A community-driven genomic investigation of *Helicobacter pylori* isolates from Indigenous communities in the Arctic of Canada.

by

Jeremiah Yarmie

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Medical Microbiology and Infectious Diseases

University of Manitoba

Winnipeg, Manitoba, Treaty 1 Territory

Copyright ©2022 Jeremiah Yarmie

Abstract

Helicobacter pylori is a Gram-negative, flagellate, microaerophilic member of the Epsilonproteobacteria, with a spiral or curved bacilli morphology (Marshall and Warren 1984). The bacterium resides in the human stomach, and infections can span decades, often going undetected due to minor pathogenesis (Kusters et al. 2006). Infections frequency is higher in the Global South, and is associated with low socioeconomic status, education, remoteness, and inaccessibility to healthcare (Hooi et al. 2017; Bruce and Maaroos 2008). This infection is a public health concern because *H. pylori* infection increases the risk of adenocarcinoma of the gastric mucosa and gastric B-cell mucosal-associated lymphoid tissue (MALT) lymphoma, as well as peptic ulcer disease (Haley and Gaddy 2015). Indigenous communities in Canada experience increased rates of *H. pylori* infection, which is associated with poverty, crowded living conditions, and inaccessible health care. Infection is a concern for Indigenous communities given the bacterium's implication in the development of associated cancers, which is a severe health burden that too disproportionately affects Indigenous peoples. This project builds upon ongoing community-driven research by the Canadian North *H. pylori* (CAN*Help*) Working Group investigating the impact of *H. pylori* infection. In this research project we present the sequence *H. pylori* genomes of 57 CAN*Help* isolates sampled from the Western Canadian Arctic communities of Aklavik, Inuvik, Old Crow, Ross River, Teslin, and Fort McPherson and a comparator cohort of 73 southern Manitoban isolates. I developed and implemented a contamination detection pipeline to ensure only *H. pylori* sequence reads were used in genome assembly. The genomes were analyzed for their presence of: known markers associated with antimicrobial resistance to clarithromycin, metronidazole, levofloxacin, and amoxicillin; virulence factors like *cagA*, the *cag*PI, *vacA*, outer membrane proteins such as *babA*, and the Type I toxin-antitoxin system

aapA1/*isoA1*; and the presence of mobile genomic elements such as plasmids, prophages, and genomic islands. The phylogenetic structure of these isolates was determined and explored, including determining genome sequence type, admixture, and patterns based on host household structures. Manitoban genomes were predominantly European in sequence type and were more likely to carry genetic features associated with virulence and antimicrobial resistance. Several distinct clades of CAN*Help* genomes were identified. Most notable were clades containing isolates belonging to a North American Indigenous sequence type. These genomes were less likely to carry genetic features associated with virulence and antimicrobial resistance. They were also smaller in genome length and in the number of coding sequences and were predominantly from Aklavik. This absence of genetic features was especially apparent when these genomes were compared to other CAN*Help* genomes belonging to European and African sequence types, which were more likely to carry these virulence- and resistance-associated features. These results suggest that there may be distinct *H. pylori* strains circulating in CAN*Help* communities with variable impact on hosts, including resistance to treatment. This project represents the first step in fully characterizing nearly 300 *H. pylori* isolates collected by the CAN*Help* Working Group.

Acknowledgements

The research presented in this thesis has been conducted in and uses data taken from people living on Treaty 1, 2, 5, 6, and 11 Territory, as well as unceded territories of First Nations, Metis, and Inuit people. The purpose of a land acknowledgement is to situate ourselves on the grounds where colonialism has and continues to interrupt the lives of Indigenous nations. It is my attempt as a settler scientist to conduct research that may be seen as a small act of reparation for the unspeakable things that have and continue to be done to displace and dispossess the First Peoples of this land.

To my advisors: Drs. Natalie Knox and Morag Graham. Honestly, you two are amazing. I couldn't have asked for a better advisory team. You are both incredible women scientists and I am so happy and glad to have had the privilege to study under your supervision. You are both full of great ideas and have such broad and deep scientific knowledge. But most importantly, you saw me as a human being during my entire project, and always remembered to consider the personal just as much as the science itself.

Thank you so much to my advisory committee, Drs. Richard Sparling, Linda Larcombe, and Grant McClarty. Your guidance was monumental in keeping us grounded and seeing our blind spots in the project. Thank you to the Genomics Core at the National Microbiology Laboratory (NML) for taking 80 odd tubes of DNA and masterfully producing DNA sequences from my 4-year rusty molecular biology products.

Thank you to Dr. Kaisa Thorell for quite literally being the world's expert on *Helicobacter pylori* genomics, and for always making yourself available to us despite being halfway across the world. Thank you to Zilia Muñoz-Ramirez for your amazing scripts. A big thank you goes out to Drs. Janis Geary and Karen Goodman, as well as the rest of the CAN*Help* Working Group, including Sari Valensky, Taylor Cromarty, Lauren Lindsey, Rabail

Shahid, and especially to all of our partner community members in Old Crow, Teslin, Ross River, Inuvik, Aklavik, Fort McPherson, and Tuktoyaktuk.

My graduate studies were funded in part by the Canadian Institute for Health Research through a Project Grant and the National Science and Engineering Research Council through the Visual and Automated Disease Analytics Graduate Training Program.

I would like to take a moment to centre myself as a non-Indigenous scientist conducting research that is about, involves, and has life-and-death stakes for several Northern communities of Inuit and First Nations peoples. I do not come from the communities in the Northwest and Yukon Territories that are burdened by *Helicobacter. pylori*. I believe that the best person to conduct this research is someone from the affected communities. I am not that person. However, I bring to this project a wealth of domain knowledge in the life sciences, professional experience, and expertise in the communication of scientific research, as well as an open heart and mind. I hope that makes me a suitable enough collaborator for the community members who have entrusted the CAN*Help* Working Group with the privilege of carrying out this project to better their communities.

I acknowledge the role that colonization has and continues to play in the problems that our partner communities currently face. *H. pylori* infection and its associated diseases, most notably stomach cancer, would likely not be as acute or pressing of an issue for these Indigenous communities were the plains not cleared, peoples not displaced, sled dogs not shot, children not separated from their families, and languages not forbidden.

It is my hope, as a settler and a scientist, that this research project may stand as a slight act of decolonization. This research project aims to help ensure that Inuit and First Nations living in Arctic Canada can have better, longer, and healthier lives, so they may be in a better position to claim sovereignty.

Table of Contents

Abstract..... 1
Acknowledgements..... 3
Table of Contents..... 5
Chapter 1: Introduction..... 10
1.1 Current study..... 11
1.1.1 Study background and context..... 11
1.1.2 Community-driven research..... 12
1.1.3 The CAN*Help* Working Group..... 14
1.1.4 Data background..... 16
1.1.5 Problem statement..... 20
1.1.6 Study purpose..... 21
1.1.7 Research aims..... 22
1.2 *Helicobacter pylori*..... 23
1.2.1 Discovery..... 23
1.2.2 Relationship with humans..... 24
1.2.3 Relationship with cancer..... 29
1.2.3.1 Outer membrane proteins..... 30
1.2.3.2 Vacuolating toxin: VacA..... 31
1.2.3.3 Cancer-associated genes: *cag*..... 32
1.2.3.4 Toxin-antitoxin systems..... 33
1.2.4 Genomics..... 34
1.2.4.1 Early *H. pylori* sequences..... 34
1.2.4.2 Variability amongst *H. pylori* genomes..... 35
1.2.4.3 *de novo* genome assembly..... 36
1.2.4.4 Addressing *H. pylori* recombination..... 37
1.2.5 Antimicrobial resistance..... 39
1.3 *H. pylori* in the Canadian Arctic..... 40
Chapter 2: Methods..... 43
2.1 Isolate collection..... 44
2.2 DNA extraction..... 45
2.3 Sequencing..... 45
2.4 Quality assessment, contamination detection, and filtering..... 46
2.5 Genome assembly..... 46
2.6 Multi-locus sequence typing..... 47
2.7 Annotation..... 47
2.8 Phylogenomic analysis..... 47
2.8.1 Reference-free phylogenomics..... 47
2.8.2 Overlaying contextual data on phylogenies..... 48
2.9 Characterization..... 48
2.9.1 Antimicrobial resistance..... 48
2.9.2 Virulence factor detection..... 49
2.9.3 Genomic islands..... 54
2.9.4 Plasmid profile..... 54
2.9.4 Bacteriophage presence..... 54
2.10 Comparative genomics..... 54

2.10.1 Core genome analysis..... 54
2.10.2 Genomic admixture and structure analysis..... 55
2.11 Community engagement..... 56
2.11.1 Knowledge exchange..... 56
2.11.2 Data visualization..... 56
2.12 Ethics Approval..... 57
Chapter 3: Results..... 60
3.1 Description of *H. pylori* genomes obtained for the analysis..... 61
3.1.1 Genomes included in current study..... 61
3.1.2 Multi-locus sequence typing..... 69
3.2 Phylogenomics..... 69
3.2.1 Reference-free phylogeny..... 69
3.2.2 Genomic admixture..... 73
3.2.3 Phylogenetic patterns of contextual data..... 77
3.3 Genome characterization..... 80
3.3.1 Antimicrobial resistance..... 80
3.3.2 Virulence factors..... 92
3.3.3 Genomic islands..... 104
3.3.4 Plasmid content..... 110
3.3.5 Bacteriophage content..... 114
Chapter 4: Discussion, Limitations, Future Directions, and Conclusions..... 117
4.1 Discussion..... 118
4.1.1 Genome quality and contamination..... 118
4.1.2 Sequence types and phylogenetic structure..... 120
4.1.3 Multihousehold group and community trends..... 123
4.1.4 Antimicrobial resistance patterns..... 125
4.1.5 *cagA* and *cag*PI..... 129
4.1.6 *vacA*..... 132
4.1.7 Outer membrane proteins..... 134
4.1.8 Toxin-antitoxin *aapA1/isoA1*..... 135
4.1.9 Horizontal mobile elements..... 135
4.1.10 Review of discussion..... 140
4.2 Limitations..... 142
4.2.1 *H. pylori* genomic plasticity..... 142
4.2.2 Sample size, specificity, and sensitivity..... 143
4.2.3 Sample selection..... 143
4.2.4 Community-driven research..... 144
4.2.5 COVID-19 pandemic..... 145
4.3 Future directions..... 146
4.4 Conclusions..... 147
References..... 149
Appendix..... 185

List of Tables

Table 2.1 CagA protein EPIYA motifs……..50

Table 2.2 Primers used for the amplification of *vacA* s and m allelic regions…52

Table 3.1. *H. pylori* isolates that underwent contamination read filtering…….63

Table 3.2 Antimicrobial resistance genotypes present in at least one *H. pylori* genome…… 83

Appendix

Table A2.1. Publicly available *H. pylori* genomes used in the present study…….. 187

Table A3.1 *H. pylori* isolate genomes first sequenced and described in the present study…… 192

Table A3.2 Multi-locus sequence typing for Manitoba and Arctic *H. pylori* isolates... …… 198

Table A3.3 fineSTRUCTURE-determined *H. pylori* sequence types……205

Table A3.4 Presence of phenotypic and genotypic antimicrobial resistance markers in Manitoban and Arctic *H. pylori* isolates……211

Table A3.5 *cagA* presence/absence, EPIYA motif type, and *cag*PI presence/absence ……..213

Table A3.6 *vacA* allele and genotype status by both *in silico* and lab-based PCR… ……... 215

Table A3.7 Genomic Island data dictionary for *H. pylori* genomes in the present study …… 217

List of Figures and Illustrations

Figure 1.1 Partner communities in this research project……. 18

Figure 1.2 The CAN*Help* working group logo……..…… 25

Figure 3.1 Genome length, G-C%, number of contigs for all assemblies first described in this study……..…… 66

Figure 3.2 Sankey diagram displaying the loss of isolates through quality control and final tally of included genomes……..……67

Figure 3.3 *k*-mer based, reference-free phylogeny computed with Mashtree (*k* = 21)..... 71

Figure 3.4 Co-ancestry matrix in the order of Manitoban, Arctic, and global *H. pylori* core genomes……..…75

Figure 3.5 Ancestry admixture plot of core genomes painted with fineSTRUCTURE determined sequence types……..…… 78

Figure 3.6 Phylogenetic tree of Arctic and Manitoba genomes annotated with household- and multihousehold group…… 81

Figure 3.7 Upset plot depicting the intersection of antimicrobial resistance genotype markers in a) Manitoban and b) Arctic *H. pylori* isolates……..85

Figure 3.8 Phylogenetic tree of Arctic and Manitoba isolates annotated with genotypic antimicrobial resistance markers……87

Figure 3.9 Comparison between presence of genomic antimicrobial resistance markers and phenotypic resistance detected in Arctic and Manitoban *H. pylori*… ……..90

Figure 3.10 *cagA* presence/absence, EPIYA motif type, and *cag*PI presence/absence… …… 93

Figure 3.11 Phylogenetic tree of Arctic and Manitoba isolates annotated with *cagA* EPIYA motif type and *cagPI* status……95

Figure 3.12 Phylogenetic tree of Arctic and Manitoba isolates annotated with *vacA* s and m genotype……97

Figure 3.13 Upset plot of *vacA* allele and genotype status by both *in silico* and lab-based PCR……..99

Figure 3.14 Phylogenetic tree of Arctic and Manitoba isolates annotated with presence or absence of the toxin-antitoxin *aapA1*/*isoA1* and OMPs……102

Figure 3.15 Presence of various genomic islands within CAN*Help* and Southern Manitoban *H. pylori*…….. 106

Figure 3.16 Alignment of ICE *tfs3* and *tfs4* containing genomic content in a selected subset of six Manitoba and CANHelp cohort genomes……..108

Figure 3.17 Presence of various plasmids within CAN*Help* and Southern Manitoban *H. pylori*…….. 112

Figure 3.18 Presence of various bacteriophages within CAN*Help* and Southern Manitoban *H.pylori*…….. 115

Appendix

Figure A3.1 Co-ancestry matrix of Manitoban, Arctic, and global *H. pylori* core genomes ordered by phylogenetic signal……203

Chapter 1: Introduction

1.1 Current Study

1.1.1 Study background and context

This research project focuses on the bacterium *Helicobacter pylori*. It is a Gram-negative member of the Epsilonproteobacteria, flagellate, microaerophilic, with a spiral or curved bacilli morphology (Marshall and Warren 1984). In other words, it uses tail-like projections for motility, tolerates air poorly, and has a wiggly, spiral shape.

The most important aspect of *H. pylori* is where it lives. Similar to its closest relatives, *Campylobacter*, *H.* species inhabit animal guts, residing exclusively in the environmental niche of mammalian stomachs. Since its discovery in the 1980s, *H. pylori* has been well characterized as a pathogen that prefers the distal region of the human stomach—the antrum, yet can also be found in the mid-stomach—the corpus, the upper stomach—the fundus, and the upper small intestine—the duodenum (Fagan-Garcia et al. 2019).

H. pylori prevalence is largely a result of impacts on health of social inequity. Infections are higher in the Global South, and are associated with low socioeconomic status, education, and inaccessibility to healthcare (Hooi et al. 2017; Bruce and Maaroos 2008). The prevalence of *H. pylori* varies drastically across geographies and populations, and can be as low as 15% to as high as 85% (McMahon et al. 2016). Prevalence of *H. pylori* is decreased in developed regions that have overcome these burdens such as North America and Europe (Mascellino et al. 2017).

In Canada, the first peoples of the land, as noted by both contemporary scientific research and community observations, are being inequitably burdened by this bacterium. *H. pylori* infection is a health threat to Indigenous communities across the country and has been identified by many of these communities as a public health concern.

An overall trend of improved sanitation and housing in Southern Canada during the 20th century was likely accompanied by a lower prevalence of *H. pylori* infection (Bruce and Maaroos 2008; Burucoa and Axon 2017). Studies vary in their estimation of *H. pylori* prevalence in the Canadian Arctic, but prevalence reported in the literature ranges from 50-95%, which is two to three times higher than the national average (Goodman et al. 2008).

While the majority of individuals infected with *H. pylori* will never develop any symptoms, for some, presence of this stomach bacterium can result in either asymptomatic or symptomatic gastritis—irritation of the stomach, even worse, peptic ulcers, and very rarely gastric cancer—adenocarcinoma of the gastric mucosa. Stomach cancer is one of the most deadly forms of cancer worldwide (Colquhoun et al. 2015). It is also one of the most painful cancers. On top of adenocarcinoma, infection is also associated with gastric B-cell mucosal-associated lymphoid tissue (MALT) lymphoma (Haley and Gaddy 2015). This relationship with cancer led the World Health Organization to classify *H. pylori* as a Class I carcinogen (WHO-IARC 1994). The cumulative evidence supports that *H. pylori* infection is a serious public health concern for inhabitants of Arctic communities such as Aklavik, Northwest Territories.

1.1.2 Community-driven research

Community-driven research adheres to an investigative partnership and collaboration between communities and researchers both from academic and non-academic institutions (Goodman et al. 2019). In Canada, many community-driven research projects focus on the concerns of Indigenous communities, often addressing issues that arose as the result of settler-colonialism. Thus, one goal of conducting community-driven research is to

work towards genuine reconciliation through establishing trust between Indigenous communities and settler research institutions (Finlay et al. 2010).

Community-driven epidemiological research is necessary in order to address health issues facing certain communities but not others (Goodman et al. 2019). In this way, collective concerns regarding a community's health are undertaken by the community alongside academic and non-academic partners. Community-driven research is not prescriptive, as communities identify subjects of research and drive the research questions of projects.

The fundamental aspects of community-based epidemiological research are: developing research questions collaboratively with communities; multi-directional exchange and donation of information and materials to collaboratively generate knowledge that serves as the basis of research data; strict ethical, professional, scientific, contractual, and legal standards of conduct; community review protocols to facilitate consensus amongst all partners; and crediting all members, both community and academic, for their participation (Goodman et al. 2019).

Ultimately, themes of Indigenous self-determination should be entwined into any community-driven research project involving Indigenous communities. Such research projects follow many aspects of the Research Self-Determination framework presented by Brian Schnarch in the *Journal of Aboriginal Health* (2004) and through the Ownership, Control, Access, and Possession (OCAP) framework developed by the First Nations Information Governance Centre (https://fnigc.ca/ocap-training).

Most importantly, this research project stemmed out of community concern about *H. pylori* infection, and perceptions of increased gastric cancer incidence in their communities. Academic researchers, health care providers, and health officials listened to

the concerns of Indigenous communities and worked with them to develop a research project to investigate *H. pylori* infection in the Canadian circumpolar North.

It is also important to acknowledge how established scientific research and output have, in many ways, reproduced the systems and structures of colonialism that we now try to rectify. Historically, scientific studies involving Indigenous peoples resulted in an imbalance of benefit between researchers and community members. In many cases, academics benefitted much more from research involving Indigenous peoples than did the communities themselves. Often research involving, concerning, and about Indigenous peoples has been done without considering the interests of Indigenous communities, failing to engage in informed consent through ineffective research designs and communication, treating Indigenous communities simply as a source of data, making assumptions on behalf of communities about what is best for them, and failing to establish and maintain meaningful relationships (Schnarch 2004). For these reasons, community-driven research is vitally important.

1.1.3 The CANHelp Working Group

Previous community-driven research work done by the Canadian North *H. pylori* (CAN*Help*) Working Group has led to a better understanding of *H. pylori* infection in Northern Indigenous communities. Established between 2006-2008, the group has expanded from research that initiated in Aklavik in 2007, to include Tuktoyaktuk (2011), Fort McPherson (2012), and Inuvik (2017) in the Northwest Territories (NT) and also expanding westward to include Old Crow (2010), Ross River (2016), Teslin (2016), Carmacks (2017), and Pelly Crossing (2017) in Yukon (YT).

As a community-driven research group, CAN*Help* projects adhere to important guiding principles. The development of research questions, the use of data, and the communication of findings are all done collaboratively with the participating communities. These principles drive research questions, projects, data use and sharing, and publication of results for transparency and mutual benefits. The CAN*Help* Working Group research goals are provided in the appendix. The group's guiding principles are presented in Section 2.12.

Research already published by the CAN*Help* Group has characterized *H. pylori* infection as disproportionately affecting Northern Indigenous populations when compared to Southern Canadians. As reported for the first four community projects, conducted in the far northern Beaufort-Delta region of NT and nearby Old Crow, Yukon, the *H. pylori* prevalence in project participants at the time of initial screening was 60%, which is higher than the prevalence below 40% typically seen in Southern Canadians (Fagan-Garcia et al. 2019). CAN*Help* projects have shown that, as a group, participants with current infections have stomach pathology consistent with a relatively high risk of developing gastric cancer.

In a study investigating adherence to treatment in participants of the Aklavik *H. pylori* project, the CAN*Help* group found that self-reported adherence to treatment regimens varied amongst participants. Adherence was positively associated with higher education and age, and men were more likely to have perfect adherence (Lefebvre et al. 2013). Despite this, even with perfect self-reported adherence, *H. pylori* clearance rates in the Canadian Arctic were notably lower than under the same treatment courses in Southern Canada.

Ultimately *H. pylori* infection in these Arctic communities is a multifactorial problem, and requires an interdisciplinary research approach to explore all community concerns. The role of the CAN*Help* Working Group is to conduct community-driven research based on

concerns and priorities established by Indigenous communities and to ensure that this research does not perpetuate colonial power dynamics that have let down many Indigenous communities.

The CAN*Help* research program adheres to principles of the Association of Canadian Universities for Northern Studies (ACUNS 2003) and the ethical requirements for research with Indigenous peoples of the Tri-Council Policy Statement (TCPS2 2014). The CAN*Help* Working Group arose from extensive dialogue among community leaders, health officials, and academic researchers to clarify community priorities and goals. Relationships were developed over time and are maintained as the CAN*Help* Working Group continues to work through its mandate.

The knowledge exchange philosophy employed by the CAN*Help* Working Group is as follows: research develops through the multi-directional flow of information between all parties involved. All parties give input, not only the dissemination of research findings, but also on the initial study design, goals, and objectives at the start of a research project, as well as the best way to implement research activities in the local setting. To participate in this research, the CAN*Help* Working Group required each community to form an *H. pylori* project planning committee comprised of local community leaders to guide the research from start to finish.

1.1.4 Data background

Between 2007 and 2019, CAN*Help* projects has enrolled more than 1421 participants from 9 predominantly Indigenous Northwest Territory and Yukon communities. Clinical information was provided for 1228, of whom 407 underwent endoscopic examination. From these endoscopy participants, more than 274 *H. pylori* isolates were cultured and stored

(biobanked) from biopsy materials. These some-280 samples represent isolates from participants who were 57% women and 95% Indigenous participants, aged 14-77 years, with an mean age of 47 and age median of 48, from the NT communities of Aklavik, Fort McPherson, Tuktoyaktuk, and Inuvik, as well as the YT communities of Old Crow, Ross River, and Teslin.

Figure 1.1 Partner communities in this research project. Image provided by Taylor Cromarty, CAN*Help* Working Group.

Aklavik
Inuvik
Old Crow
Tuktoyaktuk
Fort McPherson
Yukon
Ross River
Whitehorse
Teslin
Northwest Territories
Yellowknife

Demographically, Fort McPherson, Old Crow, and Ross River residents are predominantly Athabaskan First Nations. Tetlit Gwich'in people reside in Fort McPherson. Vuntut Gwitchin people reside in Old Crow. And Ross River is the home of Kaska Dena people. Teslin is an inland Tlingit community. Tuktoyaktuk is a coastal Inuvialuit community. Aklavik residents are predominantly Inuvialuit and Gwich'in. Lastly, about 40% of residents in Inuvik are Inuvialuit and 23% identify as Indigenous, yet most originate from another community.

There is a wealth of contextual data about the participants who contributed these *H. pylori* strains. Isolates came from individuals who provided information on demographics, with 89% reporting additionally on smoking, alcohol use, and health history, including previous *H. pylori* testing and treatment, digestive symptoms, and use of non-steroidal anti-inflammatory drugs (NSAIDs), aspirin, or antacids. For all participants who contributed isolates, CAN*Help* team members generated endoscopic and pathological diagnosis of gastric abnormalities, and for all isolates, virulence genotypes and antimicrobial susceptibility. Medical record data was obtained for 91% of this participant group. There is also data available on household and kin relationships amongst participants. These data will not be the subject of this current study, but will be investigated by the end of the grant funding period.

1.1.5 Problem statement

H. pylori infects about half of individuals worldwide according to published estimates. However, elevated prevalence stems from inequity. Indigenous communities in Canada have an increased prevalence of *H. pylori* infection, which is associated with poverty, crowded living conditions, and reduced access to health care. Infection is a concern

for Indigenous communities in particular given the bacterium's implication in the development of stomach cancer, which is a severe health burden that too disproportionately affects Indigenous Peoples. Community-driven research projects have confirmed that there is a relatively high prevalence of gastric pathology consistent with increased risk of stomach cancer, and that these pathologies are associated with *H. pylori* infection in communities of predominantly Inuit and First Nations residents in the Canadian Arctic.

1.1.6 Study purpose

The goal of this research project is to describe and characterize the genomic features of *H. pylori* strains isolated from Arctic Indigenous communities in Canada. This work has been undertaken with a specific focus on exploring the relation of these strains to geographic and demographic variation in their hosts at a community and household level. We hope that a better understanding of the genetic characteristics of these *H. pylori* strains will help explain the disproportionate health burden from this bacterium in these communities. This analysis was done through genomic comparisons of CAN*Help* strains, which all come from the western Arctic region of Canada, to strains isolated and from the southern region of Manitoba, which will henceforth be referred to as southern Manitoba isolates. The project also attempts to engage in meaningful knowledge exchange around genomics and northern Indigenous health as a means of incorporating community-based study design, participatory research, and decolonization. This project will lead to an enhanced understanding of *H. pylori* genomics and pathogenicity through a community-driven research study based on the health needs and priorities specifically identified by participating Arctic Indigenous communities.

1.1.7 Research aims

1) Characterize the phylogenetic structure of *H. pylori* isolates from Arctic Indigenous communities in relation to:
    a) Isolates from Southern Manitoba
    b) Representative genomes from across the world
    c) Isolates within and between western Canadian Arctic communities
    d) Subgroups of isolates with kinship relationships (*i.e.,* household)
2) Explore how western Canadian Arctic Indigenous *H. pylori* vary in genomic trends in relation to:
    a) Isolates from Southern Manitoba
    b) Isolates from different Indigenous communities within the western Canadian Arctic
    c) Isolates from different western Canadian Arctic Indigenous households/kinship relationships
3) Ensure that participating communities have the opportunity to review and provide input on the presentation of study results through knowledge exchange in line with principles of research self-determination, community-driven study design, and prior and informed consent.

As this is an exploratory, data-driven, hypothesis-generating project, I will refrain from making any hypotheses about *H. pylori* from Manitoba and CAN*Help* communities other than I expect them to be different ancestrally, phylogenetically, and genetically. Otherwise, I will let the data drive the story of this research project.

1.2 *Helicobacter pylori*

1.2.1 Discovery

H. pylori was first described in the early 1980s by Australian physician-researchers Barry Marshall and Robin Warren. They described the new bacterium as a Gram-negative, flagellate, microaerophilic, spiral or curved rod, and noted its phenotypic similarity with *Campylobacter*. Marshall even went as far as self-inoculation in order to help convince their skeptical contemporaries that a bacterium could survive the hostile and acidic environment of the stomach (Warren and Marshall 1983; Marshall et al. 1985; Pincock 2005; Marshall and Adams 2008). For their discovery (and unofficially for Marshall's daring to drink a 'brew' containing two whole plates worth of *H. pylori*), the duo won the Nobel Prize in Medicine in 2005 (Goodman et al. 2007).

Today, *H. pylori* is known as a common resident of the stomach, which is mostly asymptomatic for those infected. For some, carriage of *H. pylori* causes gastric disease and peptic ulcers, and more rarely, stomach cancer.

Once *H. pylori* establishes an infection in the stomach, it induces a mucosal inflammatory response known as gastritis (Cover 2016). Less than a week into his experiment Marshall experienced bloating, bad breath, and non-acidic vomiting (Marshall et al. 1985; Marshall and Adams 2008). Some evidence suggests these may be typical acute symptoms of *H. pylori* colonization (Kusters et al. 2006), but because they are common symptoms of diverse digestive disorders, no body of evidence indicates what proportion of *H. pylori* infection onsets are symptomatic. Furthermore, clinical guidelines do not list any acute symptoms as an indication to test for *H. pylori*. *H. pylori* infection can persist for decades, often asymptomatic the entire duration, and therefore often goes undetected. Marshall himself originally believed his self-inoculation would result in an asymptomatic

infection, but after two weeks his physician wife, Adrienne, demanded that he begin an antibiotic regimen, though *H. pylori* was not detected in gastric biopsies taken at 10 days, so the infection may have resolved before treatment.

Marshall and Warren's 1984 paper described *H. pylori*'s polar flagellation, typically with four or more flagella found at one end of a cell. This bacterium's morphological description is reflected in the logo for the CAN*Help* Working Group (**Figure 1.2**).

1.2.2 Relationship with humans

According to proposed human migration patterns, *H. pylori* may have accompanied humans as they left Africa, meaning that the bacterium may have infected our species for at least 60,000 years, if not longer; some believe that this relationship goes as far back as 120,000 years (Lamichhane et al. 2019). The bacterium is unique in its degree of genetic variation and the extent to which this variation aligns with geography (Moreli et al. 2010; Moodley et al. 2012; Burucoa and Axon 2017)

Figure 1.2 The CAN*Help* working group logo. Source: http://canhelpworkinggroup.ca/

CAN Help
WORKING GROUP

H. pylori strains have been classified by multilocus sequencing typing (MLST) into the following population types: hpAfrica1, hpAfrica2, hpNEAfrica, hpEastAsia2, hpEurope, and hpSahul (Lamichhane et al. 2019). Recently, two Indigenous American subpopulations were described: hspIndigenousNAmerica and hspIndigenousSAmerica (Muñoz-Ramirez et al. 2020). It should be noted that most hspIndigenousNAmerica isolates characterized to date come from CAN*Help* community projects. Variants of genes that are often assessed to classify these different strains include the outer membrane adhesin *horB* and the cytotoxin *vacA*, a known risk factor for developing *H. pylori*-associated cancer (Gutierrez-Escobar et al. 2017). Some authors have speculated that the co-evolution of *H. pylori* with humans has resulted in other distinctive bacterial genotypic variants in response to different host factors, for example location of stomach infection—stomach antrum, fundus, or corpus (Ailloud et al. 2019), as seen amongst different strains (Noto et al. 2018).

Despite evidence that *H. pylori* causes close to 90% of non-cardiac, or distal, gastric cancer, and accounts for roughly 6% of all cancers worldwide, a small fraction of those infected will develop cancer due to variation in strain genetic factors and other gastric cancer risk factors (Mendez-Tenorio et al. 2014). *H. pylori* is also responsible for the development of gastric and duodenal ulcers in some infected people, however duodenal ulcers do not typically co-occur with gastric cancer, which possibly suggests strain-associated specificity to the type of resulting disease (Ahn and Lee 2015), along with distinct co-factors for the distinct disease pathway.

In order to colonize the stomach and establish an infection, *H. pylori* must overcome several obstacles (Kao et al. 2016). Firstly, the bacterium must survive the hostile acidic environment of the stomach. Initially, stomach acid is neutralized through the activity of bacterial urease enzymes, which catalyze the conversion of urease into carbon dioxide and

ammonia. Ammonia is a strong base, and is the potent acid-neutralizing factor in *H. pylori* colonization and persistence. Notably, urease-derived ammonia is also important for amino acid biosynthesis (Kusters et al. 2006), indicating that urease activity continues even after successful stomach colonization.

The *H. pylori* urease cluster includes the main catalytic subunits, *ureA* and *ureB*, an acid-gated urea channel to gain access to urea only in acidic environments, *ureI*, and accessory proteins, which are required for enzyme assembly and activity, *ureE-ureH*. An essential protein cofactor for *H. pylori* is nickel, which is required for the function of urease (Thorell et al. 2017). However, *H. pylori* cannot survive in the low pH environment indefinitely, and must migrate to the gastric epithelium eventually (Kusters et al. 2006).

The next important factor in *H. pylori* colonization is the bacterium's 4-6 flagella, which make it highly motile. Using its flagella, *H. pylori* travels to the gastric epithelium, a region of neutral pH where it can settle in its environmental niche. Chemotaxis towards the gastric mucosa is dependent on urea, as well as sodium and bicarbonate ions (Yoshiyama and Nakazawa 2000). Once the bacterium reaches the epithelium, it attaches to host cells with the use of outer membrane proteins (OMPs), most notably adhesins.

Disease progression resulting from *H. pylori* colonization and infection is a complex result of a combination of factors including host genetics, pathogen biology, and environmental factors such as diet (Haley and Gaddy 2015). For example, smoking and high salt intake are risk factors for gastric cancer (Gaddy et al. 2013).

Carriage of *H. pylori* in the human gut may not always be a negative situation for the host. Some studies have suggested that *H. pylori* colonization may have a few health benefits, including protection against allergic airway disease, gastro-esophageal reflux disease, esophageal cancer, and obesity (Fallone et al. 2000; Arnold et al. 2011; Cover and

Blaser 2009), and even Irritable Bowel Disease (Luther et al. 2010), however it's not clear if confounding and other potential biases have been adequately taken into account in this body of studies. The philosophical belief of a potential role of *H. pylori* in human health maintenance is in line with the "hygiene hypothesis" of human immunity, which posits that when a child's environment is "too clean", lack of exposure to germs does not give the immune system a chance to develop an immunity to microbial diseases (Miftahussurur et al. 2017).

1.2.3 Relationship with cancer

The relationship between *H. pylori* and gastric cancer has long been established. Specifically, *H. pylori* is a risk factor for non-cardia gastric carcinoma—cancer of the stomach excluding the cardia, the region closest to the gastroesophageal junction (Plummer et al. 2015; Colquhoun et al. 2019*). H. pylori* infection is estimated to cause nearly 90% of all non-cardia gastric cancers (Mendez-Tenori et al. 2014).

This relationship is of concern because gastric cancer is one of the most deadly types of cancer worldwide (Colquhoun et al. 2019). Based on 2012 numbers, it is the fifth most common form of cancer worldwide (Colquhoun et al. 2015). Along with adenocarcinoma of the stomach, *H. pylori* is also associated with mucosa-associated lymphoid tissue (MALT) lymphoma (Munoz-Ramirez et al. 2017).

Something to note is that cancer-causing strains of *H. pylori* may not always be present in cases of gastric cancer, because the cancer causes alterations to the gastric mucosa that changes the ecological niche in which the bacterium resided during carcinogenesis (Berthenet et al. 2018). In fact, extensive observations of decreased *H. pylori* prevalence in gastric cancer cases at the time of cancer diagnosis relative to cancer-free

controls suggest that precancerous lesions in the stomach are less hospitable to *H. pylori* than normal gastric tissue.

1.2.3.1 Outer membrane proteins

There are a number of genes and phenotypes that are hypothesized to be causal pathways for *H. pylori*-associated oncogenesis. Some evidence shows that gastric cancer occurs more frequently in people infected with *H. pylori* strains that have the blood-antigen binding gene *babA* (Berthenet et al. 2018), whose protein product mediates *H. pylori* adherence to stomach epithelial cells (Dadaschzadeh et al. 2017). Expression of adhesins like blood group antigen binding adhesin (BabA) are crucial for the bacterium to adhere to the gastric epithelium and maintain their environmental niche (Aspholm-Hurtig et al. 2004). In particular, BabA-mediated adhesion has been shown to be associated with *H. pylori* adherence that allows for the *cag*T4SS (Type 4 secretion system) to translocate CagA into host cells. This binding is also associated with HopQ, which targets host CEACAM1, 3, 5, and 6 glycoproteins. The binding targets of BabA include Lewis B antigen, as well as Globo A and H hexaglycosylceramide (Matsuo, Kido, and Yamaoka 2017).

The *H. pylori* genome encodes for about 60 OMPs, including BabA and sialic acid-binding adhesin (SabA), adhesins that assist the bacterium in binding to the stomach epithelium (Cover 2016). SabA-mediated adhesion to host cells occurs through targeting Sialyl Lewis A and X antigens, as well as Lewis X antigens. The outer membrane protein HopH has been shown to be associated with induction of inflammatory cytokine production, and AlpA/B (also known as HopC/B) are crucial for adhesion to collagen and laminin in the host extracellular matrix (Matsuo, Kido, and Yamaoka 2017).

The majority of *H. pylori* strains are generalist, and are able to bind blood group antigens A, B, and O. Conversely, 60% of investigated isolates from South Americans belonging to hspIndigenousSAmerica obtained from South Americans are specialists, preferentially binding the blood type O antigen, which is increased in Indigenous populations in the Americas (Aspholm-Hurtig et al. 2004). A recent meta-analysis indicated that individuals with the type O blood group have increased odds of *H. pylori* carriage (Chakrani et al. 2018).

1.2.3.2 Vacuolating toxin: VacA

VacA, vacuolating cytotoxin, is a virulence factor of *H. pylori* that is used as an important marker in strain genotyping. The *vacA* gene encodes for a vacuolating toxin that contains subtypes associated with gastric cancer (Akeel at al. 2019). VacA protein causes the formation of pores in host cells, which can lead to autophagy, apoptosis, suppression of T-cell activation, and, as the name suggests, vacuolation of host cell and lysosome membranes (Cao et al. 2016). *vacA* is found in nearly all *H. pylori* isolates and contains three variable regions: m, s, and i. Genotyping of *vacA* is based on allelic differences at these three regions. For example, evidence indicates that the prevalence of *vacA* s1 and m1 alleles, relative to the presence of s2 and m2 alleles, is higher in isolates from people with gastric cancer than it is in people without gastric cancer (Sugimoto et al. 2009). The effects of these different genotypes manifest as a mosaic form, with s1/m1 being the most virulent and producing the most toxin, s1/m2 and s2/m1 combinations producing moderate toxin and resulting virulence, and finally s2/m2 producing negligible toxin and not being virulent (Xue et al 2021).

1.2.3.3 Cancer-associated genes: *cag*

The *cag* pathogenicity island (*cag*PI) is a 40kb mobile element encoding a type IV secretion system. CagE is an ATPase that is essential for proper secretion (Karlsson et al. 2016). The molecular syringe injects CagA into the epithelium of the stomach. A beautiful depiction of the genetic arrangement of the *cag*PI is presented in Figure 1. of Fischer et al.'s 2020 paper. Likewise, Figure 2. of the same paper depicts an assembled model of the *cag* T4SS in relation to the homologous *comB* T4SS. Once uptaken, CagA undergoes tyrosine phosphorylation (Backert and Blaser 2016) by host Src and Abl kinases at repetitive amino acid motifs known as EPIYA motifs (Cover 2016) and causes a downstream host signal transduction cascade (Dadaschzadeh et al. 2017). Evidence suggests that CagA signalling cascades can cause aberrant changes in epithelium cell shape and polarity, disrupt cell junctions, affect cell differentiation, enhance proliferation, inhibit apoptosis, activate gastric stem cells, and degrade p53 tumor suppressor (Cover 2016). These changes can ultimately disrupt tight and adherent junction function, cause a pro-inflammatory response, and alter mitogenesis (Dadaschzadeh et al. 2017). CagA may also trigger many pro-carcinogenic processes including an IL-8-mediated cytokine response, as well as promote host cell proliferation, scattering, and migration (Rizzato et al. 2012).

The *cagA* gene encodes for repetitive amino acid motifs of Glu-Pro-Ile-Tyr-Ala (EPIYA), of which there are four described genotypic variants, A, B, C, and D. EPIYA-C, which is observed most frequently in strains in the Western hemisphere, and EPIYA-D, which is more frequent in Eastern Asia, are both associated with increased gastric cancer incidence/risk and *in vitro* CagA activity (Rizzato et al. 2012).

Owing to evidence that it plays a role in promoting gastric cancer, CagA is known as a bacterial oncoprotein. Variants with a higher number of EPIYA-C motifs are associated with

greater gastric cancer risk (Cover 2016), presumably because EPIYA-C copy number is associated with higher rates of tyrosine phosphorylation (Basso et al. 2008; Ferreira et al. 2012). Likewise, the presence of EPIYA-D motifs is associated with higher gastric cancer risk, possibly due to stronger SHP-2 binding and downstream host signalling cascade effects (Hatakeyama and Higashi 2005). Observations of a relationship between *cagA*-positive *H. pylori* strains and gastric cancer goes back as far as 1995 (Blaser et al.).

Often people living in geographically remote communities will possess *H. pylori* strains that encode attenuated or truncated copies of *cagA* (Delahay, Croxall and Stephens 2018). Epidemiological studies have shown that high dietary salt is an increased risk factor for stomach cancer (Gaddy et al. 2013). Laboratory studies have shown that Mongolian gerbils fed a high-salt diet have increased *cagA* transcription and increased *H. pylori*-associated inflammation and cancer development. The regulation of *cagA* is thought to be influenced by a number of environmental factors including iron, salt, and acid concentration (Gaddy et al. 2013).

1.2.3.4 Toxin-antitoxin systems

The AapA1/IsoA1 toxin-antitoxin system in *H. pylori* is believed to be associated with the bacterium's resistance to treatment. The toxins expressed by this system result in morphological changes that convert spiral-shaped *H. pylori* cells into coccoid morphology. AapA1/IsoA1 is the only identified type I toxin-antitoxin systems identified in *H. pylori*. There are five characterized type II toxin-antitoxin systems also known to occur in the bacterium. Expression of activity of the AapA1 toxin is triggered by oxidative stress, reduces the number of *H. pylori* cells that are culturable, and may be associated with bacterial response

to the bactericidal antimicrobials amoxicillin, levofloxacin, and metronidazole (Krzyżek 2021).

aapA1 encodes a 30-amino acid peptide toxin, the synthesis of which is repressed by the antitoxin RNA IsoA1. The presence of this type I toxin-antitoxin system is conserved in nearly all *H. pylori* isolates examined and are associated with conserved chromosomal loci (Arnion et al. 2017; Krzyżek 2021).

