Development and Analysis of Molecular Methods for Functional Metagenomics of the Human Gut Microbiome

by

Kathy Nguyen Lam

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the the degree of
Doctor of Philosophy
in
Biology

Waterloo, Ontario, Canada, 2016

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Each of the seven chapters in this thesis begins with a section entitled **Acknowledgements and declarations**, which states whether and where the work has been published or previously written. The section also lists the individuals who contributed to the chapter as well as describes the exact nature of those contributions.

Abstract

Interest in the human microbiome has risen quickly in recent years as the microbes that live in and on our body have been implicated in a growing number of human health and disease states. This interest has been supported by advances in DNA sequencing technology that have allowed us to obtain vast amounts of sequence data, and yet we have difficulty assigning function to many of the gene sequences obtained. As research on the role of these microorganisms continues, there will be an increased need for high-throughput methods that can provide knowledge of microbial gene function. Functional metagenomics is one such method, and it relies on first cloning environmental DNA to generate metagenomic libraries that are maintained in *Escherichia coli* and second, screening the cloned DNA for particular functions of interest. This powerful function-first method allows for the isolation of genes whose role may not have been predicted using DNA sequence homology. This thesis describes the analysis of techniques used in functional metagenomics research, as well as the development of new strategies to aid in functional screening of metagenomic libraries, particularly those constructed from gut-derived DNA. The work is divided into four data chapters that each explore a distinct aspect of the functional metagenomics approach.

The first data chapter describes the evaluation of a pooled strategy for sequencing cosmid clones that were previously isolated in functional screens of metagenomic libraries. Ninety-two large-insert clones were pooled for Illumina-sequencing and the assembled sequence data were evaluated against reference sequence data that were obtained from individual barcoded Illumina sequencing of the same clones. The results indicated that a pooled strategy works well provided that sufficient sequencing depth is obtained and that pooled clones do not share sequence similarity to the extent that would be problematic for assembly of short reads that derive from those clones.

The second data chapter is an exploration of possible causes for the known cloning bias of metagenomic libraries, by comparing environmental DNA before cloning to the DNA cloned in the final metagenomic library in *E. coli*. For a human gut metagenomic library, DNA was sampled and Illumina-sequenced at three different steps during the construction of the library. Analyses of the sequence data showed that there was indeed major bias in the final library, but that the bias was not due to fragmentation of the DNA during the cloning process as has been previously suggested; rather, the data were consistent with alternative hypotheses that suggest bias occurs after the DNA is

introduced into *E. coli*, and analyses provide support for the hypothesis that spurious transcription of foreign DNA in *E. coli* may be contributing to the bias of libraries. Bias was also examined for a soil metagenomic library using 16S rRNA gene sequencing and though broad phylum-level biases were not as severe as observed for the human gut library, analyses revealed a bias in the relative abundance of individual OTUs.

The third data chapter describes efforts to develop *Bacteroides thetaiotaomicron* (*B. theta*) VPI-5482 as a surrogate host for screening metagenomic libraries constructed from human gut-derived DNA. In this strategy, metagenomic libraries that have been constructed in *E. coli* can be transferred to *B. theta* using triparental conjugation. A member of the Bacteroidetes was chosen to specifically address the likely barrier to gene expression in *E. coli* of DNA that originates from this phylum. To allow the library to be replicated in *B. theta*, a *B. theta*-compatible library cloning vector was constructed, and this vector was used to generate genomic and metagenomic clone libraries. A metagenomic library was successfully screened in *B. theta*, leading to functional complementation of a *B. theta* mutant strain unable to grow on chondroitin sulfate as sole carbon source. However, further examination of the complemented clones indicated that the library clone DNA had integrated into the *B. theta* mutant genome. To address this problem, an alternative method for screening was devised, and although this method demonstrates that screening in *B. theta* remains feasible, more work is required to optimize the conjugation efficiency and the level of throughput.

The fourth and last data chapter is an exploration of the use of transcriptional terminator elements in library cloning vectors, inspired by the results of previous chapters. Two unidirectional transcriptional terminators were added to a copy number-inducible fosmid vector, flanking the cloning site, with the intention of reducing insert-born transcription into the vector backbone. The terminators were tested using a reporter gene to confirm their functionality in this context, and derivative vectors were generated for future testing of whether or in what contexts terminators may help alleviate cloning bias in metagenomic libraries. The work described in this thesis contributes to method advancement for functional metagenomics through the analysis of a cost-effective strategy for sequencing library clones, the examination of potential causes of sequence bias in metagenomic libraries, the development of a surrogate host for more productive functional screening, and the consideration of vector elements that may improve metagenomic library stability in *E. coli*.

Acknowledgements

"Professors spend most of their time doing research? I thought they just taught classes – like regular teachers, but at a university." – Kathy Lam, circa 2006

I am grateful to my supervisor Dr. Trevor Charles for seeing my potential through my ignorance and for giving me the chance to discover the joy of experimentation. Thank you for providing me with the freedom required for science to be fun, but supplementing that with an open office door. Thank you most of all for reminding me to stay positive and to simply go where the science takes me when I am working in uncharted territory.

I thank my supervisory committee members, Drs. Josh Neufeld, David Rose, and Gabriel Moreno-Hagelsieb for their support and suggestions. I am grateful to Josh in particular for helping me improve my science communication skills through the use of thoughtful visual narratives. I also thank Dr. Eric Martens for hosting me at his lab at the University of Michigan.

I am grateful for the scholarships and funding I've held through these years, without which I would not have been able to pursue my studies. I remain indebted to the Canadian public and various funding agencies and institutions, including the Canadian Institute of Health Research, the Natural Sciences and Engineering Research Council, the Government of Ontario, and the University of Waterloo.

I thank past and present members of the Charles Lab for their friendship, scholarly discussions, and technical advice. In particular, thank you to Katja Engel for a genuine good nature, Tam Tran for being so easy-going, Ariana Marcassa for thought-provoking conversations, Maya D'Alessio for a love of orderliness, John Heil for diverse problem-solving skills, and Jiujun Cheng for a vast knowledge of molecular biology. I remain grateful to Maria Trainer for introducing me to molecular biology methods and I credit MIT's iGEM competition for igniting a lasting love of scientific pursuit.

I thank my parents, Co Lam and Luyen Nguyen, for their love and support during my higher education. Con cảm ơn Cha Mẹ yêu thương con và ủng hô con cố gắng học cao. I am grateful to my father for teaching me to be skeptical; my mother, compassionate. I also thank my acquired parents Marina Kolpatkchi and Alexandre Юрьевич Doubov for their love, consideration, and encouragement. I am grateful to my brother Nam Lam and close friends Lillian Lem and Melissa Lem for giving me breaks away from work, camping trips away from civilization, and for encouraging me with the incessant tease, "Are you ever gonna be done school?". I also thank Andre Masella for his close friendship and many long discussions about science and life.

Finally, to Alexandre Александрович Doubov, my partner, husband, best friend, and Солнышко: I thank you for your interest in my work, your sacrifices, your many hours of both IT support and moral support, and your ease of being that complements my frenzied nature and moderates my "binary" personality. Possibly most of all, I thank you for being my anchor to the equally wonderful world outside of academic research, thereby bringing balance to my life. Спасибо, Саныч! Я тебя люблю и я по тебе скучаю, когда ты не со мной.

