Genomics of Ancient Pathogenic Bacteria

Genomics of Ancient Pathogenic Bacteria: Novel Techniques & Extraordinary Substrates

By ALISON M. DEVAULT, B.A., M.A.

A Thesis Submitted to the School of Graduate Studies
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

McMaster University © Copyright by Alison M. Devault, September 2014

DOCTOR OF PHILOSOPHY (2014) McMaster University
(Anthropology) Hamilton, Ontario

TITLE: Genomics of Ancient Pathogenic Bacteria: Novel Techniques & Extraordinary Substrates

AUTHOR: Alison M. Devault, B.A., M.A. (McMaster University)

SUPERVISOR: Dr. Hendrik N. Poinar

NUMBER OF PAGES: xiv, 194

Abstract

Palaeogenetic research on human pathogenic and microbiomic bacteria has been largely restricted to bloodborne pathogens from skeletal tissue and, due to short lengths of degraded ancient DNA, small-scale single loci studies. My thesis has expanded the breadth and depth of palaeomicrobial knowledge via the study of novel specimen types with next-generation technologies. Presented in sandwich thesis format, I discuss genome-scale studies of three previously-unstudied historical pathogens: 19th century *Vibrio cholerae* (cholera) from an alcohol-preserved intestine from Philadelphia, and medieval *Staphylococcus saprophyticus* (urinary tract infections) and *Gardnerella vaginalis* (bacterial vaginosis) from calcified urogenital infections of a Trojan woman. Cholera persists as a dangerous modern disease that was also responsible for severe historic epidemics. My research confirms that 19th century pandemics were caused by an O1 classical strain that may have possessed genomic features that contributed increased virulence. *S. saprophyticus* and *G. vaginalis* are opportunistic pathogens of the urogenital microbiome, especially in reproductive-age females. Using very high endogenous DNA content of the calcified infections, I have reconstructed one of the most complete ancient bacterial genomes for *S. saprophyticus* and coding genome for *G. vaginalis.* Both ancient pathogens possess most of the virulence and urogenital adaptive genes of modern strains, indicating similar ecological roles for these species in past female health. Finally, I successfully use LLMDA microarray technology (never before utilized for ancient DNA research) to detect ancient pathogens. LLMDA provides an inexpensive and informative alternative to high-throughput sequencing for assessing the metagenomic content of ancient samples. Together, my findings provide a framework emphasizing the need to broadly study past microbiomes in conjunction with specific pathogens. Using molecular data, this work supports anthropological views of infectious disease ecology related to the first epidemiological transition and historical narratives. Taken together with the recent literature on ancient pathogen genomes, my findings indicate that palaeogenome sequences may not necessarily reveal any specific signatures of greater virulence, and interpretations of past diseases must necessarily take into account additional host, environmental, and cultural factors.

Acknowledgements

The last 6 (!!!) years has given me many ups and downs, but I am so thankful for the friends, family, and colleagues who encouraged me along the way. You guys saw the light at the end of the tunnel even when I did not. There are so many people I wish to thank, and to the probably many more who I left out, I hope you know how truly grateful I am for your support.

I would firstly like to thank the Province of Ontario, the Department of Anthropology, and Hendrik Poinar for financial support of both myself, and my thesis research.

To Mom, Dad, and Shelle, I absolutely would not be here without your enthusiastic support, both financially and emotionally. Thank you from the bottom of my heart!!!

I gratefully acknowledge the insightful wisdom and extreme patience and positivity of my stellar thesis committee, Ann, David, and Hendrik, who have patiently accommodated both my less-than-fast pace and the shifting landscape of this research over the years.

To Hendrik, you have been a truly excellent supervisor, and you have always showed faith in me, even when I felt really unsure if I deserved it. Your encouraging attitude continues to inspire me, and as a result I'm doing research that I did not even know was possible for me. Even when I felt overwhelmed, your vision of the final products helped see it through. Thank you for the countless opportunities, and for your patience and understanding.

To my fellow anthropology department members, thank you for such a welcoming community. To my students (especially 2D03, 3R03, & 4GS3), thank you for your continued enthusiasm for ancient DNA, which always made me appreciate how lucky I was!

To my co-authors, thank you for your endless patience, support, and contributions. I would especially like to thank the "cholera paper" folks, who had been waiting patiently for me to analyze & write up the data since way back in 2008 (!) when we first collected the specimens. Your unwavering enthusiasm for our project has been a true blessing. I would also especially like to thank the LLNL team who likewise patiently waited for two years, and slogged through many paper drafts and revisions, without complaint. And to my new "Troy paper" colleagues, thank you for your patience while I busted out this thesis; I'm so excited for the road ahead!

To my current/former Ancient DNA Centre members, thank you for your endless insights, support, and help along the way. It feels like a lifetime ago when I first arrived, and your encouragement and guidance helped to keep me going and teach me how things work in

the wacky world of ancient DNA (special thanks to Regis, John, Christine, & Kirsti). To the recent additions (who I wish were here the whole time), I thank Nat & Jen for bringing their fresh cheer, especially at a time when I needed it the most! And for her guidance since I arrived, I would especially like to thank Mel, who always brings the right combination of positivity and rationality!

Thank you from the bottom of my heart to Jake, for providing indescribable help and support over the last 5 years (in all categories – I literally could not have done this without you), for putting up with me, and for helping me grow.

And last but not least, to Oscar, who has been patiently waiting for me to finish writing since comps (but never judges!): thank you for always giving me something to hold on to, even when I doubted myself.

Table of Contents

Chapter 1 - Introduction **1**

1.1 Background 1

1.2 The physically invisible: Ancient extraskeletal bacteria 2

1.2.1 Extraskeletal infectious diseases 3

1.2.2 The extraskeletal microbiome 4

1.3 The epistemologically invisible: (Meta)genomic bacterial content 6

1.3.1 Genome-scale studies 6

1.3.2 Targeted approaches are biased 7

1.3.3 Metagenomic approaches 9

1.4 Goals of this thesis 10

Chapter 2 **13**

2.1 Thesis Preface 13

2.2 Abstract 14

2.3 Glossary 14

2.4 Introduction 16

2.5 Methods 17

2.6 Results 17

2.6.1 Historical *V. cholerae* genome 17

2.6.2 Genomic islands and virulence factors 18

2.6.3 Human mitochondrial and nuclear DNA 18

2.6.4 Origin and evolution of *V. cholerae* 19

2.7 Discussion 19

2.8 Acknowledgements 20

2.9 Figures 22

2.10 Supplementary Appendix 25

2.11 Supplementary Methods 25

2.11.1 Cholera victim specimen 25

2.11.2 Sampling 25

2.11.3 DNA extraction 26

2.11.4 Library preparation and indexing 26

2.11.5 *V. cholerae* DNA qPCR assessment 27

2.11.6 Array design 28

2.11.7 Array enrichment 28

2.11.8 Sequencing 29

2.11.9 Shotgun sequencing 30

2.11.10 Authorship contributions 30
2.12 Supplementary Analysis 30
2.12.1 Draft genome sequence 30
2.12.2 Other loci included on the array 34
2.12.3 Other analyses 34
2.13 Supplementary Results 35
2.13.1 PA1849 vs. O395 35
2.13.2 PA1849 vs. other strains 36
2.13.3 Human loci 36
2.14 Supplementary Figures 37
2.15 Supplementary Tables 53
2.16 References 68

Chapter 3 **71**
3.1 Thesis Preface 71
3.2 Abstract 72
3.3 Introduction 72
3.4 Methods 73
3.5 Results & Discussion 74
3.5.1 Burial 177, grave 14, quadrat x24, Troy 74
3.5.2 Snapshot of an ancient urogenital infection 75
3.5.3 *Staphylococcus saprophyticus* Troy 77
3.5.4 *Gardnerella vaginalis* Troy 80
3.6 Conclusion 82
3.7 Acknowledgements 83
3.8 Figures 85
3.9 Supplementary Appendix 89
3.10 Supplementary Methods 89
3.10.1 Samples and DNA extraction 89
3.10.2 Library preparation and indexing 91
3.10.3 Enrichment 91
3.10.4 Sequencing and read preparation 92
3.10.5 Modern *S. saprophyticus* isolates 92
3.10.6 *Staphylococcus saprophyticus* Troy 93
3.10.7 *Gardnerella vaginalis* Troy 96
3.10.8 Phylogenetic analysis 97
3.10.9 Human mitochondrial genome 98
3.10.10 Metagenomic analysis 99

3.10.11 Microbiome analysis 99
3.11 Supplementary Tables 100
3.12 Supplementary Figures 114
3.13 References 132

Chapter 4 **137**
4.1 Thesis Preface 137
4.2 Abstract 139
4.3 Introduction 140
4.4 Results 141
4.4.1 Taxa detected by both LLMDA & HTS 142
4.4.2 Taxa detected by only one method 143
4.5 Discussion 143
4.6 Methods 146
4.7 Acknowledgements 147
4.7.1 Author contributions 147
4.7.2 Additional information 147
4.8 Tables 148
4.9 Figures 150
4.10 Supplementary Appendix 155
4.11 Supplementary Methods 155
4.11.1 Sample preparation 155
4.11.2 Shotgun HTS sequencing 155
4.11.3 Pathogen HTS assemblies 155
4.11.4 HTS BLAST & MEGAN metagenomic analysis 156
4.11.5 LLMDA analysis 156
4.12 Supplementary Tables 159
4.13 Supplementary Figures 167
4.14 References 174

Chapter 5 – Conclusion **176**
5.1 Contributions of this thesis 176
5.2 Human-pathogen interactions 176
5.2.1 Diseases of the first epidemiological transition 178
5.2.2 Co-infections 181
5.3 Reproductive and female health 183
5.4 Open questions 185
5.5 The future of palaeomicrobiology 186
5.4 References (Introduction and Conclusion) 188

List of Figures

Chapter 2

Figure 1	Historical intestinal specimen	22
Figure 2	The PA1849 genome	23
Figure 3	Evolutionary analysis of *Vibrio cholerae*	24
Figure S1	qPCR amplification curves for 58bp *ompW* assay	37
Figure S2	Fragment length distributions (FLDs)	38
Figure S3	PA1849 coverage vs. GC% (100bp window)	41
Figure S4	mapDamage graphs	42
Figure S5	Map of known geographic strain locations	45
Figure S6	Maximum likelihood (ML) phylogeny of 32 strains of *V. cholerae* excluding recombinant regions	46
Figure S7	Maximum parsimony phylogeny of 32 strains of *V. cholerae* excluding recombinant regions	47
Figure S8	Assemblies supporting CTX prophage arrangements	48
Figure S9	Possible PA1849 CTX region structural configurations	49
Figure S10	No evidence for truncated CTX prophage	50
Figure S11	PA1849 coverage across the O394 CRISPR region of GI-24	51
Figure S12	Human amelogenin X and Y coverage vs. GC% (100bp window)	52

Chapter 3

Figure 1	Burial #177 at Late Byzantine Troy, a female approximately 30 years old	85
Figure 2	The nodule bacterial metagenome (family-level)	86
Figure 3	*Staphylococcus saprophyticus* phylogenetic tree and genome content comparison	87
Figure 4	*Gardnerella vaginalis* phylogenetic tree and genome content comparison	88
Figure S1	Additional photos of nodules 1 & 2	115
Figure S2	SEM image of nodule 1 at 2,000x and 10,000x magnification	116
Figure S3	XRD analysis of nodule 1	117
Figure S4	Shotgun metagenomic phylum-level comparison between the 2 nodules, bone, sediment	118
Figure S5	Principal Components Analysis (PCA) of shotgun metagenomic datasets from the nodules, bone, and sediment	119
Figure S6	Shotgun metagenomic comparison between extraction blanks	120
Figure S7	Ancient DNA damage assessment of *S. saprophyticus* and	

	G. vaginalis	121
Figure S8	Ancient DNA damage assessment of human mitochondrial reads	124
Figure S9	Fragment length distribution (FLD) for *S. saprophyticus* ATCC 15305	126
Figure S10	Fragment length distribution (FLD) for *G. vaginalis* ATCC 14019	126
Figure S11	Fragment length distributions for nonU human mitochondrial assemblies	127
Figure S12	Gene diagrams for *S. saprophyticus* novel regions	128
Figure S13	MAUVE alignment of *S. saprophyticus* Troy plasmid pSST1	129
Figure S14	Coverage diagram for *G. vaginalis* ATCC 14019	130
Figure S15	Coverage diagram for *S. saprophyticus* ATCC 15305	131

Chapter 4

Figure 1	Number of bacterial families detected by HTS and/or LLMDA	150
Figure 2	Comparison of HTS and LLMDA results for cholera victim 3090.13	151
Figure 3	Comparison of HTS and LLMDA results for plague victim 8291	152
Figure 4	HTS vs. LLMDA comparisons	153
Figure 5	HTS rarefaction analysis	154
Figure S1	LLMDA detected probe distributions for *V. cholerae* and *Y. pestis* pathogens	167
Figure S2	Flowchart of workflow	169
Figure S3	Average LLMDA probe GC% vs. average LLMDA probe log intensity, by family	170
Figure S4	Number of HTS reads vs. HTS GC%, by family	171
Figure S5	Examples of LLMDA detected probe distributions	172

List of Tables

Chapter 2

Table S1	Sequences included on enrichment array	53
Table S2	Genomes used for phylogenetic comparison	54
Table S3	Results of genomic islands of interest	55
Table S4	Results of non-classical regions included on enrichment array	56
Table S5	Results of human loci include on enrichment array	56
Table S6	Human mitochondrial genome SNPs (relative to rCRS, NC_012920)	57
Table S7	RC27 vs. O395 genomic island comparison	57
Table S8	Results of the integron region core genome and O395-specifc genes	58

Table S9	SNP calls	60

Chapter 3

Table S1	Abundances of *S. saprophyticus*, *G. vaginalis*, and human mitochondrial genome via paired-end assembly	100
Table S2	Troy sample details	101
Table S3	SEM-EDS results from nodule	102
Table S4	Troy DNA extraction details	103
Table S5	HTS Datasets	104
Table S6	New modern *S. saprophyticus* strains sequenced in this study	105
Table S7	List of gaps and novel regions >100bp in *S. saprophyticus* Troy vs. strain ATCC 15305	105
Table S8	*S. saprophyticus* Troy novel regions >100bp	106
Table S9	Novel *S. saprophyticus* Troy region #02: Gene descriptions and top BLAST hits	107
Table S10	Novel *S. saprophyticus* Troy region #03: Gene descriptions and top BLAST hits	109
Table S11	Novel *S. saprophyticus* Troy region #13: Gene descriptions and top BLAST hits	110
Table S12	Novel *S. saprophyticus* Troy region #15: Gene descriptions and top BLAST hits	111
Table S13	Novel *S. saprophyticus* Troy region #16 Gene descriptions and top BLAST hits	111
Table S14	*S. saprophyticus* Troy plasmid pSST1: Gene descriptions and top BLAST hits	112
Table S15	*G. vaginalis* modern strains	113
Table S16	Called SNPs in human mitochondrial assemblies	114

Chapter 4

Table 1	Summary of LLMDA and HTS results	148
Table S1	LLMDA and HTS analysis, full results	159

List of Abbreviations and Symbols

aDNA	Ancient DNA
ATCC	American Type Culture Collection
BCF	Binary VCF file format
BEAST	Bayesian Evolutionary Analysis Sampling Trees, software for sequence analysis
BLAST	Basic Local Alignment Search Tool, software for sequence analysis
CDS	Coding DNA sequence
CFX	Product line of qPCR equipment from Bio-Rad
CLiMax	Proprietary algorithm from LLNL for analysis of LLMDA data
CRISPR	Clustered regularly interspaced short palindromic repeats
CTX	Cholera toxin phage of *V. cholerae*
DB	Digestion buffer
DM	Demineralization
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphates
DTT	Dithiothreitol
EDS	Energy dispersive X-ray spectroscopy
FastQC	Software for quality control analysis of HTS data
FLD	Fragment length distribution
GATK	Genome analysis toolkit, software for HTS data analysis
GC%	Percentage of G+C bases in a DNA sequence
GI	Genomic islands
GTR	Generalized time reversible model of nucleotide substitution rate
HTS	High-throughput sequencing
LGT	Lateral gene transfer
LLMDA	Lawrence Livermore Microbial Detection Array
LLNL	Lawrence Livermore National Laboratory
MAFFT	Software for multiple sequence alignment
MAUVE	Software for multiple genome alignment
MCMC	Markov Chain Monte Carlo
MEGAN	MEtaGenome ANalyzer, software for metagenomic data analysis
ML	Maximum likelihood
MRSA	Methicillin-resistant *Staphylococcus aureus*
mtDNA	Mitochondrial DNA
NCBI	National Center for Biotechnology Information
O395	A 20th century strain of O1 classical *V. cholerae,* isolated in India in 1965
PA1849	1849 strain of *Vibrio cholerae* from Philadelphia, PA, USA
PCA	Principal Component Analysis
PCR	Polymerase chain reaction
PG	Phylocore genome
PhyML	Sequence analysis software for estimating maximum likelihood phylogenies
PTB	N-phenacylthiazolium bromide
PVP	Polyvinylpyrrolidone

RAxML	Randomized Axelerated Maximum Likelihood, software for sequence analysis
RC27	A 20th century strain of O1 classical *V. cholerae*, isolated in Indonesia in 1991
RDP4	Package for recombination analysis
RTX	Repeat-in toxin gene cluster of *V. cholerae*
qPCR	Quantitative polymerase chain reaction
SEM	Scanning electron microscopy
SNP	Single-nucleotide polymorphism
SXT	A conjugative genetic element conferring antibiotic resistance in *V. cholerae*
TB	Tuberculosis
TE-HTS	Targeted enrichment and high-throughput sequencing
tMRCA	Time to the most recent common ancestor
UDG	Uracil-DNA Glycosylase
UTI	Urinary tract infection
VCF	Variant call format, file format for DNA data
VPI	Vibrio pathogenicity islands
VSP	Vibrio seventh-pandemic GI
XRD	X-ray diffraction

Declaration of Academic Achievement

Included in this thesis are two papers published in the *New England Journal of Medicine* (chapter 2) and *Scientific Reports* (chapter 4) and one paper prepared for journal submission (chapter 3). The following outlines and summarizes the published scholarly contributions of myself and co-authors of the three works (published or in-preparation) that comprise this thesis. In cooperation with the co-authors, I participated in all experimental designs and data analyses, and was primarily responsible for execution of the laboratory experiments, genome reconstructions and content analyses, and manuscript preparation, including main and supplemental texts, figures, and tables. For chapter 2: AMD, JME, MK, JHT, DNF, AND, DJDE, and HNP designed the study, AMD, JME, MK, JHT, and HNP gathered the data, AMD, GBG, NW, JME, MK, MS, DNF, SF, KIB, ECH, and HNP analyzed the data, and AMD, JME, ECH, and HNP co-wrote the paper with the assistance of all co-authors. For chapter 3: HNP and CSP conceived of the research, AMD, HK, TS, MK, and AH designed and performed experiments, AMD, TDM, MK, JME, GBG, WA, HNP, and CSP analyzed and interpreted data, and AMD, TDM, HNP, and CSP contributed to manuscript preparation with assistance of co-authors. For chapter 4: HNP, MB, CJ conceived of the research, AMD, CJ, JT, JME designed and performed experiments, AMD, CJ, SG, TMP, JME, JT, JA, KM and HNP analyzed data, AND and SND provided samples, and all authors contributed to manuscript preparation.

Chapter 1

Introduction

1.1 Background

Palaeomicrobiology (1) is an interdisciplinary field concerned with the study of ancient microbes, especially human pathogens. Ancient DNA (aDNA) techniques offer a particularly apt approach for studying palaeopathogens, since sequenceable DNA has been demonstrated to survive for hundreds of thousands of years under certain conditions (e.g., 2). Since the first study in 1993 on *Mycobacterium tuberculosis* (3), a wide range of ancient bacterial diseases has been studied (see recent reviews 4, 5), albeit with a wide range of methods and reliabilities. The most abundant and prominent studies have focused on three species: *M. tuberculosis* (TB), *M. leprae* (leprosy or Hansen's disease), and *Yersinia pestis* (bubonic plague). While my focus is on bacteria, aDNA techniques have also been applied to the study of viruses such as 1918 influenza (6) and parasites such as whipworm (7) in human remains, as well as in non-human remains such as the potato blight oomycete in historic herbareum specimens (8, 9).

In the study of ancient bacteria, scholars are concerned with both (a) anthropological interests in revealing their specific roles in past human health, as well as (b) epidemiological and biological interests in reconstructing phylogenetic relationships between bacterial lineages. Over the last ~5000 years, the largest category of selective pressure faced by humans is likely to have been infectious disease (10). In the case of (a), we seek to understand what effect(s) the species had on past individuals or populations. Did ancient pathogens have virulence properties that might explain the historical prevalence of a disease? For example, Black Death *Y. pestis* is estimated to have caused millions of deaths, so a reconstruction of its unique genes would provide insights into whether the high mortality was mainly due to this particular strain, or instead was only one of a number of contributing factors (11). In the case of ancient microbiomes, can we gain a greater understanding of individual past health experiences based on the known health outcomes of modern species? For example, higher than average diversity in modern vaginal microbiomes has been linked to disease outcomes such as bacterial vaginosis, whereas lower than average diversity in gut microbiomes has been linked to inflammatory bowel disease and obesity (12). Were the overall diversities of these microbial communities similar in the past, and even if so, would it be accurate to infer similar health outcomes? In the case of (b), we wish to better understand the conditions under and rates at which features of interest (such as virulence traits) are gained, lost, or change. In what circumstances do certain traits confer a selective advantage to a species? For example, the presence of certain key genomic features such as cholera toxin are physiologically linked to

virulence in *Vibrio cholerae* (13); under what environmental scenarios (anthropogenic, climatic, etc.) were these traits favorably maintained in certain *V. cholerae* strains in the past? Additionally, can we use ancient strains to provide anchoring calibration points for reconstructing the timing of key phylogenetic events, such as a jump to humans as host or a recent global dissemination? Even just considering the past tens of thousands of years that we *know* are within the current feasible range for aDNA study, this provides an enormous landscape of observable scenarios of human-pathogen co-evolution, with the goal of eventually offering insights into combating or predicting future bacterial virulence.

However, the field of palaeomicrobiology has suffered from some critical gaps in our knowledge and methodologies that render some aspects of ancient health "invisible" to us. By the nature of time, certain aspects of the experiences and outcomes of past infectious disease necessarily remain shrouded because they cannot be directly observed, such as non-recorded attitudes, beliefs, concerns, and theories about disease, and the states of mind of those who experienced or witnessing disease in the past (14, 15). Cultural and individual experiences of disease in the past are distinct from what we today would consider the underlying biological root of disease – the pathogen and the host – as well as the associated physical and environmental factors that contribute to risk of infection and illness, such as climate, resource access, and host behavior (10, 14, 16). The difference between these "social diagnoses" as experienced by people in the past and our "biological diagnoses" as we today would interpret them (i.e., presence of a pathogen) may only be reconcilable to a limited extent, depending on the historical contexts and available materials for our study (14). However, we can (and do) attempt to reconstruct as much as possible based on human remains, writings, and material culture left behind (1, 14, 17-19). My thesis contributes to two unseen aspects of palaeomicrobiology, the physical and the epistemological. For the former, I refer to the fact that palaeogenetic investigations of ancient bacteria have overwhelmingly been restricted to specific pathogens that leave DNA residues in skeletal remains. For the latter, I refer to the wider issue of the limitations that targeted approaches place upon our inferences and conclusions about ancient health.

1.2 The physically invisible: Ancient extraskeletal bacteria

Palaeopathology has largely focused on the skeleton, which constitutes the bulk of archaeological human remains. However, inferences about infectious diseases that can be made from physical observations are limited (17). Only certain diseases, such as tuberculosis, brucellosis, leprosy, and osteomyelitis (which can be caused by a number of bacteria such as *Staphylococcus aureus*) can characteristically modify skeletal elements, though generally only after severe and chronic illness (20). This means that reconstructing disease-specific morbidity

and mortality based on skeletal indicators should not be done without careful consideration of additional factors such as statistical analysis of overall population age and sex demographic structures (and in some cases, may not be possible) (17, 21). The application of aDNA techniques to provide "diagnoses" of bacterial infection without the need for palaeopathological indicators was an important development in the field (4). However, many bacterial species that are critically important contributors to past health are simply not recoverable from skeletal remains, because they were never located there endogenously. This includes myriad non-bloodborne pathogenic and commensal species that colonize organ surfaces, such as skin or the gastrointestinal, urinary, and reproductive tracts. Cholera is one example of an extraskeletal infectious disease of immense historical and modern importance. The unseen species of the microbiome are critical to health outcomes, often functioning physiologically in roles such as digestion; for example, gut microbiome species such as *Bifidobacterium* and *Clostridum spp.* render more calories available from ingested molecules that are not initially digestible by the host, as well as alter the capacity for nutrient absorption in the intestine (22). As non-skeletal tissues are rarely observed archaeologically, aDNA studies of extraskeletal infectious diseases and the microbiome are equally rare.

1.2.1 Extraskeletal infectious diseases

To help evaluate the impacts of infectious diseases on past human societies, researchers have relied on archival research into historical accounts and demographic records of births, illness, and deaths (10, 18). However, this research is by nature restricted to what is mentioned in historical records, which are human-generated and therefore (as with any source of data) are subject to certain biases (14). Generally, we would expect historically-recorded accounts of disease to be focused on diseases with striking symptoms and/or that caused marked morbidity and mortality (15). Socially-constructed definitions of "normal" health, different understandings of pathogens/vectors, and level of impact on the social classes responsible for the records are all factors that will impact historical reports of disease (14). Therefore, it is often challenging (though not impossible) to link such diseases with known modern pathogens (14), and aDNA techniques offer one solution to establishing the identity of past historically recorded plague – for example, *Y. pestis* and the Plague of Justinian (23). However, for extraskeletal diseases, such as cholera, there are additional, unique challenges.

In the case of cholera, the disease is historically well known over the last 200 years (24) and it continues to be responsible for extensive critical modern outbreaks (25). Cholera is an acute illness caused by gastrointestinal infection with the bacterium *Vibrio cholerae* that induces severe diarrhea and can cause death if not treated (26). There have been eight historical pandemics since the 1800s, most spreading globally from cholera's endemic source in

the Indian subcontinent (24). Though it is not certain, due to ambiguous historical accounts, it is probable that the disease itself is thousands of years old (27), though perhaps only recently laterally acquired the specific traits required for pandemic-level virulence (13). Bacteria possess an ability to integrate novel genomic regions into their genomes (from free DNA or other strains) known as lateral gene transfer (LGT) (28). LGT is especially relevant to evolutionary genetics studies because it is one of the main methods of genomic change over time, as well as a major mechanism for the acquisition of pathogenic traits such as virulence or antibiotic-resistance genes (29). In *V. cholerae*, there are multiple essential virulence factors that were acquired through LGT that confer properties such as toxicity or adhesion, and the presence of different virulence factors is one of the main distinguishing features between pandemic *V. cholerae* lineages of the last century (13). However, despite cholera's extensive impacts since the early 19th century (and most likely earlier), the disease has not been studied bioarchaeologically because it does not leave any skeletal indicators of infection and does not preserve in the skeleton. If sufficient cells are ingested to survive the acidic stomach environment, *V. cholerae* colonizes the intestine and rapidly multiplies. In conjunction with a variety of virulence factors required for adhesion and life in the intestine, the cells release cholera toxin. This induces massive water loss from the intestine cells and results in severe diarrhea, releasing the multiplied cells back into the environment (30). The disease's acute nature and restriction to the gastrointestinal tract mean that infection or death due to *V. cholerae* cannot be detected in the skeletal remains of its victims. In order to recover ancient *V. cholerae* DNA, one would require past environmental, faecal, or gastrointestinal specimens in which sufficient cells had been shed or colonized during infection, though such specimens have been neither (to date) recovered and/or examined with aDNA methods.

1.2.2 The extraskeletal microbiome

For understandable reasons, outbreaks of infectious disease are simply far more likely to be noted in historical records than any information about commensal microbiome bacteria (in fact, I am aware of none of the latter). Areas of the body such as the mouth, rectum, skin, and vagina are home to distinct microbial communities, yet there is also considerable variation between individuals (31). Microbiome variation has been associated with risk of negative health outcomes such as digestive illness (32), obesity (33), or pre-term birth (34). In addition to their commensal presence, which for some includes roles in maintaining normal body functions like digestion, many species are also known to be occasionally pathogenic (12). For example, *Escherichia coli* is an enormously diverse species that includes harmless commensal strains as well as pathogenic strains that can infect the blood, brain, gastrointestinal tract, and urogenital tracts and cause serious disease such as meningitis, urinary tract infections, gastroenteritis, sepsis, hemorrhaging, or diarrhea (35). In contrast, the commensal strains are adapted to the

human gastrointestinal tract and do not harm the host, and in some ways may benefit the host through competitive exclusion of colonization of other types of pathogenic bacteria (36). The human microbiome is a very active area of modern medical genetic research, yet there is still much that is not well understood about the outcomes of particular microbiomes and their roles and relationship to major anthropologically relevant arenas such as diet and nutrition, transmission through kinship and community relationships, and states of health and disease (37). For example, the use of antibiotics can disrupt the gut microbiome community to such an extent that otherwise commensal species can take over and become detrimental to the host (22), such as is the case for *Clostridium difficile* (38). This highly diverse species, estimated to have diverged within the last 1.1-85 million years, with multiple independent emergences of pathogenic lineages (38).

The ability to examine ancient microbiomes will allow us to better understand the influence of diet, environment, and lifestyle factors that differ between communities of the past and the present. However, most body microbiomes are found in or on soft tissues, which are not often preserved in the archaeological record; as such, aDNA microbiome studies are limited to certain unique materials. For example, the skin, muscle, stomach, and colon of the frozen mummified Tyrolean Iceman (a.k.a., Ötzi) have been analyzed with PCR metagenomic techniques, revealing some probable endogenous species such as *Clostridium perfringens* in the colon, which is present in the modern faecal microbiome, but also many likely originating from contamination and decay, such as *Clostridium algidicarnis,* which is known to colonize cold-preserved tissues (39, 40). Other mummified soft tissues such as muscle from artificially preserved mummies from Egypt (41) reveal that amounts of human and putatively endogenous bacterial DNA vary greatly from sample to sample. This indicates that even if a soft tissue sample was (in life) endogenously colonized by a microbiome of interest, the survival to the present day of DNAs of these microbes will depend on both endogenous microbial load and preservation conditions, just as for skeletal specimens. Desiccated faecal specimens (coprolites) can reveal signals of the gut microbiome flora, such as *Bacteroides* spp. that are common to the modern faecal microbiome (42-44). Encouraging success in characterizing the oral microbiome has been seen in recent studies on calcified dental plaque (calculus) from prehistoric to medieval teeth (45, 46), including a *de novo* genome of *Tannerella forsythia*, a member of the 'red complex' bacteria responsible for periodontal disease (45). Another red complex species, *Treponema denticola,* was recently detected with shotgun analysis of tissue from Ötzi (47). Another example of an occasionally-pathogenic microbiome species that has been studied with aDNA techniques is the stomach pathogen *H. pylori*, found in >50% of modern individuals and responsible in some for gastric ulcers and adenocarcinoma (48). Epidemiology and outcome of *H. pylori* infection is complex, and lower socioeconomic status is associated with higher risk of infection (49). *H. pylori* amplification was attempted from 18th century alcohol-preserved

gastric ulcer and cancer specimens from the Hunterian Museum in London, but these attempts were unsuccessful (50). However, *H. pylori* genes were successfully amplified from frozen remains, dated to the 17th to 19th century, of a First Nations individual from British Columbia, Canada (51). While these studies have been few and far between, the application of high-throughput sequencing (HTS) techniques has convincingly demonstrated that ancient microbiomes can be observed under the right conditions, and are being applied with increasing frequency and success.

1.3 The epistemologically invisible: (Meta)genomic bacterial content

1.3.1 Genome-scale studies

The success of applying HTS techniques to the study of aDNA within the last decade has ushered in a revolution in the scale and quality of research questions being asked (52, 53). HTS is the "next-generation" of sequencing technology (such as the Illumina platform) that allows for the simultaneous sequencing of hundreds of millions of DNA molecules. Ancient DNA has particularly benefited from these approaches by coupling HTS with targeted enrichment (TE-HTS), which uses free-floating (e.g., 54) or fixed (e.g., 55) single-stranded baits to capture genomic sections of interest from a complex sample, increasing their abundance relative to background in the sequenced pool. The effectiveness of TE-HTS in palaeomicrobiology is due both to the generally highly degraded and damaged nature of aDNA molecules, as well as the overall generally low abundance of pathogen DNA within a specimen. We now know that endogenous aDNA molecules are often extremely short (for example, the average fragment length of all samples studied in this thesis is <100bp). This makes reconstructing large genomic regions with overlapping PCR experiments tediously difficult or impossible, since the PCR technique can only access fragments ~50bp or above, and therefore is best suited for studying individual loci (genes). However, the sheer number of sequences (even 30-40bp long) provided by HTS has allowed for large-scale bioinformatic approaches based on the partial or whole genome reconstruction from TE-HTS or "shotgun" metagenomic analysis to take over as the major theme of impactful aDNA research projects. The aDNA researcher who wishes to successfully study past pathogens from a genetic perspective will need to be comfortable and familiar with an interdisciplinary approach intersecting the fields of anthropology, bioinformatics, epidemiology, microbiology, and phylogenetics (at least!).

Recent success has been seen in HTS and TE-HTS genome-scale studies on ancient pathogens, notably *Y. pestis* from the Black Death (11) and the Plague of Justinian (23), medieval *M. leprae* (56, 57), 19th century *M. tuberculosis* (58), and very recently on medieval *Brucella melitensis* (59). Genome-scale studies do not just create catchy news headlines – such

as "*The Black Death: Plague that killed millions is able to rise from the dead*" (Steve Connor, The Independent, 2014-Jan-28) or "*One of these plagues is not (genetically) like the others*" (Melissa Healy, Los Angeles Times, 2014-Jan-28) – but are critically important for studying pathogen evolution. Evolutionary relationships between bacteria and characteristics (such as virulence) are not just determined by presence/absence of and mutations in individual genes, but also by the order and arrangement of large-scale genomic regions (60, 61). For example, overall genome reduction is believed to play a role in the contrasting human virulence of *Y. pestis* vs. ancestral *Y. pseudotuberculosis* (62), and certain genomic islands are essential for different virulence capabilities of *V. cholerae* strains, such as the presence of the genes for cholera toxin, the O1 antigen, or biofilm and adhesion (13).

1.3.2 Targeted approaches are biased

However, while highly informative, most of the genome-scale studies mentioned above still rely on inherently biased research designs. By their nature, targeted approaches (PCR and TE-HTS) restrict both the overall organism as well as the specific genomic content of study. TE-HTS experiments are necessarily restricted to whatever modern genomic content was used for bait design, which could very well leave out virulence-associated genes of the ancient strain. However, due to financial and methodological constraints, it is not currently feasible to reconstruct most bacterial genomes without using TE-HTS, due to low bacterial DNA abundance. Of the current ancient genome-scale studies mentioned above, only one (56) had sufficient endogenous pathogen content to also allow for a *de novo* genome reconstruction from one of their samples. For the other studies on *M. leprae, M. tuberculosis,* and *Y. pestis*, the abundance of pathogen DNA was probably far too low for the researchers to feasibly afford what would necessarily entail a very high amount of shotgun sequencing. For example, in this thesis, the shotgun HTS datasets analyzed in chapter 4 had abundances of only 0.03-0.08% for the two pathogens of interest. If one's goal is genome-scale reconstruction, an even greater sequencing depth is required. For example, the *V. cholerae* genome reported in chapter 2 of this thesis was obtained with TE-HTS and 5 lanes of an Illumina GAIIx sequencing run (over 170 million reads) with a cost of approximately $2,500. Without TE-HTS, it would have required a minimum of 40 full lanes (~$20,000) to achieve this same dataset! It is very computationally intense to analyze hundreds of millions of reads, and perhaps more critically, it can be prohibitively expensive for many researchers to generate these data.

But beyond the inherent observational bias of TE-HTS, the very approach of most pathogen studies necessitates an inherent bias from the outset. Research questions are typically formed around a specific pathogen(s) that has been deemed likely as the culprit of infection, according to prior archaeological or historical knowledge; for example, seven

pathogens – *Bacillus anthracis* (anthrax), *Bartonella henselae* (cat-scratch disease), cowpox virus, *M. tuberculosis*, *Rickettsia prowazekii* (typhus), *Salmonella enterica* serovar Typhi (typhoid fever), and *Y. pestis* – were targeted by PCR during an investigation of the pathogen responsible for the Plague of Athens (63). These seven pathogens were chosen from amongst an even larger list generated by modern historians based on Thucydides' account of the Plague of Athens' symptoms, which included diarrhea, fever, and rash (63). However, such an approach is implicitly self-limiting approach, which restricts us to pathogens that have already been implicated by other methods (i.e. paleopathology or historic records) and/or continue to be compelling social narratives (as for the "Black Death"). This pulls our focus towards major pathogens of interest, and effectively away from any potential co-infecting pathogens or commensal microbiomic components of health. In this, we are likely missing out on other critical components of past health, and therefore limiting any potential inferences that we can make. For example, as noted by Papagrigorakis *et al.* (2006) in their study of the Plague of Athens, Thucydides reported a rapid onset of the disease. However, this is not consistent with modern outbreaks of the disease typhoid fever (*S. enterica* Typhi), the pathogen that they detected in their palaeomicrobial diagnostic investigation. Papagrigorakis *et al.* postulate that possibly genetic differences or the presence of other unknown co-infections may be partially to blame for symptom differences between past and present.

The contributions of co-infecting pathogens in both the long-term evolutionary and short-term individual effects on human health are complex but critically important areas of research. Co-infecting pathogenic strains can lessen ("inhibit") or worsen ("enhance") each other's effects due to the competitive physiological interactions between the pathogens and their host (64, 65). Lower socioeconomic status can put an individual at even greater risk for negative health impacts due to both the cumulative intersecting impacts of infection with multiple pathogens as well as an overall higher likelihood of contracting and succumbing to infections (66, 67). This conceptual view of disease as the interaction of co-infections and the cultural and environmental circumstances that led to their occurrence is known as "syndemic" theory. In this framework, the co-occurrence of these diseases is mediated by underlying social circumstances which perpetuate the disease burden, leading to a sum of negative health impacts that is more than the additive effects of either infection alone (68). For example, co-infection with HIV is a critical risk factor determining the development of symptomatic TB in an infected individual due to reduction of the host immune response, and in turn TB is the major cause of mortality for individuals with AIDS (69). Unfortunately, infections with both pathogens are often associated with developing countries and lower socioeconomic status, which exacerbates their impacts even farther beyond their inherently complex epidemiologies and treatments (69). The syndemic nature of TB/HIV infection with disparities in health and

resource access (66) necessitates that any epidemiological, evolutionary, or public health studies of these pathogens take these factors into account in a holistic approach.

1.3.3 Metagenomic approaches

To more fully examine the true scope of microbial contributions to past/modern health, much less-biased approaches are needed for both the reconstruction of individual genomes, but more importantly, for comprehensive metagenomic reconstruction of all co-infecting microbes. Even studies with completely reconstructed ancient genomes are still missing a part of the picture without contextual information about the whole ancient microbiome. While several 'non-targeted' approaches are currently available, all have methodological and financial drawbacks, and none currently offers a complete solution to the issues noted above. One option is to target a widely conserved locus (such as 16S rRNA) that is varied but retrievable across many species (70). 16S rRNA studies were adopted as the standard for bacterial metagenomic research in the last several decades because they offer a relatively simple and cost-effective simple workflow (71). The use of a molecular genetic-only method such as 16S rRNA PCR solves one of the major issues of bacterial research for complex metagenomic samples, namely, that not all bacteria in a sample can be cultured under laboratory conditions, due to specific growth requirements (72). This creates a significant bias in the metagenomic representation of a sample. Using a PCR assay directly on the extracted DNA from a complex human or environment sample (such as a cheek swab or a water sample) enables the broad study of the metagenomic content of the sample without relying on the bottleneck of bacteria culture (71). 16S studies have surveyed a wide range of metagenomic contexts, including the core human microbiome (73) and environmental biomes such as soil (74).

However, the 16S rRNA approach unfortunately has both limited resolution (75), and relies on biased PCR amplification (76). The limited resolution of 16S studies is driven by the need for standard primers that target "universally" conserved regions of the locus (71). However, this naturally constrains the size of the internal informative regions. Also, such studies cannot possibly capture all the known or unknown diversity, because the "universally" conserved regions of the locus are not maintained in all taxa and thus will not be sampled (or only poorly) (71). Additionally, the need for PCR amplification itself can skew the relative representation of the metagenomic content of the sample (i.e., the relative percentage of different taxa) because of unequal rates amplification across varying target lengths and GC% (75-77), which could be problematic for certain research questions such as comparative microbiome studies.

A broader, less-biased approach to studying the metagenomic content is using HTS with a "shotgun" approach to sequence a large subset of DNA in a sample and assess it against large comparative databases (78). This has become in many cases the "next generation" of metagenomic research because it allows for the same non-culture benefits as for 16S PCR, but without the restrictive bottlenecks limited resolution of that locus and with the possibility of reconstructing functional information from the recovered genes (71) (and, ideally, requires fewer cycles of overall amplification that might skew the relative taxa proportions). Like with 16S, HTS metagenomic studies have evaluated a wide range of metagenomic contents, from human microbiomes such as saliva (79) or environments such as whale falls (80). However, the HTS metagenomic shotgun approach is still not ideal for palaeomicrobiology, because extremely deep (and expensive) sequencing depths are typically required to reliably assess or identify ancient pathogen DNA, which is usually in very low abundance, as noted above. The use of this strategy to screen samples for potential targets of deeper genomic study amongst dozens or hundreds of archaeological skeletal remains would be prohibitively expensive for most researchers. That said, the recent TE-free genome reconstructions of medieval *M. leprae* from a skeletal sample (56), *T. forsythia* from dental calculus (45), and *B. melitensis* from a medieval calcified nodule (59) are encouraging signs of a new wave of palaeopathological research, although such studies were conducted on specimens with relatively high endogenous content for the particular microbes of interest.

1.4 Goals of this thesis

To fully investigate microbiomes and co-infections of the past, we need to both pursue the in-depth aDNA examination of novel extraskeletal specimens, as well as apply thorough and (as much as possible) non-biased assessments of all historical and archaeological human remains. The research questions I address in this thesis are as follows:

1.) Can genome-scale data be recovered from two novel extraskeletal historical/archaeological specimens?
   - a.) Was the 1849 Philadelphia cholera outbreak caused by *V. cholerae*, and if so, what strain was responsible?
     - i.) How does historical *V. cholerae* compare genetically with modern strains?
     - ii.) Can the date and sequence of a historic *V. cholerae* strain aid in calculating the origin of pandemic lineages?
   - b.) What is the pathological origin of a calcified infection in a young woman from Late Byzantine Troy?

i.) How do Medieval strains of *Staphylococcus saprophyticus* and *Gardnerella vaginalis* compare genetically with their modern counterparts?

2.) Can the LLMDA microarray technology provide an affordable, less-biased approach to the metagenomic study of archaeological human remains?
   a.) Can the LLMDA technology detect and identify specific known pathogens in archaeological remains?
   b.) How does the LLMDA approach compare to a more traditional deep-sequencing metagenomic study using HTS?

In this thesis, I investigate three ancient bacterial species that have never before been studied with aDNA techniques. In Chapter 2, I report a draft genome of the gastrointestinal epidemic pathogen *V. cholerae* from a 19th century alcohol-preserved intestine from Philadelphia (PA, USA). In Chapter 3, I report on genome-scale data from two occasionally pathogenic urogenital species recovered from a calcified urogenital infection from medieval Troy (Turkey): an essentially-complete genome of *Staphylococcus saprophyticus* (2nd most common cause of urinary tract infections in young women) and the coding genome content of *Gardnerella vaginalis* (associated with bacterial vaginosis). In both Chapters 2 and 3, I aim to contextualize these three ancient species within their available conspecific phylogenetic datasets, and characterize their virulence properties as compared to extant strains to determine whether they may have had similar ecological disease roles in the past. In Chapter 4, I apply a metagenomic characterization tool that has never before been utilized with ancient specimens called the Microbial Detection Array. I explore whether this technique can detect low abundances of known pathogen DNA and how it compares to shotgun HTS metagenomic analysis. Lastly, in my concluding remarks in Chapter 5, I address the consequences of these studies for the anthropological study of past health and disease, particularly in the areas of human-pathogen interactions and female and reproductive health.

The order of the chapters reflects the underlying shift in my thesis research away from strictly targeted approaches towards much less-biased assessment of bacterial genomic and metagenomic content. While a variety of financial and methodological considerations still constrain these approaches, palaeomicrobiologists are certainly making great strides in the right direction. That said, genetic study of past pathogens must be approached with the perspective that the social and environmental contexts in which a disease existed are as fundamental as the physiological effects of infection. It may well be that there is little that genetically distinguishes the bacteria responsible for 'killer' pandemics (such as the Black Death, or 19th century cholera) from their modern-day counterparts. Instead, as addressed above and

in my concluding discussion, a myriad of other factors such as socioeconomic outcomes, host genetics, co-infecting pathogens, or environmental conditions contribute to disease, and may have been the driving forces behind the severity of historical pandemics (10, 16). The role of palaeomicrobiologists must be to characterize, as thoroughly as possible, the genomic properties and metagenomic communities of ancient microbes, with the goal of determining what specific contributing roles (major or minor) they played in health outcomes of the past. In this way, we can enhance both our knowledge about the past as well as our predictive power regarding biological and social contexts of infectious disease in the future.

Chapter 2

Second-Pandemic Strain of *Vibrio cholerae* from the Philadelphia Cholera Outbreak of 1849

Published **January 2014** in **New England Journal of Medicine** 370(4): 334-340
Copyright © (2014) Massachusetts Medical Society. Reprinted with permission.
doi: 10.1056/NEJMoa1308663

Alison M. Devault[12], G. Brian Golding[3], Nicholas Waglechner[4], Jacob M. Enk[13], Melanie Kuch[12], Joseph H. Tien[7], Mang Shi[8], David N. Fisman[6], Anna N. Dhody[9], Stephen Forrest[1], Kirsten I. Bos[12], David J.D. Earn[45], Edward C. Holmes[8], Hendrik N. Poinar[1234]

(1) McMaster Ancient DNA Centre, McMaster University, Hamilton, ON, Canada; (2) Department of Anthropology, McMaster University, Hamilton, ON, Canada; (3) Department of Biology, McMaster University, Hamilton, ON, Canada; (4) Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, ON, Canada; (5) Department of Mathematics and Statistics, McMaster University, Hamilton, ON, Canada; (6) Dalla Lana School of Public Health, Toronto, ON, Canada; (7) Department of Mathematics, Ohio State University, Columbus, OH, USA; (8) Marie Bashir Institute for Infectious Diseases and Biosecurity Institute, School of Biological Sciences and Sydney Medical School, University of Sydney, Sydney, Australia; (9) College of Physicians of Philadelphia, Mütter Museum, Philadelphia, PA, USA

2.1 Thesis Preface

The following paper explores the genomic content of a 19th century strain of *Vibrio cholerae* from Philadelphia, PA, which was part of the second global cholera pandemic (1829-1851). The port city of Philadelphia, a bustling industrial trade center, was struck by several cholera epidemics in the 1800s. While it was not yet recognized that a pathogen was responsible, the disease and its consequences were notorious, and health officials and practioners vigorously debated its etiology. The College of Physicians of Philadelphia commissioned a report on the intestinal morphologies of cholera victims in order to investigate if there were any physical clues that could explain cholera's severe diarrheal symptoms. This resulted in the collection of the extremely rare alcohol-preserved intestinal specimens that formed the basis of this unique ancient DNA investigation. Alcohol is an excellent DNA preservative, and the sealed jars provide a time capsule for cholera DNA from over 150 years ago. This paper describes the first ancient DNA study of *V. cholerae* – and it may be the last for a while, until more archaeological or medical specimens are located, or creative solutions for investigating environmentally preserved *V. cholerae* DNA are pursued.

Recent severe cholera epidemics such as in Zimbabwe in 2008 (1) and Haiti in 2010 (2) indicate that epidemic cholera continues to persist as a severe modern pathogen, even though classical *V. cholerae* has been replaced by newer strains O1 El Tor and O139 as the dominant pandemic lineages. In this paper, I have established the first genetic evidence that 19th century cholera was indeed caused by a classical strain. I also provide evidence that there may have been genomic features of the historical strain that conferred different (possibly greater) virulence to the epidemics. However, the historic strain is overall very similar to 20th century classical. Accordingly, this research supports the view that improvements in sanitary infrastructure and public health eradicated epidemic *V. cholerae* from North America, rather than any pathogen genetic changes. Additionally, phylogenetic dating of the historic and modern human pathogenic strains provides support for an origin of cholera hundreds or, more likely, thousands of years ago, which supports anthropological ecological theories of the disease's origin during the first epidemiological transition (3).

2.2 Abstract

In the 19th century, there were several major cholera pandemics in the Indian sub-continent, Europe, and North America. The causes of these outbreaks and the genomic strain identities remain a mystery. We used targeted high-throughput sequencing to reconstruct the *Vibrio cholerae* genome from the preserved intestine of a victim of the 1849 cholera outbreak in Philadelphia, part of the second cholera pandemic. This O1 biotype strain has 95 to 97% similarity with the classical O395 genome, differing by 203 single-nucleotide polymorphisms (SNPs), lacking three genomic islands, and probably having one or more tandem cholera toxin prophage (CTX) arrays, which potentially affected its virulence. This result highlights archived medical remains as a potential resource for investigations into the genomic origins of past pandemics.

2.3 Glossary

Ancient DNA. The term used to characterize nucleic acids isolated from an extinct or long-since dead organism. Ancient DNA is typically highly degraded and damaged.

Cholera toxin phage (CTX). One of the main virulence factors of *Vibrio cholerae*, located on a genomic island acquired through the CTX prophage. Different pathogenic strains possess different CTX numbers and variants. Cholera toxin itself interferes with the membrane pores of gastrointestinal cells, inducing the water loss that leads to severe diarrhea.

Clade. A phylogenetic group that contains an ancestor and all its descendants.

Genomic islands (GI). Unique genome regions, frequently mobile elements, acquired via lateral gene transfer; these regions can confer specialized functions, such as antibiotic resistance or pathogenicity.

High-throughput sequencing. A suite of modern sequencing technologies that generate millions to billions of unique DNA sequences per run. Also known as next generation sequencing.

Molecular-clock dating. Dating of divergence times between taxa by estimating the number of substitutions separating them and calibrating with either known times of sampling or fossil ages.

O1 classical. One of the two pathogenic cholera biotypes of the O1 serogroup. Classical cholera is believed to be responsible for the first six global pandemics (until the emergence of El Tor in the 20^{th} century).

O1 El Tor. The second of the two pathogenic cholera biotypes of the O1 serogroup. El Tor strains have been responsible for most of the cholera outbreaks during the past approximately 50 years.

Phylocore genome (PG). The clade that includes all known pandemic cholera strains of serogroups O1 and O139, plus their close relatives. PG-1 and PG-2 are two subclades.

Prophage. A viral genome that has integrated into a bacterial plasmid or chromosomal genome.

Recombination. The process in which novel genetic sequences are formed via the exchange and combination of DNA from two original chromosomes.

RTX. “Repeat-in-toxin” gene cluster, located next to the CTX locus.

Single-nucleotide polymorphism (SNP). A DNA mutation that is a single base pair variant at a position (A, T, G, or C). A SNP is a form of point mutation.

Site saturation. Over evolutionary time, mutations in a DNA sequence can reoccur at the same site, in many cases reverting to a previous nucleotide. This can reduce, or saturate, the available signal for estimating rates of change or divergence between two sequences, leading to erroneous inferences.

Targeted enrichment. Method to "enrich" for DNA of a certain desired target (e.g., a pathogen genome) by the exposure of synthesized single-stranded "bait" molecules (RNA or DNA) to total prepared single-stranded sample DNA. Baits are either free-floating or fixed to a medium such as a glass slide. Baits and sample are exposed to each other for many hours, which allows for hybridization between any complementary sample and bait DNA sequences. The bait-sample mixture is subsequently washed to remove loosely or poorly bound sample molecules. Any hybridized sample molecules that remain bound to the baits are retrieved and sequenced with the use of high-throughput sequencing.

Vibrio pathogenicity islands (VPI). VPI-1 and VPI-2 are genomic islands (see above) in pathogenic *V. cholerae* strains that are known to function in cholera virulence through the addition of important adhesion and other genes.

Vibrio seventh-pandemic (VSP) islands. VSP-I and VSP-II are genomic islands (see above) found in seventh-pandemic *V. cholerae* strains (clade 7P, in PG-1), but not other virulent *V. cholerae* strains (such as O1 classical).

2.4 Introduction

Cholera is a diarrheal disease caused by colonization of the intestines by cholera toxin-expressing strains of the waterborne enteric bacterium *V. cholerae*. An outbreak can arise suddenly, especially in vulnerable populations with compromised sanitation infrastructure, as in the devastating 2010 outbreak in Haiti (2). In 2012 alone, *V. cholerae* infected 3 million to 4 million people, killing nearly 100,000 (4). Although all pathogenic *V. cholerae* possess a similar genomic backbone that may have facilitated adaptation to human intestinal mucosa (5, 6), the predominant pathogenic strain, serogroup O1, harbors two genetically distinct biotypes: classical and El Tor (for descriptions of these and other terms, see the Glossary). In the 20th century, for unknown reasons, El Tor replaced classical as the dominant biotype. There have been seven documented pandemics since 1817 (7), but the causal *V. cholerae* strains have been genetically characterized only for the two most recent outbreaks. Therefore, although most assume that the classical biotype was responsible for the earlier pandemics (8), the identities of the strains that caused them remain unknown.

Diverse tissue specimens archived by medical practitioners at the time of an outbreak represent an essentially untapped genetic museum for pathogen research. One extraordinary collection includes a preserved intestine from a patient who died from cholera in the 1849 Philadelphia outbreak (**Figure 1**) collected by Dr. John Neill. By coupling targeted enrichment

with high-throughput sequencing (9), we use trace degraded DNA from this specimen to reconstruct a mid-19th century *V. cholerae* genome and to test the hypothesis that the O1 classical biotype was responsible for the second cholera pandemic.

2.5 Methods

We extracted DNA with the use of an organic extraction protocol and converted the extracted DNA into Illumina sequencing libraries (10) in dedicated ancient-DNA facilities. We then enriched the libraries for the O1 classical genome (strain O395; National Center for Biotechnology Information reference sequences NC_009456 and NC_009457) as well as for regions not found in classical strains (e.g., vibrio seventh pandemic islands I & II [VSP-I and VSP-II]), human mitochondrial genome and amelogenin X and Y genes, and other regions. We mapped the sequencing reads to strain O395 with the use of Burrows-Wheeler Aligner, version 0.4.9rc1 (release 1561) (11), aligned the resulting consensus sequence to 31 full genomes (**Table S2** in [the Supplementary Appendix]) using MAUVE (12), called SNPs with a custom PERL script, and visualized features of the sequence in comparison with those of strain O395 with the use of Circos, verson 0.36-4 (13) (**Figure 2**). We used Gblocks (14) to prune insertions, deletions, and potential misalignments, which resulted in a final alignment of 28,591 SNPs that we used in maximum-likelihood (15) and Bayesian (16) phylogenetic analysis.

Our full methods and results are available in [the Supplementary Appendix]. Sequences can be found at the Sequence Read Archive (www.ncbi.nlm.nih.gov/sra) under the BioProject accession number SRP029921. Ethics approval for the study was obtained from Hamilton Health Sciences and McMaster University.

2.6 Results

2.6.1 Historical V. cholerae genome

We reconstructed a draft *V. cholerae* genome (PA1849) at an average unique coverage depth of 15.0X, comprising 94.8% of the O395 reference strain with at least 1X coverage and differing by 203 SNPs (see [**Table S9**]). If regions not present in PA1849 (see below) are excluded, then 97.4% of the reference sequence is present at an average of 15.4X. Reference strain O395 regions not covered by our sequencing data could represent missing, rearranged, or highly divergent regions in the historic genome itself; could be the result of preservation or procedural biases (e.g., poorly-preserved AT-rich regions, biased amplification, and uneven enrichment (9, 17)), or both. The *V. cholerae* DNA fragments have typical ancient-DNA damage

patterns (**Figure S4**) (18). Overall, coverage correlates strongly with GC content across most regions (**Figure 2; Figure S3**).

2.6.2 Genomic islands and virulence factors

Strain PA1849 shares the following phylocore genome (PG) genomic islands (GIs) with strain O395: O1, vibrio pathogenicity islands 1 and 2 (VPI-1 and VPI-2), and GI-1 through GI-10 (**Table S3**); in addition, PA1849 possesses the PG-2 islands GI-23 (putative prophage found today only in classical strains O395 and RC27) and GI-24 (a putative prophage with CRISPR [clustered regularly interspaced short palindromic repeat]-associated proteins) (**Figure 2**) (6). Strain PA1849 does not have GI-11, GI-14, or GI-21; the absence of these GIs suggests that they were acquired after 1849 by the modern classical strains.

Strain PA1849 contains all known major virulence regions (e.g., VPI-1, VPI-2, and CTX prophage) common to classical *V. cholerae* but does not have nonclassical genomic regions or variants (e.g., VSP-I and VSP-II) (**Tables S3 and S4**). The average GC content for these loci is not substantially lower than that in the successfully recovered genomic regions, suggesting that the absence of these loci is unlikely to be an artifact of preservation. VPI-1 has lower-than-average coverage (7.5X, vs. 15.0X), which is probably a result of its relatively low GC content (35%, vs. 46.7% for the entire genome) rather than its absence in the historical genome (**Figure 2**). Relative to the region in strain O395, VPI-1 in strain PA1849 contains one synonymous SNP (in *tcpA*), and VPI-2 contains four SNPs.

Like strain O395, PA1849 contains the classical *ctxB* and *rstR* variants and the expected deletion in the large-chromosome RTX element. It is also likely to possess the same CTX positions as strain O395, because it appears to have identical chromosomal flanking regions, albeit observed at low coverage. However, its CTX prophage configuration, which varies between strains (6, 19), has not been observed in classical strains to date; read assemblies indicate that there is a tandem CTX repeat span (**Figures S8 and S9**) on one or both chromosomes, with no read assemblies supporting the presence of the truncated CTX prophage repeat that is typical of modern classical strains (**Figure S10**).

2.6.3 Human mitochondrial and nuclear DNA

The complete mitochondrial DNA genome from the patient with cholera was retrieved with a coverage depth of 149.0X, and reads exhibit a typical ancient-DNA damage profile (**Figure S4c**) (18). The consensus sequence belongs to haplogroup L3d1b3, found today in sub-

Saharan western Africa (20). Reads matching amelogenin gene X and Y alleles suggest that this patient was male, although coverage across these regions was poor.

2.6.4 Origin and evolution of V. cholerae

Our phylogenetic analysis revealed a major division between the PG-1 and PG-2 lineages (**Figure 3**). Most El Tor strains cluster into the seventh-pandemic (7P) clade, which also includes strain MO10 (6, 21). Representatives of clades L6 (strain NCTC8457), L3 (strain 2740-80), and L5 (strains M66-2 and MAK757), together with the 7P clade, make up the PG-1 clade, whereas the PG-2 clade comprises of L7 (strain V52), L1 (classical strain), and strain PA1849. Strain PA1849 sits several SNPs away from the L1 clade node, with strong bootstrap support.

Our initial attempts to date the evolutionary history of *V. cholerae* were hindered by an inability to estimate evolutionary rates from tip dates alone, even when PG-1 and PG-2 were analyzed separately. This is probably the result of a combination of site saturation and the extensive recombination typical of *V. cholerae* (6, 21, 22). We confidently detected at least 37 recombination events between PG-1 and PG-2 (**Figure 3**), as well as a number of intragroup recombination events and events that brought genetic diversity into PG-1 and PG-2 from unknown parental lineages. Such a recombination frequency makes it difficult to determine whether specific recombination events explain the recent predominance of El Tor strains.

To overcome these limitations, we imposed a strict molecular clock (1.3×10^{-3} nucleotide substitutions per SNP site per year) based on a re-analysis of a large El Tor dataset (21). Assuming this rate, we estimate that the El Tor 7P strains emerged between 1940 and 1957 (95% highest posterior density), in agreement with previous estimates (21). Similarly, we estimate that the ancestor of the classical strains originated between 1843 and 1860, with divergence of the lineage leading to strain PA1849 occurring between 1797 and 1813, close to the time of the first recognized cholera pandemic, in 1817 (23). The combined topology and temporal estimations suggest that the first five cholera pandemics were caused by *V. cholerae* possessing a common core genome, each representing a clonal reemergence with few genome-scale mutational differences.

2.7 Discussion

The PA1849 genome has a number of unique structural features but differs from strain O395 by only a few hundred SNPs across the entire 4-Mb genome. This suggests that modern *V. cholerae* evolution has been subjected to substantial selective constraint since the mid-19th century, similar to that of other pathogens that exhibit long-term core genome sequence

conservation over a period of centuries (e.g. *Yersinia pestis* (9) and *Mycobacterium leprae* (24)). One of the striking features of the PA1849 historical genome is the tandem CTX configuration; this could indicate that it was capable of producing infectious CTX virions (25), which potentially conferred greater pathogenic capacity (26). Therefore, the suggestion that the absence of CTX virion production in classical strains may have contributed to their replacement by El Tor (26) may be unfounded. However, because there is a *V. cholerae* strain (B33) with tandem CTX repeats that does not produce replicating virions (27), the functional implications of this structure in strain PA1849 cannot be confirmed without experimental expression in a model vibrio strain.

Previous attempts to date the origin of pandemic *V. cholerae* have returned very different results (e.g., approximately 130 to 50,000 years ago for the classical-El Tor split (28)). If a constant evolutionary rate of 1.3×10^{-3} substitutions per SNP site per year were applied across the entire phylogeny, then the common ancestor of all pathogenic *V. cholerae* (PG, potentially the ancestral strain first adapted to humans) would date to only 430 to 440 years ago. However, a combination of site saturation and recombination means that this date is an underestimate, and the date of PG is more likely to be on time scale of millennia, predating all historically recognized pandemics and arguing against a post-medieval origin of pathogenic *V. cholerae* (28). Our analysis therefore suggests that the PG-1 and PG-2 lineages cocirculated in humans and water sources for many centuries and potentially thousands of years before the 19th century pandemics, a finding compatible with the theory that cholera is a disease of the "first epidemiological transition," during which sedentary agriculture (beginning approximately 10,000 years ago) opened new disease niches (3).

Collections of historical pathological specimens are invaluable resources for reconstructing pathogen evolution, yet the study of these collections remains a sensitive topic, because it was common for the bodies of marginalized minorities and the poor to be retained for medical research without consent (29). We hope that by highlighting the intrinsic scientific, historical, and social values of these underappreciated collections, we can help to recognize and protect them in perpetuity.

2.8 Acknowledgements

Supported by Canada Research Chairs from the Natural Sciences and Engineering Research Council of Canada (to Drs. Poinar and Golding), a grant from the Social Sciences and Humanities Research Council of Canada (to Dr. Poinar), a National Health and Medical Research Council Australia Fellowship (to Dr. Holmes), and an Ontario Graduate Scholarship (to Ms. Devault). Disclosure forms provided by the authors are available with the full text of this article

at NEJM.org. We thank Dr. D. Ann Herring, Dr. D. Poinar, Dr. C. Yates, Dr. G. Wright, and current and former members of the McMaster Ancient DNA Centre, McMaster University, Hamilton, ON, Canada.

2.9 Figures

Figure 1. Historical Intestinal Specimen.

Panel A shows specimen 3090.13, a preserved portion of intestine from an 1849 cholera victim. The specimen was collected by Dr. John Neill and colleagues for the purpose of studying the effects of cholera on the intestinal lining. Their 1849 report was presented to the College of Physicians of Philadelphia, and the specimens were displayed, later becoming part of the Mütter Museum collections. Panel B shows examples of histologic and morphologic sketches of intestinal damage that accompanied the report; on the left is a sketch of villi structure, and on the right is a sketch of villi at greater magnification with injected capillaries visible.

A



B

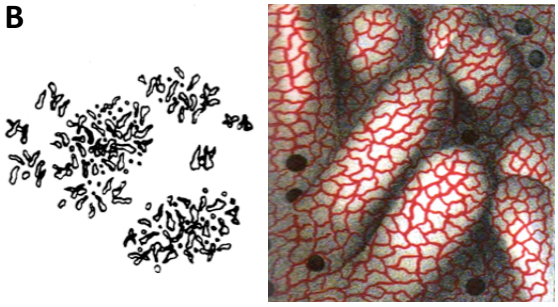


Figure 2. The PA1849 Genome.

Both chromosomes of the PA1849 draft genome, as compared with the O395 reference genome (NC_009456 and NC_009457), are illustrated. The ring closest to the center (in green and red) shows the GC content of O395 across 1000-bp consecutive windows (the gray axis line denotes the genomic average of 47.5%); the next ring shows the chromosomes (with the large chromosome shown in blue and the small chromosome shown in orange) measured in kilobases, with major genomic islands (GIs), vibrio pathogenicity islands (VPIs), cholera toxin phage (CTX), and toxin-linked cryptic (TLC) labeled; the next ring shows single-nucleotide polymorphisms (SNPs) as compared with O395 (black circles denote nonsynonymous, gray synonymous, and white noncoding SNPs); and the outer ring shows unique coverage across 100-bp consecutive windows (the gray axis lines indicate 25× and 50× coverage).

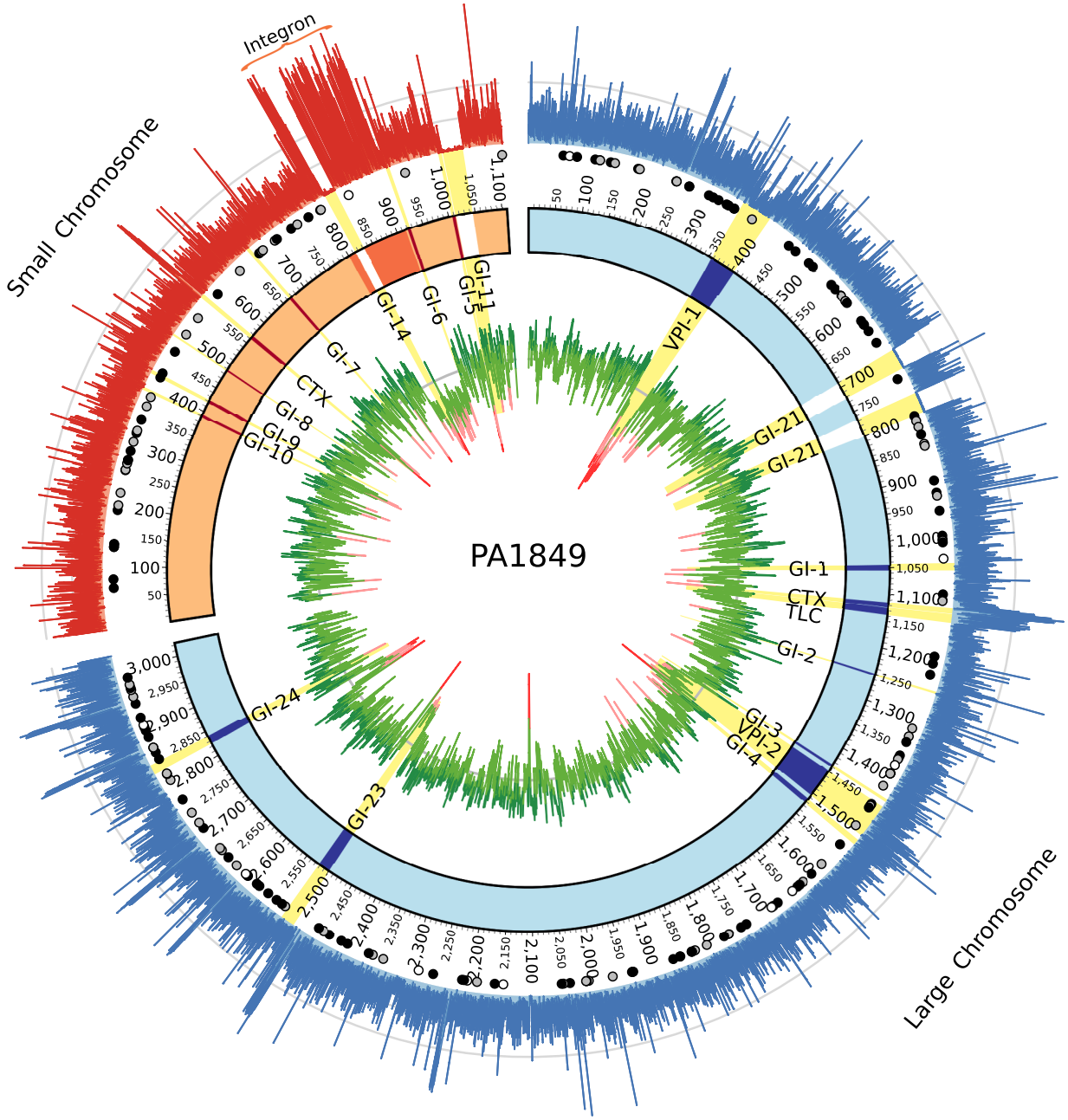
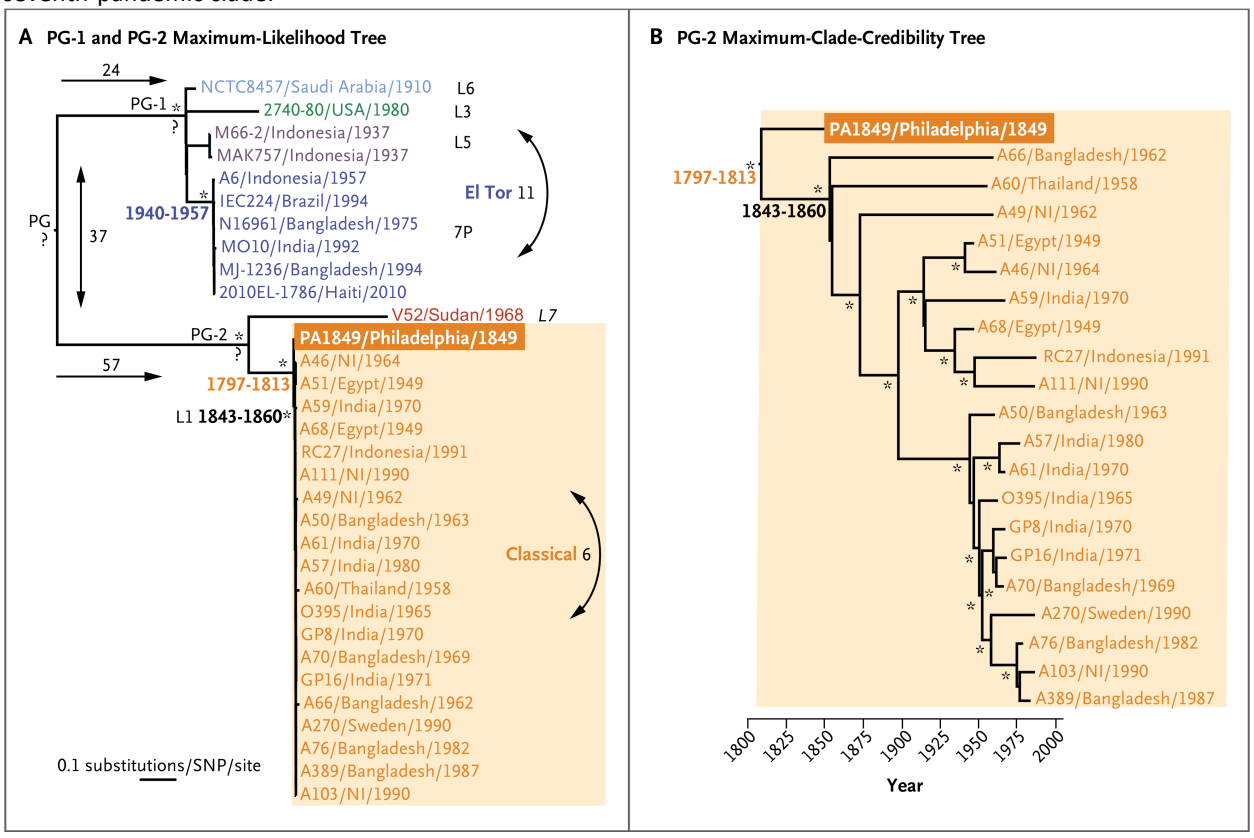


Figure 3. Evolutionary Analysis of *Vibrio cholerae*.

Panel A shows the maximum-likelihood phylogenetic tree of the El Tor and classical (including PA1849) strains (28,591 SNP sites) of *V. cholerae* present within the phylocore genome (PG) groups PG-1 and PG-2. Strain names are expanded to give the place and date of isolation. Branches are color-coded to match the geographic distribution of strains as shown in **Figure S6** in [Supplementary Appendix]. Estimates of the divergence times (95% highest posterior density intervals) of key nodes obtained from a separate molecular-clock dating analysis are shown. Question marks indicate key nodes, particularly PG nodes, in which a combination of multiple substitutions and recombinations hindered attempts to accurately estimate divergence times. Branch lengths are scaled to the number of nucleotide substitutions per SNP site, and bootstrap support values greater than 90% at key nodes are indicated with an asterisk. The numbers of recombination events within (curved double-headed arrows) and between (vertical double-headed arrow) the PG-1 and PG-2 groups, as well as into PG-1 and PG-2 from unknown parents (horizontal arrows), are also shown. Panel B shows the maximum-clade-credibility tree for the evolution of the PG-2 classical strains (excluding L7) in inferred chronologic time. All tip times are set to the time of sampling, as shown on the x axis. Posterior probability values greater than 0.90 are indicated with an asterisk. 7P denotes the seventh-pandemic clade.