1.2.4 Genomics

1.2.4.1 Early *H. pylori* sequences

The first *H. pylori* genome was published in 1997 and the second was published in 1999. *H. pylori* was the first bacterial species to have two different strains sequenced, allowing for it to be the subject of the first comparative genomic studies in the field (Kurtz et al. 2004). The *H. pylori* genome is about 1.6 Mb, with a G+C content of around 39% (Ahmed et al. 2013). Along with its small size, *the H. pylori* genome encodes only about 1500 open reading frames (van Vliet 2016). Like obligate human pathogens in general, *H. pylori* has undergone reductive evolution over its history, which explains its small genome size and number of coding sequences (Kuipers et al. 2000; Haley and Gaddy 2015). For example, the *H. pylori* genome lacks several biosynthetic pathways for amino acids (Kusters et al. 2006). It has been observed that more virulent strains of *H. pylori,* such as those isolated from people with gastric cancer, have of reduced genome size on average when compared to strains isolated from people with more mild gastritis or peptic ulcers (Fischer et al. 2014).

1.2.4.2 Variability amongst *H. pylori* genomes

The pan-genome refers to the entire gene repertoire of a given species—the overlapping genes shared by different genomes being studied (Page et al. 2015). The core genome refers to genes conserved in all strains and the accessory genome refers to the rest (Uchiyama et al. 2016). The core genome of *H. pylori* is estimated to be 250 to 500 genes depending on the number and types of strains under comparison (van Vliet 2016).

H. pylori is known as an organism with highly variable genomic differences between strains, due in part to its unique ecological niche (Tada et al. 2017). Some argue that its propensity for genetic fluidity, via recombination and mutation, is what allows it to continually reside in such a hostile environment and evade both innate and adaptive host immune responses (Robinson et al. 2005). The majority of genes with sequence variation amongst *H. pylori* strains and isolates encode for OMPs and other surface proteins that interact with the host environment supporting this suggestion (Karlsson et al. 2016).

H. pylori has a high level of genetic diversity due to elevated intraspecies recombination (Cover 2016). The bacterium's high mutation rate may be partly due to the species lacking the MutHLS1 DNA methyl-directed mismatch repair (MMR) system (Moreli et al. 2010). However, *H. pylori* is also competent for transformation and readily integrates small homologous DNA fragments into its genome (Ailloud et al. 2019). The *H. pylori* genome contains plasticity zones—regions of hypervariability that contain about half of *H. pylori*'s accessory genome (Berthenet et al. 2016). These gene clusters form genomic islands in which recombination and mutation is common (Haley and Gaddy 2015).

H. pylori has proven to be a difficult bacterium to work with because of its uniquely plastic genome. Genome sequencing via large-scale parallel pyrosequencing (454 Life Sciences, Branford, Connecticut) of two isolates from inhabitants of Aklavik examined

previously by the CAN*Help* Working Group contained around 140 chromosomal rearrangements relative to each other. Notably, these isolates also lacked the *cag*PI, which is associated with gastric cancer risk (Kersulyte et al. 2015).

H. pylori's high mutation and recombination rates means it is feasible for strains of the bacterium to undergo significant microevolution during single-person infections and within family units (Moreli et al. 2010). It is thought that *H. pylori*'s high recombination rates are high due to the occurrence of co-infection within single hosts (Kennemann et al. 2011). During co-infection, a nucleotide is three times more likely to be altered by recombination than mutation (Linz et al. 2013), and this is significant because the short-term mutation rate of *H. pylori* is already high, estimated by Morelli et al. (2010) to be between 1.46×10^{-6} to 4.56×10^{-6} per nucleotide per year. Linz et al. (2014) characterized the mutation rate at about 1.5×10^{-5} changes per site per year and the recombination rate at about 5.5×10^{-5} per initiation site per year. These estimates put mutation rates in *H. pylori* higher than any other bacterium currently being studied (Robinson et al. 2005; Haley and Gaddy 2015).

Furthermore, some evidence suggests that during the initial acute infection phase, *H. pylori* undergoes a mutation burst, seeing its mutation rate, which is already the highest among all bacteria, increase by yet another order of magnitude (Linz et al. 2014). This mutation rate burst is seen predominantly in OMP genes, which indicates that this increased rate of genetic variation is crucial for adaptation to the current host.

1.2.4.3 *de novo* genome assembly

Given how plastic and variable *H. pylori* genomes are, most researchers try to avoid reference-genome based phylogenetic approaches as much as possible. To mitigate this at the genome assembly stage of genomics analysis, one can use *de novo* genome assembly.

These *de novo* assemblers use de Bruijn graphs to identify read overlap through shared *k*-mers (Bankevich et al. 2012). However, this approach struggles with resolving regions with high densities of repetitive sequences (Heydari et al. 2019). Despite the benefits of *de novo* genome assembly, it rarely results in *de novo* genome assemblies with few large contiguous sequences due to insufficient contig overlap (Carrico et al. 2018). Reference-independent visualization of *H. pylori* aligned genomes have been used to confirm genetic relationships between strains and isolates, and have identified the presence of structural genomic inversions (Tada et al. 2017). Studies have shown that it isn't wise to trust consensus sequences and genomic assemblies of *H. pylori*, even from single colonies, as there may be genetic variability even at the strain level (Draper et al. 2017).

1.2.4.4 Addressing *H. pylori* recombination

The fact that the *H. pylori* genome is plastic and variable due to not only a high mutation rate, but also frequent intraspecies recombination means that recombination must also be considered in bioinformatic analysis to ensure it does not confound phylogenetic analyses. In prokaryotes, acts of homologous recombination influence genomic microevolution and help maintain stable population structures (González-Torres et al. 2019). Several tools exist for predicting recombination events in bacteria. ClonalFrameML uses maximum likelihood inference to determine how recombination events in ancestral individuals shaped bacterial genomes within a given phylogeny (Didelot and Wilson 2015). Gubbins is another maximum likelihood-based tool that iteratively analyzes regions of high single-nucleotide variant (SNV) density to infer ancestral recombination events (Croucher et al. 2015). Both of these approaches have been validated using *H. pylori* as a test organism

and have been used in the literature to conduct recombination detection and masking in *H. pylori* phylogenetic studies.

Rather than employing a recombination detection and masking approach as discussed above, ultimately it was decided for this study to address recombination in the *H. pylori* genome by using a genomic admixture methodology. This approach, as demonstrated by Munoz-Ramirez et al. in 2021, and first developed by Yahara et al. in 2013, determines the amount of genomic admixture present in a given individual "recipient" genome from a collection of curated "donor" genomes. By using this approach, one is able to present the recipient genome painted by donors from various sequence types, inferring geographic signal of origin for *H. pylori* genomes.

1.2.5 Antimicrobial resistance

H. pylori has seen an increase in antibiotic resistance in recent years, and with that an associated increase in failing treatments and eradication therapies (Pohl et al. 2019). In 2017, the World Health Organization listed clarithromycin-resistant *H. pylori* as among the 20 highest priority antimicrobial-resistant bacteria challenging modern medicine (Tacconelli et al. 2017). Antimicrobial resistance has the potential to prolong infections and thereby increase risk of disease progression as a result of treatment failure (Alfizah et al. 2014; Mera et al. 2005; Fuccio et al. 2009), as well as reduced susceptibility to antibiotics taken to treat other conditions.

Antimicrobials used to treat *H. pylori* infection include amoxicillin, tetracycline, clarithromycin, metronidazole, levofloxacin, tinidazole, furazolidone, and rifabutin (McMahon et al. 2016), and resistance to date is predominantly against metronidazole and clarithromycin. *H. pylori* strains are increasingly resisting first-line clinical antibiotic

treatments, as strains are becoming harder to treat with metronidazole and clarithromycin. This is due, in part, to the use of these same antibiotics for the treatment of other bacterial infections (Mascellino et al. 2017). The increasing difficulty to clear *H. pylori* infections is concerning when considering the bacterium's ability to cause cancer.

There are a number of SNVs in various genes (*gyrA, gyrB, rdxA, frxA, rpsU, omp11, dppA*) and 23S rRNA that are thought to be associated with antimicrobial resistance in *H. pylori* (Wang et al. 2018; Camorlinga-Ponce et al. 2021). The majority of mechanisms that provide *H. pylori* antimicrobial resistance arise from mutations in target genes rather than horizontal gene transfer (Camorlinga-Ponce et al. 2021).

Clarithromycin resistance is associated with the following mutations in 23S rRNA V: A2142G, A2142C, A2143G, A2146C, A2146G, A2147G, and T2182C (Pohl et al. 2019; Wang et al. 2018). These mutations to the 23S rRNA cause structural changes in the ribozyme which prevent clarithromycin from interacting (Camorlinga-Ponce et al. 2021). Resistance to clarithromycin has also been associated with efflux pump activity (Iwamoto et al. 2014).

Levofloxacin resistance is thought to be related to mutations in the quinolone resistance-determining region, with amino acid mutations in sites 87, 88, and 91 of GyrA including N87I, N87K, D91G, and D91N. Fluoroquinolones like levofloxacin inhibit DNA gyrase and topoisomerase, interfering with prokaryotic DNA replication (Nishizawa and Suzuki 2014).

Metronidazole resistance is related to inactivation of *rdxA* reductase, often through truncation or by an A118S mutation in the protein. Resistance to amoxicillin is associated with N562Y and T556S coding mutations in the penicillin-binding protein (PBP) 1A (De Palma et al. 2017). Amoxicillin binds to PBP and prevents peptidoglycan synthesis, ultimately disrupting the bacterial cell wall and causing cell lysis (Nishizawa and Suzuki 2014).

However, amoxicillin resistance rates in *H. pylori* remain low worldwide (Camorlinga-Ponce et al. 2021).

Camorlinga-Ponce and colleagues (2021) hypothesized that the low rates of antimicrobial resistance found in *H. pylori* isolates from Mexican Indigenous communities were a result of decreased antimicrobial usage and exposure. They also found that the convergence of phenotypic resistance and presence of genomic markers had moderate correlation, with the weakest correlation belonging to metronidazole resistance.

1.3 *H. pylori* in the Canadian Arctic

Worldwide, *H. pylori* infection is associated with low socioeconomic status, low education levels, and marginalized ethnicity and race. These risk factors for *H. pylori* infection are relatively more common in northern communities compared to those living in other regions of Canada.

In developed countries such as Canada, health disparities exist in the wake of colonialism. Indigenous peoples living on reserves, in northern regions, and in the Arctic, experience poorer health outcomes when compared to major urban regions (Young et al. 2016). This disparity characterizes *H. pylori* epidemiology. A 1999 study found that 95% of the 306 adult participants from Wasagamack Cree Nation in Northern Manitoba were positive for *H. pylori* (Hastings et al. 2014). In contrast, presence of the bacterium is much lower in Southern Canadian urban areas (Lefebvre et al. 2013). When compared to global rates, Inuit, Athabaskan, and Dene peoples living in the Canadian Arctic have higher cancer rates at various sites, including stomach cancer (Young et al. 2016).

Inuit Nunangat consists of 4 Inuit land-claim regions: Nunavut; Nunavik, in northern Quebec; Inuvialuit, the north-western part of the Northwest Territories; and Nunatsiavut,

on the northern coast of Labrador. The life expectancy of Inuit Nunangat is more than a decade lower than other Canadians. Inuit Nunangat experience a higher incidence of nasopharynx, lung and bronchus, colorectal, stomach, kidney, and renal pelvis cancer (Carriere et al. 2012). In a study of Native Alaskans, *H. pylori*-positive people were found to be 16 times more likely to have moderate to severe acute gastritis, and 4 times more likely to have chronic gastritis (Nolen et al. 2018). Relative to the Canadian average, Indigenous populations in Northwest Territories and Yukon are disproportionately affected by gastric cancer, especially non-cardia forms (Colquhoun et al. 2019), consistent with an elevated presence of *H. pylori* infection. These two different cancer subsites reflect different disease origins, with non-cardiac forms of gastric cancer being strongly associated with *H. pylori* infection (Colquhoun et al. 2015). Gastric cancer is the second most commonly diagnosed type of cancer in Inuit men, and third most in Dene men (Cheung et al. 2008).

The prevalence of *H. pylori* in the Northwest Territories has been studied for over a decade by the CAN*Help* Working Group, which began testing NT residents for *H. pylori* in 2008, as part of the the Aklavik *H. pylori* Project. Aklavik is a community of predominantly Inuvialuit and Gwich'in Dene people, located on the Peel Channel of the Mackenzie River Delta. The hamlet's motto is "Never Say Die", dating back to the 1960s when many in the community refused the federal government’s effort to relocate them some 50km east to a new town that became present-day Inuvik (Cheung et al. 2008), established in 1958.

Residents of Aklavik have long been concerned about *H. pylori* infection and the potential for it to develop into gastric cancer in its community members. Previous work done by the CAN*Help* Working Group has shown that *H. pylori* is highly endemic in Aklavik (Cheung et al. 2014). The Aklavik *H. pylori* Project, which enrolled over 60% of Aklavik’s population of roughly 600, had a number of findings (Carraher et al. 2013). More than half

of participants undergoing urea breath testing between 2007 and 2010 were *H. pylori*-positive, and this infection was much more common amongst Indigenous residents. For those treated for *H. pylori* infection, reinfection rates in Aklavik were around 5%. In studies conducted in countries in the Global North, observed reinfection rate are typically lower, at 2% or less (Pohl et al. 2019).

The CAN*Help* team has reported that amongst 878 participants studied between 2008 and 2013 overall prevalence was 62%, with 66% prevalence in 740 Indigenous participants and only 22% in 77 non-Indigenous participants (Fagan-Garcia et al. 2019). These data pertain to participants from four Arctic Hamlets: Aklavik, Old Crow, Fort McPherson, and Tuktoyaktuk, all communities included in this project.

Chapter 2: Methods

2.1 Isolate collection

Temporary endoscopy clinics were established by the CAN*Help* Working Group in Aklavik (2008), Old Crow (2011), Fort McPherson (2012), in Inuvik for participants residing in Tuktoyaktuk (2012) and Inuvik (2017), Ross River (2017) and Teslin (2017). There, endoscopists took 7-10 gastric mucosal biopsies from two sites, the corpus and antrum, preparing 5 (Aklavik, 2008) or 6 (all other projects, 2011-2018) samples for pathology and 2 (Aklavik, Old Crow, Tuktoyaktuk, Fort McPherson, 2008-2012) or 4 (Ross River, Teslin, Inuvik, 2017-2018) for microbiological culture. For the microbiological samples, half were taken from each of the two gastric topographical sites. Biopsies were homogenized, cultured, and colonies of *H. pylori* were subcultured and stocked in glycerol. The biobank stocks were confirmed to be *H. pylori* using urease, catalase, and oxidase testing, as well as 16s ribosomal RNA PCR testing. PCR testing was also used to categorize stocks based on genotyping at *cagA*, *vacA*, and other important loci. This work was all done prior to this current study.

To represent our Southern control samples, we cultured clinical gastric biopsy tissues previously collected and biobanked by the Government of Manitoba's Cadham Provincial Laboratory from the province's Southern Health Region. Stomach biopsies taken during gastric endoscopy were homogenized in saline, pelletized, and inoculated on Mueller-Hinton agar supplemented with 10% sheep blood. Plates were incubated at 37 °C in an airtight container with a Campy Gen-Pak or a microaerophilic gas mixture—*H. pylori* is microaerophilic—for several days. It's important to note that *H. pylori* is a fastidious and slow grower, and, according to the literature, usually requires 5 days to a week to reach optimal growth (Blanchard and Nedrud 2006). In actuality, most cultures took between one and two weeks before there was enough biomass for genomic DNA extraction.

2.2 DNA extraction

Fully cultured *H. pylori* isolates were prepared as clonal populations for genomic DNA (gDNA) extraction using Qiagen DNeasy DNA extraction kits (Qiagen, Hilden, Germany) for the Arctic isolates cultured at the University of Alberta and Epicentre MasterPure Complete DNA and RNA Purification kit (Illumina, San Diego, California) for the Manitoba isolates cultured at the NML. DNA extraction was conducted as per the manufacturers' protocols. gDNA was quantified using Qubit Fluorometric Quantification (ThermoFisher Scientific, Waltham, Massachusetts) and assessed for quality with A_{260}/A_{280} spectroscopy.

2.3 Sequencing

Library preparation and index-tagging was done using illumina Nextera XT kits (Illumina, San Diego, California). DNA fragments between 600 and 1500 bp were selected for sequencing using a BluePippin pulsed field DNA size selector (Sage Science, Beverly, Massachusetts) to avoid adaptor-dimers. Sequencing was conducted on an Illumina MiSeq (Illumina, San Diego, California) technology as v3-chemistry 2x300bp paired-end reads with 96 isolates per run. This sequencing target was expected to yield an average 95-fold base coverage of the 1.6 Mb *H. pylori* genome. A small subset of 5 Arctic genomes (CAN-002, -003, -007, -010, and -012) were sent to a collaborator, Dr. M. Constanza Camargo, at the National Institutes of Health for PacBio (Pacific Biosciences, Menlo Park, California) long read sequencing.

IRIDA was used to store genomic sequence files, as well as for data sharing, analysis, and visualization (Matthews et al. 2018). Initial quality assessment was done using FastQC v.0.11.9 (Andrews 2010) for all short-read sequences within IRIDA v20.09.3.

2.4 Quality Assessment, Contamination Detection, and Filtering

Short-reads in .fastq file format were imported into the Galaxy Bioinformatics platform v20.01 for analysis (Afgan et al. 2018). All short-reads were inputted into fastp v0.20.1 for quality filtering and adaptor trimming (Chen et al. 2018).

Short-reads were investigated for potential contamination by running RefSeq Masher v0.1.2 (Government of Canada, 2017). Runs containing both *H. pylori*-mapping reads as well as reads mapping to other bacterial genomes were selected for contamination filtering. Reads were classified based on taxonomic grouping using Kraken2 v2.0.8 with the standard Kraken database (Wood et al. 2019; Wood and Salzberg 2014), and non-*H. pylori*-mapping reads were filtered out using the KrakenTools script suite v1.2 (Lu 2018). Sequencing reads that did not contain any *H. pylori* genomic signal were dropped from analysis.

2.5 Genome assembly

Assembly of *H. pylori* genomes from short-read sequences was done *de novo* using SPAdes tool within the Shovill wrapper pipeline v1.1.0. SPAdes can assemble bacterial genomes from both single- and multi-cell samples (Bankevich et al. 2012). Assembled draft genomes were assessed using QUAST v.5.0.2. QUAST is a quality assessment tool for genome assemblies (Gurevich et al. 2013). It provides information on contigs, coverage, G-C%, gene qualities, and has a visualization component. To ensure contaminated genomes were of good quality after contamination filtering and removal, QUAST also was run pre- and post-filtering.

Some tools within the analysis pipeline do not take .fasta files as input and are only coded to accept reads in .fastq file format. For the 5 genomes sequenced using long read PacBio technology, illumina short paired-end reads in .fastq file format were simulated using ART v2014.11.03 (Huang et al. 2011).

2.6 Multi-locus sequence typing

in silico multi-locus sequence typing (MLST) was done to determine the lineages to which our isolates cluster (Jolley and Maiden 2010; Maiden et al. 2013). MLST was conducted using Torsten Seeman's MLST tool v.2.19.0 (2016), which scans contig files against the PubMLST Bacterial Isolate Genome Sequence database using the *H. pylori* MLST schema (found at https://pubmlst.org/helicobacter/). The loci typically used in MLST for *H. pylori* are *atpA, efp, ppa, mutY, ureI, trpC,* and *yphC* (Haley and Gaddy 2015).

2.7 Annotation

Genome annotation was done using the Prokka tool v1.13 (Seemann 2014) to identify the location of biologically important elements of assembled *H. pylori* genomes. Transitive annotation was guided using a manually curated reference annotation provided by our collaborator Dr. Kaisa Thorell of the University of Gothenburg.

2.8 Phylogenomic analysis

2.8.1 Reference-free phylogenomics

Due to the issues outlined in section 1.2.4, reference-based approaches in *H. pylori* genomics are often insufficient and can lead to bioinformatics tool failure for scripts

developed for more clonal organisms. Many phylogenetics tools fall under this limitation. To address this, phylogenomic analysis was done using a reference-free *k*-mer based tool, Mashtree v1.1.2 (Katz et al. 2019) using assembled whole genomes as input. The phylogenetic tree containing Arctic and Manitoban *H. pylori* genomes was supplemented with publicly available genomes published from representative genomes from across world geographies. Publicly available genomes used in this study, which will henceforth be referred to as global and world genomes, are presented in Table A2.1 in the appendix.

2.8.2 Overlaying contextual data on phylogenies

All phylogenetic tree figures presented in this current study were prepared in R (v4.0.5) using the ggtree package (Yu 2020; Yu et al. 2018; Yu et al. 2016). Contextual data from both community outreach and genomics sources were used to overlay and annotate the trees to be more informative using the package ggheat. Starting from the same Mashtree phylogeny (tree) in each instance, a heatmap overlay of the desired contextual data was created from an annotation file using the package ggheat.

2.9 Characterization

2.9.1 Antimicrobial resistance

Genes of interest from the literature to be known sources of antimicrobial resistance were investigated. Protein sequences for *rdxA* reductase, *gyrA* gyrase subunit A, and *pbp*1A penicillin binding protein, as well as the nucleotide sequence for 23S rRNA were extracted from the Prokka output annotation file. A multiple alignment of these sequences was conducted using MAFFT v.7.221.3 (Katoh et al. 2002), and variants at known resistance markers were visualized in Unipro UGENE v38.0 (Okonechnikov et al. 2012).

These results were confirmed by running the Resistance Gene Identifier (RGI) v4.2.2, which scans the Comprehensive Antibiotic Resistance Database (CARD) (McArthur et al. 2013; Alcock et al. 2020). These results were also compared to phenotypic resistance measurements previously collected by the CAN*Help* Working Group using the E test method (Cheung et al. 2009).

2.9.2 Virulence factor detection

The coding sequences for the CagA protein were extracted from the Prokka annotation and were used as the query sequences for a BLASTp search using a small database of the various CagA EPIYA motifs (Gomez 2020) under default parameters. The EPIYA motifs are presented in **Table 2.1**. The presence or absence of the *cag* pathogenicity island (*cag*PI) exists independent of *cagA* status, and so all isolates, including *cagA*- ones were investigated for *cag*PI genes. *In silico* polymerase chain reaction (PCR) was done to determine *vacA* genotype within the Unipro UGENE software (Okonechnikov et al. 2012) to detect variations of s and m alleles. The primers used for amplification were taken from Akeel et al. 2019 and are presented in **Table 2.2.**

The annotations for select OMP genes, namely *babA*/*hopS*, *sabA*/*hopP*, *hopH*/*oipA*, *hopQ*, *alpA*/*hopC*, *alpB*/*hopB*, *homA*, *homC*, and *homD* (Matsuo, Kido, and Yamaoka 2017; Xu et al 2020), as well as the toxin-antitoxin *aapA1*/*isoA1* (Arnion et al. 2017; Krzyżek 2021) were pulled from the annotation and the presence or absence of these virulence-associated genes were noted. Upset plots were created to depict the overlap of the presence of antimicrobial resistance markers and virulence factor genotypes including *vacA* s/m allele types. These figures were created in R (v4.0.5) using the UpSetR package (Conway, Lex, and Gehlenborg 2017).

Table 2.1 CagA protein EPIYA motifs

EPIYA Type	Protein Sequence	Notes
EPIYA-A	EPIYAKVNKKKAGQ	
EPIYA-B	EPIYAQVAKKVNAKI	
EPIYA-C	EPIYATIDDLGGP	-ABC associated with Western isolates -C motifs can exist in multiple copy numbers, *e.g.* ABCCC
EPIYA-D	EPIYATIDFDEANQAG	-ABD associated with Asian isolates -D motifs can exist in multiple copy numbers, *e.g.* ABDDD

Table 2.2 Primers used for the amplification of *vacA* s and m allelic regions

Allele	Forward	Reverse
s1a	CTCTCGCTTTAGTAGGAGC	CTGCTTGAATGCGCCAAAC
s1b	AGCGCCATACCGCAAGAG	CTGCTTGAATGCGCCAAAC
s1c	CTCTCGCTTTAGTGGGGYT	CTGCTTGAATGCGCCAAAC
s2	GCTAACACGCCAAATGATCC	CTGCTTGAATGCGCCAAAC
m1a	GGTCAAAATGCGGTCATGG	CCATTGGTACCTGTAGAAAC
m1b	GGCCCCAATGCAGTCATGGA	GCTGTTAGTGCCTAAAGAAGCAT
m2	GGAGCCCCAGGAAACATTG	CATAACTAGCGCCTTGCA

2.9.3 Genomic islands presence

The presence of genomic islands in the genomes was predicted using IslandViewer 4, a webserver application that integrates four genomic island prediction approaches: IslandPick, IslandPath-DIMOB, SIGI-HMM, and Islander (Bertelli et al. 2017).

2.9.4 Plasmid profile

Plasmid content was identified in assemblies using MOB-suite, a modular set of tools allowing for plasmid construction and typing with both high specificity and sensitivity (Robertson and Nash 2018). MOB-suite v3.0.0 allows for the detection, characterization, and filtering out of plasmid sequences in whole genomes.

2.9.5 Bacteriophage presence

The presence of phage sequences within our genomes was identified using PHASTER (PHAge Search Tool Enhanced Release), an upgrade to the PHAST web server (Zhou et al. 2011; Arndt et al. 2016). PHASTER allows for the identification and annotation of prophages within bacterial genomes.

2.10 Comparative genomics

2.10.1 Core genome analysis

Comparative genomics was done using Roary v3.12.0. Roary builds pan-genomes quickly and on a large scale (Page et al. 2015), allowing for the characterization of the core and accessory genomes for our collection of *H. pylori* isolates. Core genomes were built for all Arctic, Manitoban, and global isolates at three threshold levels: included in 90% of

genomes, 95% of genomes, and 99% of genomes. Hyperplastic, accessory loci, and mobile elements such as genomic islands and plasmids were excluded from the core.

2.10.2 Genomic admixture and structure analysis

Since *H. pylori* is such a highly recombinant bacterium, analysis to detect instances of recombination is required, especially in cases where individuals experience co-infection, in which they are infected with multiple strains.

To address this characteristic of the genome, a core genome admixture approach was pursued. A core genome multiFASTA alignment of 769 core genes present in >90% of Manitoban, Arctic, and world isolates was created using Roary v3.12.0 (Page et al. 2015). Genome-wide haplotype data were prepared using BEAGLE v.3.3.2 (Browning & Browning, 2009), which imputed missing base pairs and called 36,164 SNPs in the alignment. Isolate population clusters were determined using fineSTRUCTURE v.0.02 run for 100,000 iterations of both the burn-in and Markov chain Monte Carlo method and CHROMOPAINTER v.0.02 (Lawson et al. 2012). All genomes in the current study were designated as DNA donors, and all Arctic and Manitoba isolates were designated as DNA recipients. The output of fineSTRUCTURE includes a coancestry matrix, a heat map with cells indicating the proportion of DNA “chunks” a recipient receives from a given donor, and a principal component analysis (PCA). Using CHROMOPAINTER, genomic admixture was visualized through stacked bar charts using chromosome painting (Yahara et al. 2013).

2.11 Community engagement

2.11.1 Knowledge exchange

Unfortunately, a planned genomics workshop in Inuvik, NT, in April 2020 with our partner communities was cancelled due to the COVID-19 pandemic. Instead, the in-person workshop was replaced with several virtual seminars, one of which I presented on Bacterial Genomics, as well as a half-day virtual workshop, both of which were recorded. During this workshop, I attempted to describe bacterial genomics in lay language suitable for community members. For example, I used the analogy of the genome as a cookbook, with certain recipes being analogous to genes and meals as proteins.

The lessons taken from these seminars and workshop are currently being developed as a publication for which I will co-author. I have also agreed to maintain a relationship with the CAN*Help* Working Group such that, when appropriate, I may be able to continue to deliver knowledge exchange activities regarding my thesis work with our partner communities, even after my graduate studies have concluded.

In September 2020, I travelled to Edmonton, Alberta for two months to assist the CAN*Help* Working Group with various laboratory tasks to help achieve project goals related to and outside of the scope of my thesis work. While there, I also participated in knowledge exchange activities with the group focused on the topic of *H. pylori* genomics.

2.11.2 Data visualization

As stated at the beginning of this project outline, a strong data visualization aspect of this project is crucial to ensuring that this research project reflects Indigenous research frameworks of self-determination. Scientists have failed in the past to integrate Indigenous participants and communities in meaningful knowledge exchange at both the design and

results stages of a research project. This ultimately has resulted in uninformed consent and general neglect of community needs.

As noted above, a large component of this research project is conducting structured knowledge exchange activities with Indigenous participants and communities in order to develop actionable ways of improving Northern Arctic health. A strong, culturally-appropriate, and audience-focused data visualization approach is key in communicating the volume and complexity of genomic epidemiological data. This work will continue even after my graduate studies have concluded.

2.12 Ethics Approval

This study was approved under the University of Alberta HREB# Pro00007868 (Study Title: Addressing Community Concerns about Risks from *H. pylori* Infection in the Circumpolar North). As well, this project adheres to the guiding principles for stewardship and dissemination of knowledge generated collaboratively in CAN*Help* Working Group community projects (Goodman et al. 2019):

1. Research questions were developed with input from community project planning committees, who represent research participants

2. Research participants donated personal information, including that arising from biological material, to community research projects; knowledge takes shape from information donated by participants through:

    a. Scientific observation and measurement that turns information into scientific data

b. Analysis that reveals patterns in data

c. Interpretation of data patterns

3. Researchers are bound by ethical, professional, scientific, contractual, and other legal standards to be responsible stewards of the information donated by participants; this entails using this information for no other purpose than to achieve explicitly shared research goals

4. Any member of the CAN*Help* Working Group who contributes to a particular research project may contribute to the interpretation of the data patterns; when consensus is not possible, different viewpoints will be reported

5. All members of the CAN*Help* Working Group who contribute to a particular research project are entitled to be acknowledged for their contributions (either as a group or individually depending on practical considerations)

6. In keeping with professional and ethical standards, individuals who make key contributions to specific research reports will be acknowledged as authors

7. After following CAN*Help* Working Group guidelines for review of research results by relevant collaborators, results will be made accessible to the public to ensure open access to knowledge generated through the support of public resources, as required by funding agencies and the university

8. The CAN*Help* Working Group research director, Karen Goodman, has ultimate responsibility for ensuring that these guiding principles are upheld

Chapter 3: Results

3.1 Description of *H. pylori* genomes obtained for the analysis

3.1.1 Genomes included in current study

In total, 156 unique genomes were sequenced using Illumina short-read sequencing technology (68 Arctic genomes and 88 Manitoba genomes). The 88 Manitoba genomes contained 4 biological duplicate pairs (1I5A+B; 1B9A+B; 2H4A+B; 3C9A+B) and one set of biological quadruplicates (1H3A-D), giving a total of 81 unique isolates sequenced. The 68 Arctic genomes contained 4 pairs of biological duplicates (CAN-010A+B, CAN181A+B, CAN247A+B, and CAN271A+B), resulting in a total of 64 unique isolates sequenced. Five Arctic genomes (CAN-002, CAN-003, CAN-007, CAN-010, and CAN-012) were sequenced using PacBio long-read sequencing technology, including one genome that was also sequenced with Illumina sequencing technology. In total, this project began with 149 new and unique *H. pylori* genomes that have never before been sequenced.

After *de novo* genome assembly, two Manitoba isolates, 1E7 and 1I5A failed to assemble due to insufficient breadth of genome coverage. Two additional Manitoba isolates, 3E1 and 4C7 were removed for assembling in >300 contigs and were deemed poor quality. QUAST indicated that 14 isolates had G-C% that strayed (<39% and >40%) from the standard expected value of 39% for *H. pylori*. Included in those 14 were 6 isolates (1C3, 1C5, CAN-039, CAN-076, CAN-091, and CAN-288) that assembled in less than the 1.6 Mb complement for *H. pylori* genomes; thus these genomes were removed. Eleven (11) additional isolates assembled much larger (range: 2.0 to 6.4 Mb) than the 1.6 Mb genome complement expected for *H. pylori*.

Contamination detection was run for these 19 aberrant genomes. Contaminating species for these genomes are presented in **Table 3.1**. Of these contaminated data sets, 8 had insufficient *H. pylori* reads to assemble the full genome complement (3A3, CAN-023,

CAN-093, CAN-102, CAN-121, CAN-152, CAN-236, and CAN-279) after contaminant read removal and were removed from downstream processing. Seven genomes (1D2, 4C7, CAN-005, CAN-074, CAN-087, CAN-212, and CAN-261) successfully passed the contamination-filtering step and were 'rescued' for downstream *H. pylori* genomic analyses.

Finally, one pair of Manitoban isolates, 2C1 and 2C2, were suspected of being mislabelled at some point, whether it be in the clinic or during processing for the current study; this was suspected because 2C1 and 2C2 were nearly identical genetically (>99.9%) despite having no reason to be identical. These two isolates were ostensibly taken from two different, unrelated individuals at the same clinic during the same week. Despite this, these two genomes clustered as closely as biological replicates and repeat isolates, and had near-identical sequences. It is possible that these two samples were mislabelled at some point during collection, storage, resurrection, or sequencing. As a result, these two isolates were consolidated due to this unexplained redundancy, and only 2C2 was analyzed downstream. Consequently 2C1 was randomly removed from analysis.

In summary, 57 Arctic genomes and 73 Manitoba genomes of suitable quality and length remained after quality assurance. These genomes and their assembly statistics are presented in **Table A3.1** in the appendix, with genome length, contig number, and G-C% visualized in **Figure 3.1**. The 5 Arctic genomes sequenced by PacBio (CAN-002, CAN-003, CAN-007, CAN-010, and CAN-012) each yielded a single high confidence contig (**Table A3.1)**. A Sankey diagram displaying the flow of isolates through all quality control steps is presented in **Figure 3.2**.

Table 3.1. *H. pylori* isolates that underwent contamination read filtering

Isolate	Contaminating reads	*H. pylori* content	Outcome
CAN-005	*Mycoplasma hominis*	Full genome	Remained
CAN-023	*Brevundimonas diminuta*	None	Removed
CAN-074	*Mycoplasma hominis*	Full genome	Remained
CAN-087	*Mycoplasma hominis*	Full genome	Remained
CAN-093	*Mycoplasma hominis*	Some reads	Removed
CAN-102	*Streptomyces coelicolor* *Micrococcus luteus* *Frankia spp.* *Papiine alphaherpesvirus* *Kitasatospora purpeofusca* *Streptomyces spp.*	None	Removed
CAN-121	*Brevundimonas diminuta*	None	Removed
CAN-152	*Mycoplasma hominis*	Some reads	Removed
CAN-212	None	Full genome	Remained
CAN-236	*Haemophilus influenzae*	Some reads	Removed
CAN-261	*Mycoplasma hominis*	Full genome	Remained
CAN-279	*Mycoplasma hominis*	None	Removed
1D2	None	Full genome	Remained
3A3	*Capnocytophaga sputigena*	Some reads	Removed
4C7	None	Full genome	Remained

Figure 3.1 Assembled genome length, G-C%, number of contigs for all assemblies first described in this study. The expected genome size of *H. pylori* is 1.6 Mb with slight variation observed (1.5-1.7 Mb). The ideal genome is assembled in as few contigs as possible. The expected G-C% for a given *H. pylori* genome is 39% with slight variation observed (38-40%). G-C% is represented by both point size and colour. Cohort (Manitoban or CAN*Help*) is represented by point shape. The genomes depicted are only those that had suitable quality and length after quality assurance.

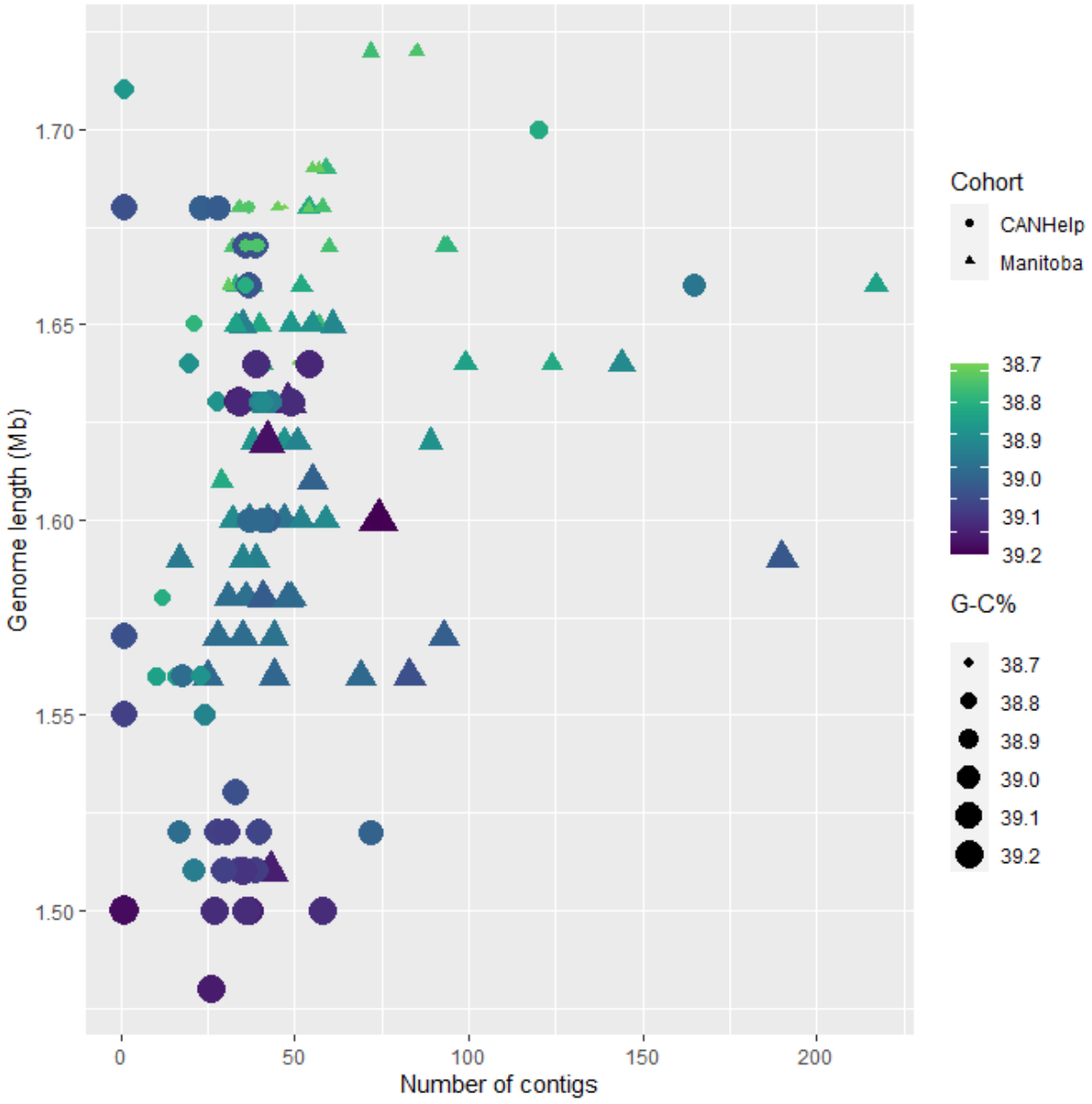
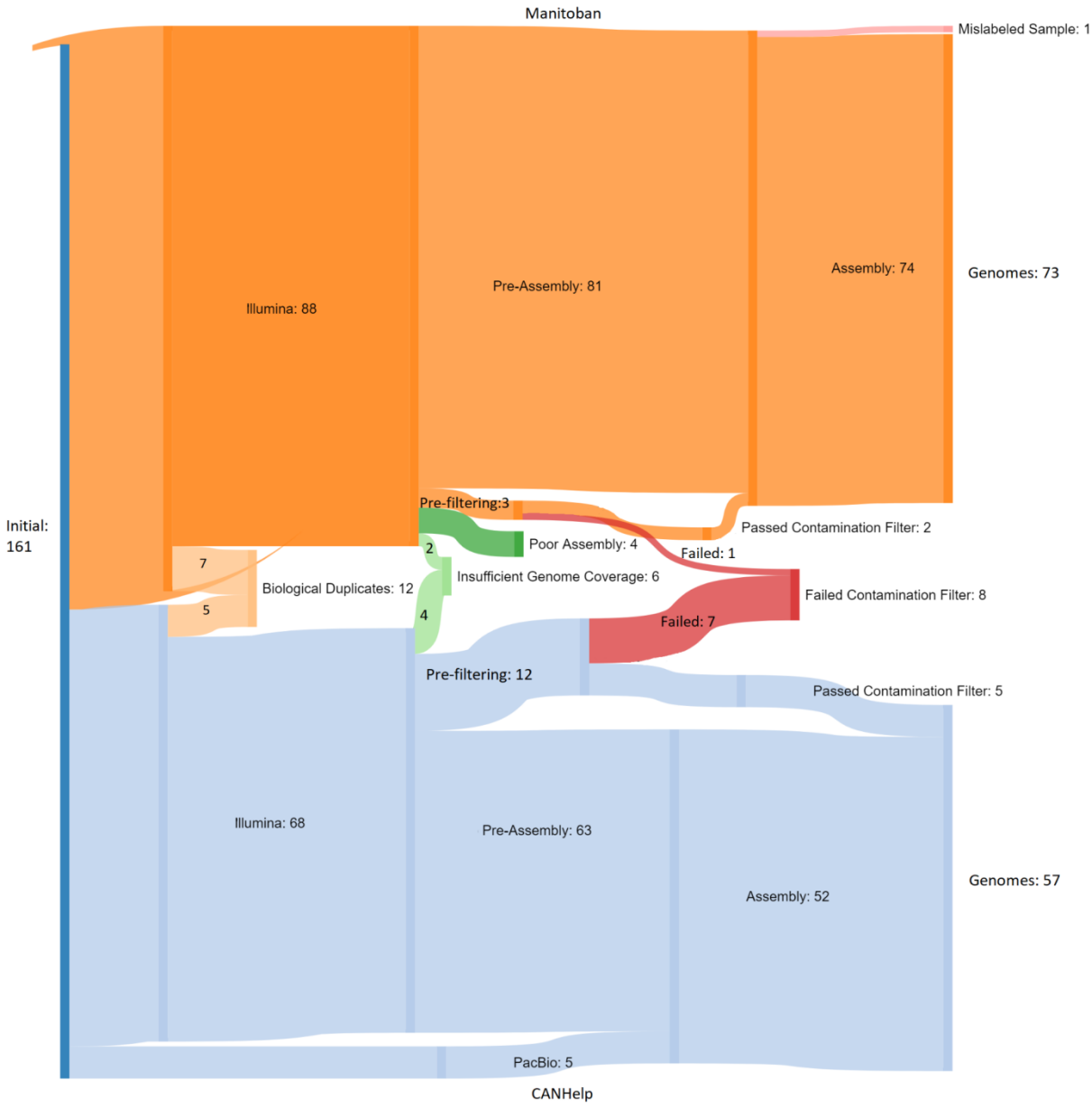

1.70
1.65
1.60
1.55
1.50
Genome length (Mb)
0
50
100
150
200
Number of contigs
Cohort
CANHelp
Manitoba
38.7
38.8
38.9
39.0
39.1
39.2
G-C%
38.7
38.8
38.9
39.0
39.1
39.2

Figure 3.2 Sankey diagram displaying the loss of genomes through quality control and final tally of included genomes. Manitoba isolates are depicted in orange while CAN*Help* isolates are depicted in sky blue. 12 genomes (7 Manitoban and 5 Arctic) represented biological duplicates and redundancies were removed. After assembling short-reads, 4 Arctic and 2 Manitoban isolates were removed for not covering the entire *H. pylori* genome (read coverage not distributed) and an additional 4 Manitoban isolates were removed for assembling in too many contigs. A total of 8 genomes (1 Manitoban and 7 Arctic) failed to pass contamination filtering and were removed. Finally, 2 unrelated Manitoban isolates were identical in genomic content and were deemed to be mislabeled; consequently one was removed due to redundancy.


