polymorphisms (SNPs). At the same time, ANI calculations will only function if the organism has greater than 70% similarity (Chapter 5). For the 40 strains of *L. fermentum*, both methods agreed that the genomes were around 96.9% similar in an all versus all comparison (Chapter 5). If the average number of predicted genes per strain was 2,147, and the genomes are 3.1% different, then each genome theoretically possesses 66.56 unique

genes (Chapter 5). These results support the pan-genome analysis that each genome provided, on average, 67 unique genes (Chapter 5). If this holds, 96.6% of the genomic content can be ignored, and the search can focus on the roughly 67 genes that are unique or rare found in AGR1487. This could be tested by aligning the data to find structural context and directly examine the genomic regions for loci of interest.

The pan-genome analysis and similar comparisons did not provide structural information about the strains. All-vs-all genome-wide alignments obtained this detail. Chapter 5 only displays six genome-wide alignment combinations of the possible 780 combinations (with all 40 strains). The six genomes selected for that Chapter were randomly selected from across the phylogenetic tree generated for the species to maximise the displayed genomic diversity and prevent biasing the image to a branch of the tree that shows a high degree of similarity. These comparisons revealed that most of each genome was similar in gene identity and order. The regions of similarity or conserved genetic sequences were broken by smaller regions of high variability (Chapter 5). The presence or absence of these high variation regions did not agree between strains, and the location of at least one of these regions had evidence of being translocated and inverted in AGR1485 when compared to AGR1487 and three other strains. In AGR1487 and AGR1485, these regions of high variability always aligned with regions of low GC content compared to the rest of the conserved genome, often called genomic islands (GEI) (Chapter 5). GEIs are considered a horizontal gene transfer (HGT) component and are believed to have a deviating GC content because they arose from another organism (Juhas et al., 2009). The presence of a GC skew on its own is insufficient evidence to conclude an HGT event. Still, the edges of the GCs, coordinated with the edges of high variability regions, offer borders within which it is possible to search for the other signs of HGT events, such as palindromic repeats and transposase genes.

The high variability regions were surrounded by palindromic repeats in AGR1485 and AGR1487. Most GEI clusters harboured transposases and showed lower GC content than conserved

genomic regions. The gene content of the high variation regions in four of the five areas in AGR1487 contained phage-related protein-coding genes. A comparison of the regions in an all vs all analysis found that they appear in several strains with an HGT. For instance, **Region 3** of the AGR1487 genome contains a phage insert not found in AGR1485 but is found in 15 of the 40 strains tested (Chapter 5). However, AGR1487 was the only genome with a predicted complete phage. The same locus in the other 15 strains was missing multiple genes; the remaining genes were not universally shared nor had the exact alignment match, which means this specific region is undergoing reductive evolution. Notably, this integrated phage clustered features a phage-associated, membrane-targeted truncated metallopeptidase gene. This truncated metallopeptidase appears to be eroded in all the other strains that contain this phage or its remnants (Chapter 5).

These high variation regions showed genomic divergence and provided evidence of being created by HGT. However, this approach did not indicate where each unique gene was located within the genome, and the exact identity of each gene was unknown. By creating a custom BLAST database of all AGR1487's genes, the unique or cloud gene from the pan-genome list could be filtered against AGR1487. This method provided the exact identity and genomic coordinate for each cloud gene that matched the AGR1487 BLAST database. AGR1487's 62 unique genes identified in Chapter 5 were aligned back to the genome along with a synteny alignment of regions of similarity and GC content. In every case, the unique genes fell within a region of high variation. This experiment confirmed that most of the genomic content was the same and that the search could be limited to these regions of high variation. Most of AGR1487's unique gene complement was within **Region 1** and **Region 3** of the genome (Chapter 5).

7.1.10. Candidate Genetic Regions for the Barrier-Disruptive Phenotype

Being present and unique does not imply that the gene is expressed during the barrier-disruptive phenotype nor that it is responsible for mediating this phenotype. Given that these are protein-coding genes, their mRNA transcripts can be measured during barrier disruptive phenotype and used to

narrow the list of potential gene candidate(s). Surprisingly, none of the unique genes were expressed during the TEER assay (Chapter 6). Both strains provided few transcripts in M199 media in the presence of the Caco-2 cells. Neither strain had detectable expression of genes associated with cell cycle control, cell division, chromosome partitioning, intercellular trafficking, secretion, or vesicle transport transcripts when exposed to M199 media during the TEER assay (Chapter 6). These results supported the findings reported in Chapter 3, which showed that cell culture media did not support or inhibit bacterial growth.