Dedication

Con, Lam Nguyễn Hoài Hương, tặng luận án nay cho Cha, Lam Co, và Mẹ, Nguyễn Thị Hồng Luyến, vì Cha Mẹ chiu đưng trương hợp rất khó khi đi vượt biển, vì Cha Mẹ cho con cơ hội được sống ở xứ tự do, vì Cha Mẹ nuôi nâng con ở một nơi xa lạ, vì Cha Mẹ ủng hộ con học cao. Con sẽ không quên là cuộc sống có lúc rất vất vả, và con rất là may mắn, và cả đời này con sẽ không thể nào trả nợ được Cha Mẹ.

I, Kathy Nguyen Lam, dedicate this thesis to my father, Co Lam, and mother, Luyen Thi Hong Nguyen, for enduring great hardship as Vietnamese boat people, for affording me life in a free and democratic society, for raising me in an unfamiliar culture, and for encouraging my higher education. Mom, Dad, I will not forget that life can be difficult, that I am very fortunate, nor that I carry a debt that can never be repaid.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

12AC	unique identifier for soil sample from agricultural corn field
anSME	anaerobic sulfatase maturase enzyme
Ap	ampicillin
AT	adenine-thymine
BAC	bacterial artificial chromosome
bp	basepair
BHI	brain heart infusion media
BHI+	brain heart infusion media with supplementation
BHIH	brain heart infusion media with 10% horse blood
BLAST	basic local alignment search tool
blastn	BLAST nucleotide to nucleotide search
blastx	BLAST translated nucleotide to protein search
BT1	pJC8-based library constructed from *B. theta* DNA
BT2	pKL3-based library constructed from *B. theta* DNA
BT3	pKL13-based library constructed from *B. theta* DNA
CLGM1	pJC8-based library constructed from human get metagenomic DNA
CLGM2	pKL3-based library constructed from human get metagenomic DNA
CLGM3	pKL13-based library constructed from human get metagenomic DNA
Cm	chloramphenicol
chuR	chondroitin sulfate utilization regulator
CTAB	cetyltrimethylammonium bromide
dH_2O	distilled water
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid

Em	erythromycin
EDTA	ethylenediaminetetraacetic acid
g	g-force
Gb	gigabase
GC	guanine-cytosine
GFP	green fluorescent protein
GFPuv	highly fluorescent variant of wild-type GFP
GH	glycoside hydrolase
GI	gastrointestinal
Gm	gentamicin
k-mer	oligonucleotide of length k
HMP	human microbiome project
HMW	high molecular weight
IPTG	isopropylb-D-1-thiogalactopyranoside
LB	lysogeny broth or Luria-Bertani media
kb	kilobases
Km	kanamycin
Mb	megabase
MM	minimal media
na	not applicable
NA	nalidixic acid
nd	not determined
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
Nx	nalidixic acid
OD	optical density
ORF	open reading frame
oriT	origin of transfer
oriV	origin of vegetative replication
OTU	operational taxonomic unit
PCR	polymerase chain reaction
PNK	polynucleotide kinase
PL	polysaccharide lyase
P_{tac}	*tac* promoter

PUL	polysaccharide utilization locus
RBS	ribosome-binding site
RDP	ribosomal database project
rRNA	ribosomal RNA
RNA-seq	RNA sequencing
rpm	revolutions per minute
SCODA	synchronous coefficient of drag alteration
SDS	sodium dodecyl sulfate
Sm	streptomycin
SNP	single nucleotide polymorphism
SRA	short read archive
SUS	starch utilization system
TAE	Tris-acetic acid-EDTA
Tc	tetracycline
TE	Tris-EDTA
Tp	trimethoprim
TT	transcriptional terminator
TYG	tryptone-yeast-glucose
w/v	weight by volume
v/v	volume by volume

List of Symbols

Δ	denotes gene deletion; precedes the name of the gene deleted
Φ	denotes a phage; precedes the name of the phage
λ	lambda phage
σ	sigma factor
$\sim$	approximately
$\times$	fold greater than, with respect to a reference

Chapter 1

Introduction

1.1 Acknowledgements and declarations

- Part of the introduction for this chapter was written as part of a review for my graduate course in bacterial molecular genetics, BIOL 608.
- A few paragraphs of this introduction, in Section 1.6.1, are from a Perspective article in the journal **Frontiers in Microbiology**. I was the primary author of this article. The citation for the article is:

  > Lam KN, Cheng J, Engel K, Neufeld JD, Charles TC (2015) Current and future resources for functional metagenomics. *Frontiers in Microbiology* 6:1196. doi:10.3389/fmicb.2015.01196

- This chapter was also proofread by my supervisor **Trevor Charles**.

1.2 Abstract

Interest in the human microbiome has risen quickly in recent years as technological advancements have allowed us to explore the microbial world in unprecedented depth, and the gut environment in particular has attracted much attention. The interaction between the human host and their microorganisms begins at birth, varies between individuals, and fluctuates throughout life with environmental influences such as diet. These microbes contribute to our health, but have also been implicated in various disease states through an altered composition of microbiota although causal links for many have yet to be shown. Moving from more correlative studies to those providing explanatory mechanisms will likely require a broader knowledge of microbial gene function, as many genes identified from shotgun metagenomic sequencing datasets lack a sequence homology-based functional annotation, interfering with our ability to understand the role of the microorganisms present as a whole. To address this lack in knowledge will require high-throughput methods to mine genes using a function-first approach, allowing function to be determined for those genes whose function could not have been predicted using sequence homology. Functional metagenomics is one such method, in which DNA is isolated from environmental samples, cloned en masse, and screened for particular enzymatic activities. This thesis describes the analysis and development of methods to advance functional metagenomics, particularly for study of the human gut microbiome.

1.3 Interest in the human microbiome

Over the past couple of decades, there has been mounting interest in the *human microbiome*, that is, the community of microorganisms living on and in the human body and the host environment with which they interact [171]. The microorganisms themselves are distinctly referred to as the *microbiota* [199]. The productivity in this research area has been largely due to technological advances in DNA sequencing, allowing researchers to deep-sequence DNA samples isolated from various parts of the body. This requires isolating the *metagenomic* DNA of these environments – a term originally coined by Jo Handelsman during studies of soil microorganisms that refers to the collective genomic DNA from an environmental sample [118].

Metagenomic methods are crucial in studies of the human microbiome, as many of these organisms may not be easily cultured using standard laboratory techniques. Some estimates of the fraction of uncultured bacteria in oliogtrophic environments have been as high as 99% [242,319]; in the nutrient-rich system of the gastrointestinal tract, however, previous studies have cited 50% uncultivated taxa in the stomach (2006) [22], 80% in the distal intestine (2006) [102], and 70% in the oral cavity (2010) [60]. Although these taxa are occasionally referred to as "unculturable" [242], recent reports have challenged this idea with the isolation hundreds of species from the human gut, including novel ones, using carefully designed and comprehensive culturing techniques [109,163,319].

In 2007, with growing interest in the scientific community and funding from the NIH, the Human Microbiome Project (HMP) was inititated – a five-year, $150 million collaborative endeavour to characterize various human microbial communities, targeting the skin, oral cavity, nasal cavity, vagina, and gastrointestinal tract [233]. Today, the list of body sites has expanded to include other body parts, such as the urogenital

tract, with the goal of providing 3,000 reference genomes, either sequenced or collected from public databases. The majority of these genomes will be sequenced only to a high-quality draft stage, which is the second of six possible stages of completion, as provisionally defined by the HMP Consortium. To be considered high-quality, the draft sequence must, among other requirements, have >90% of the genome included in contigs ≥500 bp, with >90% of bases at >5× read coverage, and >90% of Bacterial "core genes" present. At the moment, the HMP has ∼1,700 bacterial reference genomes either finished or in progress.