2.10 Supplementary Appendix

The following sections contain the Supplementary Appendix, available online at NEJM.org.

2.11 Supplementary Methods

2.11.1 Cholera Victim Specimen

The intestinal specimen (museum catalog number 3090.13) from a putative 1849 cholera victim examined in this paper is part of the collections of the Mütter Museum (College of Physicians; Philadelphia, PA, USA). The original description associated with the sample is as follows: "*Intestine. Cholera. Yellow. Presented by Dr. John Neill.*" No date is directly associated with the specimen, and no museum records appear to directly reference this specimen. However, this set of specimens (of which at least 7 remain) that purport to be from cholera victims is almost certainly part of the exhibited collection presented by Neill in 1849 to the College of Physicians (see quote below). We have been unable to locate the dry preparations, microscope preparations, or drawings from this set. Also it is unclear which (if any) of the wet preserved specimens belonged to the control set of specimens (from the victims of pleurisy).

> "In 1849, the Asiatic cholera visited us for the second time, and Neill was physician to the Southeast Cholera Hospital, which was established in the old Friends' Meeting House, in Pine Street, above Front. During his service here he made the minute injections upon which was based the Report made to this College, and published in the Transactions, as we shall state more fully hereafter. … At the meeting of Dec. 4, 1849, the Committee on Cholera, of which Professor Samuel Jackson was Chairman, and Drs. Neill, H. H. Smith and Pepper were members, presented a report upon the injections of the intestines of patients dying of cholera, which had been made by Dr. Neill, as has already been stated. The Committee reported that "the admirable manner in which he performed this duty can be judged of by the beautiful preparations now upon the table, which he has presented to the College for its Museum." The injections were made with turpentine colored with vermilion, having failed when size and Canada balsam had been used. This had led, at first, to the supposition that the capillaries of the intestinal tract were destroyed by the disease; but the method devised by the investigator showed the perfect integrity of those vessels. A portion of healthy intestine, taken from the body of a patient who had died of pleurisy, was injected with the same fluid as a standard of comparison. These preparations speak for themselves; they form a groundwork for any future investigation, and those interested may find them in the Museum of this building. The whole series consisted of eleven mounted wet, and twelve dry preparations, four fitted for the microscope, and seven drawings. An interesting discussion ensued upon the presentation of these specimens to the College." (30)

2.11.2 Sampling

Subsampling of the 3090.13 intestinal tissue specimen was performed at the Mütter Museum's curation room by AMD and JT. The workspace was thoroughly bleached prior to

subsampling. The lid of the sealed jar was removed by cutting along the circumference of the lid using a thin dremel wheel (diamond-tip) and scalpel blade. The tissue specimen was removed from the solution with tweezers and placed into a clean weighing boat. A small section (~2x2 cm) was transferred to a sterile 50-mL falcon tube by slicing off the end of the intestinal specimen with a scalpel blade. The specimen was returned to the jar and the jar opening was temporarily sealed with plastic tape. The curatory staff restored the jar's seal after subsampling was complete. A secondary subsample was also taken at the time of sampling to retain in frozen storage at the Mütter Museum for future work.

2.11.3 DNA Extraction

All laboratory methods were performed in the facilities of the McMaster Ancient DNA Centre (McMaster University, Hamilton ON, Canada), which include physically separated "ancient" clean room (extraction, library preparation, and PCR set-up) and "modern" laboratory (amplification, enrichment, and re-amplification) facilities. Subsampled 3090.13 tissue specimen was stored at -20°C upon arrival. An approximately 92 mg tissue sample was subjected to DNA extraction using a modified tissue extraction protocol (31), alongside extraction blanks (1 per set of extractions). Tissue sample was divided into small fragments using a scalpel blade and incubated with 800µl of tissue digestion buffer (pH 7.5; 1% SDS, 50mM DTT, 25 mM Tris-HCl pH 8.0, 25 mM sodium citrate, 5 mM $CaCl_2$, 2.5 mM EDTA pH 8.0, 10 mM PTB, and water) for 95°C for 5 min (1000 rpm). We added 80 µl of proteinase K (20 mg/mL, activated), and the contents were incubated with rotation for 50°C for 24 hours. Digested material was subjected to organic extraction with 2x 800 µl phenol-chloroform (spun for 10 min at 14,000 rpm and aqueous phase retained) and 2x 800 µl chloroform. The organic extraction products were subjected to column ultrafiltration with 10 KDA cut-off (via 10K Amicon Ultra-0.5mL; Millipore, MA). Column was primed with 300 µl of 0.1x TE and spun to minimum retention. Extracted material was applied stepwise to column and spun to minimum retention. Three washes of 300 µl 0.1xTE were applied to the column and spun to minimum retention. Filters were inverted into clean microcentrifuge tubes and spun at 1000 RCF for 5 minutes to capture extracted DNA solution (if necessary, volume was increased to reach a final volume of 50 µl 0.1x TE for the extracted DNA solution).

2.11.4 Library Preparation and Indexing

Preparation of adapted libraries capable of sequencing on the Illumina platform was performed as given in reference (10), with some modifications as given in reference (32). 450 µl of a 1 in 10 dilution (in 0.1x TE) of extracted 3090.13 DNA solution was used as template in multiple library preparation reactions (50 µl template DNA in each of 9 libraries). 45 µl of

extraction blank library plus 5 µl 0.1x TE was also used as template in an additional library preparation reaction. Two library preparation blanks (0.1xTE; water) were also included. Library preparation enzymes were purchased from New England Biolabs (NEB) and used in these concentrations during library preparation: T4 polynucleotide kinase, 0.5 U/µl; T4 DNA polymerase, 0.1 U/µl; T4 DNA ligase, 0.125 U/µl. MinElute (Qiagen) PCR product purification was substituted for SPRI bead clean-up between library preparation steps; instead of a final purification after the fill-in stage, a 20 minute 80°C heat inactivation was performed, and the final product was used as template in indexing PCRs directly (32). After library preparation, the 9 3090.13 libraries were pooled to a final library volume of 360 µl. The three blanks had final volumes of 40 µl.

Post-library quantitation. Primers *IS7_short_amp.P5* and *IS8_short_amp.P7* (10) were used in a quantitative PCR assay alongside a quantitative library standard (49bp template insert) to quantify "copies/µl" (c/µl) of the libraries: 2.2E07 3090.13, 4.3E03 extraction blank, 6.3E03 0.1xTE blank, 8.7E03 H_2O blank. The qPCR conditions were 200 nM each primer, 1X PCR Buffer II (Applied Biosystems), 2.5 mM $MgCl_2$ (Applied Biosystems), 250 µM dNTPs, 0.15 mg/mL bovine serum albumin, 0.05 U/ µl Amplitaq Gold polymerase (Applied Biosystems), 0.167X SYBRgreen fluorescent dye (Invitrogen), and 1 µl template DNA (1 in 100 or 1:1K dilution) in a 10 µl final qPCR volume. A 95°C initial denaturation (4 min) followed by 50 cycles of 95°C (30s), 62°C (30s), and 72°C (30s) were performed.

Indexing. Double indexing (32) PCR reactions were performed as given in reference (10), using AmpliTaq Gold polymerase instead of Phusion Hot Start High-Fidelity DNA polymerase: 14x 3090.13, 2x extraction blank, 2x each library blank, and 2x PCR blanks. PCR conditions were 200 nM each indexing primer, 1X PCR Buffer II, 2.5 mM $MgCl_2$, 250 µM dNTPs, 0.15 mg/mL bovine serum albumin, 0.05 U/ µl Amplitaq Gold polymerase, and 10 µl DNA template, in a 100 µl final PCR volume. A 95°C initial denaturation (4 min) followed by 10 cycles of 95°C (30s), 60°C (30s), and 72°C (30s) were performed. Indexed reactions were pooled by sample/blank and every 200 µl was purified over MinElute to a final volume of 140 µl 3090.13 and 20 µl each blank.

2.11.5 V. cholerae DNA qPCR Assessment

A qPCR assay specific to the cholera genome was designed to amplify a 58bp fragment of the *ompW* gene (forward primer Vchol_ompW_507F GCTCAATGATAGCTGGTTCCTCAAC; reverse primer Vchol_ompW_564R CGTTGTTTCAATATTGGCATACCACAC). A 580bp fragment of the *ompW* gene was amplified from the 3090.13 sample (results not shown) and cloned to serve as a standard control of known copy number for qPCR. The 58bp *ompW* assay was

optimized for sensitivity down to <10 c/µl under the following PCR conditions: 200 nM each primer, 1X PCR Buffer II (Applied Biosystems), 2.5 mM $MgCl_2$ (Applied Biosystems), 250 µM dNTPs, 0.05 U/µl Amplitaq Gold polymerase (Applied Biosystems), 0.5X EVAgreen fluorescent dye (Biotium), and 1 µl template DNA in a 10 µl final qPCR volume, with cycling parameters of 95°C initial denaturation (4 min) followed by 50 cycles of 95°C (30s), 65°C (30s), and 72°C (30s). Amplifications using these conditions were performed (in replicate) on the following templates at 1 in 100 (0.01x) dilutions (in 0.1xTE): libraries (3090.13, extraction blank, 2x library preparation blanks), indexed libraries (3090.13, extraction blank, 2x library preparation blanks), and re-amplified libraries (extraction blank, 2x library preparation blanks, and indexing blank). Standards of known copy number (10K, 1000, 100, 10) and PCR blanks were also included in replicate. Positive amplifications of the 58bp product were only detected in the 3090.13 templates, at the following average calculated c/µl: 9.22E03 library, 9.76E03 indexed library (see **Figure S1**). All tested blanks (as indicated above) were free of the 58bp cholera specific product.

2.11.6 Array Design

DNA enrichment was performed using a custom one-million feature Agilent SureSelect DNA Capture Array (Agilent Technologies, Inc., Santa Clara, CA). All probes were designed using software described elsewhere (33). Efforts were made to identify and mask repetitive elements based on 15-mer frequency counts in the design template, where all probes carrying an average 15-mer frequency of 100 or more were eliminated (34). In addition, probes containing a 15-mer sequence represented in either of the Illumina adapter sequences (P7, CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGT and P5, AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT) were removed to discourage non-specific binding of adapters. Probes with base pair sequences identical to those already generated were subsequently removed. Both chromosomes of the *Vibrio cholerae* O395 strain (O1 classical serotype) were included on the array at 5bp tiling, with increased tiling density (2bp offset) across several specific regions of interest (see **Table S1**). Other virulence regions associated with other modern pathogenic cholera strains such as El Tor and O139 (but NOT expected in O1 classical) were also included on the array at 2bp tiling, such as VSP-I & VSP-II, the El Tor *tcpA* variant, and the O139 antigen region. Other probes were also included on the array at 4bp tiling density to enhance the recovered information on the historical cholera victim: human mitochondrial genome and human X and Y amelogenin genes. The array also included 13,729 additional non-project probes (results not discussed here). The total number of probes included on the array was 923,153.

2.11.7 Array Enrichment

Pre-enrichment amplification. Prior to enrichment, additional indexing amplification was performed in 32 reactions (2.5 µl straight template DNA in 50 µl total reaction volume) of the 3090.13 library (see above) using 400 nM each indexing primer and 15 cycles to reach a final, post-MinElute purified total DNA concentration of 187 ng/µl.

Enrichment. Array enrichment was performed as in reference (35) and the Agilent SureSelect DNA Capture Array protocol v1.0 (36). 18.7 total µg of amplified 3090.13 DNA (100 µl) was combined with 18 µl nuclease-free water, 40 µl of 8 blocking oligos (2 µM), 50 µl Cot-1 DNA (0.1 mg/mL), 52 µl 1X Agilent Blocking Agent, and 260 µl 1X Agilent Hi-RPM Hybridization Buffer. This mixture was distributed into low-bind PCR tubes and incubated at 95°C (3 min) and 37°C (30 min), pooled, transferred to the enrichment array slide sandwich in the hybridization chamber, and rotated in the dark at 10 rpm for 65h at 65°C. The enriched fraction was recovered using the protocol of Hodges et al (2009) and Agilent, with a final recovered eluate volume of 462.5 µl. (We also recovered the post-enrichment "non-enriched" fraction by additionally performing the syringe protocol prior to slide washing).

Post-enrichment amplification. 450 µl of post-enriched eluate alongside PCR blanks was immediately amplified in 18x qPCR reactions using the re-amplification primers *IS5_reamp.P5* and *IS6_reamp.P7* (10) according to the following protocol: 400 nM each primer, 1X PCR Buffer II, 2.5 mM $MgCl_2$, 250 µM dNTPs, 0.05 U/µl Amplitaq Gold polymerase, 0.167X SYBRgreen fluorescent dye (Invitrogen), and 25 µl DNA template, in a 50 µl final PCR volume. A 95°C initial denaturation (4 min) followed by 20 cycles of 95°C (30s), 65°C (30s), and 72°C (30s) were performed. Reactions were pooled and purified over MinElute to a final volume of 135 µl (3090.13). qPCR quantitation of pre- and post-amplified eluate indicated a 7.7E04 fold increase due to post-enrichment amplification (1.0E04 c/µl vs. 7.7E08 calibrated c/µl, respectively). (The additional recovered non-enriched, non-sequenced fraction, as noted above, was 2.1E11 c/µl.)

2.11.8 Sequencing

The final 3090.13 post-enriched library sent for sequencing was 6.8 ng/µl (via Nanodrop; Thermo Scientific) and 14.5 nM (via Bioanalyzer; Agilent Technologies, Inc). Sequencing was performed across 5 lanes of Illumina GAIIx system by the Donnelly Sequencing Centre (University of Toronto). 72bp single read chemistry was used, but the total read length was extended to 80bp by the addition of 8bp worth of sequencing reagents to the sequencing read length (rather than being used towards an indexing read). Each lane yielded 34,019,321 to 34,821,855 total reads passing filter, in total yielding 172,138,676 raw reads passing filter. We

have deposited all raw sequencing reads at the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under the BioProject accession number SRP029921.

2.11.9 Shotgun Sequencing

For the purposes of comparing the overall difference between percent of *V. cholerae* DNA in the non-enriched vs. enriched state, we also subjected an additional aliquot of indexed library (from section [2.11.4] above) to re-amplification, and sequenced this non-enriched library on the Illumina platform. (It is noted that the number of cycles experienced by the enriched and non-enriched libraries are unequal, so the comparison between pre- and post-enriched percent *V. cholerae* DNA, as discussed in the main manuscript, is not ideal; however, this does not influence any of the main conclusions of this study.)

Pre-sequencing amplification. Prior to sequencing, additional indexing amplification was performed in 8 reactions (5 µl 0.1x diluted template DNA in 50 µl total reaction volume) of 3090.13 indexed library (see above) using 400nM each indexing primer and 11 cycles.

Sequencing. The purified 3090.13 library was pooled with an additional (un-related) sample in equimolar ratio on one lane of Illumina HiSeq 1000. Sequencing was performed by the Farncombe Family Digestive Health Research Institute (McMaster University). 100bp paired-end read chemistry was used, with one indexing read. The lane yielded 263,870,537 total raw reads passing filter, 141,039,627 of which belonged to sample 3090.13.

2.11.10 Authorship Contributions

Grants awarded to HNP funded the study. AMD, JME, MK, JHT, DNF, AND, DJDE, and HNP designed the study. AMD, JME, MK, JHT, and HNP gathered the data. AMD, GBG, NW, JME, MK, MS, DNF, SF, KIB, ECH, and HNP analyzed the data. All authors vouch for the data and analysis. AMD, JME, ECH, and HNP wrote the first draft and co-wrote the additional drafts of paper with the assistance of all other co-authors. AMD and HNP decided to publish the paper with the approval of all other co-authors.

2.12 Supplementary Analysis

2.12.1 Draft genome sequence

Trimming. Trimming was done using cutadapt v.1.0 (37) to eliminate adapters and to eliminate bases with a PHRED quality cutoff 20, a minimum overlap 3bp between read and

adapter, and a minimum length after trimming of 20bp. The qualities of the reads were double-checked using FASTQC.

Assemblies. Reads were mapped to *Vibrio cholerae* O395 (NC_009456, NC_009457) using BWA v.0.5.9rc1 (r1561) (11) at default settings. The resulting sam file was edited to retain entries with a minimum length of 20bp, a minimum of 98% of the read length matched perfectly, and a map quality of 20. These were then converted to bam files via samtools(38) and sorted, PCR duplicates were removed using samtools rmdup (-s), and samtools mpileup was used to generate a bcf file. The 'loose' assembly was done with default base quality and map quality (13 and 0, respectively) in mpileup. 'Strict' mappings were treated the same way but with map quality of 20. Fragment length distributions can be found in **Figures S2a & S2b**. Average coverage vs. %GC for the PA1849 genome is plotted in **Figure S3**.

Ancient DNA Damage Assessment. We subjected the post-trimmed reads to analysis using the MapDamage (v.0.3.6) program (18) (**Figures S4a & S4b**). Patterns of damage (C→T or G→A changes at the 5' and 3' fragment ends) typical to ancient DNA were seen, despite the differences in tissue, age, preservation method, etc. between this sample and other "typical" ancient DNA specimens.

SNP Calling. SNPs (n=203) were called from BCF files generated by samtools, using a minimum coverage of 5x, 90% variant frequency, a minimum quality PHRED value of 5 and a mapping quality of ≥ 20 (see above). We also generated alternate SNP datasets that show what SNPs would have been called by restricting base quality values to either minimum 10 (n=192) or 20 (n=147). All SNP calls can be found in [**Table S9**]. If by chance some reads of sufficient coverage mapped to locations that were within broader regions that were determined to be actually absent in PA1849 (e.g., GI-21), or within regions that were likely to be conserved within other species or could have similar repetitive regions within the cholera genome (e.g. 16S rRNA or tRNA loci), these were manually removed from the list of final SNPs. Several SNPs were also eliminated that relied on the ends of reads, as these were likely to contain damaged sites due to the nature of "ancient DNA" (see above).

Evolutionary Analysis. A complete genome sequence alignment of 32 strains of *V. cholerae*, including PA1849 (**Table S2**; map in **Figure S5**), was generated using MAUVE assuming default parameters. (Consensus sequences from any SRA database readsets were first generated via assembly to O395.) A SNP alignment was then generated from this complete genome alignment (see "SNP calling"). Because some regions of the complete genome, and hence SNP, alignment were uncertain, reflecting the occurrence of multiple indels, we used Gblocks (14) to purge all ambiguous regions, including all indel sites. This resulted in a final SNP

alignment of 28,591 nt that could be utilized for evolutionary analysis. Although we also undertook phylogenetic analyses with two environmental strains – 12129/Australia/1985 and LMA3984-4/Brazil/2007 – they were excluded from the final data set because of their highly divergent nature and hence likely extensive site saturation compared to PG1 and PG2 strains.

We first inferred a phylogenetic tree using the Maximum Likelihood (ML) approach available in the PhyML package (15). This analysis employed the GTR model of nucleotide substitution and a gamma distribution of among-site rate variation with four rate categories (G_4) (i.e. the GTR+G_4 nucleotide substitution model) and utilizing subtree pruning and regrafting (SPR) branch-swapping until the highest likelihood was obtained (log likelihood = - 153884.6). The final parameter values for the nucleotide substitution model were: GTR = A-C 0.962, A-G 5.548, A-T 0.749, C-G 0.609, C-T 5.591, G-T 1.000; I = 0; G_4 = infinity. To determine the robustness of individual nodes on this phylogeny, we performed a bootstrap resampling analysis, employing 1000 pseudo-replicate ML trees estimated using the same procedure as described above, optimizing parameters in each run. This tree is shown in **Figure 3** in the main text.

Next, we attempted to infer a time-scale for the evolutionary history of *V. cholerae* using the Bayesian Markov Chain Monte Carlo (MCMC) approach available in the BEAST package v.1.7.5) (39) and employing the sampling dates (i.e. years) of each sequence, which ranged from 1849-2010. Multiple runs were undertaken using both strict and relaxed (uncorrelated lognormal) molecular clocks, different nucleotide substitution models (i.e. HKY85 and GTR), and different tree priors (i.e. constant population size and Bayesian skyline). All runs were performed multiple times, with chain lengths varying from 100,000,000 to 1,000,000,000. However, the use of these time-informed (i.e. tip-dated) data to estimate rates of nucleotide substitution and times to common ancestry resulted in unstable values in all cases (i.e. poor ESS values, strongly autocorrelated traces), indicative of a lack of temporal signal in the data. A lack of temporal signal was also apparent in a regression of sampling time (year) against root-to-tip genetic distances on the ML tree performing using the Path-O-Gen program kindly provided by Andrew Rambaut, University of Edinburgh (http://tree.bio.ed.ac.uk/software/pathogen/); the correlation coefficient was very low (0.136), confirming a lack of temporal signal in the data.

The lack of temporal signal in the *V. cholerae* SNP data necessitated that we set (i.e. 'fixed'), rather than estimate, a nucleotide substitution rate. To this end we conducted a BEAST analysis of SNP data from 122 El Tor strains (sequence alignment of 1757 bp; data kindly provided by Ankur Mutreja, Wellcome Trust Sanger Institute, UK) sampled between 1937-2010. These data exhibited a strong temporal signal in a root-to-tip regression (correlation coefficient = 0.863) suggesting that they provide a robust estimate of substitution rate, at least over the

time span of sampling. Accordingly, a BEAST analysis (strict molecular clock) of these data produced an evolutionary rate of 1.3×10^{-3} nucleotide substitutions per SNP site, per year (subs/SNP site/year) (and all parameters converged after chains of 100,000,000 and with good mixing). This rate was then to estimate the time-scale of evolution of (i) the 6 El Tor strains, and (ii) the 21 classical strains, including PA1849, in our 32 sequence data set (i.e. we assume that the El Tor and classical strains have evolved at the same rate). However, because of a combination of site saturation on deep branches (note the scale bar in Figure 3) and recombination (see below) we were unable to obtain robust estimates for the age of deep nodes on the *V. cholerae* phylogeny, including the common ancestor of PG1 and PG2 (i.e. node PG).

Recombination Analysis. To identify potential recombinant regions and their evolutionary origin, we utilized seven recombination detection methods implemented in the RDP4 package; RDP, GENECONV, BOOTSCAN, MAXCHI, CHIMAERA, SISCAN, and 3SEQ (40). Recombination events significantly detected by more than two methods (bonferroni corrected *P* values < 0.05) with default parameters were counted as *bona fide*, and those sharing the same recombination pattern were merged and treated as a single recombination event. We then summarized the occurrences of recombination between lineages, including between the classical and El Tor strains, to produce a final count of the number recombination events across the data set as whole. Note that in some cases the recombinant parent (i.e. donor sequence) was unidentified, and high levels of sequence similarity made it difficult to accurately infer the number of recombination events within the classical and El Tor groups.

To determine whether recombination has compromised our attempts to reconstruct the phylogenetic history and time-scale of *V. cholerae* evolution we repeated the analyses described in section (v) on a data set in which the recombination events defined using RDP4 described above were coded as gaps (-); this allowed us to retain a sequence alignment of the same length as the original (28,591 nt; thereby maximizing phylogenetic resolution) but with all recombinant regions removed. This analysis resulted in similar phylogenetic trees to that obtained with recombinant regions included in the alignment, particularly with respect to the nodes of interest; ML and maximum parsimony trees are shown in **Figures S6 and S7**, respectively. Similarly, to assess the affect of recombination on our molecular clock dating we repeated the BEAST analysis above (with an evolutionary rate set to 1.3×10^{-3} subs/SNP site/year) on the recombination-free data set. Strikingly similar times to the Most Recent Common Ancestor (tMRCA) of the whole tree (i.e. node PG) were obtained both with (430-440 years ago) and without (468-487 years ago) recombinant regions, suggesting that site saturation is of more evolutionary importance on these data than recombination. Finally, to obtain an approximate estimate of the number of recombinant sites in our SNP alignment we

estimated a number of parsimony-based statistics on the ML tree (with all recombinant regions included). These metrics are as follows: (i) Tree length = 30954; (ii) Consistency index (CI) = 0.9253; (iii) Homoplasy index (HI) = 0.0747; (iv) CI excluding uninformative characters = 0.8744; (v) HI excluding uninformative characters = 0.1256; (vi) Retention index (RI) = 0.9811; and (vii) Rescaled consistency index (RC) = 0.9079. Hence, these results indicate that there is relatively little recombination in these data relative to divergent evolution (i.e. 92.5% of SNPs are consistent with the ML tree and 7.5% of SNPs are homoplasic, perhaps due to recombination).

2.11.2 Other Loci Included on the Array: Non-classical V. cholerae and Human

Trimming. The 172,138,676 raw sequenced reads were trimmed to remove any remaining adaptor sequence using cutadapt v.1 (37) with the settings: error rate (0.16), minimum overlap (1).

Assemblies. Assemblies to the reference sequences included on the array were done in Geneious (v.5.6.5; Biomatters, Ltd.) using "medium" assembly parameters for the non-classical cholera loci (maximum gaps per read = 15%, maximum gap size = 50, word length = 14, index word length = 12, ignore words repeated more than 10x, maximum mismatches per read = 30%, maximum ambiguity = 4), and "low" assembly parameters for the human loci (maximum gaps per read = 10%, maximum gap size = 3, word length = 24, index word length = 14, ignore words repeated more than 1x, maximum mismatches per read = 10%, maximum ambiguity = 4). Assemblies were collapsed for PCR duplicate reads using the samtools rmdup (-s) program (38).

Ancient DNA Damage Assessment. We subjected the post-trimmed human mitochondrial genome reads to analysis using the MapDamage v.0.3.6 tool (18) (**Figure S4c**). Patterns of damage (C→T or G→A changes at the 5' and 3' fragment ends) typical to ancient DNA were seen, despite the differences in tissue, age, preservation method, etc. between this sample and other "typical" ancient DNA specimens.

2.11.3 Other Analyses

Assessment of GIs Present in Strain RC27. The downloaded contigs of the scaffold genome of strain RC27 (accession NZ_ADAI00000000.1) were aligned to the O395 chromosomes (NC_009456, NC_009457) using Geneious (5.6.5) highest sensitivity assembly parameters and allowing for mapping multiple best matches. **Table S7** shows the results for the PG-2 GIs of interest. Note that we did NOT investigate for the presence of any non-O395 genomic islands of interest, so it is possible/likely that RC27 does harbor additional GIs found in other strains.

2.13 Supplementary Results

2.13.1 PA1849 vs O395

CTX Prophages. Reads support a possible tandem configuration of CTX elements on one OR both chromosomes (**Figure S8e**), as well as the immediate (**Figure S8a-d**) and more distant (not shown) chromosomal flanking regions, which demonstrate that the locations of the CTX element(s) are the same in PA1849 as in modern O395. In **Figure S8e**, only those reads not clearly supporting the other arrangements (as in **S8a-d**) are shown assembled to the CTX-CTX tandem span region from the B33 strain (accession GQ485644) (41). However, due to short fragment lengths in the historical specimen, we are unable to determine the exact configuration of CTX elements (**Figure S9**): a double tandem CTX prophage may exist on just one or both chromosomes (or, if not actually in the genome, in plasmid form).

Additionally, we have no evidence for the "truncated" CTX prophage that is seen adjacent to the full prophage on the large chromosome in modern classical strains. Instead, reads that map to this region only map to the immediate 5'/3' sequences that would correspond to a single full prophage (**Figure S10**).

"Superintegron". Due to the repetitive nature of the integron cassettes, it is not possible to completely reconstruct such a locus with an ancient sample, due to the highly fragmented nature of the DNA molecules (as short fragment lengths precludes specific assignment of reads falling into repeat regions). While it is therefore not possible to reliably reconstruct the PA1849 integron gene order, we have here reported on the genic content of the integron vs. O395. **Table S8** highlights the PA1849 coverage, O395 %GC, and percentage of the O395 reference sequence covered by at least one read for many superintegron genes of interest: those gene cluster (i.e., groups of homologous genes) that form the "core" integron genome and those genes unique to O395, as outlined in reference (42). Overall the gene content of the integron appears very similar to O395, although as noted previously, PA1849 is missing GI-14 (a large section of the integron). All the "core" genes are apparently present, as are most of the O395 genes. Several (non-GI-14) O395-specific genes MAY be missing due to lower average coverage and percent of the reference covered than the genomic average; however, %GC may be contributing to these lower statistics, as most of these are slightly lower than average %GC (these genes of unknown presence are marked "Y?" in the **Table S8** "confirmed?" column).

CRISPR Region (GI-24). Overall, the CRISPR genic region (falling in GI-24) appears to be largely similar to O395, with differences in the 'spacer' region. The CAS/CSE genic content is the same and appears the same order as O395 (*cas3', cse1, cse2, cas6e, cas7, cas5, cas1,* and *cas2*, followed by the repeat/spacer region) (43, 44). The repeat/spacer region has a visibly punctuated coverage pattern that differs from the rest of the island due to the lack of coverage at the majority of the CRISPR spacers, indicating that PA1849 likely had different spacer content than O395. Unfortunately, it is not possible to reconstruct the content with these types of data, due to short fragment lengths and lack of proper enrichment targets. **Figure S11** visually demonstrates the coverage pattern across the CRISPR repeat/spacer region.

Other Loci. Comparing other genes of interest between PA1849 and modern classical indicate that the strains are extremely similar (45). The PA1849 strain possesses the classical (deletion) variant of *hlyA,* hemolysin at all covered positions. The hemagglutinin protease (Hap) sequence is identical to the O395 reference at all covered positions. The mannose-sensitive hemagglutinin (MSHA) genes (such as *mshA*) are present and identical to O395 at all covered positions, except for 1 SNP in *mshN* (2,973,101). Regarding the antibiotic resistance element SXT, we can report that (as expected) there is no evidence of its presence because the reads falling across the SXT integration site at the 5' end of *prfC* (peptide chain release factor 3) (46) are identical to the O395 reference sequence, which does not contain SXT. SXT is currently only found among recent El Tor and related strains, and was not included as one of the several "non-classical" loci on this enrichment array (which was not intended to be an exhaustive list of all possible 'pan-genomic' cholera content).

2.13.2 PA1849 vs. Other Strains (Regions Not Expected to Occur in the Classical Strain)

All regions returned the expected outcomes, which were: low coverage and percent of the reference covered for those regions absent from the classical genome; the correct (classical) variant for those regions with significant divergence to classical (see **Table S4**).

2.13.3 Human Loci

Mitochondrial Genome. The assembly relative to the revised Cambridge Reference Sequence yielded 37 SNPs relative to the reference, with high unique coverage and variant frequencies. Mitotool.org returns the haplogroup as L3d1b3, which is consistent with the L3d1b3 SNPs recorded at phylotree.org (both accessed 2012). Coverage information can be found in **Table S5**; SNP calls can be found in **Table S6**.

Amelogenin Genes. The coverage of reads across the human AMELX and AMELY genes was highly variable (see **Table S5**). Low coverage regions tend to be low in GC% (see **Figure S12**). Due to the extreme coverage disparities and the likely inclusion of many non-amelogenin reads in this assembly (at conserved regions), it is not possible to directly reconstruct the sequence of these genes. However, several reads BLAST exclusively to the human AMELY gene (BLASTN; default parameters), suggesting that the individual was indeed a male, though more work would be necessary to be conclusive.

2.14 Supplementary Figures

Figure S1. qPCR amplification curves for 58bp *ompW* assay

One replicate of each standard of known copy number (10K, 1K, 100, and 10) are shown in black. PCR blanks are shown in grey. Library amplifications (1 in 100 dilution; 0.01x) are shown in green (3090.13 in dark green, all other blanks in light green). Indexing amplifications (0.01x) are shown in blue (3090.13 in light blue, all other blanks in very light blue). Re-amplification blanks (0.01x) are shown in violet. The threshold RFU (relative fluorescence units) level to calculate Cq threshold was automatically determined at 69.01 (BioRad CFX Manager 3.0 software). Cq values shown below for the positive template amplifications were as follows: 3090.13 library (0.01x) replicates (35.35, 36.29), 3090.13 indexed library (0.01x) replicates (35.58, 35.75). All other templates tested (blanks) were negative for the 58bp *ompW* product after 50 cycles.

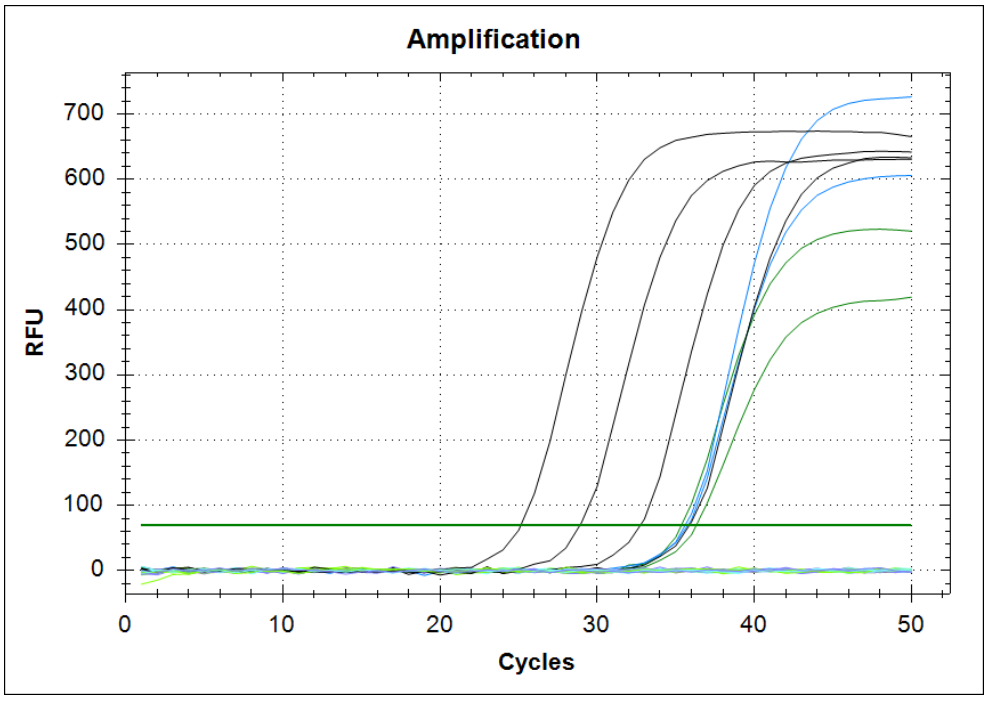


Figure S2. Fragment length distributions (FLDs)

The *V. cholerae* chromosomal FLDs reflect all unique reads. The human mitochondrial FLD reflects the 14,731 collapsed unique reads (samtools rmdup) from a 33,000 random raw read subset (out of 19.1 million total raw reads).

Figure S2a. FLD: V. cholerae large chromosome

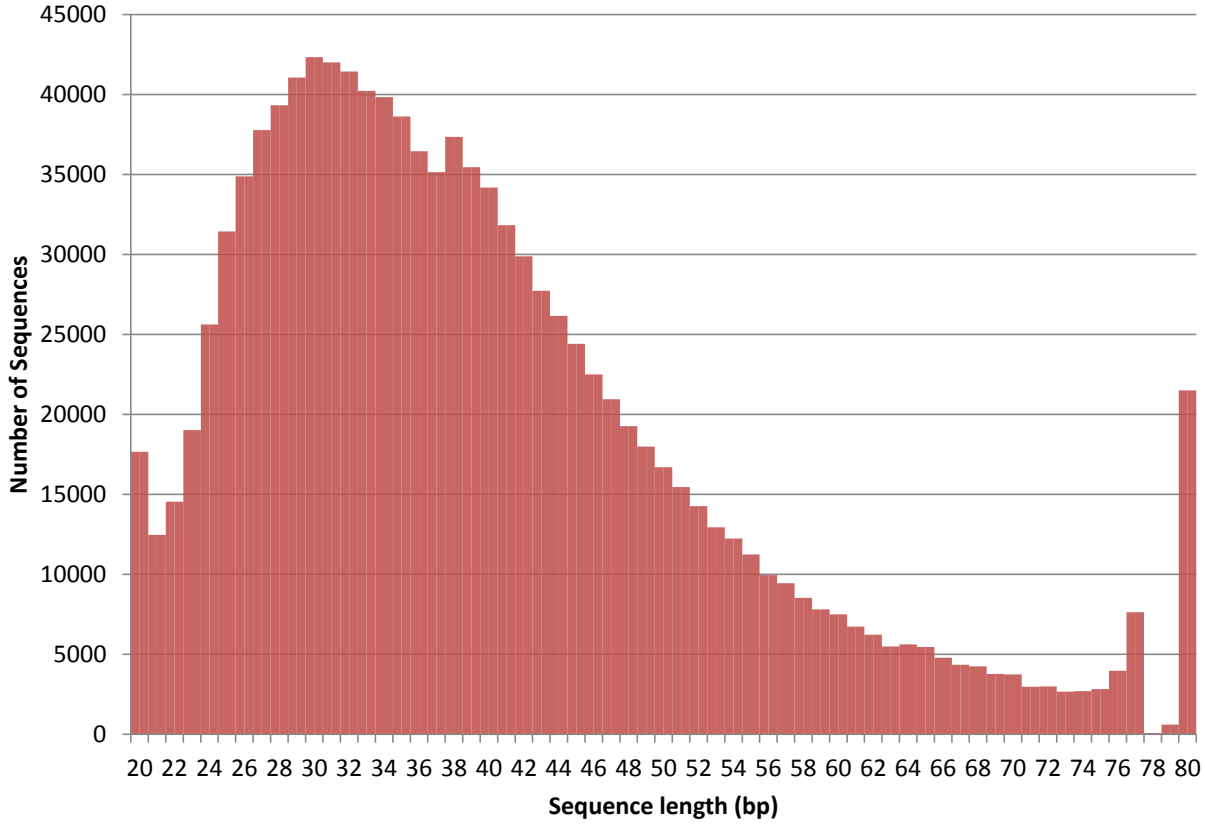


Figure S2b. FLD: V. cholerae small chromosome

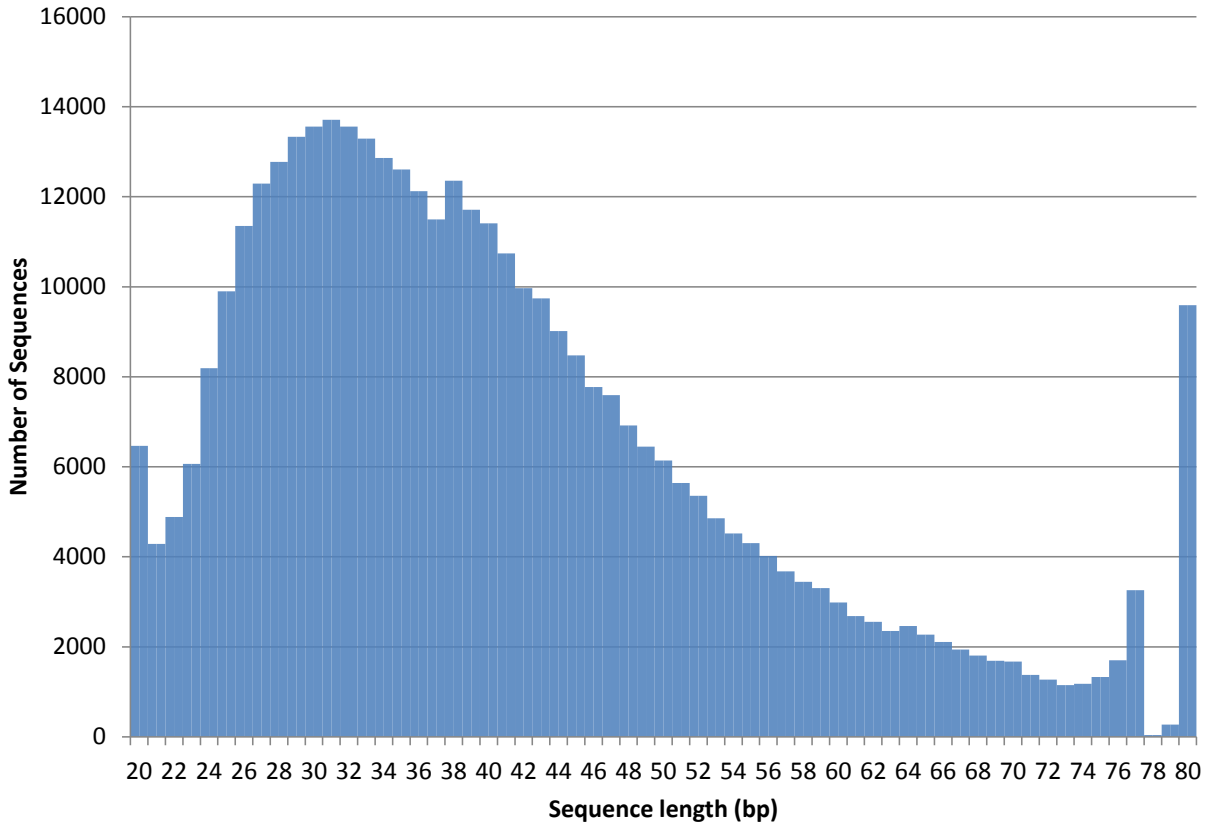


Figure S2c. FLD: Human mtDNA

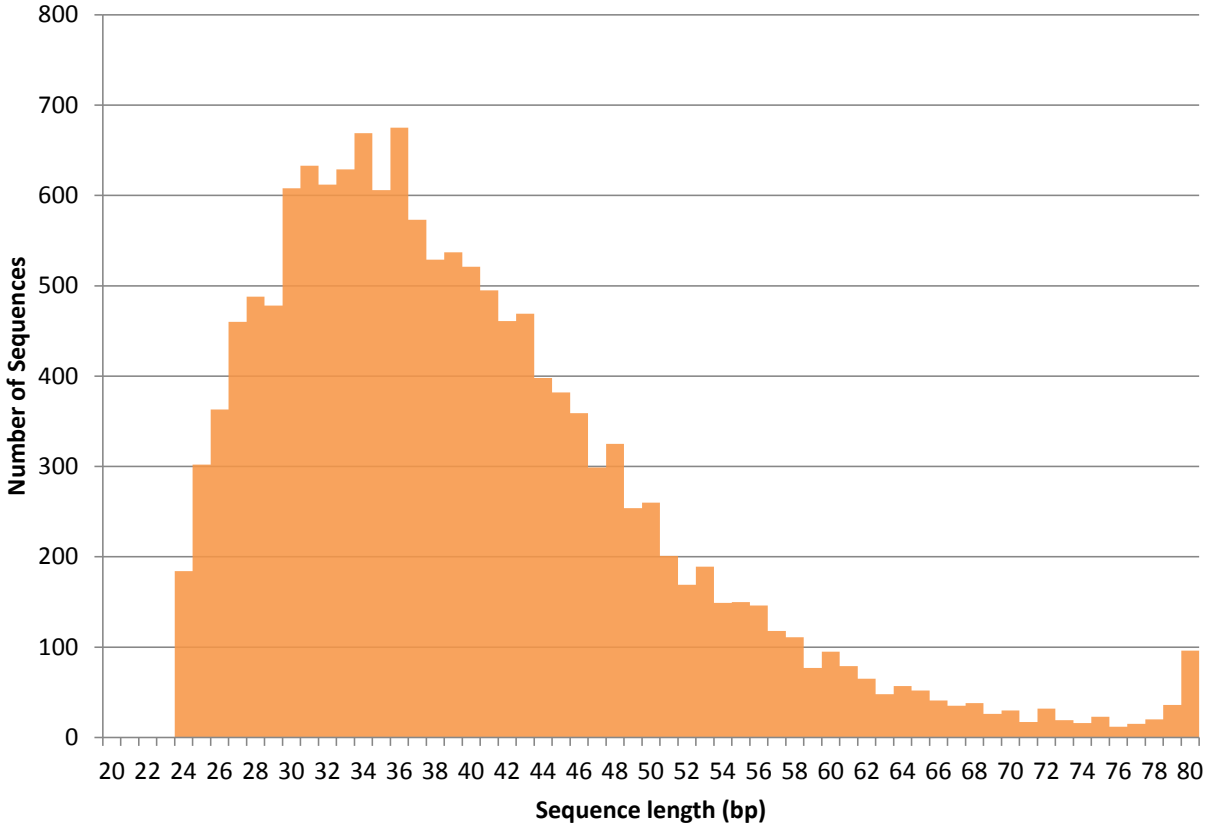


Figure S3. PA1849 coverage vs. GC% (100bp window)

Plot of average unique PA1849 coverage vs GC% of the O395 reference genome across 10bp consecutive windows. To reduce complexity, only every 10th data point is considered here. Coverage axis is scaled logarithmically. The logarithmic trendline and equation of best fit is shown (R^2 = 0.25).

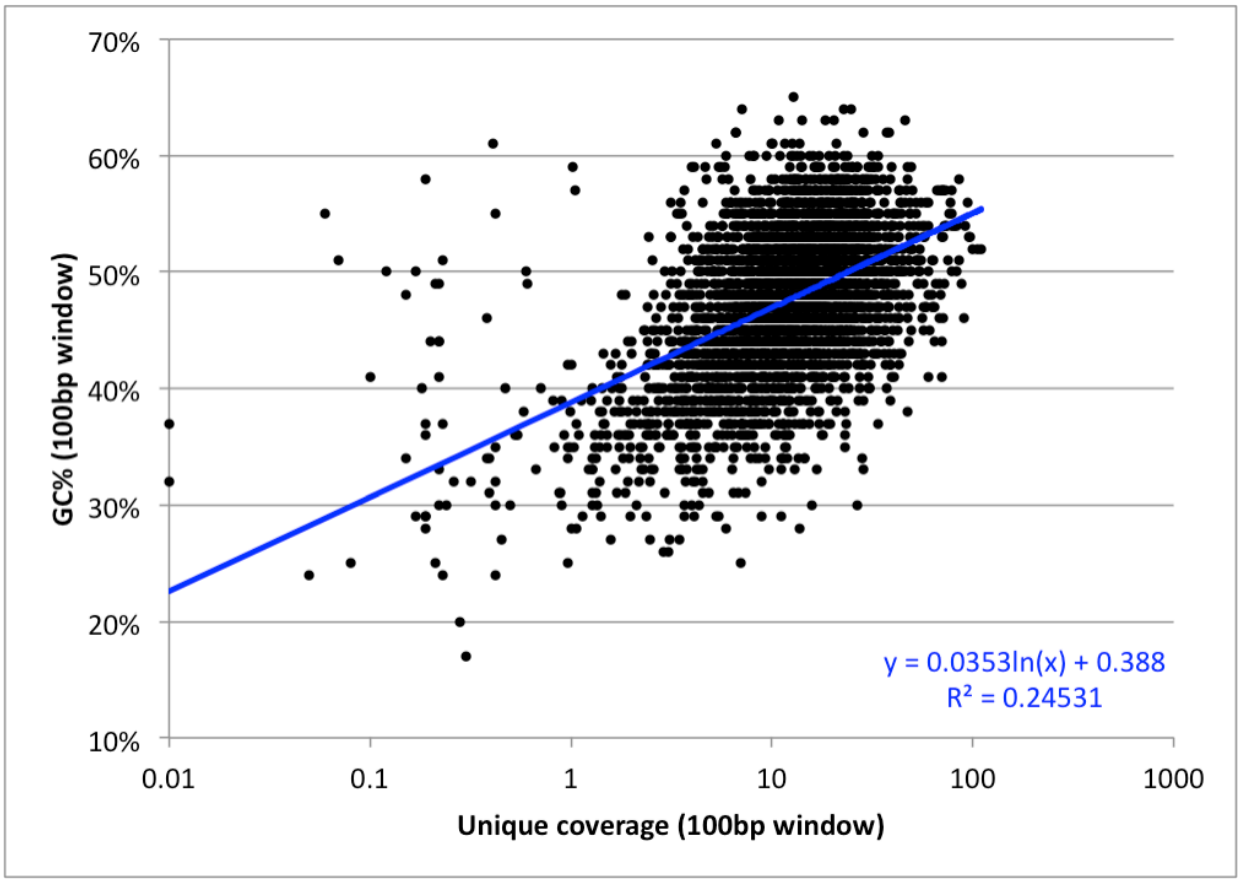


Figure S4. mapDamage graphs

The *V. cholerae* chromosomal mapDamage profiles reflect all unique reads. As in Figure S2, the human mitochondrial mapDamage profile reflects the 14,731 collapsed unique reads (samtools rmdup) from a 33,000 random raw read subset (out of 19.1 million total raw reads).

Figure 4a. mapDamage: V. cholerae large chromosome

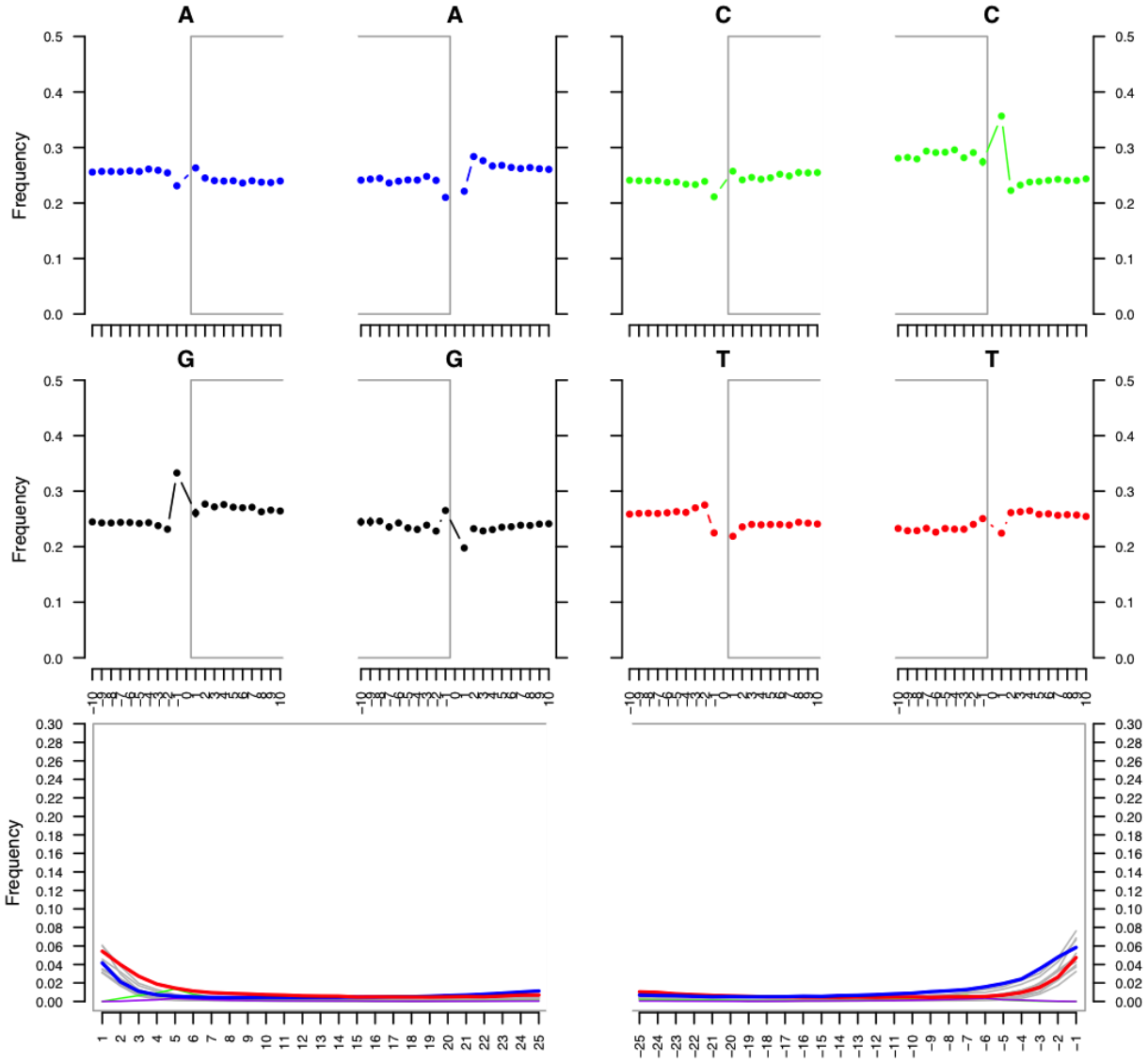


Figure S4b. mapDamage: V. cholerae small chromosome

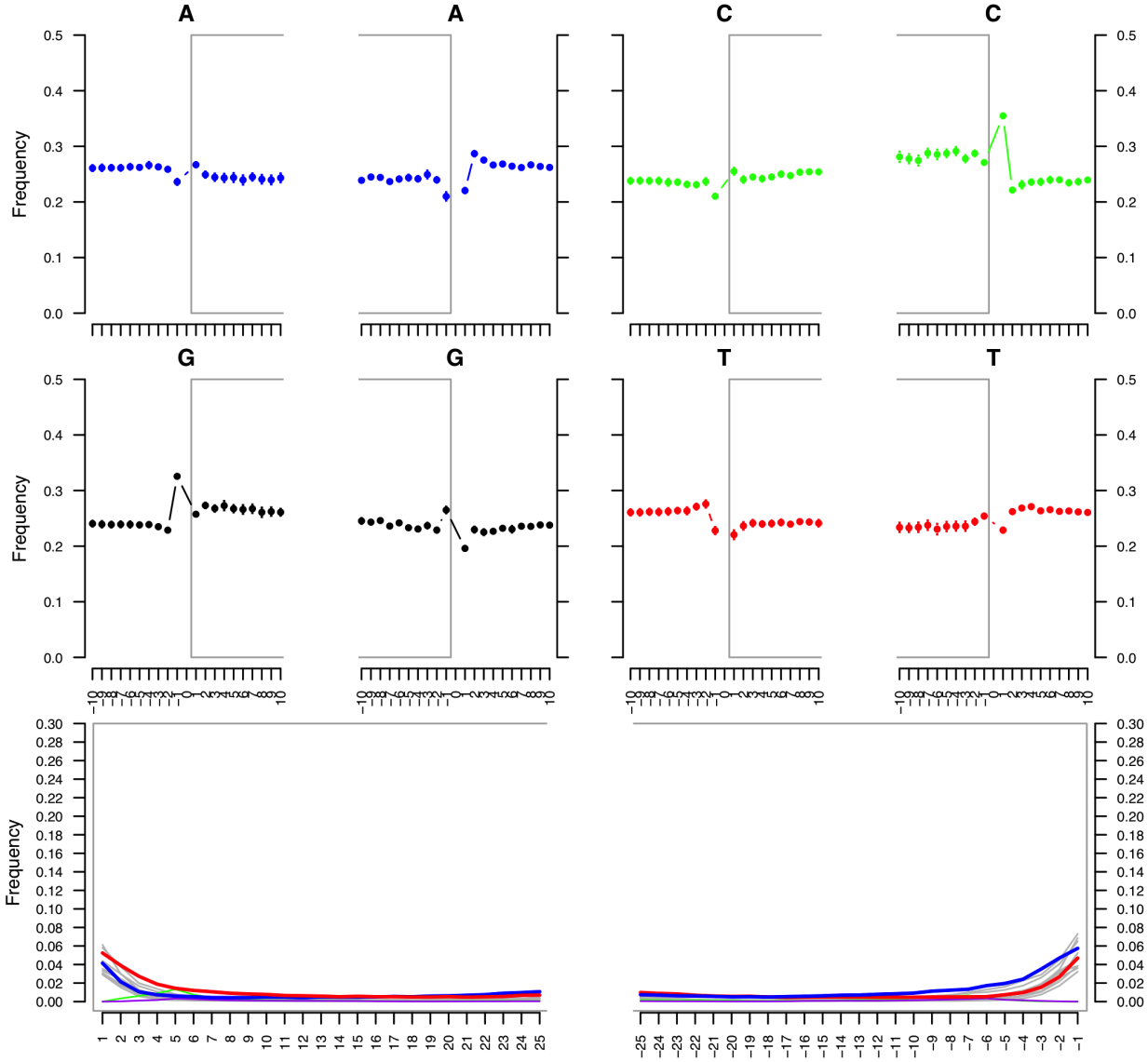


Figure S4c. mapDamage: Human mtDNA.

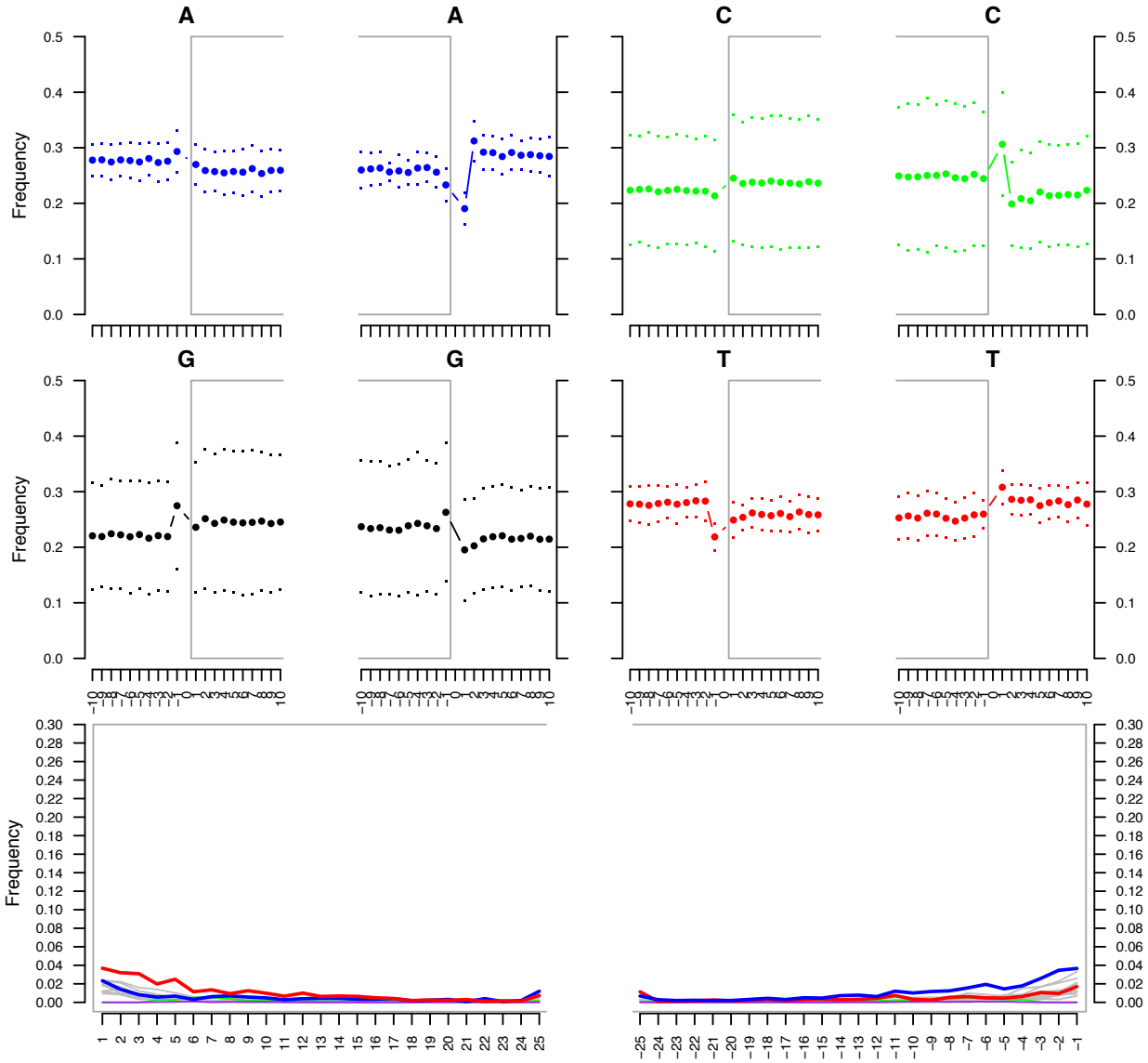


Figure S5. Map of known geographic strain locations

Map of all known geographic locations of strains used in the evolutionary analysis, grouped by geographic region. (Note that some of the strains used for the phylogenetic analysis were of unknown provenience.) Size of the nodes corresponds to number of samples from that region (from west to east: Pennsylvania USA n=1, southern North America n =2, Brazil n=1, Sweden n=1, Red Sea region n =4, Bay of Bengal region n = 14, Southeast Asia n = 5). Colors correspond to the clade colors used in the phylogenetic tree in the main text, **Figure 3**.

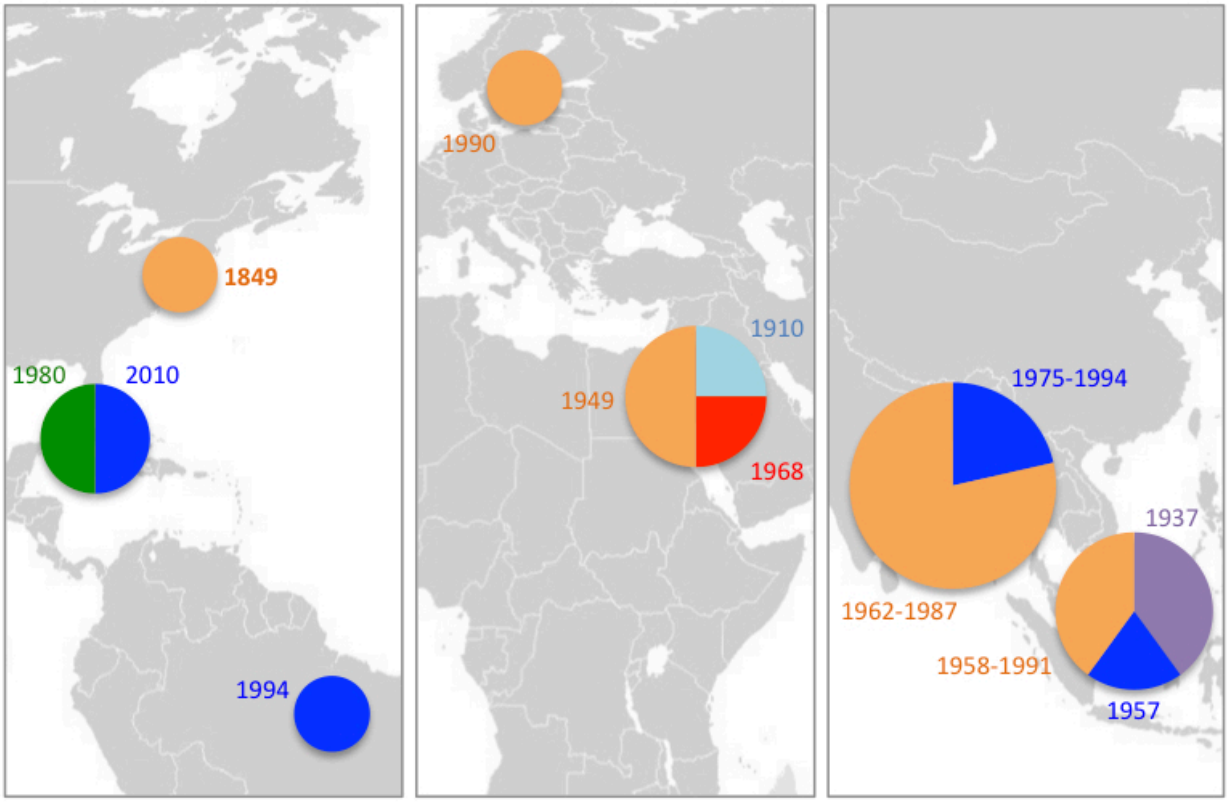


Figure S6. Maximum likelihood (ML) phylogeny of 32 strains of *V. cholerae* excluding recombinant regions.

This tree is the same as that shown in **Figure 3** in the main text, with the exception that recombinant regions (detected by RDP4; see above) have been excluded. The tree is mid-point rooted for clarity and all horizontal branch lengths are drawn to a scale of nucleotide substitutions per SNP site.

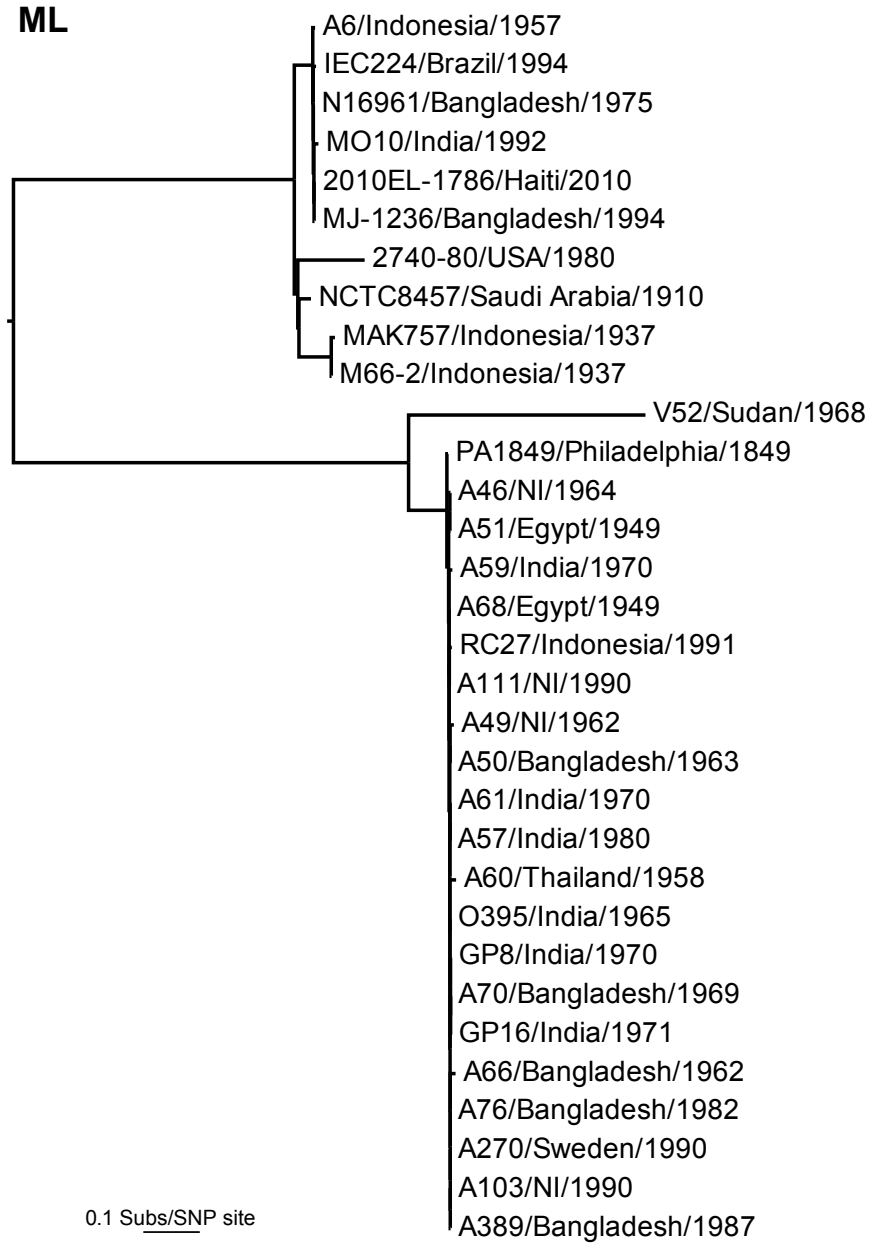


Figure S7. Maximum parsimony phylogeny of 32 strains of *V. cholerae* excluding recombinant regions.

To improve visual resolution of the branching order, branch lengths are not drawn to scale. The tree is mid-point rooted for clarity only.

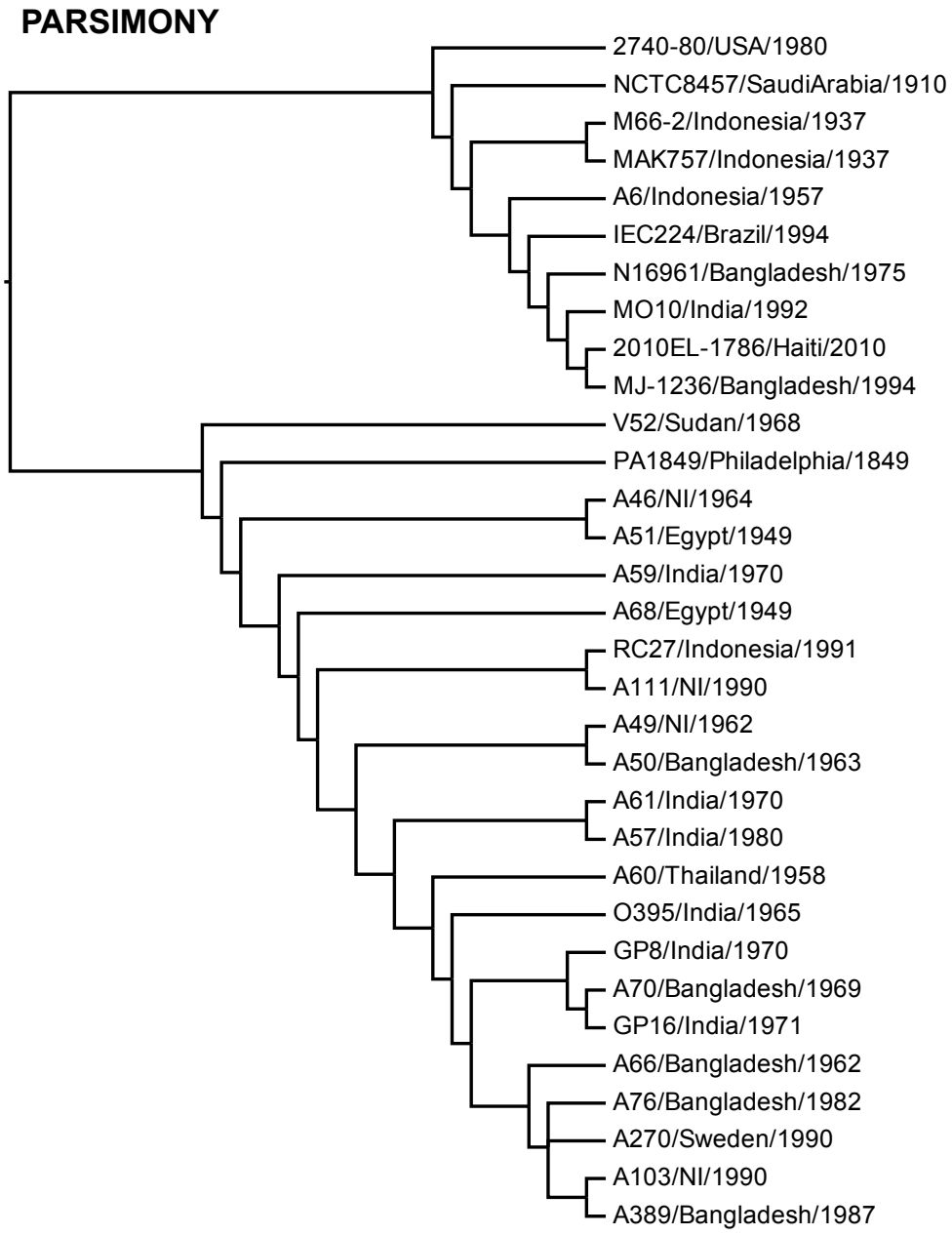


Figure S8. Assemblies supporting CTX prophage arrangements

Assemblies showing the "upstream" and "downstream" edges of CTX prophage on both chromosomes (vs. strain O395; **a-d**) and possible PA1849 tandem CTX arrangement (vs. strain B33, accession GQ485644; **e**). CTX orientation is shown as in O395 small chromosome. Both chromosomes share the same immediate "downstream" flanking at this scale (**c & d**). Red box highlights the 'ATTA' motif at the CTX edges in each assembly. Assembly, graphs, and annotations visualized using Geneious (v.6.1.6, Biomatters, Ltd).

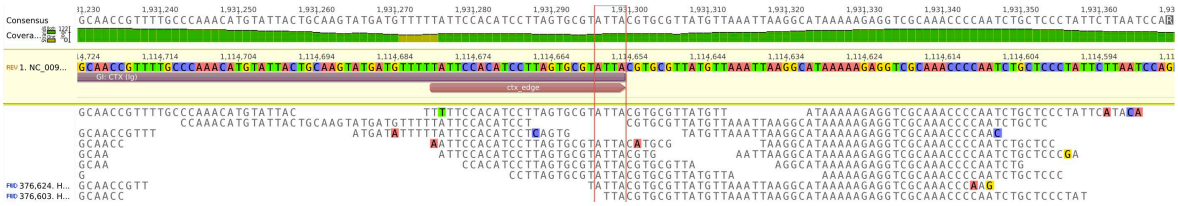
S8a. O395 Large chromosome – "upstream"



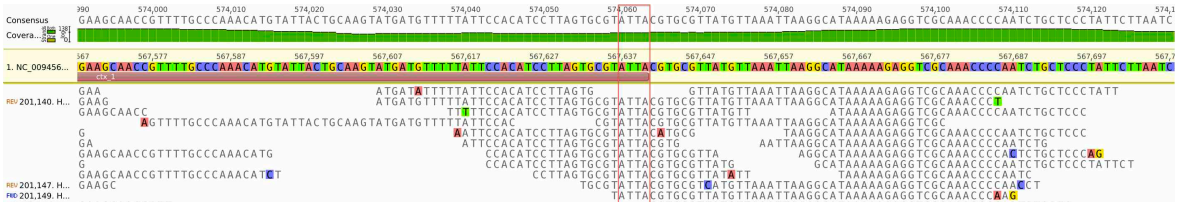
S8b. O395 Small chromosome - upstream



S8c. O395 Large chromosome – "downstream"



S8d. O395 Small chromosome – downstream



S8e. Putative tandem CTX arrangement

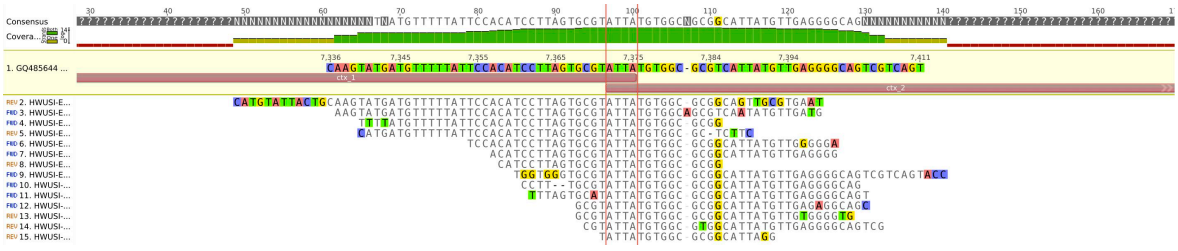


Figure S9. Possible PA1849 CTX region structural configurations

O395 CTX region arrangement has full and truncated CTX prophage sequences on the large chromosome adjacent to the TLC element, and a full CTX prophage on the small chromosome. Based on read data and assemblies (see text and Figure S7), PA1849 has at least a CTX prophage tandem arrangement on one or both chromosomes, and no evidence of a truncated/full CTX prophage structural arrangement. One of three possible chromosomal CTX arrangements exists: a single CTX on the large and tandem on the small, tandem on the large and single on the small, or tandem on both chromosomes. (It is also possible (not shown) that the tandem orientation exists only in plasmid format.) Because of the short fragment lengths in the historical specimen, we are unable to exactly determine the original configuration.