AGR1487 created the barrier-disruptive phenotype after UV sterilisation, suggesting that the components responsible for the barrier-disruptive phenotype were independent of metabolism and were created before death and persisted into cell culture and the TEER assay. During the transcriptomic analysis, the Caco-2 cell genes expressed during the TEER assay were shared between the two strains (Chapter 6). It was unlikely that these genes would create different phenotypes in the same environment. Therefore, the candidate gene(s) responsible for the barrier-disruptive phenotype were expressed in the MRS media before the TEER assay.

The differential gene expression analysis of both strains revealed that in MRS, phage-associated genes were the most highly upregulated (Chapter 6). In both strains, these phage genes exclusively were within the regions of high variation regions, which coincided with the location of the predicted unique genes. 37 of the 62 unique genes were found to be expressed in Caco-2 cells in MRS (Chapter 6). In AGR1487, of the 37 unique genes, the most highly expressed genes were in **Region 3**, a domesticated phage containing a unique surface-tagged truncated metallopeptidase gene (Chapter 6). However, not all the expressed unique genes could be associated with the barrier-disrupting phenotype.

For instance, in **Region 2**, defined by the genome atlas of AGR1487 in Chapter 6, one of the unique genes included a beta-galactosidase, a well-documented enzyme for digesting lactose (Juers,

Matthews, & Huber, 2012). The beta-galactosidase gene expression was accompanied by its associated *Lac*I transcriptional regulator (Juers et al., 2012). The other unique genes were short-chain dehydrogenases/reductases, an IS5 transposase, and a Gycoside-pentoside-hexuronide (GPH) cation symporter. These last three genes are either targeted to act on DNA or small molecules (Guan & Kaback, 2004; Humayun, Zhang, Butcher, Moshayedi, & Saier, 2017; Kavanagh, Jörnvall, Persson, & Oppermann, 2008). These candidates did not fit the inclusion criteria for the observed phenotype. Therefore, **Region 2** was disregarded except for the lone hypothetical protein.

Region 4 was also disregarded except for a single hypothetical protein-encoding gene. The other genes in this region were not expressed. The remaining genes were DNA-targeted transposases and SAM-dependent methyltransferases. SAM-dependant methyltransferases have been well-studied and understood (Struck, Thompson, Wong, & Micklefield, 2012) and were found to be cytosolic. In addition, no surface signalling sequences were found in or around the genes. However, the region contains a hypothetical protein with one-fold gene expression and could not be ruled out due to a lack of information about its identity. Internal structures of the gene sequence were not predicted to be domains or peptide motifs, leading to a complete lack of annotation. Due to this lack of annotation, this hypothetical protein cannot be ruled out and remains a gene candidate. However, because these two remaining genes from **Regions 2** and **4** did not have high expression levels, were isolated, and did not have the typical surface signals, that would indicate that the resulting gene product was surface-targeted. These two genes were less likely to be involved in the phenotype. Future work should prioritise the other regions instead.

Of the remaining three regions, **Region 5** was less likely to be involved in the barrier-disruptive phenotype than **Regions 1** and **3**. **Region 5** had nine unique genes, of which five had gene expression. As discussed in Chapter 6, the remaining genes and their associated domains were unlikely to be involved in the barrier-disruptive phenotype. However, they were included because they were targeted to the cell surface, unlike the two candidates from regions two and four.

The AGR1487 genome had five regions of high variation, as described in Chapter 5. **Regions 2**, **4**, and **5** were unlikely to be involved in the barrier-disruptive phenotype. Therefore, the gene candidate(s) responsible for the barrier-disruptive phenotype were most likely limited to regions 1 or 3.

Some unique gene hits in **Region 1** of the atlas (Figure 6.5-7) were predicted to be cytosolic gene products, as discussed in Chapter 6, further limiting the search for the most likely gene candidate(s). However, the region contains several genes that could be involved in the unique autoaggregation phenotype of AGR1487. A cluster of unique genes in the region's centre were known to be involved in extracellular matrix production, and their respective gene products are targeted at the cell surface. The first of these belonged to the GNAT family N-acetyltransferase, which is known to be involved in the modification of peptidoglycan and resistance to lytic enzymes and antibiotics (Burckhardt & Escalante-Semerena, 2020; Favrot, Blanchard, & Vergnolle, 2016).