1.4 The human gut microbiome

A fact often given to illustrate the importance of the human microbiome is that microbial cells outnumber human cells by at least a factor of 10, and their genes outnumber human genes by at least a factor of 100 [102, 253], although a more recent study has countered this widely cited claim with estimates that the bacterial cell to human cell ratio is in fact closer to one-to-one [260]. Regardless of the precise number, it is indisputable that microorganisms occupy our body sites where they play an important role; of all human microbiomes, the gut seems to have attracted the most research interest, likely because the vast majority of the microbes we harbour reside in the gastrointestinal tract, particularly in the distal gut where they aid in host metabolism [102] and influence host immunity [176, 246]. To determine which organisms form the microbiota, and in what proportion, the culture-independent approach of 16S rRNA gene sequencing is often used – sequencing either the full 16S rRNA gene length or one of the hypervariable regions. Typically, though somewhat arbitrarily, cut-offs of 95% and 97-98% identity are used to define Genus and Species (or Operational Taxonomic Unit, OTU), respectively [12].

1.4.1 Initial colonization

The gastrointestinal tract of a newborn is sterile and, in a vaginal birth, the mother's microbes serve as the initial inoculum for the newborn, along with other external contacts that may take place during birth. Initial colonizing bacteria are facultative, lowering the redox potential of the environment, allowing strict anaerobes to flourish [150]. Later in life, other microorganisms are introduced; for example, with the ingestion of food, bacterial survival through the acidic environment of the stomach is aided by the rise in pH immediately following a meal [170]. Interestingly, studies have suggested that birthing via caesarean section may have negative consequences. For example, compared to infants delivered vaginally, initial colonization of the gut of infants delivered by C-section was delayed, with persisting differences in microbiota composition. In addition, infant immune function may be affected due to lack of exposure to microorganisms [145].

With weaning and the introduction of solid food, the next major community succession brings an increase in Bacteroidetes and Firmicutes, the dominant phyla of the adult gut [145]. One study tracked the developing gut microbiota of an infant, delivered vaginally, for the first 2.5 years of life and found, as one might expect, that changes in composition were associated with life events [156]. For instance, the early microbiota provided lactate utilization functions, and later additions provided functions for plant polysaccharide metabolism.

1.4.2 Diversity, variability, and individuality

Though the human gut harbours higher bacterial density than any environment, its diversity is low when compared to that of soil [12], with fewer bacterial phyla represented [325]. Generally, the dominant phyla in the human gut are by far the Bac-

teroidetes and Firmicutes, followed by a much smaller representation of the Proteobacteria, and then others [288]. Despite being from only a handful of phyla, it is estimated that more than one thousand species are present in the human gut [164], although there can be substantial differences between individuals [310]. As one might expect, the diversity of the gut mcrobiota is greatly affected by environmental factors; diet in particular is very important in influencing gut microbial diversity [55]; such changes may reflect the different metabolic specializations of microbial species [288], and there is evidence that certain taxa can be lost over time with a long-term diet that is low in fibre [284]. There have been efforts to try to classify the microbiota of individuals into groups, called "enterotypes" [8,338], although more recent work has acknowledged that discrete groups may not exist and that variation in the microbiota appears to be continuous [155].

Interestingly, one study attempted to use the microbiota from various body sites of individuals as an identifying "code", and found that the majority of microbiota codes collected from the same individuals 30-300 days later uniquely identified their host in a group of 120 people [93], suggesting remarkable potential stability of the microbiota within an individual. Such findings naturally lead to the question of whether host genotype can influence the composition of the microbiota. Although twin studies have had conflicting results and suggest that any effect of host genotype influence on the microbiota is likely small, more systematic studies in mice suggest that there are significant associations between variations in certain host loci and variation in microbial taxa, with most loci being involved in immunity and some in metabolism [288]. Future genome-wide association studies are required, treating the gut microbiota composition as a phenotype, to elucidate the relationship between variation in host genotype and variation in the gut microbiota.

Beyond environmental and host influences, there is still a substantial amount of variability in gut microbial composition that appears to be random [288], which may confound association studies. Rather than trying to assess variation by looking at taxonomic compositions, it may be more informative to focus on the functional composition. In one study, it was found that in lean individuals, despite a large variation in microbial community, there existed a core gut microbiota at the functional level, and deviations from this core were associated with obesity [309]. This emphasizes the importance of a function-based viewpoint with respect to studies of the human microbiome.

1.4.3 Mutualism between host and microbiota

The microorganisms comprising the microbiota have in the past often been described as "commensals", but such a label is misleading as more evidence suggests that host-microbiota interactions tend to be mutualistic in nature [12]. The microbiota in the gut possess a large arsenal of enzymes for breaking down complex polysaccharides in the human diet and they contribute about 10% of the calories that are absorbed [72]. Interestingly, differences in the gut microbiota between individuals can lead to differences in the capacity to obtain energy from ingested food [311], and in addition to contributing calories, the microbiota have also been shown to be involved in the promotion of fat storage host adipocytes [11]. Although such consequences may be undesirable in this age, they may have been very advantageous to our ancestors in earlier times when food was much more scarce.

While we do not necessarily need the additional calories provided by our resident microbes, colonocytes primarily use bacterially produced butyrate as an energy source, and in its absence, these cells suffer from an energy deficit that leads them to degrade their own cellular components for survival [67]. This illustrates the important mutualistic relationship that hosts have evolved over time with the microbiota, leading to dependence on microbial metabolites. In some cases, the host may even require metabolites; for example, germ-free mice raised without gut microbes require supplementation of vitamin K and some forms of vitamin B [130]. The gut microbiota produce metabolites that otherwise would not be circulating in the body and they also change the concentrations of some that are produced [332]. Interestingly, a number of metabolites that are predicted to be produced by the microbiota are currently used as drugs, suggesting that many of these metabolites may be bioactive [134]. While the vast number of small molecules produced by the microbes in the gut at high micromolar concentrations remain to be identified, some are likely to be relevant for pharmaceutical applications once their roles in human physiology are elucidated [66].

In addition to producing drug-like compounds, resident microbes may also affect orally ingested drugs, a fact that can lead to unexpected consequences in health care. In one study that examined urine metabolites of the widely used painkiller acetaminophen, it was found that there were differences between individuals in the ratio of two metabolites, acetaminophen glucuronide and acetaminophen sulfate, and the difference was attributed to bacterially produced compounds that compete for sulfonation in the gut [46]. Another study in which the efficacy of a statin used in the treatment of high cholesterol was examined, researchers found that differences in efficacy between individuals correlated with gut-derived metabolites [140]. Interestingly, a case in which the mechanism of drug metabolism was actually demonstrated was that for *Eggerthella lenta* inactivating the drug digoxin, which is used in the treatment of cardiac disease:

a strain of *E. lenta* was known to inactivate the drug in vitro, and RNA-seq analysis revealed that exposure to digoxin led to upregulation of an operon containing two genes predicted to be cytochromes capable of using digoxin as an electron acceptor, and that the presence of this operon was higher in the guts of individuals that showed a high level of inactivation of the drug [116]. These examples illustrate the important and likely under-appreciated influence of the gut microbiota on host drug metabolism.

1.4.4 Disease and the gut microbiota

Given that the host and the microbiota share such a close interaction, it would seem to follow that in some situations, they may be able to cause harm, and indeed, microbes with whom the host participates in mutualism can sometimes take on the role of pathogen [103]. In straightforward examples, opportunitistic pathogens may traverse through broken barriers in the host such as wounds in the skin or perforations in the lining of the gut [328]. Mutualistic organisms may also incidentally aid the virulence of pathogens by generating metabolites such as sugars [53], or perhaps by harbouring a reservoir of antibiotic resistance genes that can potentially spread to more serious pathogens, although evidence suggests there may be barriers to the general transfer of these genes among members of the microbiota [283].