Manitoban
Mislabeled Sample: 1
Illumina: 88
Pre-Assembly: 81
Assembly: 74
Genomes: 73
Initial:
161
Pre-filtering:3
Passed Contamination Filter: 2
Poor Assembly: 4
Failed: 1
7
2
Insufficient Genome Coverage: 6
Biological Duplicates: 12
Failed Contamination Filter: 8
5
4
Failed: 7
Pre-filtering: 12
Passed Contamination Filter: 5
Illumina: 68
Pre-Assembly: 63
Assembly: 52
Genomes: 57
PacBio: 5
CANHelp

3.1.2 Multi-locus sequence typing

The MLST sequence typing results for all quality-filtered sequenced isolates are presented in **Table A3.2** in the appendix. It is important to note that not every isolate has a sequence type assigned, even if it has an assigned allele type at each of the 8 MLST loci. This is because of the recombinant nature of the *H. pylori* genome; some isolates, due to homologous recombination, will contain unique MLST allele combinations that have never previously been characterized in the pubMLST *Helicobacter pylori* schema database.

3.2 Phylogenomics

3.2.1 Reference-free phylogeny

To ensure that $k = 21$ was an appropriate *k*-mer size to use in Mashtree, phylogenetic trees were produced using $k = 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31$. There were no differences in the output phylogenetic trees amongst these different *k*-mer sizes. The *21-mer* based reference-free phylogeny of Arctic and Manitoba genomes, supplemented with publicly available *H. pylori* genomes from across the world is presented in **Figure 3.3.** The overall phylogenetic structure of the isolates is similar to that of other published *H. pylori* trees (Duncan et al. 2013). As expected, the tree is structured based on human geography and proposed human migration patterns, with African isolates existing as an outgroup at one end of the tree and American Indigenous isolates existing as an outgroup at the other end. Notably, an entire cluster of Arctic isolates, and a handful of Manitoban ones cluster with African global isolates. The branch length of the Arctic isolates is rather short relative to all of the other branches, implying that these genomes are very closely related

genetically, something that is seen throughout every single cluster of CAN*Help* isolates in the phylogeny.

The majority of Manitoban isolates cluster closely with world isolates from Northern and Southern Europe. The branch lengths separating these Manitoban isolates from each other are consistently longer than the branch lengths separating Arctic isolates within clades, and imply that most isolates are less closely related in Manitoba than in the Arctic. Of note, paired longitudinal Manitoba isolates sampled from the same individual months to years apart, clustered extremely close together as might be expected from longitudinal sampling from the same host. These include 1G5 & 2C5, 2B2 & 2I5, 1I7 & 3B4, and 1E9 & 4A1. Within these few longitudinal sequence pairs, there were no discernible sequence effects of timespan between samplings.

As well, two Manitoba isolates, 2C1 and 2C2 showed high genetic similarity, over 99.9% sequence similarity, despite being sampled from different individuals, albeit at the same clinic during the same week. It is suspected that these isolates are in fact the same and were likely mislabeled some time during isolate processing, either at Cadham Provincial Laboratory or at the NML. As a result, 2C1 was removed from analysis.

Notably, clusters of Arctic isolates are localized with global isolates predominantly from Northern European countries. Finally, the last large clade of Arctic isolates cluster closely with *H. pylori* genomes isolated from Indigenous communities in North and South America, including two isolates, Aklavik86 and Aklavik117, which were previously sequenced by the CAN*Help* Working Group (Kersulyte et al. 2015).

Figure 3.3. *k*-mer based, reference-free phylogeny computed with Mashtree (*k* = 21). Tree rooted with reference genome J99. Tree tips labelled with both geography of origin and fineSTRUCTURE-determined sequence type. Phylogeny was created using Mashtree with whole genomes as input. fineSTRUCTURE analysis was run using 769 core genes present in >90% of Manitoban, Arctic, and world isolates.

geography fs
fineSTRUCTURE
population (k)
geography

3.2.2 Genomic admixture

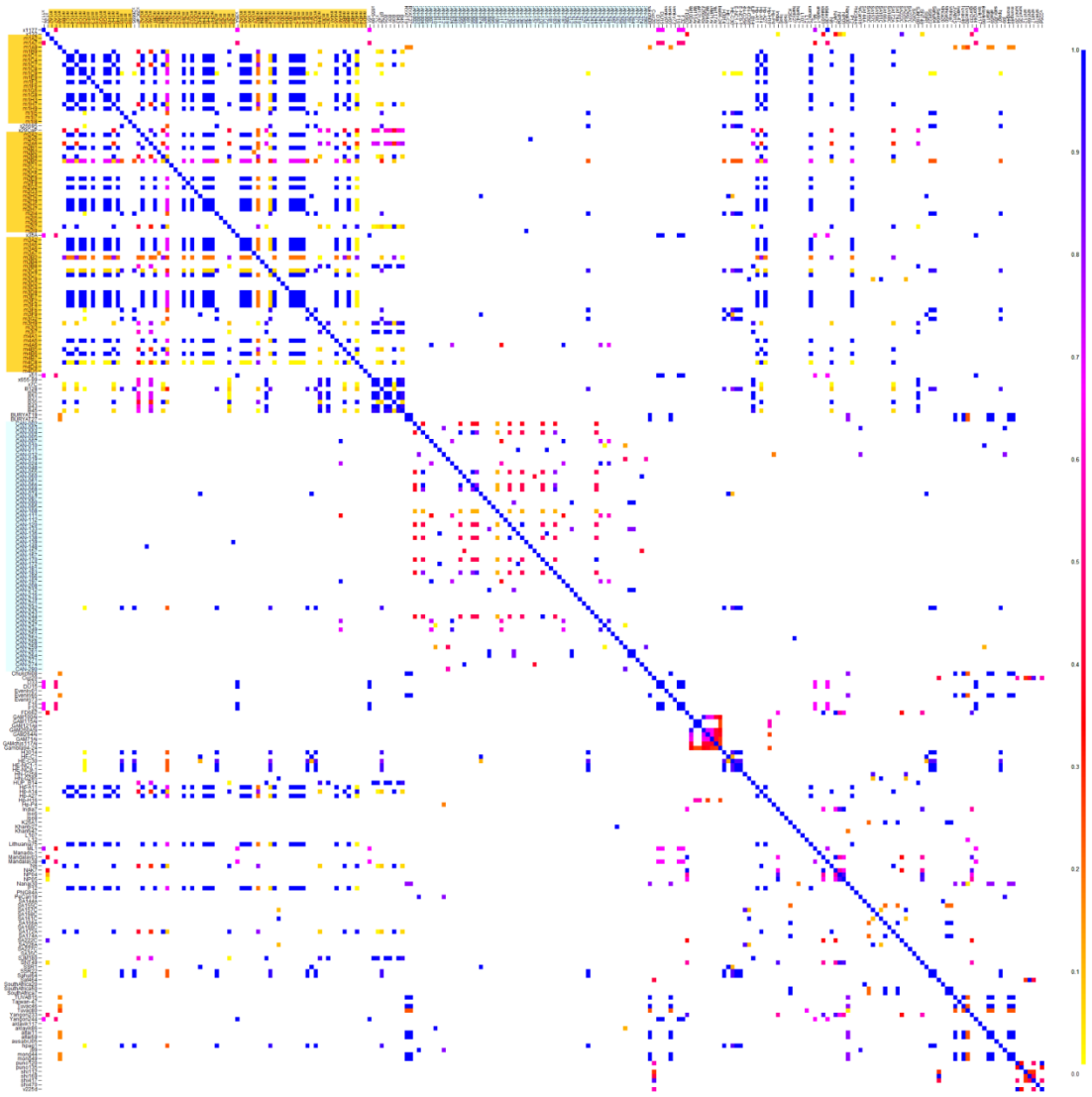
The co-ancestry matrix plot presented in **Figure 3.4** indicates broad trends within the Manitoban and Arctic *H. pylori* genomic cohorts. All three large genomic cohorts have strong self-self signals, which are to be expected, with the strongest self signals existing within the Manitoba cohort. As well, the Manitoban and global cohorts shared strong co-ancestry signal with each other. Conversely, the Arctic cohort shared little co-ancestry with either the Manitoban or global genome cohorts. The strength of the co-ancestry signal in the matrix was then used to compute the phylogenetic groupings as presented in the left side panel of **Figure 3.5**. As seen in the *k*-mer based phylogenies, the Arctic genomes tended to cluster in tight and isolated sub-clades, whereas the Manitoban isolates were predominantly dispersed in broad clades with global isolates of European-assigned sequence types. Employing chromosome painting and fineSTRUCTURE analysis allowed for exploration of *H. pylori* population structure at a finer scale of resolution than would normally be possible with draft genomes. For example, I have been able to infer the extent and direction of recombinogenic genetic flux between various subgroups.

From the phylogenetic tree produced by fineSTRUCTURE seen in **Figure A3.1** in the appendix, sequence types were then assigned to all Manitoban and Arctic isolates based on known sequence types of global isolates contained within clusters and clades. fineSTRUCTURE was run to include 10 high-quality reference donor genomes from each of the following sequence types: hpEAsia, which represent *H. pylori* from East Asia; hpAsia2, which represent *H. pylori* from South Asia; hspSiberia, which represent *H. pylori* from East Russia; hspIndigenousNAmerica, which represent *H. pylori* from Indigenous communities in North America; hspIndigenousSAmerica, which represent *H. pylori* from Indigenous communities in Central and South America; hspSEurope, which represent *H. pylori* from

Southern Europe; hspNEurope, which represent *H. pylori* from Northern Europe; hspSWEurope, which represent *H. pylori* from Southwestern Europe; hpAfrica1, which represent *H. pylori* from Central and Western Africa; and hpAfrica2, which represent *H. pylori* from Southern Africa. All 100 determined sequence types are presented in **Table A3.3** in the appendix.

The fineSTRUCTURE-determined sequence types were then used to paint the chromosomes of every Arctic and Manitoban genome to produce a stacked bar chart representing core genomic admixture for each individual genome. Here, the term "admixture" is employed to infer a flow of chunks of sequence (more generally, "genetic flux") between (sub)groups, and is presented in **Figure 3.5.** The majority of Manitoban isolates had similar chromosome painting patterns, with a significant proportion mapping to the European sequence types hspSEurope and hspNEurope, as visualized by the strong proportion of light and dark blue (respectively) in the stacked bar charts. Of note are 1A2 and 3I3, which are dominated by hpAfrica1 signal (maroon; **Figure 3.5**), as well as 2A5, 2I9, and 4A6, which are driven by strong hspIndigenousSAmerica signal (light purple; **Figure 3.5**) for the former two and hspIndigenousNAmerica (dark purple; **Figure 3.5**) for the latter. Of the isolates that are not heavily dominated by a single sequence type signal, the proportion of core genomic admixture from the various sequence types appears to be stable. Focusing on the Arctic genomes, almost every isolate has a significant hspIndigenousNAmerica and hspIndigenousSAmerica signal (dark and light purple, respectively), except for three: CAN-138, CAN-139, and CAN-172, which are dominated by hpAfrica2 (pink; **Figure 3.5**). Akin to the dual-European signal (light and dark blue) in the Manitoba isolates, this admixture signal for these 3 isolates can be seen by the strong proportion of two-toned purple in the stacked bar charts.

Figure 3.4 Co-ancestry matrix in the order of Manitoban, Arctic, and global *H. pylori* core genomes. Based on fineSTRUCTURE analysis, the colour assigned for the CAM heatmap indicates the degree of overall sequence content similarity between two genomes. Isolates labelled yellow share slight similarity (<10%), isolates labelled blue share near total similarity (>90%), and all other intervening colour assignments as depicted (at right) for 11-89% sequence similarity. The diagonal represents an isolate paired with itself (self-self). Isolates labelled on the X and Y axes with orange representing Manitoban samples and with sky blue representing CAN*Help* samples. Genomes from global sources are ordered alphabetically, but then further assigned into clusters based on their overall sequence content similarity. Interpretation for the colour of each cell in the matrix: indicates evidence of genetic flux, as the expected number of chunks imported from a donor genome (column) into a recipient genome (row). fineSTRUCTURE analysis was run using 769 core genes present in >90% of Manitoban, Arctic, and world isolates.



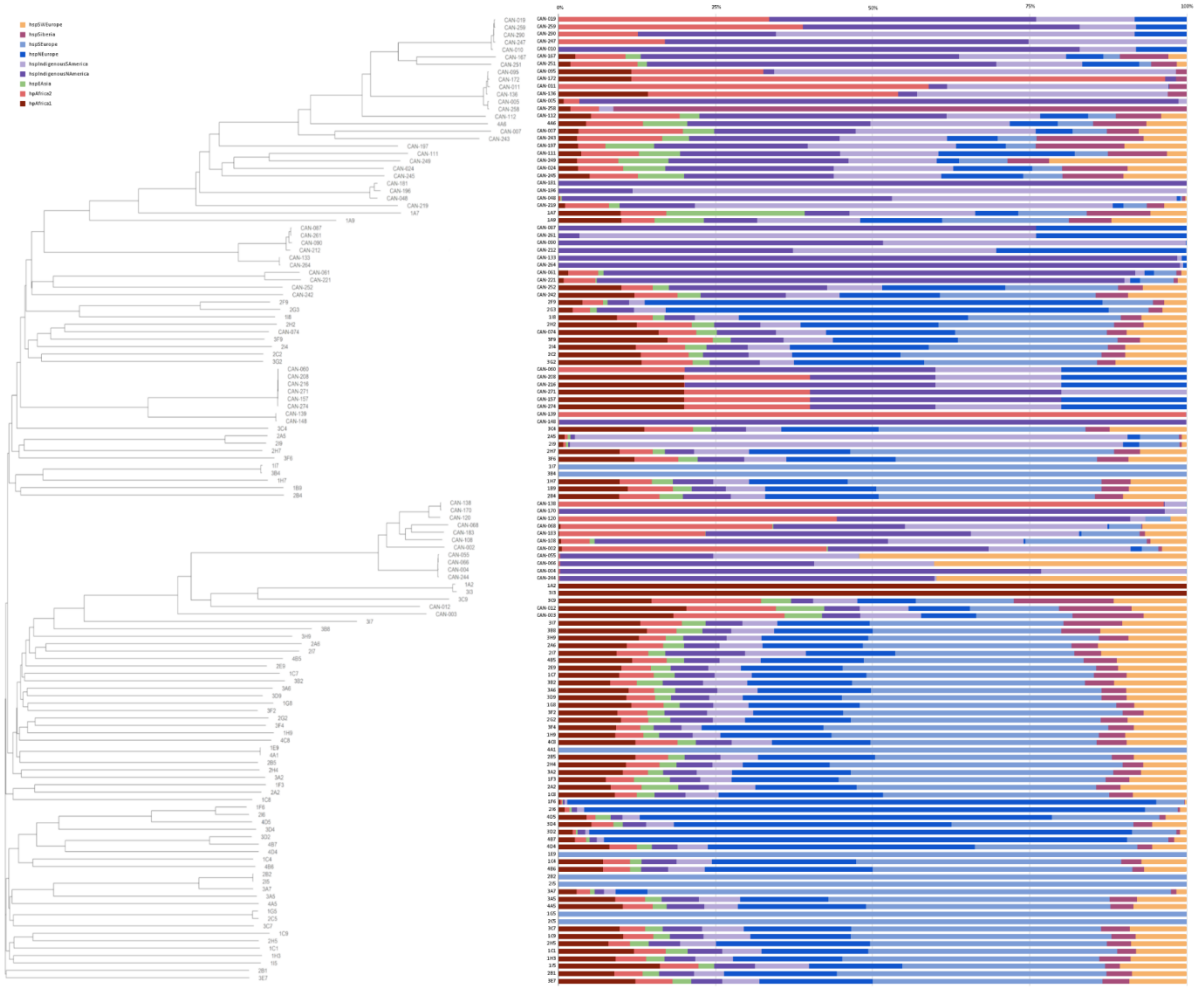
Unlike the admixture-stable Manitoban isolates, the proportion of admixture does not appear to be stable for most of the Arctic isolates (variations in the stacked bar distributions are more evident).

The Arctic clades belonging to the hpAfrica2 sequence types are of note in the phylogenetic clustering depiction (leftside **Figure 3.5**) owing to their genetic distinctness from all other isolates due to their tight clustering and long branch lengths compared to all other isolates. This distinguishes them from the clades belonging to the hspIndigenousNAmerica, hspNEurope, and hspIndigenousSAmerica sequence types, which have core genomes that, while clustered closely together indicate some ancestral bottleneck event, are more closely related to Manitoban and world isolates.

3.2.3 Phylogenetic patterns of contextual data

A large proportion of the Arctic *H. pylori* isolates included in this study come from Aklavik, NT. Of the 57 Arctic genomes sequenced, 33 were from Aklavik, 6 from Old Crow, 6 from Fort McPherson, 4 from Teslin, 4 from Inuvik, and 4 from Ross River. The two Arctic isolates belonging to hpAfrica1, CAN-003 and -012 (both sequenced to high-confidence as a resultant single contig), were from different communities, Aklavik and Old Crow, and their hosts are not closely related. However, these hpAfrica1 isolates also cluster in a broader clade with the Manitoban isolates 1A2, 3I3, and 3C9, as seen in **Figure 3.5**. The hpAfrica2 clade showed strong phylogenetic clustering based on community of origin, with two Aklavik subclades, one Inuvik subclade, and an Old Crow singleton. While one hspIndigenousSAmerica clade had a mix of community origins, the other had a very strong Aklavik-origin signal, with a mere two isolates, CAN-060 and -208 coming from a single non-Aklavik site (namely Old Crow).

Figure 3.5. Ancestry admixture plot of core genomes painted with fineSTRUCTURE determined sequence types. Genomic admixture is represented as stacked bar plots for each genome. Sequence type is represented by colour: hspSWEurope with yellow; hpAfrica1 with dark red; hpAfrica2 with orange; hspEAsia with green; hspIndigenousNAmerica with dark purple; hspIndigenousSAmerica with light purple; hspNEurope with dark blue; hspSEurope with light blue; and hspSiberia with pink. Phylogeny was created using Mashtree with whole genomes as input. fineSTRUCTURE analysis was run using 769 core genes present in >90% of Manitoban, Arctic, and world isolates.


hspEnWEurope
hspEnorra
hspEBarope
hspNEurope
hspIndigenousSAmerica
hspIndigenousNAmerica
hspEAsia
hpAfricaI
hpAfricaL
0%
25%
50%
75%
100%

The larger hspIndigenousNAmerica clade, while containing mixed origin signal, has a subclade of closely related isolates coming from either Aklavik or Fort McPherson exclusively. Notably, a single Manitoba isolate, 4A6, also clusters tightly with this subclade. The phylogenetic structure of isolates annotated with household, community of origin, sequence type, and multihousehold group (MHG), where known, is presented in **Figure 3.6**. No multihousehould group contained Arctic genomes solely from a single clade in the phylogenetic tree. Although the MHG data is sparse, it indicates unique and complex relationships based on host living conditions. For example, within the hpAfrica2 subclade quadruplet containing CAN-055, -066, -004, and -244, three belong to the MHG 10. Also within the larger hpAfrica2 clade, CAN-244 and -120 belong to the MHG 2, a group that also contains isolates within the hspIndigenousNAmerica and hspIndigenousSAmerica sequence type clades, including the distantly related genomes of CAN-172 and -290. Finally, MHG 4 contains the closely related isolates CAN-259 and -247 within the hspIndigenousNAmerica sequence type clade, as well as more distantly related hspIndigenousNAmerica isolate CAN-005, and the outgroup CAN-216 and -274.

3.3. Genome characterization

3.3.1 Antimicrobial resistance

The following resistance genotypes were identified within the Manitoba and Arctic isolates as presented in **Table 3.2**. The presence of resistance markers in the Manitoba and Arctic *H. pylori* genome populations are presented in **Figure 3.7** and **Table A3.4** in the appendix, while **Figure 3.8** represents the phylogenetic structure of the Manitoba and Arctic isolates annotated with resistance marker presence/absence.

Figure 3.6. Phylogenetic tree of Arctic and Manitoba genomes annotated with household- and multihousehold group. Tree tips are labelled with cohort (Manitoban or CAN*Help*), community of origin, fineSTRUCTURE-determined sequence type, and multihousehold group assignment. Two columns were used to depict multihousehold group assignment because some community members belonged to more than one multihousehold group. Phylogeny was created using Mashtree with whole genomes as input. fineSTRUCTURE analysis was run using 769 core genes present in >90% of Manitoban, Arctic, and world isolates.

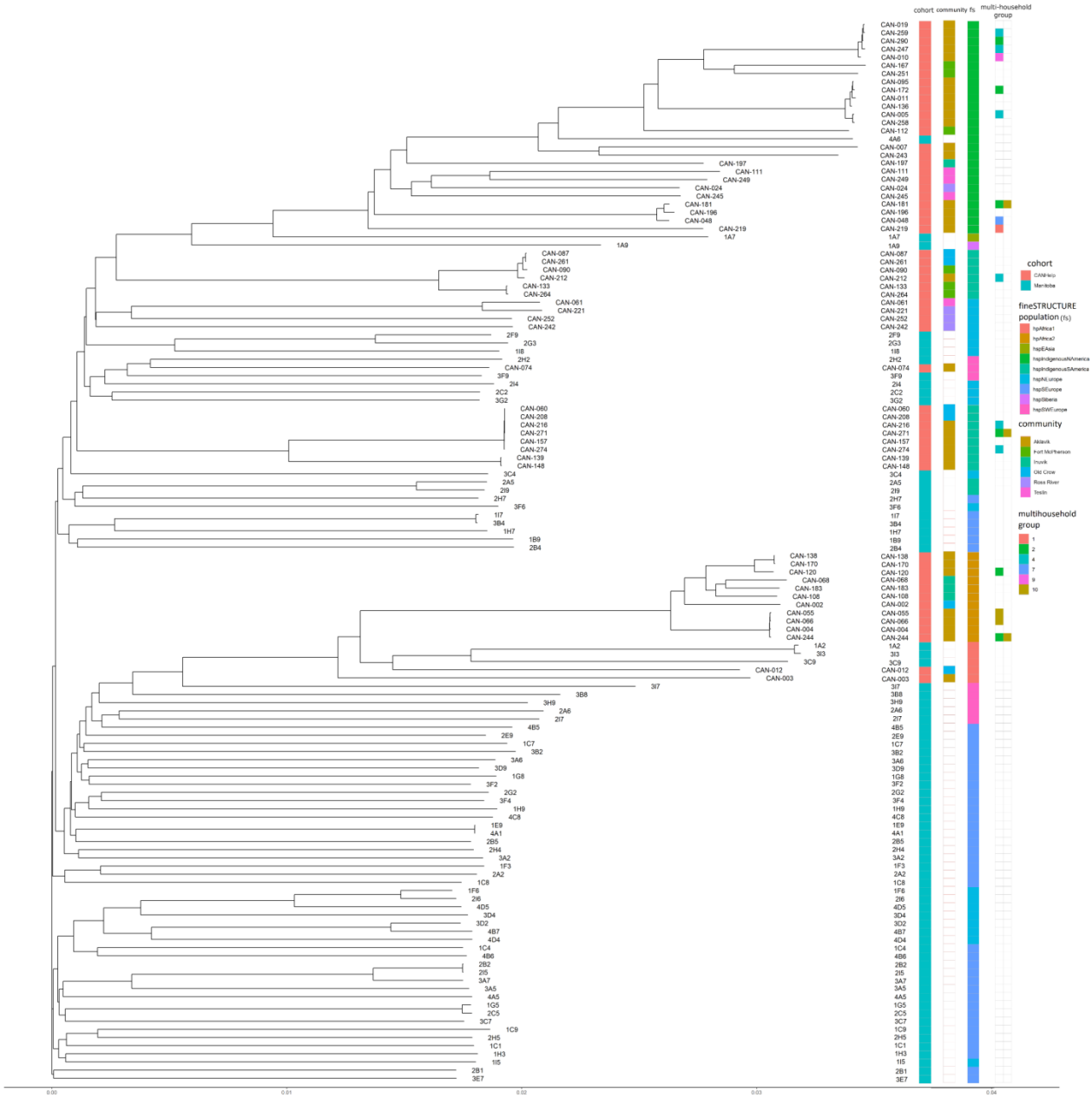

cohort
community
fs
multi-household group
CAN-019
CAN-259
CAN-290
CAN-247
CAN-010
CAN-167
CAN-251
CAN-095
CAN-172
CAN-011
CAN-136
CAN-005
CAN-258
CAN-112
4A6
CAN-007
CAN-243
CAN-197
CAN-111
CAN-249
CAN-024
CAN-245
CAN-181
CAN-196
CAN-048
CAN-219
1A7
1A9
CAN-087
CAN-261
CAN-090
CAN-212
CAN-133
CAN-264
CAN-061
CAN-221
CAN-252
CAN-242
2F9
2G3
1I8
2H2
CAN-074
3F9
2I4
2C2
3G2
CAN-060
CAN-208
CAN-216
CAN-271
CAN-157
CAN-274
CAN-139
CAN-148
3C4
2A5
2I9
2H7
3F6
1I7
3B4
1H7
1B9
2B4
CAN-138
CAN-170
CAN-120
CAN-068
CAN-183
CAN-108
CAN-002
CAN-055
CAN-066
CAN-004
CAN-244
1A2
3I3
3C9
CAN-012
CAN-003
3I7
3B8
3H9
2A6
2I7
4B5
2E9
1C7
3B2
3A6
3D9
1G8
3F2
2G2
3F4
1H9
4C8
1E9
4A1
2B5
2H4
3A2
1F3
2A2
1C8
1F6
2I6
4D5
3D4
3D2
4B7
4D4
1C4
4B6
2B2
2I5
3A7
3A5
4A5
1G5
2C5
3C7
1C9
2H5
1C1
1H3
1I5
2B1
3E7
0.00
0.01
0.02
0.03
0.04
cohort
CANHelp
Manitoba
fineSTRUCTURE population (fs)
hpAfrica1
hpAfrica2
hpEAsia
hpIndigenousNAmerica
hpIndigenousSAmerica
hpNEurope
hpSEurope
hpSiberia
hpSWEurope
community
Akiavik
Fort McPherson
Inuvik
Old Crow
Ross River
Teslin
multihousehold group
1
2
4
7
9
10

Table 3.2 Antimicrobial resistance genotypes present in at least one *H. pylori* genome

Gene	CDS Amino Acid Mutation	Resistance	Reference
23S rRNA	A2142G	Clarithromycin	Wang et al. 2018
23S rRNA	A2143G	Clarithromycin	Wang et al. 2018
rdxA oxidoreductase	truncation	Metronidazole	Wang et al. 2018
rdxA oxidoreductase	A118S	Metronidazole	Wang et al. 2018
Penicillin-binding protein 1A	T556S	Amoxicillin	De Palma et al. 2017
gyrA gyrase subunit A	N87K	Levofloxacin	Wang et al. 2018
gyrA gyrase subunit A	N87I	Levofloxacin	Wang et al. 2018
gyrA gyrase subunit A	D91G	Levofloxacin	Wang et al. 2018
gyrA gyrase subunit A	D91N	Levofloxacin	Wang et al. 2018

Figure 3.7. Upset plot depicting the intersection of antimicrobial resistance genotype markers in a) Manitoban and b) Arctic *H. pylori* isolates. Intersections of resistance marker presence are depicted by black dots connected by vertical lines. The counts of genomes containing each genetic marker known to confer phenotypic resistance, or genetic marker intersection are presented in the Intersection Size bar plot. The total counts of each genetic marker, from all intersections combined, are presented in the Set Size bar plot.

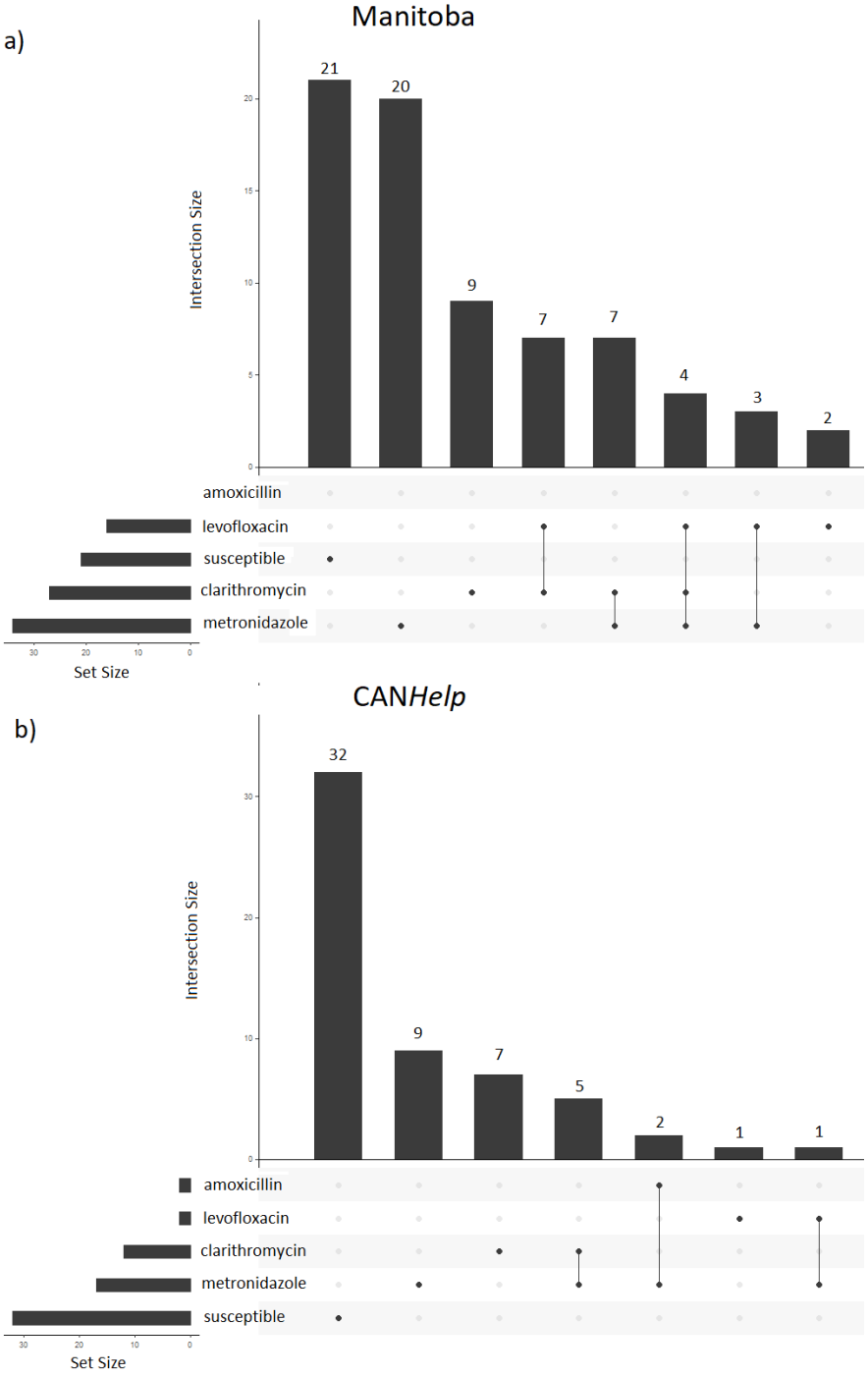
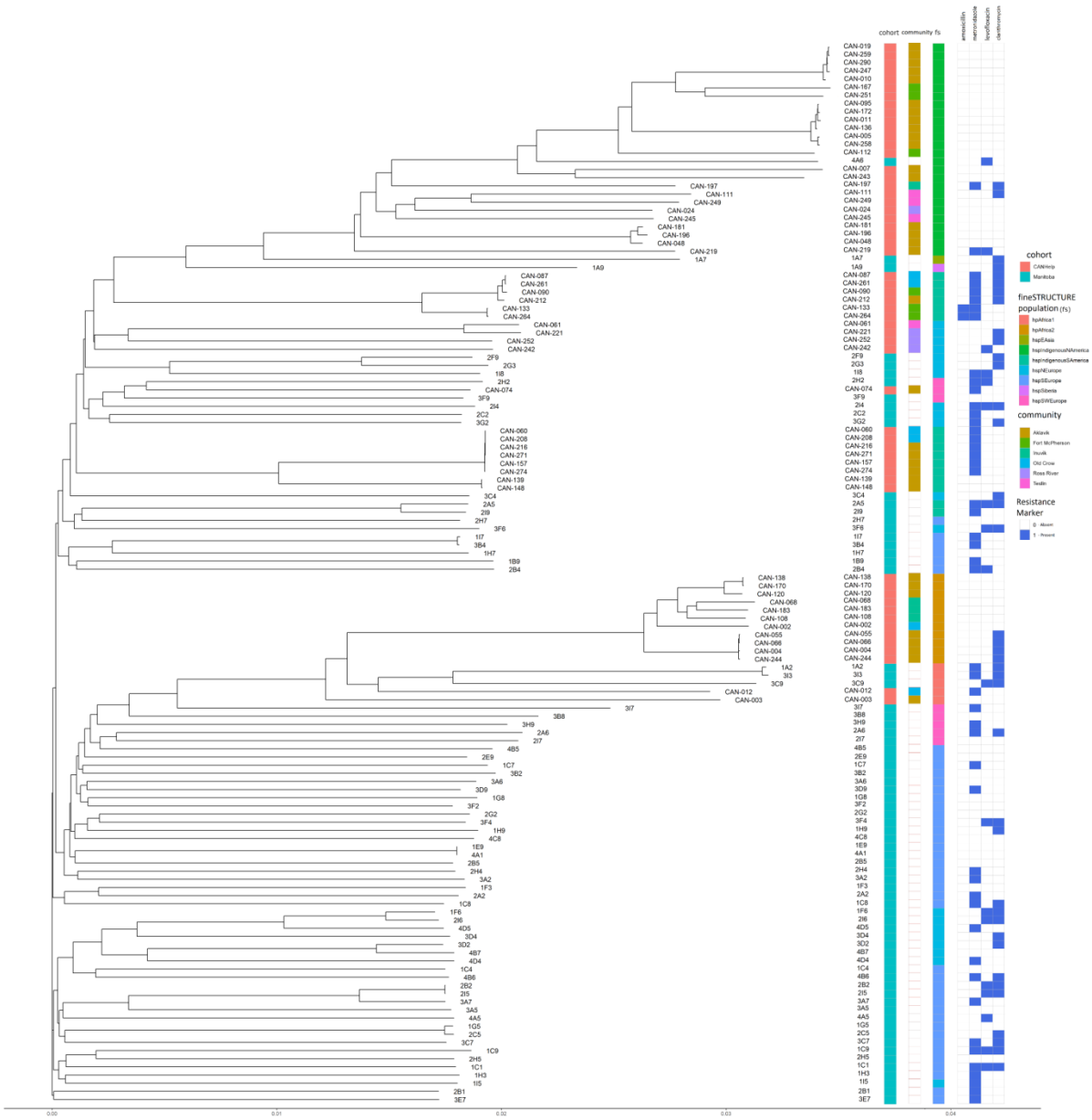

a)
Manitoba
Intersection Size
21
20
9
7
7
4
3
2
amoxicillin
levofloxacin
susceptible
clarithromycin
metronidazole
Set Size
b)
CANHelp
Intersection Size
32
9
7
5
2
1
1
amoxicillin
levofloxacin
clarithromycin
metronidazole
susceptible
Set Size

Figure 3.8. Phylogenetic tree of Arctic and Manitoba isolates annotated with genotypic antimicrobial resistance markers. Tree tips are labelled with cohort (Manitoban or CAN*Help*), community of origin, fineSTRUCTURE-determined sequence type, and presence (blue-shaded) or absence (white-shaded) of resistance markers for: amoxicillin, metronidazole, levofloxacin, and clarithromycin. Phylogeny was created using Mashtree with whole genomes as input. fineSTRUCTURE analysis was run using 769 core genes present in >90% of Manitoban, Arctic, and world isolates.

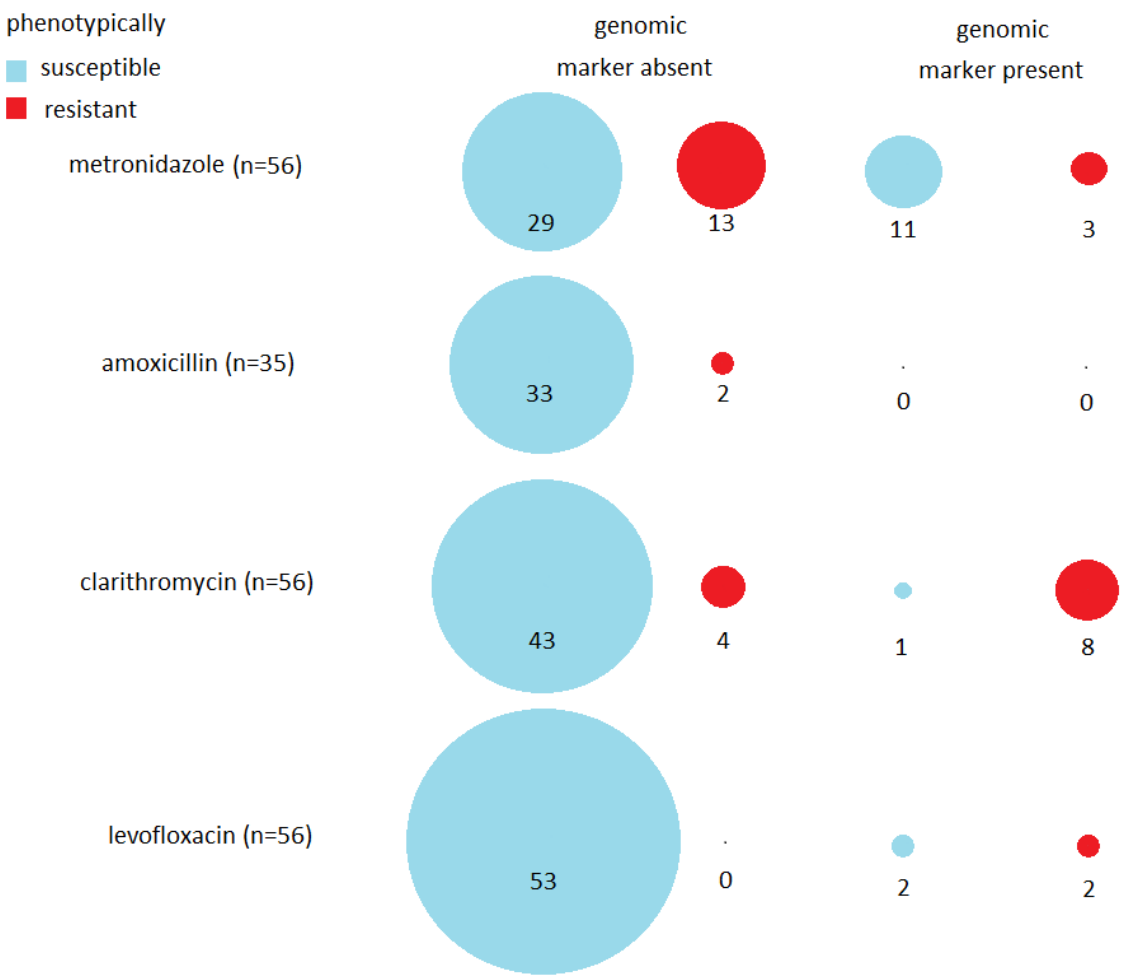

cohort
community
fs
amoxicillin
nemorubicin
levofloxacin
clarithromycin
CAN-019
CAN-259
CAN-290
CAN-247
CAN-010
CAN-167
CAN-251
CAN-095
CAN-172
CAN-011
CAN-136
CAN-005
CAN-258
CAN-112
4A6
CAN-007
CAN-243
CAN-197
CAN-111
CAN-249
CAN-024
CAN-245
CAN-181
CAN-196
CAN-048
CAN-219
1A7
1A9
CAN-087
CAN-261
CAN-090
CAN-212
CAN-133
CAN-264
CAN-061
CAN-221
CAN-252
CAN-242
2F9
2G3
1I8
2H2
CAN-074
3F9
2I4
2C2
3G2
CAN-060
CAN-208
CAN-216
CAN-271
CAN-157
CAN-274
CAN-139
CAN-148
3C4
2A5
2I9
2H7
3F6
1I7
3B4
1H7
1B9
2B4
CAN-138
CAN-170
CAN-120
CAN-068
CAN-183
CAN-108
CAN-002
CAN-055
CAN-066
CAN-004
CAN-244
1A2
3I3
3C9
CAN-012
CAN-003
3I7
3B8
3H9
2A6
2I7
4B5
2E9
1C7
3B2
3A6
3D9
1G8
3F2
2G2
3F4
1H9
4C8
1E9
4A1
2B5
2H4
3A2
1F3
2A2
1C8
1F6
2I6
4D5
3D4
3D2
4B7
4D4
1C4
4B6
2B2
2I5
3A7
3A5
4A5
1G5
2C5
3C7
1C9
2H5
1C1
1H3
1I5
2B1
3E7
0.00
0.01
0.02
0.03
0.04
cohort
CANHelp
Manitoba
fineSTRUCTURE
population (fs)
hpAfrica1
hpAfrica2
hspEAsia
hspIndigenousNAmerica
hspIndigenousSAmerica
hspNEurope
hspSEurope
hspSiberia
hspSWEurope
community
Aklavik
Fort McPherson
Inuvik
Old Crow
Ross River
Teslin
Resistance
Marker
0 - Absent
1 - Present

Only two isolates, CAN-133 and CAN-264, both Arctic isolates and clustered closely together, carry the genetic resistance marker associated with amoxicillin resistance. This is expected as amoxicillin resistance in *H. pylori* is rare. A large proportion of isolates from Manitoba, as well as the majority of Arctic isolates belonging to hspIndigenousSAmerica carried the genetic markers associated with metronidazole resistance. Notably, the Arctic clades with sequence types hspIndigenousNAmerica, hpAfrica2, and hspNEurope lacked these markers (**Figure 3.8**). The presence of levofloxacin and clarithromycin genetic resistance markers had no clear phylogenetic association in the tree.