In addition, the cluster of genes also includes two domains of unknown function (DUF). These DUF4422 family genes were accompanied by a glycosyltransferase targeted to the cell surface by an attached DUF domain predicted to have a transmembrane domain structure. When targeted towards the cell surface, their functions include increased adhesion, signalling and cell wall biosynthesis (Drickamer & Taylor, 1998). Improved adhesion has already been shown for AGR1487 (Rachel C Anderson et al., 2013). These glucosyltransferases and DUF proteins would make good targets for future experimentation.

The region also contained an acetyltransferase with a signal peptide. Again, these genes have a known function of modifying peptidoglycan that makes the cell walls of the bacteria convey an increased resistance to lytic enzymes, a phenotype of AGR1487 not expressed by AGR1485 (Sychantha, Brott, Jones, & Clarke, 2018). In addition to the aforementioned cell surface modification genes, the region contains a capsular polysaccharide synthesis gene and an EpsG gene. These genes

produce proteins that generate EPS components of the extracellular matrix during biofilm formation in *L. fermentum* strain TDS030603 (Shi, Aryantini, Uchida, Urashima, & Fukuda, 2014). While systems like these are not unique to *L. fermentum*, this specific EpsG gene is unique to AGR1487. While **Region 1** could be involved in the barrier-disruptive phenotype, it is more likely to be involved in the autoaggregation and lytic enzyme resistance of AGR1487.

Region 3, the most likely source of the barrier-disruptive phenotype, contained a domesticated phage. All the highest upregulated genes from AGR1487 were traced back to this prophage. Several prophage genes are predicted to be cytosolic. Such as the DNA-binding Arm gene, transcriptional regulator, phage anti-repressor gene and transposase genes could be ruled out due to their DNA-binding nature. Seven hypothetical proteins with unknown functions remained, many encoded for predicted signal peptides, implying a role towards the cell surface. The remaining four genes, a truncated ImmA/IrrE metalloendopeptidase, Inositolphosphorylceramide synthase, S74 domain JNU P5 phage tail fibre protein with hydrolase activity and a Mucin Binding Protein (MucBP) domain-containing gene, all contain signal peptides and should be considered. The highly upregulated truncated metalloendopeptidase nested in the middle of the prophage genome is the most immediate candidate for producing the barrier-disruptive phenotype. Metalloendopeptidase gene expression in *Enterococcus faecalis* is stimulated by collagen (Hirakawa, Sugimoto, Tsuji, Inamoto, & Maeda, 2022). However, while the effect the metalloendopeptidase has on tight junctions remains unknown, there is a reported interaction between this protein and the host. The MucBP domain is documented to increase the adhesion of bacteria to intestinal mucin (Chatterjee et al., 2018). **Region 3** contained 11 possible gene candidates for the barrier-disruptive phenotype, which represents the best opportunity to unravel its mechanism of action.

7.1.11. Implications and Broader Significance

The identification of a domesticated phage within **Region 3** of *Limosilactobacillus fermentum* AGR1487, particularly the upregulated truncated metalloendopeptidase gene, suggests a potential

mechanism underlying the observed barrier-disruptive phenotype. This discovery challenges the prevailing perception of *L. fermentum* as a universally beneficial and safe probiotic. While *L. fermentum* is widely regarded as beneficial and safe for human consumption, the presence of pathogenic traits in certain strains could have significant implications. Identifying and characterising such pathogenic loci is essential for understanding their role in disease, developing effective preventive measures, and mitigating potential health risks. Furthermore, elucidating the genetic basis of pathogenicity provides critical insights into microbial evolution, enabling targeted approaches to eliminate deleterious alleles. This knowledge could inform the development of safer probiotics, advance biotechnological applications, and contribute to medical research. Additionally, these findings may influence regulatory frameworks governing probiotic use and microbial safety assessments. Given the propensity of bacteria to exchange genetic material, the dissemination of virulence-associated loci is a pressing concern, highlighting the importance of systematic genomic surveillance to detect and monitor these elements.

7.2. Strengths and limitations

The computational approach applied in this thesis demonstrated significant strength in its reproducibility, not only for *L. fermentum* but also for other species due to the non-specific methodology. The genomic analysis relies on pre-existing datasets available through NCBI, eliminating the need for additional laboratory-based *in vitro* work. This computational methodology provided invaluable insights, narrowing down the gene candidates likely responsible for the barrier-disruptive phenotype. However, one of the drawbacks is the high computational resource requirements of this strategy.