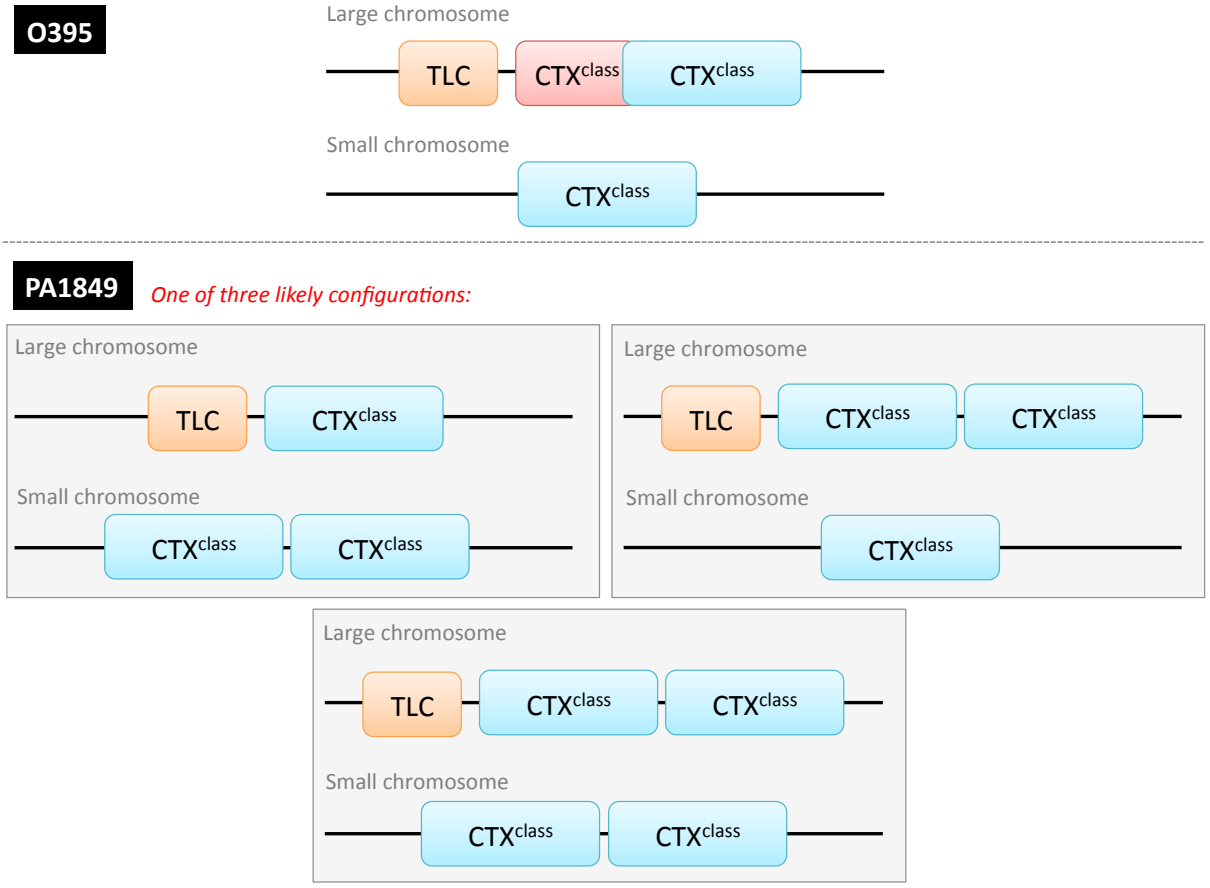


Figure S10. No evidence for truncated CTX prophage

All unique PA1849 reads mapping to the 'duplicated' rst region (truncated CTX prophage adjacent to full CTX prophage) on the large chromosome of O395 are shown. A red box outlines the overlap 'A' between rst regions. No reads span this intersection, indicating that PA1849 lacks the truncated CTX prophage. Assembly, graphs, and annotations visualized using Geneious (v.6.1.6, Biomatters, Ltd).

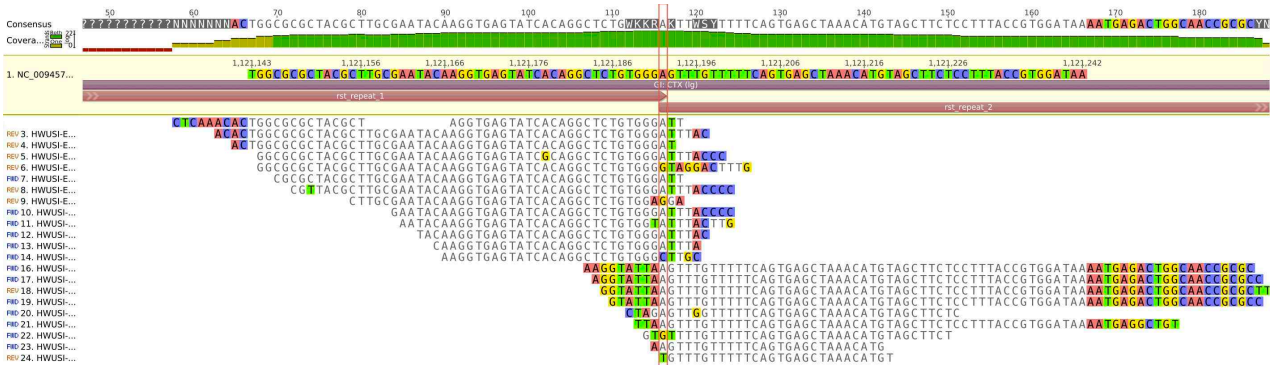


Figure S11. PA1849 coverage across the O395 CRISPR region of GI-24.

Graphical representation of PA1849 reads assembled to the O395 GI-24 CRISPR region, highlighting (left to right) genes *cas3', cse1, cse2, cas6e, cas7, cas5, cas1, cas2*, and repeat-spacer region (gene terminology from reference (44)). Coverage of the repeat-spacer region is high in the repeats but almost zero in the spacers, indicating that PA1849 does not contain similar spacer content to O395 (with a few exceptions). Assembly, graphs, and annotations visualized using Geneious (v.6.1.6, Biomatters, Ltd).

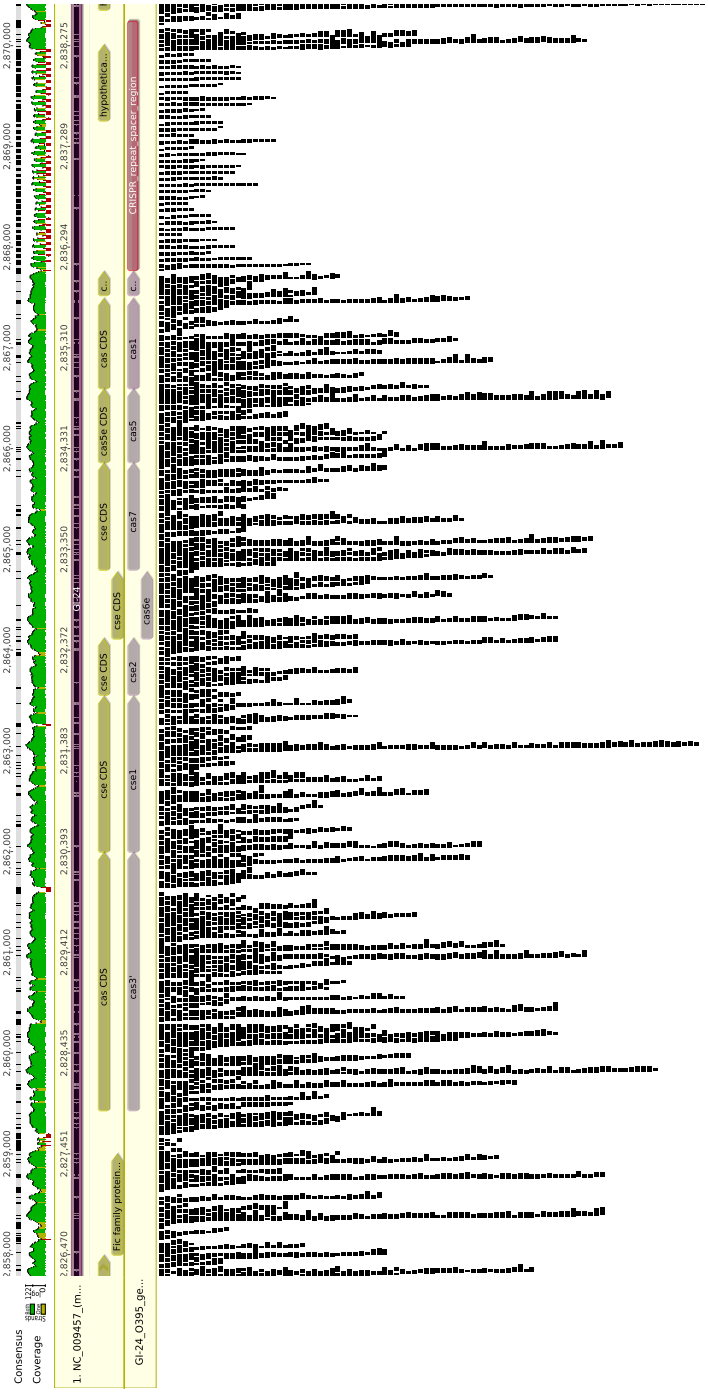
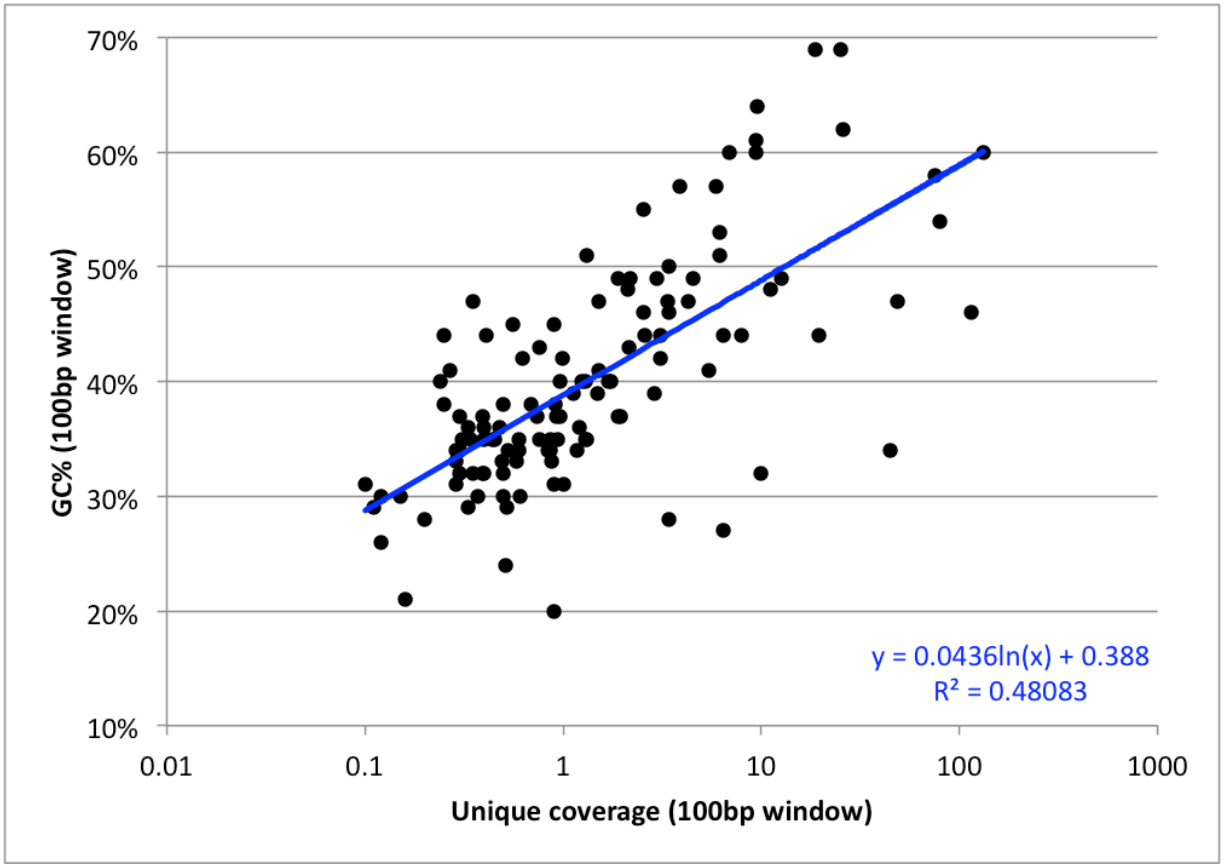


Figure S12. Human amelogenin X and Y coverage vs. GC% (100bp window)

Plot of average unique 3090.13 coverage across the human AMELX and AMELY genes vs GC% of the reference sequence across 100bp consecutive windows. Coverage axis is scaled logarithmically. The logarithmic trendline and equation of best fit is shown (R^2 = 0.48).



2.15 Supplementary Tables

Table S1. Sequences included on enrichment array

#	Accession	Description	Length	Position	Tiling	# probes	Ref
		CHOLERA - EXPECTED					
1	NC_009457	"Vibrio cholerae O395 chromosome 2, complete sequence" (large)	3,024,069	-	5	603,469	
		recA	1265	65,378-66,642	5 (2 offset)	242	
		RTX	35735	1,103,698-1,139,432	5 (2 offset)	7,136	(47)
		VPI	42290	363,273-405,562	5 (2 offset)	8,447	(48)
		VPI2	58540	1,448,788-1,507,327	5 (2 offset)	11,697	(49)
2	NC_009456	"Vibrio cholerae O395 chromosome 1, complete sequence" (small)	1,108,250	-	5	220,898	
		CTX	7540	560,458-567,997	5 (2 offset)	1,497	(19)
		hlyA	1705	986,188-987,892	5 (2 offset)	330	
		Integron (partial)	5830	911,858-917,687	5 (2 offset)	1,155	
		ompW	855	404,753-405,607	5 (2 offset)	160	
		CHOLERA - NOT EXPECTED or DIFFERENT VARIANTS					
	NC_002505	Vibrio cholerae O1 biovar El Tor str. N16961 chromosome I, complete genome	-	-	-	-	
3		VSP-I	16,946	173,969-190,914	2	8,443	(50)
4		RS1 prophage	3,182	1,563,785-1,566,966	2	1,561	(51)
5		recA (El Tor variant)	1,239	574,522-575,760	2	591	
6		VC0514-VC0516	5,048	545,074-550,121	2	2,494	(50)
7		tcpA (El Tor variant)	675	890,449-891,123	2	309	
8		VSP-II	27,790	522,397-550,186	2	13,865	(50)
9		rtxC (VC1449-VC1450)	1,190	1,548,919-1,550,108	2	565	(50)
	NC_002506	N16961 El Tor chr 2 (sm)	-	-	-	-	
10		hlyA (El Tor variant)	60	238,297-238,356	2	1	
11		VCA0300	830	315,211-316,040	2	385	(50)
12	AB012956	O-antigen synthesis, strain MO45 [O139]	35,806	-	2	17,873	(52)
		HUMAN – EXPECTED and/or UNKNOWN					
13	RefSeq accession NG_012040	Homo sapiens amelogenin, X-linked (AMELX), RefSeqGene on chromosome X	8,059	4,681-12,739	4	2001	
14	RefSeq accession NG_008011	Homo sapiens amelogenin, Y-linked (AMELY), RefSeqGene on chromosome Y	8,759	4,737-13,495	4	2176	
15	NC_012920	Homo sapiens mitochondrion, complete genome	16,569	-	4	4129	(53)
		OTHER - NOT DISCUSSED					
16	-	Additional non-project probeset	48,517	-	-	13,729	

Table S2. Genomes used for phylogenetic comparison

Strain	Description	Date	Location	Accession (lg, sm chr)	Reference
NCTC 8457	O1 Inaba - El Tor, nonpandemic	1910	Saudi Arabia	NZ_AAWD00000000.1	NCBI
M66-2	O1, pre-7th pandemic	1937	Indonesia	NC_012578.1, NC_012580.1	(28)
MAK 757	O1 Ogawa - El Tor	1937	Indonesia	NZ_AAUS00000000.2	NCBI
A51	O1 Ogawa - classical (Cairo 50)	1949	Egypt	ERS013165 [SRA dataset]	(21)
A68	O1 Inaba - classical (Cairo 48)	1949	Egypt	ERS013171 [SRA dataset]	(21)
A6	O1 – El Tor	1957	Indonesia	ERS013246 [SRA dataset]	(21)
A60	O1 Inaba - classical	1958	Thailand	ERS013168 [SRA dataset]	(21)
A49	O1 Inaba - classical	1962	unknown	ERS013161 [SRA dataset]	(21)
A66	O1 Inaba - classical	1962	Bangladesh	ERS013170 [SRA dataset]	(21)
A50	O1 Ogawa - classical	1963	Bangladesh	ERS013164 [SRA dataset]	(21)
A46	O1 Ogawa - classical	1964	unknown	ERS013160 [SRA dataset]	(21)
O395	O1 Ogawa - classical	1965	India	NC_009457.1, NC_009456.1	NCBI
V52	O37, clinical	1968	Sudan	NZ_AAKJ00000000.2	NCBI
A70	O1 Inaba - classical (G28190)	1969	Bangladesh	ERS013162 [SRA dataset]	(21)
A59	O1 Inaba - classical	1970	India	ERS013167 [SRA dataset]	(21)
A61	O1 Inaba - classical	1970	India	ERS013169 [SRA dataset]	(21)
GP8	O1 Inaba - classical	1970	India	ERS013128 [SRA dataset]	(21)
GP16	O1 Inaba - classical	1971	India	ERS013136 [SRA dataset]	(21)
N16961	O1 Inaba - El Tor	1975	Bangladesh	NC_002505.1, NC_002506.1	(54)
2740-80	O1 Inaba - El Tor, environmental	1980	USA	NZ_AAUT00000000.1	NCBI
A57	O1 Ogawa - classical (U10198)	1980	India	ERS013166 [SRA dataset]	(21)
A76	O1 Inaba - classical (X19850)	1982	Bangladesh	ERS013163 [SRA dataset]	(21)
A389	O1 Inaba - classical (VM11647)	1987	Bangladesh	ERS013203 [SRA dataset]	(21)
A111	O1 Inaba - classical (V591)	1990	unknown	ERS013176 [SRA dataset]	(21)
A279	O1 Inaba - classical (K216/92)	1990	Sweden	ERS013197 [SRA dataset]	(21)
A103	O1 Inaba - classical (V584)	1990	unknown	ERS013172 [SRA dataset]	(21)
RC27	O1 - classical	1991	Indonesia	NZ_ADAI00000000.1	NCBI
MO10	O139	1992	India	NZ_AAKF00000000.3	NCBI
MJ-1236	O1 Inaba - El Tor (Matlab variant)	1994	Bangladesh	NC_012668.1, NC_012667.1	(6)
IEC224	O1 - El Tor	1994	Brazil	NC_016944.1, NC_016945.1	(55)
2010EL-1786	O1 - El Tor	2010	Haiti	CP003069.1, CP003070.1	(56)

Table S3. Results of genomic islands of interest

Class	#§	Chr	Description§	O395 position	%GC	Avg unique coverage	% reference covered @1X	Confirmed?
"PG"§	O1	lg	O1-antigen region	2,771,546-2,795,569 #	40.4	10.0	87.10%	Y
	GI-1	lg	Motility and chemotaxis	1,043,291-1,055,071	43.4	14.4	97.80%	Y
	GI-2	lg	Oxidative stress response	1,248,952-1,252,534	44.1	12.9	93.00%	Y
	GI-3	lg	Membrane proteins	1,435,685-1,440,550	44.3	17.3	98.80%	Y
	GI-4	lg	Carbohydrates (PTS system)	1,514,201-1,523,686	46.4	18	98.00%	Y
	GI-5	sm	Site-specific DNA-methyltransferase	1,005,832-1,011,520	36	11.8	76.00%	Y*
	GI-6	sm	Putative prophage	919,877-925,639	41.6	16.7	94.20%	Y
	GI-7	sm	Sodium-solute symporter/Sugar transporter	664,079-669,638	42.9	14.2	96.20%	Y
	GI-8	sm	Transposable element	486,113-489,350	41.3	26.1	97.40%	Y
	GI-9	sm	Autolysin sensor kinase/ABC-type transport system	417,216-423,212	48.2	19	99.70%	Y
	GI-10	sm	Integrase/Non-hemolytic enterotoxin lytic component L1	388,323-393,223	40.3	11.9	91.00%	Y
	VPI-1	lg	Vibrio pathogenicity island-1 (VPI-1)	355,973-404,865	35.9	10.1	81.40%	Y*
	VPI-2	lg	Vibrio pathogenicity island-2 (VPI-2)	1,449,205-1,505,734	41.3	21	97.30%	Y
Post "PG-2"§	GI-11	sm	Kappa prophage	1012862-1045566	48	0.6	26.30%	**N**
	GI-14	sm	Hypothetical proteins	814223-832392	37.7	0.6	18.00%	**N**
	GI-21	lg	Mu-like prophage	689569-722653	50.4	0.4	19.40%	**N**
		lg	Mu-like prophage	766615-799585	50.5	0.4	19.20%	**N**
	GI-23	lg	Putative prophage	2499645-2522558	39.8	15.6	97.90%	Y
	GI-24	lg	Putative prophage (CRISPR-associated proteins)	2825203-2840525	46.1	21.8	96.70%	Y
Other	CTX (lg)	lg	Cholera toxin prophage (CTX)	1,114,655-1,123,257	41.8	64.6**	99.70%	Y
	CTX (sm)	sm	Cholera toxin prophage (CTX)	560,562-567,519	42.7	67.0**	99.60%	Y
	TLC	lg	Cryptic plasmid linked to the CTX prophage	1,123,876-1,138,038	45	41.8**	99.80%	Y

§ - From reference (6); # - From reference (57)

* - Low %GC likely contributes to the relatively lowered coverage and percent of reference covered at these loci.

** - Relatively higher coverage at these loci is possibly attributable to these loci being repeat regions of uncertain copy number in the ancient strain.

Table S4. Results of non-classical regions included on enrichment array

Accession	Description	Total length included on array	%GC	# reads assembled (raw)	# reads assembled (unique)	Avg unique coverage	% reference covered @1X	Hypothesized or expected presence or type	Confirmed?
NC_002505	*N16961 lg chr*	-	-			-	-	-	-
	VSP-I	16,946							
		14,038 (VC0175-0185 only)	38.5	17,533	2,686	0.5	21.2%	Absent	Y
	VSP-II	27,790							
		26,866 (VC0490-0516 only)	39.9	4,633	1,016	0.8	23.9%	Absent	Y
	VC0514-VC0516	5,048	35.1	94	53	0.2	15.2%	Absent	Y
	rtxC (VC1449-VC1450)	1,190	42.1	125	45	0.9	28.8%	Absent	Y
	RS1 prophage (El Tor variant)	3,182	41.7	49,975	3,013	49.7	77.4%	Classical variant rstR; lacking rstC	Y
	tcpA (El Tor variant)	675	43.0	392	100	7.0	34.8%	Classical variant	Y
	recA (El Tor variant)	1,239	45.6	4,703	993	37.1	95.8%	Classical variant	Y
NC_002506	*N16961 sm chr*	-	-			-	-	-	-
	VCA0300	830							
		630 (VCA0300 only)	39.1	1,279,820	389	4.6	27.8%	Absent	Y
	hlyA (El Tor variant)	60	46.8	77	14	6.9	100%	Classical variant	Y
AB012956	***MO45 O-antigen synthesis genes (O139)***	35,806	40.4	7,236	1,591	1.7	25.6%	Absent	Y

Table S5. Results of human loci included on enrichment array

Accession	Description	Total length included on array	%GC	# reads assembled (raw)	# reads assembled (unique)	Avg unique coverage	% reference covered @1X	Relevant findings
NG_012040	Homo sapiens amelogenin, X-linked (AMELX), RefSeqGene on chromosome X	8,059	37.6	1,347	412	2.0	39.2%	X chr present
NG_009011	Homo sapiens amelogenin, Y-linked (AMELY), RefSeqGene on chromosome Y	8,759	36.6	1,403,301	1,235	14.1	41.5%	Y chr present (?)
NC_012920	Homo sapiens mitochondrion, complete genome	16,569	44.4	19,160,771	33,151	149.3	100%	Haplogroup L3d

Table S6. Human mitochondrial genome SNPs (relative to rCRS, NC_012920)

SNP #	rCRS position	rCRS base	3090.13 base	Unique coverage	% variant frequency
1	73	A	G	149	95.30%
2	146	T	C	140	94.30%
3	152	T	C	136	98.50%
4	263	A	G	128	99.20%
5	750	A	G	150	100.00%
6	921	T	C	151	98.00%
7	1438	A	G	138	97.80%
8	2706	A	G	138	95.70%
9	4769	A	G	132	98.50%
10	4937	T	C	161	95.70%
11	5046	G	A	128	100.00%
12	5147	G	A	150	96.00%
13	6680	T	C	138	97.80%
14	7028	C	T	143	100.00%
15	7424	A	G	152	93.40%
16	8618	T	C	155	94.20%
17	8701	A	G	144	99.30%
18	8860	A	G	159	95.00%
19	9540	T	C	155	95.50%
20	10398	A	G	134	97.00%
21	10694	A	T	142	97.90%
22	10873	T	C	157	96.20%
23	11719	G	A	156	98.10%
24	12280	A	G	138	98.60%
25	12705	C	T	132	97.00%
26	13105	A	G	147	98.60%
27	13886	T	C	156	97.40%
28	14284	C	T	155	98.10%
29	14287	T	C	156	96.20%
30	14634	T	C	156	95.50%
31	14766	C	T	148	99.30%
32	15110	G	A	143	98.60%
33	15301	G	A	148	97.30%
34	15326	A	G	155	97.40%
35	16124	T	C	145	97.90%
36	16223	C	T	155	98.10%
37	16519	T	C	122	95.90%

Table S7. RC27 vs O395 genomic island comparison

Class	GI§	Chr	Description§	O395 position	Pairwise identity	Confirmed?
Post "PG-2"§	GI-11	sm	Kappa prophage	1012862-1045566	99.9%	Yes
	GI-14	sm	Hypothetical proteins	814223-832392	99.9%	Yes
	GI-21	lg	Mu-like prophage	689569-722653	100%	Yes
		lg	Mu-like prophage	766615-799585	n/a	*Unknown*
	GI-23	lg	Putative prophage	2499645-2522558	100%	Yes
	GI-24	lg	Putative prophage (CRISPR-associated proteins)	2825203-2840525	91.1%	Yes

§ - From reference (6)

Table S8. Results of the integron region core genome and O395-specific genes

Cluster§	Locus tag§	Description*	O395 position	%GC	Avg unique coverage	% reference covered @1X	Confirmed?
Full "superintegron"							
-	VC0395_0740 to VC0395_0938	PA1849 is missing GI-14 (814,223-832,392)	799827-916350	41.0%	23.3	80.2%	Y
Core genome integron genes§							
1	VC0395_0867	hypothetical protein	876486-876668	41.8%	50.4	100.0%	Y
1	VC0395_0907	hypothetical protein	898632-898814	44.8%	43.0	100.0%	Y
1	VC0395_0884	hypothetical protein	884967-885149	44.8%	42.1	100.0%	Y
2	VC0395_0802	putative acetyltransferase	838538-838972	41.7%	12.6	98.2%	Y
3	VC0395_0913	putative lipoprotein	900754-901179	41.6%	24.9	100.0%	Y
3	VC0395_0840	putative lipoprotein	863244-863669	40.9%	20.6	100.0%	Y
3	VC0395_0901	hypothetical protein	894878-895303	40.1%	21.2	100.0%	Y
3	VC0395_0819	hypothetical protein	850309-850704	40.1%	20.5	100.0%	Y
4	VC0395_0832	hypothetical protein	858808-858987	44.9%	71.9	100.0%	Y
4	VC0395_0902	hypothetical protein	895487-895666	44.3%	41.1	100.0%	Y
4	VC0395_0857	hypothetical protein	872107-872289	44.3%	48.9	100.0%	Y
5	VC0395_0833	hypothetical protein	858972-859106	43.4%	24.2	100.0%	Y
5	VC0395_0910	hypothetical protein	899789-899929	40.8%	27.8	100.0%	Y
7	VC0395_0938	intI4	915388-916350	41.6%	18.0	100.0%	Y
8	VC0395_0742	CopG family transcriptional regulator	801567-801935	44.5%	19.9	100.0%	Y
9	VC0395_0831	putative acetyltransferase	858099-858614	39.0%	18.2	97.7%	Y
9	VC0395_0741	putative acetyltransferase	800942-801535	39.9%	31.8	100.0%	Y
O395-specific genes§							
308	VC0395_0931	hypothetical protein	910606-910719	41.9%	56.5	100.0%	Y
309	VC0395_0845	hypothetical protein	864958-865884	38.3%	4.8	83.5%	**Y?**
310	VC0395_0783	GI-14; hypothetical protein	826832-826984	32.7%	0.1	13.1%	**N**
311	VC0395_0885	hypothetical protein	885210-885380	43.0%	47.7	100.0%	Y
312	VC0395_0745	hypothetical protein	803209-803358	39.1%	51.3	100.0%	Y
313	VC0395_0876	hypothetical protein	881858-881959	54.6%	91.2	100.0%	Y
314	VC0395_0756	hypothetical protein	809326-809460	40.4%	50	100.0%	Y
315	VC0395_0838	hypothetical protein	862045-862185	3450.0%	8.1	100.0%	Y
316	VC0395_0780	GI-14; hypothetical protein	825158-825496	35.8%	0.3	16.2%	**N**
317	VC0395_0777	GI-14; hypothetical protein	824612-824764	40.5%	0	0.0%	**N**
318	VC0395_0909	hypothetical protein	899193-899369	42.8%	21.3	100.0%	Y
319	VC0395_0779	GI-14; hypothetical protein	825025-825141	35.0%	0	0.0%	**N**
320	VC0395_0859	hypothetical protein	872584-873033	34.8%	5.3	74.0%	**Y?**
321	VC0395_0774	GI-14; hypothetical protein	822322-822504	35.3%	0.2	10.4%	**N**

322	VC0395_0843	hypothetical protein	864432-864818	44.9%	16.2	100.0%	Y
323	VC0395_0789	GI-14; hypothetical protein	830312-830647	34.5%	0	0.0%	**N**
324	VC0395_0782	GI-14; hypothetical protein	826271-826747	36.6%	0.1	8.4%	**N**
325	VC0395_0912	hypothetical protein	900502-900687	43.9%	75.9	100.0%	Y
326	VC0395_0863	hypothetical protein	874509-874637	52.6%	89.4	100.0%	Y
327	VC0395_0824	blc-4	853496-854011	41.5%	14.7	100.0%	Y
328	VC0395_0908	hypothetical protein	898875-899030	42.9%	15.6	100.0%	Y
329	VC0395_0933	hypothetical protein	912303-912494	44.4%	69.9	100.0%	Y
330	VC0395_0847	hypothetical protein	866735-866941	41.8%	45.1	100.0%	Y
331	VC0395_0770	GI-14; hypothetical protein	817596-817973	41.0%	0.1	5.3%	**N**
332	VC0395_0860	isochorismatase family protein	873168-873707	38.2%	6.9	94.1%	**Y?**
333	VC0395_0788	GI-14; hypothetical protein	829454-830008	37.4%	0.7	25.0%	**N**
334	VC0395_0882	hypothetical protein	884165-884278	40.2%	59.8	100.0%	Y
335	VC0395_0791	GI-14 adjacent; hypothetical protein	832392-832850	39.7%	0.1	9.2%	**N**
336	VC0395_0852	hypothetical protein	868887-869006	45.8%	51.5	100.0%	Y
337	VC0395_0878	blc-4	882450-882695	41.3%	19.7	100.0%	Y
338	VC0395_0825	hypothetical protein	853923-854087	41.3%	58.5	100.0%	Y
339	VC0395_0879	hypothetical protein	882877-883041	41.8%	53.5	100.0%	Y
340	VC0395_0773	GI-14; hypothetical protein	822079-822315	36.7%	0.2	10.5%	**N**
341	VC0395_0841	hypothetical protein	863769-863882	36.8	43.4	100.0%	Y
342	VC0395_0856	hypothetical protein	871207-871899	34.1%	3.1	58.0%	**Y?**
343	VC0395_0781	GI-14; hypothetical protein	825518-825817	38.7%	0	0.0%	**N**
344	VC0395_0854	putative acetyltransferase	869432-869545	38.5%	66.2	100.0%	Y
345	VC0395_0767	GI-14; hypothetical protein	814797-814934	40.6%	0	0.0%	**N**
346	VC0395_0769	GI-14; hypothetical protein	817180-817575	36.4%	0.4	11.9%	**N**
347	VC0395_0837	hypothetical protein	861740-861930	41.7%	79.8	100.0%	Y
348	VC0395_0771	GI-14; hypothetical protein	817990-818265	35.0%	0.1	14.5%	**N**
349	VC0395_0839	hypothetical protein	862212-862469	38.2%	11.9	89.1%	Y
350	VC0395_0846	hypothetical protein	866032-866395	38.3%	8.4	96.1%	Y
351	VC0395_0848	hypothetical protein	867226-867807	37.0%	5.4	89.3%	**Y?**
352	VC0395_0924	hypothetical protein	906739-906888	44.1%	65.2	100.0%	Y

§ - From reference (42)

*** -** Gene names, locus ID, and locus descriptions taken from NC_009456 chromosomal annotations (http://www.ncbi.nlm.nih.gov/bioproject?term=PRJNA58425).

Table S9. SNP calls.

Gene names, locus ID, and locus descriptions taken from NC_009456 [small] and NC_009457 [large] chromosomal annotations (http://www.ncbi.nlm.nih.gov/bioproject?term=PRJNA58425). Protein effects, amino acid changes, codon changes, and SNP types generated by Geneious (v.5.6.5, Biomatters, Ltd.) using bacterial translation. *Alt_10:* Alternate call PA1849 (min mapping quality 10). *Alt_20:* Alternate call PA1849 (min mapping quality 20). Ts = transition. Tv = transversion.

#	Ref	Ref Pos	Ref Base	PA1849	C>T/ G>A	Cov	Var Freq	Effect	AA Change	Codon Change	Type	Gene	VC0395 Locus	Description	Alt 10	Alt 20
1	Small	66,022	T	**G**	-	14	92.9%	Subst.	V -> G	GTC -> GGC	Tv	-	62	sulfate permease family protein	G	G
2	Small	80,201	T	**A**	-	9	88.9%	Subst.	E -> V	GAG -> GTG	Tv	hutR	74	heme receptor HutR	A	A
3	Small	132,212	G	**A**	*	24	91.7%	Subst.	P -> S	CCA -> TCA	Ts	vgrG-2	115	vgrG protein	A	A
4	Small	132,579	T	**C**	-	19	100.0%	Subst.	I -> V	ATC -> GTC	Ts	hcp-1	116	hcp protein	C	C
5	Small	136,525	C	**T**	*	10	100.0%	Subst.	G -> D	GGC -> GAC	Ts	malQ	118	4-alpha-glucanotransferase	T	T
6	Small	137,299	G	**A**	*	11	90.9%	Subst.	A -> V	GCA -> GTA	Ts	malQ	118	4-alpha-glucanotransferase	A	A
7	Small	192,164	T	**C**	-	6	100.0%	Subst.	V -> A	GTC -> GCC	Ts	pdxH	163	pyridoxamine 5'-phosphate oxidase	C	C
8	Small	200,223	T	**C**	-	9	88.9%	None	-	CTT -> CTC	Ts	putA	169	bifunctional proline dehydrogenase/pyrroline-5-carboxylate dehydrogenase	C	C
9	Small	220,488	A	**G**	-	11	90.9%	None	-	ACT -> ACC	Ts	-	186	LysR family transcriptional regulator	G	-
10	Small	257,867	T	**A**	-	10	90.0%	Subst.	S -> T	TCA -> ACA	Tv	-	222	Ada family transcriptional regulator	A	A
11	Small	258,772	A	**G**	-	10	100.0%	None	-	GCA -> GCG	Ts	-	222	Ada family transcriptional regulator	G	G
12	Small	269,103	A	**G**	-	22	90.9%	None	-	ACT -> ACC	Ts	-	236	AraC/XylS family transcriptional regulator	G	-
13	Small	274,459	A	**G**	-	18	100.0%	Subst.	C -> R	TGT -> CGT	Ts	-	242	ABC transporter ATP-binding protein	G	G
14	Small	275,155	C	**T**	*	6	100.0%	Subst.	E -> K	GAG -> AAG	Ts	-	242	ABC transporter ATP-binding protein	T	T
15	Small	290,065	T	**C**	-	8	87.5%	Subst.	T -> A	ACG -> GCG	Ts	-	253	oxidoreductase/iron-sulfur cluster-binding protein	C	C
16	Small	305,751	T	**C**	-	11	100.0%	None	-	GGT -> GGC	Ts	-	267	hypothetical protein	C	C
17	Small	320,729	T	**G**	-	13	100.0%	Subst.	S -> A	TCA -> GCA	Tv	-	284	MarR family transcriptional regulator	G	G
18	Small	324,884	T	**C**	-	12	100.0%	None	-	GGT -> GGC	Ts	-	289	SM-20-related protein	C	C
19	Small	343,763	T	**C**	-	8	100.0%	Subst.	T -> A	ACC -> GCC	Ts	-	311	hypothetical protein	C	C

20	Small	366,368	T	**C**	-	7	100.0%	None	-	TTT -> TTC	Ts	-	333	permease	C	-
21	Small	382,738	C	**T**	*	6	100.0%	None	-	GCG -> GCA	Ts	-	349	LysR family transcriptional regulator	T	T
22	Small	418,160	T	**C**	-	40	90.0%	Subst.	V -> A	GTG -> GCG	Ts	-	382	ABC transporter ATP-binding protein	C	C
23	Small	425,788	A	**G**	-	11	100.0%	Subst.	K -> E	AAA -> GAA	Ts	-	388	hypothetical protein	G	-
24	Small	465,922	A	**G**	-	9	88.9%	Subst.	C -> R	TGC -> CGC	Ts	-	421	aminopeptidase	G	G
25	Small	499,123	A	**T**	-	5	100.0%	None	-	TCT -> TCA	Tv	fruA-2	449	PTS system fructose-specific transporter subunit IIBC	T	-
26	Small	531,700	G	**A**	*	10	90.0%	None	-	CCG -> CCA	Ts	-	475	LysR family transcriptional regulator	A	A
27	Small	581,974	G	**A**	*	9	100.0%	Subst.	H -> Y	CAC -> TAC	Ts	-	526	hypothetical protein	A	A
28	Small	633,914	A	**C**	-	8	100.0%	None	-	GTT -> GTG	Tv	-	582	AcrB/AcrD/AcrF family transporter	C	C
29	Small	676,014	T	**C**	-	7	100.0%	Subst.	L -> P	CTT -> CCT	Ts	napB	618	periplasmic nitrate reductase, cytochrome c-type protein	C	-
30	Small	681,155	C	**T**	*	33	100.0%	None	-	ATC -> ATT	Ts	uhpC	624	regulatory protein UhpC	T	-
31	Small	683,932	A	**G**	-	25	96.0%	None	-	TCA -> TCG	Ts	-	626	iron(III) ABC transporter permease	G	-
32	Small	708,575	A	**G**	-	6	100.0%	None	-	TAT -> TAC	Ts	-	643	putative phosphoglycerate transport regulatory protein PgtB	G	G
33	Small	712,788	T	**A**	-	6	100.0%	Subst.	L -> Q	CTA -> CAA	Tv	pykA-2	647	pyruvate kinase	A	A
34	Small	741,285	T	**C**	-	12	91.7%	Subst.	R -> G	AGG -> GGG	Ts	luxQ	671	sensor histidine kinase LuxQ	C	-
35	Small	753,295	C	**T**	*	9	100.0%	None	-	GAC -> GAT	Ts	glpC	688	sn-glycerol-3-phosphate dehydrogenase subunit C	T	T
36	Small	774,732	T	**G**	-	11	90.9%	Subst.	S -> A	TCC -> GCC	Tv	-	713	methyl-accepting chemotaxis protein	G	G
37	Small	796,953	A	**C**	-	7	100.0%	Subst.	N -> H	AAT -> CAT	Tv	-	735	hypothetical protein	C	-
38	Small	798,190	T	**C**	-	12	91.7%	None	-	GAT -> GAC	Ts	-	736	hypothetical protein	C	C
39	Small	848,915	T	**C**	-	18	83.3%	-	-	-	Ts	-	-	-	C	C
40	Small	947,092	T	**C**	-	7	100.0%	None	-	TAT -> TAC	Ts	-	971	hypothetical protein	C	C
41	Small	1,106,917	A	**G**	-	11	90.9%	None	-	GGT -> GGC	Ts	-	1137	putative chitinase	G	G
42	Large	56,505	G	**A**	*	6	100.0%	Subst.	S -> N	AGC -> AAC	Ts	rpoS	A0062	RNA polymerase sigma factor RpoS	-	-
43	Large	66,630	T	**C**	-	8	100.0%	-	-	-	Ts	-	-	-	C	C
44	Large	78,656	A	**C**	-	9	100.0%	Subst.	D -> A	GAT -> GCT	Tv	-	A0089	insulinase family protease/insulinase family protease	C	C
45	Large	108,381	A	**G**	-	8	100.0%	Subst.	D -> G	GAT -> GGT	Ts	-	A0120	ABC transporter, ATP-binding protein	G	G
46	Large	115,828	T	**G**	-	11	90.9%	Subst.	I -> M	ATT -> ATG	Tv	-	A0129	hypothetical protein	G	-

47	Large	116,323	A	**G**	-	15	86.7%	None	-	CCA -> CCG	Ts	-	A0129	hypothetical protein	G	G
48	Large	133,808	G	**A**	*	6	100.0%	Subst.	A -> V	GCC -> GTC	Ts	-	A0140	putative phosphoglucomutase/phosphomannomutase	A	A
49	Large	140,885	A	**G**	-	7	100.0%	None	-	GCT -> GCC	Ts	-	A0144	endoglucanase-related protein	G	-
50	Large	179,324	T	**C**	-	5	100.0%	Subst.	V -> A	GTT -> GCT	Ts	truB	A0176	tRNA pseudouridine synthase B	C	-
51	Large	182,036	A	**G**	-	7	100.0%	None	-	CTA -> CTG	Ts	pnp	A0178	polynucleotide phosphorylase/polyadenylase	G	G
52	Large	244,360	T	**C**	-	10	100.0%	Subst.	M -> V	ATG -> GTG	Ts	-	A0233	signal protein domain/GGDEF domain-containing protein	C	C
53	Large	244,466	A	**C**	-	21	100.0%	Subst.	Y -> D	TAC -> GAC	Tv	-	A0234	putative spermidine/putrescine ABC transporter, periplasmic spermidine/putrescine-binding protein	C	C
54	Large	244,878	C	**T**	*	7	100.0%	None	-	AGG -> AGA	Ts	-	A0234	putative spermidine/putrescine ABC transporter, periplasmic spermidine/putrescine-binding protein	T	T
55	Large	266,668	A	**T**	-	10	100.0%	Subst.	L -> H	CTT -> CAT	Tv	ppx	A0256	exopolyphosphatase	T	-
56	Large	266,935	G	**A**	*	6	100.0%	Subst.	A -> V	GCC -> GTC	Ts	ppx	A0256	exopolyphosphatase	A	A
57	Large	301,627	T	**C**	-	18	94.4%	Subst.	I -> T	ATA -> ACA	Ts	-	A0287	hypothetical protein	C	C
58	Large	309,320	C	**T**	*	6	100.0%	Subst.	A -> T	GCC -> ACC	Ts	xseA	A0294	exodeoxyribonuclease VII large subunit	T	T
59	Large	318,928	T	**C**	-	23	95.7%	Subst.	D -> G	GAC -> GGC	Ts	vibE	A0301	vibriobactin-specific 2,3-dihydroxybenzoate-AMP ligase	C	C
60	Large	333,744	T	**C**	-	9	100.0%	Subst.	Y -> C	TAT -> TGT	Ts	dadA	A0313	D-amino acid dehydrogenase small subunit	C	C
61	Large	337,647	T	**C**	-	7	100.0%	Subst.	H -> R	CAC -> CGC	Ts	citA	A0318	sensor kinase citA	C	C
62	Large	347,443	T	**G**	-	6	100.0%	Subst.	L -> V	TTG -> GTG	Tv	citF	A0326	citrate lyase, alpha subunit	G	G
63	Large	347,853	T	**G**	-	18	94.4%	Subst.	F -> V	TTT -> GTT	Tv	citX	A0327	apo-citrate lyase phosphoribosyl-dephospho-CoA transferase	G	-
64	Large	381,304	G	**T**	-	66	86.4%	None	-	GGG -> GGT	Tv	tcpA	A0353	toxin co-regulated pilin	T	-
65	Large	454,548	T	**G**	-	8	100.0%	Subst.	C -> G	TGC -> GGC	Tv	syd	A0425	SecY interacting protein Syd	G	G
66	Large	468,576	A	**G**	-	10	100.0%	Subst.	S -> G	AGC -> GGC	Ts	-	A0439	putative multidrug resistance protein	G	G
67	Large	497,165	G	**A**	*	16	81.3%	Subst.	A -> V	GCG -> GTG	Ts	-	A0464	hypothetical protein	A	A
68	Large	506,041	T	**C**	-	13	92.3%	Subst.	K -> E	AAA -> GAA	Ts	mrdA	A0473	penicillin-binding protein 2	C	C
69	Large	546,223	A	**G**	-	12	91.7%	Subst.	V -> A	GTG -> GCG	Ts	asnB	A0512	asparagine synthetase B	G	G
70	Large	555,530	T	**C**	-	7	100.0%	Subst.	V -> A	GTC -> GCC	Ts	glnS	A0517	glutaminyl-tRNA synthetase	C	C
71	Large	570,413	C	**T**	*	12	100.0%	Subst.	G -> D	GGC -> GAC	Ts	-	A0530	hypothetical protein	T	T
72	Large	571,610	T	**C**	-	13	100.0%	None	-	GCA -> GCG	Ts	nth	A0532	endonuclease III	C	C

73	Large	582,216	T	**C**	-	21	95.2%	None	-	GCT -> GCC	Ts	luxO	A0541	LuxO repressor protein	C	C
74	Large	586,361	T	**G**	-	11	100.0%	Subst.	F -> L	TTT -> TTG	Tv	moaE	A0548	molybdenum cofactor biosynthesis protein E	G	-
75	Large	623,349	A	**G**	-	30	93.3%	Subst.	Q -> R	CAG -> CGG	Ts	-	A0584	hypothetical protein	G	G
76	Large	629,663	A	**G**	-	6	100.0%	Subst.	S -> G	AGC -> GGC	Ts	-	A0589	putative transporter	G	G
77	Large	631,340	A	**G**	-	12	100.0%	Exten.	-	-	Ts	-	A0591	putative chitinase	G	G
78	Large	643,771	T	**C**	-	7	100.0%	Subst.	V -> A	GTA -> GCA	Ts	-	A0605	response regulator	C	C
79	Large	666,437	A	**G**	-	15	100.0%	Exten.	-	-	Ts	-	A0624	tRNA-dihydrouridine synthase C	G	G
80	Large	734,478	C	**T**	*	8	100.0%	Subst.	A -> V	GCA -> GTA	Ts	icd	A0711	isocitrate dehydrogenase, NADP-dependent, monomeric type	T	T
81	Large	800,212	T	**C**	-	13	92.3%	Subst.	D -> G	GAT -> GGT	Ts	-	A0790	proton/glutamate symporter	C	C
82	Large	808,282	C	**A**	-	6	100.0%	None	-	GGC -> GGA	Tv	-	A0798	hypothetical protein	-	-
83	Large	818,387	T	**C**	-	7	85.7%	None	-	CGT -> CGC	Ts	-	A0806	GGDEF family protein	C	C
84	Large	831,319	T	**A**	-	6	100.0%	Subst.	S -> T	TCC -> ACC	Tv	-	A0818	hypothetical protein	A	-
85	Large	847,296	G	**A**	*	16	93.8%	None	-	CGC -> CGT	Ts	polB	A0831	DNA polymerase II	A	A
86	Large	850,735	T	**C**	-	22	95.5%	None	-	CCT -> CCC	Ts	uvrC	A0834	excinuclease ABC subunit C	C	-
87	Large	908,801	T	**G**	-	19	100.0%	Subst.	C -> G	TGC -> GGC	Tv	-	A0894	hypothetical protein	G	-
88	Large	928,592	A	**T**	-	6	100.0%	Subst.	L -> Q	CTG -> CAG	Tv	-	A0910	hypothetical protein	-	-
89	Large	933,802	C	**T**	*	11	90.9%	None	-	GCC -> GCT	Ts	thiD	A0914	phosphomethylpyrimidine kinase	T	-
90	Large	958,402	C	**A**	-	10	100.0%	Subst.	L -> M	CTG -> ATG	Tv	-	A0934	putative chemotaxis protein CheY	A	A
91	Large	998,888	C	**A**	-	23	100.0%	Subst.	P -> Q	CCG -> CAG	Tv	-	A0965	sensory box sensor histidine kinase/response regulator	A	A
92	Large	1,004,776	T	**C**	-	22	100.0%	Subst.	C -> R	TGC -> CGC	Ts	-	A0969	acylphosphatase	C	-
93	Large	1,006,899	G	**A**	*	14	100.0%	None	-	TTC -> TTT	Ts	-	A0972	amino acid ABC transporter, ATP-binding protein	A	-
94	Large	1,009,772	A	**C**	-	6	100.0%	Subst.	C -> W	TGT -> TGG	Tv	-	A0975	amino acid ABC transporter, periplasmic amino acid-binding protein	C	-
95	Large	1,036,037	T	**A**	-	52	86.5%	-	-	-	Tv	-	-	-	A	A
96	Large	1,093,693	T	**C**	-	6	100.0%	Subst.	T -> A	ACG -> GCG	Ts	-	A1048	cation transporter E1-E2 family ATPase	C	-
97	Large	1,103,306	T	**C**	-	10	90.0%	Subst.	N -> S	AAC -> AGC	Ts	-	A1056	putative toxin secretion transporter	-	-
98	Large	1,104,525	A	**G**	-	32	100.0%	None	-	GCT -> GCC	Ts	rtxD	A1057	RTX toxin transporter	G	G

99	Large	1,195,245	A	**G**	-	9	100.0%	Subst.	I -> M	ATA -> ATG	Ts	moeA-1	A1134	putative bifunctional molybdopterin-guanine dinucleotide biosynthesis protein MobB/MoeA	G	G
100	Large	1,206,933	T	**G**	-	12	100.0%	Subst.	V -> G	GTC -> GGC	Tv	-	A1146	multidrug efflux protein	G	G
101	Large	1,225,393	A	**G**	-	17	100.0%	Subst.	T -> A	ACC -> GCC	Ts	perA	A1166	catalase/peroxidase	G	G
102	Large	1,316,676	A	**G**	-	5	100.0%	None	-	ACA -> ACG	Ts	-	A1243	hypothetical protein	-	-
103	Large	1,330,767	T	**C**	-	8	100.0%	None	-	GCT -> GCC	Ts	-	A1256	collagenase	C	C
104	Large	1,340,672	C	**T**	*	10	100.0%	Subst.	L -> F	CTC -> TTC	Ts	mgtE-1	A1260	magnesium transporter	T	T
105	Large	1,364,732	T	**C**	-	18	94.4%	Subst.	R -> G	AGG -> GGG	Ts	pspA	A1283	phage shock protein A	C	-
106	Large	1,369,091	A	**G**	-	5	100.0%	Subst.	H -> R	CAC -> CGC	Ts	sapC	A1287	peptide ABC transporter, permease protein	G	G
107	Large	1,371,760	A	**G**	-	6	83.3%	None	-	TCA -> TCG	Ts	-	A1290	hypothetical protein	G	G
108	Large	1,382,329	T	**C**	-	10	100.0%	-	-	-	Ts	-	-	-	C	-
109	Large	1,397,204	A	**C**	-	21	95.2%	Subst.	V -> G	GTT -> GGT	Tv	-	A1313	hypothetical protein	C	-
110	Large	1,397,683	G	**T**	-	7	100.0%	Subst.	D -> E	GAC -> GAA	Tv	-	A1313	hypothetical protein	T	T
111	Large	1,397,857	A	**G**	-	5	100.0%	None	-	TGT -> TGC	Ts	-	A1313	hypothetical protein	G	G
112	Large	1,420,480	A	**G**	-	15	93.3%	None	-	GAT -> GAC	Ts	topA	A1332	DNA topoisomerase I	G	-
113	Large	1,455,601	T	**C**	-	39	97.4%	Subst.	E -> G	GAG -> GGG	Ts	-	A1360	hypothetical protein	C	-
114	Large	1,458,822	G	**A**	*	8	100.0%	None	-	AGC -> AGT	Ts	-	A1362	hypothetical protein	A	-
115	Large	1,462,768	A	**G**	-	19	94.7%	Subst.	C -> R	TGT -> CGT	Ts	-	A1364	hypothetical protein	G	G
116	Large	1,498,330	T	**C**	-	16	93.8%	None	-	GAA -> GAG	Ts	-	A1396	hypothetical protein	C	C
117	Large	1,538,329	A	**G**	-	8	100.0%	Subst.	V -> A	GTC -> GCC	Ts	ruvB	A1436	Holliday junction DNA helicase RuvB	G	G
118	Large	1,581,922	T	**C**	-	42	92.9%	None	-	GTA -> GTG	Ts	-	A1474	hypothetical protein	C	C
119	Large	1,604,639	G	**C**	-	11	100.0%	Subst.	G -> A	GGT -> GCT	Tv	dsbB	A1492	disulfide bond formation protein B	C	C
120	Large	1,623,768	T	**C**	-	20	100.0%	None	-	GTA -> GTG	Ts	lon	A1510	ATP-dependent protease LA	C	-
121	Large	1,631,000	T	**C**	-	13	100.0%	Subst.	C -> R	TGT -> CGT	Ts	dctD-1	A1516	C4-dicarboxylate transport transcriptional regulatory protein	C	C
122	Large	1,635,653	A	**G**	-	10	100.0%	Subst.	I -> T	ATC -> ACC	Ts	-	A1520	hypothetical protein	G	-
123	Large	1,646,423	G	**A**	*	8	87.5%	-	-	-	Ts	-	-	-	A	A
124	Large	1,676,446	T	**C**	-	6	100.0%	Subst.	T -> A	ACT -> GCT	Ts	menC	A1558	O-succinylbenzoate synthase	C	C

125	Large	1,685,272	T	**C**	-	6	100.0%	Subst.	C -> R	TGC -> CGC	Ts	dgt	A1566	deoxyguanosinetriphosphate triphosphohydrolase-like protein	C	C
126	Large	1,686,990	A	**G**	-	11	90.9%	None	-	GGT -> GGC	Ts	-	A1568	putative peptidase	G	G
127	Large	1,688,590	G	**A**	*	7	100.0%	Subst.	A -> V	GCG -> GTG	Ts	-	A1568	putative peptidase	A	A
128	Large	1,688,814	T	**C**	-	5	100.0%	-	-	-	Ts	-	-	-	C	C
129	Large	1,737,647	T	**C**	-	18	100.0%	Subst.	V -> A	GTG -> GCG	Ts	rne	A1616	ribonuclease E	C	C
130	Large	1,747,420	T	**C**	-	13	100.0%	Subst.	L -> P	CTA -> CCA	Ts	yqkI	A1622	Na+/H+ antiporter	C	-
131	Large	1,769,719	T	**C**	-	29	89.7%	None	-	GCA -> GCG	Ts	cheB-2	A1650	chemotaxis-specific methylesterase	C	C
132	Large	1,770,100	T	**G**	-	13	100.0%	None	-	ACA -> ACC	Tv	cheB-2	A1650	chemotaxis-specific methylesterase	G	G
133	Large	1,778,301	A	**G**	-	7	100.0%	Subst.	S -> P	TCG -> CCG	Ts	flhA	A1657	flagellar biosynthesis protein FlhA	G	G
134	Large	1,809,756	C	**T**	*	11	100.0%	None	-	GGG -> GGA	Ts	pgm	A1681	phosphoglucomutase	T	T
135	Large	1,830,663	C	**T**	*	6	100.0%	None	-	GCG -> GCA	Ts	fliP	A1705	flagellar biosynthesis protein FliP	T	-
136	Large	1,836,002	C	**T**	*	10	100.0%	None	-	GAG -> GAA	Ts	fliJ	A1713	flagellar biosynthesis chaperone	T	T
137	Large	1,836,427	A	**G**	-	5	100.0%	None	-	AAT -> AAC	Ts	fliI	A1714	flagellum-specific ATP synthase	G	G
138	Large	1,836,533	A	**G**	-	12	91.7%	Subst.	L -> P	CTG -> CCG	Ts	fliI	A1714	flagellum-specific ATP synthase	G	-
139	Large	1,856,378	C	**T**	*	10	100.0%	Subst.	A -> V	GCT -> GTT	Ts	dapE	A1734	succinyl-diaminopimelate desuccinylase	T	T
140	Large	1,857,209	G	**A**	*	18	94.4%	Subst.	G -> D	GGT -> GAT	Ts	dapE	A1734	succinyl-diaminopimelate desuccinylase	A	A
141	Large	1,868,567	A	**C**	-	17	100.0%	Subst.	D -> E	GAT -> GAG	Tv	-	A1748	hypothetical protein	C	C
142	Large	1,936,466	A	**C**	-	16	93.8%	None	-	ACT -> ACG	Tv	-	A1821	hypothetical protein	C	C
143	Large	1,937,686	G	**C**	-	17	100.0%	Subst.	C -> S	TGT -> TCT	Tv	fadE	A1823	acyl-CoA dehydrogenase	C	C
144	Large	1,938,486	C	**A**	-	8	100.0%	Subst.	H -> N	CAT -> AAT	Tv	fadE	A1823	acyl-CoA dehydrogenase	A	-
145	Large	1,970,429	T	**C**	-	13	100.0%	None	-	ACT -> ACC	Ts	map	A1852	methionine aminopeptidase	C	C
146	Large	2,009,684	A	**G**	-	17	94.1%	None	-	GAT -> GAC	Ts	-	A1891	RNA polymerase sigma factor	G	G
147	Large	2,013,030	A	**C**	-	24	95.8%	None	-	GTA -> GTC	Tv	thiJ	A1895	4-methyl-5(B-hydroxyethyl)-thiazole monophosphate biosynthesis enzyme	C	C
148	Large	2,013,619	T	**C**	-	5	100.0%	None	-	CGT -> CGC	Ts	-	A1896	aminotransferase, class V	C	C
149	Large	2,014,096	A	**T**	-	20	95.0%	None	-	GTA -> GTT	Tv	-	A1896	aminotransferase, class V	T	T
150	Large	2,040,334	C	**A**	-	16	87.5%	Subst.	P -> T	CCC -> ACC	Tv	-	A1916	LacI family transcription regulator	A	-

151	Large	2,050,946	A	**G**	-	5	100.0%	-	-	-	Ts	-	-	-	G	G
152	Large	2,053,823	A	**G**	-	25	96.0%	Subst.	T -> A	ACA -> GCA	Ts	-	A1923	hypothetical protein	G	G
153	Large	2,148,204	T	**A**	-	9	100.0%	-	-	-	Tv	-	-	-	A	A
154	Large	2,162,100	C	**T**	*	13	92.3%	Subst.	G -> S	GGT -> AGT	Ts	cpdA	A2010	cyclic 3',5'-adenosine monophosphate phosphodiesterase	T	T
155	Large	2,191,021	T	**G**	-	5	100.0%	None	-	ACA -> ACC	Tv	-	A2031	GGDEF family protein	-	-
156	Large	2,206,838	A	**G**	-	6	100.0%	-	-	-	Ts	-	-	-	G	G
157	Large	2,212,043	C	**T**	*	10	100.0%	Subst.	A -> V	GCG -> GTG	Ts	-	A2055	hypothetical protein	-	-
158	Large	2,215,334	T	**G**	-	6	100.0%	Subst.	E -> A	GAG -> GCG	Tv	ilvI	A2059	acetolactate synthase 3 catalytic subunit	G	G
159	Large	2,262,928	A	**G**	-	32	93.8%	Subst.	H -> R	CAC -> CGC	Ts	-	A2105	3-deoxy-D-manno-octulosonate 8-phosphate phosphatase	G	G
160	Large	2,287,151	A	**G**	-	8	100.0%	-	-	-	Ts	-	-	-	G	G
161	Large	2,347,304	A	**G**	-	6	100.0%	None	-	AAT -> AAC	Ts	-	A2199	hypothetical protein	G	G
162	Large	2,365,350	A	**G**	-	10	80.0%	None	-	GTA -> GTG	Ts	-	A2215	dihydrolipoamide dehydrogenase	G	G
163	Large	2,372,470	G	**A**	*	7	85.7%	Subst.	G -> S	GGT -> AGT	Ts	ppc	A2222	phosphoenolpyruvate carboxylase	-	-
164	Large	2,410,099	A	**G**	-	7	100.0%	Subst.	I -> M	ATA -> ATG	Ts	glpX	A2260	fructose 1,6-bisphosphatase II	-	-
165	Large	2,421,128	A	**T**	-	25	92.0%	Subst.	E -> V	GAG -> GTG	Tv	dcuA	A2271	anaerobic C4-dicarboxylate transporter	T	T
166	Large	2,442,940	C	**T**	*	16	100.0%	Subst.	V -> M	GTG -> ATG	Ts	ompR	A2286	osmolarity response regulator	T	T
167	Large	2,453,481	C	**T**	*	22	95.5%	None	-	CGG -> CGA	Ts	epsK	A2299	general secretion pathway protein K	T	T
168	Large	2,462,523	A	**G**	-	29	96.6%	Subst.	K -> E	AAA -> GAA	Ts	-	A2308	hypothetical protein	G	-
169	Large	2,527,414	C	**T**	*	27	100.0%	None	-	CTG -> TTG	Ts	-	A2370	soluble pyridine nucleotide transhydrogenase	T	-
170	Large	2,534,013	T	**C**	-	17	100.0%	Subst.	T -> A	ACT -> GCT	Ts	-	A2377	hypothetical protein	C	C
171	Large	2,544,839	T	**C**	-	16	100.0%	Subst.	H -> R	CAC -> CGC	Ts	-	A2389	GGDEF family protein	C	C
172	Large	2,562,078	A	**G**	-	6	100.0%	Subst.	T -> A	ACT -> GCT	Ts	engB	A2407	ribosome biogenesis GTP-binding protein YsxC	G	G
173	Large	2,580,581	A	**C**	-	15	93.3%	Subst.	S -> R	AGC -> CGC	Tv	dinF	A2425	DNA-damage-inducible protein F	C	C
174	Large	2,585,952	A	**G**	-	30	93.3%	Subst.	Y -> H	TAC -> CAC	Ts	ubiB	A2430	putative ubiquinone biosynthesis protein UbiB	G	G
175	Large	2,605,179	T	**C**	-	15	100.0%	-	-	-	Ts	-	-	-	C	C
176	Large	2,606,287	G	**A**	*	8	87.5%	Subst.	H -> Y	CAC -> TAC	Ts	thiE	A2451	thiamine-phosphate pyrophosphorylase	A	A
177	Large	2,610,145	C	**G**	-	8	100.0%	-	-	-	Tv	-	-	-	G	G

178	Large	2,630,356	A	**G**	-	12	100.0%	None	-	GTA -> GTG	Ts	-	A2478	hypothetical protein	G	G
179	Large	2,648,813	T	**C**	-	56	94.6%	Subst.	T -> A	ACG -> GCG	Ts	-	A2497	hypothetical protein	C	C
180	Large	2,664,345	A	**G**	-	13	92.3%	None	-	GTA -> GTG	Ts	-	A2511	amino acid ABC transporter, ATP-binding protein	G	-
181	Large	2,668,895	A	**G**	-	8	100.0%	None	-	GAA -> GAG	Ts	mioC	A2516	flavodoxin	G	G
182	Large	2,711,052	A	**G**	-	39	89.7%	Subst.	T -> A	ACT -> GCT	Ts	-	A2564	peptide ABC transporter, periplasmic peptide-binding protein	G	G
183	Large	2,722,127	A	**G**	-	13	100.0%	None	-	GCA -> GCG	Ts	uvrD	A2571	DNA-dependent helicase II	G	G
184	Large	2,742,112	A	**G**	-	19	100.0%	None	-	GGA -> GGG	Ts	murP	A2586	PTS system N-acetylmuramic acid transporter subunits EIIBC	G	G
185	Large	2,768,000	A	**G**	-	5	100.0%	Subst.	I -> T	ATA -> ACA	Ts	-	A2617	hypothetical protein	G	G
186	Large	2,805,015	T	**C**	-	12	83.3%	None	-	AAA -> AAG	Ts	purD	A2652	phosphoribosylamine--glycine ligase	C	-
187	Large	2,816,109	A	**G**	-	19	100.0%	Subst.	Y -> H	TAC -> CAC	Ts	-	A2659	methyl-accepting chemotaxis protein	G	G
188	Large	2,817,245	C	**T**	*	23	91.3%	None	-	GCG -> GCA	Ts	-	A2662	TonB system receptor	T	T
189	Large	2,843,557	G	**A**	*	13	100.0%	Subst.	T -> I	ACC -> ATC	Ts	prmA	A2687	ribosomal protein L11 methyltransferase	A	A
190	Large	2,855,536	A	**C**	-	81	96.3%	Subst.	Q -> P	CAG -> CCG	Tv	-	A2695	sensor histidine kinase	C	C
191	Large	2,864,037	A	**T**	-	19	94.7%	Subst.	Q -> H	CAA -> CAT	Tv	-	A2702	hypothetical protein	T	T
192	Large	2,864,400	G	**A**	*	6	100.0%	None	-	TCG -> TCA	Ts	-	A2702	hypothetical protein	-	-
193	Large	2,883,093	T	**C**	-	10	100.0%	Subst.	V -> A	GTA -> GCA	Ts	rpoB	A2730	DNA-directed RNA polymerase subunit beta	C	C
194	Large	2,894,227	C	**T**	*	30	90.0%	Subst.	H -> Y	CAT -> TAT	Ts	-	A2735	TetR family transcriptional regulator	T	T
195	Large	2,900,319	C	**G**	-	13	100.0%	Subst.	V -> L	GTA -> CTA	Tv	-	A2740	putative transporter	G	G
196	Large	2,904,164	A	**G**	-	7	100.0%	-	-	-	Ts	-	-	-	G	-
197	Large	2,916,569	A	**C**	-	8	100.0%	Subst.	L -> R	CTA -> CGA	Tv	-	A2764	putative lipoprotein	C	C
198	Large	2,940,607	T	**G**	-	6	100.0%	Subst.	D -> E	GAT -> GAG	Tv	cysJ	A2795	sulfite reductase (NADPH) flavoprotein alpha-component	G	G
199	Large	2,952,457	T	**G**	-	14	100.0%	None	-	GGT -> GGG	Tv	metH	A2808	B12-dependent methionine synthase	G	-
200	Large	2,964,016	G	**A**	*	10	100.0%	None	-	CAC -> CAT	Ts	-	A2814	LuxR family transcriptional regulator	A	A
201	Large	2,967,469	C	**T**	*	19	100.0%	Subst.	H -> Y	CAC -> TAC	Ts	mshI	A2817	MSHA biogenesis protein MshI	T	T
202	Large	2,973,101	G	**A**	*	11	90.9%	None	-	GTG -> GTA	Ts	mshN	A2822	MSHA biogenesis protein MshN	-	-
203	Large	2,984,796	T	**C**	-	19	100.0%	Subst.	F -> S	TTT -> TCT	Ts	mreB	A2834	rod shape-determining protein MreB	C	-

2.16 References

1. Mukandavire Z *et al.* 2011. Estimating the reproductive numbers for the 2008–2009 cholera outbreaks in Zimbabwe. *Proc Natl Acad Sci* 108:8767-8772.
2. Chin CS *et al.* 2011. The origin of the Haitian cholera outbreak strain. *N Engl J Med* 364:33-42.
3. Armelagos GJ, Brown PJ, Turner B. 2005. Evolutionary, historical and political economic perspectives on health and disease. *Soc Sci Med* 61:755-765.
4. Ali M *et al.* 2012. The global burden of cholera. *Bull World Health Org* 90:209-218.
5. Faruque SM *et al.* 2004. Genetic diversity and virulence potential of environmental Vibrio cholerae population in a cholera-endemic area. *Proc Natl Acad Sci USA* 101:2123-2128.
6. Chun J *et al.* 2009. Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic Vibrio cholerae. *Proc Natl Acad Sci USA* 106:15442-15447.
7. Rosenberg CE. 1987. *The Cholera Years – The United States in 1832, 1849, and 1866*, University of Chicago Press, Chicago, IL.
8. Harris JB, LaRocque RC, Qadri F, Ryan ET, Calderwood SB. 2012. Cholera. *Lancet* 379:2466-2476.
9. Bos KI *et al.* 2011. A draft genome of Yersinia pestis from victims of the Black Death. *Nature* 478:506-510.
10. Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protoc* 2010:1-10, 11-17.
11. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754-1760.
12. Darling AE, Mau B, Perna NT. 2010. progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* 5.
13. Krzywinski M *et al.* 2009. Circos: An information aesthetic for comparative genomics. *Genome Res* 19:1639-1645.
14. Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* 56:564-577.
15. Guindon S *et al.* 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst Biol* 59:307-321.
16. Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7.
17. Briggs AW *et al.* 2007. Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci USA* 104:14616-14621.
18. Ginolhac A, Rasmussen M, Gilbert MTP, Willerslev E, Orlando L. 2011. mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics* 27:2153-2155.
19. Waldor MK, Mekalanos JJ. 1996. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* 272:1910-1914.
20. Barbieri C *et al.* 2012. Contrasting maternal and paternal histories in the linguistic context of Burkina Faso. *Mol Biol Evol* 29:1213-1223.

21. Mutreja A *et al.* 2011. Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* 477**:**462-U111.
22. Kaper JB, Morris JG, Jr., Levine MM. 1995. Cholera. *Clin Microbiol Rev* 8**:**48-86.
23. Barua D. 1992. *History of Cholera*, p. 1-36. *In* Barua D, Greenough III WB (ed.), *Cholera*. Plenum Medical Book Company, New York, NY.
24. Schuenemann VJ *et al.* 2013. Genome-wide comparison of Medieval and modern Mycobacterium leprae. *Science* 341**:**179-183.
25. Davis BM, Moyer KE, Boyd EF, Waldor MK. 2000. CTX Prophages in Classical Biotype Vibrio cholerae: Functional Phage Genes but Dysfunctional Phage Genomes. *J Bacteriol* 182**:**6992-6998.
26. Davis BM, Waldor MK. 2000. CTXphi contains a hybrid genome derived from tandemly integrated elements. *Proc Natl Acad Sci USA* 97**:**8572-8577.
27. Faruque SM *et al.* 2007. Genomic analysis of the Mozambique strain of Vibrio cholerae O1 reveals the origin of El Tor strains carrying classical CTX prophage. *Proc Natl Acad Sci* 104**:**5151-5156.
28. Feng L *et al.* 2008. A recalibrated molecular clock and independent origins for the cholera pandemic clones. *PLoS ONE* 3**:**e4053.
29. Harrington JM, Blakely RL. 1995. *Rich man, poor man, beggar man, thief: The selectivity exercised by graverobbers at the Medical College of Georgia, 1837-1887*, p. 153-178. *In* Saunders SR, Herring A (ed.), *Grave Reflections: Portraying the Past through Cemetery Studies*. Canadian Scholars' Press Toronto, Toronto, ON.
30. Shippen E. 1881. *Memoir of John Neill, M.D., Late Emeritus Professor of Clinical Surgery in the University of Pennsylvania [Read October 6, 1880.]*, p. cxli-clvi. *In* Ashhurst Jr. J (ed.), *Transactions of the College of Physicians of Philadelphia*, vol. V. Lindsay & Blakiston, Philadelphia, PA.
31. Okello JBA *et al.* 2010. Comparison of methods in the recovery of nucleic acids from archival formalin-fixed paraffin-embedded autopsy tissues. *Anal Biochem* 400**:**110-117.
32. Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* 40.
33. Bos KI. 2011. *Doctoral Thesis: Genetic Investigations into the Black Death*. McMaster University, Hamilton, ON, Canada.
34. Hodges E *et al.* 2007. Genome-wide in situ exon capture for selective resequencing. *Nat Genet* 39**:**1522-1527.
35. Hodges E *et al.* 2009. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc* 4**:**960-974.
36. 2009. Agilent SureSelect DNA Capture Array Protocol, Version 1.0 ed, vol. G4458-90000. Agilent Technologies.
37. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 17.
38. Li H *et al.* 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25**:**2078-2079.
39. Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 29**:**1969-1973.