Of interest was any possible association between the presence of genetic antimicrobial resistance markers and phenotypic resistance within bacterial isolates. Unfortunately, no resistance phenotypic data was available for the Manitoban isolates. Phenotypic resistance testing for the Arctic isolates was performed when isolates were first isolated by the CAN*Help* Working Group. This data is presented in **Figure 3.9**.

There is a notable lack of concordance between marker presence and phenotypic resistance (as determined by laboratory testing) for metronidazole, amoxicillin, clarithromycin, and levofloxacin. Of the 14 isolates with metronidazole resistance genomic markers, 11 were susceptible to metronidazole. Conversely, of the 42 isolates that lacked the genomic markers, 13 were resistant. The lack of presence of PBP-1A mutations in all isolates limited the data available for amoxicillin resistance. Of the 35 isolates with phenotypic resistance testing, only 2 were resistant. Of the 9 isolates carrying clarithromycin resistance markers, only 1 was susceptible to the antimicrobial. Of the 47 that lacked clarithromycin resistance markers, 4 were resistant. Finally, while all isolates that did not carry levofloxacin resistance markers were susceptible, 2 out of the 4 isolates that did carry the marker were resistant.

Figure 3.9. Comparison between presence of genomic antimicrobial resistance markers and phenotypic resistance detected in Arctic and Manitoban *H. pylori*. The genomic resistance markers were based on the following: *rdxA* A118S or truncation (metronidazole), 23S rRNA A2142G or A2143G (clarithromycin), *gyrA* N87K, N87I, D91G, or D91N (levofloxacin), and penicillin-binding protein 1A T556S (amoxicillin).


phenotypically
susceptible
resistant
genomic
marker absent
genomic
marker present
metronidazole (n=56)
29
13
11
3
amoxicillin (n=35)
33
2
.
0
.
0
clarithromycin (n=56)
43
4
1
8
levofloxacin (n=56)
53
.
0
2
2

3.3.2 Virulence factors

cagA and *cag*PI status are presented in **Figure 3.10** and **Table A3.5** in the appendix**,** while **Figure 3.11** depicts the phylogenetic structure of all isolates annotated with *cagA* and *cag*PI status. All genomes contained the *vacA* gene. The *vacA* allele and genotype status is presented with phylogenetic structure in **Figure 3.12**, as an upset plot in **Figure 3.13** and **Table A3.6** in the appendix.

Based on the phylogenetic structure, *cagA* and *cag*PI have phylogenetic patterns in both Manitoba and Arctic isolates (**Figure 3.11**). For the majority of Manitoban isolates belonging to the hspSEurope sequence type, they lacked both the *cag*PI and *cagA* effector gene. For hspNEurope isolates, on the other hand, the majority possessed both the pathogenicity island and *cagA* effector gene. For the Arctic *H. pylori* isolates, almost every genome within the hpAfrica1 and hpAfrica2 sequence types carried both the *cagA* effector gene and *cag*PI. Every isolate within the mixed hspNEurope/hspIndigenousSAmerica clade carried both *cagA* and the *cag*PI. In contrast, only two isolates in the hspIndigenousNAmerica clade carried an intact *cag*PI, and of those two, only one carried an intact copy of *cagA*. The rest of the isolates within this clade were negative for both cancer-associated genetic elements. In both the Manitoban and CAN*Help* cohorts, the EPIYA motif that was dominant in most genomes was the less-virulent Western ABC genotype. A small number of Manitoban genomes possessed the more virulent, East Asian D EPIYA motif type, either in a mosaic ABCD genotype or in a truncated CD genotype. No CAN*Help* genomes contained the D genotype.

Figure 3.10. *cagA* presence/absence, EPIYA motif type, and *cag*PI presence/absence. Figure displays data for a) Manitoban cohort and b) CAN*Help* cohort. i) represents *cagA* status as a pie chart. ii) represents *cag*PI status as a pie chart. iii) represents combined *cagA* and *cag*PI status as a pie chart. iv) is a bar chart depicting the total number of *cagA* EPIYA motifs present in *cagA*+ genomes.

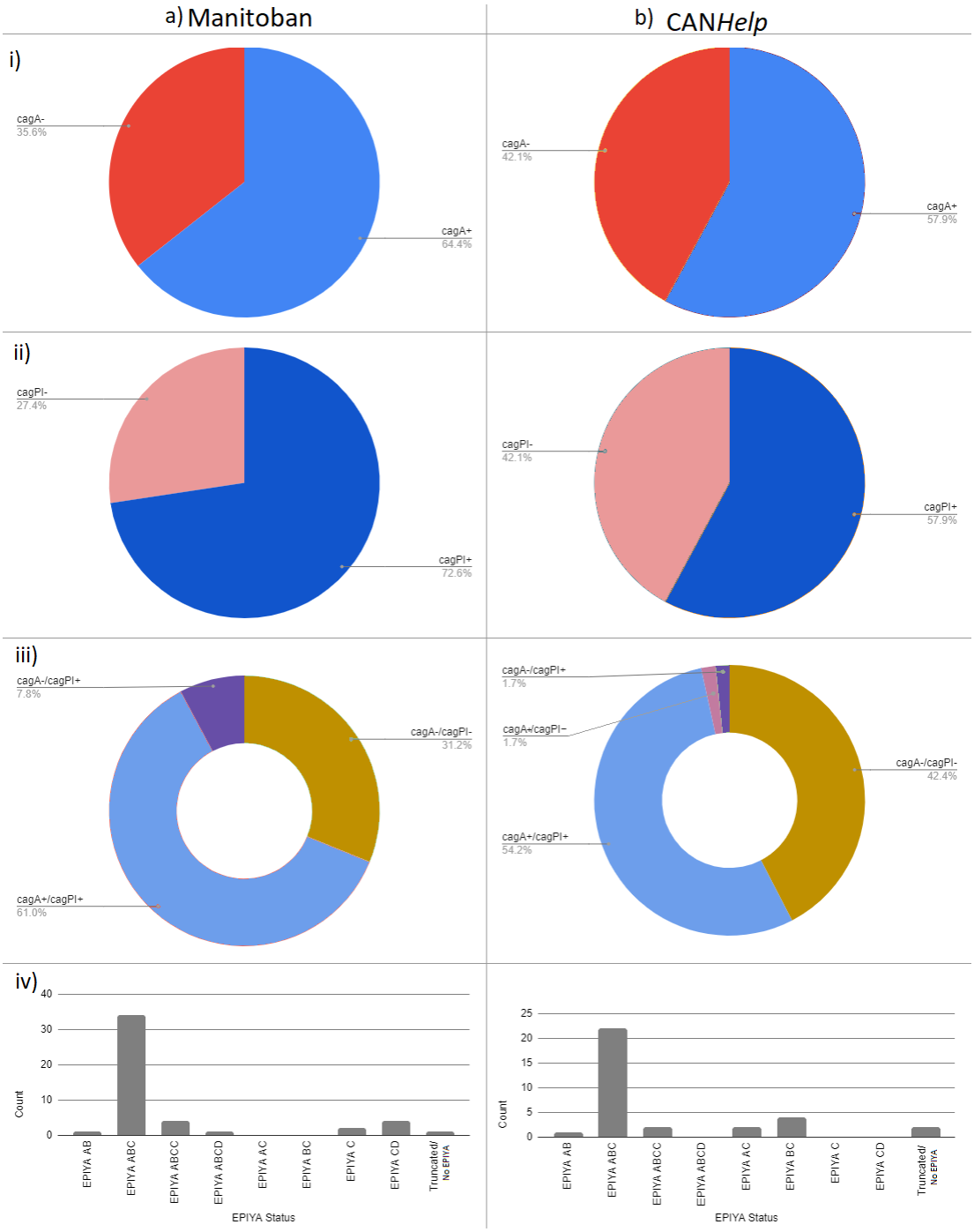

a) Manitoban
b) CANHelp
i)
cagA-
35.6%
cagA+
64.4%
cagA-
42.1%
cagA+
57.9%
ii)
cagPI-
27.4%
cagPI+
72.6%
cagPI-
42.1%
cagPI+
57.9%
iii)
cagA-/cagPI+
7.8%
cagA-/cagPI-
31.2%
cagA+/cagPI+
61.0%
cagA-/cagPI+
1.7%
cagA+/cagPI−
1.7%
cagA-/cagPI-
42.4%
cagA+/cagPI+
54.2%
iv)
Count
EPIYA Status
EPIYA AB
EPIYA ABC
EPIYA ABCC
EPIYA ABCD
EPIYA AC
EPIYA BC
EPIYA C
EPIYA CD
Truncated/ No EPIYA

Figure 3.11 Phylogenetic tree of Arctic and Manitoba isolates annotated with *cagA* EPIYA motif type and *cag*PI status. Tree tips are labelled with cohort (Manitoban or CAN*Help*), community of origin (when known), fineSTRUCTURE-determined sequence type, *cag*PI status, *cagA* status, and *cagA* EPIYA motif genotype. For *cag*PI or *cagA* to be called "present" (1-blue shading), they had to be intact sequences, without any truncations. Phylogeny was created using Mashtree with whole genomes as input. fineSTRUCTURE analysis was run using 769 core genes present in >90% of Manitoban, Arctic, and world isolates.

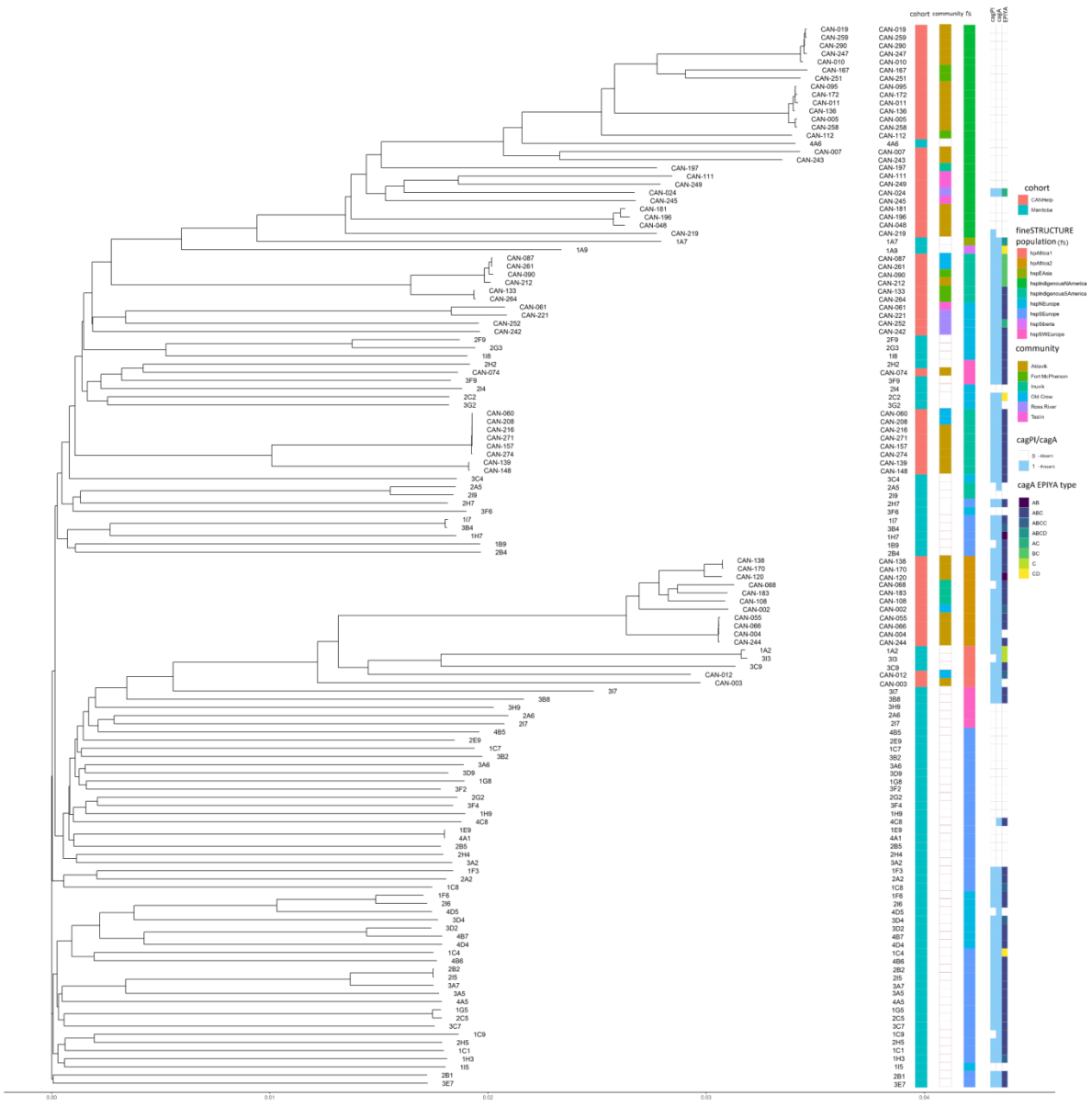

cohort community fs
cagPI cagA EPIYA
cohort
CANHelp
Manitoba
fineSTRUCTURE
population (fs)
hpAfrica1
hpAfrica2
hpEAsia
hpIndigenousNAmerica
hpIndigenousSAmerica
hpNEurope
hpSEurope
hpSiberia
hpSWEurope
community
Aklavik
Fort McPherson
Inuvik
Old Crow
Ross River
Tsiigehtchic
Tsiigehtchic
Tsiigehtchic

Figure 3.12. Phylogenetic tree of Arctic and Manitoba isolates annotated with *vacA s* and *m* genotype. Tree tips are labelled with cohort (Manitoban or CAN*Help*), community of origin, fineSTRUCTURE-determined sequence type, s genotype and m genotype. Phylogeny was created using Mashtree with whole genomes as input. fineSTRUCTURE analysis was run using 769 core genes present in >90% of Manitoban, Arctic, and world isolates.

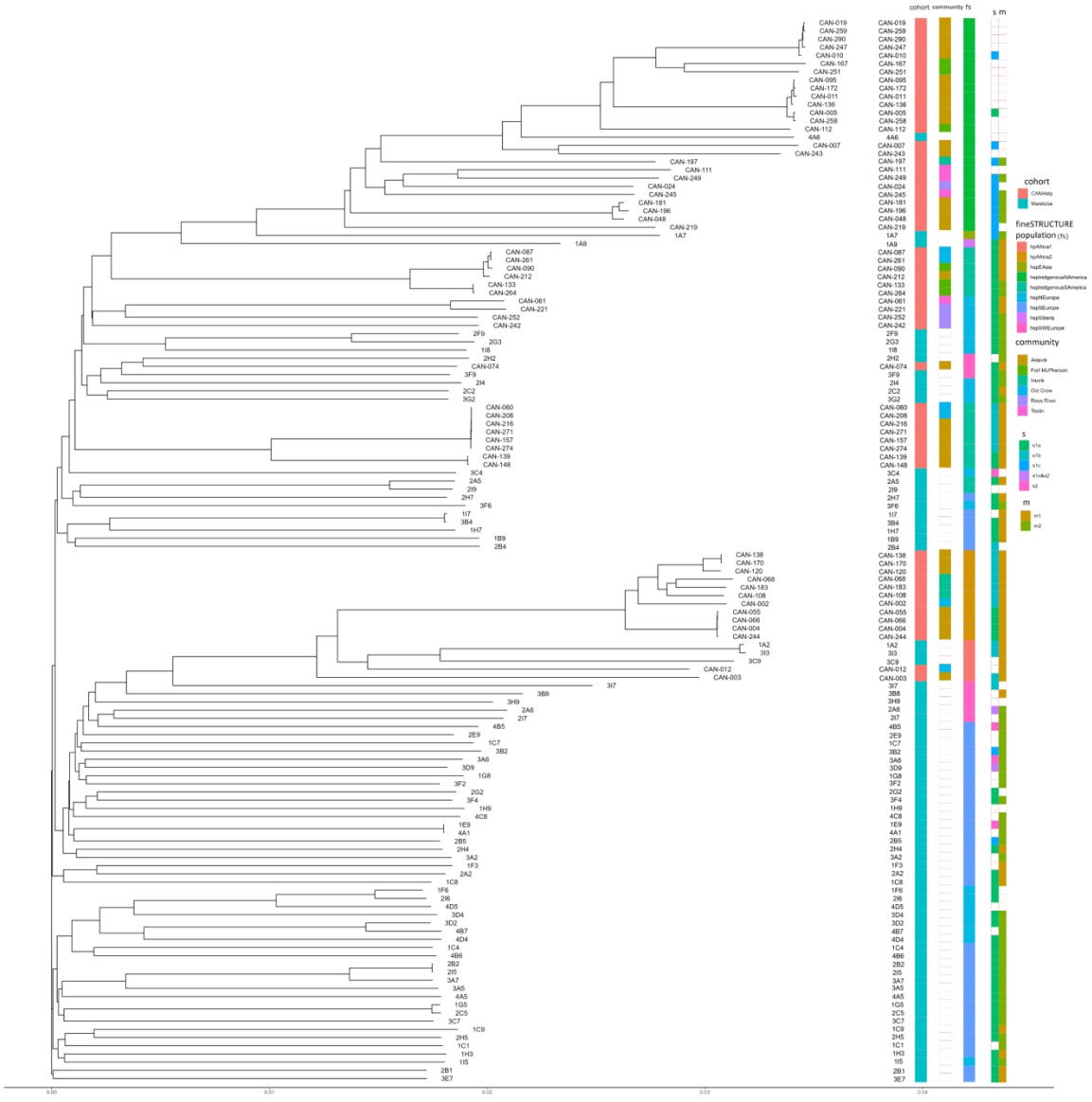
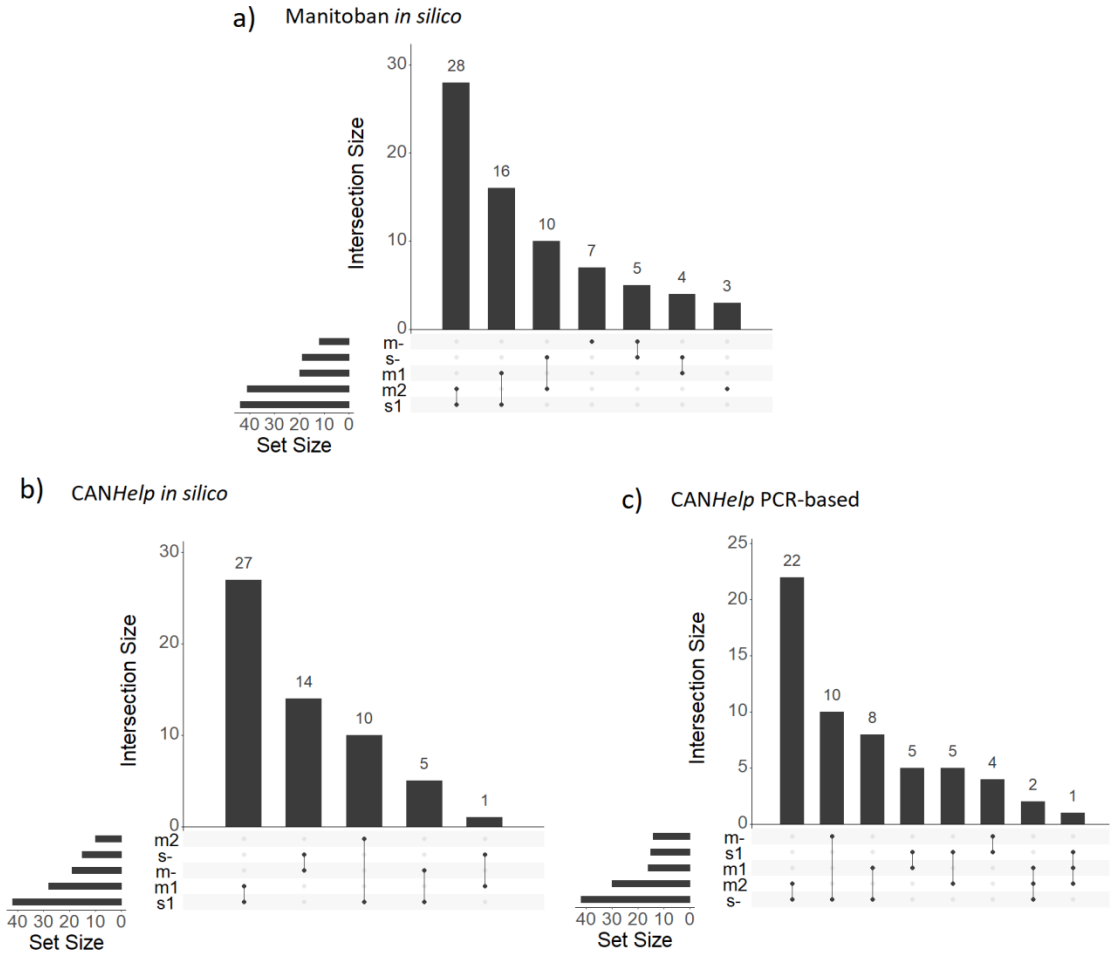

cohort
community
fs
s
m
CAN-019
CAN-259
CAN-290
CAN-247
CAN-010
CAN-167
CAN-251
CAN-095
CAN-172
CAN-011
CAN-136
CAN-005
CAN-258
CAN-112
4A6
CAN-007
CAN-243
CAN-197
CAN-111
CAN-249
CAN-024
CAN-245
CAN-181
CAN-196
CAN-048
CAN-219
1A7
1A9
CAN-087
CAN-261
CAN-090
CAN-212
CAN-133
CAN-264
CAN-061
CAN-221
CAN-252
CAN-242
2F9
2G3
1I8
2H2
CAN-074
3F9
2I4
2C2
3G2
CAN-060
CAN-208
CAN-216
CAN-271
CAN-157
CAN-274
CAN-139
CAN-148
3C4
2A5
2I9
2H7
3F6
1I7
3B4
1H7
1B9
2B4
CAN-138
CAN-170
CAN-120
CAN-068
CAN-183
CAN-108
CAN-002
CAN-055
CAN-066
CAN-004
CAN-244
1A2
3I3
3C9
CAN-012
CAN-003
3I7
3B8
3H9
2A6
2I7
4B5
2E9
1C7
3B2
3A6
3D9
1G8
3F2
2G2
3F4
1H9
4C8
1E9
4A1
2B5
2H4
3A2
1F3
2A2
1C8
1F6
2I6
4D5
3D4
3D2
4B7
4D4
1C4
4B6
2B2
2I5
3A7
3A5
4A5
1G5
2C5
3C7
1C9
2H5
1C1
1H3
1I5
2B1
3E7
cohort
CANHelp
Manitoba
fineSTRUCTURE
population (fs)
hpAfrica1
hpAfrica2
hpEAsia
hpIndigenousAfrica
hpIndigenousSAmerica
hpNEurope
hpSEurope
hpSiberia
hpSWEurope
community
Aklavik
Fort McPherson
Inuvik
Old Crow
Ross River
Teslin
s
s1a
s1b
s1c
s1c&s2
s2
m
m1
m2
0.50
0.51
0.52
0.53
0.54

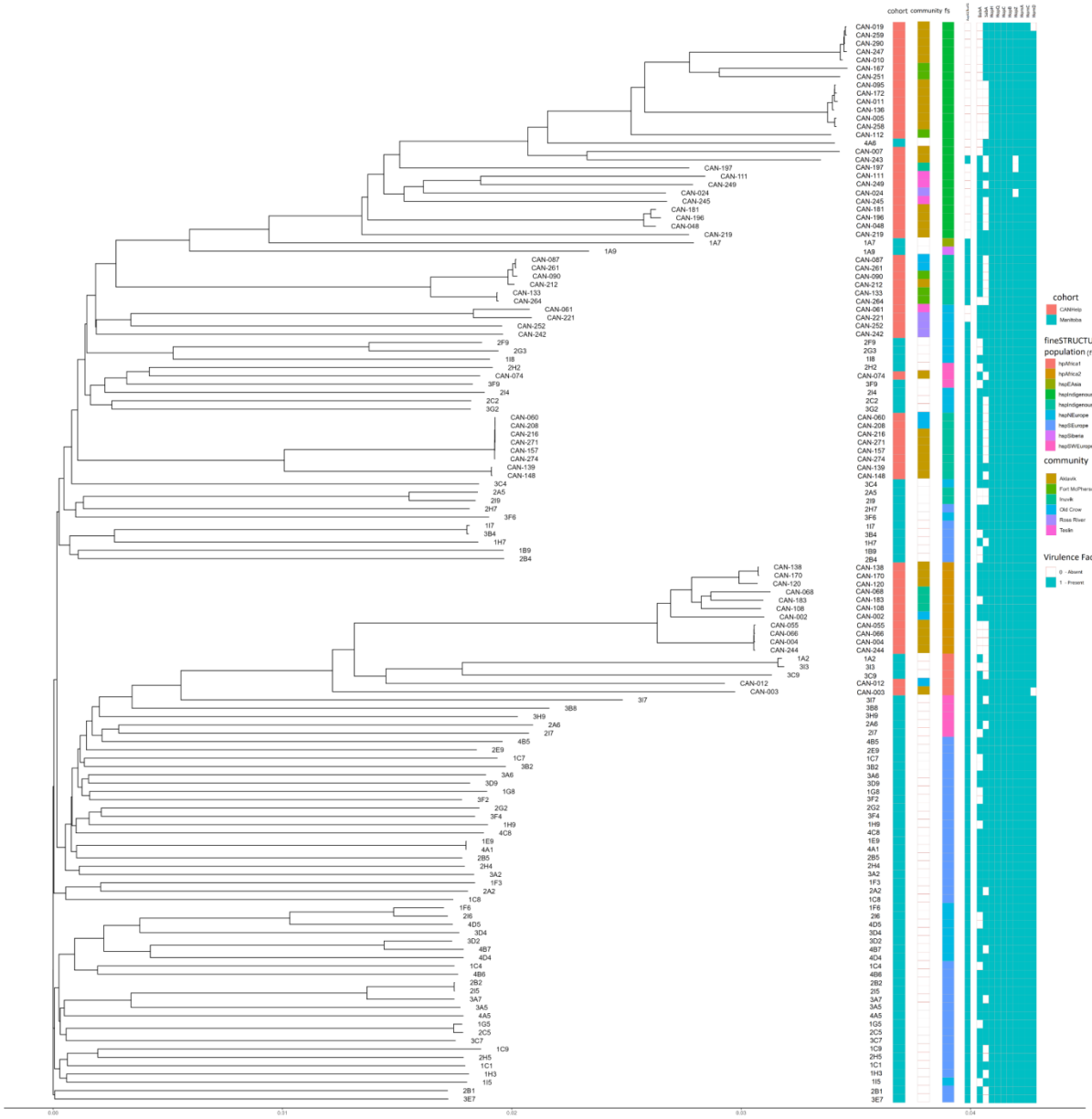
Figure 3.13. Upset plot of *vacA* allele and genotype status by both *in silico* and lab-based PCR. Plots are for a) Manitoba isolates determined *in silico* b) CAN*Help* isolates determined *in silico* c) CAN*Help* isolates determined by PCR. Intersections of *vacA* genotype allele presence are depicted by black dots connected by vertical lines. The counts of genomes containing each allele (s-; s1; m-; m1, or m2) or allele intersection (s-/m-; s-/m1; s-/m2; s1/m-; s1/m1; s1/m2, or s1/m1&m2) are presented in the Intersection Size bar plot. The total counts of each allele, from all intersections combined, are presented in the Set Size bar plot.


a) Manitoban *in silico*
Intersection Size
30
20
10
0
28
16
10
7
5
4
3
m-
s-
m1
m2
s1
40 30 20 10 0
Set Size
b) CAN*Help in silico*
Intersection Size
30
20
10
0
27
14
10
5
1
m2
s-
m-
m1
s1
40 30 20 10 0
Set Size
c) CAN*Help* PCR-based
Intersection Size
25
20
15
10
5
0
22
10
8
5
5
4
2
1
m-
s1
m1
m2
s-
40 30 20 10 0
Set Size

Of interest is the presence of the less virulent s2 and m2 *vacA* alleles within our study genomes. There was a small prevalence of the less-virulent s2 *vacA* allele in both the Arctic and Manitoba genomes. Notably, the presence of the less-virulent m2 allele was much more prevalent in Manitoba genomes (57.5%) as compared to Arctic genomes (17.5%). As seen in **Figure 3.12**, while the distribution of s and m genotypes is rather homogenous, with certain clades being either predominantly m1 or m2, the large hspIndigenousNAmerica clade contains genomes that are almost entirely s-/m-, with only a handful of genomes being s1/m- or s1/m2.

The presence or absence of both the toxin-antitoxin *aapA1*/*isoA1*, as well as a select few outer membrane proteins also showed phylogenetic patterns. As presented in **Figure 3.14**, the toxin-antitoxin *aapA1*/*isoA1* was present in almost every single phylogenetic clade except for a doublet of CAN*Help* isolates (CAN-061 and -221) and additionally the large clade of hspIndigenousNAmerica isolates. The absence of *babA* and *sabA* displayed less strong phylogenetic patterns, but still indicated an overall absence in certain CAN*Help* subclades, while being predominantly present in Manitoban isolates.

Figure 3.14. Phylogenetic tree of Arctic and Manitoba isolates annotated with presence or absence of the toxin-antitoxin *aapA1*/*isoA1* and OMPs. OMPs analyzed include *babA*/*hopS*, *sabA*/*hopP*, *hopH*/*oipA*, *hopQ*, *alpA*/*hopC*, *alpB*/*hopB*, *homA*, *homC*, and *homD.* Tree tips are labelled with cohort (Manitoban or CAN*Help*), community of origin, fineSTRUCTURE-determined sequence type, and presence (blue-shaded) or absence (white-shaded) of these virulence factors. Phylogeny was created using Mashtree with whole genomes as input. fineSTRUCTURE analysis was run using 769 core genes present in >90% of Manitoban, Arctic, and world isolates.


cohort
community
fs
CAN-019
CAN-259
CAN-290
CAN-247
CAN-010
CAN-167
CAN-251
CAN-095
CAN-172
CAN-011
CAN-136
CAN-005
CAN-258
CAN-112
4A6
CAN-007
CAN-243
CAN-197
CAN-111
CAN-249
CAN-024
CAN-245
CAN-181
CAN-196
CAN-048
CAN-219
1A7
1A9
CAN-087
CAN-261
CAN-090
CAN-212
CAN-133
CAN-264
CAN-061
CAN-221
CAN-252
CAN-242
2F9
2G3
1I8
2H2
CAN-074
3F9
2I4
2C2
3G2
CAN-060
CAN-208
CAN-216
CAN-271
CAN-157
CAN-274
CAN-139
CAN-148
3C4
2A5
2I9
2H7
3F6
1I7
3B4
1H7
1B9
2B4
CAN-138
CAN-170
CAN-120
CAN-068
CAN-183
CAN-108
CAN-002
CAN-055
CAN-066
CAN-004
CAN-244
1A2
3I3
3C9
CAN-012
CAN-003
3I7
3B8
3H9
2A6
2I7
4B5
2E9
1C7
3B2
3A6
3D9
1G8
3F2
2G2
3F4
1H9
4C8
1E9
4A1
2B5
2H4
3A2
1F3
2A2
1C8
1F6
2I6
4D5
3D4
3D2
4B7
4D4
1C4
4B6
2B2
2I5
3A7
3A5
4A5
1G5
2C5
3C7
1C9
2H5
1C1
1H3
1I5
2B1
3E7
0.00
0.01
0.02
0.03
0.04
cohort
CANHelp
Manitoba
fineSTRUCTURE
population (fs)
hpAfrica1
hpAfrica2
hspEAsia
hspIndigenousNAmerica
hspIndigenousSAmerica
hspNEurope
hspSEurope
hspSiberia
hspSWEurope
community
Aklavik
Fort McPherson
Inuvik
Old Crow
Ross River
Teslin
Virulence Factors
0 - Absent
1 - Present

3.3.3 Genomic islands

The presence of various genomic islands within the Arctic and Manitoban genomes are presented in **Figure 3.15**. What constitutes presence and any annotated genomic content that is piggybacked (within each island denoted in the figure) is detailed in appendix **Table A3.7 Genomic Island Data Dictionary**. The phage/colicin/tellurite resistance cluster *terY* protein gene was present within both Arctic hspIndigenousSAmerica subclades (comprised of CAN-087, -261, -090, -212, -133, and -264), yet lacked a clear phylogenetic pattern in all other Arctic and Manitoba isolates. A handful of isolates, predominantly Manitoban, contained a genomic island (GI) carrying the *tetR* tetracycline repressor resistance-associated gene. The Arctic isolates CAN-068 and CAN-167, however, did not display phenotypic resistance to tetracycline in the lab. Two Manitoba isolates contained genomic islands carrying the *spaB* multi-drug resistance associated ABC transporter. A handful of isolates carried a genomic island containing the *lptB* ABC transporter gene, however, like the *spaB* GIs, have no clear phylogenetic patterns. A large proportion of the identified genomic islands within Arctic and Manitoban genomes carried outer membrane proteins that may be associated with adhesion to host (Cover 2016), including *homD*, *frpB*, *hop*, *hor*, and *hof* family genes, and *babA* and *babC* (Aspholm-Hurtig et al. 2004; Dadaschzadeh et al. 2017; Berthenet et al. 2018). These GIs did not present strong phylogenetic trends, except for a family of GIs containing the *hofC* and *hofD*, which were found in the majority of one Arctic hspIndigenousNAmerica subclade (comprised of CAN-197, -111, -249, -245,196, 048) and one hspIndigenousSAmerica subclade (comprised of CAN-087, -261, -090, -212, -133, and -264). The same isolates within the hspIndigenousSAmerica subclade carrying the *terY*-containing genomic island also carried a distinct genomic island containing *leoA*, a virulence-associated toxin gene, which has the

capacity of mediating *H. pylori* virulence by dynamin-like activity resulting in membrane remodelling and potential toxin secretion and outer membrane protein localization (Michie et al. 2014).

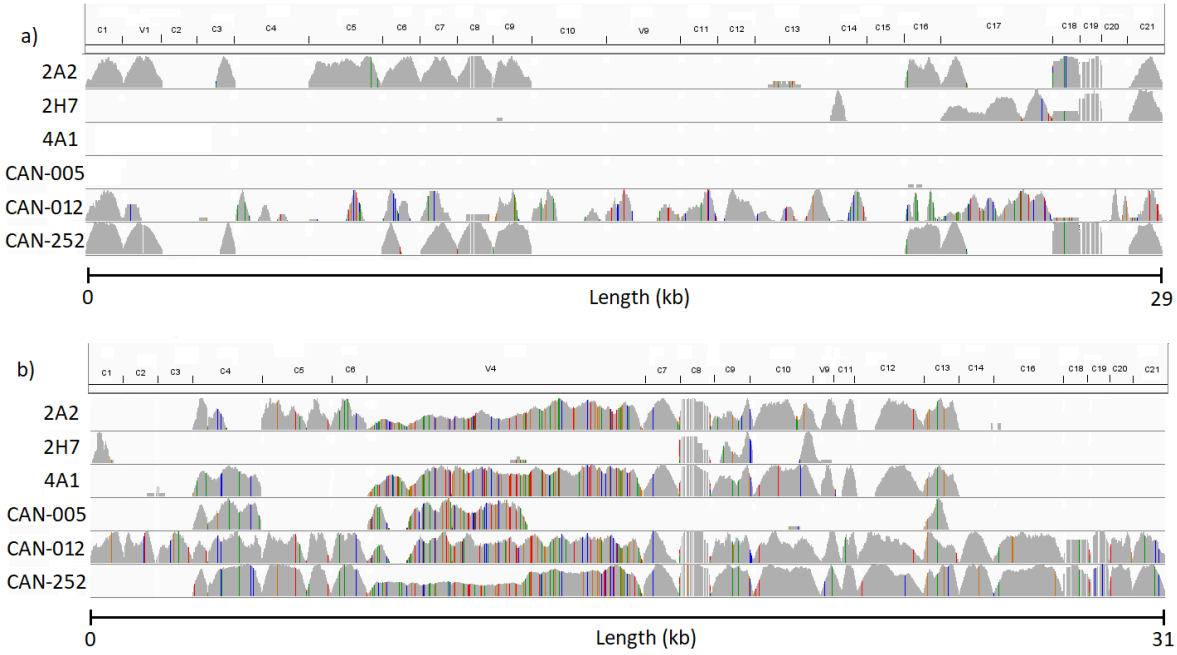
Two notable genomic islands seen were those containing the *tfs* T4SS, which is homologous to the *cag* pathogenicity island *cag*T4SS yet distinct. Notably, a large proportion of Manitoban isolates carrying the *cag*PI also carried this *tfs* GI, while most *cag*PI+ Arctic isolates lacked this GI. Based on their gene content, these identified genomic islands are full or partial (degraded) copies of the integrative and conjugative elements (ICEs) in *H. pylori* known as *tfs3* and *tfs4*. These genomic islands are identified by whether they carry alleles of ICE genes associated with *tfs3* and/or *tfs4*. A genomic island carrying a beta-lactamase gene was found in all Arctic genomes within a hspIndigenousNAmerica subclade (comprised of CAN-095, -172, -011, -136, -005, and -258), which all also carried the *tfs3/4* T4SS ICE, despite not being *cag*PI+.