A key limitation of the methodology is the need for more causal evidence. Although knockout and complementation experiments were initially part of the project design, they could not be executed

within the given time frame, and the disruption brought about by the COVID-19 pandemic. This underscores a broader challenge in computational biology: bridging the gap between *in silico* predictions and experimental validation. While computational tools are potent and flexible, their outputs require laboratory confirmation to ensure reliability and biological relevance. Without this validation, computational models remain speculative.

Some experimental work had already been completed prior to this study, providing a valuable foundation for future investigations. This overlap highlights the importance of integrating computational and laboratory approaches to maximise the impact and accuracy of genomic research.

7.3. Future work

Future research should uncover the genetic and molecular bases of AGR1487's barrier-disruptive phenotype and autoaggregation behaviour. The most immediate step involves targeted gene knockout and complementation studies to identify the gene(s) responsible for these phenotypes. Given that not all the ranked genes were expressed and that the remaining candidates all fall within the regions of variation, deletion experiments should be conducted. All remaining candidate genes can be systematically deleted from AGR1487 using CRISPR-Cas9 / Lambda Red or similar genetic engineering tools. The resulting mutant strains should be tested for their ability to disrupt the barrier integrity of Caco-2 cells. Observing a loss of the phenotype in any of these mutants will pinpoint the involvement of the corresponding gene(s). This strategy also applies to other phenotypes, modes of adaptation such as the autoaggregation phenotype of AGR1487. A ranked list of deletions is provided along with coordinate data in Table 7.3-17 that includes the RNA evidence from Chapter 6. Deletion 1 and 2 should be prioritised as they capture the majority of the candidate genes.

Table 7.3-17: Region deletion table.

DELETION #	START	END	LENGTH	GENES CAPTURED
1	723,640	769,643	46,004	RS03615, RS03620, RS03625, RS03635, RS03640, RS03645, RS03655, RS03660, RS03695, RS03700, RS03795, RS03820, RS03860, RS03875, RS03905
2	133,622	137,534	3,913	RS00580, RS00585, RS00590, RS00595
3	1,572,904	1,573,995	1,092	RS08060, RS08065
4	80,164	80,748	585	RS00340
5	1,714,184	1,714,540	357	RS08770
6	1,248,542	1,248,784	243	RS06245
7	548,139	548,309	171	RS02660
NOTE: THIS TABLE CAPTURES ALL 25 GENES OF INTEREST (WITH "Y" SIGNALS) FROM TABLE 6.5-1: UNIQUE GENES OF AGR1487, WHILE MINIMISING DELETIONS, WHICH ARE ORDERED BY DESCENDING LENGTH.				

Complementation experiments would follow to confirm the role of the identified genes. The inactivated genes could be reintroduced into AGR1487 knockout mutants via a recombinant plasmid with an inducible promoter, enabling controlled expression. The absence of the phenotype without induction and its reappearance upon induction would provide strong evidence for the gene's involvement. Additionally, introducing the same plasmid into AGR1485, a strain lacking the phenotype, could determine whether the gene alone can induce the barrier-disruptive phenotype. If successful, this approach will allow for further analysis of a strain with a well-documented background.

In addition, other strains with the same loci found in AGR1487's **Region 1** and **Region 3** could be obtained. A comparative analysis of these strains' behaviour in TEER assays could provide an additional differential analysis that could inform the mechanism of action and limit the locations for further study, simplifying the number of targets for select deletion outlined above.

Protein characterisation will build upon these findings by isolating and studying the proteins encoded by the identified genes. Expressing and purifying these proteins will facilitate structural analysis, such as X-ray crystallography or cryo-EM, to reveal their mechanisms of action. Functional assays will test whether the purified proteins can directly disrupt the Caco-2 monolayer, providing a mechanistic link between the gene products and the observed phenotype. Interaction studies will also explore how these proteins interact with host cell structures, such as tight junctions or microtubules.

In vivo studies using germ-free rodent models will extend the *in vitro* findings to a physiological context. These experiments will test the knockout and complemented strains to determine whether the barrier-disruptive phenotype or autoaggregation behaviour impacts intestinal health or immune responses. This approach will clarify whether the observed inflammatory responses are due to a single mechanism or the combined effects of both phenotypes.

Further investigations will leverage advanced genomics and proteomics approaches to complement these experimental findings. Transcriptomics studies from cells harvested during the *in vivo* experiments will identify genes and pathways active during AGR1487's interaction with Caco-2 cells, while proteomic analysis will highlight key proteins associated with the barrier-disruptive phenotype. These experiments will ensure a comprehensive understanding of AGR1487's unique behaviours.