Interestingly, there is a growing list of disease states that appear to be associated with a change in the composition of the gut microbiota. For example, in both Type I diabetic and obese individuals, the ratio of Firmicutes to Bacteroidetes has been shown to be altered. With obesity, there appears to be an increase in the relative abundance of Firmicutes [104]; the reverse is true for diabetes, in which there is both an increase in Bacteroidetes and a decrease in Firmicutes as children become autoimmune [311]. Though these descriptions of the changing gut microbiota are very broad, it has been

suggested that they may prove to be useful as diagnostic markers, for example, to identify infants at high risk for onset of Type I diabetes.

A number of other GI-related disease states have also been shown to be associated with changes in the microbiota, such as colorectal cancer [198,281], Type II diabetes [90, 322], and inflammatory bowel disease [92], including Crohn's Disease [78, 196]. Other non-gastrointestinal diseases have also implicated the microbiota, such as cardiovascular disease [129], allergies [96], multiple sclerosis [19], and neurodevelopmental disorders such as autism spectrum disorder [131]. There have also even been suggestions that the microbiota may be involved in behavioural or mood disorders [91]. The many health conditions in which the microbiota also vary are perhaps not surprising as the microbes in the gut have been recognized for their importance in host immunity [246]. The interaction between an individual and their gut microbiota is a complex one, in which both partners may influence the other. Though fascinating, the exact relationship between certain disease states and their altered microbiota remains to be elucidated [83].

1.5 Challenges in metagenomics and microbiome research

The gut microbiome has recently become a hot topic in the popular media even as the scientific community struggles to understand the specifics of how the microbiota contribute to human health and disease. Challenges in the metagenomics and microbiome fields tend to fall into two broad areas.

1.5.1 Correlation versus causation

The use of antibiotics leading to reduced diversity in the gut microbiota has been blamed by some for many different conditions [23] and while there is likely some truth to the idea that widespread antibiotic use has had broad unexpected consequences, more work is required to tease apart the many factors that contribute to complex diseases. Furthermore, although certain disease states appear to be associated with a change in the composition of microbiota based on 16S rRNA gene sequencing, the exact mechanisms will need to be determined before causality can be ascribed. Even gene function-based analyses of sequence data [189, 234], although extremely useful for generation of hypotheses, are in themselves merely correlative as there are many factors that can influence the expression of genes in a given system, including physical linkage to other genes as well as environmental and cell-to-cell interactions [2]. Knowledge pertaining to these levels of regulation will need to be integrated for the generation of meaningful and biologically relevant models of the microbiota.

Thus, a current challenge in metagenomics and microbiome research is moving beyond survey-type, correlation studies, and incorporating methods that allow causality to be determined [2], including biochemistry, genetics, and, generally, controlled hypothesis-driven experiments [256], for example using enrichment cultures or cultures of a subset of the microbiota. Recent efforts to array cultured isolates from the human gut microbiota combined with culture of these microbes in gnotobiotic mice [109] allow for tractable, combinatorial approaches to systematic identification of organisms or groups of organisms that result in specific phenotypes in the host [82]. These types of methods will likely be critical in determining whether, in which direction, and to what extent these relationships are causal.

1.5.2 Informatics and sequence data annotation

The Human Microbiome Project, along with other large-scale sequencing-heavy projects, illustrate the power of today's high-throughput, low-cost sequencing technology in aiding our study of these previously underappreciated microbial communities, as well as making such studies feasible for smaller laboratories. However, a 2011 review discussed the limitations of the current shotgun sequencing approach [293], arguing that genomes could only be assembled for the most dominant members of a complex community, citing previous work in the Sargasso Sea [315], and that the probability of capturing rare organisms, such as methylotrophs, is low [223].

The generation of large amounts of sequence data across many different labs leads to many practical issues not discussed here but which include requiring an optimized/standardized work pipeline, large quantities of computer memory as well as databases, high-quality analytical tools, and trained bioinformaticians [292]. Beyond these issues lies an additional hurdle which must happen after obtaining genomes or

metagenomes from a sequencing project: the functional annotation of genes. The research community has recognized the need for easily accessible and user-friendly computational tools to aid in the analysis of metagenomic sequence data, and many stand-alone or web-based tools and databases, such as MG-RAST [211, 334], have become available. To carry out automated functional assignments, these software use homology-based annotation, comparing metagenomic sequences to existing protein and nucleotide databases.

One obvious pitfall in a sequence homology-based strategy is genes that are similar in function but dissimilar in sequence to known genes cannot be annotated. Furthermore, the case may very likely be that we currently simply have not amassed enough sequences of known function to be able to accurately and thoroughly annotate new sets of sequences. For example, in a 2007 dataset of 480 Mb of gut metagenomic sequence data and a predicted 660,000 genes from 13 individuals [161], more than one-half of predicted genes could not be assigned to a Cluster of Orthologous Groups (COG) [300,301] and therefore could not be given a functional assignment. Indeed, a 2015 US-initiated call for a Unified Microbiome Initative has emphasized the need for characterizing genes with currently unknown function [2]. Although there are computational approaches to improve functional annotation of genes, such as inference of gene function from operon rearrangements [217], it is becoming increasingly necessary to complement sequence-based approaches with high-throughput approaches that provide proof-of-function for genes, to obtain the information necessary to carry out functional annotation. To identify novel genes whose functions may not be predicted from their sequence alone, a functional metagenomic approach can be used.

1.6 Functional metagenomics

In general, use of the term "functional metagenomics" implies a very specific function-based "wet-lab" methodology, herein described. Although the term is occasionally co-opted to mean something different – for example, to mean sequence-based metagenomics with a focus on gene function [62, 244] or even completely redefined to mean the study of functional members of the microbiota that influence human health [183] – such uses are rare in the scientific literature. In this section, a brief introduction to the overall methodology and its advantages is provided, setting the context for subsequent chapters of this thesis, in which various aspects of the functional metagenomic approach are described in greater detail.

1.6.1 General methodology

Functional metagenomics is an experimental approach that involves isolating DNA from microbial communities to study the functions of proteins encoded by that DNA, typically through cloning DNA fragments, expressing genes in a surrogate host, and screening for enzymatic activities of interest. Using such a function-based approach can be powerful for the discovery of novel enzymes whose functions could not have been predicted based on DNA sequence alone. New information from function-based analyses can then be used to annotate genomes and metagenomes derived solely from sequence-based analyses. In this way, functional metagenomics complements sequence-based metagenomics, analogous to how molecular genetics of model organisms has provided knowledge of gene function that has been widely applicable in genomics.