To better understand the completeness of these ICE elements in Manitoban and CAN*Help* cohort genomes, the genetic content of Manitoba cohort genomes 2A2, 2H7, and 4A1, as well as CAN*Help* cohort genomes CAN-005, -012, and -252 were compared to the coding sequences of the *H. pylori* reference *tfs3* from hpAfrica1-representative isolate Gambia94/24 and the reference *tfs4* from hpEurope-representative isolate G27, similar to Delahay, Croxall and Stephens (2018). These genomes were selected due to their phylogenetic separation and distinct fineSTRUCTURE sequence type assignments. The results of these selected alignments are presented in **Figure 3.16**. While coverage of the *tfs3* gene content was sparse in these genomes (panel a), coverage of *tfs4* was much more complete, with CAN-012 showing near complete coverage of the *tfs4* ICE (panel b). As well, it

Figure 3.15. Presence of various genomic islands within CAN*Help* and Southern Manitoban *H. pylori*. Tree tips labelled with both cohort (Manitoban or CAN*Help*), community of origin (when known), fineSTRUCTURE-determined sequence type, and presence (various, coloured) or absence (0-white shading) of genomic islands in the genomes. Higher values in the matrix indicate the presence of a genomic island (in part or full), alongside other genes of interest co-appearing on the mobile element. See appendix Table A3.7 Genomic Island Data Dictionary for data dictionary. Phylogeny was created using Mashtree with whole genomes as input. fineSTRUCTURE analysis was run using 769 core genes present in >90% of Manitoban, Arctic, and world isolates.

cohort
CANHelp
Montcalm
fineSTRUCTURE
population (fs)
hpAfrica1
hpAfrica2
hepEAsia
hepIncigenousNAmerica
hepIncigenousSAmerica
hepNEurope
hepSEurope
hepSWeste
hepSWEurope
community
Ardarik
Fort McPherson
Inuvik
Old Crow
Rosa River
Tsiigehtchic
genomic islands
0
1
2
3
4

Figure 3.16. Alignment of ICE a) *tfs3* and b) *tfs4* containing genomic content in a selected subset of six Manitoba and CAN*Help* cohort genomes. Read mapping alignment for the coding sequences of *tfs3* and *tfs4* ICE was visualized in the Integrative Genomics Browser (Robinson et al. 2011). The appropriate *tfs*-containing sequence (segment of genome) was selected for read mapping each isolate: for a) *tfs3* ICE from hpAfrica1-representative isolate Gambia94/24 (genes 7340–7535) for b) *tfs4* ICE from hpEurope-representative isolate G27 (genes 0959–0986) were used as reference for the ICE alignments. Grey shaded areas depict reads mapping to the reference sequence with 100% sequence identity; read map areas of colour (*i.e*. vertical coloured bars) indicate locations of nucleotide substitution, causing a single nucleotide variant SNV). Colours for SNV nucleotides: A (red); C (green); G (blue); T (yellow). Gene labels and corresponding functions are as follows (Delahay, Croxall and Stephens 2018): C1 — integrase-recombinase; V4 — DNA methylase; C7 — *virC1*/*parA*; C9 — *virD2* relaxase; C10 — *virD4* DNA transfer protein; C11 — *virB11* T4SS ATPase; C12 — *virB10*/*comB10* competence protein; C13 — *virB9*/*comB9* T4SS secretion system protein; C14 — *virB8*/*comB8* competence protein; C15 — *virB7*/*comB7* T4SS secretion system protein; C16 — *topA* topoisomerase I; C17 — *virB4*/*comB4* DNA transfer protein; C18 — *virB3*/*comB3* T4SS secretion system protein; C19 — *virB2*/*comB2* T4SS secretion system protein V1, V9, C2, C3, C4, C5, C6, C8, C20, C21 — hypothetical protein.


a)
C1 V1 C2 C3 C4 C5 C6 C7 C8 C9 C10 V9 C11 C12 C13 C14 C15 C16 C17 C18 C19 C20 C21
2A2
2H7
4A1
CAN-005
CAN-012
CAN-252
0
Length (kb)
29
b)
C1 C2 C3 C4 C5 C6 V4 C7 C8 C9 C10 V9 C11 C12 C13 C14 C16 C18 C19 C20 C21
2A2
2H7
4A1
CAN-005
CAN-012
CAN-252
0
Length (kb)
31

is interesting to note a large amount of genetic diversity within the *tfs4* ICE, particularly within the V4 - methylase gene.

3.3.4 Plasmid content

The full presence of various plasmids within the Arctic and Manitoban genomes are presented in **Figure 3.17.** The presence of certain plasmids within tightly clustered Arctic subclades, for example pNC_014161 within hspIndigenousNAmerica Arctic genomes (comprised of CAN-172, -011, -136, -005, and -258) gives further evidence of a close genetic relationship and likely recent common ancestor for these genomes. This also can be inferred from the co-presence of certain plasmid pairs, for example pAB769 and pAG532 in hspIndigenousNAmerica isolates (comprised of CAN-019, -259, -290, and -247), or pAG531 and pCP000242 in hspIndigenousSAmerica isolates (comprised of CAN-060, -208, -216, -271, -157, and -274).

Plasmid pAB503 was found in all isolates in a sub-clade of the hspAfrica1 clade of Arctic isolates (comprised of CAN-138, -170, -120, -068, -183, and -108). However, this plasmid lacked any notable coding sequences and should be further characterized with long-read Oxford Nanopore Technologies MinION or another long read sequencing method. Likewise, partial versions of plasmids pAG531 and pCP00242 were seen in the Arctic hpAfrica1-containing clade (comprised of -216, -271, -157, and -274) and Arctic hspNEurope isolates (comprised of CAN-060 and -208). The non-intact plasmid pAB769 was present in a number of hspNAmericanIndigenous isolates (namely CAN-019, -259, and -290). Each of these plasmids lacked some notable coding regions and should be further investigated.

The pAB768 plasmid is predicted to be mobilizable in Manitoban isolate 2I7, whereas non-mobilizable in CAN-219, potentially due to partial loss of content. Although

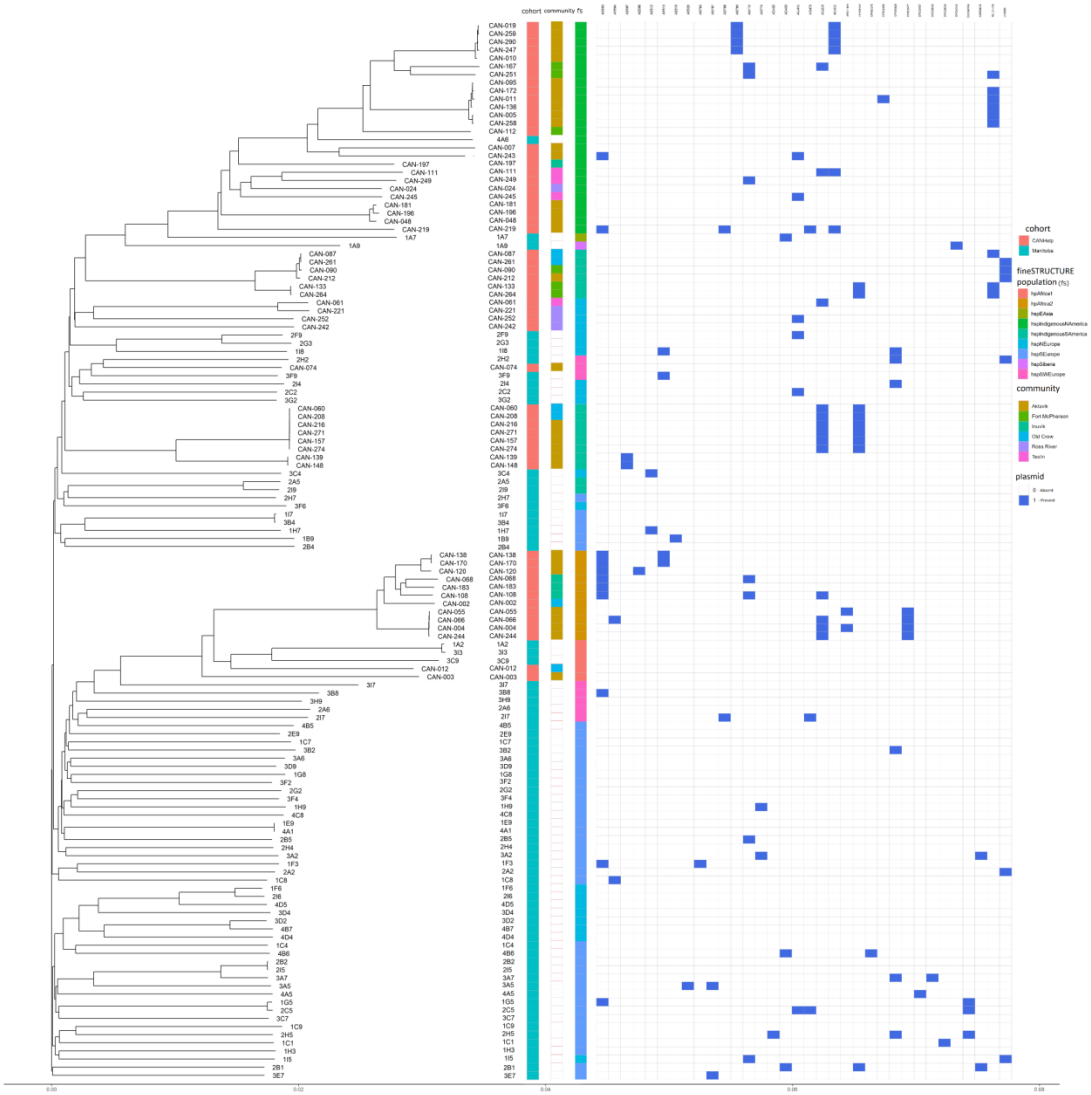
these two isolates do not share phylogenetic connections, CAN-219 is tightly clustered with aklavik117, which lacks this plasmid. These two plasmid versions contain the *Bacillus subtilis* putative multidrug resistance ABC transporter *yheH*.

The pAB513 plasmid is inferred as mobilizable in Manitoba isolate 3F9 and likely non-mobilizable (truncated) in 1I8, CAN-138, and CAN-170. Both CAN-138 and CAN-170 are closely related isolates in a subclade of Arctic hspAfrica1 isolates. The Manitoba genomes 1I8 and 3F9, however, do not share phylogenetic connections with any other isolates containing pAB513. This plasmid is of possible interest owing to presence of yet another coding sequence of unknown function related to *Bacillus subtilis* ABC transporter: *yxdL*.

Of biological interest (based on gene content) were the plasmids pAG450, pAB768, and pAB513. The complete pAG450 plasmid was only present in the Arctic isolates CAN-245 and CAN-252. Plasmid pAG450 present in Manitoba isolates 2C1 and 2C2 is inferred as mobilizable. The plasmid in Manitoba isolates 2C5 and 2F9 is inferred as non-mobilizable. Arctic isolates CAN-243, CAN-245, and CAN-242 contain a conjugative copy of the plasmid. And Manitoba isolates 1I8, 2H2, 2H5, 2I4, 3A7, and 3B2, contain a highly truncated but mobilizable copy of pAG450. These plasmids are closely related to pXZ274, characterized and sequenced from a Tibetan gastric cancer patient (Guo et al. 2012). The pXZ274 plasmid is of interest because it encodes for a T4SS, possessing the T4SS protein *ptlC*, as well as *virB10*, *virB11*, *virC1*, and *virD4*, known homologues of genes in the *cag*PI (Guo et al. 2012). Truncated plasmids found in Manitoba isolates 1I8, 2F9, 2H2, 2H5, 2I4, 3A7, and 3B2 contained few of the T4SS-associated loci, however.

A Mashtree phylogeny with masking out of the plasmid regions was re-created (figure not shown), and the resulting phylogenies had no differences in tree structure from the ones produced using unmasked genomes (*i.e*. containing the plasmid content).

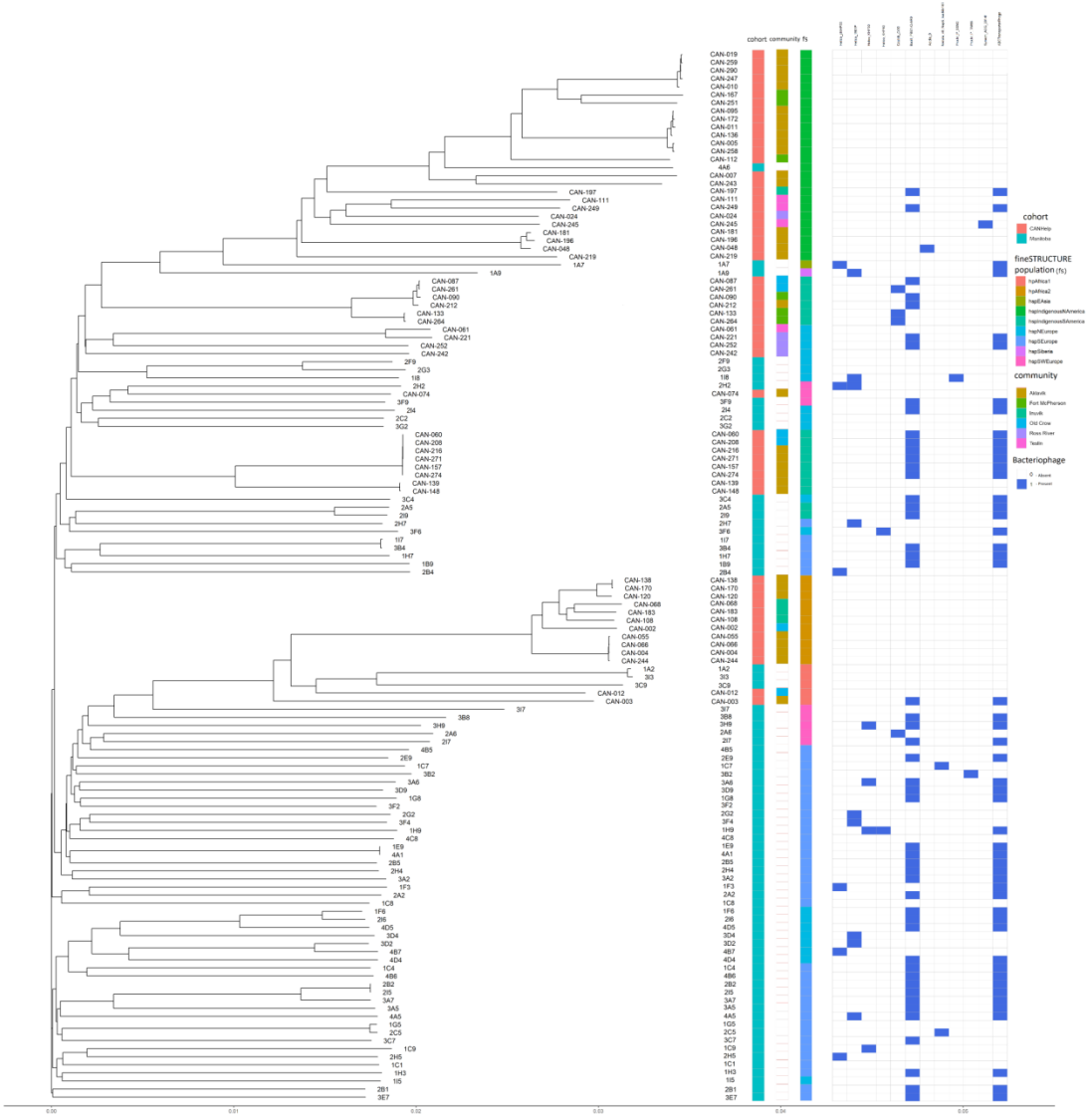
Figure 3.17. Presence of various plasmids within CAN*Help* and Southern Manitoban *H. pylori*. Tree tips labelled with both cohort (Manitoban or CAN*Help*), community of origin (when known), fineSTRUCTURE-determined sequence type, and presence (blue-shaded) or absence (white-shaded) of various plasmids. Phylogeny was created using Mashtree with whole genomes as input. fineSTRUCTURE analysis was run using 769 core genes present in >90% of Manitoban, Arctic, and world isolates.


cohort
community
fs
CAN-019
CAN-258
CAN-290
CAN-247
CAN-010
CAN-167
CAN-251
CAN-095
CAN-172
CAN-011
CAN-138
CAN-005
CAN-258
CAN-112
4A6
CAN-007
CAN-243
CAN-197
CAN-111
CAN-249
CAN-024
CAN-245
CAN-181
CAN-196
CAN-048
CAN-219
1A7
1A9
CAN-087
CAN-261
CAN-090
CAN-212
CAN-133
CAN-264
CAN-061
CAN-221
CAN-252
CAN-242
2F9
2G3
1I8
2H2
CAN-074
3F9
2I4
2C2
3G2
CAN-060
CAN-208
CAN-216
CAN-271
CAN-157
CAN-274
CAN-139
CAN-148
3C4
2A5
2I9
2H7
3F6
1I7
3B4
1H7
1B9
2B4
CAN-138
CAN-170
CAN-120
CAN-068
CAN-183
CAN-108
CAN-002
CAN-055
CAN-066
CAN-004
CAN-244
1A2
3I3
3C9
CAN-012
CAN-003
3I7
3B8
3H9
2A6
2I7
4B5
2E9
1C7
3B2
3A8
3D9
1G8
3F2
2G2
3F4
1H9
4C8
1E9
4A1
2B5
2H4
3A2
1F3
2A2
1C8
1F6
2I8
4D5
3D4
3D2
4B7
4D4
1C4
4B6
2B2
2I5
3A7
3A5
4A5
1G5
2C5
3C7
1C9
2H5
1C1
1H3
1I5
2B1
3E7
0.00
0.02
0.04
0.06
0.08
cohort
CANHelp
Manitoba
fineSTRUCTURE
population (fs)
hpAfrica1
hpAfrica2
hpEAsia
hpIndigenousNAmerica
hpIndigenousSAmerica
hpNEurope
hpSEurope
hpSiberia
hpSWEurope
community
Akiavik
Fort McPherson
Inuvik
Old Crow
Ross River
Teetn
plasmid
0 - absent
1 - Present

3.3.5 Bacteriophage content

The presence of various bacteriophages within the Arctic and Manitoban genomes are presented in **Figure 3.18.** Notably both phage types Helico_KHP40 and Bacilli_PBS1-G-AR9 carried an ABC transporter protein, which has the potential of being associated with antimicrobial resistance or substrate uptake from the environment. These two bacteriophages were present in a large number of Manitoban genomes as well as Arctic genomes belonging to the hspIndigenousSAmerica and hspNEurope sequence type clades. Arctic isolates belonging to the hspIndigenousNAmerica and hpafrica2 contained no prophage content. The presence of all other phages had neither carriage of any coding sequences of biological relevance or phylogenetic pattern within the Arctic and Manitoban isolates.

Figure 3.18. Presence of various bacteriophages within CAN*Help* and Southern Manitoban *H. pylori*. Tree tips labelled with cohort (Manitoban or CAN*Help*), community of origin, fineSTRUCTURE-determined sequence type, and presence (blue-shaded) or absence (white-shaded) of various bacteriophages. Phylogeny was created using Mashtree with whole genomes as input. fineSTRUCTURE analysis was run using 769 core genes present in >90% of Manitoban, Arctic, and world isolates.


cohort
community
fs
CAN-019
CAN-259
CAN-290
CAN-247
CAN-010
CAN-167
CAN-251
CAN-095
CAN-172
CAN-011
CAN-136
CAN-025
CAN-258
CAN-112
4A6
CAN-007
CAN-243
CAN-197
CAN-111
CAN-249
CAN-024
CAN-245
CAN-181
CAN-196
CAN-048
CAN-219
1A7
1A9
CAN-087
CAN-281
CAN-090
CAN-212
CAN-133
CAN-264
CAN-061
CAN-221
CAN-252
CAN-242
2F9
2G3
1I8
2H2
CAN-074
3F9
2I4
2C2
3G2
CAN-060
CAN-208
CAN-216
CAN-271
CAN-157
CAN-274
CAN-139
CAN-148
3C4
2A5
2I9
2H7
3F6
1I7
3B4
1H7
1B9
2B4
CAN-138
CAN-170
CAN-120
CAN-068
CAN-183
CAN-108
CAN-002
CAN-055
CAN-086
CAN-004
CAN-244
1A2
3I3
3C9
CAN-012
CAN-003
3I7
3B8
3H9
2A6
2I7
4B5
2E9
1C7
3B2
3A6
3D9
1G8
3F2
2G2
3F4
1H9
4C8
1E9
4A1
2B5
2H4
3A2
1F3
2A2
1C8
1F6
2I6
4D5
3D4
3D2
4B7
4D4
1C4
4B6
2B2
2I5
3A7
3A5
4A5
1G5
2C5
3C7
1C9
2H5
1C1
1H3
1I5
2B1
3E7
cohort
CANHelp
Manitoba
fineSTRUCTURE
population (fs)
hpAfrica1
hpAfrica2
hpEAsia
hpIndigenousAmerica
hpMEurope
hpSEurope
hpSiberia
hpSWEurope
community
Aklavik
Fort McPherson
Inuvik
Old Crow
Ross River
Yellin
Bacteriophage
0 Absent
1 Present

Chapter 4: Discussion, Limitations, Future Directions, and Conclusions

4.1 Discussion

This discussion section will be focused on the following results from this research project: genome quality and contamination; sequence types, genomic admixture, and phylogenetic structure; community and multihousehold genomic trends; the presence of antimicrobial resistance markers; various virulence factors including *cagA*, the *cag*PI, *vacA*, outer membrane proteins, and the type I toxin-antitoxin system *aapA1*/*isoA1*; and finally the presence of mobile genetic elements such as genomic islands, plasmids, and prophages.

4.1.1 Genome quality and contamination

An interesting trend depicted in **Figure 3.1** shows that larger genomes tended to have lower G-C% compared to smaller genomes having higher G-C%. It is possible that larger genomes may arise as a result from lacking certain genomic regions, which may in turn affect G-C%. However, the general plasticity of the *H. pylori* genome may also affect these trends. Most of these isolates with smaller genomes belong to a clade of CAN*Help* genomes that lack various virulence factors and horizontal mobile elements, and while it has been documented in the literature that reduction in genome size is generally associated with more virulent strains (Fischer et al. 2014), we observed the opposite.

While it is likely that some of contaminants within our *H. pylori* samples arose during the gastric endoscopic procedures, the presence of contamination may also be related to the fastidious nature of *H. pylori* growth. Culturing of *H. pylori* requires multiple instances of re-establishing a microaerophilic environment through gas flushing or replacing gas packs. These instances, as well as normal sampling and subculturing, combined with *H. pylori*'s requirement for culture on the time frame of weeks, could present multiple chances for contamination introduction (Stevenson et al. 2001). The majority of contaminated CAN*Help*

isolates contained *Mycoplasma hominis*. Owing to *M. hominis*'s small genome, a number of *H. pylori* genomes were salvageable from the read sets despite contamination. *M. hominis* is a bacterium typically associated with the male and female urogenital tracts, but has been isolated from the oral cavity of male patients with urethritis (Le et al. 2017). Two isolates (CAN-023 and CAN-121) contained no *H. pylori* reads whatsoever, but instead these isolates appeared to be pure clones of *Brevundimonas diminuta*. *Brevundimonas spp.* are typically considered opportunistic pathogens of low clinical significance, but have the potential of causing severe and invasive infections and appear to be emerging with multidrug resistance (Ryan and Pembroke 2018). *B. diminuta* is an opportunistic pathogenic species with low virulence, and has been isolated from blood, urine, and lung sputum of patients with cystic fibrosis. As a result, it is plausible for this bacterium to contaminate gastric biopsy samples if it is residing in the alimentary canal of a patient. One isolate, CAN-236 was contaminated with *Haemophilus influenzae* reads and was not salvageable. *H. influenzae* is a well-characterized pathogen that has been previously isolated from the mouth, stomach, and bowel, albeit rarely (Christensen, Kirkegaard, and Korner 1990). Finally, CAN-102 appeared to be a mixed sample of environmental soil bacteria with no *H. pylori* content whatsoever. This sample may have been compromised by any number of environmental sources along the processing pipeline of these samples. In contrast, the only contaminant in the Manitoban genomes was *Capnocytophaga sputigena*, a rare bacterium associated with dog bites and potentially periodontal disease in humans (Hazırolan 2021). It is plausible that this microbe may be colonizing the alimentary canal of the sampled individual, especially if they are a dog owner who is very close with their pet.

Together, these observations highlight some of the challenges of working with *H. pylori* in particular, and human isolates in general, particularly with samples collected from

community projects with funding limits. Despite the majority of isolates not containing contaminating reads, the results of the contamination-filtering pipeline, particularly the salvaging of 4 CAN*Help* genomes contaminated with *M. hominis* reads, displays the importance and benefits of including a contamination filtering step in any *H. pylori* genomics pipeline.

4.1.2 Sequence types and phylogenetic structure

The phylogenetic structure of *Helicobacter pylori* shows a strong geographic signal, believed partially due to the ancient association the bacterium has with humans. Although, it should be noted that this geographic signal constitutes much of that best for inferring that the association is ancient. Just based on genetic similarity, with as low resolution as MLST typing, one can infer where in the world an *H. pylori* isolate came from through comparative genomics (Thorell et al. 2017; Thorell et al. 2017). This demonstrates the power of sequence typing when it comes to *H. pylori* phylogenetics.

It should be noted, however, that while CAN*Help* community partners have a keen interest in how *H. pylori* is genetically similar or different across diverse human populations, questions about the ancestral origins of *H. pylori* and what they imply about human migration did not arise from CAN*Help* communities. To the contrary, the rationale for these questions may be viewed as being at odds with the de-colonizing principles of the CAN*Help* research program. Key among these principles is that communities should have control over investigations and narratives of their own history. Furthermore, given that inferences regarding ancestral origins of *H. pylori* are about events that occurred in the past and, therefore, cannot be confirmed empirically, it is particularly important to emphasize caution about such inferences drawn from observed patterns of *H. pylori* phylogenetics. As well,

particular care should be taken to avoid overlooking any departures from theory-driven expectations, to allow new perspectives and data to shape evolving knowledge.

When compared to the MLST sequence types, the fineSTRUCTURE-determined sequence types, which are based on whole-genome content, have higher resolution and are more informative. It is unsurprising that all but 8 of Manitoban genomes were assigned to the hspNEurope, hspSEurope, and hspSWEurope sequence types given the historical settlement of Southern Manitoba by European settlers before, during, and after the colonization of Canada. As well, Southern Manitoba has a large proportion of immigrants from European descent, which may drive *H. pylori* sequence type. Notably, 3 Manitoban and 11 CAN*Help* isolates clustered closely together with African *H. pylori* genomes and were assigned hpAfrica1 and hpAfrica2 sequence types. The majority of CAN*Help* isolates, however, were assigned to hspIndigenousNAmerica, and 14 were assigned to hspIndigenousSAmerica, which would be expected from isolated Indigenous *H. pylori* strains. However, sequence type alone does not show the full picture when it comes to the geographic signal of *H. pylori* genomes. Given the significant recombination rate of the bacterium, considerable admixture exists in all but the most remote and isolated *H. pylori* strains (Muñoz-Ramirez et al. 2020).

The overall trends of the admixture heatmap presented in **Figure 3.4** show Manitoba genomes sharing intra-cohort similarity between individual isolates along with inter-cohort similarities with some world genomes. In contrast, the CAN*Help* genomes predominantly show only intra-cohort similarities, and with weaker signals (viewed as lighter/more sparse patterns in **Figure 3.4**). When looking at the genomic admixture stacked bar plots presented in **Figure 3.5** further trends are evident. Firstly, amongst Manitoban isolates, the admixture

appears to be somewhat stable, with equal proportions of most sequence type donor genomes, albeit a predominance of hspNEurope and hspSEurope signal.

The majority of CAN*Help* genomes were dominated by hspIndigenousNAmerica and hspIndigenousSAmerica signals, with some isolates such as CAN-264 and -133 showing near-homogeneity (**Figure 3.5**). Some very closely clustered sub-clades, for example CAN-060, -208, -216, -271, -157, and -274 showed almost identical admixture proportions. Despite CAN-060 and -208 coming from Old Crow and -216, -271, -157, and -274 coming from Aklavik, these isolates appear to have clonal relation. This may be interpreted as evidence that *H. pylori* strains can migrate with community members who travel or move between CAN*Help* communities for various reasons; in particular, Old crow residents are almost exclusively Gwich'in First nations, which is the second largest ethnic group in Aklavik. It is not uncommon for residents of one of these communties to have relatives in the other, or move back and forth between the two communities.

Overall, it is clear that the Manitoban genomes are dominated by European signal, as shown by the proportion of blue in their admixture stacked bar pattern, whereas the CAN*Help* genomes are dominated by Indigenous American signal, as shown by the proportion of purple in their admixture stacked bars. It is important to consider that the results of these fineSTRUCTURE analyses are a core genome-based approach, and as a result do not reflect the considerable variability found within the accessory genome of these *H. pylori* isolates.

The admixture section of this current study was largely inspired by a paper published by Zilia Muñoz-Ramirez and colleagues in 2021. Along with Karen Goodman, the principal investigator of this grant project, who co-authored this paper, co-authors Zilia Muñoz-Ramirez and Kaisa Thorell mentored me in this analysis. In their study, they analyzed 723 *H.*

pylori genomes from 14 countries from the Americas, Iberian Peninsula, Europe, Africa, and Asia, to demonstrate how admixture varies across geographic regions and genetic backgrounds.

When compared to the admixture presented by Muñoz-Ramirez and colleagues (2021) the patterns are slightly different for *H. pylori* isolates in the CAN*Help* cohort I analyzed, but not the Manitoban cohort. Much like the non-Indigenous Canadian isolates investigated by Muñoz-Ramirez and colleagues, the Manitoban isolates showed a strong European signal with a dominance of hspSEurope, hspSWEurope, and hspNEurope signal. In fact, the proportions of the admixture signals were similar between the Manitoban cohort and these published results. What differs between the current results and those published by Muñoz-Ramirez et al. lies within the admixture of the CAN*Help* isolates. Muñoz-Ramirez and colleagues conducted the same analysis on a few dozen CAN*Help* isolates, some of which overlap with those presented in this current study. However, these isolates showed a much stronger homogenous hspIndigenousNAmerica signal than the isolates in the current study. It is likely because Muñoz-Ramirez and colleagues used a larger number of donor and recipient genomes than our study. While they looked at 723 *H. pylori* genomes, we only looked at just over 200 genomes. As a result, their analysis had more granular resolution than ours. However, in speaking with Muñoz-Ramirez and Thorell, they indicated that the number of donor isolates we used was sufficient for the analysis we wanted to conduct, relative to the number of genomes we analyzed.

4.1.3 Multihousehold group and community trends

For the most part, CAN*Help* isolates tended to cluster in subclades with other isolates from the same community. This can be seen with the quadruplet CAN-055, -066, -

004, and -244 coming from Aklavik; the triplets CAN-138, -170, and -120 coming from Aklavik, as well as CAN-068, -183, and -108 coming from Inuvik; the doublet CAN-133 and -264 coming from Fort McPherson, and so on. However, not all clades were homogenous with respect to community of origin. For example, two isolates, CAN-167 and -251 were most closely related to Aklavik isolates, and the subclade of CAN-087, -261, -090, and -212 contained isolates from Aklavik, Old Crow, and Fort McPherson, a community that is nearly exclusively Gwich'in First Nations. What these results suggest is that while the presence of unique and distinct *H. pylori* strains is stable within communities, there is migration of strains between communities, which is a result of human movement for family ties, living off the land, and other social reasons.

In terms of CAN*Help* genomes clustering tightly together in the phylogenetic tree, there are fewer closely related isolates from members of the same MHG than one might expect (**Figure 3.6**). For example, the tightly clustered subclade containing CAN-247 and CAN-259 belong to MHG 4, as does the separate subclade containing CAN-216 and CAN-274. And three members of a tight quadruplet, CAN-055, CAN-066, and CAN-244 belong to MHG 10, however, no other clusters showed a shared MHG signal. There are a number of reasons for this. The MHG data is sparse and is only available for a small subset of individuals, namely only from individuals that reside in Aklavik.

From a societal and biological perspective, it also makes sense why households would not cluster as exclusively as one might think. The very existence of these MHG implies a dynamic and flexible concept of a household within these communities. Individuals come and go, spend divided time in communities and out on the land, live with different family members during different periods of their lives, and change their lifestyles as needed based on seasonal changes and subsistence living. It is also very possible and extremely likely that

multiple strains would be present in a given household. Especially given that transmission of *H. pylori* most often occurs early in life (Stefano et al. 2018), it is extremely likely in households with multiple generations residing together that a number of distinct strains would be present within members of that household. It is also important to consider the fact that the *H. pylori* genome is highly recombinogenic and has a high mutation rate, which would drive sequence and strain diversity within households. Within-host evolution of *H. pylori* also helps drive the formation of different strains within a single individual, with bacteria varying genetically based on stomach anatomy and antibiotics exposure (Ailloud et al. 2019). Unfortunately, given the limited methodology employed in this study, namely only sequencing one culture isolate per individual, we were not able to explore the possibility of multi-strain carriage in individuals and any genotypic associations with households. However, this work will be carried out by the CAN*Help* Working Group under the current grant within the next few years.

4.1.4 Antimicrobial resistance patterns

Among resistance genomic markers, the presence of metronidazole resistance genomic markers was highest in both the Manitoban and CAN*Help* cohorts (**Figure 3.7**; **Table A3.4**). This is unsurprising as metronidazole resistance is the leading *H. pylori* AMR development across the world (Alba, Blanco, and Alarcon 2017). Additionally (and concerningly) metronidazole resistance markers were most commonly found in combination with other AMR markers in genomes possessing multidrug resistance genotypes. However, the presence of metronidazole resistant genotypes was lower in the CAN*Help* cohort compared to the Manitoban cohort. As well, clinical studies have provided substantial

evidence that phenotypic metronidazole resistance can usually be overcome with a treatment course of sufficient duration, even if the regimen includes metronidazole.

When overlaid with phylogenomic data (**Figure 3.8**), the presence of metronidazole and clarithromycin resistance genotypes shared phylogenetic connections in several CAN*Help* subclades. Several closely-grouped clusters of CAN*Help* isolates possessed metronidazole and/or clarithromycin resistance genotypes. In contrast, the distribution of levofloxacin resistance genotypes did not have a clear phylogenetic pattern in either Manitoban or CAN*Help* isolates.

Amoxicillin resistance markers were absent in all Manitoban isolates and only present in 2 of 56 (3.5%) CAN*Help* isolates, where it was carried alongside metronidazole resistance markers. This indicates that *H. pylori* in both Manitoba and the Arctic have yet to develop resistance to this beta-lactam antibiotic, which is not surprising because rates of amoxicillin resistance in *H. pylori* are low worldwide (Alba, Blanco, and Alarcon 2017). The two isolates containing amoxicillin resistance markers were closely related, pairing together as a duplet on the phylogenetic tree, both isolates from Fort McPherson.

In general the CAN*Help H. pylori* genomic cohort had much lower percentage of AMR genotypes detected *in silico* (**Table A3.4**) when compared to the Manitoban cohort, barring amoxicillin resistance. Fewer Arctic isolates contained metronidazole and clarithromycin resistance markers than the Manitoban isolates, and levofloxacin resistance detection was much lower in the CAN*Help* cohort. Notably, the phylogenetic clades containing the fewest AMR genotypes were the mostly distantly related Aklavik-dominant groups of sequence type hpAfrica2 and hspIndigenousNAmerica. This AMR genotypes observation overall appears to bode well for *H. pylori* clinical treatability in the North, at least for the foreseeable future, assuming these strains dominate within community members.

Very few scientific research groups are currently doing *in silico* antimicrobial resistance testing in *H. pylori*. Generally speaking, *H. pylori* research is lagging compared to other enteric pathogens. This is in part due to *H. pylori* being discovered relatively recently, the chronic nature of infection and disease progression, and the fact that it is not a foodborne pathogen. In general, there is a need for increased *Helicobacter pylori* genomics research worldwide, especially for a bacterium that is believed to cause cancer. For the purpose of this study, we thus needed to develop robust *in silico* antimicrobial resistance prediction approaches in order to fully understand the characteristics of our isolates, and to compare to phenotypic results as much as possible. Unfortunately, due to the COVID-19 pandemic, originally planned phenotypic testing of Manitoban *H. pylori* samples was not possible, so our ability to compare genotype to phenotype was limited to the CAN*Help* samples.

One commonly used *in silico* AMR-detection pipeline using RGI, which scans the well curated Comprehensive Antibiotic Resistance Database (CARD) (McArthur et al. 2013; Alcock et al. 2020), was explored but ultimately was only used for confirming local alignment results. This is because the CARD lacks sufficient curated entries for known antimicrobial resistance markers in *H. pylori*. There is a growing need for research groups to generate and make publicly available *H. pylori* data sets to detect emerging resistance mechanisms. Furthermore, these *in silico* results must be supported by experimental data from the lab.

In their 2021 paper, Camorlinga-Ponce et al. explored the incidence of antimicrobial resistance in Indigenous and non-Indigenous Mexican *H. pylori* isolates. They found that rates of phenotypic antimicrobial resistance, as determined by E test, were 58.6% for metronidazole, 18.5% for levofloxacin, 8.9% for clarithromycin, and only 1.8% for

amoxicillin. When stratified based on ethnicity, they found that rates of resistance in Indigenous Mexican isolates were 39.3% for metronidazole, 1.8% for clarithromycin and amoxicillin, and 0% for levofloxacin. The observed rate of multidrug resistance was 18.9% in mestizo Mexican isolates, but only 1.7% in Indigenous strains. When comparing these results to those of the current study, some interesting similarities are evident. Notably, rates of AMR genotypes (**Table A3.4**) were similar to the Mexican phenotypic results for amoxicillin, metronidazole and levofloxacin in the Manitoban cohort, and amoxicillin and metronidazole in the CAN*Help* cohort. The rate of clarithromycin resistance was drastically different between the Mexican Indigenous isolates (1.8%) and the CAN*Help* isolates (21.1%), with clarithromycin resistance genetic markers being much more prevalent in the Arctic. Where Camorlinga-Ponce et al.'s results diverge further from this current study are in the presence of multidrug resistance. In Mexico, multidrug resistance was present in 18.9% of mestizo isolates and 1.7% of Indigenous isolates. In contrast, multiple AMR markers were detected as present in 28.8% of Manitoban isolates and 12.0% of CAN*Help* isolates. It is still unknown what selective pressures influenced our result in this disparate evolution, but it is likely that access to antimicrobials and health care in general in remote and Indigenous regions is a driving factor. In both our *H. pylori* cohorts, the rate of genotypic multidrug resistance was higher compared to Camorlinga-Ponce et al.'s study. It is important to note, however, that phenotypic testing was not done for the Manitoban cohort and E testing for the CAN*Help* cohort was done as far back as a decade ago. As a result, there are caveats to consider when looking at these patterns, as experimental evidence is lacking.

As discussed in Section 3.3.1, and depicted in Figure 3.9, there was variable congruence between the presence of a genomic AMR marker and measured phenotypic resistance in CAN*Help* isolates. Most evident with metronidazole resistance, the rate of

phenotypic resistance was higher in genomes lacking known AMR markers when compared to those containing such markers. There are a number of reasons why this observation may occur. Firstly, antimicrobial resistance is a dynamic process in bacteria, with many factors at play. It is possible that resistance to metronidazole is a complex trait with interactions between variants, efflux pumps, biofilm production, and other bacterial factors and at distinct levels, *i.e.* protein, RNA, and regulation (Tshibang-Kabamba and Yamaoka 2021). Secondly, the field of *H. pylori* AMR is still largely unexplored, and there may be other genomic markers resulting in metronidazole resistance that have yet to be discovered. Finally, it is important to mention that the number of genomes under study is insufficient to make strong conclusions on genotypic-phenotypic associations.

Another important consideration is the unknown congruency of AMR phenotype in different environments. Whether a given *H. pylori* strain behaves the same in terms of resistance mechanisms and antimicrobial susceptibility is not certain between *in vitro* and *in vivo* conditions. For example, it is known that the harsh acidic nature of the human stomach affects both *H. pylori* AMR biology and the pharmacokinetics of antimicrobials (Scott, Sachs, and Marcus 2016; Erah et al. 1997). This is partially why treatment of *H. pylori* infection always includes a proton pump inhibitor and sometimes bismuth. Furthermore, as described in Chapter 1, the biology of the human stomach gradually changes over time as a result of *H. pylori* pathogenesis, along with *H. pylori* varying genetically based on stomach location, adding yet another variable of uncertainty (Berthenet et al. 2018).

4.1.5 *cagA* and *cag*PI

Both the Manitoban and CAN*Help* cohort had similar prevalence of the cancer-associated effector protein gene *cagA*. It is important to note that the risk of gastric cancer

as a result of CagA activity requires that the effector protein be introduced into host human cells using the pathogenicity island *cag*PI T4SS carried by *H. pylori*. The majority of Manitoban genomes carried the *cag*PI, whereas just over half of the CAN*Help* genomes carried the *cag*PI (**Figure 3.10**). This may indicate that the Arctic strains may be less virulent. The majority of CAN*Help* isolates were observed to either possess both *cagA* and the *cag*PI, *i.e. cagA*+/*cag*PI+ (61.0% Manitoban and 54.2% CAN*Help*), or lack both *cagA* and the *cag*PI, *i.e. cagA*-/*cag*PI- (31.2% Manitoban and 42.4% CAN*Help*), whereas in the Manitoban cohort, almost a tenth of isolates carried the pathogenicity island but not the effector protein genes (*i.e.* were *cag*PI+/*cagA*-). These results show the importance of not using the presence of the *cagA* gene as a substitute for presence of the *cag*PI itself, since 7.8% of Manitoban isolates were *cag*PI+/*cagA*-, and 1.7% of Arctic isolates were each *cag*PI+/*cagA*- and *cagA*+/*cag*PI-.

In both cohorts, the Western-associated *cagA* EPIYA-ABC motif combination was dominant (44.2% in Manitoban and 37.3% in CAN*Help*) as seen in **Figure 3.10**. Of concern are the presence of the EPIYA-ABCC motif in both cohorts, as *cagA* genes with multiple C or D motifs may trigger higher rates of phosphorylation, resulting in enhanced bacterial virulence imparted by cagA (Basso et al. 2008; Ferreira et al. 2012). Although the prevalence of these motifs is low (5.2% in Manitoban and 3.4% in CAN*Help*), the ABCC motif has been shown to be associated with peptic ulcer and gastric cancer outcomes in other populations, for example in southern Mexico (Beltrán-Anaya et al. 2014). The observation that the EPIYA-D motif was not found in any CAN*Help* genomes may indicate that despite being *cagA*+/*cag*PI+ isolates, they may confer lower cancer risk overall from the presence of these genes and their gene products than occurs in other populations. It was shown, for example

in Japan, that the presence of the East Asian-ABD EPIYA motif was significantly associated with gastric cancer compared to ulcers (Satomi et al. 2006).

There are clear phylogenetic patterns associated with *cagA* and *cag*PI status in both the Manitoban and CAN*Help* cohorts. Notably, the majority of Manitoban isolates fall within two large clades, one of which is dominated by *cagA*+/*cag*PI+ status, while the other is dominated by *cagA*-/*cag*PI- status. All of the other Manitoban cohort isolates that fall within smaller and more variable clades tend to possess both *cagA* and the *cag*PI (hence *cagA*+/*cag*PI+), except for 2I9, 2I4, 3F6, 2A5 (each of which has a truncated *cagA* gene), 1B9 (which possesses *cagA* but not the *cag*PI), 3G2 (which possesses the *cag*PI and a truncated *cagA* gene), 3I3 (which lacks the *cag*PI), and most notably 4A6, which clusters with the CAN*Help* isolates assigned to hspIndigenousNAmerica.