40. Martin DP *et al.* 2010. RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* 26**:**2462-2463.
41. Lee J *et al.* 2009. Classification of hybrid and altered Vibrio cholerae strains by CTX prophage and RS1 element structure. *J Microbiol* 47**:**783-788.
42. Marin MA, Vicente ACP. 2013. Architecture of the superintegron in Vibrio cholerae: identification of core and unique genes [v1; ref status: approved 1, http:/f1000r.es/w6]. *F1000Research* 2.
43. Seed KD, Lazinski DW, Calderwood SB, Camilli A. 2013. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* 494**:**489-491.
44. Makarova KS *et al.* 2011. Evolution and classification of the CRISPR–Cas systems. *Nat Rev Micro* 9**:**467-477.
45. Mekalanos JJ. 2011. *The evolution of Vibrio cholerae as a Pathogen*, Humana Press Inc, 999 Riverview Dr, Ste 208, Totowa, Nj 07512-1165 USA.
46. Hochhut B, Waldor MK. 1999. Site-specific integration of the conjugal Vibrio cholerae SXT element into prfC. *Mol Microbiol* 32**:**99-110.
47. Lin W *et al.* 1999. Identification of a Vibrio cholerae RTX toxin gene cluster that is tightly linked to the cholera toxin prophage. *Proc Natl Acad Sci USA* 96**:**1071-1076.
48. Karaolis DKR, Lan RT, Kaper JB, Reeves PR. 2001. Comparison of Vibrio cholerae pathogenicity islands in sixth and seventh pandemic strains. *Infect Immun* 69**:**1947-1952.
49. Jermyn WS, Boyd EF. 2002. Characterization of a novel Vibrio pathogenicity island (VPI-2) encoding neuraminidase (nanH) among toxigenic Vibrio cholerae isolates. *Microbiology* 148**:**3681-3693.
50. Dziejman M *et al.* 2002. Comparative genomic analysis of Vibrio cholerae: Genes that correlate with cholera endemic and pandemic disease. *Proc Natl Acad Sci USA* 99**:**1556-1561.
51. Davis BM, Kimsey HH, Kane AV, Waldor MK. 2002. A satellite phage-encoded antirepressor induces repressor aggregation and cholera toxin gene transfer. *Embo J* 21**:**4240-4249.
52. Yamasaki S *et al.* 1999. The genes responsible for O-antigen synthesis of Vibrio cholerae O139 are closely related to those of Vibrio cholerae O22. *Gene* 237**:**321-332.
53. Andrews RM *et al.* 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* 23**:**147-147.
54. Heidelberg JF *et al.* 2000. DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae. *Nature* 406**:**477-483.
55. de Sa Morais LLC *et al.* 2012. Complete genome sequence of a sucrose-nonfermenting epidemic strain of Vibrio cholerae O1 from Brazil. *J Bacteriol* 194**:**2772.
56. Reimer AR *et al.* 2011. Comparative genomics of Vibrio cholerae from Haiti, Asia, and Africa. *Emerg Infect Dis* 17**:**2113-2121.
57. Aydanian A, Tang L, Morris JG, Johnson JA, Stine OC. 2011. Genetic Diversity of O-Antigen Biosynthesis Regions in Vibrio cholerae. *Appl Environ Microbiol* 77**:**2247-2253.

Chapter 3

Ancient DNA analysis of a unique calcified urogenital infection from Medieval Troy reveals insights into past female health

Alison M. Devault[1,2], Tatum D. Mortimer[3,4], Henrike Kiesewetter[5], Tracy Smith[3,6], Melanie Kuch[1,2], Abrar Hossain[1], Jacob M. Enk[1,7,8], G. Brian Golding[7], William Aylward[9], Hendrik N. Poinar[1,2,7,10], Caitlin S. Pepperell[3,6]

(1) McMaster Ancient DNA Centre, McMaster University, Hamilton, ON, Canada; (2) Department of Anthropology, McMaster University, Hamilton, ON, Canada; (3) Department of Medical Microbiology and Immunology, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI, USA; (4) Microbiology Doctoral Training Program, University of Wisconsin-Madison, Madison, WI, USA; (5) Department of Anthropology, University of Tübingen; (6) Department of Medicine (Infectious Diseases), School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI, USA; (7) Department of Biology, McMaster University, Hamilton, ON, Canada; (8) MYcroarray, Ann Arbor, MI, USA; (9) Department of Classics, University of Wisconsin-Madison, Madison, WI, USA; (10) Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, ON, Canada

3.1 Thesis Preface

The following paper explores the genomic content of two pathogens from a calcified urogenital infection from the skeletal remains of a young woman from medieval Troy, in Turkey. The calcified nodules were originally thought to possibly represent tubercles from pulmonary tuberculosis. However, the age at death of the young woman and the nodules' morphologies and urogenital metagenomic content indicate that it is far more likely that they represent a reproductive-associated infection such as chorioamnionitis (infection of the placenta and amniotic membranes). Such reproductive-associated infections are today associated with maternal and fetal sickness and even fatality, and were probably associated with very poor health outcomes throughout history.

These calcified nodules offer the first-ever chance to study ancient urogenital pathogens *Staphylococcus saprophyticus* (associated today with urinary tract infections in young women) and *Gardnerella vaginalis* (associated today with bacterial vaginosis). The high quality and abundance of their DNA allowed for the reconstruction of the most complete ancient pathogen genome to date for *S. saprophyticus*, and the full coding genome of *G. vaginalis*. I found that the ancient Troy strains possessed almost all of the expected virulence factors associated with virulent modern strains of these pathogens, and provide a new anchor point to date the gain/loss of some genomic islands (such as antibiotic resistance). More interestingly, the

associated presence of these two species in the calcified nodule argues for a similar physiological role of these species in past reproductive-associated illness. Our findings imply that modern lineages of *S. saprophyticus* are host-specialized, but that ancestor of modern and Troy human *S. saprophyticus* may have been acquired zoonotically, or was a generalist.

This offers the first ancient DNA study of either species of bacteria, since tissues from the urogenital environment would not be expected to typically survive archaeologically. The very high percentage of combined *S. saprophyticus* and *G. vaginalis* DNA in the calcified nodules equals or surpasses that of the comparably well-preserved >40% *M. leprae* DNA in a medieval tooth from Denmark. This surprisingly high abundance of pathogen (and human) DNA indicates that calcified extraskeletal tissues should be targeted in future studies as rich sources of microbiome DNA. Indeed, the recent success in TE-free reconstruction of a *B. melitensis* genome (brucellosis) genome from a medieval calcified nodule indicates the great potential of such samples to provide a diverse portrait of ancient microbiomes (1). A brief survey reveals that many other calcified tissues have been reported in the archaeological literature (e.g., putative urinary stones, uterine masses) that may warrant revisiting using aDNA techniques.

3.2 Abstract

Ancient DNA studies of past commensal internal microbiome and soft tissue infectious diseases are severely limited because the requisite human remains are rarely preserved archaeologically. Examining human-associated bacteria throughout time is critical for understanding the evolution of the human microbiome, environmental and lifestyle contributors to infectious diseases, and reproductive health. We report on the remarkably well-preserved ancient DNA content of a calcified urogenital infection from the skeletal remains of a young female from the Late Byzantine period at the site of Troy, likely representing a placental or female reproductive tract abscess. The extraordinary DNA preservation allowed us to reconstruct a nearly complete genome of ancient *Staphylococcus saprophyticus*, as well as the coding gene content of ancient *Gardnerella vaginalis*. Today these two bacterial species are occasionally pathogenic components of the modern urogenital microbiome, respectively linked to urinary tract infection and bacterial vaginosis. Our findings provide an extremely unique and significant window into past female health.

3.3 Introduction

Pathogenic bacteria such as *Yersinia pestis* (2, 3), *Mycobacterium leprae* (4), and *Mycobacterium tuberculosis* (5) have been the typical targets of palaeomicrobiological genome-scale investigations because their DNA is preserved inside durable, ubiquitous skeletal remains.

In contrast, ancient DNA (aDNA) from commensal or pathogenic microbiome bacteria from soft tissues and body cavities are extremely difficult to study, since these tissues are unlikely to survive long periods of burial. That said, there has been recent success in applying new technologies to extraordinary sample types, such as in the sequencing of the *Vibrio cholerae* genome from alcohol-preserved 19th century intestine specimens (6, 7), oral bacteria partially preserved in medieval dental calculus (8, 9), intestinal flora in faecal remains (e.g. 10), and skin and organ bacteria in desiccated and frozen mummies (e.g. 11, 12).

We have identified unique and well-preserved calcified nodules caused by urogenital infection in the skeletal remains of a ~30 year old 13th century woman from Troy that allow for the first genomic study of an ancient urogenital microbiome. Extraskeletal calcified specimens have been previously identified and diagnosed with palaeopathological methods (13-16), though rarely utilizing aDNA techniques (1, 17). In addition to human DNA (>30%), the nodules are composed of two bacterial species, *Staphylococcus saprophyticus* (>40%) and *Gardnerella vaginalis* (>5%) as well as other microbiome bacteria at lower abundance. *S. saprophyticus* and *G. vaginalis* are associated today with urinary tract infections (UTIs) and bacterial vaginosis (BV) respectively, both important urogenital pathologies. Interestingly, the two species have highly contrasting genomic structural variation and evolutionary dynamics. *S. saprophyticus* has a stable core genome that has maintained much gene content and synteny even in comparison to distantly related species such as *S. aureus*, with well-defined genomic islands (18, 19). In contrast, *G. vaginalis* is highly recombinogenic and exhibits very high variation in sequence content and a small core genome (20, 21).

To achieve a detailed analysis of the ancient urogenital microbiome and firmly associate the nodule with the skeleton, we generated robust DNA sequence datasets from the nodules, an arm bone from the individual herself, as well as from the sediment surrounding her skeleton. Due to only two publically available *S. saprophyticus* genome sequences, we generated sequence data for ten additional modern *S. saprophyticus* isolates to provide a more comprehensive comparative phylogenetic assessment.

3.4 Methods

A summary of the methods follows; detailed methodological descriptions can be found in the Supplementary Appendix. Radiocarbon dates were obtained for the bone and nodules, and nodule 1 and 2 were analyzed using x-ray diffraction (XRD) and scanning electron microscopy (SEM). DNA was extracted from both nodules, a portion of an ulna bone, and sediment from the site (each alongside negative control blanks) in order to assess any external or cross- contamination. The nodules and bone were subject to typical aDNA extraction

methods for skeletal tissue and the sediment was extracted using a soil-specific kit. All DNA extracts were prepared for sequencing on the Illumina platform using both damage-removing ("UDG") and non-damage-removing ("nonU") library preparation methods: UDG for all bioinformatic genome reconstructions and metagenomic analyses, nonU for verifying aDNA damage patterns of assemblies and two rounds of human mitochondrial enrichment on the bone to link the nodules to the skeleton.

Reads were trimmed of adapter sequence and sequences <24bp were discarded. A nearly full genome sequence of *S. saprophyticus* Troy was obtained using a combination of paired-end reference-guided assembly to the modern strain ATCC 15305 and iterative assembly. Reference-guided consensus genome sequences were generated from ten new modern *S. saprophyticus* isolates (eight human, one bovine, and one canine) for the purpose of phylogenetic comparison, as only two reference genomes were publically available (human and rockfish strains). Full and core genome alignments were generated between the Troy strain, newly sequenced strains, and the available human and rockfish genomes. The gene content of *G. vaginalis* Troy was obtained using *de novo* assembly of paired-end and unpaired reads mapping to all annotated genes from 34 modern strains. A core genome alignment was generated between the Troy strain and modern strains.

Metagenomic analysis for the purpose of intersample comparison (nodules, bone, and sediment) was performed on subsets of 2 million merged reads with blastn/megablast to the nt database. Metagenomic analysis to investigate microbiome content was done on the full nodule 1 dataset (read 1) after removal of reads mapping to *S. saprophyticus* (Troy consensus) and *G. vaginalis* (ATCC 14019) genomes.

3.5 Results & Discussion

3.5.1 Burial 177, grave 14, quadrat x24, Troy

The site of Troy, also known as Troia, in western Anatolia (modern day Turkey) and the putative site of the Trojan War, was occupied from at least the Bronze Age to the Byzantine period of the Middle Ages. By the Late Byzantine (12-13th centuries AD), the site was occupied by a small village community of farmers, in contrast to the populous fortified citadel of earlier periods (22-24). Based on burials in several Late Byzantine cemeteries, life expectancies were low. The skeletal remains reflect palaeopathological indicators of hard labor as would be expected from a farming community, as well as evidence for degenerative and dental disease (22). Faunal remains of this time period include cattle, pigs, sheep, and goats, similar to earlier occupations at the site (23).

The individual in burial 177 (**Figure 1**) of grave 14 in the Late Byzantine cemetery in quadrat x24 was determined to be female and approximately 30 years old at the time of her death. Her remains were interred in a stone cist oriented west-east with no material goods. The cemetery was dated to the Late Byzantine period (12^{th}-13^{th} centuries AD) based on archaeological assessment, which is in good correspondence with the calibrated ^{14}C radiocarbon date of 1154-1224AD (0.98 probability distribution at 2 sigma) on collagen extracted from the ulna bone. She had lost all teeth during life (based on the left mandible, the only present mandibular/maxillary remains), had osteomyelitis in the left ulna and radius, and a healed fracture at the distal end of the same ulna (22, 25).

3.5.2 Snapshot of an ancient urogenital infection

The calcified nodules (**Figures 1 & S1**) were found inside the ribcage, which is consistent with being located in the abdominal cavity during life. Based on their morphologies, the nodules were initially suspected to be calcified tubercles from pulmonary tuberculosis. However, this diagnosis is now very unlikely, since the urogenital species *S. saprophyticus* and *G. vaginalis* were found in great abundance, and no *M. tuberculosis* DNA was recovered at our sequencing depth (**Figure 2**). While some Mycobacteriaceae reads are present, many species in this taxon are commonly found in soil and other natural environments (26), and therefore the presence of this taxon is not indicative of pathogenic *M. tb*. We attempted to directly ^{14}C radiocarbon date both carbonate and organic fractions from both nodules, but the results were unreliable (see Supplementary Appendix).

Extracted DNA from the nodules yielded astonishingly high proportions of endogenous DNA (**Figures 1 and S4; Table S1**): 12-26% human, 31-59% *S. saprophyticus*, and 4-6% *G. vaginalis* in metagenomic assessment. Sequencing data yielded 30X unique read coverage of the human mitochondrial genome, the consensus of which belongs to mitochondrial haplotype U3b3. An identical set of SNPS and haplotype was recovered from the ulna bone following targeted enrichment (**Table S16**), indicating that the nodules and bone are from the same individual or a close maternal relative. Haplotype U3b3 is commonly found today in the Caucasus region (27). The human and bacterial non-UDG treated reads all demonstrated typical ancient DNA damage patterns (**Figures S7 & S8**) and fragment length distributions (**Figures S9-S11**). In contrast to the nodules, the metagenomic profile of the ulna was very similar to that of the sediment (**Figures S4 and S5**). This suggests either a greater susceptibility of the skeletal remains to contamination with exogenous DNA over time or much higher starting endogenous DNA content of the nodules (thus that the same overall rate of contamination is less influential on the resulting metagenomic profile).

All Nodule 1 reads that did not map to *S. saprophyticus* and *G. vaginalis* were subjected to metagenomic analysis, and the family-level bacterial content is displayed in **Figure 2**. Despite the removal of the mappable reads, the majority of the remaining reads are still Gardnerella (in family Bifidobacteriaceae) and Staphylococcus (in Staphylococcaceae), due to incomplete removal when mapping to the single references. Overall, there is a very low diversity of other bacteria observed at this sequencing depth. However, ongoing growth by *S. saprophyticus* and/or *G. vaginalis* at the site of these presumably severe, active infections would likely have dominated the bacterial profile and disrupted the tissue microbiome flora.

The presence of *S. saprophyticus* and *G. vaginalis* almost certainly indicate a urogenital origin for the calcified infections, though they do not conclusively pinpoint the specific tissue of origin (i.e. tubo-ovarian abscess, endometrium, cervix, vagina, or renal tract). There are several possibilities for the cause of these calcified urogenital infections/abscesses. One likely candidate is chorioamnionitis, a potentially fatal infection of the amniotic membranes and placenta with urogenital bacteria that gain access to the upper reproductive tract during pregnancy (28). Reproductive-associated infections can cause premature labor and maternal and fetal sepsis (28). These and other obstetric complications were undoubtedly major factors in mortality of women throughout history (29), as they are today (30). Archaeological palaeodemographic findings from Late Byzantine Troy pointedly reveal low overall life expectancies – 32 for females, 39 for males – and higher mortality rate for females of reproductive age (22). Chorioamnionitis is a strong candidate for the calcified nodules, due to the age and sex of the young woman, the severity and morphologies of the nodules' calcification, and the specific urogenital microbiome bacteria recovered.

An alternative diagnosis may be that the nodules represent renal or urinary calculi (also known as kidney/bladder stones, or urolithiasis). These are mineral aggregates of different composition, such as calcium oxalate, calcium phosphate, or struvite, which form under a variety of different physiological conditions (31). Such calculi are also known to be serious obstetric complications (32). While both *S. saprophyticus* and *G. vaginalis* can be found in the vaginal flora (33, 34), *S. saprophyticus* is primarily associated with urinary disease (UTI) and both genera can be found in the female urinary microbiome (35, 36). *S. saprophyticus* infection has been directly associated with urinary calculi in humans (37) and rats (38). These calculi were comprised of ~52% struvite (magnesium ammonium phosphate) and ~38% carbonate apatite in the human (37) and struvite in the rats (38), accompanied in the latter by severe abscess and degradation of the bladder wall. However an XRD analysis of nodule 1 (**Figure S2**) shows the sample to be primarily composed of two phosphate phases, hydroxyapatite (bioapatite) and whitlockite (as well as small amounts of calcite), both of which have been found in calcified

concretions (e.g., 39). The noticeable absence of struvite and calcium oxalate (both common constituents of urinary stones) in the XRD and SEM-EDS analyses (average 42.3% oxygen, 42.1% calcium, 7.53% phosphate, and 3.18% magnesium, by weight; **Table S3**) supports a urogenital origin for the nodules.

3.5.3 Staphylococcus saprophyticus Troy

High coverage and overall similarity to the modern reference genome allowed for the completion of an effectively full ancient *S. saprophyticus* genome via a combination of reference-guided and iterative assembly. This now represents the most-complete ancient bacterial genome sequence to date. A small number of regions were not included in the final genome: five rRNA regions were not reconstructed due to the inability to reliably assign reads to a single locus, as were three repetitive stretches (in uafA, in aas, and between reference strain genes SSP0690 and SSP0691). These eight sections have been replaced with 100Ns in the final genome sequence. The *S. saprophyticus* Troy genome is 2,471,881bp long (including masked regions as noted above), and the final assembly includes 10,784,010 readpairs with an average unique coverage of 244.7X. Compared to modern reference strain, the Troy strain has 6,800 SNPs and 7 missing and 12 novel "inserted" regions greater than 100bp (**Figures 3** and **S12; Table S8**). We also iteratively reconstructed a 22.6Kb novel plasmid, pSST1, which is essentially a truncated version of 36.9Kb plasmid pSSAP1 (strain MS1146).

Phylogenetic analysis

The complex epidemiology of *S. saprophyticus* is not completely understood. It is the second most common cause of urinary tract infection (UTI) in women of reproductive age, and is uncommon as a cause of UTI in other groups (40, 41). It has been identified in rectal, urinary, and cervical samples from healthy, asymptomatic women, although this is relatively uncommon (42). *S. saprophyticus* has also been identified in samples from a range of mammals, as a food contaminant, and in the environment (40, 41). Strains isolated from patients and food sources have been shown to be diverse and lacking distinct patterns related to virulence or geography, though similar strains have been found in UTI patients separated by both time and location (19, 41, 43). UTIs in women have a seasonal spike in the summer months (42), and there is also a preceding seasonal rise in *S. saprophyticus* infection in cattle and pig populations (44, 45). Based on phylogenetic analysis (maximum likelihood) of the Troy strain, ten new modern isolates, and the human and rockfish genomes (**Figure 3**), the Troy strain sits basal to all human strains (which include the canine strain, likely a very recent host jump). There is a considerable diversity within the human strains, which is in accordance with previous findings (19).

The very high quality of the ancient *S. saprophyticus* Troy genome and the sequencing of additional modern strains offers the possibility of inferring the timing of events in the species' evolutionary history; namely, constraining the timeline of adaptation to the urogenital environment by calculating the time of the last common ancestor of the human strains. The evolutionary history of Staphylococcus species is of immense clinical relevance, particularly *S. aureus* (responsible for internal and external infections), and especially its antibiotic-resistant forms such as the methicillin-resistant MRSA. For *S. aureus*, others have concluded that humans were the ancestral host, and introduction into animal populations were more recent host jumps associated with the expansion of Old World cattle domestication ~5500BP (46). In an attempt to date the divergence of the bovine and human strains, along with other nodes in the phylogeny, we calculated a core genome mean substitution rate with BEAST (47) using the Troy radiocarbon date as a tip calibration point. However, these results were inconclusive, due the very large confidence interval for the mutation rate posterior (mean = 3.5×10^{-7} substitutions per site per year; 95% highest posterior density = $1.1 \times 10^{-6} – 9.3 \times 10^{-11}$).

Genome content

As the DNA of the ancient strain was found at high abundance in what was clearly a severe internal infection, we can surmise that the etiology of *S. saprophyticus* urogenital infections during the Middle Ages was probably similar to today. There are number of specific factors which contribute to modern *S. saprophyticus* virulence and its adaptation to the human urinary tract that can be examined in the Troy strain.

Urinary tract adaptation. Urease is an important factor in both the adaptation of *S. saprophyticus* to the human urinary tract as well as in UTI-related formation of renal/bladder stones. Urease enzymes hydrolyze urea, which is used by the cells for nitrogen metabolism and facilitates survival in the urinary tract (43). The breakdown of urea can induce the precipitation of struvite or calcium-apatite due to the rise in urine pH (38). The Troy strain possesses all seven genes of the urease operon (ureABCEFGD cluster), though there is a nonsynonymous SNP in the ureA start codon thus that there are an additional 33 residues at the 5' end of the Troy protein. However, other non-urinary Staphylococcus strains also possess the urease cluster, and functional studies would be required to assess the level of urease activity and its relationship to stone formation. The Troy strain also possesses the dsdA gene (D-serine dehydratase), which metabolizes D-serine in urine (which can be toxic to bacteria) and is associated with adaptation to the urinary environment (48). The Troy strain also possesses a modified version of the lipase gene ssp, believed to function in virulence and is one of the major *S. saprophyticus* surface proteins (49, 50). The Troy strain ssp region is located in novel region #02. The Troy strain also has the regulatory genes agrA-D, sarA, and rot, which are believed to contribute to virulence

and are found in all modern strains surveyed (51).

Cell adhesion. Adhesion factors are important contributors to Staphylococcal virulence, as they allow for the colonization of surfaces and the formation of biofilms (52). There are a variety of surface adhesion proteins characterized in *S. saprophyticus.* The Troy strain does not possess the collagen- and fibronectin-binding surface protein SdrI found in of ~10% of modern human strains (but not ATCC 15305) and no modern animal strains (51, 53, 54). Troy also does not possess uafB, a large gene found on plasmid pSSAP1 (strain MS1146) that functions in bladder adhesion (55). The Troy genome does contain the very large adhesion gene uafA (6,951bp in ATCC 15305), though we were only able to partially reconstruct the locus because of a long imperfect repeat section; the Troy version is only 2,777bp (includes 100bp of Ns in the masked repetitive region). UafA was found in all human and animal isolates studied (51), including all newly sequenced strains in this study (although there is considerable variation in the repetitive region). The Troy genome also contains another major *S. saprophyticus* surface protein, aas, which is an autolysin and adhesion gene that binds fibronectin and uroepithelium (51, 56).

A paralogous set of the ispD, gutB, and tagB genes was found in novel Troy region #16 (along with tagE; **Table S13**). This may represent a duplicated gene cluster functioning similarly to the duplicated tar gene clusters found in *S. aureus* (57). Tar (and other) genes produce teichoic acids that are critical structural components of bacterial cell walls (58). The duplicate clusters in *S. aureus* have been proposed to provide flexibility in adhesion states that may relate to virulence (57). It is possible that the Troy strain possessed more adhesion flexibility due to greater plasticity for potential hosts, as Late Byzantine populations lived in close association with domesticated animals such as cattle and pigs. However, *S. aureus* strains use the tar pathway rather than the tag pathway and the duplicated regions are tandemly located on the genome (58), unlike in the Troy strain. Functional analysis of these proteins would be required to assess their purpose.

Antibiotic resistance. Novomycin resistance is a class trait frequently used in *S. saprophyticus* diagnosis (40) and is conferred by the gene gyrB with 2 key positions (glysine at 85, lysine at 140) (59). The gene is present in the ancient Troy strain with both of these residues conserved (but with 1 nonsynonymous and 5 synonymous SNPs). This indicates that the ancient strain may also have been novomycin resistant. Unlike all other human *S. saprophyticus* strains surveyed (33), including the eight new modern human strains of this study, the Troy strain does not appear to be fosfomycin resistant (due to lacking the entire genomic island containing the gene, vSs15305). While none of the other human strains we sequenced contain vSs15305 in its entirety, they all contain a region that maps to the fosfomycin-resistance gene, fosB (SSP1926)

(43). Only the rockfish, bovine, and Troy strains lack the gene. Thus we can putatively argue that the acquisition of fosfomycin resistance in the species post-dates the Middle Ages. vSs15305 encodes another antibiotic resistance gene against streptomycin (43), but the Troy strain and all except one of the modern strains (SRS000013) are lacking this gene, indicating it may be a very recent acquisition.

Capsule production. In the ATCC 15305 strain, there is a large 49Kb region termed SCC15305 that is likely a product of lateral gene transfer. This island functions in part to produce a polysaccharide capsule which protects the cell from host neutrophils of the immune system (60). The Troy strain does not contain this region, and in its place is a novel 58Kb island. It does not appear that any of the newly sequenced human or canine strains in our dataset contain a full homolog to SCC15305. Park *et al.* (2010) concluded the same from surveys of other modern strains (based on PCR and functional analysis) (60), although they found that other strains might contain other types of capsule polysaccharide genes. Kleine *et al.* (2010) also found that only ~1% of human strains (and no animal strains) possessed the *cap* genes of strain 15305 (51). There are 12 annotated genes functioning in capsule synthesis in another novel Troy region (#03), with top blast hits to *S. aureus*, *S. carnosus,* and *S. haemolyticus* (**Table S11**). Therefore, the Troy strain was likely capable of producing surface capsule polysaccharides.

Plasmid pSST1. The pSST1 plasmid is essentially a truncated version of pSSAP1 plasmid. Both plasmids have partial homology to the pSSP1 (strain ATCC 15305) plasmid and a portion of the ATCC 15305 genome (similar to pSSAP1) containing repA*,* alcohol dehydrogenase, and dehydrogenase genes (61) that is missing from the *S. saprophyticus* Troy chromosomal genome. See **Figure S13** for portion of the alignment between Troy genome, ATCC genome, Troy plasmid, and pSSAP1. The Troy plasmid contains sssF (found on pSSAP1 and pSSP1), with 99.9% identity with the sssF on pSSAP1. sssF is a cell-wall anchored protein believed to contribute to linoleic acid resistance (61).

3.5.4 Gardnerella vaginalis Troy

Due to the high variability in coverage and lack of synteny in the ancient reads when compared to even the closest full genome strain ATCC 14019, it was not feasible to reconstruct a contiguous *G. vaginalis* genome. Instead, we used a *de novo* approach to reconstruct the full Troy gene content using reads that assembled to the annotated coding regions of extant *G. vaginalis* strains. In this way, we can assess the gene content of ancient *G. vaginalis* Troy as compared to the modern strains, but cannot evaluate what (if any) genome content the ancient strain may have possessed that has since been lost. Using this method, we recovered 1,187 unique contigs (total length 1,435,761bp) corresponding to 972 annotated genes of modern *G.*

vaginalis. The final assembly has 2,034,514 readpairs and an average coverage of 68.9X. The number of newly annotated genes in the modern strains (3 full and 34 scaffold genomes) ranges from 1109-1337, so if the ancient *G. vaginalis* Troy strain possessed a similar genome size, we are underrepresenting its true gene content by approximately ~200-300 genes.

Phylogenetic analysis

The highly diverse species *G. vaginalis* is phylogenetically grouped into highly distinct clusters (see **Figure 4**). These clusters have recently been postulated to represent four different species on the basis of sequence divergence, very small core genome, and low lateral gene transfer between clusters (21). *G. vaginalis* is often (but not always) associated with bacterial vaginosis (BV). This and the large extant diversity within the strains (and because different strains have been found in the microbiome of one individual) mean that likely the virulence potential and disease outcome of *G. vaginalis* infection is very complicated (21).

Based on phylogenetic analysis (maximum likelihood method) of *G. vaginalis* core genome, the Troy strain falls inside clade 1 (**Figure 4**). When only including clade 1 strains (with a clade 2 strain as outgroup), the Troy strain is the most basal of one of the two clade 1 subgroups. *G. vaginalis* Troy is most closely associated with strain 315-A, which is from a vaginal sample of unknown pathological state, though the next closest-related strains are from patients with BV. There is no clear link between specific clinical outcomes (such as BV) and the presence, nor a particular clade, of *G. vaginalis*. Therefore, it is not possible to infer the role of *G. vaginalis* in this individual's mixed urogenital infection. However, sufficient information about the Troy *G. vaginalis* genome content exists for an assessment of its functional potential.

Genome content

Multiple *G. vaginalis* genes have been identified as important contributors to virulence potential, adaptation to the urogenital environment, carbohydrate metabolism genes, natural competence, and biofilm formation.

Recombination. *G. vaginalis* is noted for its high degree of lateral gene transfer (LGT) and homologous recombination between strains (20). This was supported in our data by the large number of uncovered regions (gaps) in the Troy assembly to the closest full genome ATCC 14019, 583 (assembly in **Figure S14**), compared to only 124 gaps for *S. saprophyticus* (**Figure S15**). We can infer that there has been much LGT, recombination, and/or divergence in the modern lineages relative to the Troy strain. The ability to take in extracellular DNA into the cell (competence) is one precursor for LGT between bacteria, and the Troy strain possesses key

proteins which enable natural competence, such as comEA, comEC, cinA, and recA (62).

Virulence and adhesion. The Troy strain possesses two of the main *G. vaginalis* virulence factors, vaginolysin (a toxin) and sialidase (functioning in host invasion) (63). The Troy vaginolysin gene is 1,551bp and has 68 SNPs (22 amino acid changes) compared to ATCC 14019. This version is truncated relative to the reference due to premature stops if the coding sequence was the same size, a feature that is also found in other strains (e.g. 409-05). All the Troy vaginolysin SNPs fall in the 5' half of the gene. The Troy strain has a 2,724bp sialidase A gene of the same length with 86 SNPs (29 amino acid changes) relative to strain ATCC 14019. The Troy strain also possesses genes with homology to many biofilm production and epithelial adhesion genes (62), indicating that it was capable of urogenital colonization.

Carbohydrate metabolism. Several carbohydrate metabolism genes have only been found in clade 1 strains, many of which are endometrial isolates. This has been speculated to mean these strains are potentially associated with specialized adaptation to certain areas of the female reproductive trait (endometrium vs. vagina), though there is no direct evidence (21). The Troy strain contains a portion of a clade 1-only operon contain beta-galactosidase (lacZ) and alpha-L-fucosidase (fuc1), but not the associated ABC-type sugar transporter genes. These genes function in the degradation of glycoprotein (a component of host mucosal layers) and likely contribute to virulence (21, 62).

3.6 Conclusion

The high coverage and quality of *S. saprophyticus* DNA in these calcified nodules has enabled the reconstruction of possibly the most complete ancient bacterial genome to date. The ancient *S. saprophyticus* genome shares most of the key urinary and virulence adaptations of modern strains. The absence of fosfomycin resistance (shared by all extant strains) allows us to tentatively date its acquisition to post-Middle Ages. We were unable to use molecular dating to reconstruct the timing of divergence between human-bovine *S. saprophyticus* or the ancestor of the human lineages. However, the phylogenetic relationships between the human, canine, bovine, and rockfish strains do not indicate that humans are the ancestral or original host species. Rather, they suggest that human *S. saprophyticus* was originally acquired zoonotically from a non-human host, or that the ancestor of both human and bovine strains was a generalist. The canine-isolated strain likely represents a very recent host switch. This is in stark contrast with the pattern of *S. aureus* populations, in which all currently sequenced bovid-isolated strains fall within the diversity of the human strains, rather than outside (46). The distinct phylogenetic position of the bovine strain compared to all human/canine strains – even the ~800 year old Troy strain – strongly argues for the existence of host-specialized lineages of *S.*

saprophyticus.

The *G. vaginalis* Troy genome contains many key genes that function in modern urogenital adaptation and virulence. There is still an unclear and complicated relationship between phylogenetic clade and clinical outcomes in *G. vaginalis*, so we cannot conclude what specific role *G. vaginalis* played in the urogenital infection of the individual in burial 177. However, we clearly demonstrate that the Troy strain falls into clade 1 on the basis of its core genome, and it has clade 1-only genes that may be associated to specialization within body regions of the host. *G. vaginalis* has been identified in cases of chorioamnionitis (64), and virulent isolates could enable the ascension of other organisms (in this case, *S. saprophyticus*) to the upper reproductive tract. Our findings provide support for a putatively pathogenic role of *G. vaginalis* in this Late Byzantine individual.

The high abundance of *S. saprophyticus* and *G. vaginalis* DNA in the calcified nodules strongly argues for a urogenital origin of infection. The nodules' existence and the specific genomic characteristics of these two pathogens allow us to infer similar pathological roles for these bacterial species ~800 years ago. We are limited in observations of other microbiomic bacteria in the dataset due to the high abundance of these two pathogens and human DNA. However, it is unlikely that a pathological infection microbiome would be representative of a normal endogenous organ flora. Therefore, the nature of the samples likely precludes us from studying any deeper complexity of the microbiome even with deeper sequencing.

A recent study on high endogenous microbial and human DNA in dental calculus describes a depositional scenario of layered biofilms of large numbers of bacterial cells embedded in a calcified matrix (8). While the taphonomy is different, we can postulate that the calcified nodules formed a similarly well-preserved, high-DNA aggregate due to a proliferating, biofilm-forming bacterial community coupled with immune response from the host. The high percentage and quality of endogenous DNA indicate that if archaeological calcified extraskeletal tissues are recovered, they should be targeted for genetic study as potentially excellent sources of ancient microbiome DNA.

3.7 Acknowledgements

This research was supported by a seed grant from the Michael G. DeGroote Institute for Infectious Disease Research (IIDR), McMaster University. HNP is supported by the Canada Research Chairs program and the Natural Sciences and Engineering Research Council (NSERC) of Canada. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under grant no. DGE-1256259 to T. Mortimer. We

thank the Troia Project archaeological team, especially Gebhard Bieg, for providing insights and images. We also thank Kurt Reed, Pam Ruegg, Scott Weese, and Amanda Harrington for providing *S. saprophyticus* strains. We also thank Victoria Jarvis of the Brockhouse Institute for Materials Research (McMaster University) for help with XRD, and Marcia Reid of the Electron Microscopy Facility (McMaster University) for help with TEM/SEM-EDS analyses. We also thank John Southon (Keck Carbon Cycle AMS Facility, UC Irvine, CA) for radiocarbon dating and providing valuable analysis. We also thank Shea N. Gardner and Jonathan E. Allen of the Lawrence Livermore National Laboratory for performing ongoing related metagenomic analyses that will be published at a future date. We acknowledge helpful feedback by the members of the McMaster Ancient DNA Centre in our ongoing study of ancient pathogen genomes.

3.8 Figures

Figure 1. Burial #177 at Late Byzantine Troy, a female approximately 30 years old

(a) In-situ photograph of burial 177 (grave 14, cemetery in quadrat x24) at the site of Troy. This female (mitochondrial haplotype U3b3) was approximately 30 years old, had lost all teeth during life (based on the available skeletal remains; left mandible), suffered from osteomyelitis of the left ulna/radius, and had a healed fracture of the same ulna (22). Calcified nodules were found amongst the remains that are of urogenital origin based on XRD, SEM/TEM, and ancient DNA analysis. Photo credit Gebhard Bieg, 2005. (b) Images of nodules 1 (upper) and 2 (lower). Photo credits C. Pepperell, 2013. (c) Metagenomic shotgun read percentages based on BLAST analysis and parsing with MEGAN5 (65). "Gard" = *Gardnerella vaginalis*, "Staph" = *Staphylococcus saprophyticus*. Nodule 1 = other = 18.4%, unassigned = 6.8%, no hit = 12.4%; Nodule 2, other = 9.7%, unassigned = 3.0%, no hit = 12.5%. (d) Map of Europe showing the location of Troy, in modern Turkey (green dot).

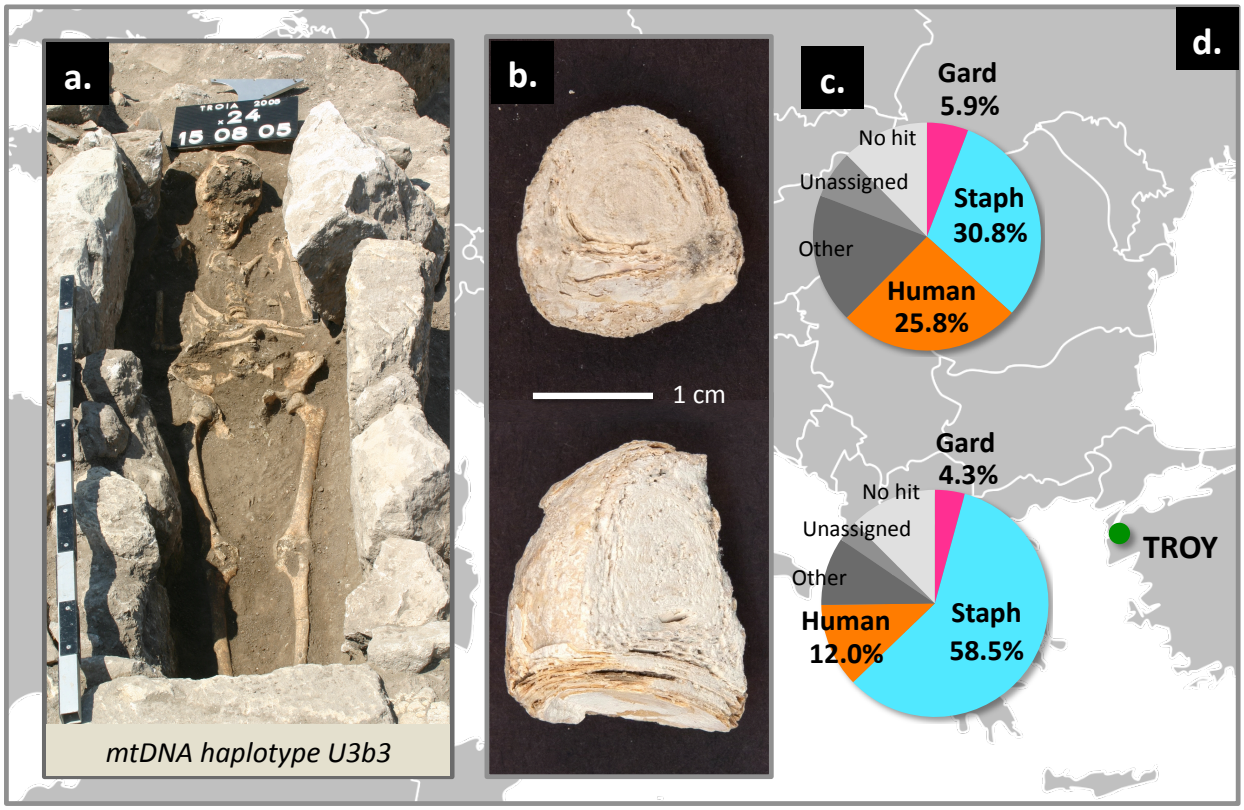


Figure 2. The nodule bacterial metagenome (family-level)

Only bacterial hits are shown. All R1 reads (trimmed, >24bp) were first curated to remove all reads mapping to the *S. saprophyticus* Troy genome and *G. vaginalis* strain ATCC 14019 with bowtie2 (66) and default settings, then subjected to blastn analysis retaining top 50 hits (67) and parsed with MEGAN5 (65) with 5 minimum reads/node. Tips (family-level only) are colored and sized by area according to their summarized read counts (log scale). All tips are <400 reads except Bifidobacteriaceae (75,463) and Staphylococcaceae (174,499). Figure was generated using FigTree (v.1.4.0).

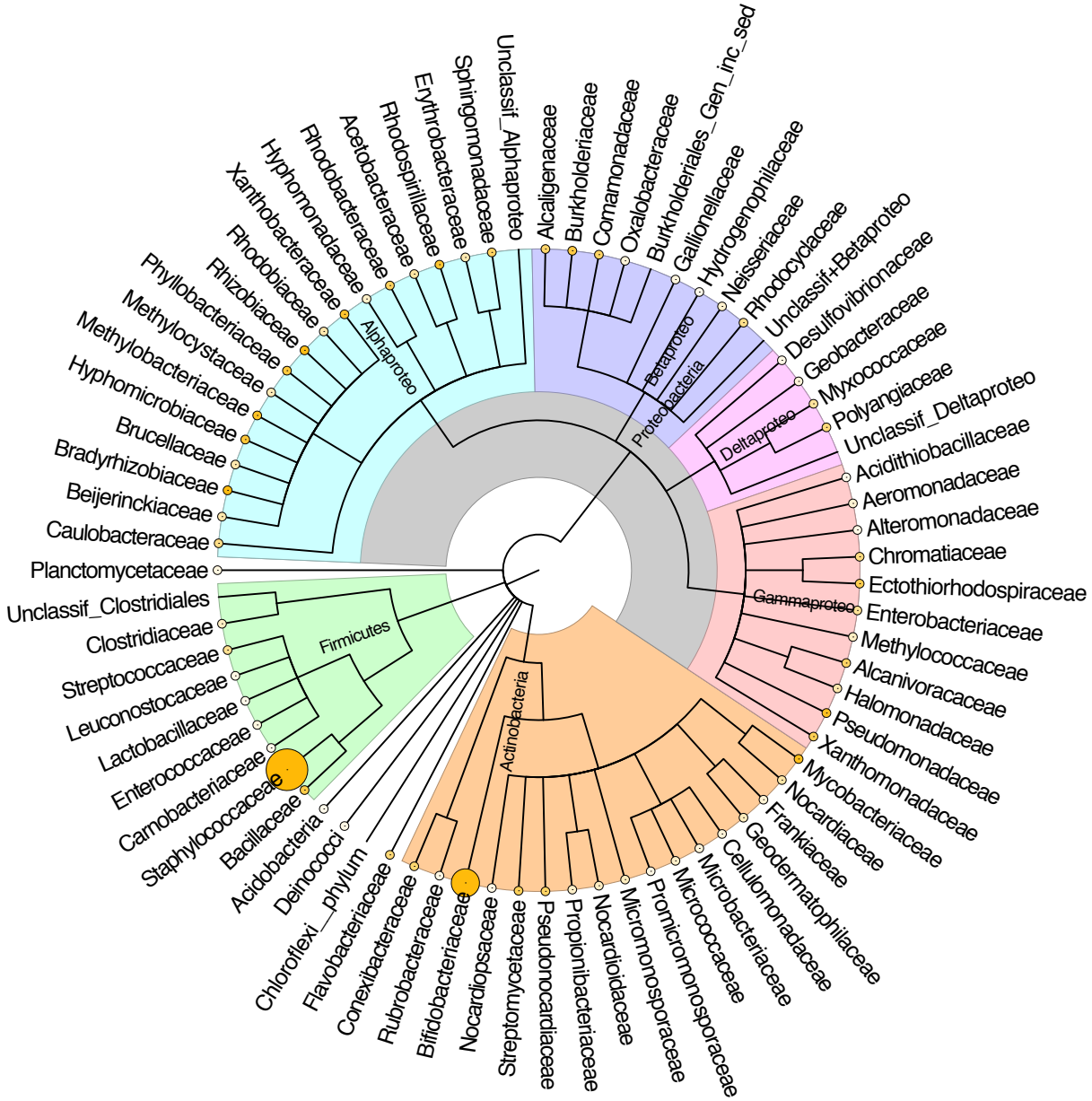


Figure 3. *Staphylococcus saprophyticus* phylogenetic tree and genome content comparison. (a) Full genome BLAST comparison generated using BRIG (68), comparing the available rockfish + ten new genomes to human reference strain ATCC 15305. Strains are ordered by host (from inside to outside: rockfish, bovine, canine, and the rest human) and colors match phylogenetic tree labels. Troy strain is the outermost ring (royal blue), above which are shown novel Troy regions (also royal blue) and labeled features of interest (red = rRNA regions; yellow = reference genes that could not be resolved in Troy; grey = truly absent Troy regions; black = genomic islands of interest in reference). (b) Maximum likelihood tree of the core genome SNP alignment of all strains, using rockfish as outgroup. Nodes are labeled with their bootstrap values. The Troy strain is the most basal of the clade that includes all humans + canine strain.

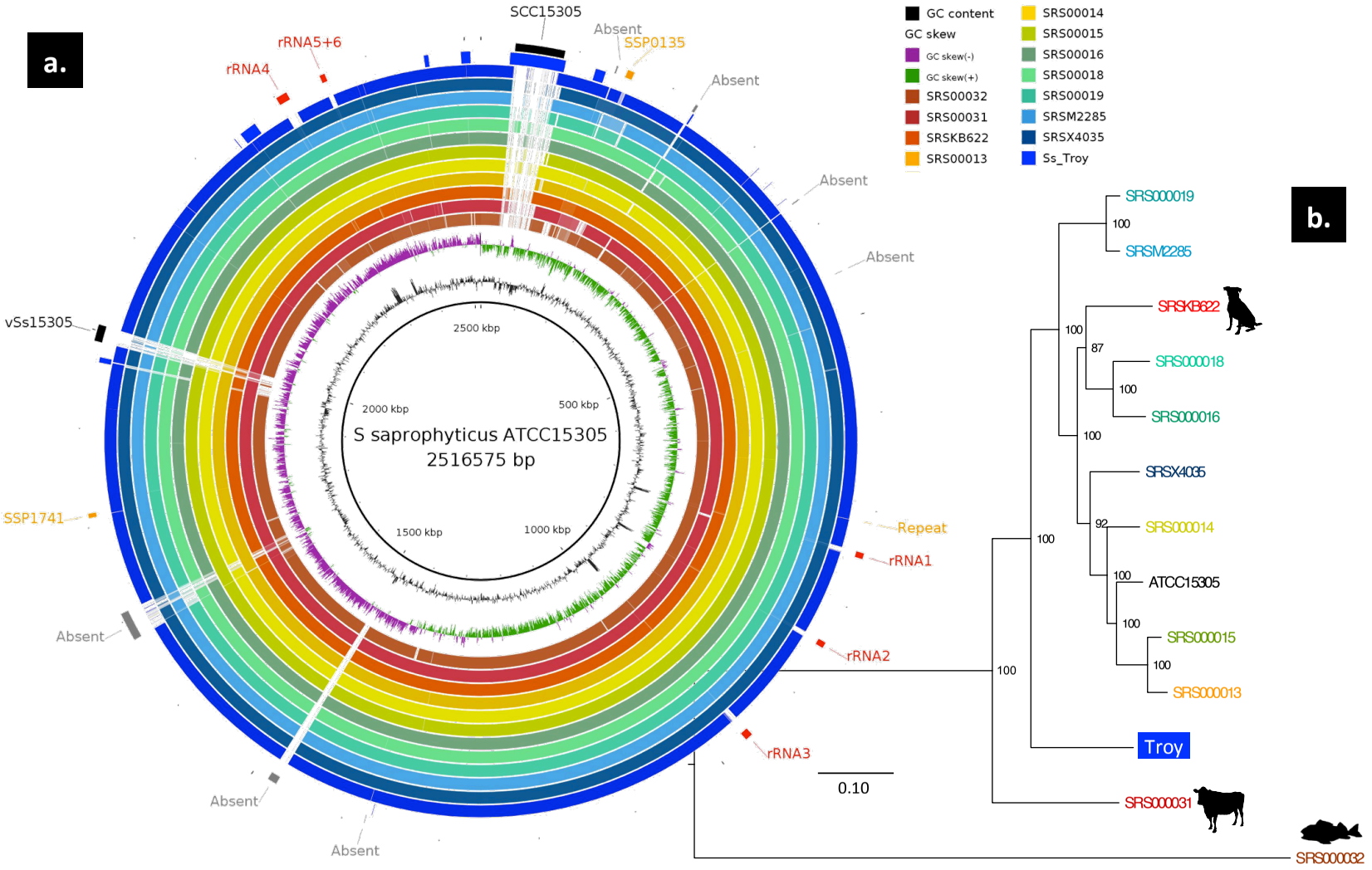
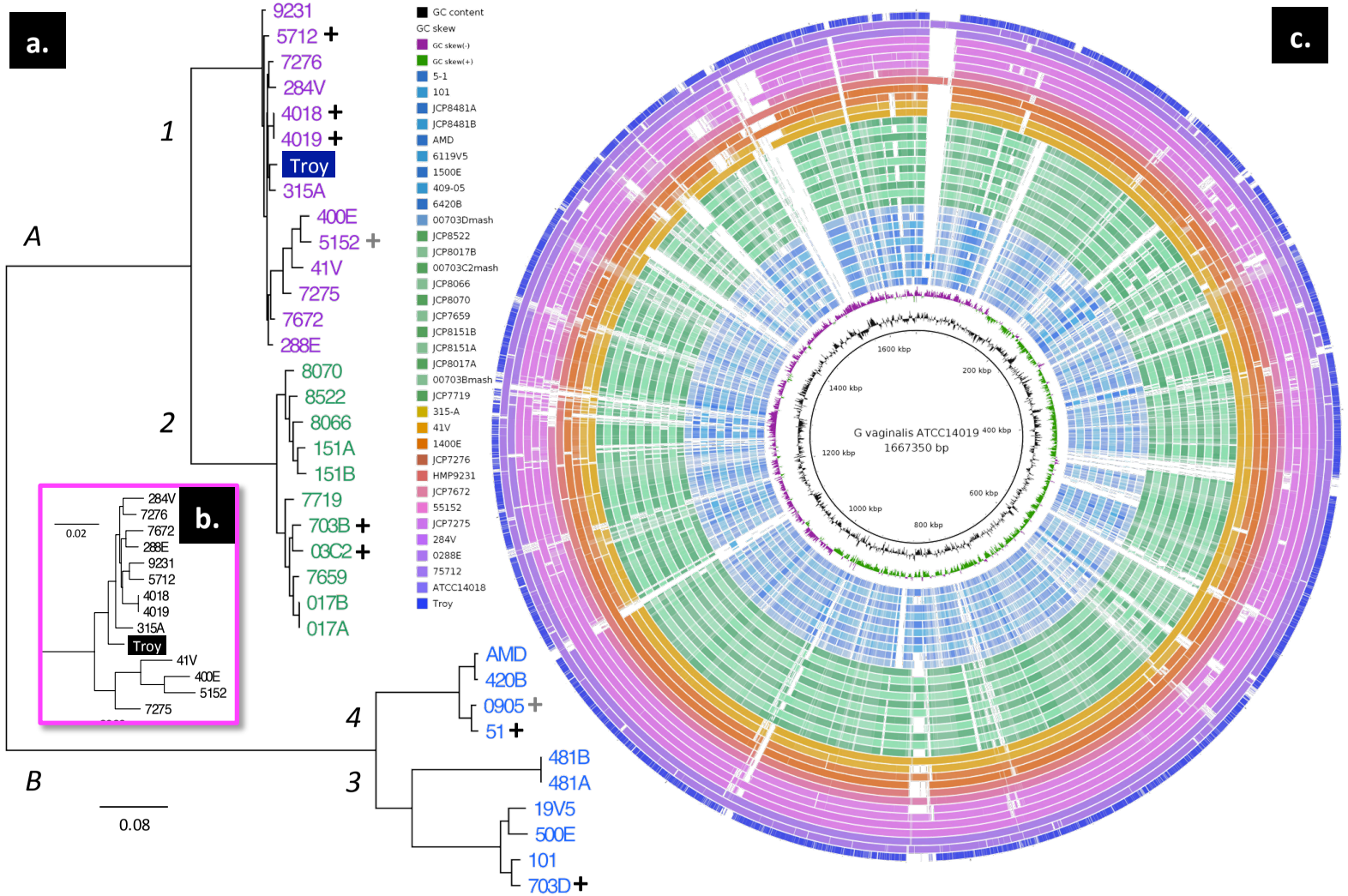


Figure 4. *Gardnerella vaginalis* phylogenetic tree and genome content comparison. (a) Core genome maximum likelihood tree, generated from alignments of single-copy genes in all genomes, determined by OrthoMCL (69). Clade labels follow ref. (21), who argue that clades 1-4 may represent distinct species. Strains marked with "+" if associated with BV (black) or asymptomatic BV (grey). (b) Inset; clade 1 core genome tree (outgroup = clade 2 strain 8066; not shown). (c) Full genome BLAST comparison generated using BRIG (68), comparing all full/scaffold genomes to clade 1 strain ATCC 14019. Troy strain is the outermost ring (dark purple). Strains are ordered and colored by clade: 3 & 4, inner, blues; 2, middle, greens; 1, outer, orange to purple.



3.9 Supplementary Appendix

The following sections provide supplementary information on Methods (1.10), followed by Supplementary Tables (1.11) and Figures (1.12).

3.10 Supplementary Methods

3.10.1 Samples and DNA extraction

Details on samples and their associated tests and treatments can be found in **Table S2.** Ethics approval for the study of the remains of the individual in burial 177 (grave 14, quadrat 24) was obtained from Hamilton Health Sciences and McMaster University (REB# 13-146-T). Sediment from the Troy site was imported to and studied at McMaster University in accordance with Canadian Food Inspection Agency guidelines, under permit number P-2012-04220.

Radiocarbon dating

Subsamples of both nodules and the ulna bone were sent to Keck Carbon Cycle AMS Facility (John Southon, Earth System Science Department, UC Irvine, CA) for radiocarbon dating. The ulna returned a calibrated date of 1154-1224AD (.98 probability distribution at 2 sigma; uncalibrated 860 ± 20BP) which is in accordance with the archaeologically established date for the quadrat x24 Troy burials in the 13th century AD (22, 25).

We attempted to ^{14}C radiocarbon date both the carbonate (with 10-30% leaching) and the organic fraction from both nodules post demineralization of carbonate with 1N HCL and gelatinization with 0.01N HCL. The diversity of ^{14}C dates (460 – 1130yrs BP uncalibrated) from the carbonate and organic fractions of the nodules (**Table S2**; for example the extremely young date on the 10% leached carbonate sample) is suggestive of possible ground water leaching and exchange. We made additional attempts to date any organic residues by base treatment (0.1N NaOH) to remove possible soil carbon contamination, post acid demineralization of the carbonate: this resulted in a very old date of 4,680 yr BP (uncalibrated), that could be the result of residual refractory soil carbon (humin). Given these range of dates, the results for the nodules must be viewed with caution, in contrast to that of the bone.

X-ray diffraction (XRD)

A subsample of nodule 2 was subjected to mineralogical analysis using XRD at the Brockhouse Institute for Materials Research (McMaster University) using the Bruker D8

DISCOVER with DAVINCI.DESIGN diffractometer (**Figure S2**). Sample flakes were piled on top of a single crystal silicon wafer, and aligned to the centre of the diffractometer using the laser-video alignment system. The detector to sample distance was calibrated with corundum to 20cm. Four frames were collected to obtain the 2θ range of 8-103°. The frames were integrated into intensity plots using DIFFRAC.EVA V.3.0 (software package from Bruker AXS). A pattern search/match was executed using the integrated ICDD PDF-2 2011 powder database. Slight mismatch in the peak positions are likely due to variation of elemental stoichiometry in the identified phase.

Scanning Electron Microscopy (SEM) with Energy Dispersive X-ray Spectroscopy (EDS)

A subsample of nodule 2 was subjected to elemental analysis using SEM-EDS at the MAX Diffraction Facility (McMaster University). Sample pieces were attached to an aluminum stub with double-sided carbon tape and sputter-coated with a thin layer of Au. The sample was viewed in a Tescan Vega II LSU operating at 20kV (**Figure S3**). Energy dispersive spectroscopy (EDS) was carried out with an IMAX detector (Oxford Instruments) and INCA software (**Table S15**).

Ancient DNA Extractions

Details of the DNA extractions and their associated libraries can be found in **Table S4**. "Bone" DNA extractions (summarily consisting of demineralization (DM), removal and freezing of DM supernatant, incubation of un-demineralized tissue with a custom digestion buffer (DB), removal and freezing of DB supernatant, organic extraction of one/both supernatants, and concentration via filtration) were performed as follows. For Set A, multiple consecutive rounds of DM (with 0.5M EDTA) and digestion were performed on a shaker (1000rpm), with the supernatant(s) from each round subjected to organic extraction and filtration. DB consisted of 20mM Tris pH 8.0, 0.5% sarcosyl, 250 µg/ml Proteinase K, 5mM $CaCl_2$, 50mM DTT, 1% PVP, and 2.5mM PTB. The breakdown of DM/DB rounds is as follows: round 1 = 1^{st} 0.5 mL overnight (ON) DM at room temperature (RT), 2^{nd} 0.5 mL DM at RT for 24h, and 7h digestion at 55°C, rounds 2-5 = ON 1 mL DM + 7h digestion at 55°C, and round 6 = ON 1mL DM only. For subsequent extraction, all round 1 supernatants were combined and all round 2-6 supernatants were treated separately: supernatants were subjected to organic extraction using half-volume of phenol-chloroform-isopropanol (centrifuged at 16,000 x g for 5m), the aqueous phase of which was extracted with 750 µl chloroform (centrifuged as before). The final aqueous phase was filtered using Amicon Ultra 0.5mL 10kDa columns (EMD Millipore Corp., Billerica, MA, USA): columns primed with 450 µl 0.1xTE, followed by sample filtration, washed twice with 450 µl 0.1xTE, and eluted in 50 µl 0.1xTE. For Set B, 1 mL DM was performed ON rotating at RT and the supernatant was frozen. Digestion was performed for 7h rotating at 55°C using 0.5 mL of DB

(same recipe as Set A) and the supernatant was frozen. These supernatants were combined and subjected to organic extraction and filtration as in Set A. Set D followed the same protocol as Set B, except they were subjected to an additional initial 30 minute demineralization with 500 µl 0.5M EDTA that was not extracted, and the DB did not contain DTT, PVP, or PTB and was performed ON.

Set C "sediment" DNA extractions were performed using the Mo Bio PowerSoil DNA Isolation Kit (MO BIO Laboratories, Inc., Carlsbad, CA) following the manufacturer's protocol, with a final elution of 100 µl in 0.1xTE.

3.10.2 Library preparation and indexing

Ancient DNA extracts were converted to double-indexed libraries for sequencing on the Illumina platform (list in **Table S4**). Prior to library preparation, all Ulna DNA extracts (E5-1, 2, 3, 4, 7, 8, 9, & 10) were pooled to homogenize library input into multiple UDG and non-UDG libraries, as were the 2 set D extraction blanks (E5-6 & 12). Libraries were prepared as in reference (2) using either regular ("non-UDG") or deaminated cytosine removal (damage repair; "UDG") protocols, and subsequently amplified using a double-indexing protocol.(70, 71) Each library set included at least one blank no-template (water) control reaction. Extract input volumes into library preparation were 10 µl (L25-L26), 20 µl (L01-L20), or 25 µl (L28-L38, 1a-1j, and 1a-blk to 1j-blk). In non-UDG libraries L25-L38, the blunt-ending step was modified with an extended (3h) T4 PNK incubation prior to adding the T4 polymerase, in the same manner as the UDG protocol. Double-indexing amplification was performed as in (2) for 10 cycles each, with 20 µl non-diluted library template input (except L25 and L26 which were used at 0.1x dilution) and included at least one no-template negative control reaction. All reactions were purified to 15 µl EB with the MinElute PCR Purification kit (Qiagen, Hilden, Germany).

3.10.3 Enrichment

Two rounds of human mitochondrion targeted enrichment were performed on the non-UDG treated ulna specimen for comparison to the nodule shotgun reads. Prior to enrichment, the 7 Ulna-non-D libraries (L31-L38) were pooled in order to homogenize input, and 9ul of this pool was used as input into 4 enrichment reactions (Ulna-D E07-E10) alongside the extraction blank (E11) and a negative control reaction (enrichment blank). The enrichments were performed just as for the human enrichments in reference (2) using the same parameters and custom MYbaits kit (MYcroarray, Ann Arbor, MI), except the protocol was modified to conform to updated manufacturer recommendations: post-hybridization bead-library binding was performed rotating at high temp (55°C), Wash Buffer 1 was eliminated, and the post-washed

beads were suspended in 20 µl EB and used directly in the post-enrichment amplification. Round 1 used 2 µM of four P5/P7 adapter sequence custom blocking oligos (corresponding to one strand of each molecule) and round 2 used the manufacturer-supplied Block #3.

Amplification after each enrichment round was performed as in reference (2), with additional re-amplifications as required due to low output molarities (all amplification reactions were purified over MinElute columns to 15 µl EB). Post-round 1 amplification used 15 µl bead mixture directly as input into each 40 µl reaction (15 cycles). 6.5 µl of this purified reaction was used as input into enrichment round 2 (E17-E21), and to increase molarity prior to sequencing, 6.5 µl was used as template for subsequent re-amplification reactions (12 cycles). Post-round 2 amplification used 10 µl bead mixture as input into two 40 µl reactions per enrichment (15 cycles), and the supernatants from each amplification (2 per sample) were purified together. Prior to sequencing, 14 µl of this purified reaction was used as the template for a subsequent re-amplification reaction (16 cycles).

3.10.4 Sequencing and read preparation

All relevant ancient samples and their associated extraction blanks were sequenced. Details on the final dataset names, their associated samples/libraries/enrichments, raw reads passing filter, and pre-sequencing pooling schemes can be found in **Table S5**. Prior to sequencing, all additional indexed libraries (shotgun and enrichment) were quantified via a qPCR assay targeting indexed molecules and pooled according to desired sequencing ratio. All pools except Nod1_all (pool “K”) were size-selected using an electrophoresis gel size selection procedure (retaining molecules ~125/150 to 500bp in length) in order to exclude as much no-insert adaptimer (and other short adapter artifacts) as possible. Pool “F” also contained 10 additional samples not considered in this paper (pooled at a ratio of “1”). Pools were sequenced across 3 paired-end runs on the HiSeq 1500, all alongside other, unrelated samples: Pool “K” (80bp final read length), Pool “F” (85bp final read length), and pools “G”-“J” (90bp final read length). On the latter run, the enrichment round 1 & 2 samples (pools “H” and “J”) were separated on two different lanes, since they have the same index sequences.

All ancient datasets were trimmed of library adapter using cutadapt with settings -e 0.16, -O 1, -a AGATCGGAAGAGC (72) and reads <24bp were removed using a custom script to retain read order.

3.10.5 Modern S. saprophyticus isolates

Ten new *S. saprophyticus* isolates (eight human, one bovine, and one canine) were sequenced for this study in order to provide a broader comparative genomic dataset (details in **Table S6**). DNA was extracted from the isolates at the University of Wisconsin. Isolates were inoculated in TSB and grown overnight at 37°C in a shaking incubator. Cultures were pelleted, resuspended in 140 µL TE, and incubated overnight with 50 units of mutanolysin. DNA was extracted using the MasterPure Gram Positive DNA Purification Kit (EpiCentre).

Extracts were prepared for sequencing with the Illumina Nextera XT library preparation kit and sequenced in two different batches: samples SRS00013-31 were pooled in equal ratio with 24 additional unrelated samples and sequenced on a MiSeq platform using a 2x250 kit; samples SRSKB622, SRSX4035, and SRSM2285 were sequenced at a desired ratio alongside pools "G"-"J" (described above) as part of the HiSeq 1500 paired-end run (85bp read length). Average insert sizes (estimated from BioAnalyzer analysis) are 849bp (SRS00013-31, pooled average), 775bp for SRS0M2285, 716bp for SRS0X4035, and 769bp for SRS0KB622.

Datasets for the new modern samples were processed with reference-guided assembly via a pipeline [https://github.com/tracysmith/RGAPepPipe]. Read quality was assessed and trimmed with TrimGalore! [www.bioinformatics.babraham.ac.uk/projects/trim_galore], a wrapper script for FastQC [www.bioinformatics.babraham.ac.uk/projects/fastqc] and cutadapt (72). Reads were mapped to the ATCC 15305 reference genome using BWA-MEM [bio-bwa.sourceforge.net] and bam files sorted using Samtools (73). Read group information was edited and duplicates removed using Picard [picard.sourceforge.net]. Reads were locally realigned and variants were called using GATK (74). VCFs produced by GATK were converted to alignments using custom Python scripts. Genomes were annotated using Prokka (75).

3.10.6 Staphylococcus saprophyticus Troy

Genome reconstruction

Draft genome was reconstructed using iterative assembly to span gaps between contigs that were created from assembly to the *S. saprophyticus* reference genome (NC_007350). This procedure was used based on both the (a) high coverage and quality of the initial assembly to NC_007350 that encouraged spanning gaps as a feasible strategy and (b) incomplete results obtained from attempting true *de novo* assembly (i.e., only very short contigs obtained using ABySS, Velvet, or SOAPdenovo), likely because of the metagenomic nature of the sample.

Trimmed reads from Nod1_all-UDG were paired-end assembled to the *S. saprophyticus* reference (NC_007350) using BWA with default settings (76) and duplicates were removed

using samtools rmdup (73). The resulting assembly was imported into Geneious (v.6.1.6, Biomatters, Ltd) and a strict (50%) consensus sequence was generated. From this consensus, 65 large contigs (880bp – 170,993bp) that corresponded to regions of average coverage (and that did not span rRNA/tRNA regions) were manually extracted, which represented the non-gap regions of the assembly. As gap regions could represent indel regions (e.g., lateral gene transfer events), rearrangements, or divergence, these contigs were subjected to an iterative assembly process using a set of unmapped reads in order to attempt to span gaps and connect the contigs. The primary set of reads used for iterative assembly was generated by trimming 100bp from each end of the contigs, assembling all original reads using the above settings to this set of trimmed contigs, and removing these assembled reads from consideration. A subset of the unmapped reads (20-100% as required, depending on assembly success) along with the full set of contigs were then subject to iterative assembly using Geneious (v.6.1.6), seeded with the first or last 50bp of a contig (settings: maximum gaps per read 10%, maximum gap size 2, word length 20, index word length 14, ignore words repeated >8x, maximum mismatches per read 1%, maximum ambiguity 4, map multiple best matches randomly). Iterative assembly uses the consensus from the current round of assembly as the reference for the subsequent round. Depending on the size of the gap, multiple rounds of iterative assembly were required to span certain gaps. Gaps were considered closed when the iterative process spanned to the end of another contig. In certain cases, assembly would dead-end without reaching another contig, likely because of insufficient reads due to homology on a contig; if so, the iterative process was completed using the full readset. All non-rRNA-adjacent gaps were spanned, except for the region that belongs on the plasmid rather than the chromosome (see below).

Resolving duplicated/dual-motif regions

SSP0354-0356. In novel region/insert #16, there is an apparent divergent paralogous/duplicated region of NC_007350 genes SSP0354 (ispD; 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase), SSP0355 (zinc-binding dehydrogenase) and SSP0356 (glycosyl glycerophosphate transferase involved in teichoic acid biosynthesis) inside of the novel GI. The duplicated region precluded straightforward iterative or *de novo* assembly due to challenges in resolving which version was found in the same location as the reference and which version was found in the novel region #16. The challenges arose from the partially conserved, partially high divergent nature of the duplicated region (which is problematic for both assembly-guided and *de novo* assembly) as well as the short reads precluding the spanning of SNPs (haplotypes) by long paired-end reads. Proper gene sequences were established via a combination of (a) manually working inwards from the shared flanking regions with liberal assembly to manually determine correct sequence and (b) iterative assembly to *de novo* reconstruct highly divergent sections. For the majority of positions, the ancient versions of SSP0354-56 do match the

reference genome and the novel copies are the 'unmatching' positions; however, in several instances, the reads support that the 'reference matching' positions are in fact on the paralogous copies instead of the homologous ancient SSP0354-56 genes.

SSP0117, SSP0118, & SSP0788. As for SSP0354-0356 and their paralogs discussed above, the sequences of the Troy versions of the SSP0117, 0118, & 0788 paralogous hypothetical proteins were obtained through a combination of manually working inwards from the shared flanking regions with liberal assembly to manually determine correct sequence and iterative assembly to *de novo* reconstruct highly divergent sections.

Genome verification. All potential reads were assembled to a penultimate draft version of the genome (a combined product of reference-guided assembly consensus contigs, iterative *de novo* assembly, and manual curation as noted). Ambiguous positions (from the contig and iterative assembly consensus sequences) or positions heterogeneous in the reads (>40%) were resolved with the reference base (if that motif was present in the reads >40%). All Nod1_1h-UDG reads were assembled to the final genome draft using paired-end bowtie2 assembly (66) with default settings and samtools rmdup, and all positions were confirmed with at least 35X unique coverage. The final genome is 2,471,881bp long (including eight masked regions as noted below) and the final assembly includes 10,784,010 readpairs with an average unique coverage of 244.7X. The genome was annotated using Prokka (75).

Unresolved regions. There were three repetitive regions of the genome that were unable to be reliably resolved using either reference guided or iterative assembly due to the length and sequence structure of the region, and were replaced by 100 N's in the final genome sequence. These regions correspond to strain ATCC15305 positions 1810158-1810757 in SSP1741, 153694-158013 in SSP0135, and 725232-725331 between SSP0690 and SSP0691.

Regarding rRNA regions, no attempt was made to reconstruct or span these regions with iterative assembly, as they are repeat regions and would be impossible to assemble with short reads. All rRNA regions (16S, 23S, 5S genes & any flanking tRNAs) plus 200bp of flanking sequence from the ends of the rRNA genes have been replaced by 100 N's in the final ancient genome sequence (corresponding to strain 15035 positions 743525-749373, 841186-847217, 959438-967218, 2349930-2355284, and 2301048-2312686). Any tRNAs that were not flanked by rRNA genes were included in the draft genome. The iterative assembly process established the presence of at least the same number of bordering RNA regions as in the reference sequence, although it cannot be determined whether there is any rearrangement of the intervening regions in the Troy strain versus the reference.

S. saprophyticus Troy strain compared to ATCC 15305

A list of gaps and inserted regions >100bp in *S. saprophyticus* Troy as compared to strain ATCC 15305 was generated from manual assessment of a MAUVE alignment with default settings (**Table S7**). Details of the annotated CDS sequences in these 16 inserts are listed in **Table S8**. **Figure S12** displays the gene diagram of five large novel regions of particular interest. For each of these regions, each annotated CDS sequence was BLASTed with default settings with blastn to the nr database (May 2014) retaining the top 100hits, and the top hit is reported (**Tables S9-S13**).

S. saprophyticus Troy plasmid, pSST1

A partially-novel plasmid was detected during the assembly process and closed using iterative assembly. Large portions of this pSST1 plasmid have homology to the pSSAP1 (strain MS1146) and pSSP1 (strain ATCC 15305) plasmids, and a portion of the ATCC 15305 genome (similar to pSSAP1) containing repA*,* alcohol dehydrogenase, and dehydrogenase genes (61) which is missing in the Troy chromosomal genome (but located on the plasmid). The plasmid was annotated using Prokka (75), and each annotated CDS sequence was BLASTed with default settings with blastn to the nr database (May 2014) retaining the top 100hits, and the top hit is reported (**Table S14**). **Figure S13** displays a MAUVE alignment (default settings) using Geneious (v.6.1.6) of Troy plasmid pSST1, the Troy genome, the ATCC 15305 modern genome, and modern plasmid pSSAP1 from strain MS1146.

3.10.7 Gardnerella vaginalis Troy

Gene reconstruction

Ancient gene sequences were reconstructed *de novo,* via assembly of a pool of reads that mapped to annotated *G. vaginalis* CDS sequences. First, CDS annotations were extracted from 34 modern *G. vaginalis* full and scaffold annotated genome sequences (**Table S15**; all except for strains 41V and 101) and concatenated into 1 reference/genome (100 N's in between each CDS sequence). Then trimmed paired end reads from Nod1_1h-UDG reads were assembled to the set of 34 CDS concatenated references. All paired and unpaired reads that mapped were extracted and subjected to *de novo* assembly using Velvet (version 1.2.1)(77) with settings kmer 23, insert length 51, expected coverage 75, minimum contig length 24, and coverage cutoff auto (parameters for expected coverage were chosen based on previous observation of assembly to strain ATCC 14019). This generated 1,207 contigs that were analyzed using blastn (default settings) to the nr database (April 2014) to detect any non-*G.*

vaginalis sequences. 20 contigs were excluded due to the top hit being *S. saprophyticus* (indicates partially conserved regions), leaving a final set of 1,187 *G. vaginalis* contigs (total length 1,435,761bp). The final set of genomic contigs was annotated using Prokka (75), producing 972 genes. Paired-end assembly of Nod1_1h-UDG reads to the final set of contigs with bowtie2 (66) and removal of duplicates with samtools rmdup (73) consists of 2,034,514 readpairs with an average coverage of 68.9X.