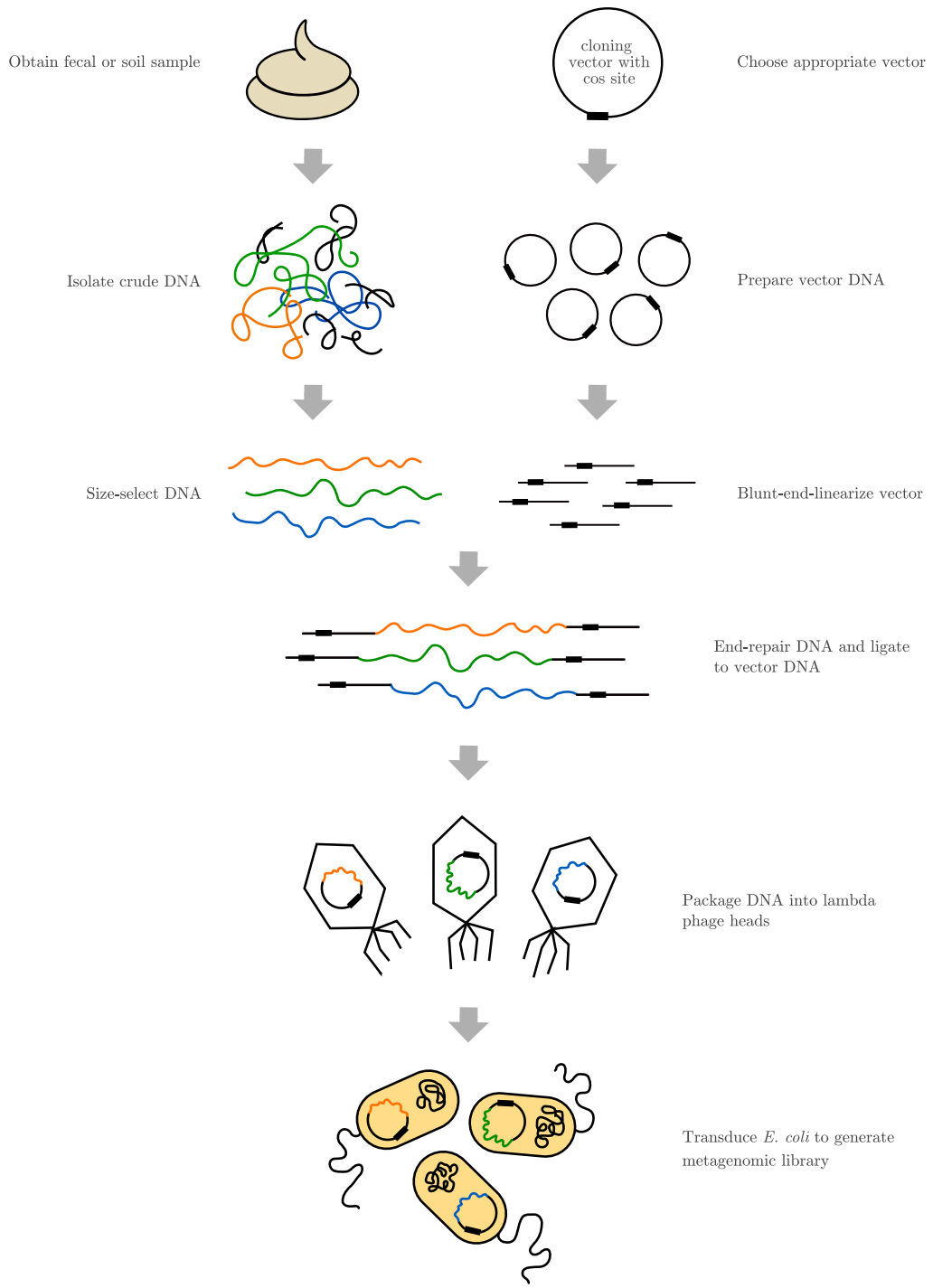


Figure 1.1: Summary of metagenomic library construction. Steps involved in the construction of a metagenomic library, from original environmental sample to the final library in the *E. coli* host. Adapted from [166].

Functional metagenomics begins with the construction of a metagenomic library, the steps of which are summarized in Figure 1.1. Cosmid- or fosmid-based libraries are preferred due to their large and consistent insert size and high cloning efficiency. DNA is extracted from the environmental sample of interest, such as soil or feces. After DNA is extracted from the sample, it is then size-selected through pulsed-field gel electrophoresis to enrich for high-molecular weight fragments. The fragments are subsequently end-repaired and ligated to a linearized and blunt-ended *cos*-based vector. The ligation mixture is then packaged into λ phage heads through recognition of the *cos* site, and the phage are used to transduce *E. coli* to generate the metagenomic library (Figure 1.1). The library contains relatively large insert DNA, typically 25 to 40 kb for *cos*-based vectors. There are two major advantages to using a *cos*-based vector and phage transduction to construct clone libraries: the high efficiency of transduction as well as the reduced likelihood of insert concatemers.

Once the metagenomic library has been constructed in *E. coli*, functional screening can be carried out. In the most straightforward approach, screening of the library can be done in the same *E. coli* host in which library construction took place. For example, to isolate clones conferring antibiotic resistance genes, the host cells can simply be plated on selective media containing antibiotics (Figure 1.2). This example, while simple, has been useful for exploring the antibiotic resistance gene reservoir harboured by our gut microbiota. Interestingly, in one study, it was found that resistance genes isolated through a culture-independent approach were substantially more novel compared to those that had been isolated through an aerobic culture-dependent approach, with on average, ∼61% versus ∼90% identity at the nucleotide level to the best hit in Genbank, respectively [283]. As this example illustrates, functional screening in *E. coli* can be productive, although there may be limitations.

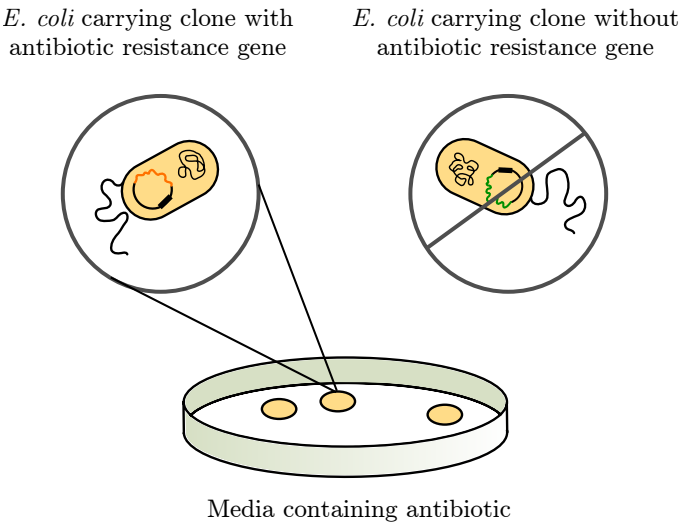


Figure 1.2: Example of a functional screen in *E. coli*. The library in *E. coli* is plated onto media with antibiotics to select for library clones that confer resistance.

Screening in hosts other than the *E. coli* library host, however, may provide additional hits from functional screens due to possible differences in elements required for gene expression between the original organism and *E. coli*. Though it is arguably difficult to quantify, one estimate of how much of the metagenome is accessible by screening in *E. coli* is 40%, based on analysis of 32 genomes from different bacteria and archea, counting ORFs with ribosome-binding sites and promoters that would be recognized in *E. coli* [97]. The fraction of "inaccessible" genes depends of course on the particular environmental DNA sample. Regardless, to address this problem, metagenomic libraries can be transferred from the *E. coli* library host to other surrogate hosts that may be more suitable for screening; this may be done efficiently using conjugation or, if the recipient species is amenable, transformation or electroporation. The issue of possible barriers to transcription and translation in *E. coli* is a particularly important methodological limitation in functional metagenomics and will be discussed in greater detail below and in subsequent chapters.

1.6.2 The power of a function-based approach

In this section, several examples from the scientific literature have been specifically chosen to highlight the strengths of a functional metagenomics approach.

Avoiding sequence-based biases

Functional metagenomics offers an avenue to finding novel proteins by functional enrichment or selection of metagenomic material. For example, one study identified three clones from activated sludge and soil samples that each carried novel genes of a *luxI-luxR*-type quorum sensing system [119]: when these gene sequences were compared to the NCBI protein database, the novel *luxI* and *luxR* genes had only ∼30-50% similarity to known *lux* proteins. It may be difficult to predict the function of genes with such low sequence similarity, illustrating the utility of a function-based approach.