When looking at the phylogenetic patterns of *cagA* and the *cag*PI (**Figure 3.11)** within the CAN*Help* cohort, it is clear that all of the subclades containing isolates not belonging to the hspIndigenousNAmerica sequence type predominantly carry both *cagA* and the *cag*PI, whereas members of the hspIndigenousNAmerica sequence type clade all lack both of these genetic elements, except for two outlier isolates: CAN-024 (which has both; hence *cagA*+/*cag*PI+) and CAN-219 (which has the *cag*PI but not *cagA*). The previously sequenced aklavik86 and aklavik117 also lacked the *cag*PI (Kersulyte et al. 2015). While the risk of cancer that *H. pylori* confers depends on multiple factors acting in complex pathways, the absence of *cagA* and *cag*PI may indicate that people infected by such strains are at lower risk of gastric cancer than people infected with strains containing these genetic elements, which are the most studied and understood virulence trait in *H. pylori*. The observation that isolates belonging to the hspIndigenousNAmerica sequence type are typically *cagA-/cag*PI- provides a rationale for investigating whether CAN*Help* participants

with *cagA-/cag*PI- isolates have gastric pathology consistent with a lower risk of gastric cancer as well.

4.1.6 *vacA*

As discussed in section 1.2.3.2, there exist various alleles within the *vacA* gene that are associated with virulence and subsequent risk of developing cancer. It has been proposed that s1/m1 *vacA* genotypes confer the highest cancer risk, followed by s1/m2 and s2/m1, and that genotype s2/m2 poses almost no risk of cancer development (Sugimoto et al. 2009; Xue et al 2021). Almost every genome genotyped *in silico* in the current study possessed the s1 allele, with only 6 Manitoban isolates possessing the s2 allele. Both the m1 and m2 alleles were present in both the Manitoban and CANHelp cohorts, however the m2 allele was dominant in the Manitoban genomes and the m1 allele was dominant in the CAN*Help* genomes. Taken together, the observation that the most virulent s1/m1 *vacA* genotype was more prevalent in the CAN*Help* cohort genomes indicates that these Arctic *H. pylori* strains may be more virulent and likely to cause cancer than the m2-containing Manitoban strains. Since there are different modalities through which *H. pylori* can cause cancer, it is unsurprising that these isolates may have mosaicism within their genotypes across different loci such as *cagA*, the *cag*PI, *vacA*, etc. The possession of non-virulent *cag* genotypes and virulent *vacA* genotypes may indicate that these isolates may require different targeted treatment and disease management compared to other isolates from different regions.

Appendix **Table A3.6** and **Figure 3.13** depict an interesting paradox within CAN*Help* *vacA* genotyping. Contrasting the *in silico* results and the lab-based PCR results, the m1 and m2 genotype counts are seemingly reversed. Unfortunately, due to limited access to

CAN*Help* laboratory notes from several years ago, I was unable to confirm what primers were used for the PCR amplification and what generated amplicons were reported for these two different genotypes. It is certainly possible that the primers were mislabelled along with the subsequently collected data. Regardless, this incongruency displays the benefit of applying *in silico* local-alignment based genotyping of *vacA* alleles. This is of concern, because overestimating the proportion of m2 genotypes by PCR could give the false sense that Arctic *H. pylori* strains are less virulent than they actually might be.

As shown in Appendix **Table A3.6**, the CAN*Help* PCR-based *vacA* genotyping workflow only identified the s1a allele; yet there are three known s1 variants, s1a, s1b, and s1c that are easily identified using *in silico* methods. As a result, it is likely that the PCR scheme used by the CAN*Help* working group had limitations and greatly underestimated the number of genomes that had the s1 genotype. This is further supported by the number of s1b and s1c genotypes uniquely identified by the *in silico* local alignment approach. Moving forward it would be prudent to employ the whole genome *in silico* approaches for genotyping *vacA* alleles to ensure that isolate genomic information is not lost nor misinterpreted.

As shown in **Figure 3.12**, certain clades of both Manitoban and CAN*Help* isolates typically will show homogeneity in both s and m genotype, for example, the clade containing CAN-060, -208, -216, -271, -157, -274, -139, and -148 being entirely s1/m1. This pattern is also seen in the clade containing CAN-138, -170, -120, -068, -183, -108, -002, -055, -066, -004, and -244. In contrast, the large clade containing isolates assigned to the hspIndigenousNAmerica sequence type predominantly possess truncated *vacA* genes that lack content at the s and m sites. Because of these frameshift truncations associated with gene expression and virulence attenuation, it is likely that these isolates have less to non-

existent VacA activity and lower risk of developing cancer. Truncation of the *vacA* gene is common in non-vacuolating *H. pylori* strains but is less commonly seen in vacuolating ones (Ito et al. 1998). This directly contrasts these isolates to the aforementioned clades containing intact *vacA* possessing the more virulent s1/m1 genotype.

4.1.7 Outer membrane proteins

As **Figure 3.14** depicts, there are no clear phylogenetic patterns for the presence or absence of *hopH, hopQ, hopC, hopB, hopZ, homA, homC,* and *homD*. With a few exceptions, these outer membrane proteins were present in the genomes of all Manitoban and CAN*Help* isolates. It is interesting to note that *babA* and *sabA*, two of the better characterized *H. pylori* outer membrane proteins, did show phylogenetic patterns. Notably, *babA* was absent from all isolates in a subclade of hspIndigenousNAmerican isolates, including the Manitoban isolate 4A6. The OMP *sabA* was absent from a number of subclades of CAN*Help* isolates, including those belonging to hspIndigenousNAmerica, hspIndigenousSAmerica, and hpAfrica2. As well, both *babA* and *sabA* were absent from the quadruplet CAN-055, -066, -004, and -244, and both outer membrane proteins were absent from a small number of Manitoban isolates, although their absences did not show a strong phylogenetic pattern.

The absence of *babA* and *sabA* in a subset of the hspIndigenousNAmerica isolates further support the notion that this subset of predominantly Aklavik isolates are less virulent than the other strains of *H. pylori* that are circulating within CAN*Help* communities. It is known that Indigenous Latin American *H. pylori* isolates have a unique and distinct *babA* variant, although this loci in general is difficult to sequence using short read technology (Thorell et al. 2017). Work that has yet to be undertaken, however, includes allelic analysis

of these outer membrane proteins, especially given that a large proportion of the mutations that accumulate in the *H. pylori* genome occur within outer membrane protein genes (Karlsson et al. 2016).

4.1.8 Toxin-antitoxin *aapA1*/*isoA1*

The absence of the *aapA1*/*isoA1* toxin-antitoxin system has a striking phylogenetic signal, as depicted in **Figure 3.14**. The toxin-antitoxin loci is absent from nearly all isolates belonging to the hspIndigenousNAmerica sequence type (except for CAN-243 and including 4A6), along with the doublet CAN-061 and -221. Given the toxin-antitoxin's proposed role in response to bactericidal antimicrobials (Arnion et al. 2017; Krzyżek 2021), its missingness in these isolates is congruent with members of this clade not carrying antimicrobial resistance markers associated with resistance to amoxicillin, levofloxacin, and metronidazole. It is likely that these isolates have not been exposed to the environmental pressure from exposure to these antimicrobials which would lead to evolutionary advantage of both the toxin-antitoxin system as well as those resistance mechanisms.

Type I toxin-antitoxin systems, such as *aapA1*/*isoA1* can sometimes be disrupted by point mutations, genomic rearrangements, or disruption via an insertion sequence, and therefore must be further investigated beyond simple presence-absence (Arnion et al. 2017; Krzyżek 2021). There is much still to be explored with both this and other toxin-antitoxin systems. It is a growing area of research within *H. pylori* genomics, and as such there are not any studies of its presence in Indigenous cohorts for comparison.

4.1.9 Horizontal mobile elements

In keeping with other regions and studies, this *H. pylori* pathogenomic study (the largest to date that we know ever conducted in Canada) has similarly revealed presence of a variety of classical mobile genetic elements such as plasmids, mobile genomic islands or bacteriophages in the *H. pylori* population of Canada-sourced isolates. It is likely that the presence of these horizontal mobile elements drives the genetic diversity of *H. pylori*, and subsequently allows for the acquisition of adaptive traits that allow it to survive in such a harsh environmental niche like the human stomach. As the *H. pylori* genome has such an open and broad pan-genome, it is no surprise that some isolates possess mobile elements and specific virulence factors while others do not. Isolates are just as likely to gain genetic content as they are to lose it, as evident by the presence of ancestral prophages, for example (Vale et al. 2015).

While increasing reports of cryptic and transferable plasmids in *H. pylori* are occurring globally (Rohrer et al. 2012), there were no real clear phylogenetic patterns for the distribution of plasmids in both the Manitoban and CAN*Help* cohorts (**Figure 3.17**). There were only a few small sub-clades of CAN*Help* isolates that seemed to inherit two intact plasmids, such as: CAN-019, -259, -290, and -247 with pAB769 and pAG531; CAN-060, -208, -216, -271, -157, and -274 with pAG531 and pCP000242; and CAN-066, -004, and -244 with pAG532 and pCP003477. It is important to note here that the presence of plasmids *in vitro*, and subsequently *in silico* after genomic sequencing may not reflect the plasmids present *in vivo* owing to the requirement for intermediate isolate processing steps. Unfortunately, plasmids can be lost during bacterial culture of these fastidious organisms, and these isolates, from both the CAN*Help* and Manitoban cohorts, have been cultured, stocked, and subcultured since their initial isolations. It is also important to consider that

plasmids are notoriously difficult to sequence with short read technology, and the ones identified in this study may appear to miss genetic content that they might actually possess. This is especially relevant for the plasmids with homology to pXZ274, as they were predominantly truncated in their *tfs* T4SS content. Despite our plan to sequence a subset of isolates with longer read sequencing technology (namely Oxford Nanopore Technologies' MinION), the COVID-19 pandemic proved insurmountable with respect to this objective.

Along with the T4SS encoded by the *cag*PI, *H. pylori* genomes can contain the *comB* T4SS, through which naked DNA is taken up by the bacterium, along with the ICE element *tfs3* T4SS and the *tfs4* T4SS (Yuan, Wang, and Wang 2018). Owing to sequence similarity, it was interesting to see that there was limited overlap with isolates carrying both the *cag*PI and the *tfs* T4SS genomic islands. The only strong phylogenetic patterns sharing these two mobile elements observed was a single subclade of Manitoban isolates. While the *tfs* T4SS genomic islands were mainly absent from the large clade of isolates assigned to hspIndigenousNAmerica (**Figure 3.15**), a subclade containing isolates CAN-095, -172, -011, -136, -005, and -258 contained both it and another genomic island encoded for a beta-lactamase (*i.e.* were *cagA*-/*cag*PI-/*tfs* T4SS+/beta-lactamase+). The concurrence of these two genomic islands may indicate that they were inherited together in an ancestral isolate, and perhaps even have some shared virulence function, despite the fact that these isolates currently lack a CagA effector protein to inject using their encoded T4SS.

Correct assignment of gene annotation/function is of paramount importance in genomic characterization studies. There is, however, a limitation to gene annotation by similarity, especially when considering the limitations of the database being used. As much as possible I attempted to assign appropriate gene annotation (and potentially function), however the functional and experimental evidence indicating the reality of these loci and

their expression is missing. As such, one must be careful in not interpolating the presence of any genes too far without experimental evidence.

As plasmids pAG450-, pAB768-, and pAB513- carrying T4SS genes belong to *virB*, *virC*, and *virD* gene families, for accurate assignment it was important to ensure that care be taken to avoid inappropriate 'over' inclusion of these mobile genetic features during genomic island and plasmid identification. However, when comparing the presence of these mobile genetic elements within individual study isolates, there was minimal feature overlap, which indicates that these elements are distinct, despite encoding similar functions. Some redundancy of *cag* T4SS function is to be expected, especially given the crucial virulence factor transport role T4SS plays in the *cag*PI because of how critical this virulence factor is in *H. pylori* pathogenesis. As well, as mentioned in Section 3.3.4, a number of these plasmids were truncated and did not encode for an entire *comB* T4SS system, which further adds to their perceived redundancy. When investigated further, these mobile genomic islands were identified as known sources of *H. pylori* genomic variability, namely the ICEs *tfs3* and *tfs4*, also known as plasticity zones (Delahay, Croxall and Stephens 2018). These ICEs encode for *xerT* recombinase and *virD2* relaxase, along with genes homologous to the *comB* T4SS. Within a given genome, the presence of the *tfs* T4SS-containing mobile element was identified based on whether they carried *tfs3* and/or *tfs4* alleles, as depicted in **Figure 3.15.** The *tfs3* and *tfs4* alignments depicted in **Figure 3.16** show that coverage of *tfs3* coding sequences is sparse in a subset of CAN*Help* and Manitoba genomes when compared to coverage of *tfs4* coding sequences. It is important to note that the presence of either ICE element presented in **Figure 3.15** only relied on the presence of a single *tfs3* or *tfs4* gene. Only six genomes were presented in **Figure 3.16** due to time constraints and to ensure the figure was legible. However, out of all six genomes investigated, only CAN-012 had near-

complete coverage across all *tfs4* coding sequences. This indicates that even if genes from a given ICE are present within a genomic island, it does not mean that the entirety of *tfs3* or *tfs4* genetic content may be actually present and expressed within a genome of interest. ICE elements are known to transfer and degrade, and *tfs3* and *tfs4* have been demonstrated to undergo erosion in certain *H. pylori* populations. Erosion of the T4SS genes by frameshifts or deletions is visible in numerous strains from different populations (Fischer et al., 2014; Delahay et al., 2018), suggesting that evolutionary advantages may be conferred upon the harboring strains by the additional genes rather than the T4SS genes themselves. Of note is the large degree of genetic dissimilarity within the V4 - methylase region of *tfs4* in within the investigated isolates except 2H7. As noted by the presence of vertical lines indicated variants, each isolate inspected had variants along the length of this gene. While variants were also present in other regions of *tfs4*, none were the same level as the V4 - methylase region.

Few isolates from both CAN*Help* and Manitoba cohorts assigned to hpAfrica1, hpAfrica2, and hspIndigenousNAmerica had notable bacteriophage content (**Figure 3.18**). Bacteriophage presence is important because it is one of the drivers of genetic diversity within *H. pylori* and at least one prophage is present in about 20% of all isolates (Muñoz et al. 2020). Some researchers have even shown that the presence of prophages in European *H. pylori* drive distinct subpopulations in hpNEurope and hpSWEurope isolates (Vale et al. 2015). Of interest is the presence of a Bacilli-like bacteriophage containing an ABC transporter gene present in a number of isolates from both cohorts spread across various clades and sequence types. The presence of this bacteriophage has some limited amount of phylogenetic patterns, with it being present in all isolates of the close clade containing CAN-060, -208, -216, -271, -157, and -274. None of the other bacteriophages identified had

phylogenetic patterns in the tree, however a few of them also contained ABC transporter genes, which is of note.

There is no clear relationship between genomic islands or integrated bacteriophages carrying an ABC transporter and the presence of any of the AMR markers discussed prior. It is important to note that the presence of a mobile element-encoded ABC transporter does not necessarily have biological relevance to virulence or AMR. Unfortunately, like with many aspects of the *H. pylori* genome, the genes in question that have been identified have yet to be characterized *in silico* or studied experimentally *in vivo*.

All of the isolates within the hspIndigenousNAmerica subclade, including the CAN*Help* isolates as well as Manitoba isolate 4A6, had smaller genome length (average: 1.52 Mb) and fewer coding sequences (average: 1423) relative to other isolates. This is not surprising given that these isolates had no bacteriophage content, minimal plasmid content, and fewer genomic islands than other clades, including the lack of the *cag*PI. While it is possible that this missing genomic content may be a result of sequencing technology limitations, the fact that all of these isolates cluster together makes this less likely. Therefore, it is probable that these isolates do in fact lack *cagA* and the *cag*PI, have truncated *vacA* genes, and have fewer horizontal mobile elements than genomes belonging to other clades. It is probable that such observed genomic differences result from being a more geographically-isolated and distinct *H. pylori* population.

4.1.10 Review of discussion

As stated in Section 1.1.7, this research project is hypothesis generating rather than hypothesis testing. That being said, I observed exactly what I expected. The strains of *H. pylori* that were sequenced from Manitoba and CAN*Help* Arctic communities were

genomically and phylogenetically distinct. Ultimately, there were clear patterns in virulence and pathogenic factors within the CAN*Help* isolates. In particular, the clade containing CAN*Help* isolates belonging to the hspIndigenousNAmerica sequence type (alongside the Manitoban isolate 4A6), generally: had smaller genomes with higher G-C%; lacked the resistance markers to amoxicillin, metronidazole, levofloxacin, and clarithromycin; possessed truncated *vacA* genes or *vacA* genes containing lower-risk s/m genotypes; lacked the *aapA1*/*isoA1* toxin-antitoxin; were less likely to carry the outer membrane proteins *babA* and *sabA*; lacked the *cag*PI and *cagA*; carried fewer genomic islands (albeit a subclade did carry the *tfs3* and *tfs4* ICEs); and contained relatively no prophage content. All of these results suggest that this strain of *H. pylori* circulating in CAN*Help* communities may pose lower risk for disease progression in carriers.

Likewise, subclades of CAN*Help* isolates belonging to the hpAfrica1, hpAfrica2, hspIndigenousSAmerica, and hspNEurope sequence types showed clade-wide similarities including: being more likely to carry one or more AMR markers; possessing higher-risk *vacA* s/m genotypes; carrying *cagA* and the *cag*PI; and containing the *aapA1*/*isoA1* toxin-antitoxin. These results suggest that these other, non-hspIndigenousNAmerica strains circulating within CAN*Help* communities may be of higher risk when it comes to disease outcome and resisting treatment.

Manitoban isolates predominantly belonged to the hspNEurope, hspSEurope, and hspSWEurope sequence types, with some isolates being classified as hpAfrica1, and a single isolate being classified as each of hspEAsia, hspSiberia, and hspIndigenousNAmerica. Of note is the isolate 4A6, which shares many genomic characteristics with Aklavik isolates belonging to the same sequence type. Generally, Manitoban isolates belonging to the three European sequence types: were more likely to carry one or more AMR markers; generally

carried the *aapA1*/*isoA1* toxin-antitoxin; possessed the outer membrane proteins *babA* and *sabA*; about half carried *cagA*, the *cag*PI, and the ICEs *tfs3* and *tfs4*; and were more likely to carry prophages.

All of this is to say that this current study is small, and I don't feel confident to make any broad claims without further sequencing and analysis being done moving forward. Unfortunately the project was not as expansive as we had hoped, given the COVID-19 pandemic, but I am nonetheless proud of the work I completed. Despite being a small study, we characterized several important genomic features in 2 geographically distinct Canadian populations. Findings from this study will provide important contributions to the field of *H. pylori* genomics, which remains in its infancy. It will also provide key guidance for the CAN*Help* Working Group's emerging *H. pylori* genomics research program.

4.2 Limitations

4.2.1 *H. pylori* genomic plasticity

The high rate of recombination within *H. pylori* makes it a particularly difficult bacterium for comparative genomic studies without employing short- and long-read sequencing technology. The use of any single given reference genome within bioinformatics approaches can give rise to missing genomic content, structural variations or propagating errors in annotation (Tada et al. 2017).

Several typical reference genome-based approaches were explored but excluded for project analyses owing to this issue of reference genome bias. For example, the NML's flagship core genome SNV-analysis and phylogenetic pipeline, SNVPhyl (Petkau et al. 2017), resulted in a large proportion of the *H. pylori* genome being 'filtered out' owing to the organism's high rates of recombination and mutation (*i.e.* small core genome). As much as

possible, it was essential to employ 'reference-free' bioinformatics tools such as Mashtree (Katz et al. 2019) and the *de novo* assembler SPades (Bankevich et al. 2012). When reference sequences were requisite for certain analysis, such as when characterizing antimicrobial resistance or virulence factor genotypes, a local alignment approach was used to limit the impact of recombination. As well, the recombinant nature of *H. pylori* was addressed in the analytical methodology itself, by devoting analyses to characterize and describe genomic admixture within our genomes.

4.2.2 Sample size, specificity and sensitivity

Unfortunately the study's sample size is not large enough to estimate precisely any measures of causal effect of *H. pylori* genomes on disease severity or progression. As a Master's Thesis, it was decided early that the scope of this research project would focus on genomic patterns and variations at the community and household level. This early scope definition was prudent in hindsight, especially given the reduced number of sequenced and analyzed genomes in light of the unpredictable COVID-19 pandemic, as any estimated measures of association between genotype and disease outcome would be imprecise and would require re-analysis with a larger dataset. Analyses for genome-wide associations require many more genomes than we could garner to provide sufficient statistical power to produce statistically precise association results.

4.2.3 Sample selection

As discussed in Sections 1.1.1 and 1.2.2, *H. pylori* demonstrates genetic variability based on host stomach anatomy. Evidence suggests that *H. pylori* isolated from the antrum, corpus, and fundus of the human stomach possess genetic variability as a result of the

unique microenvironmental niches they find themselves in within the anatomical regions of the human stomach. Unfortunately, this study is limited by how the *H. pylori* samples from the CAN*Help* communities were isolated previously. Some of the culture methods were not optimal for a bacterial genomics project, in particular, the pooling of gastric biopsies from the various regions of participant stomachs were pooled together for bacterial isolation. As a result we have no ability to discern the differences that might exist between isolates taken from different stomach regions. Despite this limitation, there is the possibility that this sampling method would make future projects investigating the presence of multiple strains within individuals more fruitful just based on the potential for the isolation of more genetically distinct isolates from a given individual.

4.2.4 Community-driven research

There are unique challenges to consider when working with communities and human host health data, especially to appropriately balance privacy protections and the interests of individual participants while respecting the values and goals of community partners. On one hand, these types of data are crucial for gathering evidence to identify public health issues and developing strategies to improve public health. Conversely, there is the significant issue of protecting the privacy of an individual when using information that has even limited potential to identify participants (Colquhoun et al. 2012) such as bacterial genomic data. Ultimately this project, as part of a research program that engages community partners at all stages of research, followed the guidance of community project planning committees in balancing the interests of the individual and community at large.

Another obstacle to acquired knowledge exchange we encountered during this project is in the disparate ways different individuals conceptualize scientific research and

interpret scientific results (Colquhoun et al. 2012). And furthermore sometimes researchers are unable to release all data back to the participating communities due to data disclosure and release rules and restrictions. This is made difficult by cultural differences in understanding everything from *how science is talked about*, *thought about*, *understanding the timeline and turnaround of research*, and even simply beliefs regarding *what is research* (Colquhoun et al. 2012). As well, misunderstanding the purpose and goals of a study can occur if not directly addressed. Ultimately these issues can be mitigated via context- and culturally-appropriate, bidirectional communication and knowledge exchange. These issues were identified by the CAN*Help* Working Group in the past, and while I only had limited interaction with community members during my Master's research, in one particular conversation I did notice differences between how the community member and I viewed the work that I was doing. Surprisingly she was much less concerned about the potential harms of Indigenous bacterial genomics work than I was, and only focused on the benefits of understanding as much as we can as collaborators.

One marked difficulty with community outreach attempts while carrying out this Master's project pertains to community priority, interest and engagement. Given the timing of unforeseen COVID-19 pandemic events, many community members were (justifiably) preoccupied with other concerns. Active, in-person engagement with communities had to be postponed. At the time of this thesis writing, a Bacterial Genomics workshop originally planned in Inuvik for April 2020 has been postponed indefinitely.

4.2.5 COVID-19 pandemic

Unfortunately, the COVID-19 pandemic affected the scale of the project, along with some of the goals and deliverables and certainly the scale of this Master's project.

Institutional shutdowns and lockdowns aimed at limiting COVID-19 spread prevented sequencing an expanded number of isolates. After completing sequencing of the initial batch of Manitoba isolates, the NML was closed for all graduate students access for roughly six months. As well, when I was allowed access to pursue my non-COVID wet lab work, the COVID-19 diagnostic testing burden at the Cadham Provincial Laboratory, where all Manitoban isolates were sourced, precluded any *H. pylori* stock subculturing. This was also partially due to material transfer agreements related to non-COVID-19 projects being deprioritized during the height of the pandemic. Furthermore, lab work at the University of Alberta was also slowed, and later halted, owing to provincial public health restrictions, and many cultures were ultimately lost or delayed in their sequencing. Consequently, decreased wet lab access and sample access during the pandemic significantly altered the scope of the project, particularly the number of *H. pylori* genomes and the extent of their characterization.

4.3 Future directions

The most important future direction is to complete the sequencing of all Manitoban and CAN*Help H. pylori* isolates. There remain more than 100 Manitoban and more than 200 CAN*Help* isolates stored in freezers at the Cadham Provincial Laboratory and the University of Alberta (respectively) that require sequencing as described in the goals of the funded grant. Sequencing these genomes is top priority in order to address further goals of this grant. As a result, after sequencing these isolates, analyses similar to what has been described herein must also be done to characterize and place them in the global context.

Importantly, phenotypic antimicrobial resistance *in vitro* should be performed on all Manitoban isolates that underwent sequencing. This work will be important for replicating

what was done in this genotype-phenotype correlations analysis for the CAN*Help* isolates, and increasing our sample size to determine with more certainty the congruency between the presence of AMR genotypes and actual phenotypic resistance observed *in vitro*.

One important group of virulence factors that were not explored in detail in this project are the OMPs that *H. pylori* uses for adherence to the stomach epithelium. As discussed in Section 1.2.3.1, various OMPs are associated with increased disease severity and gastric cancer development, including BabA, SabA, and members of the Hop, Hof, and Hom families. While these genes were noted for their presence in various genomic islands within our study genomes, none of these genes were characterized and analyzed in detail. This is an important next step for the analysis of the genomes already collected, as well as any new Canadian *H. pylori* genomes.

An important next step would be to further characterize the ICEs *tfs3* or *tfs4*. While the relationships between these ICEs and *H. pylori* virulence are less understood than the *cag*PI, it is likely that the presence of these mobile elements in some isolates (but not all) may influence disease status and differential outcome in *H. pylori* carriers.

Most importantly, the sequencing of all remaining Manitoban and CAN*Help* isolates will allow for work to begin regarding the biggest goal of the funded grant, estimating associations between *H. pylori* genomic elements and disease outcomes. Looking more broadly, it is still clear that the field of *H. pylori* genomics is in its infancy. I believe that this project will fall in line with expanding genomics work focused on this interesting and difficult bacterium. As antimicrobial resistance becomes more prevalent within *H. pylori* isolates around the world, the speed and ease of whole genome sequencing will be an important tool in the treatment of this chronic infection, especially in regions where more virulent strains are present and where gastric cancer is of large concern.

4.4 Conclusions

This foundational Canadian *H. pylori* genomics project has added to the body of knowledge for this organism, which is understudied in Canada, and significantly enhanced our understanding of *H. pylori* population structure and genome evolution for two geographic regions. This project focussed on defining genomic patterns and variations at the community and household level. I have characterized the phylogenetic structure of *H. pylori* isolates from CAN*Help* Arctic Indigenous communities and clinical samples from Southern Manitoba in comparison to representative genomes from around the world. By doing so, I have identified a clade of CAN*Help* strains belonging to the hspIndigenousNAmerica sequence type that appear to have reduced virulence and lower resistance to treatment. Although the kinship data was limited to a few multihousehold groups in Aklavik, I identified a handful of clonally related CAN*Help* pairs that belonged to the same multihousehold group. There were clear genomic differences between Southern Manitoban and CAN*Help* isolates. Generally Manitoban isolates possessed more virulence factors and AMR markers. Ultimately there are several patterns that are notable with regard to the CAN*Help* isolates including the distribution of AMR markers, high-risk *vacA* genotypes, the presence of *cagA* and the *cag*PI, the mobile element ICEs *tfs3* and *tfs4*, as well as ancestral genomic admixture and sequence type designation. Findings from this study will help broaden our understanding of *H. pylori* genomics from this under-sequenced and under-characterized cluster of communities and will serve as a framework for further work done funded by this research grant.

References

Achtman, M., Azuma, T., Berg, D. E., Ito, Y., Morelli, G., Pan, Z.-J., Suerbaum, S., Thompson, S. A., van der Ende, A., and van Doorn, L.-J. (1999). Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions. *Molecular Microbiology*, *32*(3), 459–470, DOI:10.1046/j.1365-2958.1999.01382.x

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Grüning, B., Guerler, A., Hillman-Jackson, J., Von Kuster, G., Rasche, E., Soranzo, N., Turaga, N., Taylor, J., Nekrutenko, A., and Goecks, J. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, *46*(W1), W537–W544, DOI:10.1093/nar/gky379

Ahmed, N., Loke, M. F., Kumar, N., and Vadivelu, J. (2013). *Helicobacter pylori* in 2013: Multiplying Genomes, Emerging Insights. *Helicobacter*, *18*, 1–4, DOI:10.1111/hel.12069

Ahn, H. J., and Lee, D. S. (2015). *Helicobacter pylori* in gastric carcinogenesis. *World J. Gastrointest. Oncol.* 7, 455–465. DOI: 10.4251/wjgo.v7.i12.455

Ailloud, F., Didelot, X., Woltemate, S., Pfaffinger, G., Overmann, J., Bader, R. C., Schulz, C., Malfertheiner, P., and Suerbaum, S. (2019). Within-host evolution of *Helicobacter pylori* shaped by niche-specific adaptation, intragastric migrations and selective sweeps. *Nature Communications*, *10*(1), 2273, DOI:10.1038/s41467-019-10050-1

Akeel, M., Shehata, A., Elhafey, A., Elmakki, E., Aboshouk, T., Ageely, H., and Mahfouz, M. (2019) *Helicobacter pylori vacA, cagA and iceA* genotypes in dyspeptic patients from southwestern region, Saudi Arabia: distribution and association with clinical outcomes and histopathological changes. *BMC Gastroenterol 19*(16). DOI:10.1186/s12876-019-0934-z

Alba, C., Blanco, A., & Alarcón, T. (2017). Antibiotic resistance in *Helicobacter pylori. Current opinion in infectious diseases, 30*(5), 489–497. DOI:10.1097/QCO.0000000000000396

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

Alfizah, H., Norazah, A., Hamizah, R., and Ramelah, M. (2014). Resistotype of *Helicobacter pylori* isolates: the impact on eradication outcome. *Journal of medical microbiology,* 63(Pt 5), 703–709. DOI:10.1099/jmm.0.069781-0

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Arnion, H., Korkut, D. N., Masachis Gelo, S., Chabas, S., Reignier, J., Iost, I., and Darfeuille, F. (2017). Mechanistic insights into type I toxin antitoxin systems in *Helicobacter pylori*: the importance of mRNA folding in controlling toxin expression. *Nucleic acids research*, *45*(8), 4782–4795. DOI:10.1093/nar/gkw1343

Arnold, I. C., Dehzad, N., Reuter, S., Martin, H., Becher, B., Taube, C., and Müller, A. (2011) *Helicobacter pylori* infection prevents allergic asthma in mouse models through the induction of regulatory T cells. *J Clin Invest.* 121(8):3088-93. DOI:10.1172/JCI45041.

Aspholm-Hurtig, M., Dailide, G., Lahmann, M., Kalia, A., Ilver, D., Roche, N., Vikström, S., Sjöström, R., Lindén, S., Bäckström, A., Lundberg, C., Arnqvist, A., Mahdavi, J., Nilsson, U. J., Velapatiño, B., Gilman, R. H., Gerhard, M., Alarcon, T., López-Brea, M., Nakazawa, T., Fox, J. G., Correa, P., Dominguez-Bello, M. G., Perez-Perez, G. I., Blaser, M. J., Normark, S., Carlstedt, I., Oscarson, S., Teneberg, S., Berg, D. E., and Borén, T. (2004). Functional Adaptation of BabA, the H. pylori ABO Blood Group Antigen Binding Adhesin. *Science, 305*(5683), 519–522, DOI:10.1126/science.1098801

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart, D.S. (2016) PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res. 8*;44(W1):W16-21. DOI:10.1093/nar/gkw387.

Azevedo, N. F., Huntington, J., and Goodman, K. J. (2009). The Epidemiology of *Helicobacter pylori* and Public Health Implications. *Helicobacter*, *14*, 1–7. DOI:10.1111/j.1523-5378.2009.00703.x

Backert, S., and Blaser, M. J. (2016). The Role of CagA in the Gastric Biology of *Helicobacter pylori*. *Cancer Research*, *76*(14), 4028–4031, DOI:10.1158/0008-5472.CAN-16-1680

Baltrus, D. A., Amieva, M. R., Covacci, A., Lowe, T. M., Merrell, D. S., Ottemann, K. M., Stein, M., Salama, N. R., and Guillemin, K. (2009). The Complete Genome Sequence of Helicobacter pylori Strain G27. *Journal of Bacteriology*, *191*(1), 447–448, DOI:10.1128/JB.01416-08

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and 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, DOI:10.1089/cmb.2012.0021

Basso, D., Zambon, C-F., Letley, D. P., Stranges, A., Marchet, A., Rhead, J. L., Schiavon, S., Guariso, G., Ceroti, M., Nitti, D., Rugge, M., Plebani, M. and Atherton, J. C. (2008) Clinical Relevance of *Helicobacter pylori cagA* and *vacA* Gene Polymorphisms. *Gastroenterology* 135(1), 91-99. DOI:10.1053/j.gastro.2008.03.041

Beltrán-Anaya, F.O., Poblete, T.M., Román-Román, A., Reyes, S., de Sampedro, J., Peralta-Zaragoza, O., Rodríguez, M. A., del Moral-Hernández, O., Illades-Aguiar, B., and Fernández-Tilapa, G. (2014) The EPIYA-ABCC motif pattern in *CagA* of *Helicobacter pylori* is associated with peptic ulcer and gastric cancer in Mexican population. *BMC Gastroenterol 14*, 223 (2014). DOI:10.1186/s12876-014-0223-9

Bertelli, C., Laird, M. R., Williams, K. P., Simon Fraser University Research Computing Group, Lau, B. Y., Hoad, G., Winsor, G. L., and Brinkman, F. S. L. (2017) IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Research 45* (W1), W30–W35. DOI:10.1093/nar/gkx343

Berthenet, E., Sheppard, S., and Vale, F. F. (2016). Recent "omics" advances in *Helicobacter pylori*. *Helicobacter*, *21*, 14–18, DOI:10.1111/hel.12334

Berthenet, E., Yahara, K., Thorell, K., Pascoe, B., Meric, G., Mikhail, J. M., Engstrand, L., Enroth, H., Burette, A., Megraud, F., Varon, C., Atherton, J. C., Smith, S., Wilkinson, T. S., Hitchings, M. D., Falush, D., and Sheppard, S. K. (2018). A GWAS on *Helicobacter pylori* strains points to genetic variants associated with gastric cancer risk. *BMC Biology*, *16*(1), 84. DOI:10.1186/s12915-018-0550-3

Blaser, M. J., Perez-Perez, G. I., Kleanthous, H., Cover, T. L., Peek, R. M., Chyou, P. H., Stemmermann, G. N., and Nomura, A. (1995). Infection with *Helicobacter pylori* Strains Possessing cagA Is Associated with an Increased Risk of Developing Adenocarcinoma of the Stomach. *Cancer Research* 55(10):2111-5

Bruce, M. G., and Maaroos, H. I. (2008). Epidemiology of *Helicobacter pylori* Infection. *Helicobacter*, *13*, 1–6. DOI:10.1111/j.1523-5378.2008.00631.x

Burucoa, C., and Axon, A. (2017). Epidemiology of *Helicobacter pylori* infection. *Helicobacter*, *22*, e12403. DOI:10.1111/hel.12403

CAN*Help* Statement on Knowledge Generated Collaboratively (2018) Retrieved from http://canhelpworkinggroup.ca/wp-content/uploads/CANHelp-Statement-on-Knowledge-Generated-Collaboratively_2018-11-06.pdf

Camorlinga-Ponce, M., Gomez-Delgado, A., Aguilar-Zamora, E., Torres, R. C., Giono-Cerezo, S., Escobar-Ogaz, A., and Torres, J. (2021) Phenotypic and Genotypic Antibiotic Resistance Patterns in *Helicobacter pylori* Strains From Ethnically Diverse Population in Mexico. *Frontiers in Cellular and Infection Microbiology* 10:539115 DOI:10.3389/fcimb.2020.539115

Cao, D.-M., Lu, Q.-F., Li, S.-B., Wang, J.-P., Chen, Y.-L., Huang, Y.-Q., and Bi, H.-K. (2016). Comparative Genomics of *H. pylori* and Non-Pylori *Helicobacter* Species to Identify New Regions Associated with Its Pathogenicity and Adaptability. *BioMed Research International*, *2016*, 1–15. DOI:10.1155/2016/6106029

Cao, Q., Didelot, X., Wu, Z., Li, Z., He, L., Li, Y., Ni, M., You, Y., Lin, X., Li, Z., Gong, Y., Zheng, M., Zhang, M., Liu, J., Wang, W., Bo, X., Falush, D., Wang, S., and Zhang, J. (2015).

Progressive genomic convergence of two *Helicobacter pylori* strains during mixed infection of a patient with chronic gastritis. *Gut*, *64*(4), 554–561. DOI:10.1136/gutjnl-2014-307345

Carraher, S., Chang, H.-J., Munday, R., Goodman, K. J., and the CANHelp Working Group. (2013). *Helicobacter pylori* incidence and re-infection in the Aklavik *H. pylori* Project. *International Journal of Circumpolar Health*, *72*(1), 21594. DOI:10.3402/ijch.v72i0.21594

Carriço, J. A., Rossi, M., Moran-Gilad, J., Van Domselaar, G., and Ramirez, M. (2018). A primer on microbial bioinformatics for nonbioinformaticians. *Clinical Microbiology and Infection*, *24*(4), 342–349. DOI:10.1016/j.cmi.2017.12.015

Carrière, G. M., Tjepkema, M., Pennock, J., and Goedhuis, N. (2012). Cancer patterns in Inuit Nunangat: 1998–2007. *International Journal of Circumpolar Health*, *71*(1), 18581. DOI:10.3402/ijch.v71i0.18581

Chakrani, Z., Robinson, K., and Taye, B. (2018). Association Between ABO Blood Groups and Helicobacter pylori Infection: A Meta-Analysis. *Scientific reports, 8*(1), 17604. DOI:10.1038/s41598-018-36006-x

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics (Oxford, England), 34*(17), i884–i890. DOI:10.1093/bioinformatics/bty560

Cheung, J., Goodman, K. J., Girgis, S., Bailey, R., Morse, J., Fedorak, R. N., Geary, J., Fagan-Garcia, K., van Zanten, S. V., and the CANHelp Working Group. (2014). Disease manifestations of *Helicobacter pylori* infection in Arctic Canada: Using epidemiology to address community concerns. *BMJ Open*, *4*(1), e003689. DOI:10.1136/bmjopen-2013-003689

Cheung, J., Morse, A., Goodman, K. J., van Zanten, S. V., Bailey, R., Morse, J., Fedorak, R. N., and Keelan, M. M. (2009) Prevalence of *Helicobacter pylori* and Antibiotic Resistance in an Aboriginal Population in Canada's Arctic: Preliminary Results from the Aklavik *H. pylori* Project. *Gastroenterology, 136(Suppl 1)*:A551. DOI:10.1016/S0016-5085(09)61561-4

Cheung, J., Goodman, K., Munday, R., Heavner, K., Huntington, J., Morse, J., Veldhuyzen van Zanten, S., Fedorak, R. N., Corriveau, A., Bailey, R. J., and the CANHelp working group. (2008). *Helicobacter Pylori* Infection in Canada's Arctic: Searching for the Solutions. *Canadian Journal of Gastroenterology*, *22*(11), 912–916. DOI:10.1155/2008/614538

Christensen, J. J., Kirkegaard, E., and Korner, B. (1990). *Haemophilus* Isolated from Unusual Anatomical Sites. *Scandinavian Journal of Infectious Diseases, 22*(4), 437–444. DOI:10.3109/00365549009027075

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}, *iso*-2; *iso*-3. *Fly* 6(2), 80-92 DOI:10.4161/fly.19695

Colquhoun, A, Arnold, M., Ferlay, J., Goodman, K. J., Forman, D., and Soerjomataram, I. (2015). Global patterns of cardia and non-cardia gastric cancer incidence in 2012. *Gut*, *64*(12), 1881–1888. DOI:10.1136/gutjnl-2014-308915

Colquhoun, A., Aplin, L., Geary, J., Goodman, K. J., and Hatcher, J. (2012). Challenges created by data dissemination and access restrictions when attempting to address community concerns: Individual privacy versus public wellbeing. *International Journal of Circumpolar Health*, *71*(1), 18414. DOI:10.3402/ijch.v71i0.18414

Colquhoun, A., Geary, J., and Goodman, K. J. (2013). Challenges in conducting community-driven research created by differing ways of talking and thinking about science: A researcher's perspective. *International Journal of Circumpolar Health*, *72*(1), 21232. DOI:10.3402/ijch.v72i0.21232

Colquhoun, A., Hannah, H., Corriveau, A., Hanley, B., Yuan, Y., Goodman, K. J., and the CANHelp Working Group. (2019). Gastric Cancer in Northern Canadian Populations: A Focus on Cardia and Non-Cardia Subsites. *Cancers*, *11*(4), 534. DOI:10.3390/cancers11040534

Conway, J. R., Lex, A., and Gehlenborg, N. (2017) UpSetR: An R Package for the Visualization of Intersecting Sets and their Properties. *Bioinformatics 33*: 2938–2940. DOI:10.1093/bioinformatics/btx364

Cover, T. L. (2016). *Helicobacter pylori* Diversity and Gastric Cancer Risk. *MBio*, *7*(1), e01869-15, /mbio/7/1/e01869-15.atom. DOI:10.1128/mBio.01869-15

Cover, T. L. and Blaser, M. J. **(2009)** *Helicobacter pylori* in Health and Disease. *Gastroenterology* 136(6), P1863-1873. DOI:10.1053/j.gastro.2009.01.073

Croucher, N. J., Page, A. J., Connor, T. R., Delaney, A. J., Keane, J. A., Bentley, S. D., Parkhill, J., and Harris, S. R. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research*, *43*(3), e15–e15. DOI:10.1093/nar/gku1196

Dadashzadeh, K., Peppelenbosch, M. P., and Adamu, A. I. (2017). *Helicobacter pylori* Pathogenicity Factors Related to Gastric Cancer. *Canadian Journal of Gastroenterology and Hepatology*, *2017*, 1–6. DOI:10.1155/2017/7942489

de Bernardi Schneider, A., Ford, C. T., Hostager, R., Williams, J., Cioce, M., Çatalyürek, Ü. V., Wertheim, J. O., and Janies, D. (2019). StrainHub: A phylogenetic tool to construct pathogen transmission networks. (*preprint*), DOI:10.1101/650283

Didelot, X., and Wilson, D. J. (2015). ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLOS Computational Biology*, *11*(2), e1004041. DOI:10.1371/journal.pcbi.1004041

Doern, G. V. (2011). Antimicrobial Susceptibility Testing. *Journal of Clinical Microbiology*, *49*(9 Supplement), S4–S4. DOI:10.1128/JCM.00803-11

Draper, J. L., Hansen, L. M., Bernick, D. L., Abedrabbo, S., Underwood, J. G., Kong, N., Huang, B. C., Weis, A M., Weimer, B. C., van Vliet, A. H. M., Pourmand, N., Solnick, J. V., Karplus, K., and Ottemann, K. M. (2017). Fallacy of the Unique Genome: Sequence Diversity within Single *Helicobacter pylori* Strains. *MBio*, *8*(1), e02321-16, /mbio/8/1/e02321-16.atom. DOI:10.1128/mBio.02321-16

Duncan, S. S., Valk, P. L., McClain, M. S., Shaffer, C. L., Metcalf, J. A., Bordenstein, S. R., and Cover, T. L. (2013) Comparative Genomic Analysis of East Asian and Non-Asian *Helicobacter pylori* Strains Identifies Rapidly Evolving Genes. *PLOS ONE,* 8(1): e55120. DOI:10.1371/journal.pone.0055120

Erah, P. O., Goddard, A. F., Barrett, D. A., Shaw, P. N., and Spiller, R. C. (1997) The stability of amoxycillin, clarithromycin and metronidazole in gastric juice: relevance to the treatment of *Helicobacter pylori* infection. *Journal of Antimicrobial Chemotherapy*, 39*,* 5-12

Fagan-Garcia, K., Geary, J., Chang, H. J., McAlpine, L., Walker, E., Colquhoun, A., van Zanten, S. V., Girgis, S., Archie, B., Hanley, B., Corriveau, A., Morse, J., Munday, R., Goodman, K. J., and the CANHelp Working Group (2019). Burden of disease from *Helicobacter pylori* infection in western Canadian Arctic communities. *BMC Public Health*, *19*(1), 730. DOI:10.1186/s12889-019-7065-x

Fallone, C., Barkun, A., Göttke, M., Best, L., Loo, V., Zanten, S.V., Nguyen, T., Lowe, A., Fainsilber, T., Kouri, K., and Beech, R. (2000). Association of *Helicobacter pylori* genotype with gastroesophageal reflux disease and other upper gastrointestinal diseases. *American Journal of Gastroenterology*, 95, 659-669. DOI:10.1111/j.1572-0241.2000.01970.x.