Integrating these research strategies can elucidate the genetic, molecular, and functional basis of AGR1487's phenotypes. These studies promise to deepen our understanding of *L. fermentum's* interactions with host systems and may offer insights into potential probiotic or therapeutic applications. This multifaceted approach lays the groundwork for future exploration of these complex and intriguing phenotypes.

7.4. Conclusion

Strain AGR1487 disrupted the barrier integrity between the Caco-2 cells during TEER assays. The results indicated that neither AGR1485 nor AGR1487 grew in cell culture media. AGR1487's barrier-disruptive phenotype occurred whether the bacterial cells were alive or dead. Past research also showed that the supernatants from both strains did not affect the Caco-2 cells either. These data indicated that the barrier-disruptive phenotype was likely mediated by a cell surface structure that persisted past the death of AGR1487 cells.

These results were supported by transcriptomic evidence, which indicated that both strains had similar transcriptomics profiles in cell culture media, and there was a lack of gene expression associated with cell growth, division, or metabolism. Given that the same genes in AGR1487 were expressed in AGR1485 during the TEER assay, these genes were unlikely to be involved in the barrier-disruptive phenotype.

The transcriptomic and pH data and past results ruled out a mechanism that relied on changes in pH, excreted molecules, carry-over from bacterial cell broth supernatant to cell culture or metabolic products as mechanism(s) of action. Therefore, the mechanism is more likely related to the AGR1487 cell surface. The genomes of the entire *L. fermentum* species were highly similar (96.9%). Given that AGR1485 did not express the barrier disruptive phenotype, regions that AGR1485 had in common with AGR1487 and other strains could be ruled out to carry the causative genetic element. Genomic alignments revealed the location of regions of divergence between strains. Some of these regions of divergence appeared to be caused by HGT, but universally, they all contained unique or rare gene content.

A closer analysis of the unique or rare genomics and transcriptomics data highlighted 17 gene candidates potentially responsible for the barrier-disruptive phenotype, which is found mainly in two regions of the AGR1487 genome (**Region 1** and **Region 3**). All of these genes show expression during

normal growth in the MRS medium, and the gene products would be present after the death of the bacterial cell. These two regions have candidates potentially involved in AGR1487's higher intestinal retention, unusual autoaggregation, and barrier disruption phenotype during the TEER assay.

These gene candidates are not unique to AGR1487 but are found in numerous other strains. The loci in these other strains revealed that they are undergoing reductive evolution. This process eroded the genes within these causative loci, changing the gene content. For instance, the most highly upregulated region in AGR1487 (**Region 3**) appeared in other strains. However, AGR1487 had the longest gene alignments of any other strain and the highest number of genes between the palindromic repeats that surrounded the region. Where a gene from this locus was found in another strain, the gene order, identity and orientation were conserved. Thus, AGR1487 contained the most complete domesticated phage at locus "**Region 3**" of any strain. In **Region 3**, AGR1487 contains a strain specific metallopeptidase that is eroded or lost in all the other strains with this locus. This truncated metallopeptidase stands out as the most likely candidate for the barrier-disruptive phenotype and should be the focus of future research before exploring **Region 1**.

These results progress the understanding of AGR1487 and indicate the regions that could give rise to this barrier-disruptive phenotype. There is sufficient evidence that *L. fermentum* AGR1487 decreased intestinal barrier integrity in TEER assays via a genetically mediated cell surface structure created by gene(s) found in **Region 1** or **Region 3** of AGR1487's genome.

7.5. References

Adegoke, G. (2004). *Understanding food microbiology* (2nd ed.). Alleluia Ventures.

Alfano, A., Perillo, F., Fusco, A., Savio, V., Corsaro, M. M., Donnarumma, G., . . . Cimini, D. (2020). *Lactobacillus brevis* CD2: Fermentation strategies and extracellular metabolites characterisation. *Probiotics and Antimicrobial Proteins, 12*(4), 1542-1554. https://doi.org/10.1007/s12602-020-09678-8

Altermann, E., Lu, J., & McCulloch, A. (2017). GAMOLA2, a comprehensive software package for the annotation and curation of draft and complete microbial genomes. *Frontiers in Microbiology, 8*, 346. https://doi.org/10.3389/fmicb.2017.00346

Anderson, R. C., Ulluwishewa, D., Young, W., Ryan, L. J., Henderson, G., Meijerink, M., . . . Roy, N. C. (2016). Human oral isolate *Lactobacillus fermentum* AGR1487 induces a pro-inflammatory response in germ-free rat colons. *Scientific Reports, 6*, 20318-20318. https://doi.org/10.1038/srep20318