In another study, the authors screened soil libraries containing a total of 3.6×10^9 bp for antibiotic resistance genes, and identified clones conferring resistance to ampicillin, gentamicin, chloramphenicol, and trimethoprim [307]. Of particular interest was the discovery of a novel trimethoprim resistance gene. Trimethoprim inhibits the enzyme dihydrofolate reductase (DHFR), and resistance to it is most commonly conferred by a mutant DHFR. However, the authors found that their trimethoprim resistance gene was very different from known *dhfr* genes; from biochemical analyses, it was found to be distinctly different in its mechanism and properties, and was therefore deemed to represent a novel group of DHFRs. Furthermore, its closest matches were to reductases involved in lipid metabolism, not *dhfr* genes, illustrating that function cannot always be surmised from sequence alone. Currently, we simply may not have enough data to functionally annotate new sequences accurately.

Enrichment of desired sequences

Not only can functional selections find novel proteins, they can also greatly reduce the sheer quantity of genetic material to be sequenced. In one study, a high-throughput functional metagenomic approach was used to find enzymes in the human gut involved in dietary fiber catabolism, reducing the amount of metagenomic DNA to be sequenced from 5.4×10^9 bp to 8.4×10^5 bp, a reduction of almost four orders of magnitude, simply by selecting for the growth of library clones on different polysaccharides [299]. Using this approach, the authors identified 73 carbohydrate-active enzymes, corresponding to a five-fold enrichment in the target-gene identification over random sequencing. If enrichment can be performed prior to sequencing, a great deal of time and resources can be saved, not to mention the value of having experimental data regarding function.

High-throughput functional screening strategies

In addition to straightforward functional screens, it is possible to design more complex screens that can still be high-throughput. An example of such a screen was one carried out to identify metagenomic clones that could modulate NF-κB activity in human intestinal epithelial cells [164]. NF-κB is a transcription factor involved in immunity and inflammation in the gut. Using a reporter system in human cells, they screened over 2,600 clones and identified 171 clones that either up- or down-regulated NF-κB in human cells. They went on to analyze one stimulatory clone, using transposon mutagenesis to identify two genes necessary for the stimulatory effects. These genes were predicted to encode a permease and putative lipoprotein, which allowed the authors to surmise a putative mechanism for the clone's modulatory activity. Again, there is an important feedback loop to be appreciated here: functional annotations help function-based studies, which in turn help future functional annotations, and so on.

1.6.3 Important considerations

These several examples illustrate the wide applicability of functional screens. There are important considerations, however, in undertaking a functional metagenomic approach. First, consideration must be given to choosing an appropriate environment for the desired target genes; for instance, a rumen sample from a grass-fed cow may be ideal for generating a metagenomic library that is enriched with genes encoding enzymes for cellulose degradation [108]. Second, an appropriate vector must be selected for the library backbone, and the choice depends on various factors, such as whether a small-insert or large-insert library is desired, and in the former case, whether expression vectors would be advantageous to help drive gene expression in *E. coli* [141]. Third, surrogate host(s) other than *E. coli* may be considered, for either an attempt to increase the hit rate [302, 312] or for the complementation of specific phenotypes [320]. Alternative expression hosts that have been used include *Agrobacterium tumefaciens*, *Caulobacter vibrioides*, *Rhizobium leguminosarum*, *Ralstonia metallidurans*, *Pseudomonas fluorescens*, *Pseudomonas putida*, *Xanthomonas campestris*, *Burkholderia graminis*, *Sinorhizobium meliloti*, and *Bacillus subtilis* [1, 50, 186, 254, 302, 308, 312].

Finally, other logistics in the screening strategy have to be considered, such as whether to pool clones for screening or to instead keep clones arrayed and carry out individual clone screening; in the latter case, the achievable throughput must be very carefully considered because, depending on the particular screen, clone-by-clone screening may not be a feasible strategy, although the design of automated microfluidic screening strategies is an exciting area of development [47, 313]. There are of course limitations and biases in this method [71], as there are with all methods. Nevertheless, functional metagenomics is a powerful experimental strategy that can help improve our understanding of the mechanisms that underlie biological phenomena as well as aid in the functional annotation of the exponentially increasing number of metagenomes.

1.7 Thesis outline

This thesis centres on methods to aid in the determination of gene function. The objective of this work was to advance the methods used in functional metagenomics research, through both the analysis of existing techniques as well as the development of new strategies and systems for functional screening. The results of this work are presented in four data chapters, each of which concerns a specific method or system:

- Chapter 3 evaluates the feasibility of using a pooled method for sequencing large-insert metagenomic clones. A set of 92 clones, isolated from various functional screens, was sequenced using Illumina in two ways: first, experimentally as a pool, and second, individually using barcodes. The latter was done to generate reference data for evaluation of the former pooled strategy. The results from pooled sequencing were analyzed for their accuracy and completeness to determine whether such a strategy was worthwhile.

- Chapter 4 explores the sequence bias of a human gut metagenomic library, particularly the point at which bias is introduced during the cloning process. The metagenomic DNA was sampled and sequenced at three points during library construction, and the sequence data were analyzed for bias and potential causes.

- Chapter 5 describes the development of *B. theta* as a host for screening of metagenomic libraries constructed from gut-derived DNA. A species from the Bacteroidetes phylum was chosen to help combat the likely barrier to transcription that may limit hit rates when screening gut metagenomic libraries in *E. coli*, as well as to open the door to new possibilities of phenotypes that can be complemented. This chapter describes the modification of vectors for use in *B. theta*; the generation of *B. theta*-compatible clone libraries, including genomic as well

as metagenomic libraries; and, importantly, the successful proof-of-principle functional complementation of a *B. theta* polysaccharide degradation mutant using a human gut metagenomic library.

- Chapter 6 concerns the transcriptional terminators that were designed into the *B. theta*-compatible vector that was constructed for Chapter 5. This chapter provides the rationale for including the terminators; describes the design, synthesis, and cloning of the fragment carrying the terminators; and presents the results of testing the functionality of the terminators.

Though each data chapter above concerns a distinct topic, all are explorations of various aspects of the function-based approach. Together, the work described in this thesis furthers knowledge of the methods and techniques currently used in functional metagenomics as well as those that may potentially be used in the future of this field.

Chapter 2

General materials and methods

2.1 Acknowledgements and declarations

I acknowledge the following contributions:

- Methods and techniques for the construction of metagenomic libraries, summarized in Section 2.6, were generously shared by **Jiujun Cheng** and technical advice was provided by **Katja Engel**.
- Protocols and technical assistance for pulsed-field gel electrophoresis, described in Section 2.5.9, were provided by **Katja Engel** and **Lee Pinnell**.
- This chapter was proofread by my supervisor **Trevor Charles**.

2.2 Strains, plasmids, and oligonucleotides

2.2.1 Bacterial strains

All *E. coli* and *Bacteroides* strains used in this study are summarized in Table 2.1. Genotypes and descriptions as well as literature references where applicable are provided for each strain. All strains can be found in the Charles Lab main frozen culture collection. *B. theta* strains were archived as 25% glycerol stocks and *E. coli* strains were archived as either 25% glycerol or 7% DMSO stocks.

2.2.2 Plasmids

All plasmids used in this study are summarized in Table 2.2. Descriptions and literature references where applicable are provided for each plasmid. All plasmids can be found in the Charles Lab *E. coli* frozen culture collection.

2.2.3 Oligonucleotide sequences

All oligonucleotides used in this study are summarized in Table 2.3. Descriptions and DNA sequences are provided for each. Oligos were synthesized by either Integrated DNA Technologies, Inc. or Bio Basic Inc. Lyophilized DNA was dissolved to a concentration of 100 µM and stored at -20℃.