Ferlay, J., Shin, H. R., Bray, F., Forman, D., Mathers, C., and Parkin, D. M. (2010) Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int. J. Cancer* 127:2893–2917. DOI:10.1002/ijc.25516.

Ferreira, R. M., Machado, J. C., Leite, M., Carneiro, F., and Figueiredo, C. (2012) The number of *Helicobacter pylori* CagA EPIYA C tyrosine phosphorylation motifs influences the pattern of gastritis and the development of gastric carcinoma. *Histopathology* 60(6), 992-998. DOI:10.1111/j.1365-2559.2012.04190.x

Fischer, W., Breithaupt, U., Kern, B., Smith, S. I., Spicher, C., & Haas, R. (2014). A comprehensive analysis of *Helicobacter pylori* plasticity zones reveals that they are integrating conjugative elements with intermediate integration specificity. *BMC Genomics, 15*(1), 310. DOI:10.1186/1471-2164-15-310

Fischer, W., Tegtmeyer, N., Stingl, K., and Backert, S. (2020) Four Chromosomal Type IV Secretion Systems in *Helicobacter pylori*: Composition, Structure and Function. *Front. Microbiol*., *11*:1592. DOI:10.3389/fmicb.2020.01592

Fuccio, L., Eusebi, L. H., Zagari, R. M., and Bazzoli, F. (2009). *Helicobacter pylori* eradication treatment reduces but does not abolish the risk of gastric cancer. *The American journal of gastroenterology*, 104(12), 3100–3102. DOI:10.1038/ajg.2009.516

Gaddy, J. A., Radin, J. N., Loh, J. T., Zhang, F., Washington, M. K., Peek, R. M., Jr, Algood, H. M., and Cover, T. L. (2013). High Dietary Salt Intake Exacerbates *Helicobacter pylori*-Induced Gastric Carcinogenesis. *Infection and Immunity, 81*(6), 2258–2267. DOI:10.1128/iai.01271-12

Guo, Y., Wang, H., Li, Y., Song, Y., Chen, C., Liao, Y., Ren, L., Guo, C., Tong, W., Shen, W., Chen, M., Mao, X., Guo, G., and Zou, Q. (2012). Genome of *Helicobacter pylori* strain XZ274, an isolate from a tibetan patient with gastric cancer in China. *Journal of Bacteriology*, *194*(15), 4146–4147. DOI:10.1128/JB.00804-12

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

González-Torres, P., Rodríguez-Mateos, F., Antón, J., and Gabaldón, T. (2019). Impact of Homologous Recombination on the Evolution of Prokaryotic Core Genomes. *MBio*, *10*(1), e02494-18, DOI:10.1128/mBio.02494-18

Goodman, K. J., Jacobson, K., and van Zanten, S. V. (2008). *Helicobacter Pylori* Infection in Canadian and Related Arctic Aboriginal Populations. *Canadian Journal of Gastroenterology*, *22*(3), 289–295. DOI:10.1155/2008/258610

Government of Canada (2017). RefSeqMasher: Mash MinHash search your nucleotide sequences against a NCBI RefSeq genomes database. Available online at: https://github.com/phac-nml/refseq_masher

Gutiérrez-Escobar, A. J., Trujillo, E., Acevedo, O., and Bravo, M. M. (2017). Phylogenomics of Colombian *Helicobacter pylori* isolates. *Gut Pathogens*, *9*(1), 52. DOI:10.1186/s13099-017-0201-1

Haley, K. P., and Gaddy, J. A. (2015). *Helicobacter pylori* : Genomic Insight into the Host-Pathogen Interaction. *International Journal of Genomics*, *2015*, 1–8. DOI:10.1155/2015/386905

Hashi, K., Imai, C., Yahara, K., Tahmina, K., Hayashi, T., Azuma, T., Miyabe-Nishiwaki, T., Sato, H., Matsuoka, M., Niimi, S., Okamoto, M., and Hatakeyama, M. (2018). Evaluating the origin and virulence of a *Helicobacter pylori* cagA-positive strain isolated from a non-human primate. *Scientific Reports*, *8*(1), 15981. DOI:10.1038/s41598-018-34425-4

Hastings, E. V., Yasui, Y., Hanington, P., Goodman, K. J., and the CANHelp Working Group (2014). Community-driven research on environmental sources of *H. pylori* infection in arctic Canada. *Gut Microbes*, *5*(5), 606–617. DOI:10.4161/19490976.2014.969639

Hatakeyama, M. and Higashi, H. (2005) *Helicobacter pylori* CagA: a new paradigm for bacterial carcinogenesis. *Cancer Science* 96(12), 835-843. DOI:10.1111/j.1349-7006.2005.00130.x

Hazırolan, G. (2021). *Capnocytophaga sputigena*: A rarely encountered gram-negative bacterium in microbiology laboratories. *Journal of infection and public health, 14*(5), 683–684. DOI:10.1016/j.jiph.2020.09.021

Heydari, M., Miclotte, G., Van de Peer, Y., and Fostier, J. (2019) Illumina error correction near highly repetitive DNA regions improves de novo genome assembly. *BMC Bioinformatics*, *20*, 209. DOI:10.1186/s12859-019-2906-2

Hooi, J. K. Y., Lai, W. Y., Ng, W. K., Suen, M. M. Y., Underwood, F. E., Tanyingoh, D., Malfertheiner, P., Graham, D. Y., Wong, V. W. S., Wu, J. C. Y., Chan, F. K., L., Sung, J. J. Y., Kaplan, G. G., and Ng, S. C. (2017). Global Prevalence of *Helicobacter pylori* Infection: Systematic Review and Meta-Analysis. *Gastroenterology*, *153*(2), 420–429. DOI:10.1053/j.gastro.2017.04.022

Huang, W., Li, L., Myers, J. R., and Marth G. T. (2011) ART: a next-generation sequencing read simulator. *Bioinformatics, 28* (4), 593--594. DOI:10.1093/bioinformatics/btr708

Ito, Y., Azuma, T., Ito, S., Suto, H., Miyaji, H., Yamazaki, Y., Kohli, Y., and Kuriyama, M. (1998) Full-length sequence analysis of the *vacA* gene from cytotoxic and noncytotoxic *Helicobacter pylori*. *J. Infect. Dis*, *178*(5):1391-1398. DOI:10.1086/314435.

Iwamoto, A., Tanahashi, T., Okada, R., Yoshida, Y., Kikuchi, K., Keida, Y., Murakami, Y., Yang, L., Yamamoto, K., Nishiumi, S., Yoshida, M. and Azuma, T. (2014) Whole-genome sequencing of clarithromycin resistant *Helicobacter pylori* characterizes unidentified variants of multidrug resistant efflux pump genes. *Gut Pathogens* 6(27), DOI:10.1186/1757-4749-6-27

Jolley, K. A. and Maiden, M. C. J. (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics, 11*, 595. DOI:10.1186/1471-2105-11-595

Kalaghatgi, P., and Lengauer, T. (2017). Computing phylogenetic trees using topologically related minimumspanning trees. *Journal of Graph Algorithms and Applications*, *21*(6), 1003–1025. DOI:10.7155/jgaa.00447

Kao, C.-Y., Sheu, B.-S., and Wu, J.-J. (2016). *Helicobacter pylori* infection: An overview of bacterial virulence factors and pathogenesis. *Biomedical Journal*, *39*(1), 14–23. DOI:10.1016/j.bj.2015.06.002

Karlsson, R., Thorell, K., Hosseini, S., Kenny, D., Sihlbom, C., Sjöling, Å., Karlsson, A., and Nookaew, I. (2016). Comparative Analysis of Two *Helicobacter pylori* Strains using Genomics

and Mass Spectrometry-Based Proteomics. *Frontiers in Microbiology*, *7*. DOI:10.3389/fmicb.2016.01757

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research, 30*(14), 3059-3066. DOI:10.1093/nar/gkf436

Katz, L. S., Griswold, T., Morrison, S. S., Caravas, J. A., Zhang, S., den Bakker, H. C., Deng, X., and Carleton, H. A. (2019) Mashtree: a rapid comparison of whole genome sequence files. *Journal of Open Source Software, 4*(44), 1762, DOI:10.21105/joss.01762

Kennemann, L., Didelot, X., Aebischer, T., Kuhn, S., Drescher, B., Droege, M., Reinhardt, R., Correa, P., Meyer, T. F., Josenhans, C., Falush, D., and Suerbaum, S. (2011). *Helicobacter pylori* genome evolution during human infection. *Proceedings of the National Academy of Sciences*, *108*(12), 5033–5038. DOI:10.1073/pnas.1018444108

Kersulyte, D., Bertoli, M. T., Tamma, S., Keelan, M., Munday, R., Geary, J., Veldhuyzen van Zanten, S., Goodman, K. J., and Berg, D. E. (2015). Complete Genome Sequences of Two *Helicobacter pylori* Strains from a Canadian Arctic Aboriginal Community. *Genome Announcements*, *3*(2), e00209-15, /ga/3/2/e00209-15.atom. DOI:10.1128/genomeA.00209-15

Khalifa, M., Sharaf, R., and Aziz, R. (2010). *Helicobacter pylori*: A poor man's gut pathogen? *Gut Pathogens*, *2*(1), 2. DOI:10.1186/1757-4749-2-2

Kodaman, N., Pazos, A., Schneider, B. G., Piazuelo, M. B., Mera, R., Sobota, R. S., Sicinschi, L. A., Shaffer, C. L., Romero-Gallo, J., de Sablet, T., Harder, R. H., Bravo, L. E., Peek, R. M., Jr, Wilson, K. T., Cover, T. L., Williams, S. M., and Correa, P. (2014). Human and *Helicobacter pylori* coevolution shapes the risk of gastric disease. *Proceedings of the National Academy of Sciences*, *111*(4), 1455–1460. DOI:10.1073/pnas.1318093111

Krzyżek, P. (2021) Toxin-Antitoxin Systems - A New Player in Morphological Transformation of Antibiotic-Exposed *Helicobacter pylori*? *Front. Cell. Infect. Microbiol., 11:* 670667. DOI: 10.3389/fcimb.2021.670677

Kuipers, E. J., Israel, D. A., Kusters, J. G., Gerrits, M. M., Weel, J., van der Ende, A., van der Hulst, R. W. M., Wirth, H. P., Hook-Nikanne, J., Thompson, S. A., and Blaser, M. J. (2000) Quasispecies Development of *Helicobacter pylori* Observed in Paired Isolates Obtained Years Apart from the Same Host. *Journal of Infectious Diseases* 181(1), 273-282, DOI:10.1086/315173

Kusters, J. G., van Vliet, A. H. M., and Kuipers, E. J. (2006) Pathogenesis of *Helicobacter pylori* infection. *Clinical Microbiology Reviews* 19(3), 449-490, DOI:10.1128/CMR.00054-05

Lamichhane, B., Chua, E.-G., Wise, M. J., Laming, C., Marshall, B. J., and Tay, C.-Y. (2019). The complete genome and methylome of *Helicobacter pylori* hpNEAfrica strain HP14039. *Gut Pathogens*, *11*(1), 7, DOI:10.1186/s13099-019-0284-y

Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. (2012) Inference of Population Structure using Dense Haplotype Data. *PLoS Genetics, 8(1),* e1002453, DOI:10.1371/journal.pgen.1002453

Le, P. T., Hamasuna, R., Matsumoto, M., Furubayashi, K., Hatanaka, M., Kawai, S., Yamaguchi, T., Uehara, K., Murakami, N., Yoshioka, M., Nakayama, K., Shiono, Y., Muraoka, K., Suzuki, M., Fujimoto, N., and Matsumoto, T. (2017). The detection of microorganisms related to urethritis from the oral cavity of male patients with urethritis. *Journal of infection and chemotherapy : official journal of the Japan Society of Chemotherapy, 23*(10), 668–673. DOI:10.1016/j.jiac.2017.06.011

Lefebvre, M., Chang, H.-J., Morse, A., van Zanten, S. V., Goodman, K. J., and CAN*Help* Working Group. (2013). Adherence and barriers to *H. pylori* treatment in Arctic Canada. *International Journal of Circumpolar Health*, *72*(1), 22791, DOI:10.3402/ijch.v72i0.22791

Lin, M., and Kussell, E. (2019). Inferring bacterial recombination rates from large-scale sequencing datasets. *Nature Methods*, *16*(2), 199–204, DOI:10.1038/s41592-018-0293-7

Linz, B., Balloux, F., Moodley, Y., Manica, A., Liu, H., Roumagnac, P., Falush, D., Stamer, C., Prugnolle, F., van der Merwe, S. W., Yamaoka, Y., Graham, D. Y., Perez-Trallero, E., Wadstrom, T., Suerbaum, S., and Achtman, M. (2007). An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature*, *445*(7130), 915–918. DOI:10.1038/nature05562

Linz, B., Windsor, H. M., Gajewski, J. P., Hake, C. M., Drautz, D. I., Schuster, S. C., and Marshall, B. J. (2013). *Helicobacter pylori* Genomic Microevolution during Naturally Occurring Transmission between Adults. *PLOS ONE*, *8*(12), e82187, DOI:10.1371/journal.pone.0082187

Linz, B., Windsor, H. M., McGraw, J. J., Hansen, L. M., Gajewski, J. P., Tomsho, L. P., Hake, C. M., Solnick, J. V., Schuster, S. C., and Marshall, B. J. (2014). A mutation burst during the acute phase of *Helicobacter pylori* infection in humans and rhesus macaques. *Nature Communications, 5*(1). DOI:10.1038/ncomms5165

Lloyd, T. (1977). Association of Canadian Universities for Northern Studies. *Environmental Conservation*, *4*(2), 114–114, DOI:10.1017/S0376892900025340

Lu, J. (2018). KrakenTools. Available online at: https://github.com/jenniferlu717/KrakenTools

Luther, J., Dave, M., Higgins, P. D. R., and Kao, J. Y. (2010) Association Between *Helicobacter pylori* Infection and Inflammatory BowelDisease: A Meta-analysis and Systematic Review of the Literature. *Inflammatory Bowel Disease, 16*(6), 1077-1084

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

Maixner, F., Krause-Kyora, B., Turaev, D., Herbig, A., Hoopmann, M. R., Hallows, J. L., Kusebauch, U., Vigl, E. E., Malfertheiner, P., Megraud, F., O'Sullivan, N., Cipollini, G., Coia, V., Samadelli, M., Engstrand, L., Linz, B., Moritz, R. L., Grimm, R., Krause, J., Nebel, A., Moodley, Y., Rattei, T., and Zink, A. (2016). The 5300-year-old *Helicobacter pylori* genome of the Iceman. *Science*, *351*(6269), 162–165. DOI:10.1126/science.aad2545

Marshall, B. J. and Warren, J. R. (1984) Unidentified curved Bacilli in the stomach of patients with gastritis and peptic ulceration. *The Lancet, 323*(8390), 1311-131, DOI:10.1016/S0140-6736(84)91816-6

Marshall, B. J., Armstrong, J. A., McGechie, D. B., and Glancy, R. J. (1985) Attempt to fulfil Koch's postulates for pyloric campylobacter. *The Medical Journal of Australia, 142*, 436-439

Marshall, B. J. and Adams, P. C. (2008) *Helicobacter pylori*: A Nobel pursuit? *Canadian Journal of Gastroenterology, 22*(11), 895-896, DOI:10.1155/2008/459810

Mascellino, M. T., Porowska, B., De Angelis, M., and Oliva, A. (2017). Antibiotic susceptibility, heteroresistance, and updated treatment strategies in *Helicobacter pylori* infection. *Drug Design, Development and Therapy*, *11*, 2209–2220, DOI:10.2147/DDDT.S136240

Matsuo, Y., Kido, Y., and Yamaoka, Y. (2017). Helicobacter pylori Outer Membrane Protein-Related Pathogenesis. *Toxins*, *9*(3), 101. DOI:10.3390/toxins9030101

Matthews, T. C., Bristow, F. R., Griffiths, E. J., Petkau, A., Adam, J., Dooley, D., Kruczkiewicz, P., Curatcha, J., Cabral, J., Fornika, D., Windsor, G. L., Courtot, M., Bertelli, C., Roudgar, A., Feijao, P., Mabon, P., Enns, E., Thiessen, J., Keddy A., Isaac-Renton, J., Gardy, J. L., Tang, P., The IRIDA Consortium, Carrico, J. A., Chindelevitch, L., Chauve, C., Graham, M. R., McArthur, A. G., Taboada, E. N., Beiko, R. G., Brinkman, F. S. L., Hsiao, W. W. L., and Van Domselaar, G. (2018). The Integrated Rapid Infectious Disease Analysis (IRIDA) Platform, (*preprint)*, DOI:10.1101/381830

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., Bhullar, K., Canova, M. J., De Pascale, G., Ejim, L., Kalan, L., King, A. M., Koteva, K., Morar, M., Mulvey, M. R., O'Brien, J. S., Pawlowski, A. C., Piddock, L. J. V., Spanogiannopoulos, P., Sutherland, A. D., Tang, I., Taylor, P. L., Thaker, M., Wang, W., Yan, M., Yu, T., and Wright, G. D. (2013) The comprehensive antibiotic resistance database. *Antimicrobial agents and chemotherapy*, *57*(7), 3348–3357. DOI:10.1128/AAC.00419-13

McMahon, B. J., Bruce, M. G., Koch, A., Goodman, K. J., Tsukanov, V., Mulvad, G., Borresen, M. L., Sacco, F., Barrett, D., Westby, S., and Parkinson, A. J. (2016). The diagnosis and treatment of *Helicobacter pylori* infection in Arctic regions with a high prevalence of infection: Expert Commentary. *Epidemiology and Infection*, *144*(2), 225–233. DOI:10.1017/S0950268815001181

Méndez-Tenorio, A., Larios-Serrato, V., Olguín-Ruiz, G. E., Sánchez-Vallejo, C. J., Torres-López, R. C., Avilés-Jiménez, F., Camorlinga-Ponce, M., and Torres, J. (2014). Genome Sequence of a *Helicobacter pylori* Strain Isolated from a Mexican Patient with Intestinal

Gastric Cancer. *Genome Announcements*, *2*(1), e01214-13, 2/1/e01214-13. DOI:10.1128/genomeA.01214-13

Mera, R., Fontham, E. T., Bravo, L. E., Bravo, J. C., Piazuelo, M. B., Camargo, M. C., and Correa, P. (2005). Long term follow up of patients treated for *Helicobacter pylori* infection. *Gut,* 54(11), 1536–1540. DOI:10.1136/gut.2005.072009

Mera, R. M., Bravo, L. E., Goodman, K. J., Yepez, M. C., and Correa, P. (2012). Long-term Effects of Clearing *Helicobacter pylori* on Growth in School-age Children: *The Pediatric Infectious Disease Journal*, *31*(3), 263–266, DOI:10.1097/INF.0b013e3182443fec

Michie, K. A., Boysen, A., Low, H. H., Møller-Jensen, J., and Löwe, J. (2014) LeoA, B and C from Enterotoxigenic *Escherichia coli* (ETEC) Are Bacterial Dynamins. *PLoS ONE*, *9*(9):e107211. DOI:10.1371/journal.pone.0107211

Miftahussurur, M., Nusi, I. A., Graham, D. Y., and Yamaoka, Y. (2017). Helicobacter, Hygiene, Atopy, and Asthma. *Frontiers in microbiology, 8*, 1034. DOI:10.3389/fmicb.2017.01034

Moodley, Y., Linz, B., Bond, R. P., Nieuwoudt, M., Soodyall, H., Schlebusch, C. M., Bernhöft, S., Hale, J., Suerbaum, S., Mugisha, L., van der Merwe, S. W., and Achtman, M. (2012). Age of the Association between *Helicobacter pylori* and Man. *PLOS Pathogens*, *8*(5), e1002693, DOI:10.1371/journal.ppat.1002693

Morelli, G., Didelot, X., Kusecek, B., Schwarz, S., Bahlawane, C., Falush, D., Suerbaum, S., & Achtman, M. (2010). Microevolution of *Helicobacter pylori* during prolonged infection of single hosts and within families. *PLoS genetics, 6*(7), e1001036. DOI:10.1371/journal.pgen.1001036

Morse, A. L., Goodman, K. J., Munday, R., Chang, H. J., Morse, J. W., Keelan, M., Geary, J., van Zanten, S. V., & CANHelp Working Group (2013). A randomized controlled trial comparing sequential with triple therapy for Helicobacter pylori in an Aboriginal community in the Canadian North. *Canadian Journal of Gastroenterology, 27*(12), 701–706, DOI:10.1155/2013/143047

Muñoz, A. B., Stepanian, J., Trespalacious, A. A., and Vale, F. F. (2020) Bacteriophages of *Helicobacter pylori*. *Front. Microbiol., 11*:549084. DOI:10.3389/fmicb.2020.549084

Muñoz-Ramírez, Z. Y., Mendez-Tenorio, A., Kato, I., Bravo, M. M., Rizzato, C., Thorell, K., Torres, R., Aviles-Jimenez, F., Camorlinga, M., Canzian, F., & Torres, J. (2017). Whole Genome Sequence and Phylogenetic Analysis Show *Helicobacter pylori* Strains from Latin America Have Followed a Unique Evolution Pathway. *Frontiers in Cellular and Infection Microbiology*, *7*(50), DOI:10.3389/fcimb.2017.00050

Muñoz-Ramirez, Z. Y., Pascoe, B., Mendez-Tenorio, A., Mourkas, E., Sandoval-Motta, S., Perez-Perez, G., Morgan, D. R., Dominguez, R. L., Ortiz-Princz, D., Cavazza, M. E., Rocha, G., Queiroz, D., Catalano, M., Palma, G. Z., Goldman, C. G., Venegas, A., Alarcon, T., Oleastro, M., Vale, F. F., Goodman, K. J., Torres, R. C., Berthenet, E., Hitchings, M. D., Blaser, M. J.,

Sheppard, S. K., Thorell, K., and Torres, J. (2020). A 500-year tale of co-evolution, adaptation, and virulence: Helicobacter pylori in the Americas. *The ISME Journal*, DOI:10.1038/s41396-020-00758-0

Nishizawa, T. and Suzuki, H. (2014). Mechanisms of *Helicobacter pylori* antibiotic resistance and molecular testing. *Frontiers in molecular biosciences*, 1, 19. DOI:10.3389/fmolb.2014.00019

Nitharwal, R. G., Verma, V., Dasgupta, S., and Dhar, S. K. (2011). *Helicobacter pylori* chromosomal DNA replication: Current status and future perspectives. *FEBS Letters*, *585*(1), 7–17, DOI:10.1016/j.febslet.2010.11.018

Nolen, L. D., Bruden, D., Miernyk, K., McMahon, B. J., Sacco, F., Varner, W., Mezzetti, T., Hurlburt, D., Tiesinga, J., & Bruce, M. G. (2018). *H. pylori* -associated pathologic findings among Alaska native patients. *International Journal of Circumpolar Health*, *77*(1), 1510715, DOI:10.1080/22423982.2018.1510715

Noto, J. M., Chopra, A., Loh, J. T., Romero-Gallo, J., Piazuelo, M. B., Watson, M., Leary, S., Beckett, A. C., Wilson, K. T., Cover, T. L., Mallal, S., Israel, D. A., & Peek, R. M. (2018). Pan-genomic analyses identify key *Helicobacter pylori* pathogenic loci modified by carcinogenic host microenvironments. *Gut*, *67*(10), 1793–1804, DOI:10.1136/gutjnl-2017-313863

O'Donoghue, S. I., Gavin, A. C., Gehlenborg, N., Goodsell, D. S., Hériché, J. K., Nielsen, C. B., North, C., Olson, A. J., Procter, J. B., Shattuck, D. W., Walter, T., Wong, B. (2010).

Visualizing biological data—now and in the future. *Nature Methods*, *7*(S3), S2–S4, DOI:10.1038/nmeth.f.301

Ogura, M., Perez, J. C., Mittl, P. R. E., Lee, H.-K., Dailide, G., Tan, S., Ito, Y., Secka, O., Dailidiene, D., Putty, K., Berg, D. E., and Kalia, A. (2007). *Helicobacter pylori* Evolution: Lineage- Specific Adaptations in Homologs of Eukaryotic Sel1-Like Genes. *PLoS Computational Biology*, *3*(8), 13, DOI:10.1371/journal.pcbi.0030151

Okonechnikov, K., Golosova, O., Fursov, M., and the UGENE team (2012) Unipro UGENE: a unified bioinformatics toolkit, *Bioinformatics, 28*(8), 1166–1167, DOI:10.1093/bioinformatics/bts091

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, *31*(22), 3691–3693, DOI:10.1093/bioinformatics/btv421

Perkins, T. T., Tay, C. Y., Thirriot, F., and Marshall, B. (2013). Choosing a Benchtop Sequencing Machine to Characterise *Helicobacter pylori* Genomes. *PLoS ONE*, *8*(6), e67539, DOI:10.1371/journal.pone.0067539

Pincock, S. (2005) Nobel Prize winners Robin Warren and Barry Marshall. *The Lancet* 366(9495), 1429, DOI:10.1016/S0140-6736(05)67587-3

Plummer, M., Franceschi, S., Vignat, J., Forman, D., and de Martel, C. (2015). Global burden of gastric cancer attributable to *Helicobacter pylori*: *Helicobacter pylori* in gastric cancer. *International Journal of Cancer*, *136*(2), 487–490, DOI:10.1002/ijc.28999

Pohl, D., Keller, P. M., Bordier, V., & Wagner, K. (2019). Review of current diagnostic methods and advances in Helicobacter pylori diagnostics in the era of next generation sequencing. *World Journal of Gastroenterology*, *25*(32), 4629–4660, DOI:10.3748/wjg.v25.i32.4629

Rizzato, C., Torres, J., Plummer, M., Muñoz, N., Franceschi, S., Camorlinga-Ponce, M., Fuentes-Pananá, E. M., Canzian, F., & Kato, I. (2012). Variations in *Helicobacter pylori* cytotoxin-associated genes and their influence in progression to gastric cancer: implications for prevention. *PloS one, 7*(1), e29605, DOI:10.1371/journal.pone.0029605

Robertson, J. and Nash, J. H. E. (2018) MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microbial Genomics* 4(8), DOI10.1099/mgen.0.000206

Robinson, K., Loughlin, M. F., Potter, R. and Jenks, P. J. (2005) Host Adaptation and Immune Modulation Are Mediated by Homologous Recombination in *Helicobacter pylori*. *Journal of Infectious Diseases* 191, 579-587, DOI:10.1086/427657

Rodrıguez Gomez, E. R., Otero Regino, W., Monterrey, P. A., Trespalacios Rangel, A. A. (2020) *cagA* gene EPIYA motif genetic characterization from Colombian *Helicobacter pylori isolates*: Standardization of a molecular test for rapid clinical laboratory detection. *PLoS ONE 15(1)*:e0227275. DOI:10.1371/journal.pone.0227275

Rohrer, S., Holsten, L., Weiss, E., Benghezal, M., Fischer, W., and Haas R. (2012) Multiple Pathways of Plasmid DNA Transfer in *Helicobacter pylori*. *PLoS One, 7*(9): e45623. DOI:10.1371/journal.pone.0045623

Ryan, M. P., & Pembroke, J. T. (2018). *Brevundimonas spp*: Emerging global opportunistic pathogens. *Virulence, 9*(1), 480–493. DOI:10.1080/21505594.2017.1419116

Salama, N., Guillemin, K., McDaniel, T. K., Sherlock, G., Tompkins, L., and Falkow, S. (2000). A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proceedings of the National Academy of Sciences*, *97*(26), 14668–14673, DOI:10.1073/pnas.97.26.14668

Salipante, S. J., and Hall, B. G. (2011). Inadequacies of Minimum Spanning Trees in Molecular Epidemiology. *Journal of Clinical Microbiology*, *49*(10), 3568–3575, DOI:10.1128/JCM.00919-11

Salipante, S. J., SenGupta, D. J., Cummings, L. A., Land, T. A., Hoogestraat, D. R., and Cookson, B. T. (2015). Application of Whole-Genome Sequencing for Bacterial Strain Typing

in Molecular Epidemiology. *Journal of Clinical Microbiology*, *53*(4), 1072–1079, DOI:10.1128/JCM.03385-14

Satomi, S., Yamakawa, A., Matsunaga, S., Masaki, R., Inagaki, T., Okuda, T., Suto, H., Ito, Y., Yamazaki, Y., Kuriyama, M., Keida, Y., Kutsumi, H., and Azuma, T. (2006). Relationship between the diversity of the *cagA* gene of *Helicobacter pylori* and gastric cancer in Okinawa, Japan. *Journal of gastroenterology*, *41*(7), 668–673. DOI:10.1007/s00535-006-1838-6

Schnarch, B. (2004). Ownership, Control, Access, and Possession (OCAP) or Self-Determination Applied to Research. *Journal of Aboriginal Health* 1(1), DOI:10.3138/ijih.v1i1.28934

Scott, D. R., Sachs, G., and Marcus, E. A. (2016). The role of acid inhibition in *Helicobacter pylori* eradication. *F1000Research, 5, F1000 Faculty* Rev-1747. DOI:10.12688/f1000research.8598.1

Secka, O., Berg, D. E., Antonio, M., Corrah, T., Tapgun, M., Walton, R., Thomas, V., Galano, J. J., Sancho, J., Adegbola, R. A., & Thomas, J. E. (2013). Antimicrobial Susceptibility and Resistance Patterns among *Helicobacter pylori* Strains from The Gambia, West Africa. *Antimicrobial Agents and Chemotherapy*, *57*(3), 1231–1237, DOI:10.1128/AAC.00517-12

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

Seeman, T. (2016) MLST: Scan contig files against PubMLST typing schemes. Available at: https://github.com/tseemann/mlst

Shetty, V., Lamichhane, B., Chua, E.-G., Ballal, M., and Tay, C.-Y. (2018). Draft Genome Sequences of 42 *Helicobacter pylori* Isolates from Rural Regions of South India. *Genome Announcements*, *6*(5), e01486-17, DOI:10.1128/genomeA.01486-17

Sims, G. E., and Kim, S.-H. (2011). Whole-genome phylogeny of *Escherichia coli/Shigella* group by feature frequency profiles (FFPs). *Proceedings of the National Academy of Sciences*, *108*(20), 8329–8334, DOI:10.1073/pnas.1105168108

Sims, G. E., Jun, S.-R., Wu, G. A., and Kim, S.-H. (2009). Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences*, *106*(8), 2677–2682, DOI:10.1073/pnas.0813249106

Skidmore, Z. L., Wagner, A. H., Lesurf, R., Campbell, K. M., Kunisaki, J., Griffith, O. L., and Griffith, M. (2016). GenVisR: Genomic Visualizations in R. *Bioinformatics*, *32*(19), 3012–3014, DOI:10.1093/bioinformatics/btw325

Stefano, K., Marco, M., Federica, G., Laura, B., Barbara, B., Gioacchino, L., Franceso, d. M., Gian, L. dA. (2018) *Helicobacter pylori*, transmission routes and recurrence of infection: state of the art. *Acta bio-medica : Atenei Parmensis*, *89*(8-S), 72–76. DOI:10.23750/abm.v89i8-S.7947

Stevenson, T. H., Castillo, A., Lucia, L. M., and Acuff, G. R. (2001) Growth of *Helicobacter pylori* in various liquid and plating media. *Letters in Applied Microbiology*, *30*(3): 192-196. DOI:10.1046/j.1472-765x.2000.00699.x

Stothard, P., Grant, J. R., and Van Domselaar, G. (2017). Visualizing and comparing circular genomes using the CGView family of tools. *Briefings in Bioinformatics*. *20*(4), 1576-1582 DOI:10.1093/bib/bbx081

Sugimoto, M., Zali, M. R., and Yamaoka, Y. (2009) The association of *vacA* genotypes and *Helicobacter pylori*-related gastroduodenal diseases in the Middle East. *Eur J Clin Microbiol Infect Dis*, 28(10), 1227-1236. DOI:10.1007/s10096-009-0772-y

Tacconelli, E., Carrara, E., Savoldi, A., Harbarth, S., Mendelson, M., Monnet, D. L., Pulcini, C., Kahlmeter, G., Kluytmans, J., Carmeli, Y., Ouellette, M., Outterson, K., Patel, J., Cavaleri, M., Cox, E. M., Houchens, C. R., Grayson, M. L., Hansen, P., Singh, N., Theuretzbacher, U., … WHO Pathogens Priority List Working Group (2018). Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *The Lancet. Infectious diseases, 18*(3), 318–327. DOI:10.1016/S1473-3099(17)30753-3

Tada, I., Tanizawa, Y., and Arita, M. (2017). Visualization of consensus genome structure without using a reference genome. *BMC Genomics*, *18*(S2), 208, DOI:10.1186/s12864-017-3499-7

Teixeira, A. S., Monteiro, P. T., Carriço, J. A., Ramirez, M., and Francisco, A. P. (2015). Not Seeing the Forest for the Trees: Size of the Minimum Spanning Trees (MSTs) Forest and Branch Significance in MST-Based Phylogenetic Analysis. *PLOS ONE*, *10*(3), e0119315, DOI:10.1371/journal.pone.0119315

Thorell, K., Lehours, P., and Vale, F. F. (2017). Genomics of *Helicobacter pylori*. *Helicobacter*, *22*, e12409, DOI:10.1111/hel.12409

Thorell, K., Yahara, K., Berthenet, E., Lawson, D. J., Kato, I., Mendez, A., and Rizzato, C. (2017). Rapid evolution of distinct *Helicobacter pylori* subpopulations in the Americas. *PLOS Genetics*, *21*, DOI:10.1371/journal.pgen.1006546

Tomb, J.-F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H. G., Glodek, A., McKenney, K., Fitzegerald, L. M., Lee, N., Adams, M. D., Hickey, E. K., Berg D. E., Gocayne, J. D., Utterback, T. R., Peterson, J. D., Kelley, J. M., Cotton, M. D., Weidman, J. M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W. S., Borodovsky, M., Karp, P. D., Smith, H. O., Fraser, C. M., Venter, J. C. (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*, *388*(6642), 539–547, DOI:10.1038/41483

Tshibangu-Kabamba, E., and Yamaoka, Y. (2021) *Helicobacter pylori* infection and antibiotic resistance - from biology to clinical implications. *Nature REviews Gastroenterology & Hepatology*, *18*, 613-629.

Uchiyama, I., Albritton, J., Fukuyo, M., Kojima, K. K., Yahara, K., and Kobayashi, I. (2016). A Novel Approach to *Helicobacter pylori* Pan-Genome Analysis for Identification of Genomic Islands. *PLOS ONE*, *11*(8), e0159419, DOI:10.1371/journal.pone.0159419

Vale, F. F., Vadivelu, J., Oleastro, M., Breurec, S., Engstrand, L., Perets, T. T., Megraud, F., and Lehours, P. (2015) Dormant phages of *Helicobacter pylori* reveal distinct populations in Europe. *Scientific Reports, 5*: 14333. DOI:10.1038/srep14333

van Vliet, A. H. M. (2017). Use of pan-genome analysis for the identification of lineage-specific genes of *Helicobacter pylori*. *FEMS Microbiology Letters*, *364*(2), DOI:10.1093/femsle/fnw296

van Zanten, S. V. and Sherman, P. M. (1994) *Helicobacter pylori* infection as a cause of gastritis, duodenal ulcer, gastric cancer and nonulcer dyspepsia: a systematic overview. *Canadian Medical Association Journal*, *150*(2), 177-185

Wang, D., Guo, Q., Lv, Z., Yuan, Y., and Gong, Y. (2018). Molecular Detection of *H. pylori* Antibiotic-Resistant Genes and Bioinformatics Predictive Analysis. *(preprint)* DOI:10.1101/325654

Wen, S., and Moss, S. F. (2009). *Helicobacter pylori* virulence factors in gastric carcinogenesis. *Cancer Letters*, *282*(1), 1–8, DOI:10.1016/j.canlet.2008.11.016

Williams, K. F. (2016) Investigating the Effect of Antibiotic Exposure on the Prevalence of Antibiotic-resistant *H. pylori* Infection and the Incidence of Anti-*H. pylori* Treatment Failure in Northern Canadian Communities [Master's Thesis]. University of Alberta

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

Wood, D. E. and Salzberg, S. L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology, 15*, R46 DOI:10.1186/gb-2014-15-3-r46

World Health Organization: International Agency for Research on Cancer. (1994) Schistosomes, live flukes and *Helicobacter pylori*. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, Lyon, 7–14 June 1994. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. 61*:1–241

Xu, C., Soyfoo, D. M., Wu, Y., and Xu, S. (2020) Virulence of *Helicobacter pylori* outer membrane proteins: an updated review. *Eur J Clin Microbiol Infect Dis 39*, 1821–1830. DOI:10.1007/s10096-020-03948-y

Xue, Z., Yang, H., Su, D., Song, X., Deng, X., Yu, C., Sun, C., He, L., You, Y., Gong, Y., Fan, D., Sun, L., Han, X., Fan, R., Zhang, M., Yan, X., Qian, J., and Zhang, J. (2021) Geographic distribution of the *cagA, vacA, iceA, oipA* and *dupA* genes of *Helicobacter pylori* strains isolated in China. *Gut Pathogens* 13(39). DOI: doi.org/10.1186/s13099-021-00434-4

Yahara, K., Furuta, Y., Oshima, K., Yoshida, M., Azuma, T., Hattori, M., Uchiyama, I., and Kboayashi, I. (2013) Chromosome Paintin In Silico in a Bacterial Species Reveals Fine Population Structure. *Molecular Biology and Evolution, 30(6)*, 1454-1464. DOI:10.1093/molbev/mst055

Yang, J., Cao, Z., Chen, H., Long, K., Li, G., and Zhao, L. (2011). A Method for Constructing Phylogenetic Tree Based on the Minimum Spanning Tree of the Complete Graph. *Communications in Mathematical and Computer Chemistry* 65, 469-476.