3.10.8 Phylogenetic analysis

Staphylococcus saprophyticus Troy

Alignments were made between all sequenced *S. saprophyticus* genomes (human strain 15305, NC_007350 (43); rockfish strain M1-1, AHKB00000000 (78)), the 10 newly sequenced strains, and Troy genome. A whole genome alignment was obtained with Mugsy (79). Core SNPs were obtained from the whole genome alignment using a python script, excluding regions corresponding to a gap in any strain. RAxML was used to create a maximum likelihood phylogenetic tree from the core SNP alignment with the rockfish isolate as an outgroup. Bootstrap support for the tree was calculated using 100 bootstrap replicates.

A core genome substitution rate was calculated with BEAST (47) implementing several different substitution, clock, and demographic models, using the Troy date as a calibration point. MCMC were run for 100 million generations with sampling every 10,000 generations, and the first 30 million generation were discarded as burn-in. These analyses did not return a reliable rate, due to the large confidence interval (3.5×10^{-7} [95% highest posterior density: $1.1 \times 10^{-6} – 9.3 \times 10^{-11}$] with General Time Reversible with gamma distributed site variation substitution model, lognormal relaxed clock with an exponential rate distribution, and Bayesian Skyline Plot demographic model.

Gardnerella vaginalis Troy

A core genome alignment was obtained with the Troy contigs and all strains in **Table S15** except 6240LIT and JCP8108 via a pipeline [https://github.com/tatumdmortimer/core-genome-alignment]. The protein sequences from the annotations by Prokka (72) were grouped using OrthoMCL (69) using all vs. all BLAST and clustering. Groups were filtered for those containing all the genomes of interest and no proteins present in more than one copy. The DNA sequences for the genes encoding these proteins were individually aligned with MAFFT (80) and trimmed of messy portions of the alignments using trimAl (81). Alignments were concatenated to form the core genome alignment. To generate a phylogenetic tree, core SNPs were obtained from

the core genome alignment using a python script, excluding regions corresponding to a gap in any strain. RAxML was used to create a maximum likelihood phylogenetic tree from the core SNP alignment. Bootstrap support for the tree was calculated using 100 bootstrap replicates.

3.10.9 Human mitochondrial genome

Reads from the two rounds of enriched human mitochondrial reads from the ulna (Ulna_Enr1-nonU & Ulna_Enr2-nonU) and UDG-treated Nodule 1 shotgun reads (Nod1_1h-UDG) and their respective extraction blanks were analyzed to compare their human mitochondrial content. In order to facilitate straightforward coverage comparison between these datasets, the paired end reads <24bp were first re-trimmed and merged using SeqPrep [https://github.com/jstjohn/SeqPrep] with an 11bp overlap (settings -L 24 -o 11 -A AGATCGGAAGAGC). All merged + non-merged R1 reads from each datasets were mapped to the human mitochondrial reference genome (Cambridge Reference Sequence; accession number NC_012920) using bwa (settings -n 0.02 -l 9999) (76). Duplicate reads were collapsed using a custom script that retains reads with unique 5' and 3' positions.

Numbers of unique (and raw pre-collapsed) merged reads mapping to the mitochondrial genome were as follows: Nod1_1h-UDG = 9,593 (9,765), Nod2-UDG = 557 (565), Ulna_Enr1-nonU = 8,085 (32,755), and Ulna_Enr2-nonU = 11,577 (3,693,050). These correspond to mean unique coverage values of 29.7 (Nod1), 2.0 (Nod2), 25.7 (Ulna_Enr1), and 39.7 (Ulna_Enr2). This was a successful enrichment, if we infer a high enrichment rate from the 1 (1) mitochondrial read in the similar (but not identical) Ulna-UDG shotgun library. Few to none extraction blank reads mapped (and none at SNP locations): EblkA-UDG = 1 (1), EblkB-UDG = 10 (10), EblkD_Enr1-nonU = 0, EblkD_Enr2-nonU = 1 (147).

Both rounds of enriched reads display a typical ancient DNA damage profile (**Figure S8a and S8b**) with higher C→T or G→A damage rates closer to fragment ends, generated using mapDamage2 (82). A pooled dataset of Nod1_1h-nonU and Nod2-nonU reads (merged and assembled the same as described above) also displays a typical ancient DNA damage profile (**Figure S8c**). This dataset was used for the damage assessment instead of the assembly used for haplogroup analysis because the nodule mitochondrial consensus was generated using a UDG-treated dataset (which removes the signal of damage) and because there were comparatively few reads from either Nod1- or 2-nonU datasets alone (**Table S5**).

Table S16 reports the coverage and variant frequency for all SNPs called for all samples with at least 3x coverage and 85% (nodule) or 60% variant frequency (ulna). At this variant frequency, the same 30 SNPs are called in both rounds of enriched reads and in nodule 1, but

fewer (plus 1 new SNP) are called nodule 2. This is most likely because the overall coverage in nodule 2 is insufficient to generate a complete mitochondrial profile with these coverage and variant frequency calling parameters. The lowered variant frequency necessary in the enriched ulna specimens to call the same SNPs as in the nodule is probably due to higher modern human contamination in these samples versus nodule 1 (though it is unfortunately not possible to infer this mathematically using damage level on this dataset, since the nodule 1 reads are UDG-treated, and the pooled Nod1+2_nonU dataset used for the mapDamage profile is low coverage, <1X). The mitochondrial haplotype called from both the enriched datasets and nodule 1 is U3b3, called using the whole mitochondria haplogroup assessment tool at [www.mitotool.org] (which uses data from PhyloTree build 16, available at [www.phylotree.org]).

3.10.10 Metagenomic analyses

Subsets from the 2 nodules, sediment, ulna and their respective extraction blanks were analyzed to compare their metagenomic profiles. The paired end reads <24bp were first re-trimmed and merged using SeqPrep [https://github.com/jstjohn/SeqPrep] with an 11bp overlap (settings -L 24 -o 11 -A AGATCGGAAGAGC). 2 million read subsets of the merged + non-merged R1 reads of Nod_1h-UDG, Nod2-UDG, Sediment-UDG, Ulna-UDG and all reads from their extraction blanks were subjected to blastn-megablast (v.2.2.29) against the nt database using default settings and retaining the top 100 hits per read (67). Blast results were parsed using MEGAN5 requiring minimum 5 reads/taxon (65). Profiles comparing the specimens (**Figure S4**) (with PCA plot, **Figure S5**) and extraction blanks (**Figure S6**) at phylum-level are provided.

3.10.11 Microbiome analysis

The largest full nodule 1 dataset (Nod1_1h_UDG, R1, trimmed of adapter, <24bp) was subjected to microbiome analysis. This dataset was first curated to remove all reads mapping to the *S. saprophyticus* Troy genome and *G. vaginalis* strain ATCC 14019 with bowtie2 (66) and default settings, in order to increase speed of analysis. The final dataset used for analysis was comprised of 19,197,285 reads. We performed a BLAST 2.2.29+ blastn analysis using default settings and retaining the top 50 hits. The blast results were parsed with MEGAN5 (minimum reads/node = 5) (65) and the bacterial results are displayed in **Figure 2**.

3.11 Supplementary Tables

Table S1. Abundances of *S. saprophyticus*, *G. vaginalis*, and human mitochondrial genome via paired-end assembly
Paired end-assemblies with bwa with default settings ("raw") and collapsed with samtools rmdup ("uniq"). Readcounts for mapped reads are for the first (or only) read of a pair.

Dataset ID	Sample	Reads, raw (R1)	Reads, trimmed >24bp (R1)	*S. saprophyticus*				*G. vaginalis*				Human mtDNA			
				Raw		Uniq		Raw		Uniq		Raw		Uniq	
				#	% R1t>24	#	% R1t>24	#	% R1t>24	#	% R1t>24	#	% R1t>24	#	% R1t>24
Nod1_1h-UDG	Nodule 1	32,230,769	**31,979,481**	10,508,649	32.9%	10,038,389	31.4%	2,005,305	6.3%	1,964,443	6.1%	9,695	0.0%	9,598	0.0%
EblkA-UDG	Ext. blk A	753	**684**	47	6.9%	47	6.9%	5	0.7%	5	0.7%	1	0.1%	1	0.1%
Nod2-UDG	Nodule 2	6,483,539	**6,468,263**	3,699,618	57.2%	3,626,390	56.1%	256,815	4.0%	254,311	3.9%	562	0.0%	557	0.0%
EblkB-UDG	Ext. blk B	6,009	**3,501**	2	0.1%	2	0.1%	2	0.1%	2	0.1%	10	0.3%	10	0.3%
Sediment-UDG	Sediment	3,969,139	**3,940,255**	427	0.0%	397	0.0%	314	0.0%	305	0.0%	6	0.0%	6	0.0%
EblkC-UDG	Ext. blk C	3,840	**2,680**	3	0.1%	3	0.1%	1	0.0%	1	0.0%	1	0.0%	1	0.0%
Ulna-UDG	Ulna	2,885,479	**2,850,132**	1,077	0.0%	1,067	0.0%	457	0.0%	453	0.0%	1	0.0%	1	0.0%
EblkD-UDG	Ext. blk D	488	**429**	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%

Table S2. Troy sample details

Name	Description	^{14}C Results			DNA Extractions
		Sample description	**Age (BP)**	**Age (BP), calibrated**	
Nodule #1	White colored, chalky	-	-	-	Set A
		Carbonate (no leach)	745 ± 25	-	
		Carbonate (10% leach; 0.27 mg C)	460 ± 20	-	
		Carbonate (30% leach)	765 ± 20	-	
Nodule #2	White colored, chalky	-	-	-	Set B
		"Organics", no pretreatment	765 ± 20	-	
		Organics (0.081 mg C)	1130 ± 25	-	
		Organics	1055 ± 20	-	
Sediment	Light brown, sandy, dry	-	-	-	Set C
Ulna	Distal portion	-	860 ± 20	1154-1224AD (0.98 prob dist at 2 sigma)	Set D

Table S3. SEM-EDS results from nodule

For each replicate, upper value is weight %, lower value is atomic %.

Frame	Repl.	O K	Mg K	Ca K	P K	Ni K	Au M	Al K	F K	Na K	Si K	Fe K
A	1	30.55	2.69	66.76	-	-	-	-	-	-	-	-
		5.8	3	45.19	-	-	-	-	-	-	-	-
	2	26.35	3.05	60.6	9.99	-	-	-	-	-	-	-
		45.66	3.48	41.92	8.94	-	-	-	-	-	-	-
	3	41.33	2.94	37.97	-	1.35	16.4	-	-	-	-	-
		68.74	3.22	25.21	-	0.61	2.22	-	-	-	-	-
B	1	50.33	2.37	46.06	1.25	-	-	-	-	-	-	-
		70.97	2.2	25.93	0.91	-	-	-	-	-	-	-
	2	26.69	-	62.28	10.12	-	-	0.91	-	-	-	-
		46.57	-	43.37	9.12	-	-	0.95	-	-	-	-
	3	41.17	1.56	30.16	13.63	-	-	0.57	10.66	2.24	-	-
		57.05	1.42	16.68	9.76	-	-	0.47	12.44	2.17	-	-
	4	38.67	1.83	32.47	14.35	-	-	0.77	8.97	2.95	-	-
		55	1.71	18.44	10.54	-	-	0.65	10.75	2.92	-	-
C	1	36.2	2.37	32.8	16.51	-	-	-	9.08	3.05	-	-
		52.35	2.25	18.94	12.33	-	-	-	11.06	3.07	-	-
	2	19.62	2.61	53.6	22.83	-	-	-	-	1.34	-	-
		35.38	3.09	38.58	21.26	-	-	-	-	1.69	-	-
	3	37.2	1.9	33.33	15.83	-	-	0.81	8.37	2.56	-	-
		53.72	1.8	19.22	11.81	-	-	0.69	10.18	2.58	-	-
D	1	58.68	3.38	35.21	2.73	-	-	-	-	-	-	-
		76.84	2.91	18.4	1.85	-	-	-	-	-	-	-
	2	52.01	2.71	43.56	1.72	-	-	-	-	-	-	-
		72.17	2.47	24.13	1.23	-	-	-	-	-	-	-
	3	53.29	2.66	44.18	-0.13	-	-	-	-	-	-	-
		73.39	2.41	24.29	-0.09	-	-	-	-	-	-	-
	4	56.03	3.26	33.67	5.03	-	-	-	-	2.01	-	-
		74.1	2.84	17.78	3.44	-	-	-	-	1.85	-	-
E	1	63.76	2.87	32.88	0.49	-	-	-	-	-	-	-
		80.68	2.39	16.61	0.32	-	-	-	-	-	-	-
	2	49.77	2.95	46.92	0.35	-	-	-	-	-	-	-
		70.47	2.75	26.52	0.26	-	-	-	-	-	-	-
	3	47.05	5.51	35.98	4.46	-	-	1.93	-	-	3.34	1.74
		66.37	5.12	20.26	3.25	-	-	1.61	-	-	2.68	0.7
	4	32.01	9.48	29.3	1.28	-	-	6.25	-	-	15.59	6.09
		49.29	9.61	18.01	1.02	-	-	5.7	-	-	13.68	2.69

Table S4. Troy DNA extraction details

Set	Type	Sample	Weight	ID	Assoc. Libraries		Assoc. Enrichments	
					UDG	non-UDG	Round 1	Round 2
Set A	"Bone"	Nodule 1	296 mg	Round 1 (both)	1a	-	-	-
				Round 2 (demin.)	1b	-	-	-
				Round 2 (digest)	1c	-	-	-
				Round 3 (demin.)	1d	-	-	-
				Round 3 (digest)	1e	-	-	-
				Round 4 (demin.)	1f	-	-	-
				Round 5 (demin.)	1h	L20	-	-
				Round 5 (digest)	1i	-	-	-
				Round 6 (demin.)	1j	-	-	-
		Extraction blank A	-	Round 1 (both)	1a-blk	-	-	-
				Round 2 (demin.)	1b-blk	-	-	-
				Round 2 (digest)	1c-blk	-	-	-
				Round 3 (demin.)	1d-blk	-	-	-
				Round 3 (digest)	1e-blk	-	-	-
				Round 4 (demin.)	1f-blk	-	-	-
				Round 5 (demin.)	1h-blk	L21	-	-
				Round 5 (digest)	1i-blk	-	-	-
				Round 6 (demin.)	1j-blk	-	-	-
Set B	"Bone"	Nodule 2	42.5 mg	E3-1	L01	L25	-	-
		Extraction blank B	-	E3-3	L03	L26	-	-
Set C	"Sediment"	Sediment	250 mg	E4-1	L04	L14	-	-
		Extraction blank C	-	E4-2	L05	L15	-	-
Set D	"Bone"	Ulna	63.4 mg	E5-1	*L28*	*L31-L37*	*E07-E10*	*E17-E20*
		Ulna	56.6 mg	E5-2	*L28*	*L31-L37*	*E07-E10*	*E17-E20*
		Ulna	54.7 mg	E5-3	*L28*	*L31-L37*	*E07-E10*	*E17-E20*
		Ulna	74.1 mg	E5-4	*L28*	*L31-L37*	*E07-E10*	*E17-E20*
		Extraction blank D	-	E5-6	*L29*	*L38*	*E11*	*E21*
		Ulna	92.4 mg	E5-7	*L28*	*L31-L37*	*E07-E10*	*E17-E20*
		Ulna	65.8 mg	E5-8	*L28*	*L31-L37*	*E07-E10*	*E17-E20*
		Ulna	50.7 mg	E5-9	*L28*	*L31-L37*	*E07-E10*	*E17-E20*
		Ulna	70.4 mg	E5-10	*L28*	*L31-L37*	*E07-E10*	*E17-E20*
		Extraction blank D	-	E5-12	*L29*	*L38*	*E11*	*E21*

**Demin. = demineralization; digest = digestion buffer*

Table S5. HTS Datasets

Dataset ID	Type	Sample	Library or Enrichment ID	Reads, raw (R1)	Reads, trimmed >24bp (R1)	Pool#	Within pool ratio
SHOTGUN DATASETS							
Nod1_all-UDG	UDG	Nodule 1	1a to 1j	8,346,525	8,255,071	K	1 (each lib equal, x9)
Nod1_1h-UDG	UDG	Nodule 1	1h	32,230,769	31,979,481	F	20
Nod1_1h-nonU	non-UDG	Nodule 1	L20	840,160	837,940	F	1
EblkA-UDG	UDG	Ext. blk A	1h-blk	753	684	F	1
EblkA-nonU	non-UDG	Ext. blk A	L21	748	266	F	1
Nod2-UDG	UDG	Nodule 2	L01	6,483,539	6,468,263	F	1
Nod2-nonU	non-UDG	Nodule 2	L25	728,551	727,608	F	1
EblkB-UDG	UDG	Ext. blk B	L03	6,009	3,501	F	1
EblkB-nonU	non-UDG	Ext. blk B	L13	1,664	1,139	F	1
Sediment-UDG	UDG	Sediment	L04	3,969,139	3,940,255	F	20
Sediment-nonU	Non-UDG	Sediment	L14	216,542	214,354	F	1
EblkC-UDG	UDG	Ext. blk C	L05	3,840	2,680	F	1
EblkC-nonU	Non-UDG	Ext. blk C	L15	7,297	3,122	F	1
Ulna-UDG	UDG	Ulna	L28	2,885,479	2,850,132	G	1
EblkD-UDG	UDG	Ext. blk D	L29	488	429	G	1
ENRICHED DATASETS							
Ulna_Enr1-nonU	Non-UDG	Ulna	E07/09/10	5,033,999	4,826,582	H	1 each lib
EblkD_Enr1-nonU	Non-UDG	Ext. blk D	E11	87	76	H	3
Ulna_Enr2-nonU	Non-UDG	Ulna	E17/19/20	5,518,292	5,496,111	J	1 each lib
EblkD_Enr2-nonU	Non-UDG	Ext. blk D	E21	707	670	J	3

Table S6. New modern *S. saprophyticus* strains sequenced in this study

Name	Host	Location	Raw reads (R1)
SRS000013	Human	USA	657,388
SRS000014	Human	USA	320,860
SRS000015	Human	USA	331,316
SRS000016	Human	USA	856,259
SRS000018	Human	USA	451,474
SRS000019	Human	USA	277,673
SRS000031	Bovine	USA	417,555
SRSKB622	Canine	Canada	5,591,696
SRSX4035	Human	USA	4,571,270
SRSM2285	Human	USA	4,342,040

Table S7. List of gaps and novel regions >100bp in *S. saprophyticus* Troy vs. strain ATCC 15305

Troy ID	Troy pos	Troy length	Ref pos	Ref length	Ref ID
Gap01	159,508+	0 bp	139,450-140,935	1,486 bp	-
Gap02	381,903+	0 bp	368,186-369,298	1,113 bp	-
Gap03	465,264+	0 bp	452,653-452,796	144 bp	-
Gap04	1,369,911+	0 bp	1,376,914-1,377,412	499 bp	-
Gap05	1,466,548+	0 bp	1,473,932-1,483,625	9,694 bp	-
Gap06	1,661,522+	0 bp	1,678,667-1,707,957	26,291 bp	-
Gap07	1,946,318+	0 bp	1,989,287-2,006,320	17,034 bp	vSs15305
Insert01	34,048-34,270	223 bp	34,047-34,051	5 bp	-
Insert02	35,290-93,091	57,802 bp	35,071-83,724	48,654 bp	SCC15305
Insert03	134,843-145,552	10,710 bp	125,472+	0 bp	-
Insert04	242,210-244,271	2,062 bp	227,805-230,552	3,156 bp	-
Insert05	331,411-331,915	505 bp	318,147-318,475	329 bp	-
Insert06	351,503-351,603	101 bp	338,063+	0 bp	-
Insert07	373,975-374,220	246 bp	360,501+	0 bp	-
Insert08	1,364,807-1,364,993	187 bp	1,372,001+	0 bp	-
Insert09	1,384,300-1,384,426	127 bp	1,391,800+	0 bp	-
Insert10	1,923,896-1,929,320	5,425 bp	1,970,966-1,972,289	1,324 bp	-
Insert11	2,045,420-2,045,643	224 bp	2,105,429-2,105,652	224 bp	-
Insert12	2,182,269-2,183,017	749 bp	2,242,501-2,242,514	14 bp	-
Insert13	2,199,713-2,217,902	18,190 bp	2,259,545-2,259,624	80 bp	-
Insert14	2,323,945-2,324,174	230 bp	2,382,493-2,382,623	131 bp	
Insert15	2,401,046-2,405,657	4,612 bp	2,459,870+	0 bp	-
Insert16	2,446,603-2,456,301	9,699 bp	2,500,767-2,500,993	227 bp	-

Table S8. *S. saprophyticus Troy* novel regions >100bp

Insert#	Length	Gene#	Genes description*	Notes*
01	223	0	-	-
02	57,802	49	Type IIS restriction enzyme Eco571; modification methylase Eco571B; HP; HP; HP; HP; divergent AAA domain protein; HP; recombination factor RarA; HP; HP; hsdR; putative type I restriction enzymeP M protein; hsdS; putative restriction endonuclease; HP; copB; sporulation lipoprotein YhcN/YlaJ family; sulfite exporter TauE/SafE; csoR; pspE; pksB; chrA; arsC; arsB; sdpR; HP; HP; HP; cycA; katE; lip; lip; zinc-responsive transcriptional regulator; hchA; yhfK; yfmJ; putative membrane protein; isaB; hopanoid-associated sugar epimerase; TspO/MBR family protein; HP: crtP; crtQ; crtM; crtN; ydaG; phrB; hmoB	Integrated in the site of SCC15305
03	10,710	10	3-phytase; Gram-positive signal peptide YSIRK family; putative membrane protein; HP; cbiO; putative ATPase ABC class; HP; HP; calcineurin-like phosphoesterase; HP	Extends HP (SSP0106)
04	2,062	3	HP; putative transposase OrfB; Transposase	Replaces 4 genes in 15305 (2 HPs; oxidoreductase; major facilitator superfamily permease)
05	505	1	Putative transglycosylase IsaA precursor (partial)	Alters HP (SSP0293)
06	101	0	-	-
07	246	1	nhaC (partial)	Alters Na+ H+ antiporter (SSP0329)
08	187	1	HP (partial)	-
09	127	0	-	-
10	5,425	9	7 HPs; YolD-like protein; Phage integrase family protein	Truncates/replaces haloacid dehalogenase (SSP1904); replaces HP (SSP1905)
11	224	1	HP	Alters HP (SSP2037)
12	749	1	kdgK (2-dehydro-3-deoxygluconokinase)	Alters HP (SSP2182)
13	18,190	20	LytR; epsN; eps; carbamoyl phosphate synthase-like protein; speG; epsL; tuaC; epsD; wbpA; putative adenylate-forming enzyme; rffG; HP; tuaG; HP; colanic acid biosynthesis protein; gtaB; pglF; ywqE; ywqD; cap8A	
14	230	1	lip (lipase precursor)	Alters lipase precursor (SSP2297)
15	4,612	5	yxeP; ydhP; ytcD; putative membrane protein; csbC	Extends carboxypeptidase (SSP2392)
16	9,699	6	tagE; HP; tagB; gutB; ispD2; HP	Contains paralogs of SSP0354-0456

* HP = hypothetical protein

Table S9. Novel *S. saprophyticus* Troy region #02: Gene descriptions and top BLAST hits

Troy CDS				Top BLAST hit			
#	Description	Locus#	Length	Description	Species*	Accession	Evalue
1	Type IIS restriction enzyme Eco571	00030	2,958	Putative type II restriction enzyme	L. mon.	FR733650	8.96e-68
2	Modification methylase Eco571B	00031	1,638	Adenine-specific DNA methylase	L. mon.	FR733650	3.78e-25
3	Hypothetical protein	00032	186	hsdR	S. aur.	FR821779	7.91e-60
4	Hypothetical protein	00033	120	*NO HIT*	-	-	-
5	Hypothetical protein	00034	477	Hypothetical protein	D. pur.	XM_003294330	2.30e-02
6	Hypothetical protein	00035	354	Hypothetical protein	S. sap.	AP008934	3.20e-138
7	Divergent AAA domain protein	00036	1,182	Hypothetical protein	S. aur.	AB505628	0
8	Hypothetical protein	00037	915	-	V. vin.	AM427540	2.08e-06
9	Recombination factor RarA	00038	933	DHHA1 domain-containing protein, DHH family protein	Alpha proteo. HIMB5	CP003809	9.02e-05
10	Hypothetical protein	00039	1,425	16S rRNA	S. inc.	KF751706	4.88e-04
11	Hypothetical protein	00040	750	Unnamed protein product ORF004	S. aur.	HF569109	5.15e-127
12	hsdR	00041	3,069	Type IC restriction subunit	L. car.	CP003851	0
13	Putative type I restriction enzymeP M protein	00042	1,545	Type I restriction-modification system subunit M	L. car.	CP003851	0
14	hsdS	00043	1,224	hsdS	S. aur.	CP006706	8.59e-32
15	Putative restriction endonuclease	00044	774	-	S. aur.	CP007176	0
16	Hypoth. protein	00045	621	Membrane protein	S. aur.	CP007176	0
17	copB	00046	2,064	Putative copper-transporting P-type ATPase B	S. cap.	KF049201	0
18	Sporulation lipoprotein YhcN/YlaJ family	00047	546	Hypothetical protein	S. aur.	CP007176	0
19	Sulfite exporter TauE/SafE	00048	756	Hypothetical protein	S. 693-2, plasmid	DQ390456	0
20	csoR	00049	261	Protein of unknown function DUF156	S. aur., plasmid	GQ900448	1.53e-121
21	pspE	00050	1,068	Hypothetical protein	S. 693-2, plasmid	DQ390456	0
22	pksB	00051	1,326	Hypothetical protein	S. 693-2, plasmid	DQ390456	0
23	chrA	00052	1,164	Chromate transport protein	S. sci.	HG515014	0
24	arsC	00053	396	Thioredoxin-coupled arsenate	S. cap.	KF049201	2.42e-

				reductase			178
25	arsB	00054	1,290	arsB (arsenical pump membrane protein)	S. xyl.	M80565	0
26	sdpR	00055	315	arsR (arsenical resistance operon repressor)	S. aur.	HF569116	6.17e-153
27	Hypoth. protein	00056	402	Hypothetical protein	S. sap.	AP008934	0
28	Hypothetical protein	00057	441	hsdR (Type I restriction-modification system endonuclease)	S. aur.	FR821779	1.32e-169
29	Hypothetical protein	00058	183	Type I site-specific deoxyribonuclease restriction subunit	S. aur.	CP002114	3.52e-83
30	cycA	00059	1,377	aapA (D-serine/D-alanine/glycine transporter)	S. hae.	AP006716	0
31	katE	00060	2,010	Catalase C	S. xyl.	EU475909	0
32	lip	00061	1,101	Lipase	S. sap.	AY551101	0
33	lip	00062	870	Lipase	S. sap.	AY551101	0
34	Zinc-responsive transcriptional regulator	00063	348	Transcriptional regulator	S. cap.	KF049201	2.75e-164
35	hchA	00064	684	Hypothetical protein	S. cap.	KF049201	0
36	yhfK	00065	669	NAD dependent epimerase/dehydratase	S. hae.	KF006347	0
37	yfmJ	00066	1,005	Putative oxidoreductase	S. cap.	KF049201	0
38	Putative membrane protein	00067	366	Hypothetical protein	S. hae.	KF006347	1.15e-175
39	isaB	00068	390	Hypothetical protein	S. epi.	CP000029	2.90e-158
40	Hopanoid-associated sugar epimerase	00069	1,500	Conserved domain protein	S. xyl.	HE993886	0
41	TspO/MBR family protein	00070	501	Hypothetical protein	S. xyl.	HE993885	0
42	Hypothetical protein	00071	474	Glycosyl-4,4'-diaponeurosporenoate acyltransferase precursor	S. xyl.	HE993884	0
43	crtP	00072	1,491	Phytoene desaturase, neurosporene or lycopene producing / 4,4'-diapolycopene oxidase	S. xyl.	HE993884	0
44	crtQ	00073	1,104	4,4'-diaponeurosporenoate glycosyltransferase	S. xyl.	HE993886	0
45	crtM	00074	882	Dehydrosqualene synthase	S. xyl.	HE993886	0
46	crtN	00075	1,506	Dehydrosqualene desaturase	S. xyl.	HE993884	0
47	ydaG	00076	426	General stress protein 26	S. xyl.	HE993886	0
48	phrB	00077	1,374	Deoxyribocipyrimidine photolyase	S. xyl.	HE993886	0
49	hmoB	00078	432	Hypothetical protein	S. xyl.	HE993885	0

* Alpha proteo. = Alpha proteobacterium, D. pur. = *Dictyostelium purpureum,* L. car. = *Leuconostoc carnosum*, L. mon. = *Listeria monocytogenes,* S. 693-2 = *S. sp. 693-2*, S. aur. = *S. aureus*, S. cap. = *S. capitis*, S. epi. = *S. epidermidis*, S. hae. = *S. haemolyticus,* S. sap. = *S. saprophyticus,* S. sci. = *S. sciuri*, S. xyl. = *S. xylosus,* V. vin. = *Vitis vinifera*

Table S10. Novel *S. saprophyticus* Troy region #03: Gene descriptions and top BLAST hits

Troy CDS				Top BLAST hit			
#	Description	Locus#	Length	Description	Species*	Accession	Evalue
1	Gram-positive signal peptide, YSIRK family	00117	804	*NO HIT*	-	-	-
2	Putative membrane protein	00118	1,077	Hypothetical protein	Paeni. sp. JDR-2	CP001656	9.86e-06
3	Hypothetical protein	00119	636	Hypothetical protein	S. xyl.	Y14599	4.76e-92
4	cbiO	00120	432	Hypothetical protein, similar to ABC transporter ATP-binding protein	S. hae.	AP006716	3.67e-49
5	Putative ATPase, ABC class	00121	1,710	Hypothetical protein, similar to ABC transporter ATPase domain	S. hae.	AP006716	0
6	Hypothetical protein	00122	303	-	A. tha.	CP002686	1.78e-02
7	Hypothetical protein	00123	216	Hypothetical protein	S. war.	CP003668	1.05e-49
8	Calcineurin-like phosphoesterase	00124	1,986	N-acetylmuramoyl-L-alanine amidase	F. mag.	AP008971	3.67e-11
9	Hypothetical protein	00125	1,230	Hypothetical protein	S. pas.	CP004014	6.17e-173

* A. tha. = *Arabidopsis thaliana,* F. mag. = *Finegoldia magna*, Paeni = *Paenibacillus,* S. hae. = *S. haemolyticus,* S. pas. = *S. pasteuri,* S. war. = *S. warneri,* S. xyl. = *S. xylosus*

Table S11. Novel *S. saprophyticus* Troy region #13: Gene descriptions and top BLAST hits

Troy CDS				Top BLAST hit			
#	Description	Locus#	Length	Description	Species*	Accession	Evalue
1	lytR	02172	942	capM (hypothetical protein, similar to lyt expression attenuator LytR)	S. hae.	AP006716	4.12e-85
2	epsN	02173	732	capL (hypothetical protein, similar to spore coat polysaccharide biosynthesis protein)	S. hae.	AP006716	7.96e-112
3	epsN	02174	396	capL (hypothetical protein, similar to spore coat polysaccharide biosynthesis protein)	S. hae.	AP006716	1.26e-42
4	Carbamoyl phosphate synthase-like protein	02175	984	Hypothetical protein, similar to carbamoylphosphate synthase large subunit	S. hae.	AP006716	4.93e-59
5	speG	02176	513	Hypothetical protein, similar to spermidine N1-acetyltransferase	S. hae.	AP006716	5.47e-36
6	epsL	02177	591	Hypothetical protein, similar to UDP-galactose phosphate transferase	S. hae.	AP006716	1.08e-64
7	tuaC	02178	552	Putative capsular polysaccharide biosynthesis protein	S. car.	AM295250	2.87e-46
8	epsD	02179	432	Putative capsular polysaccharide biosynthesis protein	S. car.	AM295250	4.54e-36
9	wbpA	02180	1,275	cap1L (type 1 capsule synthesis gene)	S. aur.	U10927	0
10	Putative adenylate-forming enzyme	02181	1,341	cap1K (type 1 capsule synthesis gene)	S. aur.	U10927	0
11	rffG	02182	1,002	cap1i (type 1 capsule synthesis gene)	S. aur.	U10927	0
12	Hypothetical protein	02183	735	-	P. tri.	AC182665	1.05e-02
13	tuaG	02184	780	Putative glycosyltransferase	S. sap.	AB353724	4.41e-33
14	Hypothetical protein	02185	1,274	(Repeat region)	Human	AC091062	1.11e-05
15	Colonic acid biosynthesis protein	02186	1,026	16S rRNA	P. pun.	FJ171325	3.48e-04
16	gtaB	02187	822	UTP-glucose-1-phosphate uridylyltransferase	B. wei.	CP000903	5.67e-70
17	pglF	02188	1,812	cap1D (type 1 capsule synthesis gene)	S. aur.	U10927	0
18	ywqE	02189	768	cap1C (type 1 capsule synthesis gene)	S. aur	U10927	0
19	ywqD	02190	687	cap1B (type 1 capsule synthesis gene)	S. aur.	U10927	0
20	cap8A	02191	666	cap1A (type 1 capsule synthesis gene)	S. aur.	U10927	5.90e-170

* B. wei. = *Bacillus weihenstephanensis,* S. aur. = *S. aureus,* S. car. = *S. carnosus,* S. hae. = *S. haemolyticus,* S. sap. = *S. saprophyticus,* P. pun. = *Polystoechotes punctatus*, P. tri. = *Populus trichocarpa*

Table S12. Novel *S. saprophyticus* Troy region #15: Gene descriptions and top BLAST hits

Troy CDS				Top BLAST hit			
#	Description	Locus#	Length	Description	Species*	Accession	Evalue
1	yxeP	02377	1,203	Peptidase M20D amidohydrolase B	S. xyl.	EU475913	4.48e-146
2	ydhP	02378	1,182	Major facilitator family protein	S. car.	AM295250	4.86e-174
3	ytcD	02379	381	Putative transcriptional regulator	S. car.	AM295250	1.48e-41
4	Putative membrane protein	02380	360	Hypothetical protein	S. war.	CP003668	2.52e-19
5	csbC	02381	1,356	Putative permease of the major facilitator superfamily	S. sap.	AP008934	0

* S. car. = *S. carnosus,* S. sap. = *S. saprophyticus,* S. war. = *S. warneri,* S. xyl. = *S. xylosus*

Table S13. Novel *S. saprophyticus* Troy region #16: Gene descriptions and top BLAST hits

Troy CDS				Top BLAST hit			
#	Description	Locus#	Length	Description	Species*	Accession	Evalue
1	tagE	02418	4,821	cps2D	A. ple.	AY357726	2.66e-07
2	Hypothetical protein	02419	279	(Repeat region)	Human	AC015528	1.27e-02
3	tagB	02420	1,689	Putative glycosyl glycerophosphate transferase involved in teichoic acid biosynthesis	S. sap.	AP008934	0
4	gutB	02421	1,029	Putative zinc-binding dehydrogenase	S. sap.	AP008934	0
5	ispD	02422	717	Putative 4-diphosphocytidyl-2-methyl-D-erithritol synthase	S. sap.	AP008934	0
6	Hypothetical protein	02423	117	-	S. jap.	FN328684	1.51e-02

* A. ple. = Actinobacillus pleuropneumoniae, S. sap. = *S. saprophyticus,* S. jap. = *Schistosoma japonicum*

Table S14. *S. saprophyticus* Troy plasmid pSST1: Gene descriptions and top BLAST hits

Troy CDS				Top BLAST hit			
#	Description	Locus#	Length	Description	Species*	Accession	Evalue
1	Hypothetical protein	P0001	219	Dehydrogenase (pseudogene)	S. sap., pSSAP2	HE616681	5.66e-107
2	gutB, alcohol dehydrogenase, zinc-containing	P0002	426	Dehydrogenase (pseudogene)	S. sap., pSSAP2	HE616681	0
3	Hypothetical protein	P0003	375	Hypothetical protein	S. sap., pSSAP2	HE616681	0
4	Putative acyl-CoA hydrolase	P0004	525	Putative acyl-CoA hydrolase	S. sap., pSSAP2	HE616681	0
5	Transcriptional regulator	P0005	873	Putative transcriptional regulator	S. sap., pSSAP2	HE616681	0
6	Putative site-specific recombinase	P0006	552	Putative site-specific recombinase	S. sap., pSSAP2	HE616681	0
7	Alcohol dehydrogenase	P0007	1,014	Putative alcohol dehydrogenase	S. sap., pSSAP2	HE616681	0
8	Transcriptional regulator	P0008	480	Transcriptional regulator	S. sap., pSSAP2	HE616681	0
9	Dehydrogenase	P0009	1,344	Putative dehydrogenase	S. sap., pSSAP2	HE616681	0
10	Hypothetical protein	P0010	276	Hypothetical protein	S. sap., pSSAP2	HE616681	3.59e-136
11	Small heat shock protein	P0011	429	Putative small heat shock protein	S. sap., pSSAP2	HE616681	0
12	Hypothetical protein	P0012	123	Hypothetical protein	S. sap., pSSAP2	HE616681	1.25e-54
13	Replication initiator protein A	P0013	939	repA	S. sap., pSSAP2	HE616681	0
14	Replication-associated protein	P0014	780	Replication-associated protein	S. sap., pSSAP2	HE616681	0
15	Hypothetical protein	P0015	258	Zinc-binding oxidoreductase	S. sap., pSSAP2	HE616681	4.62e-128
16	Putative alcohol-dehydrogenase, zinc-containing	P0016	219	Zinc-binding oxidoreductase	S. sap., pSSAP2	HE616681	2.41e-105
17	3-hexulose-6-phosphate synthase	P0017	294	putative truncated 3-hexulose-6-phosphate synthase	S. sap., pSSP1	AP008935	6.53e-146
18	Putative transcriptional regulator	P0018	456	Transcriptional regulator	S. sap., pSSAP2	HE616681	0
19	Phage infection protein	P0019	2,850	Hypothetical protein	S. sap., pSSAP2	HE616681	0
20	N-acetylmuramoyl-L-alanine amidase	P0020	738	N-acetylmuramoyl-L-alanine amidase	S. sap., pSSAP2	HE616681	0
21	N-acetylmuramoyl-L-alanine amidase	P0021	1,407	Putative mannosyl-glycoprotein endo-beta-N-acetylglucosamidase	S. sap., pSSP1	AP008935	0
22	Cell wall surface anchor family protein	P0022	1,965	sssF	S. sap., pSSAP2	HE616681	0

Table S15. *G. vaginalis* modern strains

Strain	Ref	Scale*	Accession#	Organ, health	Location	Notes**	Genes
409-05	(62)	F	NC_013721	Vagina, asymptomatic BV	?	HMP	1267
ATCC_14019	(62)	F	NC_014644	Vagina, BV	?	HMP	1295
HMP9231	-	F	NC_017456	Endometrium	?	HMP	1319
5-1	(34)	S	ADAN01000000	Vagina, BV	Boston, MA	BWH	1306
41V	-	S	AEJE00000000	Unknown, healthy	?	?	1282
101	(34)	S	AEJD00000000	?	?	?	1159
284V	(21)	S	ADEL00000000	Endometrium, discharge	Pittsburgh, PA	MWH	1253
0288E	(21)	S	ADEN00000000	Endometrium, discharge	Pittsburgh, PA	MWH	1317
315-A	-	S	AFDI00000000	Vagina	Pittsburgh, PA	HMP	1274
00703Bmash	(21)	S	ADET00000000	Vagina, BV	Pittsburgh, PA	MWH	1206
00703C2mash	(21)	S	ADEU00000000	Vagina, BV	Pittsburgh, PA	MWH	1187
00703Dmash	(21)	S	ADEV00000000	Vagina, BV	Pittsburgh, PA	MWH	1123
1400E	(21)	S	ADER00000000	Endometrium	Pittsburgh, PA	MWH	1337
1500E	(21)	S	ADES00000000	Endometrium	Pittsburgh, PA	MWH	1176
61/19V5	(21)	S	ADEW00000000	Vagina, healthy	Pittsburgh, PA	MWH	1134
64/20B	(21)	S	ADEP00000000	Endometrium, healthy	Pittsburgh, PA	MWH	1112
64/20LIT	(21)	S	ADEO00000000	Endometrium, healthy	Pittsburgh, PA	MWH	-
55/15-2	(21)	S	ADEQ00000000	Endometrium, asymptomatic BV	Pittsburgh, PA	MWH	1278
7571-2	(21)	S	ADEM00000000	Vagina, BV	Pittsburgh, PA	MWH	1278
AMD	(34)	S	ADAM00000000	Vagina, healthy	Richmond, VA	VCU	1206
ATCC_14018	(62)	S	ADNB00000000	Vagina, BV	?	?	1245
JCP7275		S	ATJS00000000	Vagina	Unknown	HMP	1203
JCP7276		S	ATJR01000000	Vagina	Unknown	HMP	1268
JCP7659		S	ATJQ00000000	Vagina	Unknown	HMP	1195
JCP7672		S	ATJP00000000	Vagina	Unknown	HMP	1217
JCP7719		S	ATJO00000000	Vagina	Unknown	HMP	1233
JCP8017A		S	ATJN00000000	Vagina	Unknown	HMP	1289
JCP8017B		S	ATJM00000000	Vagina	Unknown	HMP	1286
JCP8066		S	ATJL00000000	Vagina	Unknown	HMP	1155
JCP8070		S	ATJK00000000	Vagina	Unknown	HMP	1138
JCP8108		S	ATJJ00000000	Vagina	Unknown	HMP	-
JCP8151A		S	ATJI00000000	Vagina	Unknown	HMP	1210
JCP8151B		S	ATJH00000000	Vagina	Unknown	HMP	1218
JCP8481A		S	ATJG00000000	Vagina	Unknown	HMP	1187
JCP8481B		S	ATJF00000000	Vagina	Unknown	HMP	1189
JCP8522		S	ATJE00000000	Vagina	Unknown	HMP	1109

** F = full, S = scaffold; ** BWH = Brigham & Women's Hospital (34), HMP = Human Microbiome Project, MWH = Magee-Womens Hospital (21), VCU = VCU Women's Health Clinic (34)*

Table S16. Called SNPs in human mitochondrial genome assemblies

Nod1_1h-UDG: SNPs called at 85% variant frequency

Nod2-UDG, Ulna_Enr1-nonU, & Ulna_Enr2-nonU: SNPs called at 60% variant frequency

Pos.	Ref	Troy	Nod1_1h-UDG	Nod2-UDG*	Ulna_Enr1-nonU	Ulna_Enr2-nonU
73	A	G	27, 100%	-	14, 92.9%	24, 95.8%
150	C	T	31, 100%	3, 100%	18, 88.9%	32, 65.6%
152	T	C	30, 100%	3, 100%	17, 82.4%	31, 61.3%
263	A	G	33, 100%	-	22, 86.4%	26, 88.5%
310	T	C	10, 100%	-	3, 100%	8, 75.0%
750	A	G	28, 100%	4, 100%	27, 100%	43, 100.0%
1438	A	G	29, 89.7%	-	20, 90%	28, 89.3%
1811	A	G	31, 96.8%	3, 100%	21, 85.7%	28, 85.7%
2706	A	G	23, 95.7%	-	24, 91.7%	39, 92.3%
3397	A	G	29, 100%	3, 100%	21, 76.2%	36, 75.0%
4188	A	G	29, 100%	-	23, 82.6%	35, 77.1%
4640	C	A	30, 100%	-	28, 92.9%	46, 91.3%
4769	A	G	26, 100%	3, 100%	14, 100%	16, 100.0%
7028	C	T	26, 100%	-	20, 90%	24, 100.0%
8860	A	G	33, 100%	3, 100%	23, 95.7%	29, 96.6%
9426	C	T	24, 95.8%	7, 100%	17, 76.5%	32, 68.8%
9656	T	C	32, 100%	-	29, 82.8%	42, 78.6%
10775	G	A	34, 97.1%	-	24, 83.3%	37, 78.4%
11467	A	G	24, 95.8%	4, 100%	15, 80%	20, 65.0%
11719	G	A	27, 100%	-	26, 96.2%	34, 91.2%
12308	A	G	29, 93.1%	-	29, 93.1%	45, 91.1%
12372	G	A	22, 100%	-	15, 86.7%	24, 75.0%
13743	T	C	27, 100%	-	27, 66.7%	33, 63.6%
14139	A	G	20, 100%	4, 100%	27, 85.2%	48, 70.8%
14766	C	T	32, 100%	-	17, 94.1%	29, 96.6%
15326	A	G	37, 100%	3, 100%	31, 90.3%	39, 89.7%
15454	T	C	29, 100%	3, 100%	17, 88.2%	30, 66.7%
15954	A	G	23, 100%	4, 100%	31, 80.6%	46, 78.3%
16168	C	T	31, 100%	-	16, 81.3%	25, 76.0%
16343	A	G	36, 100%	6, 100%	25, 84%	32, 84.4%

* NOD2-UDG has 1 additional SNP (G -> A) called at 11,693 (coverage 3, frequency 66.7%)

3.12 Supplementary Figures

Figure S1. Additional photos of nodules 1 & 2

a. Nodule 1 (left) and nodule 2 (right). Photo credit C. Pepperell, 2013.

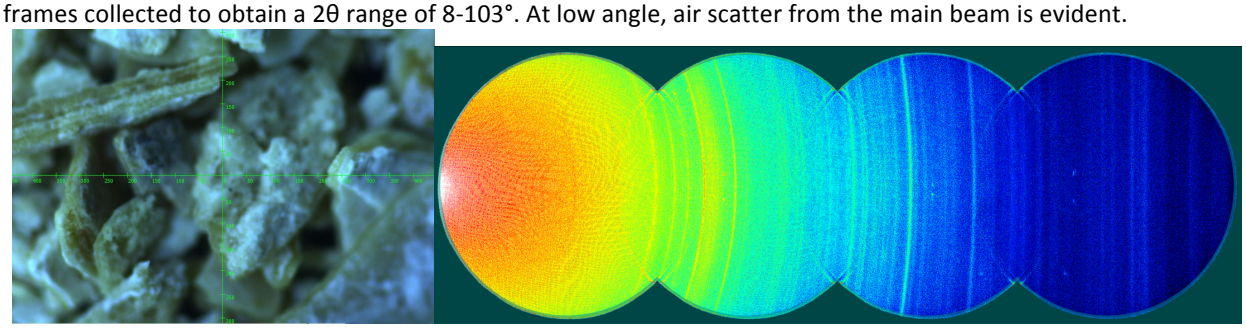


b. Nodule 2 (left) and nodule 1 (right). Photo credit H. Kiesewetter, 2013.

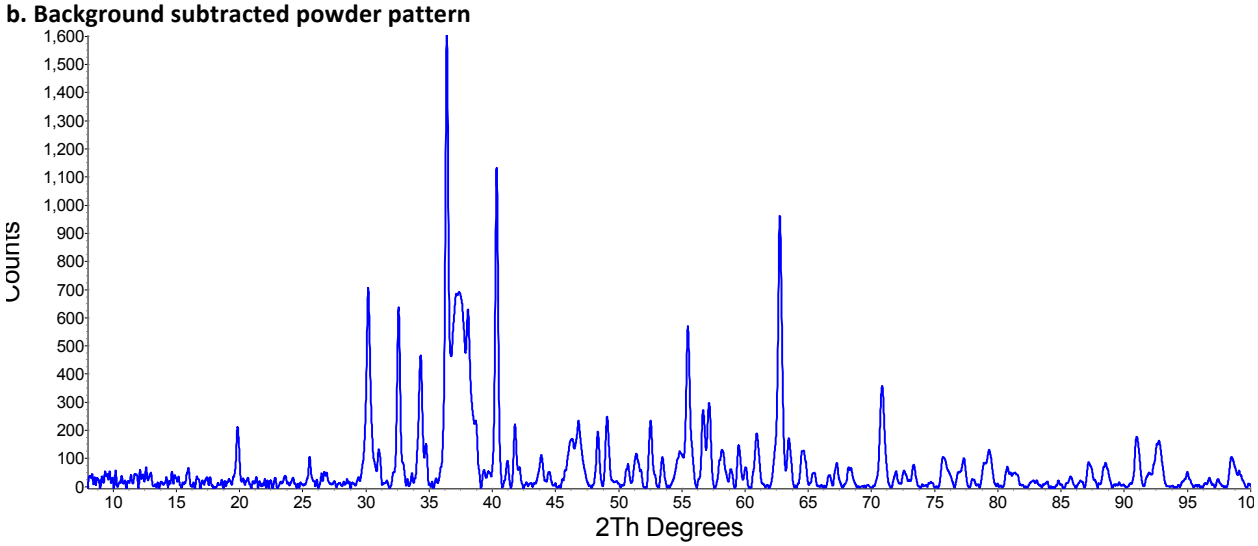


Figure S2. XRD analysis of nodule 1

a. Video alignment & XRD frames. Left; crosshairs indicate the centre of the region examined. Right; the four frames collected to obtain a 2θ range of 8-103°. At low angle, air scatter from the main beam is evident.



b. Background subtracted powder pattern



c. Search/match results

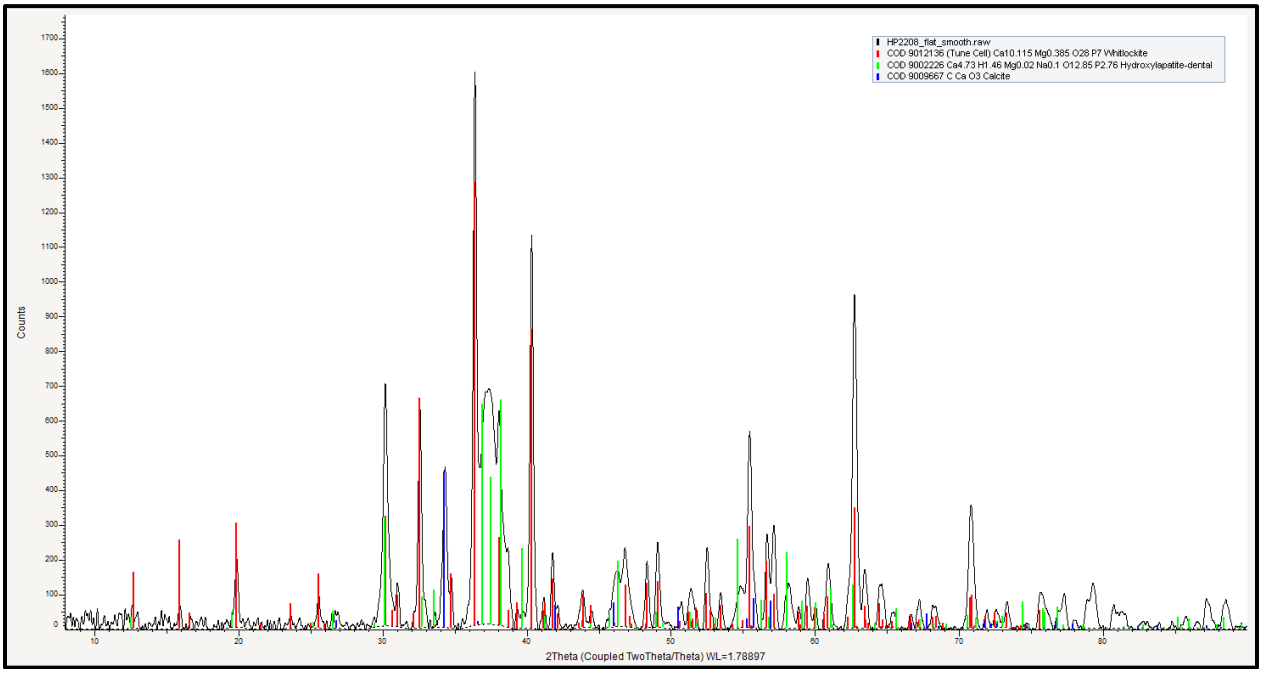


Figure S3. SEM image of nodule 1 at 2,000x and 10,000x magnification

Bacterial cells indicated with arrow diameters are ~ 1uM (within range expected for Staphylococcus).

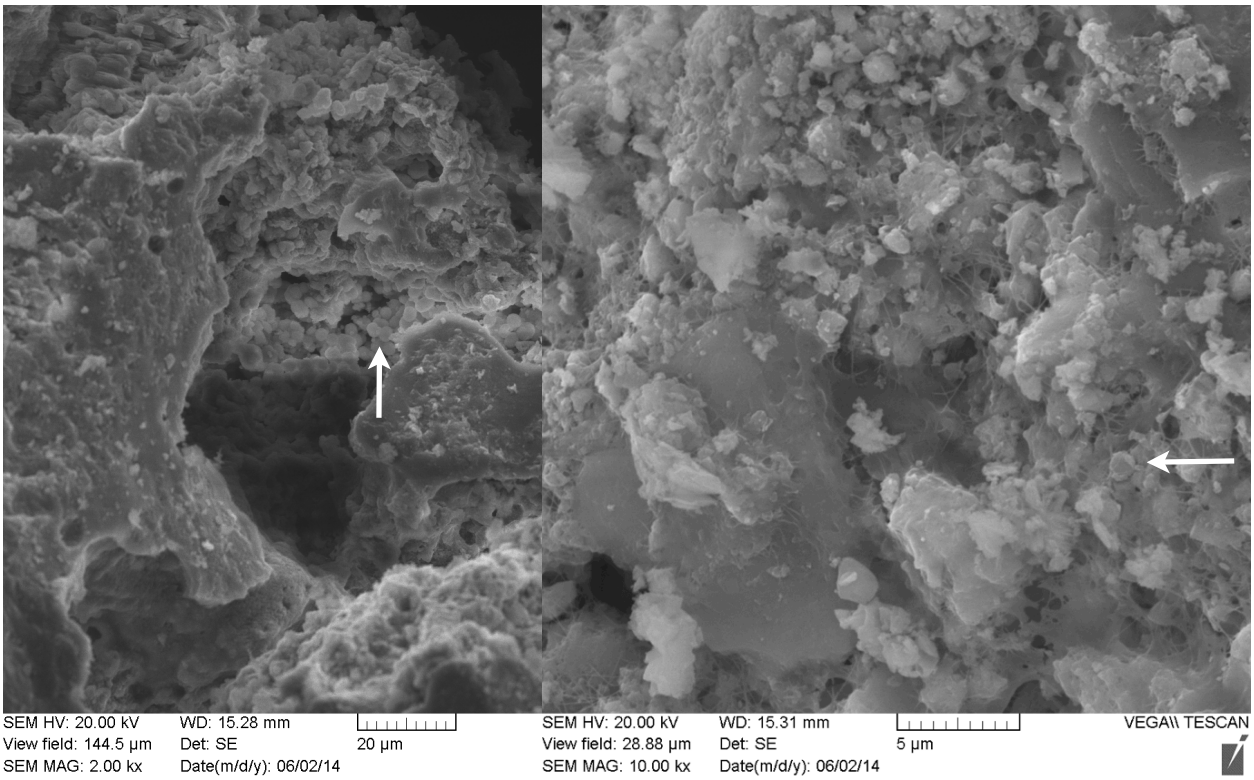


Figure S4. Shotgun metagenomic phylum-level comparison between the 2 nodules, bone, sediment. MEGAN5 parsing of blastn-megablast results for 2 million (trimmed, >24bp) UDG reads: NOD1_1h (green), NOD2 (purple), Ulna (orange), Sediment (beige). Colored bars are log-scale; readcounts shown to right.

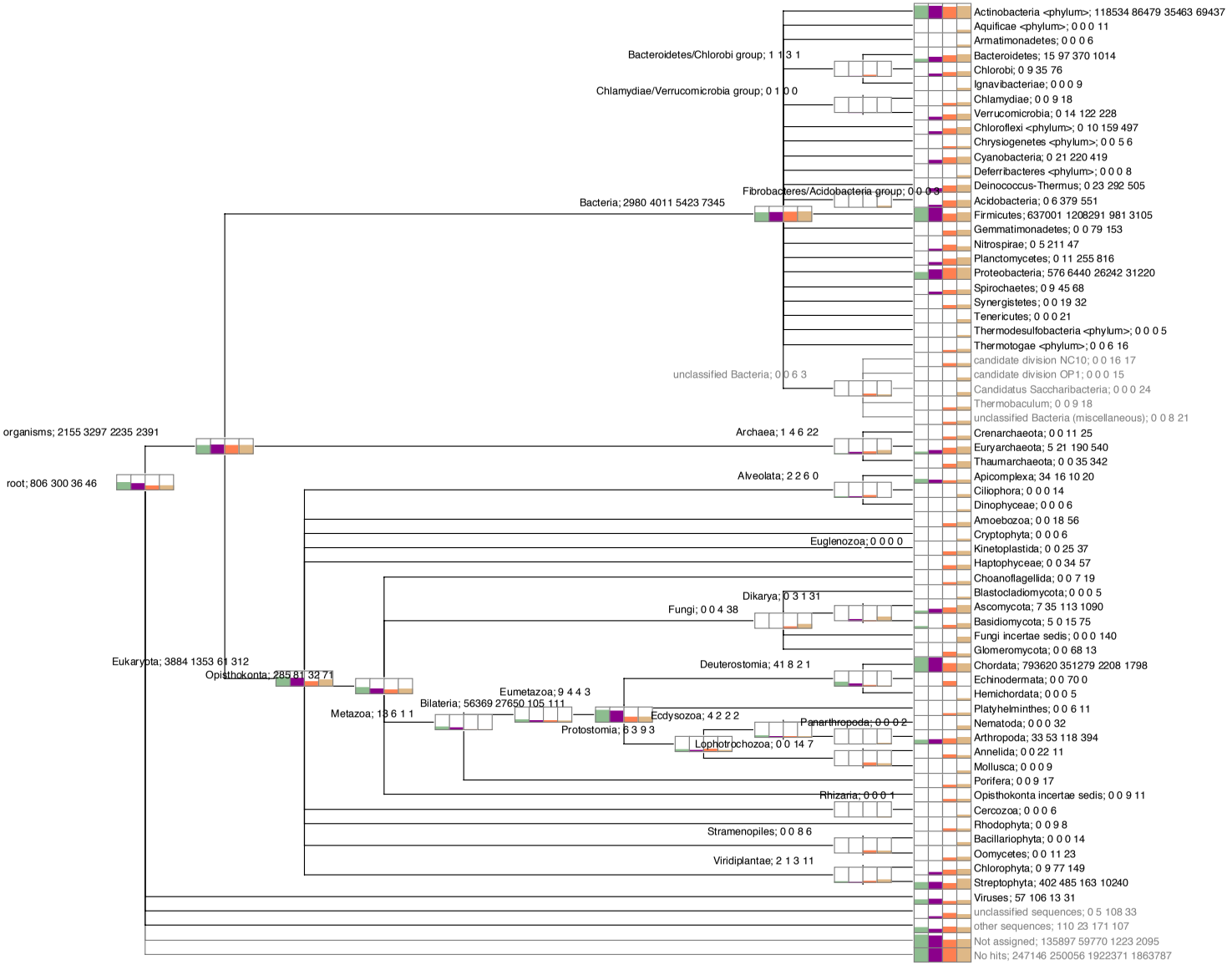


Figure S5. Principal Components Analysis (PCA) of shotgun metagenomic datasets from the nodules, bone, and sediment

PCA plot of the first two principal components generated from MEGAN5 comparison of the datasets in the previous figure (S3), showing the strong similarity between the Sediment and Ulna datasets, likely due to contamination of the bone with soil-associated organisms. The dissimilarity between the two nodules may be due to inherent differences in the endogenous metagenomic content of the two specimens, or due to slight differences in the DNA extraction methods between the two samples, which may impact the resulting metagenomic profiles.

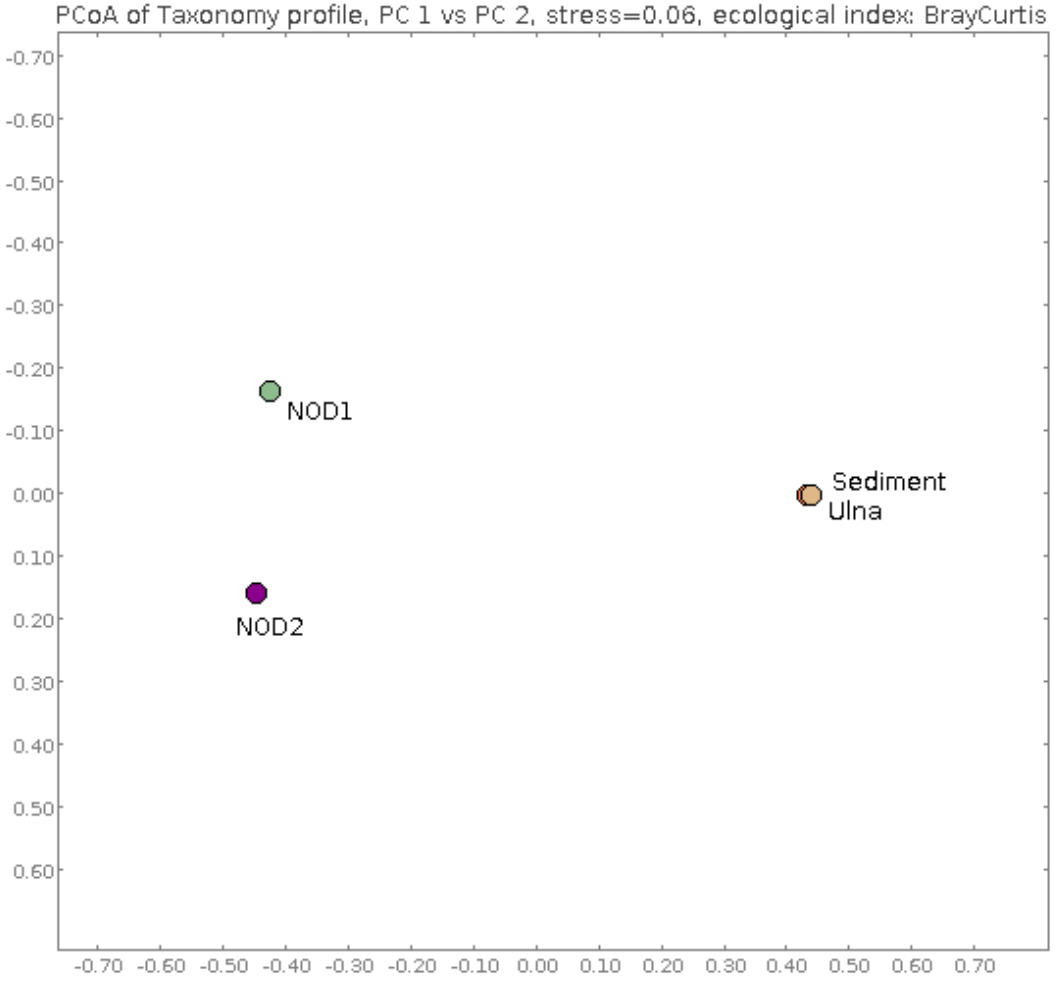


Figure S6. Shotgun metagenomic comparison between extraction blanks

Phylum-level comparison of EblkA-UDG (light blue), EblkB-UDG (dark blue), Eblk3-UDG (orange), and Eblk4-UDG (beige), based on MEGAN5 parsing of blastn-megablast results for all (trimmed, >24bp) reads from each sample.