Table 2.1: Bacterial strains used in this study.

Strain	Genotype or description	Ref./Source
E. coli DH5α	F- *supE44* Δ*lacU169 hsdR17 recA1 endA1 gyrA96* (Nx^R) *thi-1 relA1* (Φ80*lacZ* ΔM15)	[21]
E. coli EPI300	F- *mcrA* Δ*(mrr-hsdRMS-mcrBC)* Φ80d*lacZ* ΔM15 Δ*lacX74 recA1 endA1 araD*139 Δ(*ara, leu*)7697 *galU galK* λ- *rpsL* (Sm^R) *nupG trfA dhfr*	Epicentre
E. coli HB101	F- *mcrB mrr hsdS20*(rB- mB-) *recA13 leuB6 ara-14 proA2 lacY1 galK2 xyl-5 mtl-1 rpsL20* (Sm^R) g*lnV44* λ-	[25]
E. coli S17-1	F- *recA thi pro hsdR rspL* (Sm^R) RP4-2-Tc::Mu-aphA::Tn7 (Km^S)	[88, 271]
E. coli S17-1 λ-pir	λ lysogen of S17-1, providing pir protein required for plasmids with R6K origin of replication	[271]
B. fragilis NCTC 9343	*Bacteroides fragilis* type strain; same as ATCC 25285	[126]
B. theta VPI-5482	*Bacteroides thetaiotaomicron* type strain; same as ATCC 29148	[339]
B. theta BtUW24	VPI-5482 carrying deletion of *tdk* (BT_2275)	[159]
B. theta BtUW25	BtUW24 carrying deletion of *anSME* (BT_0238); *anSME* is also known as *chuR*	[17]
B. theta BtUW1	VPI-5482 *thrC*::pKNOCK-*bla-tetQ*; threonine single recombinant auxotroph in BT_2401	This study
B. theta BtUW2	VPI-5482 *trpD*::pKNOCK-*bla-tetQ*; tryptophan single recombinant auxotroph in BT_0530	This study
B. theta BtUW3	BtUW25 carrying presumably integrated clone from BT3 genomic library designated chuR1	This study
B. theta BtUW4	BtUW25 carrying presumably integrated clone from BT3 genomic library designated chuR2	This study

Continued on next page

Table 2.1 – *Continued from previous page*

Strain	**Genotype/description**	**Ref./Source**
B. theta BtUW5	BtUW25 carrying presumably integrated clone from BT3 genomic library designated chuR3	This study
B. theta BtUW6	BtUW25 carrying presumably integrated clone from BT3 genomic library designated chuR4	This study
B. theta BtUW7	BtUW25 carrying presumably integrated clone from BT3 genomic library designated chuR5	This study
B. theta BtUW8	BtUW25 carrying presumably integrated clone from BT3 genomic library designated chuR6	This study
B. theta BtUW9	BtUW25 carrying presumably integrated clone from BT3 genomic library designated chuR7	This study
B. theta BtUW10	BtUW25 carrying presumably integrated clone from BT3 genomic library designated chuR8	This study
B. theta BtUW11	BtUW25 carrying presumably integrated clone from BT3 genomic library designated chuR9	This study
B. theta BtUW12	BtUW25 carrying presumably integrated clone from BT3 genomic library designated chuR10	This study
B. theta BtUW13	BtUW25 carrying presumably integrated clone from BT3 genomic library designated chuR11	This study
B. theta BtUW14	BtUW25 carrying presumably integrated clone from CLGM3 metagenomic library designated chuR1	This study
B. theta BtUW15	BtUW25 carrying presumably integrated clone from CLGM3 metagenomic library designated chuR2	This study
B. theta BtUW16	BtUW25 carrying presumably integrated clone from CLGM3 metagenomic library designated chuR3	This study

Continued on next page

Table 2.1 – *Continued from previous page*

Strain	**Genotype/description**	**Ref./Source**
B. theta BtUW17	BtUW25 carrying presumably integrated clone from CLGM3 metagenomic library designated chuR4	This study
B. theta BtUW18	BtUW25 carrying presumably integrated clone from CLGM3 metagenomic library designated chuR5	This study
B. theta BtUW19	BtUW25 carrying presumably integrated clone from CLGM3 metagenomic library designated chuR6	This study
B. theta BtUW20	BtUW25 carrying presumably integrated clone from CLGM3 metagenomic library designated chuR8	This study
B. theta BtUW21	BtUW25 carrying presumably integrated clone from CLGM3 metagenomic library designated chuR9	This study
B. theta BtUW22	BtUW25 carrying presumably integrated clone 5B2 from arrayed CLGM3 metagenomic library; EPI300 clone from Plate 5 Row B, Well 2	This study
B. theta BtUW23	BtUW25 carrying presumably integrated clone 5B9 from arrayed CLGM3 metagenomic library; EPI300 clone from Plate 5 Row B, Well 9	This study

Table 2.2: Plasmids used in this study.

Plasmid	Description	Ref.
R751	Mobilizer plasmid used for triparental matings; Tp^R	[137, 212]
pRK2013	Mobilizer plasmid used for triparental matings; ColEI origin and Km^R (Nm^R)	[124]
pRK600	Derivative of pRK2013; Km^R::Tn9; Cm^R	[89]
pHC79	Cosmid vector derived from pBR322	[127]
pJC8	Cosmid vector with RK2 origin of replication; Genbank accession KC149513	[43]
pAFD1	*E. coli-Bacteroides* shuttle vector with pUC origin of replication; received from Nadja Shoemaker	[249]
pKNOCK-*bla-tetQ*	*B. theta* suicide vector with *E. coli* R6K *ori*; Ap^R in *E. coli*; Tc^R in *B. theta*	[200]
pJET1.2	Vector for blunt end PCR product cloning kit (Thermo Fisher K1231); Genbank accession EF694056	[194]
pCC1FOS	Copy-number inducible fosmid vector; Genbank accession EU140751	Epicentre
pKL1	pAFD1 with *cos* sequence cloned in the BamHI site using BglII fragment from pHC79; see Figure 5.8	This study
pKL2	pKL1 with polylinker between the EcoR1 and KpnI sites (EcoR1-NotI-Eco72I-NdeI-KpnI linker); see Figure 5.8	This study
pKL3	pKL2 with gentamicin resistance stuffer cloned as Eco72I fragment from pJC8; see Figure 5.8	This study
pKL4	pCC1FOS with gentamicin resistance stuffer cloned as Eco72I fragment from pJC8; see Figure 5.12	This study
pKL5	pKL4 with RK2 *oriT* from pJC8 cloned in the HindIII site; see Figure 5.12	This study
pKL6	pKL5 with *ermF-repA* fragment from pKL8 cloned in the EcoRI site; see Figure 5.12	This study