Anderson, R. C., Young, W., Clerens, S., Cookson, A. L., McCann, M. J., Armstrong, K. M., & Roy, N. C. (2013). Human oral isolate *Lactobacillus fermentum* AGR1487 reduces intestinal barrier integrity by increasing the turnover of microtubules in Caco-2 cells. *PLOS ONE, 8*(11), e78774. https://doi.org/10.1371/journal.pone.0078774

Ashoori, Y., Mohkam, M., Heidari, R., Abootalebi, S. N., Mousavi, S. M., Hashemi, S. A., . . . Gholami, A. (2020). Development and in vivo characterization of probiotic lysate-treated chitosan nanogel as a novel biocompatible formulation for wound healing. *BioMed Research International, 2020*(1), 8868618. https://doi.org/10.1155/2020/8868618

Bekiaridou, A., Karlafti, E., Oikonomou, I. M., Ioannidis, A., & Papavramidis, T. S. (2021). Probiotics and their effect on surgical wound healing: A systematic review and new insights into the role of nanotechnology. *Nutrients, 13*(12), 4265. Retrieved from https://www.mdpi.com/2072-6643/13/12/4265

Burckhardt, R. M., & Escalante-Semerena, J. C. (2020). Small-molecule acetylation by GCN5-related N-acetyltransferases in bacteria. *Microbiology and Molecular Biology Reviews, 84*(2). https://doi.org/10.1128/mmbr.00090-19

Chatterjee, M., Pushkaran, A. C., Vasudevan, A. K., Menon, K. K. N., Biswas, R., & Mohan, C. G. (2018). Understanding the adhesion mechanism of a mucin binding domain from *Lactobacillus fermentum* and its role in enteropathogen exclusion. *International Journal of Biological Macromolecules, 110*, 598-607. https://doi.org/10.1016/j.ijbiomac.2017.10.107

Di Cagno, R., Surico, R. F., Siragusa, S., De Angelis, M., Paradiso, A., Minervini, F., . . . Gobbetti, M. (2008). Selection and use of autochthonous mixed starter for lactic acid fermentation of carrots, French beans or marrows. *International Journal of Food Microbiology, 127*(3), 220-228. https://doi.org/10.1016/j.ijfoodmicro.2008.07.010

Drickamer, K., & Taylor, M. E. (1998). Evolving views of protein glycosylation. *Trends in Biochemical Sciences, 23*(9), 321-324. https://doi.org/10.1016/S0968-0004(98)01228-5

Favrot, L., Blanchard, J. S., & Vergnolle, O. (2016). Bacterial GCN5-related N-acetyltransferases: From resistance to regulation. *Biochemistry, 55*(7), 989-1002. https://doi.org/10.1021/acs.biochem.5b01269

Glotfelty, L. G., Zahs, A., Iancu, C., Shen, L., & Hecht, G. A. (2014). Microtubules are required for efficient epithelial tight junction homeostasis and restoration. *American Journal of Physiology-Cell Physiology, 307*(3), C245-C254. https://doi.org/10.1152/ajpcell.00336.2013

Golkar, N., Ashoori, Y., Heidari, R., Omidifar, N., Abootalebi, S. N., Mohkam, M., & Gholami, A. (2021). A novel effective formulation of bioactive compounds for wound healing: Preparation, in vivo characterization, and comparison of various postbiotics cold creams in a rat model. *Evidence-Based Complementary and Alternative Medicine, 2021*(1), 8577116. https://doi.org/10.1155/2021/8577116

Guan, L., & Kaback, H. R. (2004). Glucose/sugar transport in bacteria. In W. J. Lennarz & M. D. Lane (Eds.), *Encyclopedia of Biological Chemistry* (pp. 204-207). New York: Elsevier.

Hirakawa, Y., Sugimoto, S., Tsuji, N., Inamoto, T., & Maeda, H. (2022). Analysis of gene expression profiles of *Enterococcus faecalis* induced by type I collagen. *World Journal of Advanced Research and Reviews, 13*(1), 129-139.