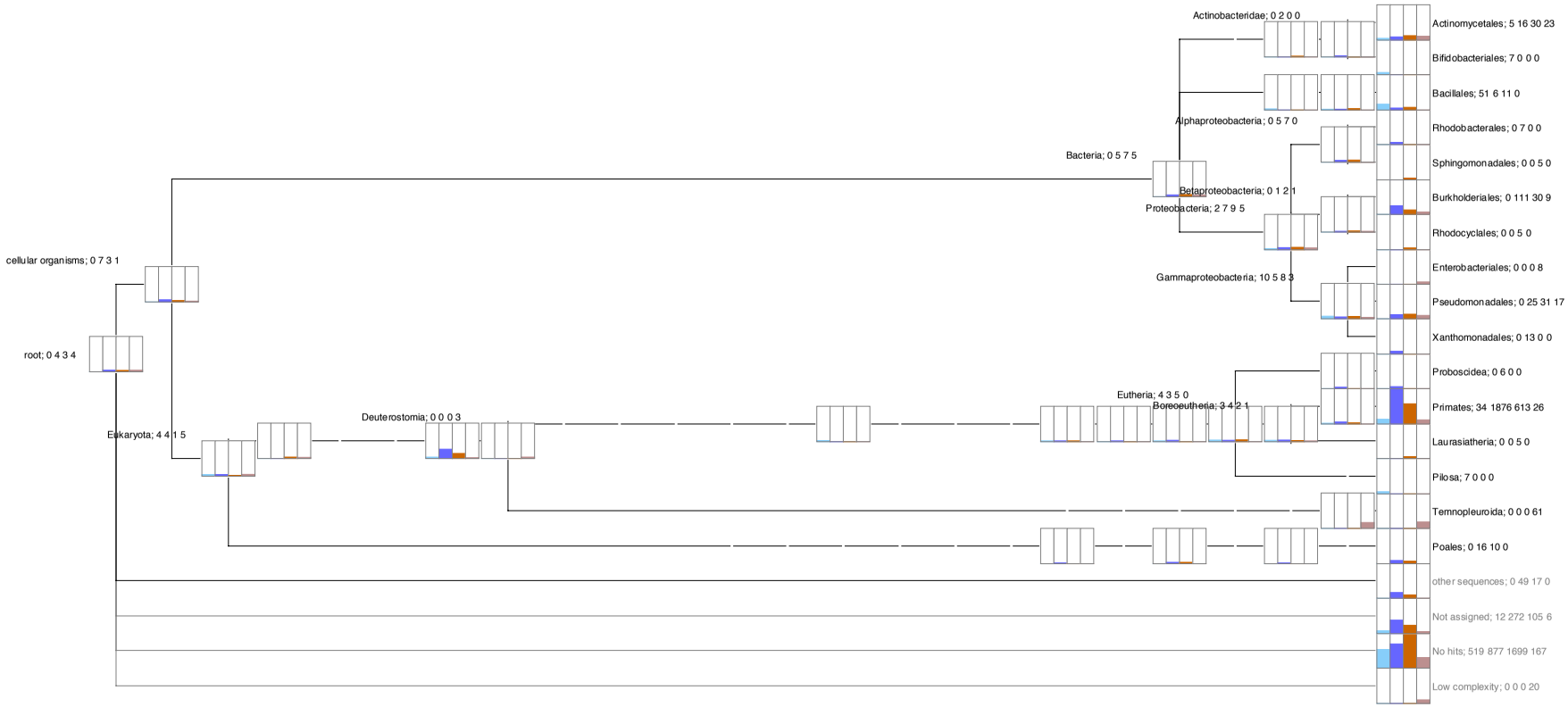
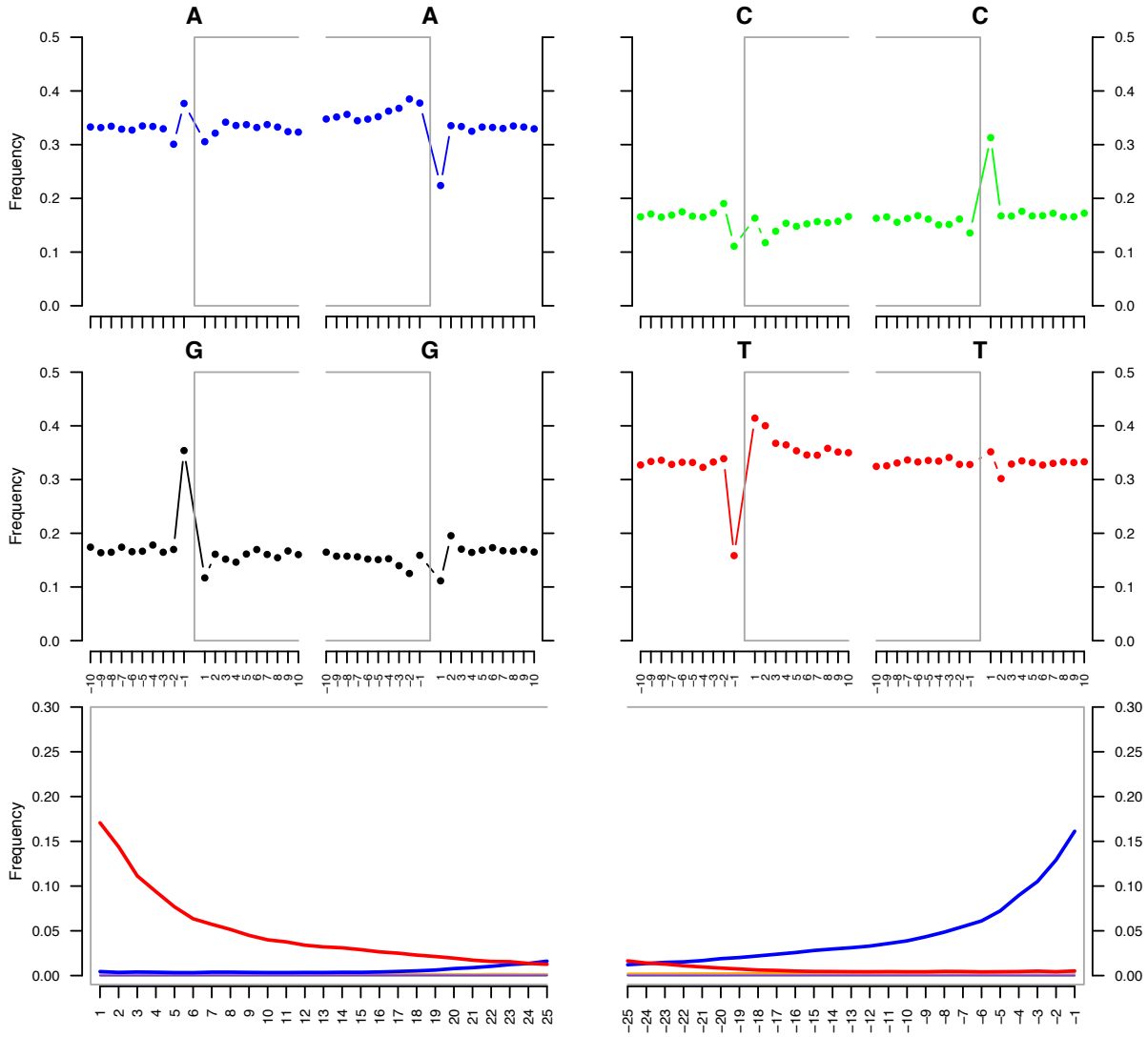


Figure S7. Ancient DNA damage assessment of *S. saprophyticus* and *G. vaginalis*

Damage profiles of non-UDG treated ('nonU') reads from a pooled NOD1_nonU and NOD2_nonU dataset (total of 1,565,548 trimmed reads >24bp) mapping to (a) *S. saprophyticus* strain ATCC 15305 and (b) *G. vaginalis* strain ATCC 14019. Paired end reads were mapped using bwa with default settings and duplicates were removed with samtools rmdup. Damage profiles were generated using mapDamage2 (82).

a. mapDamage profile for NOD1+2_nonU reads mapping to *S. saprophyticus* ATCC 15305



b. mapDamage profile for NOD1+2_nonU reads mapping to *G. vaginalis* ATCC 14019

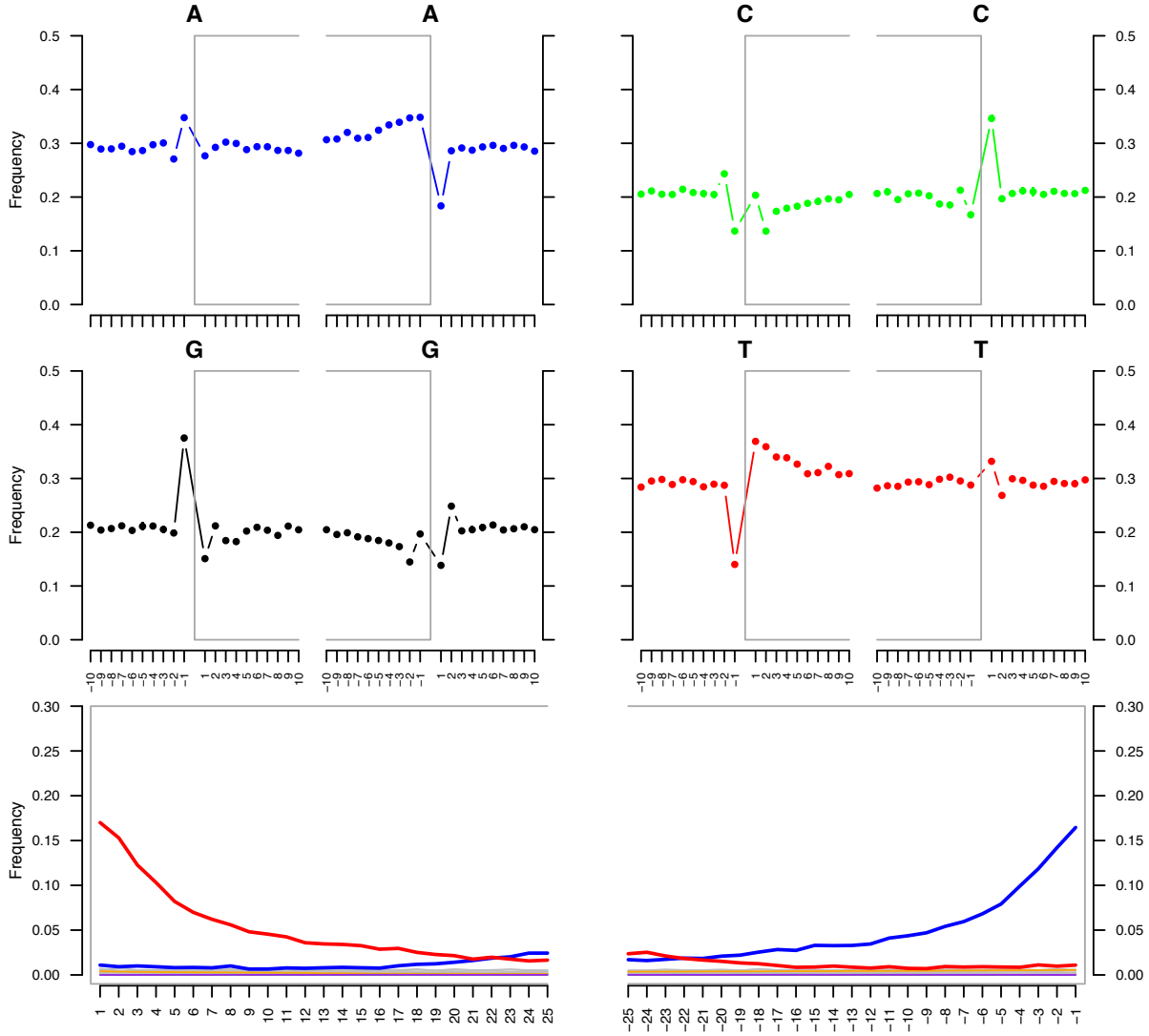
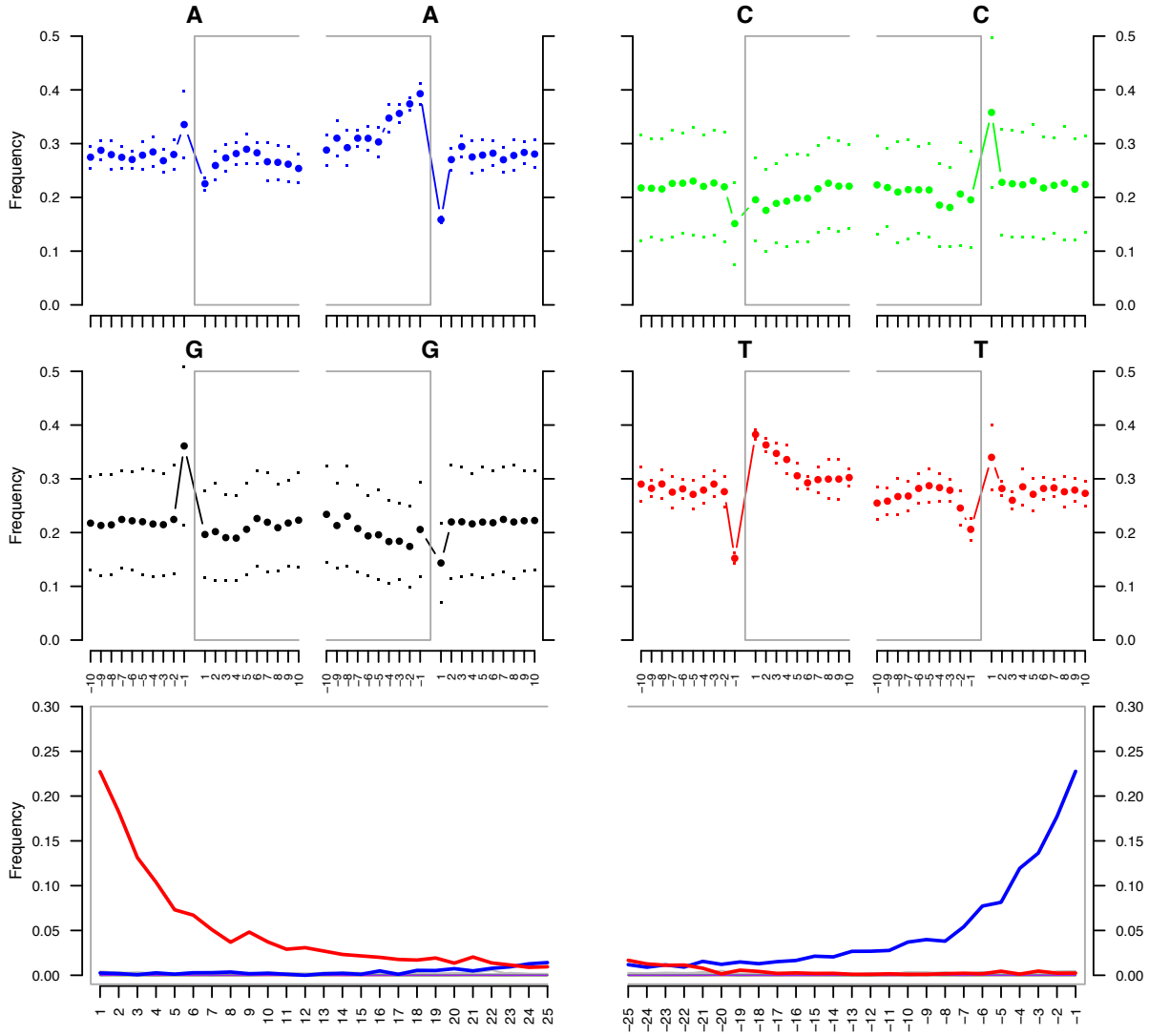


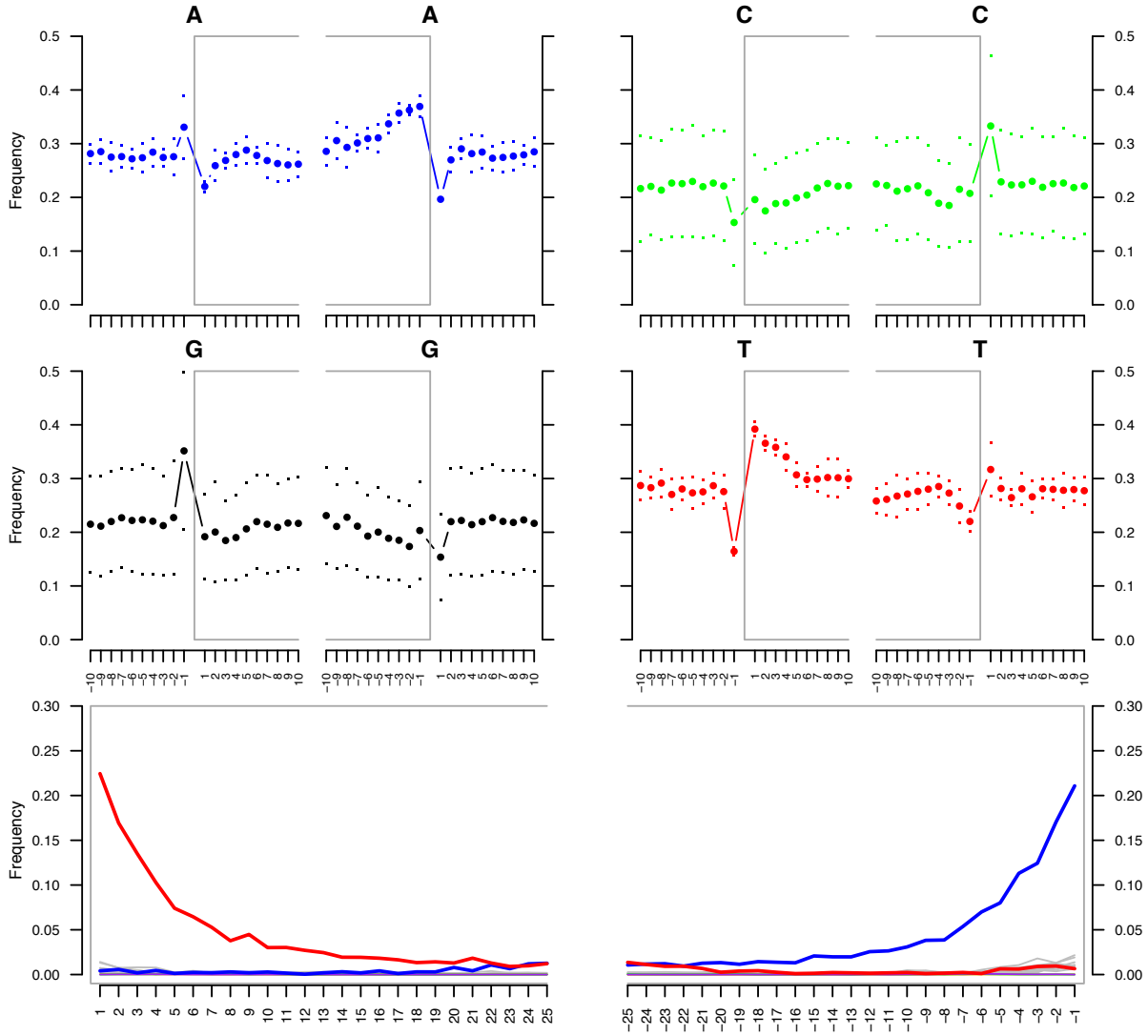
Figure S8. Ancient DNA damage assessment of human mitochondrial reads

Damage profiles of non-UDG treated ('nonU') merged reads assembled to the human mitochondrial CRS reference genome (NC_012920) for (a) Ulna_Enr1-nonU round 1 human mitochondrial enrichment, (b) Ulna_Enr2-nonU round 2, and (3) a pool of NOD1_1h-nonU and NOD2-nonU reads. Damage profiles were generated using mapDamage2 (82).

a. mapDamage profile for Ulna_Enr1-nonU reads mapping to mitochondrial genome



b. mapDamage profile for Ulna_Enr2-nonU reads mapping to mitochondrial genome



c. mapDamage profile for NOD1+2-nonU reads mapping to mitochondrial genome

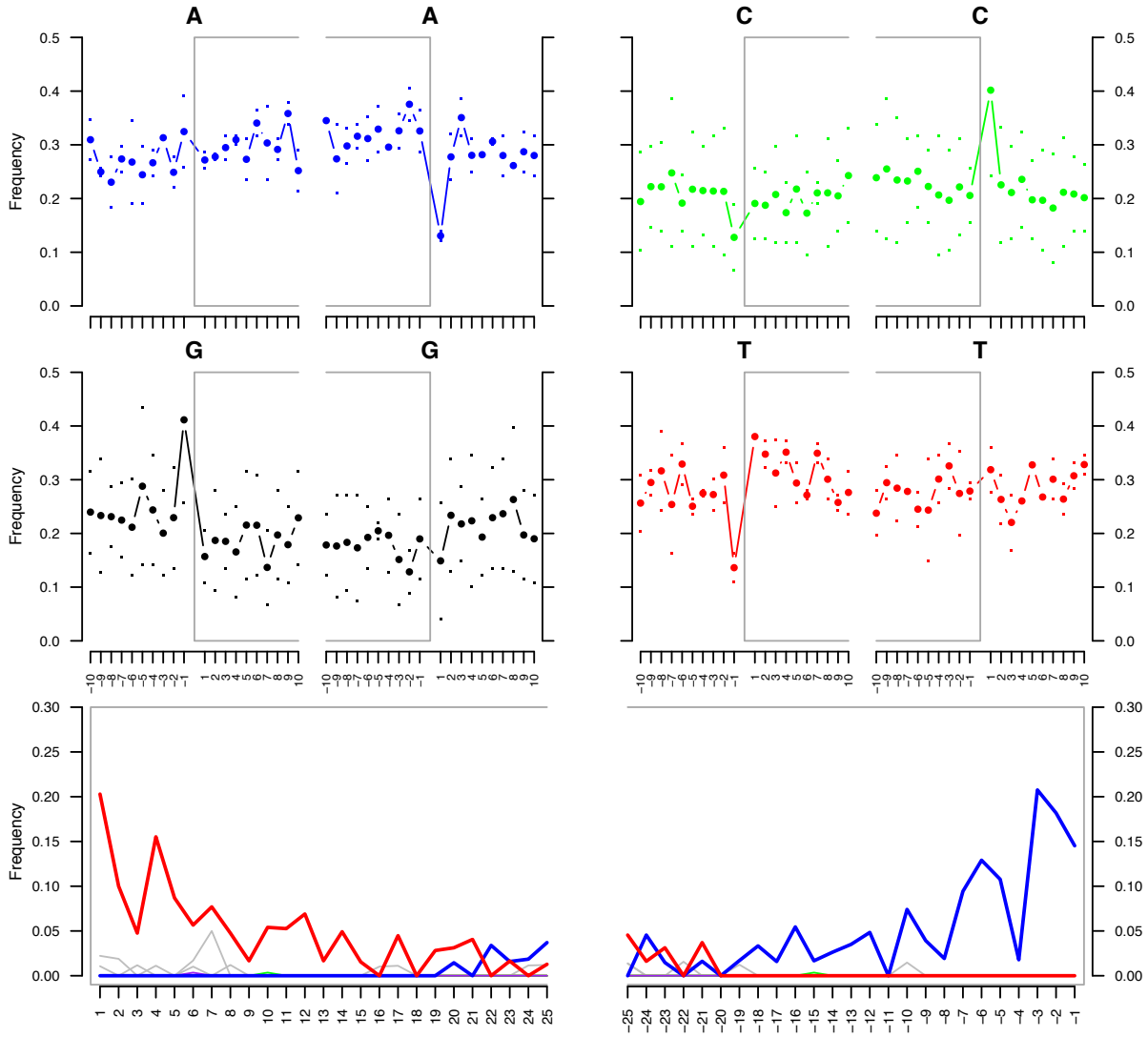


Figure S9. Fragment length distribution (FLD) for *S. saprophyticus* ATCC 15305

Nod1_1h-UDG paired-end reads (10,038,389 readpairs) mapped to the modern human *S. saprophyticus* reference strain (ATCC 15305) using bwa (76) with default settings and duplicates collapsed using samtools rmdup (73). Figure generated using Geneious (v.6.1.6) from 8,991,704 pairs (excluding 63,900 incorrectly orientated pairs and 139,899 distant pairs).

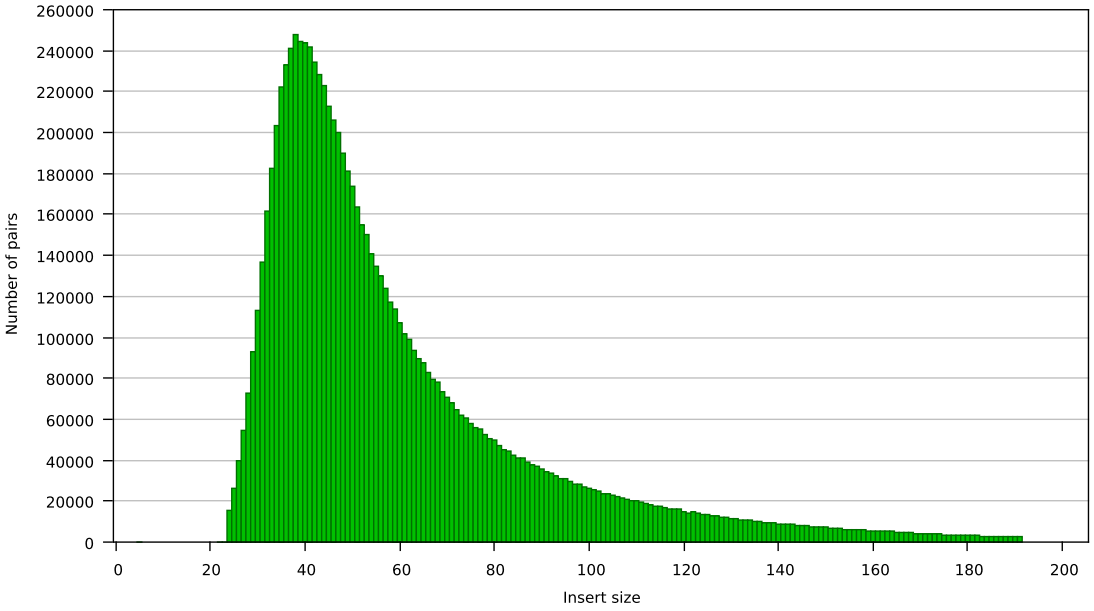


Figure S10. Fragment length distribution (FLD) for *G. vaginalis* ATCC 14019

Nod1_1h-UDG paired-end reads (1,964,443 readpairs) mapped to the modern human *G. vaginalis* reference strain (ATCC 14019) using bwa (76) with default settings and duplicates collapsed using samtools rmdup (73). Figure generated with Geneious (v.6.1.6) from 1,803,969 pairs (excluding 3,518 incorrectly orientated pairs and 34,464 distant pairs).

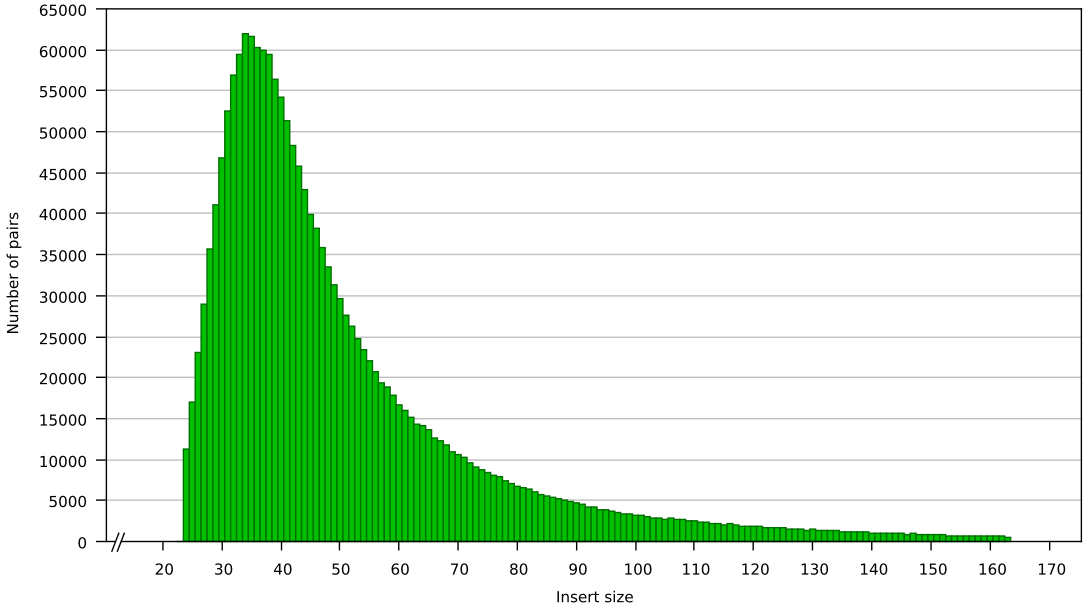
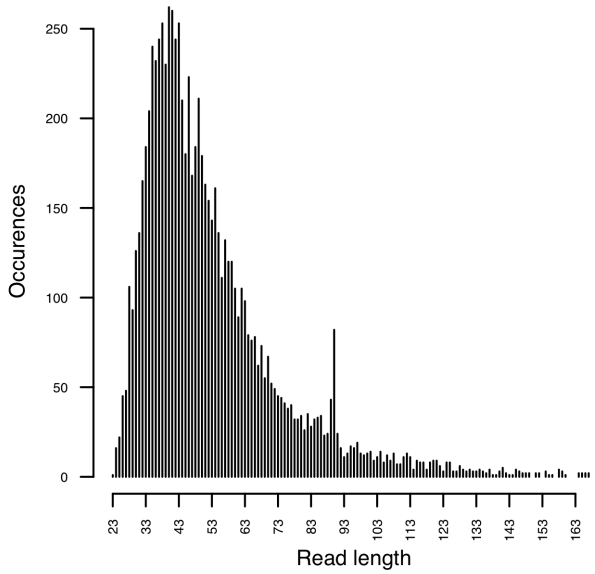


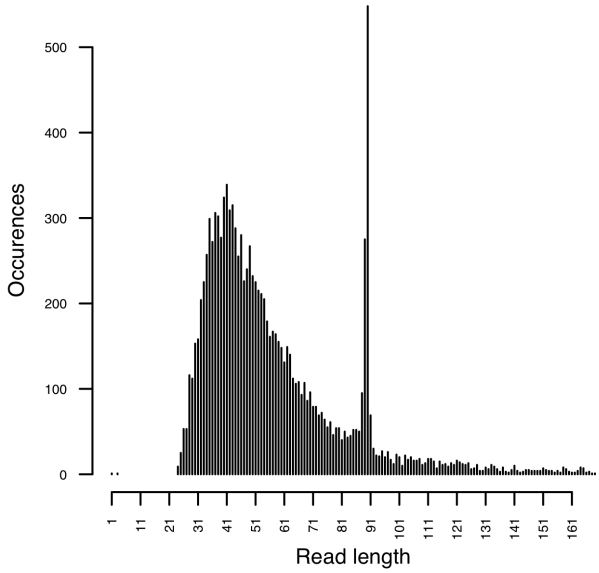
Figure S11. Fragment length distributions for nonU human mitochondrial assemblies

These FLDs were generated using mapDamage2 (82) from merged nonUDG datasets assembled to the human mitochondrial reference genome, as described in the Enrichment section: (a) Ulna_Enr1-nonU, (b) Ulna_Enr2-nonU, and (c) NOD1+2-nonU.

a. Ulna_Enr1-nonU



b. Ulna_Enr2-nonU



c. NOD1+2-nonU

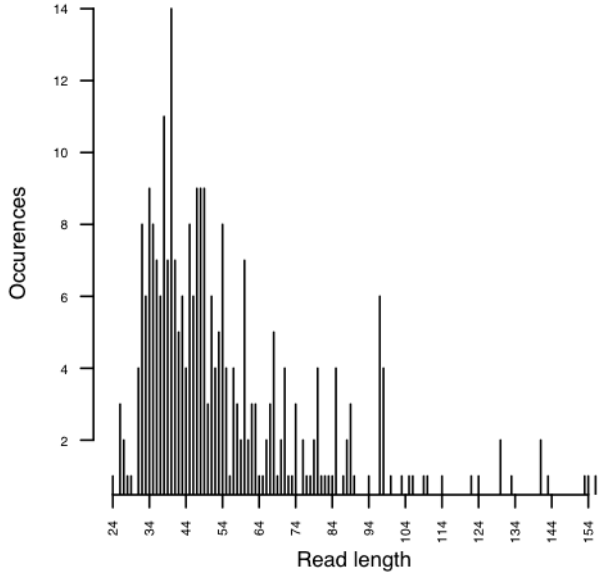


Figure S12. Gene diagrams for *S. saprophyticus* novel regions

Gene figures generated in Geneious (v.6.1.6). Note that scale for Insert 02 is different from inserts 03, 13, 15, & 16. Color of gene labels gives orientation (green = forward, red = reverse).

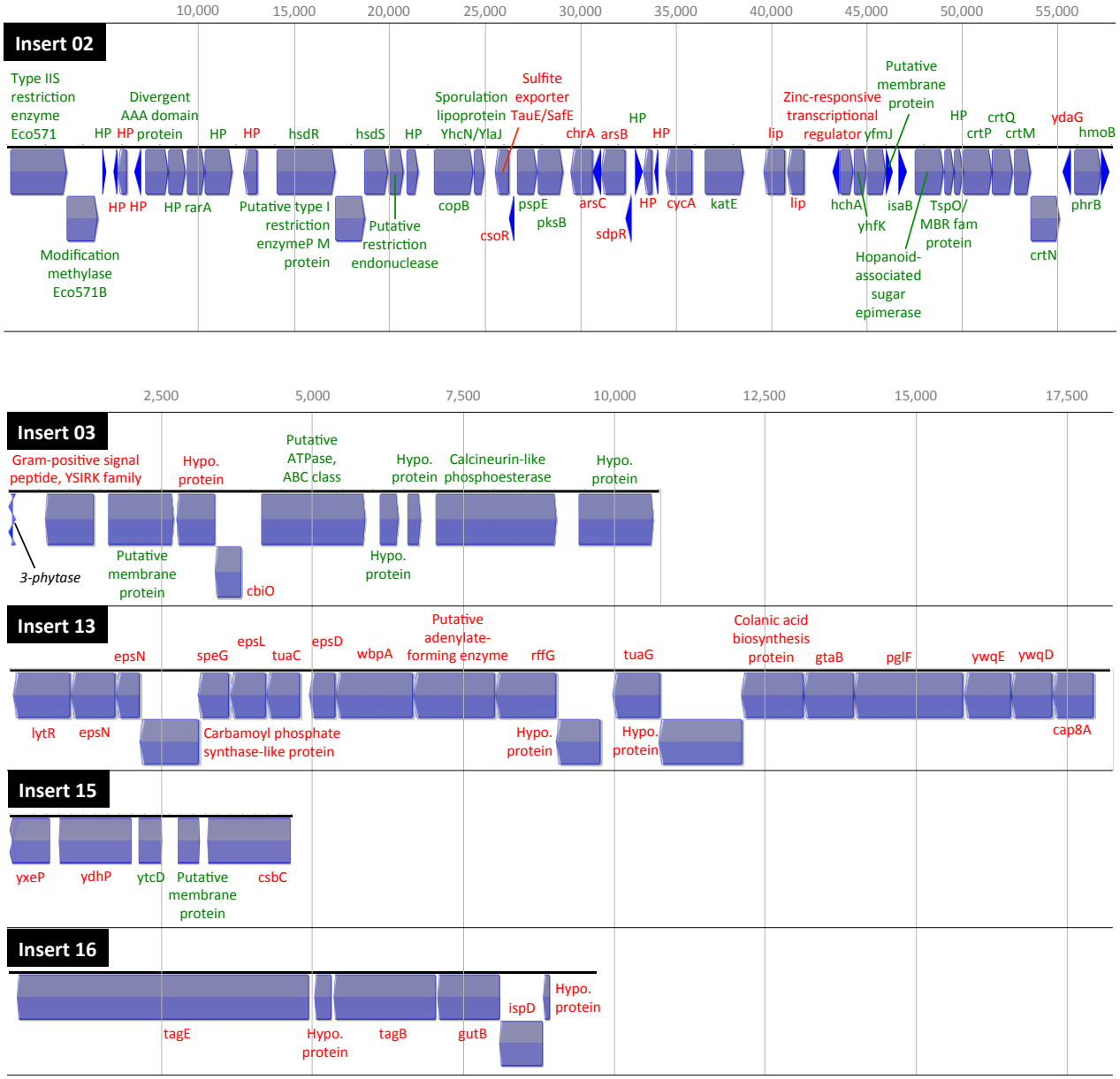


Figure S13. MAUVE alignment of *S. saprophyticus Troy* plasmid pSST1

Portion of MAUVE whole-genome alignment of *S. saprophyticus* Troy genome, plasmid pSST1, ATCC 15305 genome, and plasmid pSSAP1 from strain MS1146 using the MAUVE plugin in Geneious (v.6.1.6) with default settings.

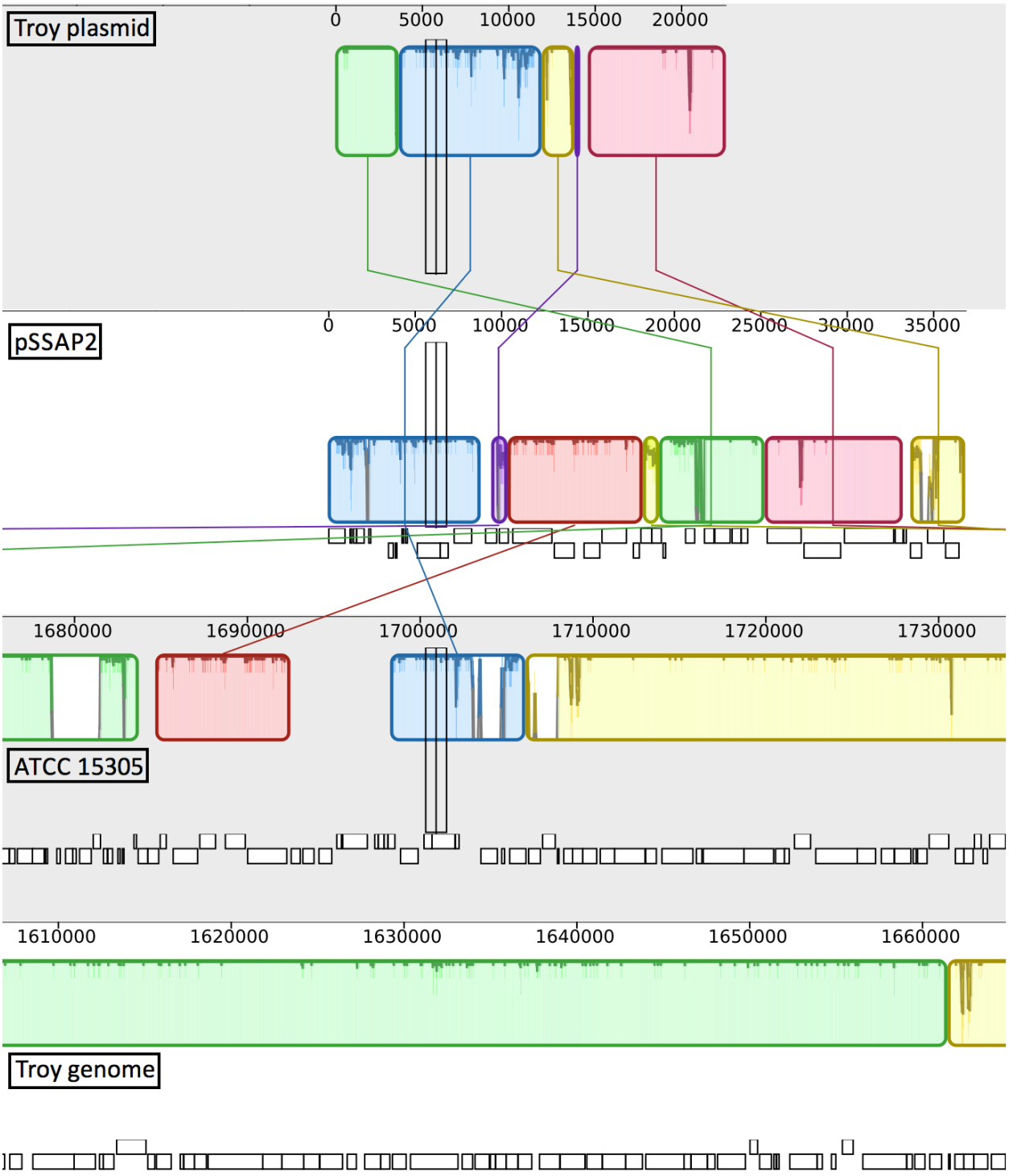


Figure S14. Coverage diagram for *G. vaginalis* ATCC 14019

Nod1_1h-UDG paired-end reads (1,964,443 readpairs) mapped to the modern human *G. vaginalis* reference strain (ATCC 14019) using bwa (76) with default settings and duplicates collapsed using samtools rmdup (73). Image generated using BRIG (68). Maximum coverage displayed is 400X (outermost tickmarks).

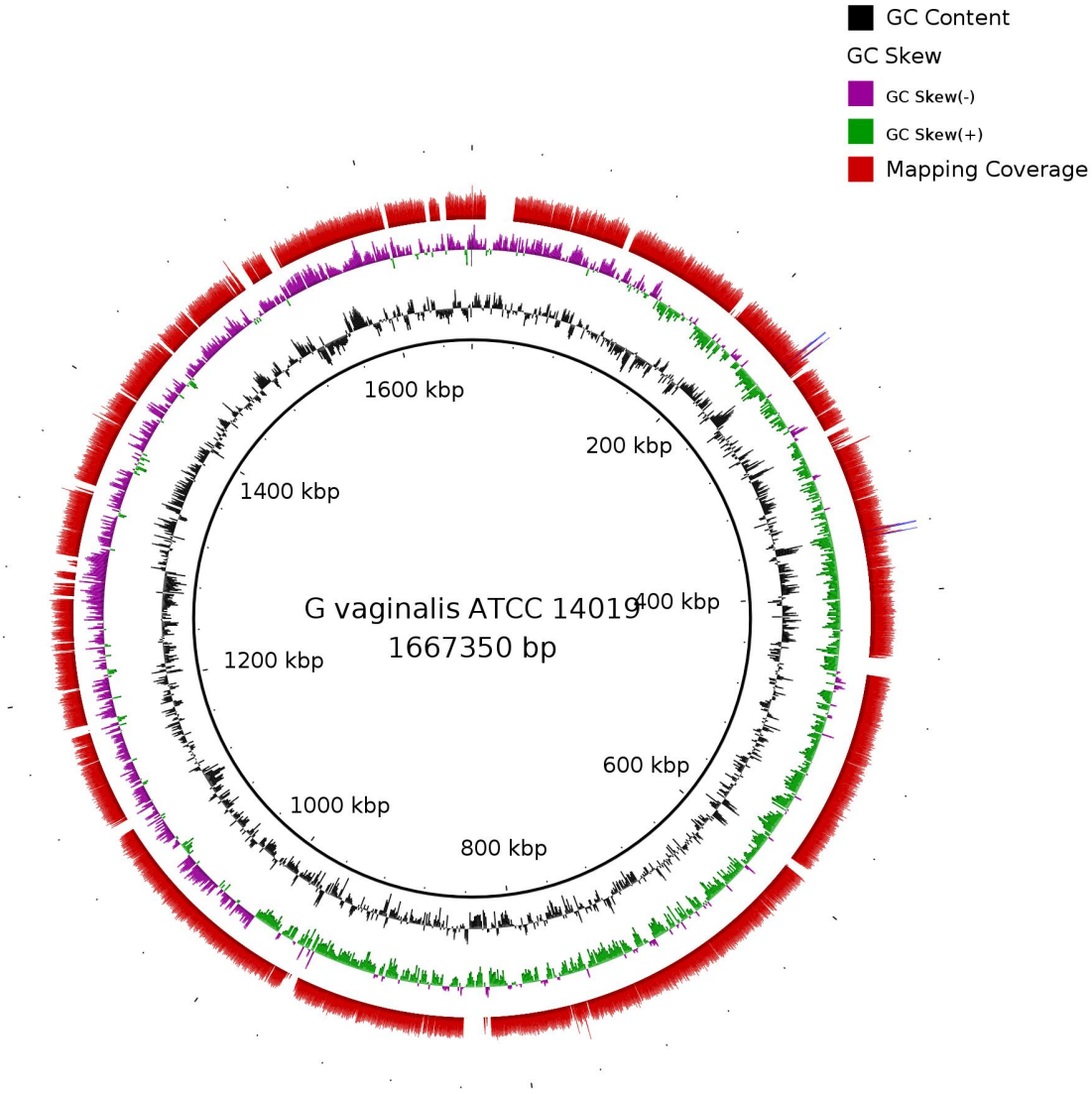
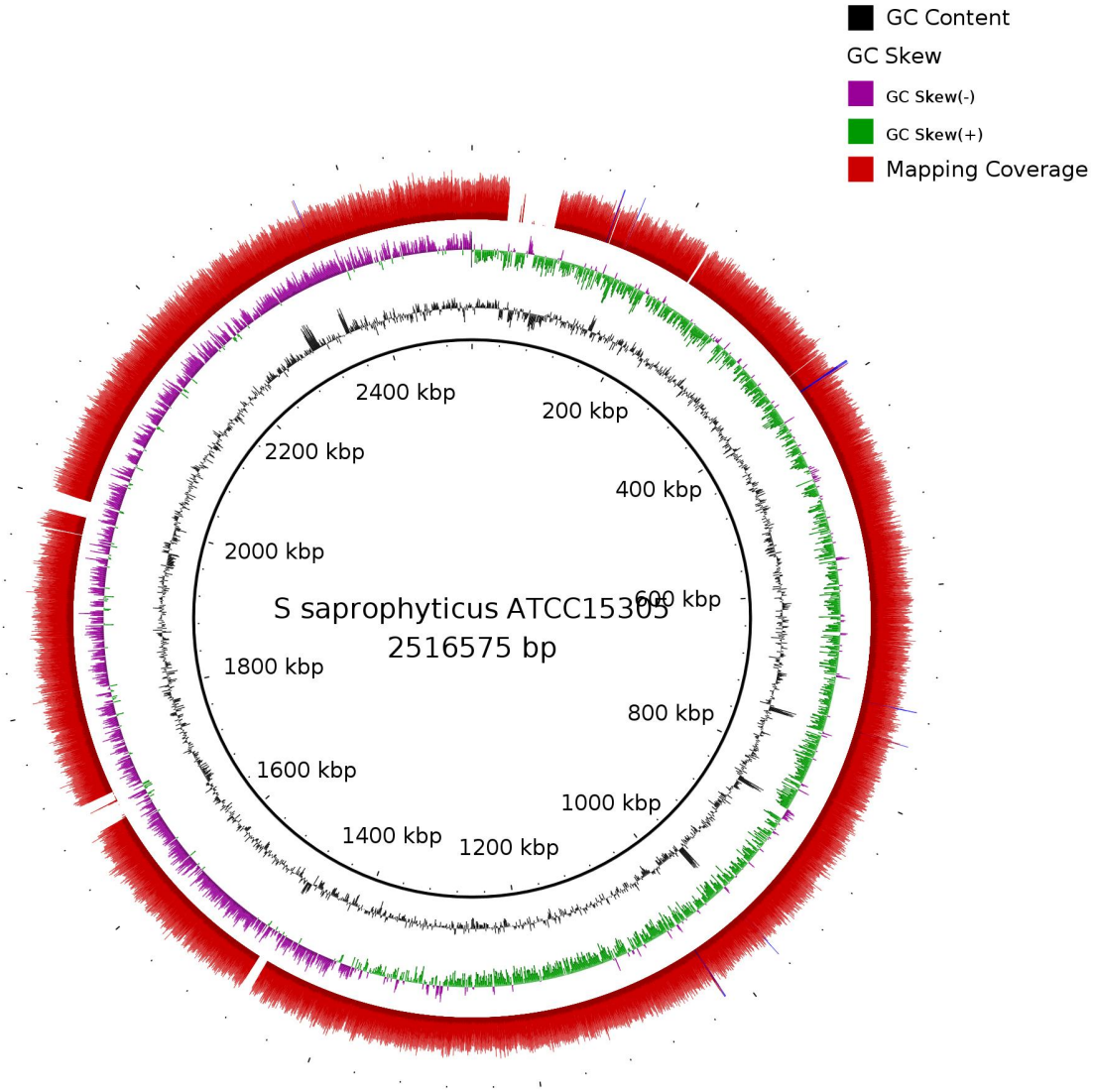


Figure S15. Coverage diagram for *S. saprophyticus* ATCC 15305

Nod1_1h-UDG paired-end reads (10,038,389 readpairs) mapped to the modern human *S. saprophyticus* reference strain (ATCC 15305) using bwa (76) with default settings and duplicates collapsed using samtools rmdup (73). Image generated using BRIG (68). Maximum coverage displayed is 800X (outermost tickmarks).



3.13 References

1. Kay GL *et al.* 2014. Recovery of a Medieval Brucella melitensis Genome Using Shotgun Metagenomics. *mBio* 5.
2. Wagner DM *et al.* 2014. Yersinia pestis and the Plague of Justinian 541-543 AD: a genomic analysis. *Lancet Infect Dis* 14**:**319-326.
3. Bos KI *et al.* 2011. A draft genome of Yersinia pestis from victims of the Black Death. *Nature* 478**:**506-510.
4. Schuenemann VJ *et al.* 2013. Genome-wide comparison of Medieval and modern Mycobacterium leprae. *Science* 341**:**179-183.
5. Bouwman AS *et al.* 2012. Genotype of a historic strain of Mycobacterium tuberculosis. *Proc Natl Acad Sci USA* 109**:**18511-18516.
6. Devault AM *et al.* 2014. Ancient pathogen DNA in archaeological samples detected with a Microbial Detection Array. *Sci Rep* 4.
7. Devault AM *et al.* 2014. Second-Pandemic Strain of *Vibrio cholerae* from the Philadelphia Cholera Outbreak of 1849. *N Engl J Med* 370**:**334-340.
8. Warinner C *et al.* 2014. Pathogens and host immunity in the ancient human oral cavity. *Nat Genet* 46**:**336-344.
9. Adler CJ *et al.* 2013. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nat Genet* 45**:**450-455.
10. Tito RY *et al.* 2012. Insights from Characterizing Extinct Human Gut Microbiomes. *PLoS ONE* 7**:**e51146.
11. Khairat R *et al.* 2013. First insights into the metagenome of Egyptian mummies using next-generation sequencing. *J Appl Genetics* 54**:**309-325.
12. Keller A *et al.* 2012. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat Commun* 3**:**698.
13. Fornaciari G, Giuffra V. 2012. Soft Tissue Tumors in Palaeopathology: A Review. *Pathobiology* 79**:**257-267.
14. Quintelier K. 2009. Calcified uterine leiomyomata from a post-medieval nunnery in Brussels, Belgium. *Int J Osteoarchaeol* 19**:**436-442.
15. Komar D, Buikstra JE. 2003. Differential diagnosis of a prehistoric biological object from the Koster (Illinois) Site. *Int J Osteoarchaeol* 13**:**157-164.
16. Anderson T. 2003. A medieval bladder stone from Norwich, Norfolk. *Int J Osteoarchaeol* 13**:**165-167.
17. Donoghue HD, Spigelman M, Zias J, Gernaey-Child AM, Minnikin DE. 1998. Mycobacterium tuberculosis complex DNA in calcified pleura from remains 1400 years old. *Lett Appl Microbiol* 27**:**265-269.
18. Ben Zakour NL, Guinane CM, Fitzgerald JR. 2008. Pathogenomics of the staphylococci: insights into niche adaptation and the emergence of new virulent strains. *FEMS Microbiol Lett* 289**:**1-12.

19. Widerström M, Wiström J, Ferry S, Karlsson C, Monsen T. 2007. Molecular Epidemiology of Staphylococcus saprophyticus Isolated from Women with Uncomplicated Community-Acquired Urinary Tract Infection. *J Clin Microbiol* 45**:**1561-1564.
20. Pleckaityte M, Zilnyte M, Zvirbliene A. 2012. Insights into the CRISPR/Cas system of Gardnerella vaginalis. *BMC Microbiology* 12**:**301.
21. Ahmed A *et al.* 2012. Comparative Genomic Analyses of 17 Clinical Isolates of Gardnerella vaginalis Provide Evidence of Multiple Genetically Isolated Clades Consistent with Subspeciation into Genovars. *J Bacteriol* 194**:**3922-3937.
22. Kiesewetter H. n.d. Site Report: Analyses of Human Remains from Late Byzantine Burials at Troy, p. 29.
23. Uerpmann H-P. 2003. *Environmental Aspects of Economic Changes in Troia*, p. 251-262. *In* Wagner GA, Pernicka E, Uerpmann H-P (ed.), *Troia and the Troad*. Springer, New York, NY.
24. 2003. *Troia and the Troad*, Wagner GA, Pernicka E, Uerpmann H-P (ed.), Springer, New York, NY.
25. Jablonka P. 2006. Preliminary Report on Work at Troia 2005. *Studia Troica* 16**:**3-26.
26. Hruska K, Kaevska M. 2012. Mycobacteria in water, soil, plants and air: a review. *Veterinarni Medicina* 57.
27. Derenko M *et al.* 2013. Complete Mitochondrial DNA Diversity in Iranians. *PLoS ONE* 8**:**e80673.
28. Redline R. 2007. *Placental Inflammation*, p. 90-101. *In* Keeling J, Khong TY (ed.), *Fetal and Neonatal Pathology*. Springer London.
29. Sayer D, Dickinson SD. 2013. Reconsidering obstetric death and female fertility in Anglo-Saxon England. *World Archaeol* 45**:**285-297.
30. WHO Report**.** 2013. Trends in maternal mortality: 1990 to 2013 - Estimates by WHO, UNICEF, UNFPA, The World Bank and the United Nations Population Division. World Health Organization.
31. Balaji KC, Menon M. 1997. Mechanism of stone formation. *Urol Clinics of N Am* 24**:**1-11.
32. Cormier CM *et al.* 2006. Urolithiasis in Pregnancy: Current Diagnosis, Treatment, and Pregnancy Complications. *Obstet Gynecol Surv* 61**:**733-741.
33. Higashide M *et al.* 2008. Methicillin-Resistant Staphylococcus saprophyticus Isolates Carrying Staphylococcal Cassette Chromosome mec Have Emerged in Urogenital Tract Infections. *Antimicrob Agents Chemother* 52**:**2061-2068.
34. Harwich MD, Jr. *et al.* 2010. Drawing the line between commensal and pathogenic Gardnerella vaginalis through genome analysis and virulence studies. *BMC Genomics* 11.
35. Lewis DA *et al.* 2013. The human urinary microbiome; bacterial DNA in voided urine of asymptomatic adults. *Front Cell Infect Microbiol* 3.
36. Fouts D *et al.* 2012. Integrated next-generation sequencing of 16S rDNA and metaproteomics differentiate the healthy urine microbiome from asymptomatic bacteriuria in neuropathic bladder associated with spinal cord injury. *J Translat Med* 10**:**174.
37. Fowler JJE. 1985. Staphylococcus saprophyticus as the Cause of Infected Urinary Calculus. *Ann Intern Med* 102**:**342-343.

38. Gatermann S, John J, Marre R. 1989. Staphylococcus saprophyticus urease: characterization and contribution to uropathogenicity in unobstructed urinary tract infection of rats. *Infect Immun* 57**:**110-116.
39. Lagier R, Baud CA. 2003. Magnesium Whitlockite, a Calcium Phosphate Crystal of Special Interest in Pathology. *Pathology - Research and Practice* 199**:**329-335.
40. Raz R, Colodner R, Kunin CM. 2005. Who Are You—Staphylococcus saprophyticus? *Clin Infect Dis* 40**:**896-898.
41. Widerstrom M, Wistrom J, Sjostedt A, Monsen T. 2012. Coagulase-negative staphylococci: update on the molecular epidemiology and clinical presentation, with a focus on Staphylococcus epidermidis and Staphylococcus saprophyticus. *Eur J Clin Microbiol Inf* 31**:**7-20.
42. Rupp ME, Soper DE, Archer GL. 1992. Colonization of the female genital tract with Staphylococcus saprophyticus. *J Clin Microbiol* 30**:**2975-2979.
43. Kuroda M *et al.* 2005. Whole genome sequence of Staphylococcus saprophyticus reveals the pathogenesis of uncomplicated urinary tract infection. *Proc Natl Acad Sci USA* 102**:**13272-13277.
44. Eriksson A, Giske CG, Ternhag A. 2013. The relative importance of Staphylococcus saprophyticus as a urinary tract pathogen: distribution of bacteria among urinary samples analysed during 1 year at a major Swedish laboratory. *Apmis* 121**:**72-78.
45. Hedman P, Ringertz O, Lindstrom M, Olsson K. 1993. The origin of Staphylococcus saprophyticus from cattle and pigs. *Scand J Infect Dis* 25**:**57-60.
46. Weinert LA *et al.* 2012. Molecular dating of human-to-bovid host jumps by Staphylococcus aureus reveals an association with the spread of domestication. *Biol Lett* 8**:**829-832.
47. Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 29**:**1969-1973.
48. Korte-Berwanger M *et al.* 2013. Significance of the d-Serine-Deaminase and d-Serine Metabolism of Staphylococcus saprophyticus for Virulence. *Infect Immun* 81**:**4525-4533.
49. Sakinç T, Kleine B, Gatermann SG. 2007. Biochemical characterization of the surface-associated lipase of Staphylococcus saprophyticus. *FEMS Microbiol Lett* 274**:**335-341.
50. Sakinc T, Woznowski M, Ebsen M, Gatermann SG. 2005. The Surface-Associated Protein of Staphylococcus saprophyticus Is a Lipase. *Infect Immun* 73**:**6419-6428.
51. Kleine B, Gatermann S, Sakinc T. 2010. Genotypic and phenotypic variation among Staphylococcus saprophyticus from human and animal isolates. *BMC Research Notes* 3**:**163.
52. von Eiff C, Peters G, Heilmann C. 2002. Pathogenesis of infections due to coagulasenegative staphylococci. *Lancet Infect Dis* 2**:**677-685.
53. Sakınç T, Kleine B, Michalski N, Kaase M, Gatermann SG. 2009. SdrI of Staphylococcus saprophyticus is a multifunctional protein: localization of the fibronectin-binding site. *FEMS Microbiol Lett* 301**:**28-34.
54. Sakinc T, Kleine B, Gatermann SG. 2006. SdrI, a Serine-Aspartate Repeat Protein Identified in Staphylococcus saprophyticus Strain 7108, Is a Collagen-Binding Protein. *Infect Immun* 74**:**4615-4623.

55. King NP *et al.* 2011. UafB is a serine-rich repeat adhesin of Staphylococcus saprophyticus that mediates binding to fibronectin, fibrinogen and human uroepithelial cells. *Microbiology* 157**:**1161-1175.
56. Hell W, Reichl S, Anders A, Gatermann S. 2003. The autolytic activity of the recombinant amidase of Staphylococcus saprophyticus is inhibited by its own recombinant GW repeats. *FEMS Microbiol Lett* 227**:**47-51.
57. Swoboda JG, Campbell J, Meredith TC, Walker S. 2010. Wall Teichoic Acid Function, Biosynthesis, and Inhibition. *ChemBioChem* 11**:**35-45.
58. Qian Z *et al.* 2006. Genomic characterization of ribitol teichoic acid synthesis in Staphylococcus aureus: genes, genomic organization and gene duplication. *BMC Genomics* 7**:**74.
59. Vickers AA, Chopra I, O'Neill AJ. 2007. Intrinsic Novobiocin Resistance in Staphylococcus saprophyticus. *Antimicrob Agents Chemother* 51**:**4484-4485.
60. Park S *et al.* 2010. Characterization of the Structure and Biological Functions of a Capsular Polysaccharide Produced by Staphylococcus saprophyticus. *J Bacteriol* 192**:**4618-4626.
61. King N *et al.* 2012. Characterisation of a cell wall-anchored protein of Staphylococcus saprophyticus associated with linoleic acid resistance. *BMC Microbiology* 12**:**8.
62. Yeoman CJ *et al.* 2010. Comparative Genomics of Gardnerella vaginalis Strains Reveals Substantial Differences in Metabolic and Virulence Potential. *Plos One* 5.
63. Pleckaityte M, Janulaitiene M, Lasickiene R, Zvirbliene A. 2012. Genetic and biochemical diversity of Gardnerella vaginalis strains isolated from women with bacterial vaginosis. *FEMS Immunol Med Microbiol* 65**:**69-77.
64. Lee W, Phillips LE, Carpenter RJ, Martens MG, Faro S. 1987. Gardnerella vaginalis chorioamnionitis: A report of two cases and a review of the pathogenic role of G. vaginalis in obstetrics. *Diagnostic Microbiology and Infectious Disease* 8**:**107-111.
65. Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC. 2011. Integrative analysis of environmental sequences using MEGAN4. *Genome Res* 21**:**1552-1560.
66. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9**:**357-359.
67. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. *J Mol Biol* 215**:**403-410.
68. Alikhan N-F, Petty N, Ben Zakour N, Beatson S. 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12**:**402.
69. Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res* 13**:**2178-2189.
70. Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* 40.
71. Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protoc* 2010**:**1-10, 11-17.
72. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 17.
73. Li H *et al.* 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25**:**2078-2079.

74. McKenna A *et al.* 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20**:**1297-1303.
75. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*.
76. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25**:**1754-1760.
77. Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18**:**821-829.
78. Kim BS *et al.* 2012. Draft Genome Sequence of Staphylococcus saprophyticus subsp saprophyticus M1-1, Isolated from the Gills of a Korean Rockfish, Sebastes schlegeli Hilgendorf, after High Hydrostatic Pressure Processing. *J Bacteriol* 194**:**4441-4442.
79. Angiuoli SV, Salzberg SL. 2011. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* 27**:**334-342.
80. Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* 30**:**772-780.
81. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25**:**1972-1973.
82. Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. 2013. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29**:**1682-1684.

Chapter 4

Ancient pathogen DNA in archaeological samples detected with a Microbial Detection Array

Published **March 6, 2014** in **Scientific Reports** 4:4245
Reprinted with permission.
doi: 10.1038/srep04245

Alison M. Devault[1], Kevin McLoughlin[2], Crystal Jaing[2], Shea Gardner[2], Teresita M. Porter[3], Jacob M. Enk[1,3],James Thissen[2], Jonathan Allen[2], Monica Borucki[2], Sharon N. DeWitte[4], Anna N. Dhody[5], Hendrik N. Poinar[1,3,6]

(1) McMaster Ancient DNA Centre, Department of Anthropology, McMaster University, 1280 Main St W, Hamilton, Ontario L8S4L9, Canada; (2) Lawrence Livermore National Laboratory, Livermore, CA 94551, USA;(3) Department of Biology, McMaster University, 1280 Main St W, Hamilton, Ontario L8S 4K1, Canada; (4) Departments of Anthropology and Biological Sciences, University of South Carolina, Columbia, SC, USA; (5) The College of Physicians of Philadelphia, Mütter Museum, 19 S 22nd St, Philadelphia, PA 19103, USA; (6) Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, 1280 Main St W, Hamilton, Ontario L8S4L8, Canada.

4.1 Thesis Preface

Co-infections and the microbiome are important components of health, now as in the past. Even though there is still much to be established about their roles in modern health, archaeologists may still want to directly assess the metagenomic content of their specimens for particular co-infection research questions. For example, archaeological specimens have been examined for the co-infection of *M. tuberculosis* and *M. leprae* to test the hypothesis that cross-immunity from TB infection may have contributed to the historical decline of leprosy (1). However, targeted approaches such as these are necessarily biased, due to the need to choose predetermined pathogenic species for study, as discussed in Chapter 1. The advent of HTS technology allows for the large-scale assessment of metagenomic content, but the requisite sequencing depths for ancient bacterial identification unfortunately necessitate a high price tag. Some shotgun approaches have been used to assess the metagenomic content of ancient samples, for example tissues from mummified remains from Egypt (2) and a bone microbiome from Ötzi, the Tyrolean Iceman (3). However, the necessity of broad and deep sequencing for large sample sets (e.g., a burial population from an archaeological site) would likely be prohibitively expensive for most researchers. This is especially true for exploratory studies where any particular microbes of interest are unknown and a deep microbial survey is

necessary. Alternatively, targeted sequencing approaches (PCR and TE-HTS) could be used to assess the microbial content and save on some (but not all) sequencing costs, due to the enrichment of the metagenomic locus of interest out of a pool of other endogenous and exogenous DNA content. The most likely candidate for a targeted metagenomic survey of ancient microbial content would be the 16S rRNA locus, which has been applied to ancient specimens such as the oral microbiome from Neolithic dental calculus (4), and ancient coprolites (5). However, such targeted approaches such as 16S PCR metagenomic survey have additional tradeoffs in terms of limited resolution and the biases of amplification, as discussed in Chapter 1. Additionally, all targeted approaches are necessarily slower than shotgun HTS due to the requirement for additional laboratory steps.

In modern clinical work, metagenomic methods have been used in the modern clinical "diagnosis" of tricky specimens where the target(s) of interest are unknown, in addition to their benefit of increasing speed of throughput by bypassing biased and time-consuming bacterial culture (6). For example, HTS strategies have been used to detect known pathogenic viruses, such as influenza or norovirus, in nasal swab and faecal samples from patients during disease outbreaks (7). Pre-designed microarrays have also been used to parse complex samples for metagenomic content, for example pathogen presence amongst background microbes (8). Microarrays represent a targeted enrichment of DNA from a sample; however, the abundance of annealed molecules to each variant of probe/bait sequence is then directly visualized with fluorescence, rather than being eluted off the baits for HTS. This offers a very rapid turnaround without the need for handling large amounts of sequencing data (such as with metagenomic or 16S HTS projects). While interpretations might be complicated by novel, unknown taxa and/or highly complex samples, they offer an attractive rapid avenue for diagnosticians (6), even more so when traditional diagnostic methods fail, or are contradictory (9). In many ways, this is similar to the predicament of palaeopathological researchers attempting to "diagnose" past infections from historical records and skeletal indicators – we have far less information to go on, and an enormous depth of time spanning a huge amount of microbial diversity.

Such a middle-ground method between HTS and targeted PCR, which can thoroughly assess microbial content in ancient specimens with degraded DNA, yet be applied quickly and without extra laboratory preparation, would be ideal. This would allow for the broad survey of the pathogenic and microbiomic content of many bioarchaeological specimens. This information could be used in statistical analysis to infer wide-ranging patterns about population health across time and space, or more immediately, to screen for candidate samples for deeper sequencing projects. We chose to evaluate an existing microarray design – Microbial Detection Array (LLMDA) technology from the Lawrence Livermore National Laboratory) – because of the demonstrated success of the technology (8, 10, 11), which includes a large specially designed

set of probes specific for metagenomic analysis as well as the likelihood algorithm which parses this information.

Our evaluation of the LLMDA reveals that despite the degradation and damage of ancient DNA molecules, *V. cholerae* can be detected in the 19th century cholera victim specimen (Chapter 2) and *Y. pestis* can be detected in a medieval tooth that had previously yielded a draft genome of that pathogen. The overall microbial profiles provided by LLMDA and deep HTS metagenomic analysis are similar, but distinct, which is likely attributable to the differences in identification of these methods. LLMDA utilizes hybridization to capture matching sequences from the samples, with carefully designed probes that are analyzed with algorithms designed for complex metagenomic mixtures. Because of this, the LLMDA is still a type of targeted approach. Therefore, despite its demonstrated ability for a broad metagenomic assessment, it is not the ideal long-term option. However, there are tradeoffs to choosing one approach over the other, and because of the current very high costs to generate and analyze HTS datasets, the viable LLMDA option is likely to be an extremely attractive tool for rapid and affordable ancient metagenomic analysis. Additionally, we did not alter the standard hybridization or washing protocols from that used for modern samples; therefore, it is possible that with further optimization for aDNA, we might achieve even better results on specimens with high complexity and low target abundance.

4.2 Abstract

Ancient human remains of paleopathological interest typically contain highly degraded DNA in which pathogenic taxa are often minority components, making sequence-based metagenomic characterization costly. Microarrays may hold a potential solution to these challenges, offering a rapid, affordable, and highly informative snapshot of microbial diversity in complex samples without the lengthy analysis and/or high cost associated with high-throughput sequencing. Their versatility is well established for modern clinical specimens, but they have yet to be applied to ancient remains. Here we report bacterial profiles of archaeological and historical human remains using the Lawrence Livermore Microbial Detection Array (LLMDA). The array successfully identified previously-verified bacterial human pathogens, including *Vibrio cholerae* (cholera) in a 19th century intestinal specimen and *Yersinia pestis* ("Black Death" plague) in a medieval tooth, which represented only minute fractions (0.03% and 0.08% alignable high-throughput shotgun sequencing reads) of their respective DNA content. This demonstrates that the LLMDA can identify primary and/or co-infecting bacterial pathogens in ancient samples, thereby serving as a rapid and inexpensive paleopathological screening tool to study health across both space and time.

4.3 Introduction

Research into the origins of infectious diseases and population health through time faces many challenges, such as biased archival records and ambiguous paleopathological skeletal indicators of actual pathogen infection levels (12). Despite its inherent fragility, ancient DNA (aDNA) remains a highly informative paleopathological study target, having been recovered and characterized from a variety of contexts, age depths and specimen types (13). Recently, high-throughput sequencing (HTS), often coupled with targeted enrichment (TE), has allowed for the recovery of large genomic targets from archaeological specimens, including full pathogen genomes (14-16). However, TE-HTS is only useful when the primary pathogen(s) are known or suspected to be present, and necessarily ignores non-targeted taxa and genomic loci. This is problematic because the primary pathogenic agent in an ancient paleopathological specimen can be elusive, and furthermore the entire microbiome likely played a significant role in past human health, as it does today (17). Therefore establishing detailed levels of commensal and co-infecting pathogens is essential for accurately reconstructing past epidemics, population health, and disease susceptibility. As such, for paleopathologists wishing to examine changes in microbial co-infection levels across space and time, more comprehensive metagenomic characterization is necessary. One way to achieve this is by sequencing amplicons of conserved loci (such as 16S rRNA) that can to a degree measure the metagenomic content of a sample. However, by design, amplicon datasets ignore potential taxonomically-informative diversity in more variable genomic regions, and for that matter can be biased by polymerase or disparate target abundances (18, 19). Metagenomic “shotgun” HTS on the other hand is arguably the most comprehensive and least biased method currently available for total microbial characterization for modern and aDNA specimens (2, 3), but very deep sequencing is often required to identify pathogens confidently. While certainly powerful, both of these metagenomic approaches can be labour- and time-intensive, thereby representing significant barriers for groups that would like to thoroughly profile or screen the microbial content of large or difficult paleopathological sample sets.

One potential technological solution to this issue is the microarray, which over the past two decades has been used for the large-scale study of gene expression and genic content of simple and complex samples (20). Microarrays are glass slides densely spotted with clusters of single-stranded synthetic oligonucleotides that are allowed to hybridize with fluorophore-labeled DNA from a sample, and the resulting fluorescence signals are interpreted to determine sequence composition and/or taxonomic content. Recently, microarrays designed specifically for characterizing the microbial content of complex samples have been successfully used, e.g., (8-11, 21, 22), particularly in cases where traditional clinical methods are inconclusive, time-consuming, and/or expensive (9). Microarrays can contain up to millions of unique

oligonucleotides and their use and analysis involve low processing time and cost (8). Therefore, they potentially provide a more practical alternative to metagenomic HTS for characterizing the microbial content of paleopathological specimens. However, microarray detection techniques have not yet been applied to aDNA extracts, which due to short fragment length and base damage may present challenges.

To assess the potential value of microarrays for pathogen detection in ancient samples, we compared microbial profiles of two archaeological human specimens generated with a recently-developed pathogen detection microarray to profiles generated with standard metagenomic HTS analysis. For microarray analysis, we used the Lawrence Livermore Microbial Detection Array (LLMDA) designed by the Lawrence Livermore National Laboratory (11), one of several array platforms developed in the last decade to identify pathogens in experimental mixtures and clinical samples (8). The LLMDA v5 12x135K array contains probes designed from all published vertebrate-infecting pathogen genomes. LLMDA probes target conserved regions amongst all known species/strains of a family (or equivalent unit), but due to the high number and overall diversity of probes, unique combinations of matching probes across an individual genome sequence allow for species or strain identification. Florescence data are analysed using a likelihood maximization algorithm to identify the combination of species that best explains the resulting signal. To achieve this, each signal set is compared against a current database of full microbial genomes and analysed for the expected vs. detected combined probe fluorescence signal, resulting in a species list ranked by likelihood of presence. If desired, these results can then be parsed to calculate overall likelihoods of higher taxa presence by summing the likelihoods of relevant species-level hits (see Supplementary Information for full description). The specimens we analysed here were a preserved medical intestine sample from an 1849AD cholera victim (specimen 3090.13) (14) and a tooth from a 1348AD Black Death plague victim (specimen 8291) (15). Both were previously confirmed with TE-HTS to contain their relevant pathogens, though they constitute very low levels in shotgun HTS datasets (3090.13: 0.03% alignable with bowtie (23) to *Vibrio cholerae*, the etiological agent of cholera; 8291: 0.08% alignable to *Yersinia pestis,* the etiological agent of the plague or Black Death). Both of these pathogens' families (Vibrionaceae and Enterobacteraceae) have probes on the LLMDA, and therefore species in these families should be detectable. We specifically assessed (a) whether LLMDA would detect the previously determined pathogens in our ancient samples, (b) which additional bacterial taxa were detectable by both LLMDA and HTS, and (c) which bacterial taxa were detected by either LLMDA or HTS alone.

4.4 Results

We have restricted our taxonomic comparisons to bacteria, since the sequencing libraries were built from DNA only and thus not appropriate for a complete viral survey. While both HTS and LLMDA analyses are capable of species-level identification (as LLMDA analysis calculates the likelihood of presence for individual species/strain genomes), for the purposes of this paper we have focused on family-level identification for ease of comparison. Note that the v5 12x135K LLMDA probes were derived from all complete vertebrate-infecting pathogen genome sequences available at the time of design (December 2011); however, as the hybridization patterns were interpreted using an updated database (April 2012), probes may match new genomes from other taxa (even non-vertebrate infecting species) and therefore potential taxonomic calls are not limited to those used for probe design. For the metagenomic HTS data, taxonomic assignments were identified by BLAST (blastn-megablast) (24) and MEGAN4 (25) analysis against the National Center for Biotechnology Information (NCBI) RefSeq genome database (26) (October 2012). A schematic comparison is provided in **Fig. 1** and results for both methods are given in **Table 1** and **Tables S1** and **S2**.

4.4.1 Taxa detected by both LLMDA & HTS

For cholera victim 3090.13, twenty-one bacterial families were detected by both LLMDA and the 118 million BLASTed HTS reads from the sample (**Fig. 1**), representing 36.8% and 40.4% of the families called by each respective method. For plague victim 8291, fifty-three families were detected by both approaches, representing 89.8% and 27.9% of the families called by LLMDA and 83 million HTS reads, respectively. When we considered only the families with specific probes designed for them on the LLMDA array, we detected 19 families in the cholera victim 3090.13 by both LLMDA and HTS, representing 41.3% and 79.2% of the families called by each respective method and 46 families for the plague victim 8291, representing 92.0% and 56.8% of the families with probes on the array (LLMDA and HTS).

The taxa detected by both methods included many groups with relatively high read counts in the HTS data (e.g. Aeromonadaceae and Enterobacteriaceae for 3090.13; Burkholderiaceae, Comamonadaceae, and Pseudomonadaceae for 8291). In addition, both methods detected the previously confirmed significant pathogens to the species level within the calls for their respective families (see **Fig. S1** and **Table S3** for the LLMDA probe data supporting these pathogen calls). For 3090.13, 10,379 (0.009% of BLAST reads) were *V. cholerae,* and LLMDA called the family Vibrionaceae with *V. cholerae* chromosomal sequences at a high log odds value (4,470.7). The detection of only *V. cholerae* chromosomal sequences amongst Vibrionaceae is not unexpected, given that Vibrionaceae spp. are aquatic and only opportunistically infect humans (27). For 8291, 1,272 (0.001% of BLAST reads) were *Y. pestis,* and LLMDA called the family Enterobacteriaceae including *Y. pestis* pPCP1 plasmid sequences

(among multiple other species) at a high odds value (1,640.8). *Y. pestis* pPCP1 was detected amongst a variety of other Enterobacteriaceae, which is a large family containing many normal microbiomic species such as *E. coli* (28). The detection of pPCP1 is not surprising, given that the plasmid is both highly specific to *Y. pestis* and is found at high copy number with the bacterium (29), which may have facilitated detection.

4.4.2 Taxa detected by only one method

BLAST analyses of HTS reads identified many bacterial families that were not detected by LLMDA analyses (for cholera victim 3090.13, n = 10, 32.3% of all HTS; for plague victim 8291, n = 137, 72.1% of all HTS), such as Neisseriaceae and Shewanellaceae in sample 3090.13, Cellulomonadaceae and Rhizobiaceae in sample 8291, and Fusobacteriaceae and Peptostreptococcaceae in both samples. Likewise, LLMDA analysis identified many families that HTS did not (for 3090.13, n = 36, 63.1% of all LLMDA; for 8291, n = 6, 10.2% of all LLMDA). However, when excluding taxonomic groups without specific probes represented on the array, only 5 families were detected by HTS alone (20.8% of all HTS) for sample 3090.13 and 35 (43.2%) for 8291, while for LLMDA alone, 27 (58.7% of all LLMDA) were detected for 3090.13 and 4 (8.0%) for 8291.

4.5 Discussion

Figure 2 (cholera victim 3090.13)and **Figure 3** (plague victim 8291) display the MEGAN4 output of the NCBI taxonomy for all family-level taxa identified with BLAST analysis of the HTS data and whether they were also detected with LLMDA. Overall, the LLMDA profiles reflect the major HTS-identified components well. Not only were the previously-identified pathogen bacterial species/families detected via both methods, but a number of major environmental, microbiomic and pathogenic taxa were identified to at least the order level (e.g., Actinomycetales, Bacilliales, Clostridiales, or Rhizobiales). While promising, a number of disparities between the profiles generated by each method encourage further investigation into their origin, discussed below.

When comparing metagenomic profiles generated by each method, it is important to be aware of the fundamental differences in their taxonomic identification strategies. For the analysis of BLAST output from HTS data, default parameters in MEGAN4 require five sequence reads to assign a taxon as being present; furthermore, the reads do not have to be assigned to the same species for family-level calls (25). MEGAN4 also gives equal weight to read mappings that are concentrated in narrow regions of a target genome, which are inherently less specific as indicators of the target's presence. A common possible scenario leading to false positive

taxon assignments could occur in both HTS and microarray analysis, when reads or probes map to ribosomal RNA or housekeeping genes that are relatively conserved between related taxa. Microarray probes can be designed to avoid these conserved regions, but in general sequence reads mapping to such regions are not filtered out in metagenomic analysis. Therefore, BLAST/MEGAN4 analysis of HTS data emphasizes sensitivity at the expense of specificity. The CLiMax algorithm used for LLMDA analysis requires that a family satisfy more stringent criteria to be considered present. The initial CLiMax analysis is performed at the target genome level rather than the family level; for a target to be called present, a minimum of 4 probes or 20% of the probes matching a specific genome (whichever is greater) must have intensities above an array-specific significance threshold. In addition, targets for which the high intensity probes are concentrated in narrow genomic regions are filtered out as potential false positives (see Supplementary Information for description of methods). When this filtering is removed, or if the minimum probe criteria are relaxed, CLiMax predicts the presence of several previously undetected families (data not shown). However, our previous experiments in which the LLMDA was hybridized to samples of known microbial content indicate that stringent filtering is necessary to avoid false positives (11). Therefore, the CLiMax analysis is much more conservative in its predictions than BLAST/MEGAN4 analysis, emphasizing specificity over sensitivity, and possibly explaining some of the apparent undetected taxa in the LLMDA data.

Several taxa detected with HTS were not detected with LLMDA. Many of these are unsurprising, as no probes designed from their genomes were present on the array. However, for those taxa with probes on the array, one possibility is that the LLMDA is simply not as sensitive as HTS at these sequencing depths: in plague victim 8291, taxa not detected with LLMDA had significantly fewer HTS reads than those that were (one-tailed, unequal variance Student's t-test, $p = 0.002$; **Fig. 4a**), though these variables are not significantly related for the cholera victim sample 3090.13 ($p = 0.076$). Furthermore, several taxa with relatively high read counts and with probes designed on the array were surprisingly not called (e.g., Sphingomonadaceae in sample 8291; Peptostreptococcaceae in both samples). That said, in the majority of cases where a family with probes designed on the array was declared present by BLAST/MEGAN4, but not called with LLMDA, a closely-related taxon was called (e.g., in both samples, Clostridiaceae was called though its close relative Peptostreptococcaceae was not).

To better understand the data used by MEGAN4 to call the family Peptostreptococcaceae as present, we examined the ribosomal RNA gene, and other feature annotations for the mapped read positions in *Clostridium difficile* strain 630 (RefSeq accession NC_009089.1), one of the fully sequenced genomes in this family. Notably, 915 of 1328 (69%) reads mapping to this genome from cholera victim 3090.13 and 146 of 319 (46%) from plague victim 8291 were within rRNA genes. Since rRNA genes only cover 1.1% of the *C. difficile* 630

genome, these read counts are much higher than would be expected by chance alone. Consequently, we suspect that a large part of the data used by MEGAN4 to call this family as present is based on reads that map to highly conserved genes, and could also support the presence of a related taxon. Although a detailed analysis of MEGAN4 performance is beyond the scope of this study, our preliminary results suggest that its relative non-specificity could underlie some of the discrepancies between HTS and microarray identifications.

We also considered the possibility that relatively low GC content of the targets could compromise hybridization-based LLMDA detection. Average log (fluorescence) intensity of probes for a given taxon is strongly correlated with the average GC% of that probe set ($r=0.56$, $p = 0.0028$, $R^2 = 0.368$ for cholera victim 3090.13; $r=0.65$, $p = 2.5 \times 10^{-13}$, $R^2 = 0.653$ for plague victim 8291; **Fig. S3**), but LLMDA detected taxa across the range of average log intensities. Furthermore, for taxa used for probe design, there was no significant difference in GC content between LLMDA-positive and LLMDA-negative HTS reads (two-tailed, unequal variance Student's t-test, $p = 0.252$ for 3090.13, $p = 0.779$ for 8291; **Fig. 4b**). This indicates that GC content alone cannot explain a taxon's presence or absence from the LLMDA calls. We also considered the possibility that confident LLMDA identification may be compromised if regional preservation or amplification biases reduce the evenness of genomic representation amongst the reads. However, for taxa with probes on the array, there was no significant difference between the proportions of unique genomic bases covered by HTS reads for LLMDA-positive and LLMDA-negative taxa (two-tailed, unequal variance Student's t-test, $p = 0.365$ for 3090.13, $p = 0.843$ for 8291; **Fig. 4c**).