Yoshiyama, H. and Nakazawa, T. (2000) Unique mechanism of *Helicobacter pylori* for colonizing the gastric mucus. *Microbes and Infection* 2, 55-60, DOI:10.1016/s1286-4579(00)00285-9

Young, T. K., Kelly, J. J., Friborg, J., Soininen, L., and Wong, K. O. (2016). Cancer among circumpolar populations: An emerging public health concern. *International Journal of Circumpolar Health*, *75*(1), 29787, DOI:10.3402/ijch.v75.29787

Yu, G. (2020) Using ggtree to Visualize Data on Tree-Like Structures. *Current Protocols in Bioinformatics, 69*(1), e96. DOI:10.1002/cpbi.96

Yu, G., Lam, T. T-Y., Zhu, H., and Guan, Y. (2018) Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using *ggtree*. *Molecular Biology and Evolution*, *35*(12), 3041–3043. DOI:10.1093/molbev/msy194

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T-Y (2016) ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28-36. DOI:10.1111/2041-210X.12628

Yuan, X. Y., Wang, Y., and Wang, M. Y. (2018). The type IV secretion system in *Helicobacter pylori. Future Microbiology*, 13, 1041–1054. DOI:10.2217/fmb-2018-0038

Zhang, X., Peng, C., Zhang, G., and Gao, F. (2015). Comparative analysis of essential genes in prokaryotic genomic islands. *Scientific Reports*, *5*(1), 12561, DOI:10.1038/srep12561

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. and Wishart, D. S. (2011) PHAST: a fast phage search tool. *Nucleic Acids Res.* 39, W347–W352. DOI:10.1093/nar/gkr485

Appendix

CANHelp Working Group Goals

We use a community-driven research approach in northern communities to:

- Characterize the burden of disease from *H. pylori* infection
- Exchange knowledge between scientists and concerned community members and health care providers to work together to reduce health risks from this infection

Our research goals, based on the priorities of the participating communities and their health care providers, are to:

1. Describe the disease burden and risk factors associated with *H. pylori* infection;
2. Develop clinical management approaches to reduce health risks from *H. pylori* infection;
3. Develop knowledge exchange strategies that help foster a greater understanding of health risks associated with *H. pylori* infection, as well as solutions and unsolved challenges for reducing these risks.

Table A2.1. Publicly available *H. pylori* genomes used in the present study

Strain	Source	Country	Hp Type
K26A1	GCF_001653455	Angola	hpAfrica2
SA144A	GCF_000475715	South Africa	hpAfrica2
SA155C	GCF_000475915	South Africa	hpAfrica2
SA166A	GCF_000476255	South Africa	hpAfrica2
SA169C	GCF_000475955	South Africa	hpAfrica2
SA172C	GCF_000476815	South Africa	hpAfrica2
SA174A	GCF_000477015	South Africa	hpAfrica2
SouthAfrica20	GCF_000590775	South Africa	hpAfrica2
SouthAfrica50	GCF_000448525	South Africa	hpAfrica2
SouthAfrica7	GCF_000185245	South Africa	hpAfrica2
132	GCF_001939635	Singapore	hpAsia2
FD662	GCF_000296095	Malaysia	hpAsia2
India7	GCF_000185185	India	hpAsia2
Mandalay03	https://datadryad.org/resource/doi:10.5061/dryad.8qp4n/2	Myanmar	hpAsia2
NAK7	GCF_000401295	India	hpAsia2
NP04	https://datadryad.org/resource/doi:10.5061/dryad.8qp4n/2	Nepal	hpAsia2
NP05	https://datadryad.org/resource/doi:10.5061/dryad.8qp4n/2	Nepal	hpAsia2
SA222C	GCF_000476435	South Africa	hpAsia2
SNT49	GCF_000224575	India	hpAsia2
Yangon233	https://datadryad.org/resource/doi:10.5061/dryad.8qp4n/2	Myanmar	hpAsia2
PeCan18	GCF_000277425	Peru	hspAfrica1MiscAmerica
J99	GCF_000008785	USA	hspAfrica1NAmerica
SA161C	GCF_000475575	South Africa	hspAfrica1NAmerica
HN-G258	HEL_AA2633AA	Honduras	hspAfrica1SAfrica
HN-G285	HEL_AA2664AA	Honduras	hspAfrica1SAfrica
SA157C	GCF_000476775	South Africa	hspAfrica1SAfrica

SA158C	GCF_000476695	South Africa	hspAfrica1SAfrica
SA226A	GCF_000475655	South Africa	hspAfrica1SAfrica
SA227C	GCF_000604405	South Africa	hspAfrica1SAfrica
SA35C	GCF_000604265	South Africa	hspAfrica1SAfrica
GAM100Ai	GCF_000310005	Gambia	hspAfrica1WAfrica
GAM115Ai	GCF_000345045	Gambia	hspAfrica1WAfrica
GAM121Aii	GCF_000345145	Gambia	hspAfrica1WAfrica
GAM260ASi	GCF_000345405	Gambia	hspAfrica1WAfrica
GAM264Ai	GCF_000345485	Gambia	hspAfrica1WAfrica
GAM71Ai	GCF_000345585	Gambia	hspAfrica1WAfrica
Gambia94-24	GCF_000185205	Gambia	hspAfrica1WAfrica
GAMchjs117Ai	GCF_000346835	Gambia	hspAfrica1WAfrica
Hp-H16	GCF_000275245	USA	hspAfrica1WAfrica
Hp-P4	GCF_000274845	USA	hspAfrica1WAfrica
1177	GCF_001940175	Singapore	hspEAsia
35A	GCF_000178935	Japan	hspEAsia
51	GCF_000011725	Korea	hspEAsia
D33	GCF_000709945	China	hspEAsia
F16	GCF_000270005	Japan	hspEAsia
Manado-1	GCF_000766035	Indonesia	hspEAsia
Mandalay38	https://datadryad.org/resource/doi:10.5061/dryad.8qp4n/2	Myanmar	hspEAsia
ML1	GCF_001549715	Taiwan	hspEAsia
Taiwan-47	GCF_000766025	Taiwan	hspEAsia
Yangon244	https://datadryad.org/resource/doi:10.5061/dryad.8qp4n/2	Myanmar	hspEAsia
aklavik117	GCF_000315955	Canada	hspIndigenousNAmerica
aklavik86	GCF_000317875	Canada	hspIndigenousNAmerica
ausabrJ05	GCF_001653435	Australia	hspIndigenousSAmerica
Cuz20	GCF_000148895	Peru	hspIndigenousSAmerica
PNG84A	GCF_001653475	Papa New Guinea	hspIndigenousSAmerica

Puno120	GCF_000224535	Peru	hspIndigenousSAmerica
Puno135	GCF_000224555	Peru	hspIndigenousSAmerica
Sat464	GCF_000148915	Peru	hspIndigenousSAmerica
Shi112	GCF_000277405	Peru	hspIndigenousSAmerica
Shi169	GCF_000277385	Peru	hspIndigenousSAmerica
Shi470	GCF_000020245	Peru	hspIndigenousSAmerica
v225d	GCF_000093185	Venezuela	hspIndigenousSAmerica
26695	GCF_000307795	UK	hspNEurope
H3014	GCF_000824865	UK	hspNEurope
HE-C1	HEL_AA2685AA	Sweden	hspNEurope
HE-C30	HEL_AA2667AA	Sweden	hspNEurope
HE-NC1-1	HEL_AA2935AA	Sweden	hspNEurope
HE-NC9-1	HEL_AA2915AA	Sweden	hspNEurope
HPAG1	GCF_000013245	Sweden	hspNEurope
Sahul64	GCF_000513515	Australia	hspNEurope
SSR1	GCF_003059845	Ireland	hspNEurope
SSR22	https://datadryad.org/resource/doi:10.5061/dryad.8qp4n/2	Ireland	hspNEurope
B128	GCF_000172955	USA	hspSEurope
B35	GCF_003059535	France	hspSEurope
Hp-A11	GCF_000359645	USA	hspSEurope
Hp-A14	GCF_000274005	USA	hspSEurope
Hp-A27	GCF_000274045	USA	hspSEurope
Iso6	GCF_002900195	Germany	hspSEurope
Iso8	GCF_002900175	Germany	hspSEurope
Lithuania75	GCF_000185225	Lithuania	hspSEurope
N6	GCF_000285895	France	hspSEurope
P12	GCF_000021465	Germany	hspSEurope
altai11	HEL_AA2519AA	Russia	hspSiberia1
altai59	HEL_AA2389AA	Russia	hspSiberia1
BURYAT19	HEL_AA2546AA	Russia	hspSiberia1
BURYAT27	HEL_AA3181AA	Russia	hspSiberia1

Chukchi08	HEL_AA3161AA	Russia	hspSiberia1
Evenky01	HEL_AA3037AA	Russia	hspSiberia1
Evenky65	HEL_AA3133AA	Russia	hspSiberia1
Evenky73	HEL_AA3164AA	Russia	hspSiberia1
Khanty27	HEL_AA3084AA	Russia	hspSiberia1
mong49	HEL_AA2533AA	Mongolia	hspSiberia1
DU15	GCF_001653395	South Korea	hspSiberia2
F32	GCF_000270045	Japan	hspSiberia2
Khanty47	HEL_AA3065AA	Russia	hspSiberia2
L107	HEL_AA3185AA	India	hspSiberia2
L12	HEL_AA3134AA	India	hspSiberia2
mong44	HEL_AA2391AA	Mongolia	hspSiberia2
Nanai30	HEL_AA3125AA	Russia	hspSiberia2
TUVAB15	HEL_AA2485AA	Russia	hspSiberia2
Tuvac46	HEL_AA3051AA	Russia	hspSiberia2
Tuvac80	HEL_AA3157AA	Russia	hspSiberia2
29CaP	GCF_001433495	Mexico	hspSWEurope
3699	GCF_003059245	France	hspSWEurope
655-99	GCF_000802575	Portugal	hspSWEurope
7C	GCF_001433515	Mexico	hspSWEurope
B25	GCF_003060065	France	hspSWEurope
B31	GCF_003060025	France	hspSWEurope
B43	GCF_003070765	France	hspSWEurope
B45	GCF_000234645	France	hspSWEurope
HUP-B14	GCF_000259235	Spain	hspSWEurope
SJM180	GCF_000148855	Peru	hspSWEuropeMexico

Table A3.1. *H. pylori* isolate genomes first sequenced and described in the present study

Isolate	Source	Size (Mb)	G-C%	Contigs	Coding Sequences (CDS)
1A2	Manitoba	1.63	39.2	48	1532
1A7	Manitoba	1.64	38.68	50	1572
1A9	Manitoba	1.64	38.81	41	1541
1B9	Manitoba	1.60	38.92	32	1482
1C1	Manitoba	1.65	38.77	57	1568
1C4	Manitoba	1.68	38.82	54	1578
1C7	Manitoba	1.57	39	93	1482
1C8	Manitoba	1.62	38.89	38	1511
1C9	Manitoba	1.72	38.76	72	1643
1E9	Manitoba	1.57	39.01	28	1476
1F3	Manitoba	1.66	38.79	33	1570
1F6	Manitoba	1.68	38.75	58	1566
1G5	Manitoba	1.68	38.71	54	1560
1G8	Manitoba	1.57	38.99	44	1461
1H3	Manitoba	1.68	38.73	34	1602
1H7	Manitoba	1.65	38.83	40	1555
1H9	Manitoba	1.58	39	36	1494
1I5	Manitoba	1.62	38.87	43	1524
1I7	Manitoba	1.67	38.78	93	1600
1I8	Manitoba	1.64	38.85	99	1545
2A2	Manitoba	1.65	38.88	49	1539
2A5	Manitoba	1.59	38.95	39	1492
2A6	Manitoba	1.58	39.01	31	1491
2B1	Manitoba	1.66	38.79	38	1563
2B2	Manitoba	1.67	38.74	36	1577
2B4	Manitoba	1.62	38.85	41	1509
2B5	Manitoba	1.60	38.92	59	1497
2C2	Manitoba	1.66	38.82	52	1579

2C5	Manitoba	1.64	38.92	144	1515
2E9	Manitoba	1.56	39.09	83	1456
2F9	Manitoba	1.62	38.93	47	1514
2G2	Manitoba	1.58	39.01	49	1473
2G3	Manitoba	1.62	38.87	43	1517
2H2	Manitoba	1.65	38.93	61	1562
2H4	Manitoba	1.59	38.97	17	1490
2H5	Manitoba	1.68	38.7	45	1594
2H7	Manitoba	1.60	38.99	47	1516
2I4	Manitoba	1.60	38.96	37	1515
2I5	Manitoba	1.67	38.74	32	1573
2I6	Manitoba	1.67	38.72	35	1574
2I7	Manitoba	1.62	38.89	47	1517
2I9	Manitoba	1.56	39.02	69	1467
3A2	Manitoba	1.59	38.94	35	1510
3A5	Manitoba	1.68	38.68	47	1589
3A6	Manitoba	1.60	38.94	52	1526
3A7	Manitoba	1.67	38.75	60	1563
3B2	Manitoba	1.64	38.82	124	1543
3B4	Manitoba	1.67	38.79	94	1582
3B8	Manitoba	1.61	39.04	55	1506
3C4	Manitoba	1.58	39.06	41	1472
3C7	Manitoba	1.61	38.82	29	1512
3C9	Manitoba	1.62	39.24	42	1518
3D2	Manitoba	1.69	38.78	59	1565
3D4	Manitoba	1.69	38.7	57	1587
3D9	Manitoba	1.56	39.02	25	1459
3E7	Manitoba	1.72	38.74	85	1601
3F2	Manitoba	1.57	39.05	93	1464
3F4	Manitoba	1.60	38.96	42	1523
3F6	Manitoba	1.58	39.02	48	1517

3F9	Manitoba	1.65	38.9	55	1536
3G2	Manitoba	1.62	38.94	51	1511
3H9	Manitoba	1.65	38.95	35	1532
3I3	Manitoba	1.60	39.27	74	1516
3I7	Manitoba	1.62	38.9	89	1510
4A1	Manitoba	1.57	39.01	35	1472
4A5	Manitoba	1.69	38.71	55	1587
4A6	Manitoba	1.51	39.21	43	1415
4B5	Manitoba	1.59	39.07	190	1498
4B6	Manitoba	1.67	38.9	36	1548
4B7	Manitoba	1.66	38.85	217	1554
4C8	Manitoba	1.56	39.03	44	1461
4D4	Manitoba	1.66	38.71	31	1571
4D5	Manitoba	1.65	38.85	33	1541
CAN-002	Old Crow	1.68	39.09	1*	1549
CAN-003	Aklavik	1.57	39.09	1*	1464
CAN-004	Aklavik	1.64	39.19	54	1525
CAN-005	Aklavik	1.51	39.14	39	1416
CAN-007	Aklavik	1.55	39.13	1*	1451
CAN-010	Aklavik	1.50	39.25	1*	1406
CAN-011	Aklavik	1.52	39.12	40	1432
CAN-012	Old Crow	1.71	38.88	1*	1588
CAN-019	Aklavik	1.50	39.18	58	1410
CAN-024	Ross River	1.58	38.81	12	1478
CAN-048	Aklavik	1.56	38.85	10	1466
CAN-055	Aklavik	1.63	39.18	49	1519
CAN-060	Old Crow	1.68	38.73	37	1583
CAN-061	Teslin	1.63	38.89	28	1531
CAN-066	Aklavik	1.64	39.18	39	1523
CAN-068	Inuvik	1.63	39.23	34	1524
CAN-074	Aklavik	1.60	39.05	42	1480

CAN-087	Old Crow	1.63	38.94	43	1525
CAN-090	Fort McPherson	1.66	38.98	165	1546
CAN-095	Aklavik	1.52	39.13	28	1431
CAN-108	Inuvik	1.68	39.05	23	1561
CAN-111	Teslin	1.51	38.96	21	1418
CAN-112	Fort McPherson	1.48	39.22	26	1386
CAN-120	Aklavik	1.66	39.08	37	1533
CAN-133	Fort McPherson	1.63	38.92	39	1525
CAN-136	Aklavik	1.51	39.13	34	1427
CAN-138	Aklavik	1.67	39.08	39	1541
CAN-139	Aklavik	1.60	39.02	37	1506
CAN-148	Aklavik	1.60	39.02	41	1503
CAN-157	Aklavik	1.67	38.74	39	1577
CAN-167	Fort McPherson	1.51	39.16	35	1432
CAN-170	Aklavik	1.67	39.09	36	1538
CAN-172	Aklavik	1.52	39.14	31	1433
CAN-181	Aklavik	1.56	38.85	16	1455
CAN-183	Inuvik	1.68	39.06	28	1554
CAN-196	Aklavik	1.52	39.01	17	1426
CAN-197	Inuvik	1.56	39.01	18	1455
CAN-208	Old Crow	1.67	38.73	36	1580
CAN-212	Aklavik	1.63	38.96	43	1525
CAN-216	Aklavik	1.67	38.74	36	1580
CAN-219	Aklavik	1.70	38.82	120	1590
CAN-221	Ross River	1.66	38.81	36	1562
CAN-242	Ross River	1.64	38.9	20	1518
CAN-243	Aklavik	1.52	39.04	72	1417
CAN-244	Aklavik	1.63	39.19	34	1523
CAN-245	Teslin	1.55	38.94	24	1449
CAN-247	Aklavik	1.50	39.21	37	1416
CAN-249	Teslin	1.56	38.89	23	1461

CAN-251	Fort McPherson	1.53	39.09	33	1430
CAN-252	Ross River	1.65	38.79	21	1545
CAN-258	Aklavik	1.51	39.13	30	1415
CAN-259	Aklavik	1.50	39.18	27	1408
CAN-261	Old Crow	1.63	38.95	40	1519
CAN-264	Fort McPherson	1.63	38.92	41	1527
CAN-271	Aklavik	1.67	38.74	39	1580
CAN-274	Aklavik	1.67	38.74	40	1581
CAN-290	Aklavik	1.50	39.18	36	1411

*all genomes with a contig number = 1 were sequenced using PacBio long read technology

Table A3.2 Multi-locus sequence typing for Manitoba and Arctic *H. pylori* isolates

Isolate	Sequence Type	*atpA*	*efp*	*mutY*	*ppa*	*trpC*	*ureI*	*yphC*
1A2	NA	531	1981	202	2217	2825	2639	345
1A7	NA	2680	1787	697	2545	1540	2812	2843
1A9	NA	432	1167	1088	1958	1420	1601	511
1B9	NA	176	2469	696	163	111	2563	1325
1C1	NA	1243	1632	1739	1181	666	1089	1313
1C4	NA	1629	619	1267	2045	1313	1700	699
1C7	NA	1672	1580	2560	2069	1083	1444	620
1C8	NA	1531	2245	2560	2498	236	823	1186
1C9	NA	2687	2265	2672	2733	740	679	1828
1E9	NA	1022	2267	111	998	1288	1767	518
1F3	NA	883	984	615	442	1741	750	149
1F6	NA	1619	615	1591	1053	1709	2663	937
1G5	NA	1052	1617	1620	1925	2855	166	747
1G8	NA	1661	2323	1269	2580	1640	857	226
1H3	NA	2002	1221	2717	544	2691	929	702
1H7	NA	1629	449	540	1523	116	2731	2697
1H9	NA	1412	1017	2160	783	2691	2594	898
1I5	NA	645	757	2628	2570	684	1662	350
1I7	NA	1391	663	562	2211	1185	2719	663
1I8	NA	2846	2658	2593	582	598	1589	2059
2A2	NA	112	247	713	244	632	2838	2909
2A5	NA	1661	1589	1530	2427	236	106	1081
2A6	NA	2859	112	2508	2381	540	671	2964
2B1	NA	1258	1023	876	2107	303	2594	1669
2B2	NA	1301	2302	1297	1006	1999	633	1819
2B4	NA	2183	449	2865	313	805	4	1307
2B5	NA	1742	2502	825	2444	1058	29	2609
2C1	NA	29	29	29	31	29	29	230
2C5	NA	1052	1617	1620	1925	2855	166	747
2E9	NA	931	2323	2561	2484	635	2759	1731
2F9	NA	1632	526	1406	594	2895	92	837

2G2	NA	1375	649	2911	444	853	2594	2393
2G3	NA	240	1132	489	594	2842	92	2138
2H2	NA	2855	1019	550	625	681	2946	658
2H4	NA	541	2663	2757	1600	1999	103	1307
2H5	NA	2471	1388	1122	795	2947	114	654
2H7	NA	848	181	885	785	1289	155	2629
2I4	NA	2771	615	105	723	273	943	1360
2I5	NA	1301	2302	1297	1006	1999	633	1819
2I6	NA	1619	615	1591	1053	1709	2663	119
2I7	NA	1718	552	2843	263	2172	2952	2941
2I9	NA	1661	1589	1530	2427	236	2838	1081
3A2	NA	63	947	1623	2493	941	994	415
3A5	NA	332	989	1503	104	834	561	844
3A6	NA	1672	2481	1771	594	576	2571	2826
3A7	NA	1301	1533	2516	815	1999	633	1819
3B2	NA	2188	712	683	2750	116	623	824
3B4	NA	1391	663	562	2211	1185	2719	663
3B8	NA	1703	2461	1743	2650	1802	1101	1550
3C4	NA	2843	434	575	2438	939	669	115
3C7	NA	1306	2154	1390	109	913	723	2748
3C9	NA	531	1483	203	2285	2393	2915	2866
3D2	NA	1385	1597	2827	1694	682	2663	2945
3D4	NA	1178	615	563	1053	450	2838	1334
3D9	NA	884	588	2490	1602	1138	1745	1299
3E7	NA	200	1633	929	2504	89	1131	1020
3F2	NA	2479	2527	668	743	2228	1329	690
3F4	NA	2592	1529	549	783	805	2546	1149
3F6	NA	2813	2369	581	548	1714	1233	841
3F9	NA	613	598	736	113	2954	2969	673
3G2	NA	255	601	642	2477	2642	2974	590
3H9	NA	2459	1624	1608	1677	2406	1889	2224
3I3	NA	531	1981	202	2217	2825	2639	345
3I7	NA	174	174	2456	2364	2355	2795	2409
4A1	NA	1022	2267	111	998	1288	1767	518
4A5	NA	2697	1232	884	2525	849	2722	2870

4A6	NA	2749	1442	97	1067	1525	1212	2881
4B5	NA	2771	115	2672	2580	1288	2593	1074
4B6	NA	2471	1510	894	999	1288	817	1334
4B7	NA	1398	2111	1531	1694	682	2663	1089
4C8	NA	2560	567	2624	1763	1771	1701	1304
4D4	NA	1385	510	1531	1053	1316	2663	2568
4D5	NA	804	1085	112	1053	1667	946	2629
aklavik117	2935	2583	1824	2536	179	2604	2628	2640
aklavik86	2945	2528	245	2539	2414	2602	2616	2628
CAN-002	NA	2506	2373	2661	237	2584	2587	2606
CAN-003	2934	2517	2384	213	1980	213	2605	2618
CAN-004	2925	2506	2373	2522	2396	2584	2594	2606
CAN-005	2929	2514	2380	2530	2405	2593	2599	2614
CAN-007	NA	239	238	2570	238	238	238	1394
CAN-010	2945	2528	245	2539	2414	2602	2616	2628
CAN-011	2929	2514	2380	2530	2405	2593	2599	2614
CAN-012	2994	2574	1719	2582	1735	2646	2665	2677
CAN-019	2945	2528	245	2539	2414	2602	2616	2628
CAN-024	NA	2751	231	2584	1067	2590	233	2032
CAN-048	2928	2511	2379	2527	2402	2590	2598	2612
CAN-055	2925	2506	2373	2522	2396	2584	2594	2606
CAN-060	NA	2066	2382	17	2407	2595	1837	591
CAN-061	NA	138	92	432	2459	2617	2671	2681
CAN-066	2925	2506	2373	2522	2396	2584	2594	2606
CAN-068	NA	2506	2373	2522	2426	2584	1812	2606
CAN-074	NA	560	526	94	2049	1459	2561	2561
CAN-087	NA	2571	94	2578	90	2642	2661	2657
CAN-090	NA	2571	94	2578	90	2642	2661	2657
CAN-095	2929	2514	2380	2530	2405	2593	2599	2614
CAN-108	NA	2506	2373	602	237	2584	406	2606
CAN-111	NA	2562	1879	1284	1209	2645	2851	2673
CAN-112	NA	2514	2380	2539	1109	2652	2602	2628
CAN-120	2960	2542	2373	2551	2426	2584	2631	2606
CAN-133	2992	2571	94	2578	90	2642	2663	2675
CAN-136	2929	2514	2380	2530	2405	2593	2599	2614

CAN-138	2983	2542	2373	2570	2426	2584	2631	2606
CAN-139	2932	1152	2382	2532	2407	2595	2594	2616
CAN-148	2932	1152	2382	2532	2407	2595	2594	2616
CAN-157	NA	2066	2382	17	2407	2595	1837	591
CAN-167	NA	2514	2380	2539	2056	2844	1082	2687
CAN-170	2983	2542	2373	2570	2426	2584	2631	2606
CAN-172	2929	2514	2380	2530	2405	2593	2599	2614
CAN-181	2933	2511	2379	2527	2408	2590	2598	2612
CAN-183	NA	2506	2373	2661	2426	2584	1693	2606
CAN-196	2936	2519	2379	2527	2402	2590	2598	2612
CAN-197	NA	233	1750	1221	1831	2590	2599	2678
CAN-208	NA	2066	2382	17	2407	2595	1837	591
CAN-212	NA	2571	94	2578	90	2642	2661	2657
CAN-216	NA	2066	2382	17	2407	2595	1837	591
CAN-219	NA	2583	2373	1059	179	2584	2628	2640
CAN-221	NA	138	92	575	866	2617	2890	2681
CAN-242	NA	632	2426	2579	2625	2643	2840	2679
CAN-243	NA	2585	2413	239	2439	2626	2646	2657
CAN-244	2925	2506	2373	2522	2396	2584	2594	2606
CAN-245	NA	2753	2516	724	179	2590	748	2675
CAN-247	2945	2528	245	2539	2414	2602	2616	2628
CAN-249	NA	2748	522	2530	1209	250	2851	2879
CAN-251	NA	2514	2380	2539	2414	2602	231	2687
CAN-252	NA	2750	2441	575	1045	1114	624	3021
CAN-258	2929	2514	2380	2530	2405	2593	2599	2614
CAN-259	2945	2528	245	2539	2414	2602	2616	2628
CAN-261	NA	2571	94	2578	90	2642	2661	2657
CAN-264	2992	2571	94	2578	90	2642	2663	2675
CAN-271	NA	2066	2382	17	2407	2595	1837	591
CAN-274	NA	2066	2382	17	2407	2595	1837	591
CAN-290	2945	2528	245	2539	2414	2602	2616	2628

MLST type determined by Torsten Seeman's MLST tool v.2.19.0 using the PubMLST Bacterial Isolate Genome Sequence database using the *H. pylori* MLST schema (found at https://pubmlst.org/helicobacter/ and accessed on January 2nd 2021)

Figure A3.1 Co-ancestry matrix of Manitoban, Arctic, and global *H. pylori* core genomes ordered by phylogenetic signal. Based on fineSTRUCTURE analysis, the colour assigned for the CAM heatmap indicates the degree of overall sequence content similarity between two genomes. Isolates labelled yellow share slight similarity (<10%), isolates labelled blue share near total similarity (>90%), and all other intervening colour assignments as depicted (at right) for 11-89% sequence similarity. The diagonal represents an isolate paired with itself (self-self). Interpretation for the colour of each cell in the matrix: indicates evidence of genetic flux, as the expected number of chunks imported from a donor genome (column) into a recipient genome (row). fineSTRUCTURE analysis was run using 769 core genes present in >90% of Manitoban, Arctic, and world isolates.

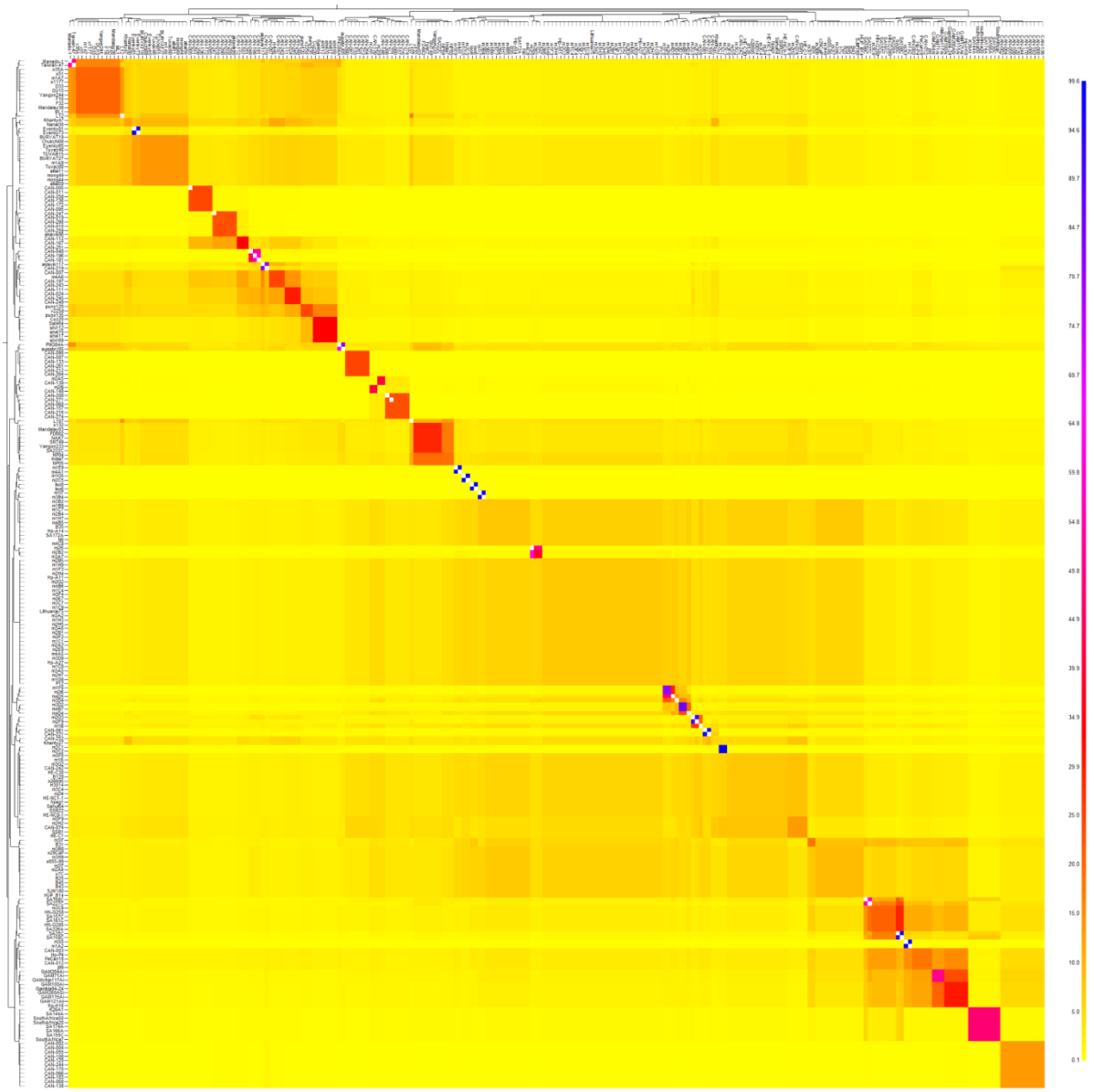


Table A3.3 fineSTRUCTURE-determined *H. pylori* sequence types

Isolate	Sequence Type
1A2	hpAfrica1
1A7	hspEAsia
1A9	hspSiberia
1B9	hspSEurope
1C1	hspSEurope
1C4	hspSEurope
1C7	hspSEurope
1C8	hspSEurope
1C9	hspSEurope
1E9	hspSEurope
1F3	hspSEurope
1F6	hspNEurope
1G5	hspSEurope
1G8	hspSEurope
1H3	hspSEurope
1H7	hspSEurope
1H9	hspSEurope
1I5	hspNEurope
1I7	hspSEurope
1I8	hspNEurope
2A2	hspSEurope
2A5	hspIndigenousSAmerica
2A6	hspSWEurope
2B1	hspSEurope
2B2	hspSEurope
2B4	hspSEurope
2B5	hspSEurope
2C2	hspNEurope
2C5	hspSEurope

2E9	hspSEurope
2F9	hspNEurope
2G2	hspSEurope
2G3	hspNEurope
2H2	hspSWEurope
2H4	hspSEurope
2H5	hspSEurope
2H7	hspSEurope
2I4	hspNEurope
2I5	hspSEurope
2I6	hspNEurope
2I7	hspSWEurope
2I9	hspIndigenousSAmerica
3A2	hspSEurope
3A5	hspSEurope
3A6	hspSEurope
3A7	hspSEurope
3B2	hspSEurope
3B4	hspSEurope
3B8	hspSWEurope
3C4	hspNEurope
3C7	hspSEurope
3C9	hpAfrica1
3D2	hspNEurope
3D4	hspNEurope
3D9	hspSEurope
3E7	hspSEurope
3F2	hspSEurope
3F4	hspSEurope
3F6	hspNEurope
3F9	hspSWEurope

3G2	hspNEurope
3H9	hspSWEurope
3I3	hpAfrica1
3I7	hspSWEurope
4A1	hspSEurope
4A5	hspSEurope
4A6	hspIndigenousNAmerica
4B5	hspSEurope
4B6	hspSEurope
4B7	hspNEurope
4C8	hspSEurope
4D4	hspNEurope
4D5	hspNEurope
CAN-002	hpAfrica2
CAN-003	hpAfrica1
CAN-004	hpAfrica2
CAN-005	hspIndigenousNAmerica
CAN-007	hspIndigenousNAmerica
CAN-010	hspIndigenousNAmerica
CAN-011	hspIndigenousNAmerica
CAN-012	hpAfrica1
CAN-019	hspIndigenousNAmerica
CAN-024	hspIndigenousNAmerica
CAN-048	hspIndigenousNAmerica
CAN-055	hpAfrica2
CAN-060	hspIndigenousSAmerica
CAN-061	hspNEurope
CAN-066	hpAfrica2
CAN-068	hpAfrica2
CAN-074	hspSWEurope
CAN-087	hspIndigenousSAmerica

CAN-090	hspIndigenousSAmerica
CAN-095	hspIndigenousNAmerica
CAN-108	hpAfrica2
CAN-111	hspIndigenousNAmerica
CAN-112	hspIndigenousNAmerica
CAN-120	hpAfrica2
CAN-133	hspIndigenousSAmerica
CAN-136	hspIndigenousNAmerica
CAN-138	hpAfrica2
CAN-139	hspIndigenousSAmerica
CAN-148	hspIndigenousSAmerica
CAN-157	hspIndigenousSAmerica
CAN-167	hspIndigenousNAmerica
CAN-170	hpAfrica2
CAN-172	hspIndigenousNAmerica
CAN-181	hspIndigenousNAmerica
CAN-183	hpAfrica2
CAN-196	hspIndigenousNAmerica
CAN-197	hspIndigenousNAmerica
CAN-208	hspIndigenousSAmerica
CAN-212	hspIndigenousSAmerica
CAN-216	hspIndigenousSAmerica
CAN-219	hspIndigenousNAmerica
CAN-221	hspNEurope
CAN-242	hspNEurope
CAN-243	hspIndigenousNAmerica
CAN-244	hpAfrica2
CAN-245	hspIndigenousNAmerica
CAN-247	hspIndigenousNAmerica
CAN-249	hspIndigenousNAmerica
CAN-251	hspIndigenousNAmerica

CAN-252	hspNEurope
CAN-258	hspIndigenousNAmerica
CAN-259	hspIndigenousNAmerica
CAN-261	hspIndigenousSAmerica
CAN-264	hspIndigenousSAmerica
CAN-271	hspIndigenousSAmerica
CAN-274	hspIndigenousSAmerica
CAN-290	hspIndigenousNAmerica

Sequence typing based on a core genome of >90% as determined by Roary v3.12.0 analyzed by BEAGLE v.3.3.2, fineSTRUCTURE v.0.02, and CHROMOPAINTER v.0.02. Ten representative donor genomes from each sequence type were used as input as shown in **Table A2.1**. All Manitoban and CAN*Help* isolates were designated as genome recipients.

Table A3.4 Presence of phenotypic and genotypic antimicrobial resistance markers in Manitoban and Arctic *H. pylori* isolates

	Manitoba (n=73)		Arctic (n=57)		
Resistance	**Count**	**%**	**Count**	**% Genomics**	**% Lab***
Metronidazole	34	46.6	17	29.8	36.1
Clarithromycin	27	37.0	12	21.1	16.2
Amoxicillin	0	0.0	2	3.5	0.0
Levofloxacin	16	21.9	2	3.5	4.2
Only Metronidazole	20	27.4	9	15.8	NA
Only Clarithromycin	9	12.3	7	12.3	NA
Only Amoxicillin	0	0.0	0	0.0	NA
Only Levofloxacin	2	2.7	1	1.8	NA
Metronidazole + Amoxicillin	0	0.0	2	3.5	NA
Metronidazole + Levofloxacin	3	4.1	1	1.8	NA
Metronidazole + Clarithromycin	7	9.6	5	8.8	NA
Levofloxacin + Clarithromycin	7	9.6	0	0.0	NA
Metronidazole + Levofloxacin + Clarithromycin	4	5.5	0	0.0	NA
Susceptible	21	28.8	32	56.1	55.0
Resistant to 1	31	42.5	17	29.8	32.0
Resistant to 2	17	23.3	8	14.0	10.0
Resistant to 3	4	5.5	0	0.0	2.0
Resistant to 1 or more	52	71.2	25	43.9	45.0

MetR = Metronidazole resistance marker, ClariR = Clarithromycin resistance marker, AmoxR = Amoxicillin resistance marker, LevoR = Levofloxacin resistance marker, *data from Williams (2016), **data is for Ciprofloxacin resistance. % Genomics are resistance predictions inferred from the genomic sequence. % Lab are phenotypes determined by E test (Cheung et al. 2009).

Table A3.5 *cagA* presence/absence, EPIYA motif type, and *cag*PI presence/absence

Status	Manitoba (n=73)		Arctic (n=57)	
	Count	%	Count	%
cagA+	**47**	**61.0**	**33**	**55.9**
***cag*PI+**	**53**	**68.8**	**33**	**55.9**
***cagA- / cag*PI-**	**24**	**31.2**	**25**	**42.4**
***cagA+ / cag*PI+**	**47**	**61.0**	**32**	**54.2**
***cagA+ / cag*PI-**	0	0.0	**1**	**1.7**
***cagA- / cag*PI+**	**6**	**7.8**	**1**	**1.7**
EPIYA AB	**1**	**1.3**	**1**	**1.7**
EPIYA ABC	**34**	**44.2**	**22**	**37.3**
EPIYA ABCC	**4**	**5.2**	**2**	**3.4**
EPIYA ABCD	**1**	**1.3**	0	0.00
EPIYA AC	0	0.0	**2**	**3.4**
EPIYA BC	0	0.0	**4**	**6.8**
EPIYA C	**2**	**2.6**	0	0.0
EPIYA CD	**4**	**5.2**	0	0.0
Truncated/No EPIYA	**1**	**1.3**	**2**	**3.4**

Table A3.6 *vacA* allele and genotype status by both *in silico* and lab-based PCR

	Manitoba *in silico* (n=73)		Arctic *in silico* (n=57)		Arctic PCR (n=57)	
	Count	**%**	**Count**	**%**	**Count**	**%**
Allele Presence						
s-	**20**	**27.4**	**15**	**26.3**	**42**	**73.7**
s1a	**40**	**54.8**	**18**	**31.6**	**15**	**26.3**
s1b	**4**	**5.5**	**14**	**24.6**	0	0.0
s1c	**5**	**6.8**	**10**	**17.5**	0	0.0
s2	**6**	**8.2**	0	0.0	0	0.0
m-	**10**	**13.7**	**19**	**33.3**	**14**	**24.6**
m1	**20**	**27.4**	**28**	**49.1**	**13**	**22.8**
m2	**42**	**57.5**	**10**	**17.5**	**27**	**47.4**
m1&m2	0	0.0	0	0.0	**3**	**5.3**
Genotype Presence						
s-/m-	**5**	**6.8**	**14**	**24.6**	**10**	**17.5**
s1a/m-	**3**	**4.1**	**1**	**1.8**	**4**	**7.0**
s1b/m-	**2**	**2.7**	0	0.0	0	0.0
s1c/m-	0	0.0	**4**	**7.0**	0	0.0
s2/m-	**2**	**2.7**	0	0.0	0	0.0
s-/m1	**4**	**5.5**	**1**	**1.8**	**8**	**14.0**
s-/m2	**10**	**13.7**	0	0.0	**22**	**38.6**
s-/m1&m2	0	0.0	0	0.0	**2**	**3.5**
s1a/m1	**14**	**19.2**	**13**	**22.8**	**5**	**8.8**
s1a/m2	**23**	**31.5**	**4**	**7.0**	**5**	**8.8**
s1a/m1&m2	0	0.0	0	0.0	**1**	**1.8**
s1b/m1	**2**	**2.7**	**14**	**24.6**	0	0.0
s1c/m2	**3**	**4.1**	**6**	**10.5**	0	0.0
s1c&s2/m2	**2**	**2.7**	0	0.0	0	0.0
s2/m2	**3**	**4.1**	0	0.0	0	0.0

Table A3.7 Genomic Island data dictionary for *H. pylori* genomes in the present study

Genomic Island	1	2	3	4
LptB	Present	+HomA		
SpaB	Present	+ABC Transporter +BabA		
TetR	+BabA	+Addiction module toxin	+HopJ +HomA +SabB +Addiction module toxin	
TerY	Present	+lysozyme	+lysozyme +HopK	
BabC/HopU	Present	+HopK		
HomD	Present	+BabC +HopU		
FrpB	Present	+HopK		
HofC/HofD	Present	+FrpB	+FrpB +HopH	
HopJ	Present	+HopZ	+BabA	+HomA +SabA
LeoA	Present	+LptB		
Beta-lactamase	Present	+lysozyme	+SpaB +SabA	
BabA	Present		+SabA +HopJ	

The presence of genomic islands in the genomes was predicted using IslandViewer 4. "Present" denotes that a given genomic island was present in entirety, without sequence/gene truncation. Further annotation, *e.g*. +TerY, indicates that a given genomic island was present and carried extra genes as indicated by annotation.