Humayun, M. Z., Zhang, Z., Butcher, A. M., Moshayedi, A., & Saier, M. H., Jr. (2017). Hopping into a hot seat: Role of DNA structural features on *IS5*-mediated gene activation and inactivation under stress. *PLOS ONE, 12*(6), e0180156. https://doi.org/10.1371/journal.pone.0180156

Juers, D. H., Matthews, B. W., & Huber, R. E. (2012). *LacZ* β-galactosidase: Structure and function of an enzyme of historical and molecular biological importance. *Protein Science, 21*(12), 1792-1807. https://doi.org/10.1002/pro.2165

Juhas, M., van der Meer, J. R., Gaillard, M., Harding, R. M., Hood, D. W., & Crook, D. W. (2009). Genomic islands: Tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiology Reviews, 33*(2), 376-393. https://doi.org/10.1111/j.1574-6976.2008.00136.x

Kavanagh, K. L., Jörnvall, H., Persson, B., & Oppermann, U. (2008). Medium- and short-chain dehydrogenase/reductase gene and protein families: The SDR superfamily: Functional and structural diversity within a family of metabolic and regulatory enzymes. *Cellular and Molecular Life Sciences, 65*(24), 3895-3906. https://doi.org/10.1007/s00018-008-8588-y

Leuschner, R. G., Robinson, T. P., Hugas, M., Cocconcelli, P. S., Richard-Forget, F., Klein, G., . . . Richardson, M. (2010). Qualified presumption of safety (QPS): A generic risk assessment approach for biological agents notified to the European Food Safety Authority (EFSA). *Trends in Food Science & Technology, 21*(9), 425-435.

Lin, J.-H., Lin, C.-H., Kuo, Y.-W., Liao, C.-A., Chen, J.-F., Tsai, S.-Y., . . . Ho, H.-H. (2024). Probiotic *Lactobacillus fermentum* TSF331, *Lactobacillus reuteri* TSR332, and *Lactobacillus*

plantarum TSP05 improved liver function and uric acid management—A pilot study. *PLOS ONE, 19*(7), e0307181. https://doi.org/10.1371/journal.pone.0307181

McClung, L. S. (1987). *Bergey's Manual of Systematic Bacteriology, Volume 2.* Edited by Peter H. A. Sneath with Nicholas S. Mair and M. Elisabeth Sharpe. The Williams & Wilkins Co., Baltimore, 1986, pp. 962-1599. *International Journal of Systematic Bacteriology, 37*(1), 91-91. https://doi.org/10.1099/00207713-37-1-91

Park, J.-S., Shin, E., Hong, H., Shin, H.-J., Cho, Y.-H., Ahn, K.-H., . . . Lee, Y. (2015). Characterisation of *Lactobacillus fermentum* PL9988 isolated from healthy elderly Koreans in a longevity village. *Journal of Microbiology and Biotechnology, 25*(9), 1510-1518.

Pederson, C. S. (1938). The gas-producing species of the genus *Lactobacillus*. *Journal of Bacteriology, 35*(2), 95-108.

Sanchez, I., Palop, L., & Ballesteros, C. (2000). Biochemical characterisation of lactic acid bacteria isolated from spontaneous fermentation of 'Almagro' eggplants. *International Journal of Food Microbiology, 59*(1-2), 9-17.

Sengupta, R., Anderson, R. C., Altermann, E., McNabb, W. C., Ganesh, S., Armstrong, K. M., . . . Roy, N. C. (2015). *Lactobacillus fermentum* AGR1487 cell surface structures and supernatant increase paracellular permeability through different pathways. *MicrobiologyOpen, 4*(4), 541-552. https://doi.org/10.1002/mbo3.260

Shi, T., Aryantini, N. P., Uchida, K., Urashima, T., & Fukuda, K. (2014). Enhancement of exopolysaccharide production of *Lactobacillus fermentum* TDS030603 by modifying culture conditions. *Bioscience of Microbiota, Food and Health, 33*(2), 85-90. https://doi.org/10.12938/bmfh.33.85

Srinivasan, B., & Kolli, A. R. (2019). Transepithelial/transendothelial electrical resistance (TEER) to measure the integrity of the blood-brain barrier. In T. Barichello (Ed.), *Blood-Brain Barrier* (pp. 99-114). New York, NY: Springer New York.