Several taxa were detected only with LLMDA. This may suggest that the LLMDA is more sensitive than HTS to certain taxa, as a rarefaction analysis of the HTS data suggests that in neither sample have all the HTS-detectable families likely been observed at these sequencing depths (**Fig. 5**). Cholera victim 3090.13 in particular shows a near-linear rarefaction curve, potentially explaining why it has so many more LLMDA-only calls than does plague victim 8291. However, taxa detected by both HTS and LLMDA still have significantly higher LLMDA log odds scores than taxa detected by LLMDA alone (one-tailed, unequal variance Student's t-test, $p = 0.015$ for 3090.13, $p = 0.007$ for 8291; **Fig. 4d**). This difference likely reflects the fact that LLMDA calls with smaller log-odds scores are supported by fewer detected probes, and are thus inherently less reliable. However, the relationship between log odds scores and HTS observations is imperfect, as several taxa with relatively high read counts have maximum log odds score values within the range of LLMDA-only calls (e.g., Caulobacteraceae for sample 8291 and Moraxellaceae for sample 3090.13). Again as noted above, there is no significant difference between the proportion of unique genomic bases covered by HTS reads for LLMDA-positive and LLMDA-negative taxa (**Fig. 4c**).

We have demonstrated that the LLMDA provides similar bacterial family-level metagenomic profiles of archaeological and archival specimens as HTS, especially for the most abundant taxa, and successfully detected the previously-verified infecting pathogen species in both specimens. Furthermore, as demonstrated with cholera victim 3090.13, it is potentially capable of detecting bacterial taxa that are insufficiently or unable to be detected even with very large HTS datasets, due to the very deep sequencing depths required to observe low abundance HTS taxa, likely common for many co-infecting pathogens in complex aDNA extracts. This is encouraging, since LLMDA analysis is at least one order of magnitude less expensive and labor-intensive than metagenomic HTS. As such, the technique could be productively applied in a number of research settings, depending on the specific question and the nature of the specimens. For instance, dozens of samples could be rapidly assessed for the most abundant pathogen constituents. Use of the LLMDA may also integrate well into TE-HTS studies not only by narrowing the range of targets for hybridization capture, but also by generating enriched libraries via elution from the microarray itself, which can be later sequenced. However the profiles generated by the LLMDA and HTS are not identical, and criteria for confident taxon identification with both platforms remains imperfect. We have shown that no simple variable completely explains the signal disparities, and it is likely that a combination of analysis techniques, sequence factors such as GC content and probe design drive the disagreements between the LLMDA and HTS. Further methodological evaluation may be able to refine these disparities. We expect that microarrays will progress in the near future to become an excellent screening tool for archaeological samples where microbial profiles can be swiftly, cheaply, and accurately reconstructed thereby aiding the elucidation of population health through deep time.

4.6 Methods

Libraries from these specimens were both shotgun HTS sequenced (divided across one HiSeq 1000 lane: 141,039,627 reads for cholera victim 3090.13, 122,830,910 reads for plague victim 8291 and utilized for LLMDA analysis. HTS datasets were compared to the NCBI RefSeq database (26) using BLAST 2.2.26+ (24) and the resulting BLAST reports were parsed using MEGAN4 v.4.70.4 with the default settings (25). Taxonomic trees were illustrated manually using FigTree (v.1.4.0; http://tree.bio.ed.ac.uk/software/figtree) based on MEGAN4 results. Indexed libraries were sent to Lawrence Livermore National Laboratory (LLNL) for blind analysis using the 12-plex 135K Roche NimbleGen version of the LLMDA v5 array, which is designed to target 3521 vertebrate-infecting species from 215 families (including bacteria, archaea, viruses, protozoa and fungi). A brief summary of the LLMDA workflow is as follows: libraries are linearly amplified via random hexamers (Cy-3 labeled) to add the necessary fluorescent signal,

hybridized to the LLMDA array for 65h, washed, scanned, and analysed. Unlike other aDNA experiments utilizing in-solution or array hybridization, “blocking oligonucleotides” were not used, as this is not a standard component of the LLMDA procedure. Arrays were analysed using the CLiMax algorithm (11) with probe intensity threshold set to the 95th percentile of negative controls. See Supplementary Information for all further details.

4.7 Acknowledgements

We would like to thank David J.D. Earn, David Fisman, Joseph Tien, G. Brian Golding, Nicholas Waglechner, Debi Poinar, Melanie Kuch, D. Ann Herring, Kirsten Bos, and Johannes Krause, and the members of the McMaster Ancient DNA Centre for their ongoing insights and contributions to this work in the field of paleopathological research. We thank Rebecca Redfern and Jelena Bekvalac at the Museum of London Centre for Human Bioarchaeology for providing access to the 8291 specimen. AMD was supported by an Ontario Graduate Scholarship. HNP was supported by a CRC (Canada Research Chair) and NSERC grants. TMP was funded by the Government of Canada through Genome Canada and the Ontario Genomics Institute through the Biomonitoring 2.0 project (OGI-050).

4.7.1 Author Contributions

HNP, MB, CJ conceived of the research. AMD, CJ, JT, JME designed and performed experiments. AMD, CJ, SG, TMP, JME, JT, JA, KM and HNP analysed data. AND and SND provided samples. All authors contributed to the manuscript preparation.

4.7.2 Additional Information

Accession Codes: Sequencing data has been deposited to the National Center for Biotechnology Information under BioProjects PRJNA239173 (sample 3090.13) and PRJNA239147 (sample 8291). Supplementary information accompanies this paper at http://www.nature.com/scientificreports. Competing financial interests: The authors declare no competing financial interests.

4.8 Tables

Table 1. Summary of LLMDA and HTS results

Only taxa with probes designed on the LLMDA array are shown (see Table S1 for full results). Only taxa with at least 5 reads are called with HTS-MEGAN4 analysis. Reads = number of HTS reads assigned to that taxonomic level (- = not found in HTS dataset). LO score = LLMDA log odds score (- = not called with LLMDA). Phyla abbreviations = Act, Actinobacteria; Bac, Bacteroidetes; Chla, Chlamydiae; Chlo, Chlorobi; Chl, Chloroflexi; Fib, Fibrobacteres; Fir, Firmicutes; Fus, Fusobacteria; Pro, Proteobacteria; Spi, Spirochaetes; Syn, Synergistetes; Ten, Tenericutes; The, Thermotogae; Ver, Verrucomicrobia.

Cholera victim specimen 3090.13				Plague victim specimen 8291			
Phylum	**Family**	**Reads**	**LO score**	**Phylum**	**Family**	**Reads**	**LO score**
Pro	Vibrionaceae	10,600	4,470.7	Pro	Enterobacteriaceae	15,062	1,640.8
Pro	Aeromonadaceae	1,877	480.0	Pro	Alcaligenaceae	11,976	880.0
Pro	Enterobacteriaceae	1,072	4,944.3	Pro	Bradyrhizobiaceae	8,189	174.4
Fir	Erysipelotrichaceae	1,039	561.7	Pro	Burkholderiaceae	7,298	10,155.0
Fir	Clostridiaceae	989	2,023.6	Fir	Clostridiaceae	5,188	1,861.8
Fir	Streptococcaceae	387	486.6	Act	Pseudonocardiaceae	4,876	474.1
Pro	Comamonadaceae	233	496.6	Pro	Comamonadaceae	3,704	466.2
Fir	Peptostreptococcaceae	216	-	Pro	Pseudomonadaceae	2,778	3,461.5
Pro	Pseudomonadaceae	178	4,313.1	Pro	Xanthomonadaceae	2,720	197.3
Pro	Moraxellaceae	122	105.2	Act	Streptomycetaceae	2,135	506.1
Pro	Xanthomonadaceae	93	228.0	Pro	Methylobacteriaceae	1,195	118.3
Pro	Burkholderiaceae	22	11,233.8	Pro	Oxalobacteraceae	1,045	119.6
Fir	Veillonellaceae	22	130.2	Pro	Neisseriaceae	903	232.0
Act	Corynebacteriaceae	19	309.5	Pro	Sphingomonadaceae	747	-
Fir	Staphylococcaceae	14	273.6	Act	Mycobacteriaceae	642	1,368.6
Pro	Pasteurellaceae	11	-	Pro	Caulobacteraceae	606	106.0
Act	Micrococcaceae	8	358.5	Pro	Acetobacteraceae	492	222.6
Pro	Neisseriaceae	8	-	Fir	Peptostreptococcaceae	324	-
Fir	Enterococcaceae	6	204.0	Act	Nocardiaceae	310	282.7
Bac	Flavobacteriaceae	6	-	Pro	Brucellaceae	274	-
Fir	Bacillaceae	5	3,077.2	Pro	Halomonadaceae	204	-
Act	Streptomycetaceae	5	523.1	Pro	Aeromonadaceae	167	-
Act	Coriobacteriaceae	5	123.6	Pro	Desulfovibrionaceae	158	218.4
Fus	Fusobacteriaceae	5	-	Fir	Lachnospiraceae	131	707.8
Fir	Paenibacillaceae	-	1,100.2	Fir	Eubacteriaceae	122	74.3
Fir	Lachnospiraceae	-	1,016.1	Act	Micrococcaceae	111	349.9
Act	Propionibacteriaceae	-	947.8	Fus	Fusobacteriaceae	99	-
Pro	Alcaligenaceae	-	745.0	Fir	Peptococcaceae	97	116.1
Fir	Lactobacillaceae	-	677.9	Act	Propionibacteriaceae	95	950.6
Pro	Desulfovibrionaceae	-	390.9	Act	Cellulomonadaceae	92	-
Act	Actinomycetaceae	-	231.2	Pro	Sutterellaceae	85	-
Act	Bifidobacteriaceae	-	225.6	Act	Gordoniaceae	84	-
Act	Micrococcineae	-	213.2	Pro	Piscirickettsiaceae	82	-
Fir	Carnobacteriaceae	-	207.6	Fir	Streptococcaceae	79	104.2
Act	Mycobacteriaceae	-	185.0	Act	Coriobacteriaceae	77	112.0
Fir	Listeriaceae	-	164.0	Act	Actinomycetaceae	74	216.8
Fir	Planococcaceae	-	157.1	Pro	Cardiobacteriaceae	70	-

Fir	Aerococcaceae	-	135.2	Fir	Lactobacillaceae	66	378.8
Pro	Deferribacteraceae	-	128.3	Fir	Veillonellaceae	65	228.7
Fir	Peptococcaceae	-	127.7	Fir	Bacillaceae	63	2,764.1
Ver	Verrucomicrobiaceae	-	127.4	Pro	Moraxellaceae	62	203.2
Act	Jonesiaceae	-	126.6	Act	Corynebacteriaceae	54	562.3
Pro	Helicobacteraceae	-	124.7	Act	Intrasporangiaceae	53	-
Pro	Caulobacteraceae	-	117.6	Act	Bifidobacteriaceae	52	748.5
Chl	Herpetosiphonaceae	-	112.8	Spi	Spirochaetaceae	52	-
Act	Brevibacteriaceae	-	112.7	Pro	Erythrobacteraceae	44	-
Act	Dermabacteraceae	-	111.3	Fir	Staphylococcaceae	43	176.7
Fir	Leuconostocaceae	-	111.0	Syn	Synergistaceae	40	108.5
Pro	Campylobacteraceae	-	107.9	Fir	Ruminococcaceae	30	100.2
Fir	Eubacteriaceae	-	95.6	Pro	Pasteurellaceae	30	80.1
Fib	Fibrobacteraceae	-	90.5	Bac	Flavobacteriaceae	25	108.9
				Pro	Vibrionaceae	24	-
				Act	Dermabacteraceae	21	108.1
				Act	Dermacoccaceae	21	-
				Act	Segniliparaceae	21	-
				Act	Tsukamurellaceae	21	-
				Spi	Leptospiraceae	20	-
				Bac	Rikenellaceae	19	-
				Fir	Erysipelotrichaceae	17	322.5
				Bac	Bacteroidaceae	17	-
				Pro	Campylobacteraceae	17	-
				Pro	Succinivibrionaceae	17	-
				The	Thermotogaceae	16	-
				Pro	Desulfomicrobiaceae	15	-
				Bac	Prevotellaceae	14	-
				Pro	Helicobacteraceae	11	120.6
				Pro	Bartonellaceae	11	-
				Ten	Mycoplasmataceae	10	-
				Fir	Leuconostocaceae	9	191.4
				Fir	Enterococcaceae	9	177.6
				Fir	Listeriaceae	9	164.6
				Bac	Porphyromonadaceae	9	-
				Pro	Legionellaceae	9	-
				Fir	Aerococcaceae	8	257.5
				Pro	Anaplasmataceae	7	-
				Pro	Coxiellaceae	7	-
				Fir	Carnobacteriaceae	6	184.4
				Pro	Bdellovibrionaceae	6	85.0
				Chla	Parachlamydiaceae	6	-
				Pro	Francisellaceae	6	-
				Pro	Rickettsiaceae	5	-
				Act	Jonesiaceae	-	108.3
				Chlo	Herpetosiphonaceae	-	101.2
				Act	Brevibacteriaceae	-	101.1
				Fib	Fibrobacteraceae	-	79.3

4.9 Figures

Figure 1. Number of bacterial families detected by HTS and/or LLMDA

Number of bacterial families (or less-specific higher taxonomic level) detected by HTS sequencing (green circles) and LLMDA analyses (blue circles). Families detected by both methods are indicated where the circles overlap. Values above the midline include all detected families, whereas values below the midline include only those families used for LLMDA probe design. Image of cholera victim specimen 3090.13 was taken by AMD; image of plague victim specimen 8291 was taken by HNP.

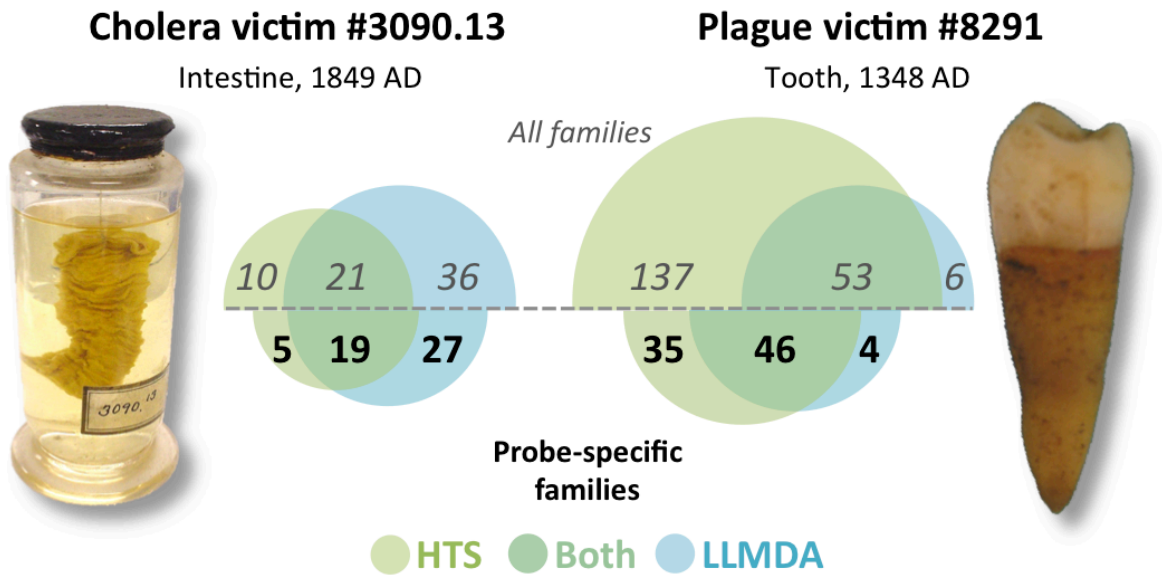


Figure 2. Comparison of HTS and LLMDA results for cholera victim 3090.13.

Cladogram based on NCBI Genbank taxonomy indicating results of the BLASTN/MEGAN4 HTS analysis at the family level and above compared to LLMDA results. At the leaves, circle size reflects the relative number of reads assigned to those taxa (internal node sizes only indicated if >10 reads). Colors of taxon names indicate whether that taxon had (1) reads present in the HTS data, (2) probes designed for that family on the LLMDA, and (3) LLMDA call for that taxon. Bacterial phyla and major clades are highlighted.

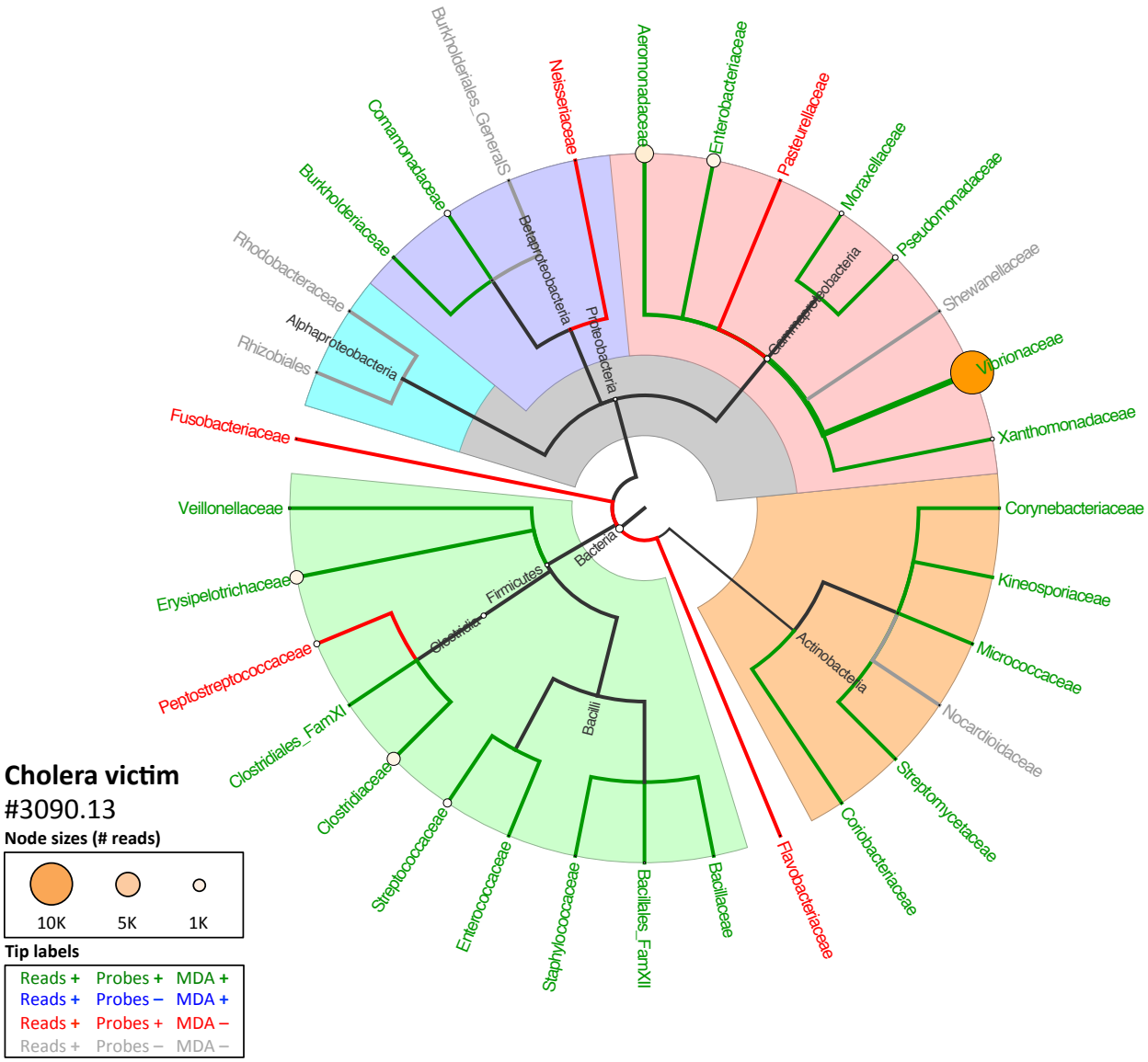


Figure 3. Comparison of HTS and LLMDA results for plague victim 8291.

See caption for Figure 2.

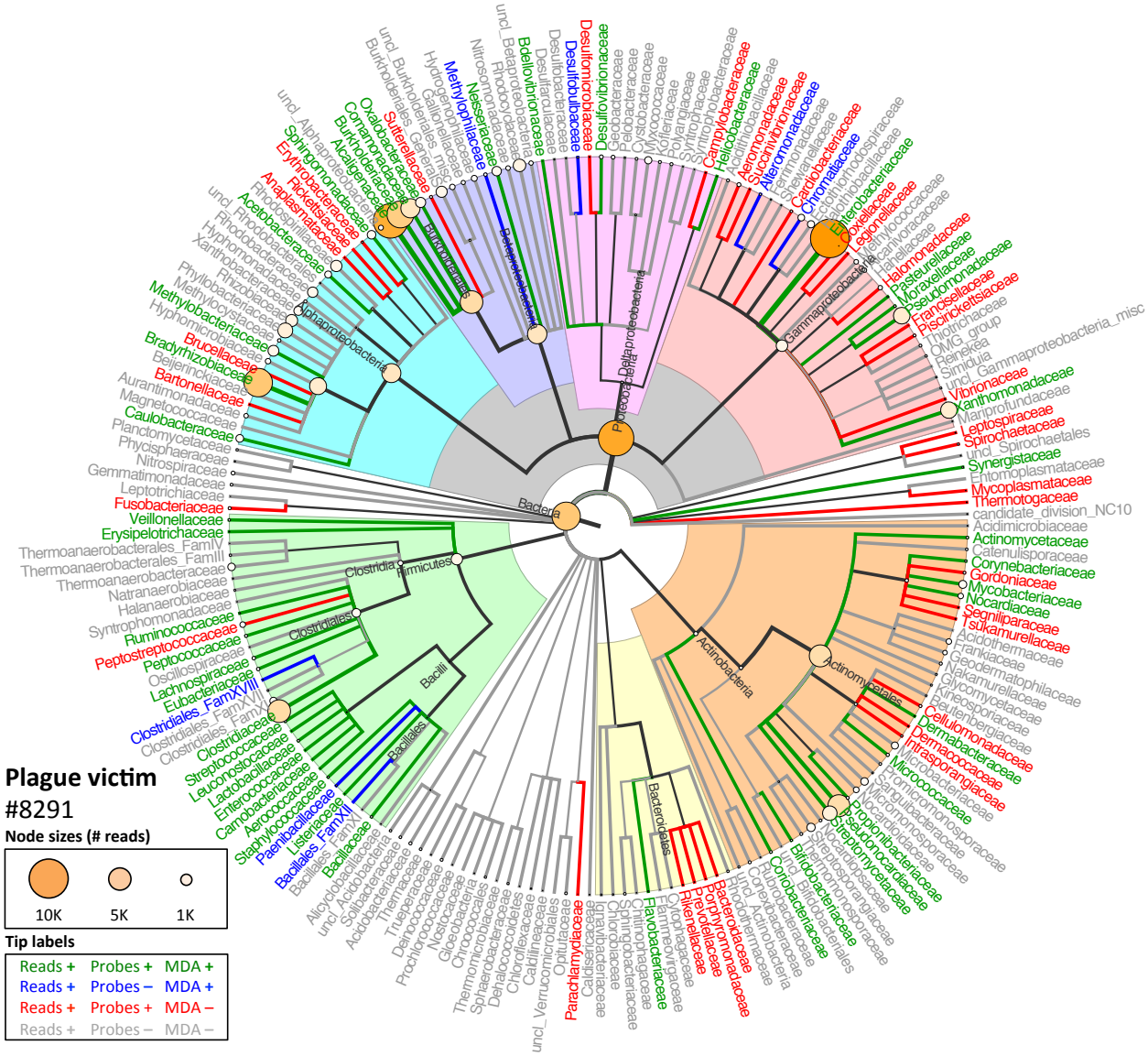


Figure 4. HTS vs. LLMDA comparisons

HTS readcounts, GC content, unique genomic positions sequenced, and maximum log odds scores for both specimens plotted against whether they were detected (+) or not detected (-) with LLMDA (a-c) or HTS (d). For HTS read counts, all HTS-identified families are analyzed (a); GC content and unique genomic positions are analyzed only for families that were used for LLMDA probe design (b,c); log odds scores are only analyzed for families detected with LLMDA.

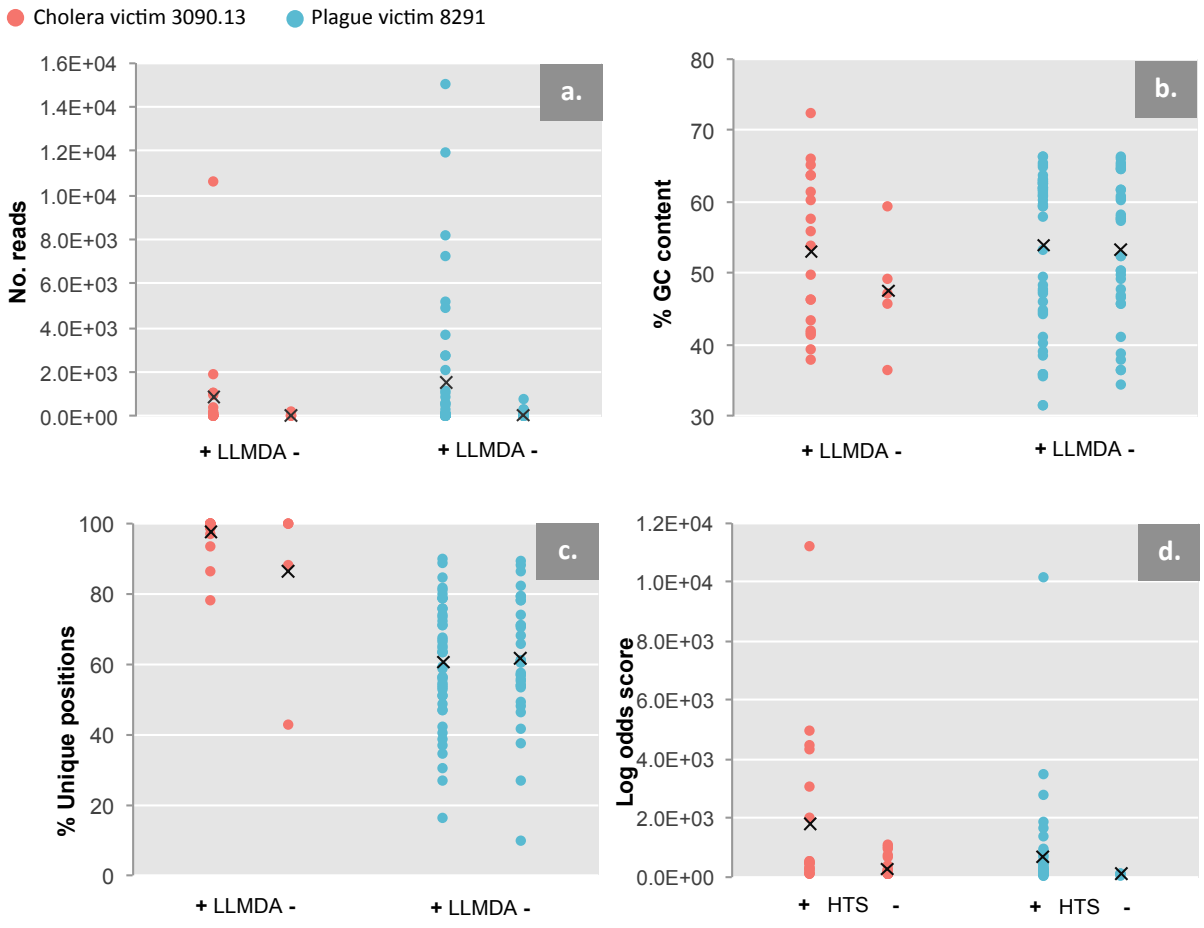
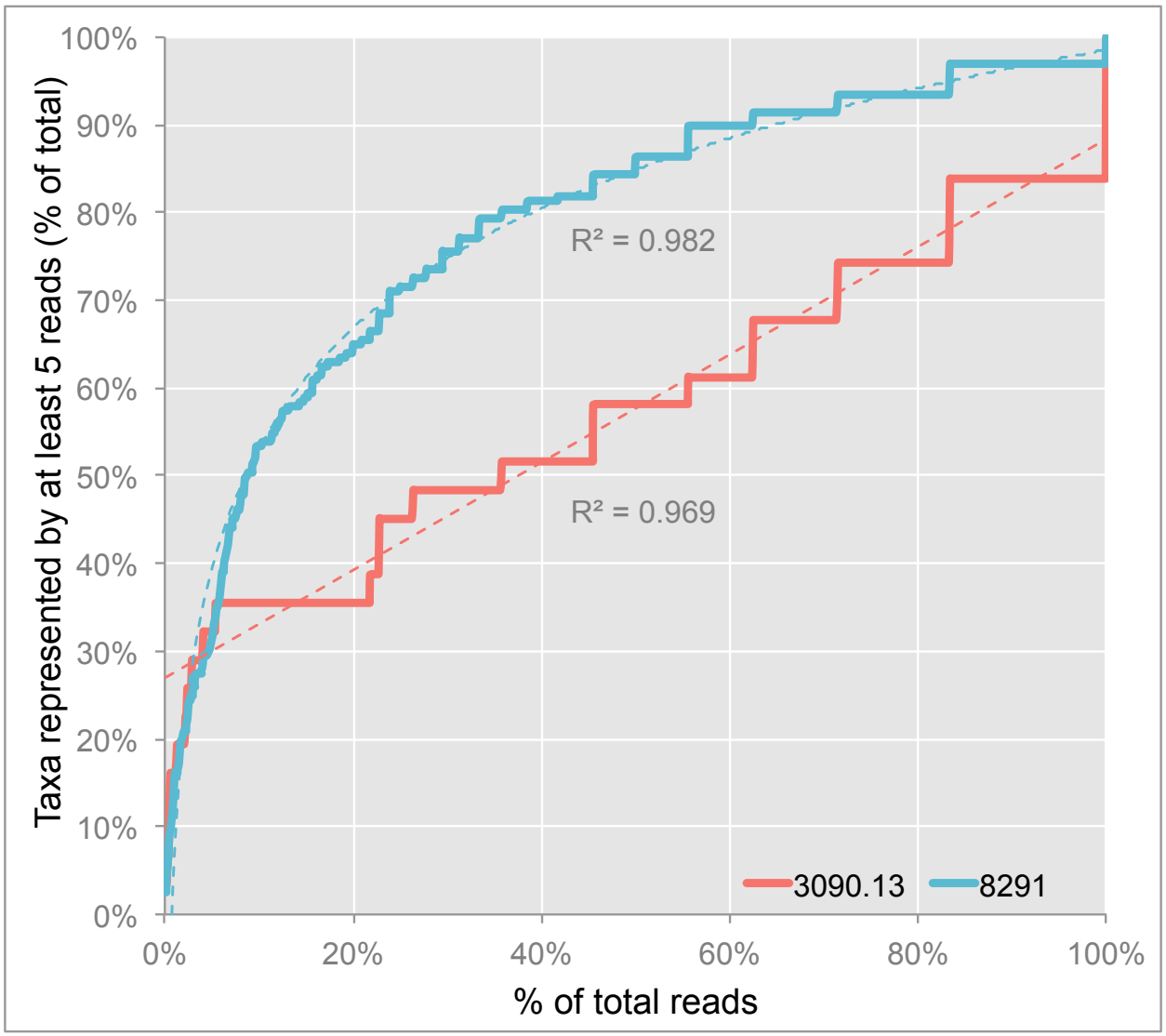


Figure 5. HTS rarefaction analysis

Rarefaction curves showing the number of bacterial families represented by at least 5 reads as a percent of the total observed families per sample with increasing read depth (0.1% increments). Dashed lines represent lines of best-fit; cholera victim specimen 3090.13 is a linear curve (R^2 = 0.96936), plague victim specimen 8291 is a logarithmic curve (R^2 = 0.98217).



4.10 Supplementary Appendix

The following sections contain the Supplementary material published online at nature.com/scientificreports.

4.11 Supplementary Methods

4.11.1 Sample preparation

Sample 3090.13 is a preserved intestinal specimen from a victim of the 1849 Philadelphia cholera epidemic, sealed in a glass jar with alcohol, and stored in the collections of the Mütter Museum (Philadelphia, PA, USA)(14).This specimen was sub-sampled, extracted, and libraries suitable for sequencing on the Illumina platform were prepared as described inreference (14). Specimen 8291 is a tooth from a victim of the Black Death buried at the East Smithfield cemetery in London in 1348-1349(15). This specimen was sampled and extracted using the same methods as described in reference(15). Libraries suitable for sequencing on the Illumina platform were prepared just as for 3090.13 (above), as described inreference (14).

4.11.2 Shotgun HTS sequencing

Prior to sequencing, additional indexing amplification was performed in 8 reactions each sample (5 µl 0.1x diluted template DNA in 50 µl total reaction volume) of indexed library, using 400nM each indexing primer, and 11 cycles for 3090.13 and 20 cycles for 8291. The purified libraries were pooled in equimolar ratio on one lane of Illumina HiSeq 1000. Sequencing was performed by the Farncombe Family Digestive Health Research Institute (McMaster University). 100bp paired-end read chemistry was used, with one indexing read. The lane yielded 141,039,627 reads each direction from 3090.13 and 122,830,910 reads each direction from 8291.

4.11.3 Pathogen HTS assemblies

Raw R1 reads from each sample were trimmed to remove residual adaptor sequence using cutadapt (v.1)(30) with the parameters: error rate (0.16), minimum overlap (1). Reads <28bp were removed from a 24,000,000 subset of each sample, leaving 12,946,441 for 3090.13 and 12,076,222 for 8291. To calculate HTS pathogen percentages, remaining reads were aligned using *bowtie* v.0.12.7 (23) with default settings to the O395 strain *V. cholerae* reference genome (NC_009456, NC_009457) for sample 3090.13 and to the CO92 strain *Y. pestis* reference genome and 3 plasmids pCD1, pPCP1, and pMT1 (NC_003143, NC_003131,

NC_003132, NC_003134) for sample 8291. For 3090.13, 6,938 aligned (0.054% of reads ≥28bp, 0.029% of total reads), and for 8291, 18,931 aligned (0.157% of reads ≥28bp, 0.079% of total reads).

4.11.4 HTS BLAST & MEGAN metagenomic analysis

Raw reads from each sample were trimmed using cutadapt (v.1.1) with the parameters -b (13bp adaptor sequence), –e (errors allowed) 0, – m (minimum length, bp) 20, -q (Phred scaled quality cutoff) 20, and -O (overlap, bp) 13, leaving 118,859,751 reads from 3090.13 and 89,321,997reads from 8291 for further processing. Reads were subjected to local BLASTN-megablast analysis(24) (v.2.2.26+) using a local copy of the refseq_genomic database(26) (downloaded October 16, 2012), using the parameters: -task megablast, -word_size 28, -evalue 1e-10, -num_descriptions 100, -num_alignments 100. BLAST reports were parsed using MEGAN4 (v.4.70.4) using the default lowest common ancestor (LCA) parameters(25). Full results of this analysis can be found in Table S2.

4.11.5 LLMDA Analysis

LLMDA v5 design

All completely sequenced genomes or elements (chromosomes, mitochondria, plasmids) as of December 20, 2011 were obtained from public sources (NCBI, J. Craig Venter Institute, etc.). These included assembled draft and finished sequences for viruses, bacteria, archaea, fungi, and the subset of protozoa known to be human pathogens or their near neighbors. These were grouped by kingdom and family. LLMDAv5 was designed using substantially the same approach as previous versions,(11) namely, finding family-specific regions in the available complete sequences, and selecting probes within those regions such that all targets are represented by both conserved and discriminating probes. The LLMDAv5 135K design has approximately 135,000 unique target probes. Conserved probes were selected favoring the most within-family conserved, thermodynamically optimal probes, so that all targets were represented by at least 15 conserved probes. Discriminating probes were selected favoring the least conserved probes for each sequence, with at least 2 per genome or sequence element. On the 135K design, only probes from families containing at least one species known to infect vertebrates were included for the viruses, bacteria, and fungi. All archaea families were included since there were few enough probes to include them all, as well as all the pathogenic protozoa previously selected for probe design. Vertebrate infecting bacterial, viral, and fungal families were selected based on literature (PubMed) and web searches to determine whether any members of a family have been found to infect vertebrates or were involved in

clinical infections, and all members of a family were included even if only some of them were vertebrate-infecting. The array also included several thousand negative control probes with random sequences designed to match the length and GC% distribution of the target probes. The following numbers of species were represented: 3,521 microbial species total, including 1,856 viral species, 1,398 bacterial species, 125 archaeal species, 94 protozoan species, and 48 fungal species.

LLMDA analyses

LLMDA arrays were analyzed using the CLiMax (Composite Likelihood Maximization) algorithm, described in detail previously(11), followed by some additional processing steps. We measured probe intensities on each array using NimbleScan software (Roche NimbleGen) and reduced them to vectors of binary probe detection indicators, by comparing each target probe intensity to the 95th percentile of the negative control probe intensities. The CLiMax software processes this indicator data using a greedy iterative procedure to predict a series of targets likely to be present in the sample. In the first iteration, a target is selected by computing, for each genome in a reference target database, the log-odds of the observed probe detection data if that genome were present in the sample; the target with the highest log-odds score becomes the first element of the series. In each subsequent iteration, a conditional log-odds score is computed for each remaining target, representing the likelihood of the data if the target were added to the series, relative to the likelihood given the previously predicted targets. The target with the largest conditional log-odds score is then appended to the series. Iterations continue until there are no additional targets with positive conditional log-odds scores, meaning that no further improvement in the likelihood can be obtained by predicting additional targets.

After the initial CLiMax analysis, we filtered the list of genomes predicted to be present by rejecting those for which the array detected only a small subset of the genome regions covered by probes. In our past experience, targets with this pattern of detected probes are likely to be false positives, resulting from cross-hybridization to a similar region in another genome. **Figure S5** shows examples of targets that were accepted and rejected under this filtering strategy. We aligned probes matching each selected target sequence to genome positions using BLAST. We used Gaussian kernel density estimates to approximate the positional distribution functions for all probes matching the target (with predicted detection probabilities greater than 0.85), and for the subset of these probes with intensities above the 95th percentile of negative controls, taking care to use the same bandwidth for both estimates. To quantify the difference between these two distributions, we computed the Kullback-Leibler divergence (D_{KL}) between the two density estimates. If $f_{pred}(x)$ and $f_{det}(x)$ are, respectively, the estimated density functions for the probes predicted to bind the target and the probes actually

detected, evaluated at discrete positions *x*, then the K-L divergence is computed as $D_{KL}(f_{pred} || f_{det}) = \sum_x f_{pred}(x) \log (f_{pred}(x) / f_{det}(x))$. Targets with $D_{KL} > 4\times10^{-4}$ were removed from the predicted set; this threshold was chosen by analysis of samples of known composition, to provide a reasonable compromise between sensitivity and specificity. The numbers of target sequences predicted to be present in sample 8291 were 398 total and 204 after filtering; for sample 3090.13 the target counts were 430 total and 217 after filtering.

Finally, to enable comparison of the LLMDA results with the family-level results produced by BLAST and MEGAN analysis of HTS data, we grouped the filtered targets by family, and summed log-odds scores over targets to produce an aggregate score for each family.

4.12 Supplementary Tables

Due to space constraints, two additional tables (**Table S2,** "HTS BLASTN/MEGAN metagenomic profiles, full results"and **Table S3,** "LLMDA probe data supporting *V. cholerae* and *Y. pestis* hits") areaccessible and freely available online as Supplementary Information Excel Files from:

[http://www.nature.com/srep/2014/140306/srep04245/full/srep04245.html#supplementary-information]

Table S1. LLMDA and HTS analysis, full results

S1a. Cholera victim 3090.13

				Shotgun sequencing results				LLMDA results				LLMDA summary
Phylum	Class	Order	Family	Read count	Mean frag len	%GC	Has reads	Log odds	Expected probes	Detected probes	KL divergence	Has probes
Actinobacteria	Actinobacteria	Actinomycetales	Streptomycetaceae	5	63.2	72.5	1	523.1	38	6	4.69E-05	1
Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	8	61.1	66.1	1	358.5	24	10	2.17E-05	1
Actinobacteria	Actinobacteria	Actinomycetales	Corynebacteriaceae	19	53.8	57.6	1	309.5	18	7	2.10E-04	1
Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	5	54.6	56	1	123.6	13	4	1.02E-04	1
Actinobacteria	Actinobacteria	Actinomycetales	Nocardioidaceae	5			1	0	0	0	0.00E+00	0
Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	6	49	45.9	1	0	0	0	0.00E+00	1
Firmicutes	Bacilli	Bacillales	Bacillaceae	5	49.4	46.2	1	3077.2	21	7	6.53E-05	1
Firmicutes	Bacilli	Bacillales	Staphylococcaceae	14	49.6	38	1	273.6	20	4	1.31E-04	1
Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	387	51.8	41.9	1	486.6	24	5	7.62E-05	1
Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	6	55.2	46.2	1	204	16	4	2.33E-05	1
Firmicutes	Clostridia	Clostridiales	Clostridiaceae	989	52.1	39.3	1	2023.6	16	6	7.35E-06	1
Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	216	52.9	49.3	1	0	0	0	0.00E+00	1
Firmicutes	Bacilli	Bacillales	Bacillales Family XII. Incertae Sedis	23			1	320.8	6	5	3.45E-07	0
Firmicutes	Erysipelotrichi	Erysipelotrichales	Erysipelotrichaceae	1039	50.1	41.7	1	561.7	11	6	1.11E-04	1
Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	22	55.8	43.5	1	130.2	14	6	3.30E-05	1
Firmicutes	Clostridia	Clostridiales	Clostridiales Family XI. Incertae Sedis	7	64.3	47.8	1	109.3	19	4	7.00E-05	0
Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	5	46	36.4	1	0	0	0	0.00E+00	1
Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	22	52.5	63.8	1	11233.8	90	65	3.99E-05	1
Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	233	54.5	63.7	1	496.6	20	8	9.14E-05	1
Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	11			1	0	0	0	0.00E+00	0
Proteobacteria	Alphaproteobacteria	Rhizobiales	unclassified Rhizobiales	9			1	0	0	0	0.00E+00	0
Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	8	51.6	59.3	1	0	0	0	0.00E+00	1
Proteobacteria	Gammaproteobacteria	Aeromonadales	Aeromonadaceae	1877	54.7	60.3	1	480	21	11	1.55E-05	1
Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiales Genera incertae sedis	7			1	0	0	0	0.00E+00	0
Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	1072	55.1	54	1	4944.3	304	149	8.67E-07	1
Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	11	54.7	47.3	1	0	0	0	0.00E+00	1
Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	6			1	0	0	0	0.00E+00	0
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	178	56.2	61.5	1	4313.1	36	22	6.13E-06	1
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	122	53.2	41.3	1	105.2	63	17	1.50E-04	1
Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	10600	51.8	49.8	1	4470.7	155	114	2.17E-06	1
Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	93	55.7	65.3	1	228	19	5	2.60E-04	1
Fibrobacteres	Fibrobacteria	Fibrobacterales	Fibrobacteraceae	0			0	90.5	26	6	5.12E-05	1
Firmicutes	Clostridia	Clostridiales	Eubacteriaceae	0			0	95.6	25	6	1.33E-05	1

Proteobacteria	Epsilonproteobacteria	Campylobacterales	Campylobacteraceae	0	0	107.9	25	5	2.77E-04	1
Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	0	0	111	17	4	3.43E-04	1
Actinobacteria	Actinobacteria	Actinomycetales	Dermabacteraceae	0	0	111.3	17	4	1.01E-04	1
Actinobacteria	Actinobacteria	Actinomycetales	Brevibacteriaceae	0	0	112.7	20	5	4.69E-05	1
Chloroflexi	Chloroflexi	Herpetosiphonales	Herpetosiphonaceae	0	0	112.8	17	5	6.74E-06	1
Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	0	0	117.6	17	6	3.19E-04	1
Proteobacteria	Epsilonproteobacteria	Campylobacterales	Helicobacteraceae	0	0	124.7	16	5	2.80E-04	1
Actinobacteria	Actinobacteria	Actinomycetales	Jonesiaceae	0	0	126.6	16	6	7.04E-05	1
Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Verrucomicrobiaceae	0	0	127.4	12	4	5.75E-05	1
Firmicutes	Clostridia	Clostridiales	Peptococcaceae	0	0	127.7	13	4	1.18E-06	1
Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Deferribacteraceae	0	0	128.3	14	4	1.50E-04	1
Firmicutes	Bacilli	Bacillales	Planococcaceae	0	0	157.1	7	5	-6.54E-08	1
Firmicutes	Bacilli	Bacillales	Listeriaceae	0	0	164	29	5	3.67E-05	1
Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	0	0	207.6	17	4	3.50E-05	1
Firmicutes	Bacilli	Lactobacillales	Aerococcaceae	0	0	135.2	11	4	7.06E-05	1
Actinobacteria	Actinobacteria	Actinomycetales	Mycobacteriaceae	0	0	185	25	5	2.77E-04	1
Actinobacteria	Actinobacteria	Actinomycetales	Micrococcineae	0	0	213.2	20	5	2.28E-04	1
Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	0	0	225.6	26	8	8.39E-05	1
Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	0	0	231.2	22	7	5.61E-05	1
Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	0	0	390.9	15	6	5.08E-05	1
Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	0	0	677.9	28	7	3.80E-04	1
Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	0	0	745	18	6	3.69E-05	1
Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	0	0	947.8	31	10	1.01E-04	1
Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	0	0	1016.1	25	8	2.21E-04	1
Firmicutes	Bacilli	Bacillales	Paenibacillaceae	0	0	1100.2	6	5	-5.19E-09	1
Firmicutes	Clostridia	Clostridiales	Clostridiales Family XVIII. Incertae Sedis	0	0	108.6	18	4	1.06E-04	0
Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	0	0	111.8	18	5	1.22E-04	0
Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae	0	0	131.1	13	5	5.07E-05	0
Proteobacteria	Deltaproteobacteria	Desulfobacterales	Desulfobulbaceae	0	0	152.7	5	4	1.50E-05	0
Proteobacteria	Gammaproteobacteria	Chromatiales	Chromatiaceae	0	0	152.9	5	5	2.92E-05	0
Firmicutes	Clostridia	Clostridiales	Heliobacteriaceae	0	0	161.4	6	5	5.61E-08	0
Proteobacteria	Betaproteobacteria	Methylophilales	Methylophilaceae	0	0	172.4	8	7	7.15E-05	0
Actinobacteria	Actinobacteria	Actinomycetales	Pseudonocardiaceae	0	0	190.1	13	4	3.01E-04	0
Synergistetes	Synergistia	Synergistales	Synergistaceae	0	0	320.1	10	5	1.26E-04	0

S1b. Plague victim #8291

				Shotgun sequencing results					LLMDA results			LLMDA summary
Phylum	Class	Order	Family	Read count	Mean frag len	%GC	Has reads	Log odds	Expected probes	Detected probes	KL divergence	Has probes
Acidobacteria	Acidobacteriia	Acidobacteriales	Acidobacteriaceae	104			1	0	0	0	0.00E+00	0
Acidobacteria	Solibacteres	Solibacterales	Solibacteraceae	74			1	0	0	0	0.00E+00	0
Actinobacteria	Actinobacteria	Actinomycetales	Mycobacteriaceae	642	56.9	65.1	1	1368.6	28	9	2.53E-05	1
Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	95	56.1	61.9	1	950.6	31	10	1.03E-04	1
Actinobacteria	Actinobacteria	Actinomycetales	Corynebacteriaceae	54	54.9	60.9	1	562.3	12	4	9.44E-05	1
Actinobacteria	Actinobacteria	Actinomycetales	Streptomycetaceae	2135	57.6	66.2	1	506.1	17	5	1.57E-05	1
Actinobacteria	Actinobacteria	Actinomycetales	Pseudonocardiaceae	4876	56.7	66.2	1	474.1	13	6	2.77E-04	1
Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	111	56.8	63.2	1	349.9	24	10	2.65E-05	1
Actinobacteria	Actinobacteria	Actinomycetales	Nocardiaceae	310	58.1	65.4	1	282.7	21	6	1.42E-04	1
Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	74	58.4	63.8	1	216.8	22	7	5.61E-05	1
Actinobacteria	Actinobacteria	Actinomycetales	Jonesiaceae	0	NA		0	108.3	16	5	7.04E-05	1
Actinobacteria	Actinobacteria	Actinomycetales	Dermabacteraceae	21	53.8	65	1	108.1	17	5	1.01E-04	1
Actinobacteria	Actinobacteria	Actinomycetales	Brevibacteriaceae	0	NA		0	101.1	20	5	4.00E-05	1
Actinobacteria	Actinobacteria	Actinomycetales	Streptosporangiaceae	710			1	0	0	0	0.00E+00	0
Actinobacteria	Actinobacteria	Actinomycetales	Micromonosporaceae	577			1	0	0	0	0.00E+00	0
Actinobacteria	Actinobacteria	Actinomycetales	Microbacteriaceae	510			1	0	0	0	0.00E+00	0
Actinobacteria	Actinobacteria	Actinomycetales	Nocardioidaceae	339			1	0	0	0	0.00E+00	0
Actinobacteria	Actinobacteria	Actinomycetales	Frankiaceae	314			1	0	0	0	0.00E+00	0
Actinobacteria	Actinobacteria	Actinomycetales	Geodermatophilaceae	215			1	0	0	0	0.00E+00	0
Actinobacteria	Actinobacteria	Actinomycetales	Nakamurellaceae	98			1	0	0	0	0.00E+00	0
Actinobacteria	Actinobacteria	Actinomycetales	Cellulomonadaceae	92	57	66.2	1	0	0	0	0.00E+00	1
Actinobacteria	Actinobacteria	Actinomycetales	Thermomonosporaceae	87			1	0	0	0	0.00E+00	0
Actinobacteria	Actinobacteria	Actinomycetales	Gordoniaceae	84	53.4	65	1	0	0	0	0.00E+00	1
Actinobacteria	Actinobacteria	Actinomycetales	Nocardiopsaceae	80			1	0	0	0	0.00E+00	0
Actinobacteria	Actinobacteria	Actinomycetales	Promicromonosporaceae	76			1	0	0	0	0.00E+00	0
Actinobacteria	Actinobacteria	Actinomycetales	Catenulisporaceae	73			1	0	0	0	0.00E+00	0
Actinobacteria	Actinobacteria	Actinomycetales	Intrasporangiaceae	53	57.3	64.5	1	0	0	0	0.00E+00	1
Actinobacteria	Actinobacteria	Actinomycetales	Kineosporiaceae	42			1	0	0	0	0.00E+00	0
Actinobacteria	Actinobacteria	Actinomycetales	Glycomycetaceae	41			1	0	0	0	0.00E+00	0
Actinobacteria	Actinobacteria	Actinomycetales	Sanguibacteraceae	32			1	0	0	0	0.00E+00	0
Actinobacteria	Actinobacteria	Actinomycetales	Beutenbergiaceae	31			1	0	0	0	0.00E+00	0
Actinobacteria	Actinobacteria	Actinomycetales	Dermacoccaceae	21	59.4	64.7	1	0	0	0	0.00E+00	1
Actinobacteria	Actinobacteria	Actinomycetales	Segniliparaceae	21	59.1	65.6	1	0	0	0	0.00E+00	1
Actinobacteria	Actinobacteria	Actinomycetales	Tsukamurellaceae	21	53.7	65.2	1	0	0	0	0.00E+00	1
Actinobacteria	Actinobacteria	Actinomycetales	Acidothermaceae	6			1	0	0	0	0.00E+00	0
Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	52	53.3	63.1	1	748.5	28	12	3.08E-04	1
Actinobacteria	Actinobacteria	Bifidobacteriales	unclassified Bifidobacteriales	7			1	0	0	0	0.00E+00	0

Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	77	51.4	60.8	1	112	13	4	1.02E-04	1
Actinobacteria	Actinobacteria	Rubrobacterales	Rubrobacteraceae	15			1	0	0	0	0.00E+00	0
Actinobacteria	Actinobacteria	Solirubrobacterales	Conexibacteraceae	93			1	0	0	0	0.00E+00	0
Actinobacteria	Actinobacteria	unclassified Actinobacteria	unclassified Actinobacteria	49			1	0	0	0	0.00E+00	0
Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	19	53.9	60.3	1	0	0	0	0.00E+00	1
Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	17	49.9	50.4	1	0	0	0	0.00E+00	1
Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	14	54.5	46.7	1	0	0	0	0.00E+00	1
Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	9	52.8	49.7	1	0	0	0	0.00E+00	1
Bacteroidetes	Cytophagia	Cytophagales	Cytophagaceae	22			1	0	0	0	0.00E+00	0
Bacteroidetes	Cytophagia	Cytophagales	Flammeovirgaceae	6			1	0	0	0	0.00E+00	0
Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	25	57.4	38.4	1	108.9	17	5	6.67E-05	1
Bacteroidetes	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae	70			1	0	0	0	0.00E+00	0
Bacteroidetes	Sphingobacteriia	Sphingobacteriales	Chitinophagaceae	13			1	0	0	0	0.00E+00	0
Bacteroidetes		Bacteroidetes Order II. Incertae sedis	Rhodothermaceae	62			1	0	0	0	0.00E+00	0
Caldiserica	Caldisericia	Caldisericales	Caldisericaceae	14			1	0	0	0	0.00E+00	0
Chlamydiae	Chlamydiia	Chlamydiales	Parachlamydiaceae	6	54.3	45.7	1	0	0	0	0.00E+00	1
Chlorobi	Chlorobia	Chlorobiales	Chlorobiaceae	27			1	0	0	0	0.00E+00	0
Chloroflexi	Caldilineae	Caldilineales	Caldilineaceae	9			1	0	0	0	0.00E+00	0
Chloroflexi	Chloroflexi	Chloroflexales	Chloroflexaceae	25			1	0	0	0	0.00E+00	0
Chloroflexi	Chloroflexi	Herpetosiphonales	Herpetosiphonaceae	0	NA		0	101.2	17	5	6.74E-06	1
Chloroflexi	Thermomicrobia	Sphaerobacterales	Sphaerobacteraceae	52			1	0	0	0	0.00E+00	0
Chloroflexi	Thermomicrobia	Thermomicrobiales	Thermomicrobiaceae	8			1	0	0	0	0.00E+00	0
Cyanobacteria		Chroococcales	unclassified Chroococcales	54			1	0	0	0	0.00E+00	0
Cyanobacteria		Nostocales	Nostocaceae	5			1	0	0	0	0.00E+00	0
Cyanobacteria		Prochlorales	Prochlorococcaceae	11			1	0	0	0	0.00E+00	0
DeinococcusThermus	Deinococci	Deinococcales	Deinococcaceae	86			1	0	0	0	0.00E+00	0
DeinococcusThermus	Deinococci	Deinococcales	Trueperaceae	15			1	0	0	0	0.00E+00	0
DeinococcusThermus	Deinococci	Thermales	Thermaceae	73			1	0	0	0	0.00E+00	0
Fibrobacteres	Fibrobacteria	Fibrobacterales	Fibrobacteraceae	0	NA		0	79.3	26	6	5.12E-05	1
Firmicutes	Bacilli	Bacillales	Bacillaceae	63	53.5	47.1	1	2764.1	21	10	7.39E-05	1
Firmicutes	Bacilli	Bacillales	Paenibacillaceae	34			1	1019.4	6	5	-5.48E-09	0
Firmicutes	Bacilli	Bacillales	Bacillales Family XII. Incertae Sedis	5			1	297.6	6	5	3.77E-07	0
Firmicutes	Bacilli	Bacillales	Planococcaceae	0			0	155.3	7	6	-8.31E-08	0
Firmicutes	Bacilli	Bacillales	Staphylococcaceae	43	62.1	44.8	1	176.7	20	4	1.31E-04	1
Firmicutes	Bacilli	Bacillales	Listeriaceae	9	54.2	31.6	1	164.6	29	6	5.07E-05	1
Firmicutes	Bacilli	Bacillales	Alicyclobacillaceae	11			1	0	0	0	0.00E+00	0
Firmicutes	Bacilli	Bacillales	Bacillales Family XI. Incertae Sedis	10			1	0	0	0	0.00E+00	0

Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	66	52.8	44.6	1	378.8	28	7	3.78E-04	1
Firmicutes	Bacilli	Lactobacillales	Aerococcaceae	8	55.4	35.9	1	257.5	14	7	1.90E-04	1
Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	79	50.8	35.6	1	104.2	21	4	2.36E-04	1
Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	9	49.2	48.3	1	191.4	17	4	3.43E-04	1
Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	6	61.5	31.7	1	184.4	17	4	3.50E-05	1
Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	9	50.2	39.2	1	177.6	16	4	2.33E-05	1
Firmicutes	Clostridia	Clostridiales	Clostridiaceae	5188	51.4	35.9	1	1861.8	16	6	7.37E-06	1
Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	131	52.3	40.2	1	707.8	25	8	2.21E-04	1
Firmicutes	Clostridia	Clostridiales	Heliobacteriaceae	0			0	149.8	6	5	5.61E-08	0
Firmicutes	Clostridia	Clostridiales	Peptococcaceae	97	52.4	49.4	1	116.1	13	4	5.62E-07	1
Firmicutes	Clostridia	Clostridiales	Clostridiales Family XVIII. Incertae Sedis	19	55.9	61	1	104.1	18	5	1.06E-04	0
Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	30	53.5	60.2	1	100.2	18	5	1.22E-04	1
Firmicutes	Clostridia	Clostridiales	Eubacteriaceae	122	50.3	47.8	1	74.3	25	5	1.21E-05	1
Firmicutes	Clostridia	Clostridiales	Clostridiales Family XI. Incertae Sedis	528	52.1	43.6	1	0	0	0	0.00E+00	0
Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	324	51.7	36.4	1	0	0	0	0.00E+00	1
Firmicutes	Clostridia	Clostridiales	Clostridiales Family XVII. Incertae Sedis	35	58.2	64.8	1	0	0	0	0.00E+00	0
Firmicutes	Clostridia	Clostridiales	Oscillospiraceae	11			1	0	0	0	0.00E+00	0
Firmicutes	Clostridia	Clostridiales	Syntrophomonadaceae	9			1	0	0	0	0.00E+00	0
Firmicutes	Clostridia	Halanaerobiales	Halanaerobiaceae	40			1	0	0	0	0.00E+00	0
Firmicutes	Clostridia	Natranaerobiales	Natranaerobiaceae	21			1	0	0	0	0.00E+00	0
Firmicutes	Clostridia	Thermoanaerobacterales	Thermoanaerobacteraceae	571			1	0	0	0	0.00E+00	0
Firmicutes	Clostridia	Thermoanaerobacterales	Thermoanaerobacterales Family III. Incertae Sedis	60			1	0	0	0	0.00E+00	0
Firmicutes	Clostridia	Thermoanaerobacterales	Thermoanaerobacterales Family IV. Incertae Sedis	8			1	0	0	0	0.00E+00	0
Firmicutes	Erysipelotrichi	Erysipelotrichales	Erysipelotrichaceae	17	54.9	44.4	1	322.5	11	5	1.06E-04	1
Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	65	55.6	61.1	1	228.7	14	7	3.30E-05	1
Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	99	58	34.4	1	0	0	0	0.00E+00	1
Fusobacteria	Fusobacteriia	Fusobacteriales	Leptotrichiaceae	7			1	0	0	0	0.00E+00	0
Gemmatimonadetes	Gemmatimonadetes	Gemmatimonadales	Gemmatimonadaceae	84			1	0	0	0	0.00E+00	0
Ignavibacteria	Ignavibacteria	Ignavibacteriales	Ignavibacteriaceae	32			1	0	0	0	0.00E+00	0
Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	157			1	0	0	0	0.00E+00	0
Planctomycetes	Phycisphaerae	Phycisphaerales	Phycisphaeraceae	23			1	0	0	0	0.00E+00	0
Planctomycetes	Planctomycetia	Planctomycetales	Planctomycetaceae	59			1	0	0	0	0.00E+00	0
Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	606	56.3	62.5	1	106	17	6	3.19E-04	1
Proteobacteria	Alphaproteobacteria	Magnetococcales	Magnetococcaceae	5			1	0	0	0	0.00E+00	0
Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	8189	58.9	61.6	1	174.4	17	5	4.62E-05	1
Proteobacteria	Alphaproteobacteria	Rhizobiales	Methylobacteriaceae	1195	57.5	62.7	1	118.3	19	7	1.93E-04	1
Proteobacteria	Alphaproteobacteria	Rhizobiales	Phyllobacteriaceae	2252			1	0	0	0	0.00E+00	0

Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	1468			1	0	0	0	0.00E+00	0
Proteobacteria	Alphaproteobacteria	Rhizobiales	Xanthobacteraceae	825			1	0	0	0	0.00E+00	0
Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	306			1	0	0	0	0.00E+00	0
Proteobacteria	Alphaproteobacteria	Rhizobiales	Brucellaceae	274	55.1	57.3	1	0	0	0	0.00E+00	1
Proteobacteria	Alphaproteobacteria	Rhizobiales	Aurantimonadaceae	207			1	0	0	0	0.00E+00	0
Proteobacteria	Alphaproteobacteria	Rhizobiales	Beijerinckiaceae	170			1	0	0	0	0.00E+00	0
Proteobacteria	Alphaproteobacteria	Rhizobiales	Methylocystaceae	123			1	0	0	0	0.00E+00	0
Proteobacteria	Alphaproteobacteria	Rhizobiales	Bartonellaceae	11	57.7	45.7	1	0	0	0	0.00E+00	1
Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	1089			1	0	0	0	0.00E+00	0
Proteobacteria	Alphaproteobacteria	Rhodobacterales	Hyphomonadaceae	105			1	0	0	0	0.00E+00	0
Proteobacteria	Alphaproteobacteria	Rhodobacterales	unclassified Rhodobacterales	22			1	0	0	0	0.00E+00	0
Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	492	56	62.9	1	222.6	13	5	8.53E-05	1
Proteobacteria	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	1301			1	0	0	0	0.00E+00	0
Proteobacteria	Alphaproteobacteria	Rickettsiales	Anaplasmataceae	7	57.7	52.5	1	0	0	0	0.00E+00	1
Proteobacteria	Alphaproteobacteria	Rickettsiales	Rickettsiaceae	5	50	41.2	1	0	0	0	0.00E+00	1
Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	747	55.2	61.7	1	0	0	0	0.00E+00	1
Proteobacteria	Alphaproteobacteria	Sphingomonadales	Erythrobacteraceae	44	54.1	60.6	1	0	0	0	0.00E+00	1
Proteobacteria	Alphaproteobacteria	unclassified Alphaproteobacteria	unclassified Alphaproteobacteria	252			1	0	0	0	0.00E+00	0
Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	7298	55.9	61.9	1	10155	90	63	3.65E-05	1
Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	11976	54.8	59.7	1	880	18	6	1.19E-05	1
Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	3704	54.8	62.6	1	466.2	20	10	6.91E-05	1
Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae	1045	54.9	59.3	1	119.6	13	5	5.07E-05	1
Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiales Genera incertae sedis	1265			1	0	0	0	0.00E+00	0
Proteobacteria	Betaproteobacteria	Burkholderiales	unclassified Burkholderiales (miscellaneous)	313			1	0	0	0	0.00E+00	0
Proteobacteria	Betaproteobacteria	Burkholderiales	Sutterellaceae	85	51.7	57.5	1	0	0	0	0.00E+00	1
Proteobacteria	Betaproteobacteria	Gallionellales	Gallionellaceae	182			1	0	0	0	0.00E+00	0
Proteobacteria	Betaproteobacteria	Hydrogenophilales	Hydrogenophilaceae	364			1	0	0	0	0.00E+00	0
Proteobacteria	Betaproteobacteria	Methylophilales	Methylophilaceae	124			1	160.8	8	7	8.96E-05	0
Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	903	55.5	61.7	1	232	12	4	1.80E-04	1
Proteobacteria	Betaproteobacteria	Nitrosomonadales	Nitrosomonadaceae	218			1	0	0	0	0.00E+00	0
Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	1604			1	0	0	0	0.00E+00	0
Proteobacteria	Betaproteobacteria	unclassified Betaproteobacteria	unclassified Betaproteobacteria	199			1	0	0	0	0.00E+00	0
Proteobacteria	Deltaproteobacteria	Bdellovibrionales	Bdellovibrionaceae	6	51.5	53.4	1	85	21	5	2.20E-04	1
Proteobacteria	Deltaproteobacteria	Desulfarculales	Desulfarculaceae	23			1	0	0	0	0.00E+00	0
Proteobacteria	Deltaproteobacteria	Desulfobacterales	Desulfobulbaceae	44			1	141.2	5	4	1.50E-05	0
Proteobacteria	Deltaproteobacteria	Desulfobacterales	Desulfobacteraceae	22			1	0	0	0	0.00E+00	0
Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	158	55	62.8	1	218.4	15	6	1.14E-04	1
Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfomicrobiaceae	15	50.8	66	1	0	0	0	0.00E+00	1
Proteobacteria	Deltaproteobacteria	Desulfuromonadales	Geobacteraceae	152			1	0	0	0	0.00E+00	0

Proteobacteria	Deltaproteobacteria	Desulfuromonadales	Pelobacteraceae	29			1	0	0	0	0.00E+00	0
Proteobacteria	Deltaproteobacteria	Myxococcales	Myxococcaceae	538			1	0	0	0	0.00E+00	0
Proteobacteria	Deltaproteobacteria	Myxococcales	Polyangiaceae	100			1	0	0	0	0.00E+00	0
Proteobacteria	Deltaproteobacteria	Myxococcales	Kofleriaceae	91			1	0	0	0	0.00E+00	0
Proteobacteria	Deltaproteobacteria	Myxococcales	Cystobacteraceae	80			1	0	0	0	0.00E+00	0
Proteobacteria	Deltaproteobacteria	Syntrophobacterales	Syntrophaceae	12			1	0	0	0	0.00E+00	0
Proteobacteria	Deltaproteobacteria	Syntrophobacterales	Syntrophobacteraceae	10			1	0	0	0	0.00E+00	0
Proteobacteria	Epsilonproteobacteria	Campylobacterales	Helicobacteraceae	11	58.6	46	1	120.6	16	6	2.80E-04	1
Proteobacteria	Epsilonproteobacteria	Campylobacterales	Campylobacteraceae	17	54.7	38.9	1	0	0	0	0.00E+00	1
Proteobacteria	Gammaproteobacteria	Acidithiobacillales	Acidithiobacillaceae	88			1	0	0	0	0.00E+00	0
Proteobacteria	Gammaproteobacteria	Aeromonadales	Aeromonadaceae	167	55.4	60.7	1	0	0	0	0.00E+00	1
Proteobacteria	Gammaproteobacteria	Aeromonadales	Succinivibrionaceae	17	49.6	47.9	1	0	0	0	0.00E+00	1
Proteobacteria	Gammaproteobacteria	Alteromonadales	Alteromonadaceae	59			1	144.2	5	4	-1.45E-06	0
Proteobacteria	Gammaproteobacteria	Alteromonadales	Ferrimonadaceae	22			1	0	0	0	0.00E+00	0
Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	18			1	0	0	0	0.00E+00	0
Proteobacteria	Gammaproteobacteria	Cardiobacteriales	Cardiobacteriaceae	70	54.5	57.6	1	0	0	0	0.00E+00	1
Proteobacteria	Gammaproteobacteria	Chromatiales	Chromatiaceae	465			1	141.3	5	5	7.46E-06	0
Proteobacteria	Gammaproteobacteria	Chromatiales	Ectothiorhodospiraceae	529			1	0	0	0	0.00E+00	0
Proteobacteria	Gammaproteobacteria	Chromatiales	Halothiobacillaceae	33			1	0	0	0	0.00E+00	0
Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	15060	48.6	47.5	1	1640.8	304	95	3.71E-06	1
Proteobacteria	Gammaproteobacteria	Legionellales	Legionellaceae	9	54.9	47	1	0	0	0	0.00E+00	1
Proteobacteria	Gammaproteobacteria	Legionellales	Coxiellaceae	7	59	52.3	1	0	0	0	0.00E+00	1
Proteobacteria	Gammaproteobacteria	Methylococcales	Methylococcaceae	218	55.9	58.9	1	0	0	0	0.00E+00	0
Proteobacteria	Gammaproteobacteria	Oceanospirillales	Halomonadaceae	204	55.2	61.8	1	0	0	0	0.00E+00	1
Proteobacteria	Gammaproteobacteria	Oceanospirillales	Alcanivoracaceae	204			1	0	0	0	0.00E+00	0
Proteobacteria	Gammaproteobacteria	Oceanospirillales	Hahellaceae	32			1	0	0	0	0.00E+00	0
Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	30	48.2	57.8	1	80.1	25	4	2.35E-04	1
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	2778	55.8	61.3	1	3461.5	36	20	8.16E-06	1
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	62	55.6	41.2	1	203.2	21	6	1.52E-04	1
Proteobacteria	Gammaproteobacteria	Thiotrichales	Thiotrichaceae	83			1	0	0	0	0.00E+00	0
Proteobacteria	Gammaproteobacteria	Thiotrichales	Piscirickettsiaceae	82	49.1	58.1	1	0	0	0	0.00E+00	1
Proteobacteria	Gammaproteobacteria	Thiotrichales	Francisellaceae	6	53.7	37.9	1	0	0	0	0.00E+00	1
Proteobacteria	Gammaproteobacteria	unclassified Gammaproteobacteria	OMG group	57			1	0	0	0	0.00E+00	0
Proteobacteria	Gammaproteobacteria	unclassified Gammaproteobacteria	unclassified Gammaproteobacteria (misc)	38			1	0	0	0	0.00E+00	0
Proteobacteria	Gammaproteobacteria	unclassified Gammaproteobacteria	Reinekea	6			1	0	0	0	0.00E+00	0
Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	24	51.4	49.3	1	0	0	0	0.00E+00	1
Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	2720	55.3	61.9	1	197.3	19	4	1.97E-04	1
Proteobacteria	Zetaproteobacteria	Mariprofundales	Mariprofundaceae	16			1	0	0	0	0.00E+00	0
Spirochaetes	Spirochaetia	Spirochaetales	Spirochaetaceae	52	52.9	50.4	1	0	0	0	0.00E+00	1

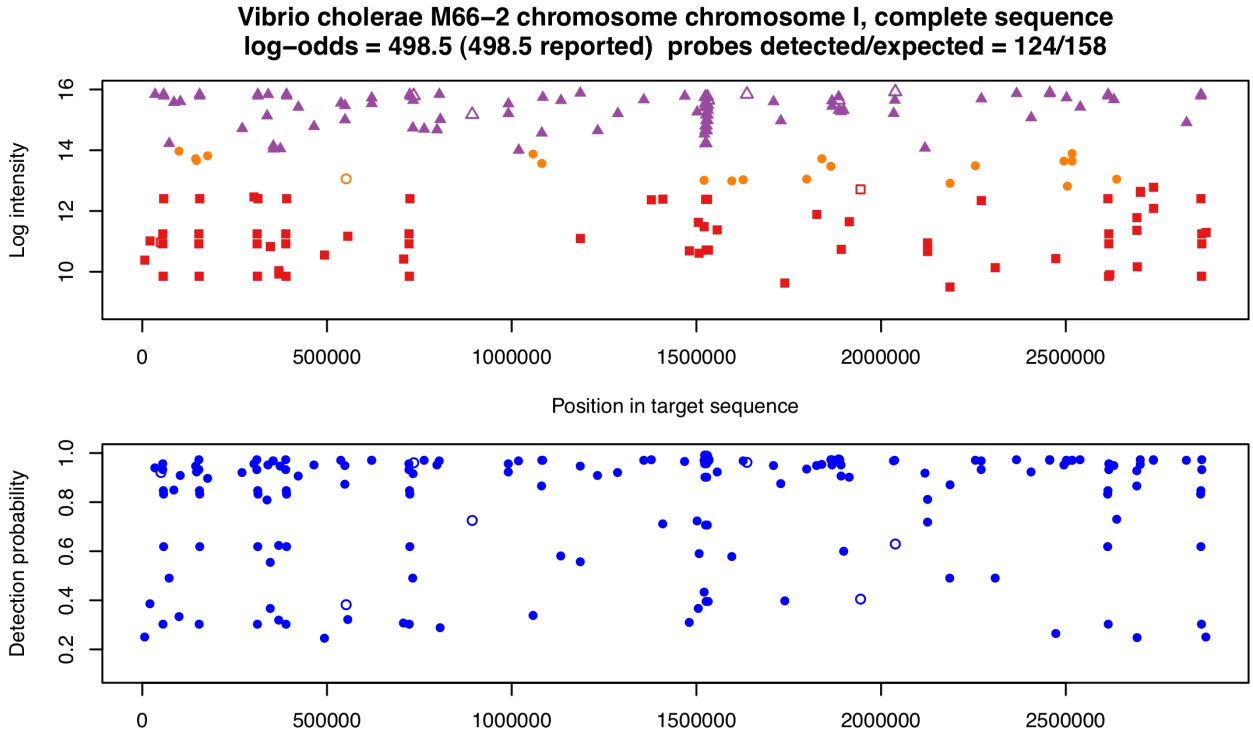
Spirochaetes	Spirochaetia	Spirochaetales	Leptospiraceae	20	54.9	57.9	1	0	0	0	0.00E+00	1
Spirochaetes	Spirochaetia	Spirochaetales	unclassified Spirochaetales	16			1	0	0	0	0.00E+00	0
Synergistetes	Synergistia	Synergistales	Synergistaceae	40	55.1	59.3	1	108.5	17	6	8.12E-05	1
Tenericutes	Mollicutes	Mycoplasmatales	Mycoplasmataceae	10	54.1	36.6	1	0	0	0	0.00E+00	1
Thermotogae	Thermotogae	Thermotogales	Thermotogaceae	16			1	0	0	0	0.00E+00	1
Verrucomicrobia	Opitutae	Opitutales	Opitutaceae	75			1	0	0	0	0.00E+00	0
Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	unclassified Verrucomicrobiales	5			1	0	0	0	0.00E+00	0

4.13 Supplementary Figures

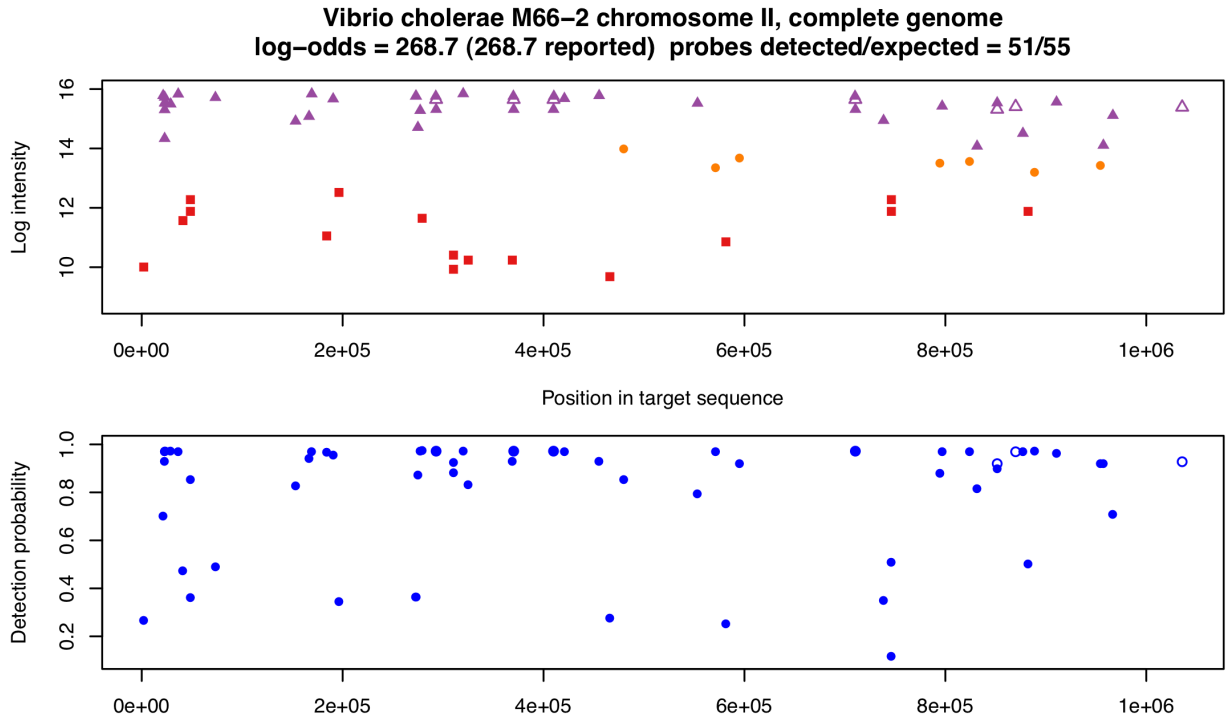
Figure S1. LLMDA detected probe distributions for *V. cholerae* and *Y. pestis* pathogens

Log intensities vs. genome position (upper graphs) and probe detection probabilities (based on similarity to target sequence) vs. position (lower graphs) for probes targeting (A) *V. cholerae* M66-2 chromosome I on array hybridized to sample 3090.13, (B) *V. cholerae* M66-2 chromosome II in sample 3090.13, and (C) *Y. pestis* CO92 plasmid pPCP1 in sample 8291. Purple triangles indicate that intensity was above the 99th percentile of the negative controls; orange circles indicated intensities between the 99th and the 95th percentiles; red squares indicate intensities below the 95th percentile. Open symbols represent probes that were excluded from the score computation, because they light up non-specifically even when there is no sample present in the hybridization mixture.