Struck, A. W., Thompson, M. L., Wong, L. S., & Micklefield, J. (2012). S-adenosyl-methionine-dependent methyltransferases: Highly versatile enzymes in biocatalysis, biosynthesis, and other biotechnological applications. *Chembiochem, 13*(18), 2642-2655. https://doi.org/10.1002/cbic.201200556

Sychantha, D., Brott, A. S., Jones, C. S., & Clarke, A. J. (2018). Mechanistic pathways for peptidoglycan O-acetylation and de-O-acetylation. *Frontiers in Microbiology, 9*, 2332. https://doi.org/10.3389/fmicb.2018.02332

Wiertsema, S. P., van Bergenhenegouwen, J., Garssen, J., & Knippels, L. M. J. (2021). The interplay between the gut microbiome and the immune system in the context of infectious diseases throughout life and the role of nutrition in optimising treatment strategies. *Nutrients, 13*(3). https://doi.org/10.3390/nu13030886

Zhao, Y., Hong, K., Zhao, J., Zhang, H., Zhai, Q., & Chen, W. (2019). *Lactobacillus fermentum* and its potential immunomodulatory properties. *Journal of Functional Foods, 56*, 21-32. https://doi.org/10.1016/j.jff.2019.02.044

Appendix: Statements of Contribution

Included in this appendix are the DRC 16 forms describing the contribution of published material to this thesis.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Marc Alex Bailie
Name and title of main supervisor:	Professor Warren McNabb
In which chapter is the manuscript/published work?	Chapter 4
What percentage of the manuscript/published work was contributed by the student?	25%

Describe the contribution that the student has made to the manuscript/published work:

The candidate conducted research, wrote, and edited a section of the book chapter. All authors participated in revising the manuscript and approved the final submitted version.

Please select one of the following three options:

- (●) **The manuscript/published work is published or in press**

  Please provide the full reference of the research output:

  Altermann, E., Bailie, M., Fraser, Karl., E., Young, W. (2021). NexGen Sequencing Data: Bioinformatic Tools for Visualisation and Analysis. Elsevier, Comprehensive Foodomics, pages 47 - 90. https://doi.org/10.1016/B978-0-12-816395-5.00001-0

- ( ) **The manuscript is currently under review for publication**

  Please provide the name of the journal:

- ( ) **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Marc Bailie — Digitally signed by Marc Bailie Date: 2026.01.17 17:26:21 +13'00'	Main supervisor's signature:	Warren McNabb — Digitally signed by Warren McNabb Date: 2025.02.25 08:42:45 +13'00'

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Marc Alex Bailie
Name and title of main supervisor:	Professor Warren McNabb
In which chapter is the manuscript/published work?	Chapter 4
What percentage of the manuscript/published work was contributed by the student?	75%

Describe the contribution that the student has made to the manuscript/published work:

The candidate procured and quality-controlled the data, then performed genome assembly and annotation before drafting the manuscript. All authors contributed to manuscript revision as well as reading and approving the final submitted manuscript.

Please select one of the following three options:

- (●) **The manuscript/published work is published or in press**

  Please provide the full reference of the research output:

  Bailie, M., Altermann, E., Young, W., Roy, N., & McNabb, W. (2020). Complete Genome Sequence of Lactobacillus fermentum Strain AGR1485, a Human Oral Isolate. Microbiology Resource Announcements, 9(36). doi: 10.1128/mra.00841-20

- ( ) **The manuscript is currently under review for publication**

  Please provide the name of the journal:

- ( ) **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Marc Bailie Digitally signed by Marc Bailie Date: 2026.01.17 17:41:06 +13'00'	Main supervisor's signature:	Warren McNabb Digitally signed by Warren McNabb Date: 2025.02.25 08:46:26 +13'00'

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Marc Alex Bailie
Name and title of main supervisor:	Professor Warren McNabb
In which chapter is the manuscript/published work?	Chapter 4
What percentage of the manuscript/published work was contributed by the student?	75%

Describe the contribution that the student has made to the manuscript/published work:

The candidate procured and quality-controlled the data, then performed genome assembly and annotation before drafting the manuscript. All authors contributed to manuscript revision as well as reading and approving the final submitted manuscript.

Please select one of the following three options:

(●) **The manuscript/published work is published or in press**

Please provide the full reference of the research output:

Bailie, M., Altermann, E., Young, W., Roy, N., & McNabb, W. (2020). Complete Genome Sequence of Lactobacillus fermentum Strain AGR1485, a Human Oral Isolate. Microbiology Resource Announcements, 9(36). doi: 10.1128/mra.00841-20

( ) **The manuscript is currently under review for publication**

Please provide the name of the journal:

( ) **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Marc Bailie — Digitally signed by Marc Bailie Date: 2026.01.17 17:41:30 +13'00'	Main supervisor's signature:	Warren McNabb — Digitally signed by Warren McNabb Date: 2025.02.25 08:41:24 +13'00'

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.