A. *V. cholerae* M66-2 chromosome 1 in sample 3090.13



B. *V. cholerae* M66-2 chromosome 2in sample 3090.13



C. Y. pestis CO92 plasmid pPCP1 in sample 8291

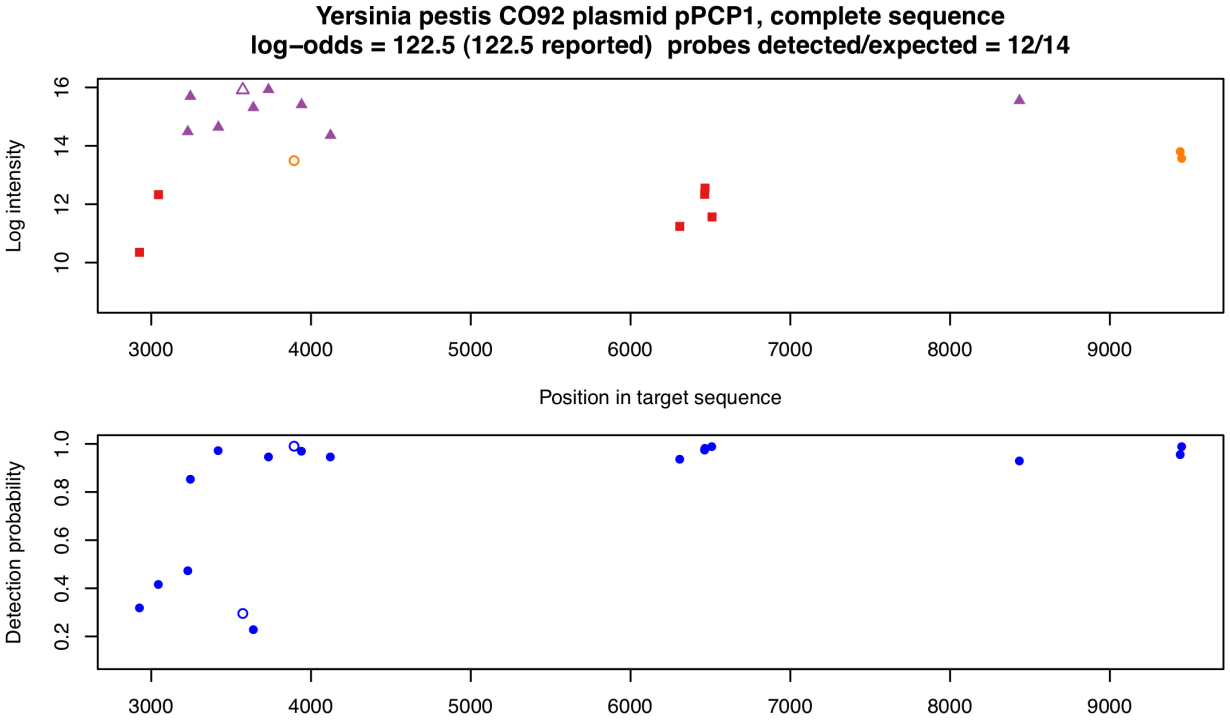


Figure S2. Flowchart of workflow

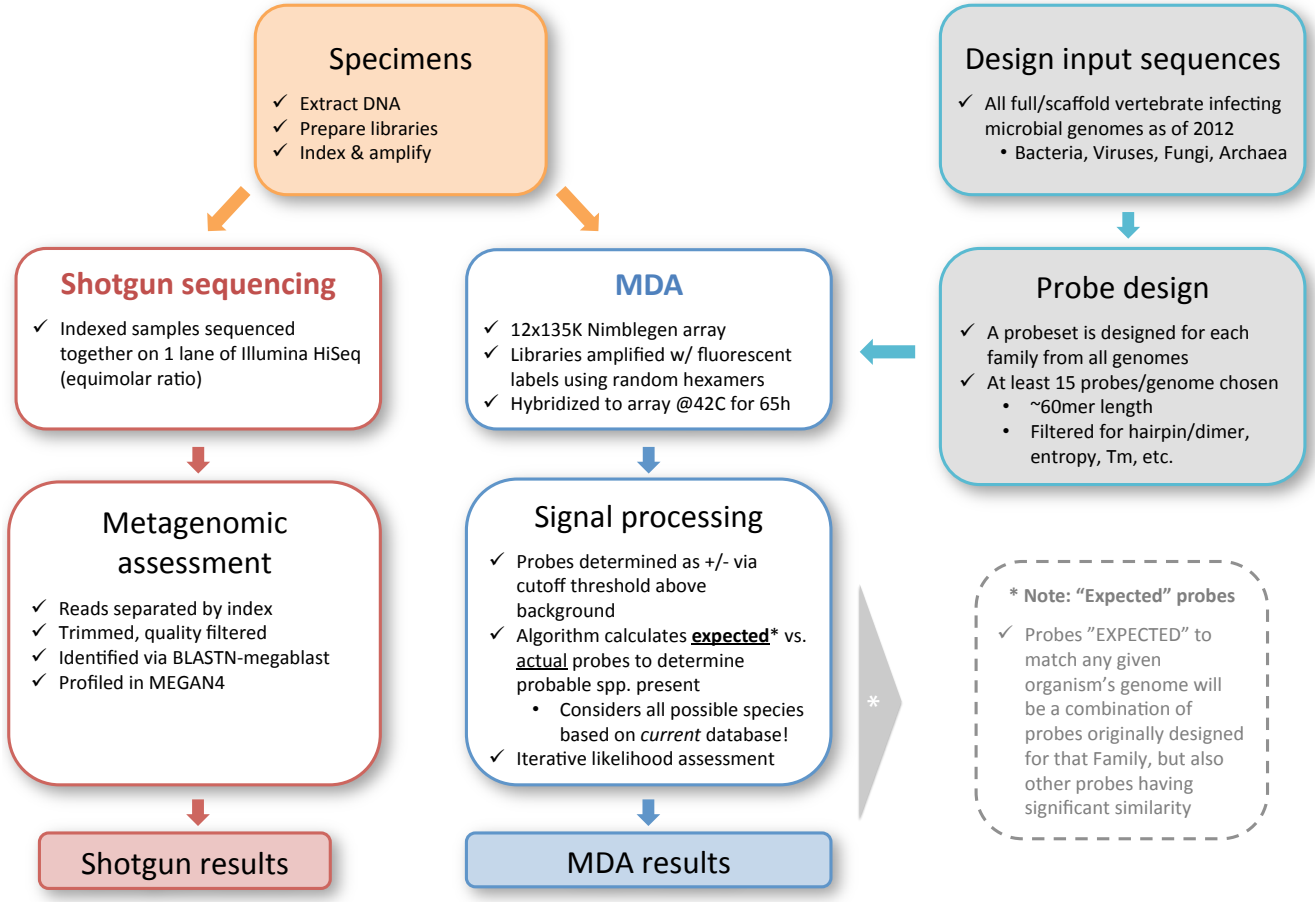


Figure S3. Average LLMDA probe GC% vs average LLMDA probe log intensity, by family.

Analysis is restricted to families used for LLMDA probe design. Families not detected with HTS are represented with blue triangles. Families detected with both methods are represented with red circles.

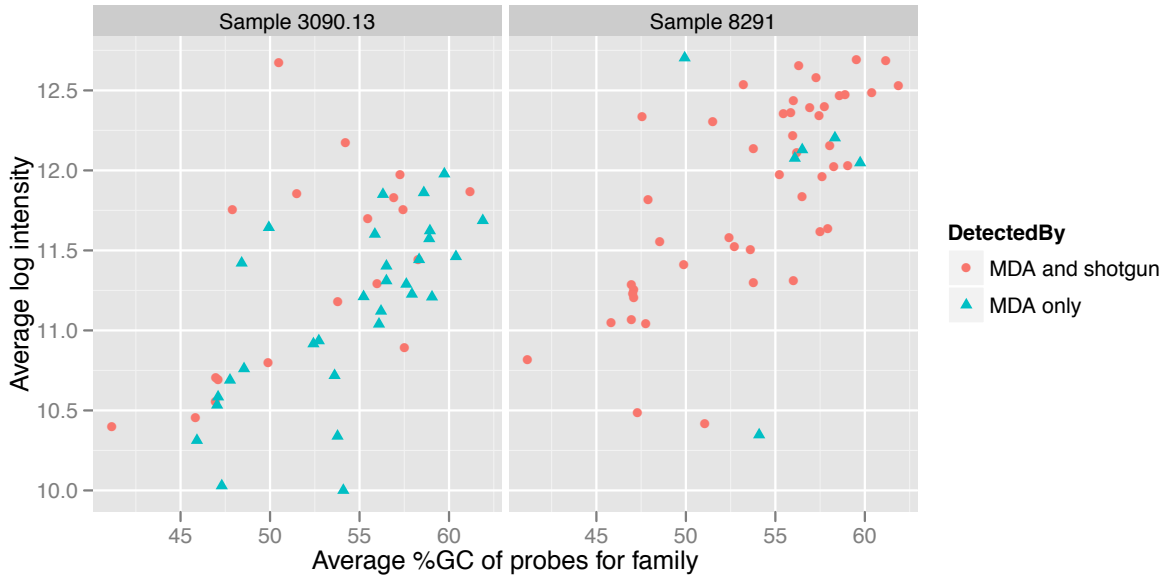


Figure S4. Number of HTS reads vs. HTS GC%, by family

For each bacterial familydetected by HTS with probes present on the MDA v5, plots of the total number of HTS reads assigned to that family versus GC% of the HTS reads. Data only shown for those families with HTS representation (3090.13 = 24; 8291 = 81). Blue triangles = families not detected via LLMDA. Red circles = families detected byLLMDA.

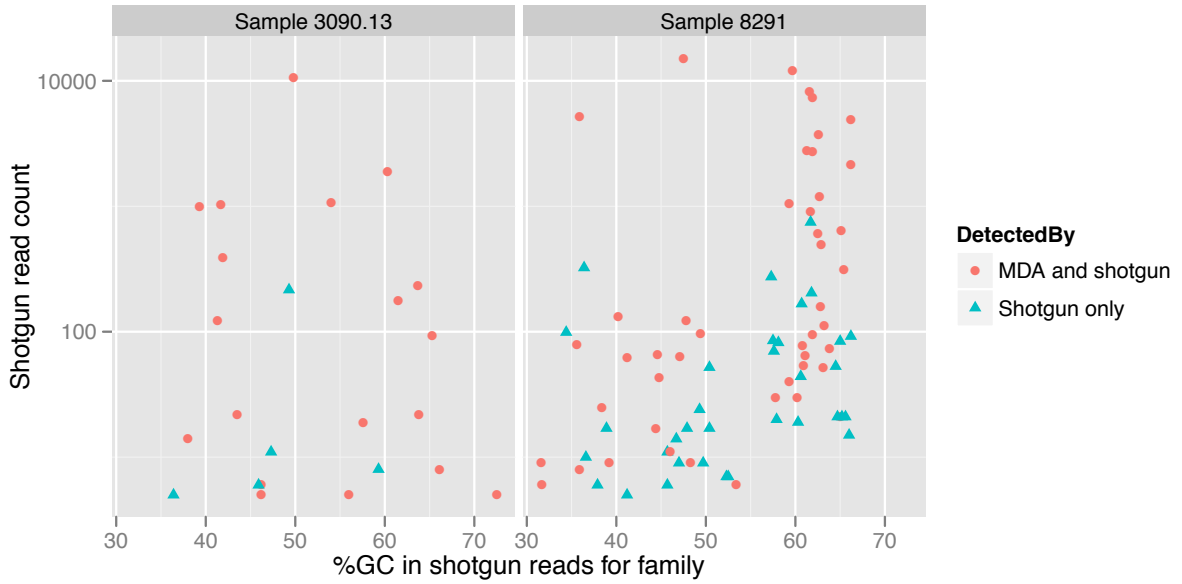
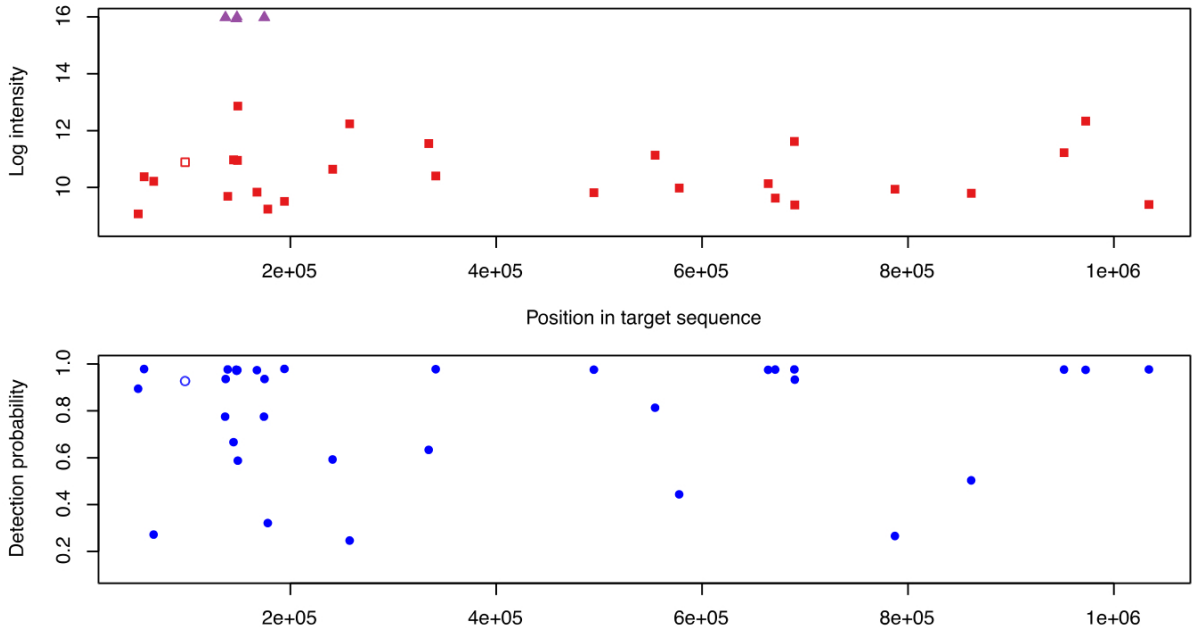


Figure S5. Examples of LLMDA detected probe distributions

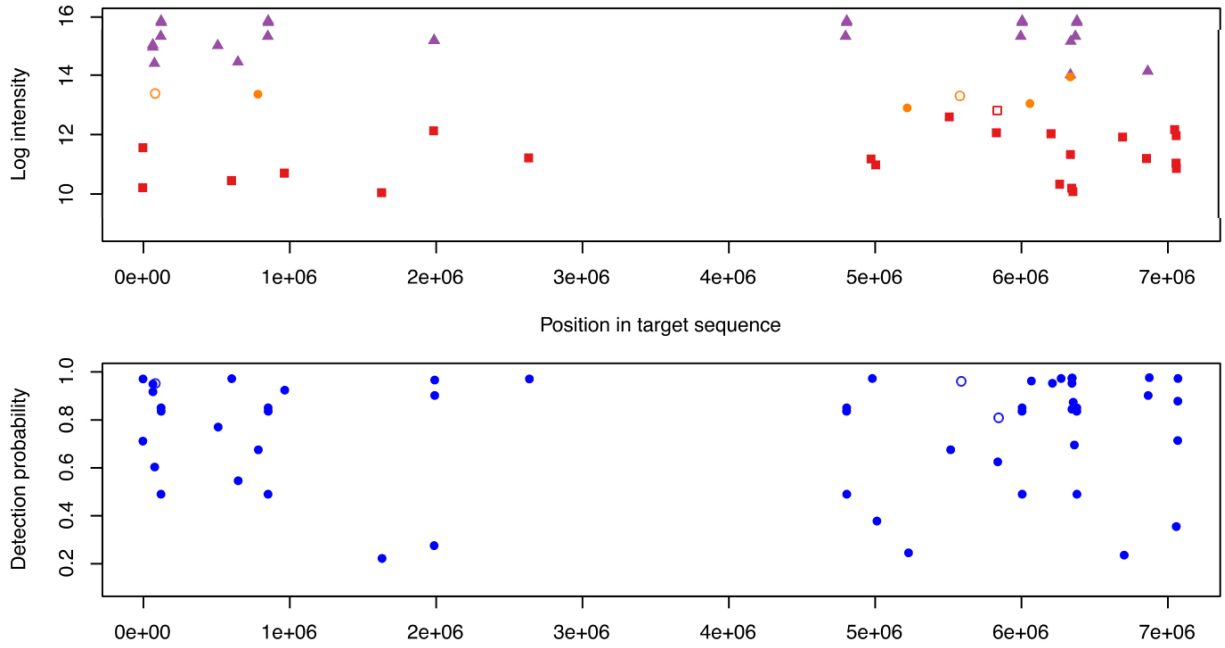
A. *Chlamydia muridarum* in sample 8291

Log intensities vs. genome position for probes targeting *Chlamydia muridarum* on array hybridized to sample 8291, and probe detection probabilities (based on similarity to target sequence) vs. position. Purple triangles indicate that intensity was above the 99th percentile of the negative controls; orange circles indicated intensities between the 99th and the 95th percentiles; red squares indicate intensities below the 95th percentile. Open symbols represent probes that were excluded from the score computation, because they light up non-specifically even when there is no sample present in the hybridization mixture. This target was removed from the predicted set because the only high-intensity probes came from a narrow region of the genome.



B. *Pseudomonas fluorescens* in sample 3090.13

Log intensities and detection probabilities vs. genome position for probes targeting *Pseudomonas fluorescens* Pf-5 on array hybridized to sample 3090.13. This target was included in the predicted set, since high-intensities are found from most regions of the genome that are covered by high-probability probes.



4.14 References

1. Donoghue HD *et al.* 2005. Co-infection of Mycobacterium tuberculosis and Mycobacterium leprae in human archaeological samples: a possible explanation for the historical decline of leprosy. *Proc R Soc B-Biol Sci* 272**:**389-394.
2. Khairat R *et al.* 2013. First insights into the metagenome of Egyptian mummies using next-generation sequencing. *J Appl Genetics* 54**:**309-325.
3. Keller A *et al.* 2012. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat Commun* 3**:**698.
4. Adler CJ *et al.* 2013. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nat Genet* 45**:**450-455.
5. Appelt S, Armougom F, Le Bailly M, Robert C, Drancourt M. 2014. Polyphasic Analysis of a Middle Ages Coprolite Microbiota, Belgium. *Plos One* 9.
6. Joseph SJ, Read TD. 2010. Bacterial population genomics and infectious disease diagnostics. *Trends Biotechnol* 28**:**611-618.
7. Nakamura S *et al.* 2009. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS ONE* 4**:**e4219.
8. McLoughlin KS. 2011. Microarrays for pathogen detection and analysis. *Brief Funct Genom* 10**:**342-353.
9. Victoria JG *et al.* 2010. Viral nucleic acids in live-attenuated vaccines: Detection of minority variants and an adventitious virus. *J Virol* 84**:**6033-6040.
10. Erlandsson L, Rosenstierne MW, McLoughlin K, Jaing C, Fomsgaard A. 2011. The Microbial Detection Array combined with random Phi29-amplification used as a diagnostic tool for virus detection in clinical samples. *PLoS ONE* 6**:**e22631.
11. Gardner S, Jaing C, McLoughlin K, Slezak T. 2010. A microbial detection array (MDA) for viral and bacterial detection. *BMC Genomics* 11**:**668.
12. Ortner DJ. 2011. Human skeletal paleopathology. *Int J Paleopathol* 1**:**4-11.
13. Rizzi E, Lari M, Gigli E, De Bellis G, Caramelli D. 2012. Ancient DNA studies: new perspectives on old samples. *Genet Sel Evol* 44**:**21.
14. Devault AM *et al.* 2014. Second-Pandemic Strain of *Vibrio cholerae* from the Philadelphia Cholera Outbreak of 1849. *N Engl J Med* 370**:**334-340.
15. Bos KI *et al.* 2011. A draft genome of Yersinia pestis from victims of the Black Death. *Nature* 478**:**506-510.
16. Schuenemann VJ *et al.* 2013. Genome-wide comparison of Medieval and modern Mycobacterium leprae. *Science* 341**:**179-183.
17. Brogden KA, Guthmiller JM, Taylor CE. 2005. Human polymicrobial infections. *Lancet* 365**:**253-255.
18. Gonzalez JM, Portillo MC, Belda-Ferre P, Mira A. 2012. Amplification by PCR artificially reduces the proportion of the rare biosphere in microbial communities. *PLoS ONE* 7**:**e29973.

19. Dabney J, Meyer M. 2012. Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques* 52**:**87-+.
20. Hoheisel JD. 2006. Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet* 7**:**200-210.
21. Wang D *et al.* 2002. Microarray-based detection and genotyping of viral pathogens. *Proc Natl Acad Sci USA* 99**:**15687-15692.
22. Palacios G *et al.* 2007. Panmicrobial oligonucleotide array for diagnosis of infectious diseases. *Emerg Infect Dis* 13**:**73-81.
23. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10**:**R25.
24. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. *J Mol Biol* 215**:**403-410.
25. Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC. 2011. Integrative analysis of environmental sequences using MEGAN4. *Genome Res* 21**:**1552-1560.
26. Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* 40**:**D130-D135.
27. Reen FJ, Almagro-Moreno S, Ussery D, Boyd EF. 2006. The genomic code: inferring Vibrionaceae niche specialization. *Nat Rev Microbiol* 4**:**697-704.
28. Brenner DJ, Farmer JJ, III. 2005. *Family I. Enterobacteriaceae*, p. 587-850. *In* Brenner DJ, Krieg NR, Staley JR (ed.), *Bergey's Manual of Systematic Bacteriology - Volume 2: The Proteobacteria, Part B: The Gammaproteobacteria*, 2nd ed. Springer, New York, NY.
29. Parkhill J *et al.* 2001. Genome sequence of Yersinia pestis, the causative agent of plague. *Nature* 413**:**523-527.
30. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 17.

Chapter 5

Conclusion

5.1 Contributions of this thesis

In this thesis, I reconstruct genome-scale sequences for three ancient bacteria never before studied with aDNA techniques: 19th century gastrointestinal pathogen *V. cholerae* and medieval pathogenic urogenital species *S. saprophyticus* and *G. vaginalis*. These reconstructions are accomplished using the cutting-edge aDNA toolkits of both TE-HTS and shotgun approaches. With the recent publication of multiple TE-free ancient pathogen genomes (45, 56, 59), we have now firmly entered the era of high-throughput, low-biased genomic studies of ancient pathogens. The novel success of the LLMDA microbial detection array tool to detect pathogens in complex ancient archaeological/historical specimens is an important methodological step for palaeomicrobiology. This affordable technique can be used by future researchers in order to assess the microbial contents of large bioarchaeological assemblages of human remains throughout time and space, or to screen large sets of specimens in order to choose samples for more in-depth sequencing analysis. These findings have relevance to the anthropological study of past health and disease, particularly in the arenas of (a) the long-time evolutionary and short-term infective interactions between humans and pathogens in the past, as well as (b) female and reproductive health.

5.2 Human-pathogen interactions

Although they are clearly linked, the state of bacterial infection must necessarily be distinguished from a state of symptomatic and/or socially-recognized disease (10). For example, infection with *M. tuberculosis* is unlikely to result in full TB disease in a healthy host that has sufficient immune response to sequester the bacteria in pulmonary tubercles (81). The pathway from bacterial infection to a particular disease is impacted by factors such as the physiological properties of the bacterium (e.g., ability to adequately colonize the requisite tissues) as well as the human immune response (10). The capacity for the latter is dependent on the phenotype of the individual, which is partially based on genotype. For example, infection with the bloodborne parasite *Plasmodium falciparum* can result in the disease malaria, but is impacted by the specific genotype of the individual. Possessing a heterozygous hemoglobin gene genotype (and having the resulting sickle-cell trait phenotype) significantly diminishes the chance of *P. falciparum* infection (and therefore reduces the outcome of the disease malaria) because the parasites cannot colonize the red blood cells (82). In another example, the ability for *V. cholerae* to infect the host's intestine is impacted by host immune response, which has been partially

linked to genetic factors (such as increased risk of infection in blood type O) as well as partial conferred immunity based on past exposure to a similar *V. cholerae* biotype (30).

However, phenotype is not merely a reflection of genotype, but also critically depends on confounding environmental factors that shape the physical manifestation of the genotype – for example, a diminished immune response due to malnutrition, pregnancy or other co-infections. These 'environmental' factors are inextricably linked to health resource access and socioeconomic status, today as well as in the past, and must be considered when examining a "holistic" picture of infection and disease (10, 16). For example, co-infection with TB and HIV is associated with poor health outcomes and is syndemically associated with disparities in health and resource access (66, 69). In another example, the chance of infection with *V. cholerae* is also increased by impaired stomach acidity, which can be linked to lower socioeconomic status (30, 83). And critically, the chance of drinking *V. cholerae*-infested water in the first place is increased in situations of poor sanitary infrastructure (84).

Past, current, and future pathogens are all situated at the intersection of multiple scales of long- and short-term processes that have shaped the interaction between humans and these microbial species: from long-term evolutionary dynamics to the specific host immune and pathogen responses/dynamics that occur in an individual during a pathogen infection (85, 86). It is not possible to study *in isolation* the short- and long-term intersecting roles of reservoirs, human and zoonotic transmission, environmental effects and cycles, specific co-infection interactions, the overall microbiome, and human behavior, because these factors are inexorably linked. Palaeopathologists are interested in all scales of these dynamics, and must make inferences from the specific archaeological outcomes of individual pathogen infections (e.g. death and/or skeletal trauma) in order to build conclusions about the overall role(s) of past pathogens in human health (21). In such cases, instance of infection must be contextualized with the information noted above such as social roles, socioeconomic status, host genotype, or co-infections and pathogen interactions (discussed below) in order to understand the factors that may (or may not) have resulted in an outcome of disease (10).

Accurate reconstruction of the timing of key events in pathogen diversification (such as the onset of human pathogenic strains from environmental precursors, or recent radiations/replacements of selectively-advantaged strains) using phylogenetic molecular dating methods is a key research theme of modern and ancient microbiology (87). These timed events can then be correlated to preceding, contemporaneous, or subsequent natural or anthropogenic changes in the environment or lifeways. For example, *Shigella sonnei* is a pathogenic *E. coli* strain that traditionally causes dysentery in developed nations, but is increasing in frequency in developing nations (88). A recent global genomic phylogenetic study

revealed that all current diversity arose from an ancestor <500 years ago. This is in accordance with an epidemiological model of increasing *S. sonnei* infections as water supply qualities have historically improved, due to the loss of cross-immunity with *Plesiomonas shigelloides* (a species found in contamined water from which *S. sonnei* acquired a virulence plasmid via LGT) (88). This has significant public health implications for balancing the design of water supply quality improvements with coordinated vaccination programs (88). However, as extant or cultured microbes can only provide a shallow time depth for analysis (<100 years), aDNA studies are very relevant for the study of much longer-term processes and/or very old events. For example, phylogenetic analysis of the recently published genomes of the Black Death and Justinian Plague strains of *Y. pestis* revealed that these epidemics were from separate pathogenic lineages, indicating that there is potential for novel virulent lineages to emerge from zoonotic reservoirs (11, 23).

5.2.1 Diseases of the first epidemiological transition

Major changes in human lifeways occurred during and after the "Neolithic Revolution" as humans adopted sedentary lifestyles due to agriculture. This brought with it corresponding dietary shifts, increased population densities, and new and different contacts with domesticated animals (89). This type of anthropogenic change resulting in new niches for infectious disease is known as "epidemiological transition theory," in which the consequences of human-pathogen interactions are considered broadly, and in the context of economic, social, and environmental effects of human behavior on infectious disease (89-91). In the first epidemiological transition that occurred during and after the transition to agriculture, such shifts caused an emergence of new zoonotic and/or 'crowd' infectious diseases, such as cholera spread by contaminated water, domesticated animal associated diseases such as tuberculosis, or vector diseases such as malaria and plague that were spread as humans created new niches for their vectors (89). As a result of these infectious diseases and lifeway changes, mortality rates were high, life expectancies were low, and birth rates were high (92). Subsequent second and third epidemiological transitions are respectively associated with higher life expectancies and higher rates of degenerative or chronic diseases due to lower infectious disease deaths and increased industrialization, and with reemergence of infectious diseases with traits such as antibiotic resistance due to continued globalization and environmental disruption (89). Although they are very different in terms of transmission and virulence, the new ancient genome-scale bacterial strains presented in this thesis may all be linked in some way to the first epidemiological transition.

In analysis of the 1849 cholera strain of Chapter 2, we attempted to date the split between the human pandemic strains (the minimum most recent common ancestor between

the O1 El Tor and O1 classical lineages). Because those lineages both possess genomic regions important for human adaptation (such as the O1 antigen cluster, cholera toxin, and adhesion genes necessary for biofilm formation and intestinal colonization) in a similar background genome (13), this earliest human adapted strain would likely have also possessed these qualities. This would have occurred when the bacterium possessed (through mutation and LGT) the necessary traits for sufficient colonization of the intestine and production of cholera toxin, but also people were living in close enough proximity that a *V. cholerae* infection would have resulted in contamination of the water supply, and a subsequent enrichment process of more virulent strains can occur (93). The emergence of cholera as a human disease therefore likely occurred during or after the transition to settled sedentary agricultural communities which would have afforded more opportunities for waterborne transmission due to increased population densities (89).

Although we were unable to calculate a timing for the most recent common ancestor of the human pathogenic *V. cholerae* strains (likely due to recombination between lineages and site saturation), our best estimate for this node is at least 400 years ago but almost certainly thousands of years older, which does support a supposed origin during the Neolithic. Based on at least our minimum calculated date of the O1 node, it is likely that the O1 El Tor-like strains were co-circulating with classical strains for hundreds or thousands of years prior to the historically recorded cholera outbreaks of the 1800s. Historic records indicate that cholera outbreaks were probably occurring in the endemic homeland of cholera for thousands of years before it emerged as a global pandemic threat (94, 95). What were the factors responsible for this emergence in the 19th century?

It is likely that similar suite of first epidemiological transition selective conditions were acting on *V. cholerae* populations both at the time of the initial emergence of cholera as a pathogen during the Neolithic as well as during the 19th century emergence of the disease: dense sedentary populations with contaminated water supplies (89). Ultimately, it is the chance, but specific, combination of host intestine, pathogen, and the phages responsible for some virulence factors (e.g., CTX) that will ultimately determine the emergence of virulent strains (94), and the conditions in the early 19th century (or earlier) were apparently conducive for this. Additionally, specific social or behavioral conditions would have been necessary for creating conditions appropriate for transmission of the novel virulent strain once it arose. During the early 19th century, the impact of what may have been a novel variant of O1 classical *V. cholerae* was exacerbated by the social structures of the relatively recent expansion of British colonial rule in India. The crowded living conditions of colonial troops, the susceptibility of European soldiers to the disease, and their extensive military movements throughout India contributed greatly to the spread of the disease during the first 19th century cholera pandemic

of 1817 (96). Stark class divisions resulted in contrasting disease outcomes for different social groups, such as a relatively lower incidence of cholera for the well-off European colonists and higher class native Indians, versus a high incidence among lower classes inhabiting crowded urban slums or rural areas (96). Additional factors such as increasing or changing trade routes (27) and particular environmental conditions such as water temperatures (known to impact modern cholera dynamics, e.g. (97)) during the 19th century would have facilitated the ultimate global spread of the disease.

While we cannot directly assess the physiological or virulence effects of the tandem CTX array discovered in the 1849 *V. cholerae* strain, we can speculate that it may have been able to produce infective virions of the CTX phage, which may have been able to excise from the genome and be transmittable to other cells (98, 99). This may have had consequences for the lateral spread of cholera toxin to other lineages, or perhaps individual cells may have been more likely have multiple copies of the CTX prophage integrated in a genome. Could this have impacted the epidemiological properties of historic pandemic *V. cholerae*? And why did O1 classical cells later lose this property? Was it in any way related to the pandemic emergence of cholera in the 1800s, or conversely to the eventual takeover of O1 El Tor as the dominant pathogenic lineage in the 1900s? Such research questions would be served not only by additional genetic studies on ancient *V. cholerae* lineages, but also epidemiological and physiological assessments of modern strains to establish a link, if any, to virulence properties of the disease.

Aside from the finding of the tandem CTX prophage, the lack of three genomic islands, and 203 SNPs, the 1849 strain is largely the same as the modern O1 classical reference. By the nature of our targeted approach, it is of course possible that the ancient strain possessed additional genomic features that we were unable to observe, and future shotgun *de novo* analysis may help to resolve this. But considering the current state of information, the 19th and 20th century versions of O1 classical are quite similar overall. In North America, pandemic cholera flourished in several epidemics during the 19th century, particularly in crowded cities with contaminated water sources that often linked to each other via riverways, facilitating spread of the disease (100). It is believed that the eradication of epidemic cholera from North America in the beginning of the 20th century was fostered by improvements in public health and sanitary infrastructure that combatted previously high rates of urban mortality due to crowding, inadequate waste disposal, lack of water treatment, wealth and racial disparities, and other factors (24, 95, 101, 102). While not directly quantifying this transition, our study on the similarity of historic and modern classical cholera indicates no significant genetic findings that would contradict this historical medical narrative. This is an example of how genetic information can support historical interpretations of public health and infectious disease, even

in those cases in which the palaeomicrobial data does not reveal any obvious "smoking guns" that explain the high recorded virulence of ancient plagues. In this way, our genetic findings on historical cholera were surprisingly similar to those of the Black Death (11), which similarly did not reveal any major significant genetic differences between the historic and modern strains, despite obviously different epidemiological outcomes. It is true that both studies may be similarly "blind" to characteristics of the past strains due to the biases of TE-HTS. However, it is also possible that these findings of large genotypic similarity between past and present infectious diseases will typify future studies of ancient pathogens. If so, this even more critically underlines the importance of future contexualization of these past epidemics in terms of additional factors which may have contributed to their impact, such as co-infections (if any), host genetics, and environmental or social conditions.

If genomic data from other ancient *V. cholerae* strains were recovered, it would greatly benefit the calibration of rates and temporal estimates in phylogenetic analysis, as well as whether genetic adaptations (if any) were responsible for the emergence of cholera as a pandemic disease in the 1800s. Such specimens could perhaps be located in the form of additional historical preserved museum specimens, mummified gastrointestinal remains, pathogen DNA adhered to sediments or burial deposits from known cholera victims, faecal or latrine remains, or sediment environmental DNA accumulated during cholera outbreaks. In a best-case scenario, one would include specimens from not only O1 classical strains (which were likely responsible for all the large known historical pandemics of the 19th century) but also of other ancient diversity from other O1 or non-O1 lineages.

5.2.2 Co-infections

In Chapter 4, our shotgun HTS and LLMDA results on the 1849 cholera victim and the Black Death victim revealed a wide multitude of additional bacterial taxa beyond their two major known pathogens. Families Pseudomonadaceae and Burkholderiaceae were detected with high abundance/likelihood in both specimens. Both families do include some human-infecting pathogens but can also be isolated from a wide range of hosts and environment (103, 104), so higher resolution analyses would be needed to determine if any potential co-infecting pathogens are present. For the cholera victim specimen #3090.13, some of the top taxa recovered by both HTS and LLMDA methods included other gut microbiome-associated families such as Clostridiaceae (105) (although this taxa is very diverse), Enterobacteriaceae (106), and Erysipelotrichaceae (107). As expected in aDNA research, many of these taxa are likely to be post-depositional contaminants, especially for Black Death victim specimen #8291, which was from a Medieval burial ground. For example, after Enterobacteriaceae (the family which contains *Y. pestis* as well as other human associated species (106)), the second-most abundant

taxa (by HTS) in the Black Death victim tooth are Alcaligenaceae, a broad group that contains environmental, animal, and human microbes (108), and Bradyrhizobiaceae which also contains many soil-dwelling species (109). Both taxa were also detected with the LLMDA. Mycobacteriaceae was also detected in the Black Death victim specimen, which is a taxon that contains bloodborne pathogens such as *M. tuberculosis* and *M. leprae* but also a variety of environmental/soil species (110). Further analysis would be needed to determine if any mycobacterial co-infections were present.

S. saprophyticus and *G. vaginalis* are both associated with urogenital disease, and the presence of both in a Medieval internal calcified infection seems to indicate similar ecological roles for these pathogens as human infections in the past. However, the outcome of co-infections with these two occasional pathogens is not clear. It is possible that the prevalence of *S. saprophyticus* and *G. vaginalis* was the result of opportunistic growth of both species due to disturbance of the healthy urinary or vaginal microbiome, which can normally act to suppress the overgrowth of pathogenic strains (111). The complicated (and sometimes nebulous) roles of occasional pathogens such as *S. saprophyticus* and *G. vaginalis* in female (and male) health typify the need for improved understanding of human-associated microbes from both genomic and metagenomic perspectives. We will require broad genomic-scale data on ancient pathogens and commensal bacteria in order to reconstruct both their physiological effects on the host as well as the phylogenetic evolutionary relationships over time, but such studies are not enough to fully understand past microbial contributions to health. We must also use metagenomic approaches in order to illuminate patterns of co-infections and pathogen interactions in the past, in order to evaluate any roles that these factors played, both specifically in the past as well as compared to today. Co-infections are critical to health outcomes due to the synergistic interaction between pathogens, and risk of contracting these pathogens is often associated with negative socioeconomic conditions that additionally contribute to health risks (68). Co-infecting pathogens can directly impact each other, but also with the host as well as other commensal species (35). Unfortunately, it is challenging with these current datasets to deeply assess the microbiomic or pathogenic significance of these findings, due to the lack of comparative contemporaneous specimens (as n=1 for each of these studies) as well as the paucity of other published ancient metagenomic datasets. However, it is encouraging that the additional metagenomic findings are somewhat representative of the depositional contexts that we would predict for these samples: apparent intestinal microbiome for the cholera victim, apparently more septicemic or soil contaminant communities for the Black Death victim, and clear presence of co-infecting urogenital pathogens for the Late Byzantine calcified female urogenital infection.

5.3 Reproductive and female health

Sex and gender differences between individuals are known to result in unequal health outcomes, some of which (e.g., iron deficiency anemia or osteoporosis) are directly accessible by traditional osteological study because they leave palaeopathological indicators on the skeleton (112). Young females also experience higher mortality rates due to the biological stresses of pregnancy and childbirth (113). This mortality risk was especially high in the past, and can be considered a significant source of ongoing selective pressure (114, 115). Childbirth and pregnancy are particularly challenging to study osteologically, due to the rarity of clearly associated mother-child remains and signatures of past pregnancy on the skeleton (113, 114, 116, 117). However, even in the absence of osteological evidence, archaeologists can also examine other remains such as birthing dwellings, aids, and equipment, as well as visual depictions in imagery or written accounts (117). The experiences and dynamics of childbirth/pregnancy have wide-ranging ties to tangible/intangible social systems and patterns such as cultural beliefs and rituals, structures of power within the community and family, and gender roles (117). Understanding fertility, birth rates, infant/child mortality is not only necessary for reconstructing palaeodemographic patterns, but more importantly for building a larger foundation for understanding the socioeconomic factors that led to those differential health outcomes in space and time in the past.

Infectious diseases can contribute directly to reproductive health but are also part of a larger social system of consequence for the individual. For example, syphilis is an infectious disease that is caused by the bacterium *Treponema pallidum pallidum*, which is transmitted either sexually or from mother to child (118). While today it can often be treated with antibiotics, the disease can result in devastating physical and neurological symptoms and high perinatal mortality (119). Syphilis is an excellent example of how multiple lines of evidence will be required to unravel the diseases' epidemiological history. While diagnoses of venereal and congenital syphilis can be made from palaeopathological lesions, related treponemal subspecies *T. pallidum pertenue* (yaws) and *T. pallidum endemicum* (bejel) found in hot climates can also leave similar skeletal lesions (118). Ultimately, the evolutionary history of the whole Treponema genus is almost certainly complex and interwoven over time, creating a critical area of research where aDNA studies have and will continue to help illuminate these phylogenetic pathways. For example, one of the major debated research questions in syphilis evolution is its history between the New and Old Worlds – was venereal syphilis brought back to Europe from the Americas with return of Columbus' crew, or was it already present in Europe? (120) Reconstructing the origin(s), evolution, and historical epidemiology of infectious diseases such as syphilis are relevant to both the evolutionary and physiological study of the pathogen itself

as well as for assisting in the resolution of significant events in human history (such as the patterns of historical contact between the New and Old Worlds).

The reproductive microbiome, and its links to other microbiomes of the body, are clearly important components of health and disease. For example, higher incidence of poor oral health (e.g. dental caries) in women have been linked in some studies to female reproductive history including pregnancy, though mechanisms and factors are complex and not entirely understood (121). Although the full implications are unclear, the recent surprising metagenomic findings of large similarity between the placental and oral microbiomes (122) demonstrates the continued need to explore the links between reproductive health and microbial communities of other areas of the body. The vaginal microbiome is one of the areas of focus for the Human Microbiome Project, which aims to examine the roles of human microbiomes in health and disease across healthy and disease individuals with a variety of analytical approaches [http://www.hmpdacc.org]. Although vaginal florae are typically dominated by lactobacilli, there is significant variation in vaginal microbiomes when surveyed metagenomically, including association between self-reported ethnicity/race, vaginal flora, and pH (12, 123-125). While there are clearly states of disease (such as bacterial vaginosis) associated with disrupted vaginal microbiomes, pinpointing exactly what the conditions are that constitute "healthy" and "unhealthy" is a significant challenge – especially considering some aspects of vaginal flora even change rapidly between serial observations (126). Understanding what aspects of environment, genetics, lifestyle, or co-infection shape these communities has critical implications for public health, since vaginal health plays a role in pregnancy outcomes such as chorioamnionitis, preterm birth, and maternal and fetal infections (127).

Although genetic information of reproductive microbiomes is generally invisible in the archaeological record, such samples should be considered for future aDNA research, such as from other calcified urogenital infections or mummified remains. There are specific research questions that would be benefit from additional archaeological specimens of reproductive and associated microbiomes. For example, I have laid out some of the unanswered questions about the relationship between vaginal microbial communities and reproductive health – to what extent are these patterns of diversity reflective of the impact of modern lifestyles, diet, and/or perturbation with antibiotics? With broad and deep enough assessment of palaeomicrobial microbiome loads in archaeological specimens spanning before and during the first epidemiological transition, we may be able to capture scenarios of increasing (or changing) frequencies of infections with opportunistic microbes like *S. saprophyticus*, and the genetic signatures of adaptation in such pathogens.

5.4 Open questions

Despite the increasing frequency and success of ancient extraskeletal and microbiome studies, there are still continuing challenges for this field and areas for further research and focus. Primarily, as already discussed, the relative rarity of the necessary preserved soft tissue or faecal specimens (compared to skeletal remains) will continue to limit such studies. There are many unresolved factors that shape how and whether evidence of endogenous bacteria is left in a skeletal or extraskeletal tissue. Even continued improvements in our breadth of samples, DNA recovery, TE-HTS, sequencing depths, and/or bioinformatic techniques may be insufficient, if the requisite DNA was never in sufficient endogenous amounts to be observable and/or lacks adequate preservation. For example, some have concluded it is unlikely or impossible for ancient syphilis DNA to be recovered from skeletal samples, due to the physiological properties of the stages of treponemal infection (128, 129). *Borrelia burgdorferi* (Lyme disease) was found during metagenomic survey of a bone from Ötzi (130), but it is very interesting that even though this species is bloodborne during infection, it has only been found in a uniquely preserved mummified bone. Perhaps this indicates that ancient *B. burgdorferi* may not preserve well in bone (or, simply, that other specimens from borreliosis infections have not yet been recovered). Perhaps in contrast, the abundance and success of mycobacterial aDNA studies has been postulated to result from good DNA preservation due to their robust lipid-rich cell walls; indeed, the ancient *M. leprae* genome study showed a higher average fragment length for *M. leprae* vs. human DNA (56). Investigators must keep in mind these putative different rates of preservation when interpreting palaeomicrobial data, an issue that will remain relevant even with the great breadth and depth of HTS studies.

Additionally, obtaining sufficient sequencing resolution for reliable identification of specific pathogenic, microbiomic, or co-infecting species is an ongoing issue. This will be aided by deeper sequencing technologies and less-biased rapid enrichment approaches such as the LLMDA. Additionally, there will be an ongoing need for thorough, genome-scale, curated modern comparative databases that hopefully will continue to expand the extent of known extant microbial diversity. The central issue is that while bacteria are extremely numerous, only a fraction of all likely species have been studied genetically, and even fewer at the full genome scale. This is due both to difficulties in establishing exact criteria for the definition of a bacterial species as well as methodological limitations in culturing bacteria or studying them only through metagenomic methods (131). While the numbers of genomic-scale studies continue to rise due to the concerted efforts of microbial researchers, there is still a very long way to go. The need for large comparative databases is especially critical for identifying *past* pathogens, since there is no guarantee that lineages of past microbiome and pathogenic bacteria will have survived to the present day – and without a close enough extant relative, it would be difficult to

evaluate the ancient species. However, with deep enough observations of modern diversity, we may be better able to assess the identities of such species, and with that, their functional roles in past health.

Deeper observations of modern diversity will necessarily need to focus on both human-associated commensal and pathogenic species from across the globe, since there are likely to be genetic, cultural, and/or environmental differences that influence both the presence and impact of different bacterial species. For example, as discussed above, studies have shown that there are significant variation (including ethnic/racial differences) in the vaginal microbiome, and effects of a particular microbiome can lead to measurable differences in factors such as pH (123-125). This as, of yet, not well understood wide variation creates challenges for the interpretation of "healthy" vs. "unhealthy" states of disease. Deeper observations of modern diversity should also focus on environmental bacteria, because of the potential for dynamics between bacterial populations via acquisition of novel genomic regions with LGT (132). The potential for bacteria to horizontally acquire factors such as antibiotic-resistance from non-human strains means that to truly understand the depths of the bacterial 'pan-genome', we will need to continue to push and explore the full range of species.

5.5 The future of palaeomicrobiology

Piers Mitchell argues (14) that much of the perceived difficulty in tackling "retrospective diagnoses" of past diseases from historical accounts is due to the differing perspectives of science and humanities researchers, but that this difficulty disappears when skills from both fields are utilized. As biological definitions of past infection continue to be established through aDNA methods, such interpretations must be reconciled with social definitions grounded in their past social contexts (14). Additionally, palaeopathological approaches typically entail an inherent assumption of a degree of uniformity between past and present, thus that past pathogens are considered to have similar physiological and epidemiological consequences to their modern counterparts (133). The *de novo* reconstruction of unique pathogen genomes from past epidemics may create opportunities for the *in vitro* testing of individual novel mutations, although restraint must be exercised against reconstructing certain past virulent strains in their entirety. Such *in vitro* testing may provide a new paradigm for past infectious disease that helps to disentangle them from reliance on modern comparative analogs.

Together, the papers that make up this thesis exemplify the fundamental shift towards deeper studies of past individual and meta-genomes that is currently occurring in the aDNA field, though there is still much progress to be made. It is true that recent examples of 'non-biased' (meta)genomic palaeopathology – e.g., the medieval oral microbiome, including the de

novo genome of the "red complex" bacterium *T. forsythia* (45), the medieval *M. leprae* genome from Denmark (56), and the medieval *B. melitensis* genome from Italy (59) – relied on specimens with sufficiently high endogenous bacterial abundances, and it goes without saying that the contributions of this thesis in Chapter 3 in this regard rely similarly on finding "that" rare extraordinary specimen. However, such outcomes are not merely the product of luck. Instead, the nature of intersecting fields of science means that new specimens and approaches are constantly changing the landscape of knowledge, and the power of HTS provides a fundamentally new understanding of ancient and modern genetics. I fully expect that the coming years will continue to provide myriad new genome-scale ancient microbial studies, both from skeletal and novel extraskeletal archaeological and historical substrates. For example, many questions about the evolution of *M. tuberculosis* (81, 134) provide a range of testable hypotheses for large-scale global ancient genomic studies— such as, did the bacterium evolve during the Neolithic transition, or is it much older? Where did it originate? To what extent is modern diversity a product of ancient human migration and/or more recent patterns of globalization? Given the demonstrated potential for *M. tuberculosis* aDNA preservation, its global modern and historical range, and major ongoing relevance to modern public health, it is both likely and desirable that we will soon see more genome-scale studies of this critical pathogen. Moreover, the study of ancient human genetic loci over time will provide key insights into the co-evolutionary dynamics of humans and *M.* tuberculosis and related species (135). In addition, the researchers will continue to explore thousands of new modern (meta)genomes that will allow for even better contextualization of these findings. If DNA from even a single exceptionally small finger bone can reveal the presence of an entirely new, previously unknown hominin lineage (136) – and yet microbiome bacteria outnumber our own cells by an order of magnitude – what will ultimately be revealed by the future study of ancient human-associated bacteria?

References – Chapter 1 (Introduction) & Chapter 5 (Conclusion)

1. Drancourt M, Raoult D. 2005. Palaeomicrobiology: Current issues and perspectives. *Nat Rev Microbiol* 3**:**23-35.
2. Orlando L *et al.* 2013. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* 499**:**74-78.
3. Spigelman M, Lemma E. 1993. The use of the polymerase chain reaction (PCR) to detect Mycobacterium tuberculosis in ancient skeletons. *Int J Osteoarchaeol* 3**:**137-143.
4. Anastasiou E, Mitchell PD. 2013. Palaeopathology and genes: Investigating the genetics of infectious diseases in excavated human skeletal remains and mummies from past populations. *Gene* 528.
5. Tsangaras K, Greenwood AD. 2012. Museums and disease: Using tissue archive and museum samples to study pathogens. *Ann Anat* 194**:**58-73.
6. Tumpey TM *et al.* 2005. Characterization of the reconstructed 1918 Spanish influenza pandemic virus. *Science* 310**:**77-80.
7. Oh CS *et al.* 2010. Amplification and sequencing of Trichuris trichiura ancient DNA extracted from archaeological sediments. *J Archaeol Sci* 37**:**1269-1273.
8. Martin MD *et al.* 2013. Reconstructing genome evolution in historic samples of the Irish potato famine pathogen. *Nat Commun* 4.
9. Ristaino JB, Groves CT, Parra GR. 2001. PCR amplification of the Irish potato famine pathogen from historic specimens. *Nature* 411**:**695-697.
10. Inhorn MC, Brown PJ. 1990. The Anthropology of Infectious Disease. *Ann Rev Anthropol* 19**:**89-117.
11. Bos KI *et al.* 2011. A draft genome of Yersinia pestis from victims of the Black Death. *Nature* 478**:**506-510.
12. The Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* 486**:**207-214.
13. Chun J *et al.* 2009. Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic Vibrio cholerae. *Proc Natl Acad Sci USA* 106**:**15442-15447.
14. Mitchell PD. 2011. Retrospective diagnosis and the use of historical texts for investigating disease in the past. *Int J Paleopathol* 1**:**81-88.
15. Roberts C, Manchester K. 2005. *The Archaeology of Disease*, 3rd ed, Cornell University Press, Ithaca, NY.
16. Brown PJ, Armelagos GJ, Maes KC. 2011. *Humans in a World of Microbes: The Anthropology of Infectious Disease*, p. 253-270. *In* Singer M, Erickson PJ (ed.), *A Companion to Medical Anthropology*, 1st ed. Blackwell Publishing Ltd., West Sussex, UK.
17. Ortner DJ. 2011. Human skeletal paleopathology. *Int J Paleopathol* 1**:**4-11.
18. Swedlund AC, Herring DA. 2003. *Human biologists in the archives: demography, health, nutrition and genetics in historical populations*, p. 1-10. *In* Herring DA, Swedlund AC (ed.), *Human Biologists in the Archives: Demography, Health, Nutrition and Genetics in Historical Populations*. Cambridge University Press, Cambridge, UK.

19. Mitchell P. 2003. The archaeological study of epidemic and infectious disease. *World Archaeol* 35**:**171-179.
20. Waldron T. 2009. *Palaeopathology*, Cambridge University Press, Cambridge, UK.
21. Wood JW *et al.* 1992. The Osteological Paradox: Problems of Inferring Prehistoric Health from Skeletal Samples [and Comments and Reply]. *Curr Anthropol* 33**:**343-370.
22. Sekirov I, Russell SL, Antunes LCM, Finlay BB. 2010. Gut Microbiota in Health and Disease. *Physiol Rev* 90**:**859-904.
23. Wagner DM *et al.* 2014. Yersinia pestis and the Plague of Justinian 541-543 AD: a genomic analysis. *Lancet Infect Dis* 14**:**319-326.
24. Rosenberg CE. 1987. *The Cholera Years – The United States in 1832, 1849, and 1866*, University of Chicago Press, Chicago, IL.
25. Griffith DC, Kelly-Hope LA, Miller MA. 2006. Review of reported cholera outbreaks worldwide, 1995-2005. *Am J Trop Med Hyg* 75**:**973-977.
26. Kaper JB, Morris JG, Jr., Levine MM. 1995. Cholera. *Clin Microbiol Rev* 8**:**48-86.
27. Barua D. 1992. *History of Cholera*, p. 1-36. *In* Barua D, Greenough III WB (ed.), *Cholera*. Plenum Medical Book Company, New York, NY.
28. Thomas CM, Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* 3**:**711-721.
29. Hacker J, Kaper JB. 2000. Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol* 54**:**641-679.
30. Sack DA, Sack RB, Nair GB, Siddique AK. 2004. Cholera. *Lancet* 363**:**223-233.
31. Li K, Bihan M, Yooseph S, Methé BA. 2012. Analyses of the Microbial Diversity across the Human Microbiome. *PLoS ONE* 7**:**e32118.
32. Qin J *et al.* 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464**:**59-65.
33. Ley RE, Turnbaugh PJ, Klein S, Gordon JI. 2006. Microbial ecology - Human gut microbes associated with obesity. *Nature* 444**:**1022-1023.
34. Hyman RW *et al.* 2014. Diversity of the Vaginal Microbiome Correlates With Preterm Birth. *Reproduct Sci* 21**:**32-40.
35. Croxen MA, Finlay BB. 2010. Molecular mechanisms of Escherichia coli pathogenicity. *Nat Rev Micro* 8**:**26-38.
36. Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The population genetics of commensal Escherichia coli. *Nat Rev Microbiol* 8**:**207-217.
37. Benezra A, DeStefano J, Gordon JI. 2012. Anthropology of microbes. *Proc Natl Acad Sci USA* 109**:**6378-6381.
38. He M *et al.* 2010. Evolutionary dynamics of Clostridium difficile over short and long time scales. *Proc Natl Acad Sci USA* 107**:**7527-7532.
39. Rollo F, Luciani S, Canapa A, Marota I. 2000. Analysis of bacterial DNA in skin and muscle of the Tyrolean iceman offers new insight into the mummification process. *Am J Phys Anthropol* 111**:**211-219.
40. Cano RJ *et al.* 2000. Sequence analysis of bacterial DNA in the colon and stomach of the Tyrolean Iceman. *Am J Phys Anthropol* 112**:**297-309.
41. Khairat R *et al.* 2013. First insights into the metagenome of Egyptian mummies using next-generation sequencing. *J Appl Genetics* 54**:**309-325.

42. Tito RY *et al.* 2012. Insights from Characterizing Extinct Human Gut Microbiomes. *PLoS ONE* 7:e51146.
43. Santiago-Rodriguez TM *et al.* 2013. Microbial Communities in Pre-Columbian Coprolites. *PLoS ONE* 8.
44. Appelt S, Armougom F, Le Bailly M, Robert C, Drancourt M. 2014. Polyphasic Analysis of a Middle Ages Coprolite Microbiota, Belgium. *PLoS ONE* 9.
45. Warinner C *et al.* 2014. Pathogens and host immunity in the ancient human oral cavity. *Nat Genet* 46:336-344.
46. Adler CJ *et al.* 2013. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nat Genet* 45:450-455.
47. Maixner F *et al.* 2014. Metagenomic Analysis Reveals Presence of Treponema denticola in a Tissue Biopsy of the Iceman. *PLoS ONE* 9:e99994.
48. Atherton JC. 2006. The pathogenesis of Helicobacter pylori-induced gastro-duodenal diseases. *Annual Review of Pathology: Mechanisms of Disease* 1:63-96.
49. Ford AC, Axon ATR. 2010. Epidemiology of Helicobacter pylori infection and Public Health Implications. *Helicobacter* 15:1-6.
50. Barnes I, Holton J, Vaira D, Spigelman M, Thomas MG. 2000. An assessment of the long-term preservation of the DNA of a bacterial pathogen in ethanol-preserved archival material. *J Pathol* 192:554 - 559.
51. Swanston T, Haakensen M, Deneer H, Walker EG. 2011. The Characterization of Helicobacter pylori DNA Associated with Ancient Human Remains Recovered from a Canadian Glacier. *PLoS ONE* 6:e16864.
52. Knapp M, Hofreiter M. 2010. Next Generation Sequencing of Ancient DNA: Requirements, Strategies and Perspectives. *Genes* 1:227-243.
53. Brown KA, Brown TA. 2013. *Biomolecular Archaeology*, p. 159-174. *In* Brenneis D, Strier KB (ed.), *Ann Rev Anthropol*, vol. 42.
54. Briggs AW *et al.* 2009. Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* 325:318-321.
55. Burbano HA *et al.* 2010. Targeted Investigation of the Neandertal Genome by Array-Based Sequence Capture. *Science* 328:723-725.
56. Schuenemann VJ *et al.* 2013. Genome-wide comparison of Medieval and modern Mycobacterium leprae. *Science* 341:179-183.
57. Mendum TA *et al.* 2014. Mycobacterium leprae genomes from a British medieval leprosy hospital: towards understanding an ancient epidemic. *BMC Genomics* 15.
58. Bouwman AS *et al.* 2012. Genotype of a historic strain of Mycobacterium tuberculosis. *Proc Natl Acad Sci USA* 109:18511-18516.
59. Kay GL *et al.* 2014. Recovery of a Medieval Brucella melitensis Genome Using Shotgun Metagenomics. *mBio* 5.
60. Rocha EPC. 2008. Evolutionary patterns in prokaryotic genomes. *Curr Op Microbiol* 11:454-460.
61. Ehrlich G, Hiller NL, Hu F. 2008. What makes pathogens pathogenic. *Genome Biol* 9:225.

62. Chain PSG *et al.* 2004. Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis. *Proc Natl Acad Sci USA* 101:13826-13831.
63. Papagrigorakis MJ, Yapijakis C, Synodinos PN, Baziotopoulou-Valavani E. 2006. DNA examination of ancient dental pulp incriminates typhoid fever as a probable cause of the Plague of Athens. *Int J Infect Dis* 10:206-214.
64. Eswarappa SM, Estrela S, Brown SP. 2012. Within-Host Dynamics of Multi-Species Infections: Facilitation, Competition and Virulence. *PLoS ONE* 7:e38730.
65. Griffiths EC, Pedersen AB, Fenton A, Petchey OL. 2011. The nature and consequences of coinfection in humans. *J Infect* 63:200-206.
66. Singer M. 2009. *Introduction to Syndemics: A Critical Systems Approach to Public and Community Health*, Jossey-Bass, San Francisco, CA.
67. Farmer P. 1996. Social Inequalities and Emerging Infectious Diseases. *Emerg Infect Dis* 2:259-269.
68. Singer M, Clair S. 2003. Syndemics and Public Health: Reconceptualizing Disease in Bio-Social Context. *Med Anthrop Quart* 17:423-441.
69. Pawlowski A, Jansson M, Sköld M, Rottenberg ME, Källenius G. 2012. Tuberculosis and HIV Co-Infection. *PLoS Pathog* 8:e1002464.
70. Clarridge JE. 2004. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. *Clin Microbiol Rev* 17:840-862.
71. Tringe SG, Hugenholtz P. 2008. A renaissance for the pioneering 16S rRNA gene. *Curr Op Microbiol* 11:442-446.
72. Pace NR. 1997. A Molecular View of Microbial Diversity and the Biosphere. *Science* 276:734-740.
73. Huse SM, Ye Y, Zhou Y, Fodor AA. 2012. A Core Human Microbiome as Viewed through 16S rRNA Sequence Clusters. *PLoS ONE* 7:e34242.
74. Rondon MR *et al.* 2000. Cloning the Soil Metagenome: a Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms. *Appl Environ Microbiol* 66:2541-2547.
75. Gonzalez JM, Portillo MC, Belda-Ferre P, Mira A. 2012. Amplification by PCR artificially reduces the proportion of the rare biosphere in microbial communities. *PLoS ONE* 7:e29973.
76. Dabney J, Meyer M. 2012. Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques* 52:87-+.
77. Aird D *et al.* 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 12:R18.
78. Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J. 2009. Metagenomic Pyrosequencing and Microbial Identification. *Clin Chem* 55:856-866.
79. Hasan NA *et al.* 2014. Microbial Community Profiling of Human Saliva Using Shotgun Metagenomic Sequencing. *PLoS ONE* 9:e97699.
80. Tringe SG *et al.* 2005. Comparative metagenomics of microbial communities. *Science* 308:554-557.
81. Russell DG. 2007. Who puts the tubercle in tuberculosis? *Nat Rev Micro* 5:39-47.

82. Verra F, Mangano VD, Modiano D. 2009. Genetics of susceptibility to Plasmodium falciparum: from classical malaria resistance genes towards genome-wide association studies. *Parasite Immunol* 31**:**234-253.
83. Sack GH, Jr. *et al.* 1972. Gastric acidity in cholera and noncholera diarrhoea. *Bull World Health Org* 47**:**31-36.
84. Root ED, Rodd J, Yunus M, Emch M. 2013. The Role of Socioeconomic Status in Longitudinal Trends of Cholera in Matlab, Bangladesh, 1993–2007. *Plos Neglect Trop Dis* 7**:**e1997.
85. Wilson DJ. 2012. Insights from Genomics into Bacterial Pathogen Populations. *PLoS Pathog* 8.
86. Barreiro LB, Quintana-Murci L. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat Rev Genet* 11**:**17-30.
87. Kuo CH, Ochman H. 2009. Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria. *Biol Direct* 4**:**10.
88. Holt KE *et al.* 2012. Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat Genet* 44**:**1056-+.
89. Armelagos GJ, Brown PJ, Turner B. 2005. Evolutionary, historical and political economic perspectives on health and disease. *Soc Sci Med* 61**:**755-765.
90. Omran AR. 1983. The epidemiologic transition theory: A preliminary update. *J Trop Pediatrics* 29**:**305-316.
91. Omran AR. 1971. The epidemiologic transition: A theory of the epidemiology of population change. *Milbank Meml Fund Q* 49**:**509-538.
92. McKeown RE. 2009. The Epidemiologic Transition: Changing Patterns of Mortality and Population Dynamics. *Am J Lifestyle Med* 3**:**19S-26S.
93. Faruque SM *et al.* 2004. Genetic diversity and virulence potential of environmental Vibrio cholerae population in a cholera-endemic area. *Proc Natl Acad Sci USA* 101**:**2123-2128.
94. Mekalanos JJ, Rubin EJ, Waldor MK. 1997. Cholera: molecular basis for emergence and pathogenesis. *FEMS Immunol Med Microbiol* 18**:**241-248.
95. Lacey SW. 1995. Cholera - Calamitous past, ominous future. *Clin Infect Dis* 20**:**1409-1419.
96. Arnold D. 1986. Cholera and Colonialism in British India. *Past & Present***:**118-151.
97. Lobitz B *et al.* 2000. Climate and infectious disease: Use of remote sensing for detection of Vibrio cholerae by indirect measurement. *Proc Natl Acad Sci* 97**:**1438-1443.
98. Davis BM, Waldor MK. 2000. CTXphi contains a hybrid genome derived from tandemly integrated elements. *Proc Natl Acad Sci USA* 97**:**8572-8577.
99. Davis BM, Moyer KE, Boyd EF, Waldor MK. 2000. CTX Prophages in Classical Biotype Vibrio cholerae: Functional Phage Genes but Dysfunctional Phage Genomes. *J Bacteriol* 182**:**6992-6998.
100. Pyle GF. 1969. The Diffusion of Cholera in the United States in the Nineteenth Century. *Geographical Analysis* 1**:**59-75.
101. Haines MR. 2001. The urban mortality transition in the United States, 1800-1940. *Annales de démographie historique***:**33-64.

102. Condran GA, Crimmins-Gardner E. 1978. Public Health Measures and Mortality in U.S. Cities in the Late Nineteenth Century. *Human Ecology* 6**:**27-54.
103. Palleroni N. 1981. *Introduction to the Family Pseudomonadaceae*, p. 655-665. *In* Starr M, Stolp H, Trüper H, Balows A, Schlegel H (ed.), *The Prokaryotes*. Springer Berlin Heidelberg.
104. Coenye T, Vandamme P. 2003. Diversity and significance of Burkholderia species occupying diverse ecological niches. *Environ Microbiol* 5**:**719-729.
105. Stackebrandt E, Kramer I, Swiderski J, Hippe H. 1999. Phylogenetic basis for a taxonomic dissection of the genus Clostridium. *FEMS Immunol Med Microbiol* 24**:**253-258.
106. Brenner DJ, Farmer JJ, III. 2005. *Family I. Enterobacteriaceae*, p. 587-850. *In* Brenner DJ, Krieg NR, Staley JR (ed.), *Bergey's Manual of Systematic Bacteriology - Volume 2: The Proteobacteria, Part B: The Gammaproteobacteria*, 2nd ed. Springer, New York, NY.
107. Zhang H *et al.* 2009. Human gut microbiota in obesity and after gastric bypass. *Proc Natl Acad Sci* 106**:**2365-2370.
108. Whiteson K *et al.* 2014. A genomic perspective on a new bacterial genus and species from the Alcaligenaceae family, Basilea psittacipulmonis. *BMC Genomics* 15**:**169.
109. Moulin L, Béna G, Boivin-Masson C, Stępkowski T. 2004. Phylogenetic analyses of symbiotic nodulation genes support vertical and lateral gene co-transfer within the Bradyrhizobium genus. *Mol Phylogen Evol* 30**:**720-732.
110. Hruska K, Kaevska M. 2012. Mycobacteria in water, soil, plants and air: a review. *Veterinarni Medicina* 57.
111. Ruiz F *et al.* 2009. Antimicrobial Activity, Inhibition of Urogenital Pathogens, and Synergistic Interactions Between Lactobacillus Strains. *Current Microbiology* 59**:**497-501.
112. 1998. *Sex and Gender in Paleopathological Perspective*, 1st ed, Grauer AL, Stuart-Macadam P (ed.), Cambridge University Press, Cambridge, UK.
113. Willis A, Oxenham MF. 2013. A Case of Maternal and Perinatal Death in Neolithic Southern Vietnam, c. 2100–1050 BCE. *Int J Osteoarchaeol* 23**:**676-684.
114. Pfeiffer S *et al.* 2014. Discernment of mortality risk associated with childbirth in archaeologically derived forager skeletons. *Int J Paleopathol* 7**:**15-24.
115. Trevathan WR. 2010. *Ancient Bodies, Modern Lives: How Evolution Has Shaped Women's Health*, Oxford University Press, New York, NY.
116. Sayer D, Dickinson SD. 2013. Reconsidering obstetric death and female fertility in Anglo-Saxon England. *World Archaeol* 45**:**285-297.
117. Beausang E. 2000. Childbirth in Prehistory: An Introduction. *European Journal of Archaeology* 3**:**69-87.
118. de Melo FL, de Mello JCM, Fraga AM, Nunes K, Eggers S. 2010. Syphilis at the Crossroad of Phylogenetics and Paleopathology. *PLoS Negl Trop Dis* 4**:**e575.
119. Singh AE, Romanowski B. 1999. Syphilis: Review with Emphasis on Clinical, Epidemiologic, and Some Biologic Features. *Clin Microbiol Rev* 12**:**187-209.
120. Armelagos GJ, Zuckerman MK, Harper KN. 2012. The Science Behind Pre-Columbian Evidence of Syphilis in Europe: Research by Documentary. *Evolutionary Anthropology: Issues, News, and Reviews* 21**:**50-57.

121. Lukacs JR. 2012. *Oral Health in Past Populations: Context, Concepts and Controversies*, p. 553-581. *In* Grauer AL (ed.), *A Companion to Paleopathology*, vol. 1. Blackwell Publishing Ltd., West Sussex, UK.
122. Aagaard K *et al.* 2014. The Placenta Harbors a Unique Microbiome. *Sci Translat Med* 6**:**237ra265.
123. Forney LJ, Foster JA, Ledger W. 2006. The Vaginal Flora of Healthy Women Is Not Always Dominated by Lactobacillus Species. *J Infect Dis* 194**:**1468-1469.
124. Zhou X *et al.* 2004. Characterization of vaginal microbial communities in adult healthy women using cultivation-independent methods. *Microbiology* 150**:**2565-2573.
125. Ravel J *et al.* 2011. Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci USA* 108**:**4680-4687.
126. Gajer P *et al.* 2012. Temporal Dynamics of the Human Vaginal Microbiota. *Sci Translat Med* 4**:**132ra152.
127. Donati L *et al.* 2010. Vaginal microbial flora and outcome of pregnancy. *Archives of Gynecology and Obstetrics* 281**:**589-600.
128. Bouwman AS, Brown TA. 2005. The limits of biomolecular palaeopathology: ancient DNA cannot be used to study venereal syphilis. *J Archaeol Sci* 32**:**703-713.
129. von Hunnius TE, Yang D, Eng B, Waye JS, Saunders SR. 2007. Digging deeper into the limits of ancient DNA research on syphilis. *J Archaeol Sci* 34**:**2091-2100.
130. Keller A *et al.* 2012. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat Commun* 3**:**698.
131. Staley JT. 2006. The bacterial species dilemma and the genomic–phylogenetic species concept. *Phil Trans R Soc B* 361**:**1899-1909.
132. Wiedenbeck J, Cohan FM. 2011. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol Rev* 35**:**957-976.
133. Grauer AL. 2012. *Introduction: The Scope of Paleopathology*, p. 1-14. *In* Grauer AL (ed.), *A Companion to Paleopathology*, vol. 1. Blackwell Publishing Ltd., West Sussex, UK.
134. Gagneux S. 2012. Host–pathogen coevolution in human tuberculosis. *Phil Trans R Soc B* 367**:**850-859.
135. Stone AC, Wilbur AK, Buikstra JE, Roberts CA. 2009. Tuberculosis and leprosy in perspective. *Am J Phys Anthropol* 140 Suppl 49.
136. Reich D *et al.* 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468**:**1053-1060.