Continued on next page

Table 2.2 – *Continued from previous page*

Plasmid	Description	Ref.
pKL7	pKL6 with removal of the Eco72I stuffer carrying the gentamicin resistance gene; see Figure 5.12	This study
pKL8	pJET1.2 with *ermF-repA* PCR product amplfied from pAFD1	This study
pKL9	pJET1.2 with synthesized transcriptional terminator (TT) fragment; sequence verified	This study
pKL10A	pKL7 with TT fragment blunt-end cloned in the Eco72I site of pKL7; note that this clone has deletion of a single base A from *ilvGEDA* terminator sequence; see Figure 6.5	This study
pKL10B	pKL7 with TT fragment blunt-end cloned in the Eco72I site of pKL7, in reverse orientation to pKL10A; note that this clone has deletion of a single base A from *ilvGEDA* terminator sequence	This study
pKL11	pKL10 with the Eco72I stuffer removed; note that this plasmid was constructed prior to determining that pKL10A had a deletion of a single base A from the *ilvGEDA* terminator sequence	This study
pKL13	pKL7 with TT fragment blunt-end cloned in the Eco72I site of pKL7; see Figure 5.12	This study
pKL14	pKL13 with removal of the Eco72I stuffer carrying P_{tac} and gentamicin resistance gene; see Figure 6.9	This study
pKL15	pKL13 with GFPuv cloned in as PacI-SgsI fragment; see Figure 6.7	This study
pKL16	pKL15 with removal of the PacI-NheI fragment containing the transcriptional terminator (*ilvGEDA* TT) by double digestion, blunting, and ligating; see Figure 6.7	This study
pKL17	pKL13 with flipped Eco72I stuffer, so that P_{tac} driving transcription in the opposite orientation to pKL13; see Figure 6.7	This study
pKL18	pKL17 with GFPuv cloned in as CpoI-SfaAI fragment; see Figure 6.7	This study

Continued on next page

Table 2.2 – *Continued from previous page*

Plasmid	**Description**	**Ref.**
pKL19	pKL18 with removal of the NsiI-CpoI fragment containing the transcriptional terminator (*rnpB* T1 TT) by double digestion, blunting, and ligating; see Figure 6.7	This study
pKL20	pKL14 with gentamicin resistance stuffer cloned as Eco72I fragment from pJC8; see Figure 6.9	This study
pKL21	pKNOCK-*bla-tetQ* with ∼600 bp *thrC* fragment (BT_2401) cloned as SalI-KpnI fragment; see Figure 5.15A	This study
pKL22	pKNOCK-*bla-tetQ* with ∼350 bp *trpD* fragment (BT_0530) cloned as SalI-KpnI fragment; see Figure 5.15A	This study
BT2	random clone from BT1 genomic library; see Table 3.7	This study
BF4	random clone from BF1 genomic library; see Table 3.7	This study
PO3	random clone from CLGM1 metagenomic library; see Table 3.7	This study
CLGM3 5B2	*chuR* complementing clone from CLGM3 metagenomic library	This study
CLGM3 5B9	*chuR* complementing clone from CLGM3 metagenomic library	This study

Table 2.3: Oligonucleotides used in this study.

Oligo	Description	Sequence (5' to 3')
KL10	Oligo 1 to generate EcoRI-NotI-Eco72I-NdeI-KpnI polylinker	AATTCGCGGCCGCCACGTGCA TATGGGTAC
KL11	Oligo 2 to generate EcoRI-NotI-Eco72I-NdeI-KpnI polylinker	CCATATGCACGTGGCGGCCGC G
KL12	F primer to amplify ∼800 bp containing RK2 *oriT* from pJC8, with HindIII adaptor	CCT AAGCTT TCGGTCTTGC CTTGCTCGTCGG
KL13	R primer to amplify ∼800 bp containing RK2 *oriT* from pJC8, with HindIII adaptor	CCT AAGCTT GCGCTTTTCC GCTGCATAACCC
KL14	F to amplify ∼4 kb containing *ermF*-IS4351-ori-*repA* from pAFD1, with EcoR1 adaptor	CCT GAATTC ACTTTTGTGC AATGTTGAAGATTAGTAATTC TATTC
KL15	R to amplify ∼4 kb containing *ermF*-IS4351-ori-*repA* from pAFD1, with EcoR1 adaptor	CCT GAATTC ATAACAGCCG GTGACAGCCGGC
KL16	Primer walking round #2 of *ermF*-IS4351-ori-*repA* fragment (#1 used KL14)	GTTCAACCAAAGCTGTGTCGT TTTCAATAGC
KL33	Primer walking round #3 of *ermF*-IS4351-ori-*repA* fragment	CAGGTATGCCAAACGTGGTTC TAAAAATGC
KL42	Primer walking *ermF*-IS4351-ori-*repA* fragment; check second A of round #2 results	GGAACTGCAAAATTCCTAAAA TCACAACC
KL43	Primer walking round #4 of *ermF*-IS4351-ori-*repA* fragment	CAAGCCCGTCAGGGCGCGTCA GCGGGTGTTGG
KL44	Check orientation of 778 bp *oriT* in *B. theta* compatible pCC1FOS derivatives	GGATCCTCTAGAGTCGACCTG CAGGCATGC
KL45	Primer walking round #5 of *ermF*-IS4351-ori-*repA* fragment	AACAGACAAAGCCGTTTATAA AGGACTTGC
KL46	Primer walking round #6 of *ermF*-IS4351-ori-*repA* fragment	GTCAGCAACAAAGGTAGTACT TTATTATCG

Continued on next page

Table 2.3 – *Continued from previous page*

Oligo	Description	Sequence (5' to 3')
KL47	F primer for GFPuv ORF +50 base upstream, with PacI adapter	CCT TTAATTAA TGCATGCC TGCAGGTCGACTCTAGAGGAT CCCC
KL48	R primer for GFPuv ORF +100 base downstream, with SgsI adapter	CCT GGCGCGCC CGCGCGAG ACGAAAGGGCCCGTACGGCCG
KL49	F primer for GFPuv ORF +50 base upstream, with CpoI adapter	CCT CGGACCG TGCATGCCT GCAGGTCGACTCTAGAGGATC CCC
KL50	R primer for GFPuv ORF +100 base downstream, with SfaAI adapter	CTCCT GCGATCGC CGCGCG AGACGAAAGGGCCCGTACGGC CG
KL51	Sequence TT fragment primer 1	GGCAAATTGGCGATGGAGCCG ACTTTTAGC
KL52	Sequence TT fragment primer 2	TATTTGCAGTACCAGCGTACG GCCCACAG
KL53	Sequence TT fragment primer 3	ATCCTGCCACGTCGCCCGTTA CACCGGACC
KL54	Sequence TT fragment primer 4	TCAGAAGGAAGGTCCAGTCGG TCATGCCTTTGC
KL55	Sequence TT fragment primer 5 (for pKL10A)	AATCTTCAACATTGCACAAAA GTGAATTCG
KL56	Sequence TT fragment primer 6 (for pKL10A)	GATAACAATTTCACACCCTAA GGCACGTGG
KL57	Sequence TT fragment primer 7 (for pKL10B)	ATTGCACTCCACCGCTGATGA CATCAGTCG
KL58	Sequence TT fragment primer 8 (for pKL10B)	AAATCCTGTATATCGTGCGAA AAAGGATGG
KL59	Sequence TT fragment primer 9 (for pKL9B)	CATTCGTATTGCACGACATTG CACTCCACC

Continued on next page

Table 2.3 – *Continued from previous page*

Oligo	Description	Sequence (5' to 3')
KL60	Sequence TT fragment primer 10 (for pKL9B)	CCTACAACGGTTCCTGATGAG GTGGTTAGC
KL61	F primer for *B. theta chuR* ORF (BT_0238)	ATGAAAGCAACAACTTATGCA CCTTTTGCCAAACC
KL62	R primer for *B. theta chuR* ORF (BT_0238)	TTAATATTCTATTTTTAAACT TCCGTCTTTTAGTGCTTTC
KL63	F primer for primer for *B. theta chuR* ORF (BT_0238) 300 bp upstream	TCTCCATCCCTCAAAGTCTTC AGATATAACATTTTTCC
KL65	R primer for primer for *B. theta chuR* ORF (BT_0238) 300 bp upstream	TAACCGCAGTGATGGTTAGTC AGGATCAAGC
KL66	Sequence *chuR* ORF from CLGM chuR5, toward ORF start (nt 265 relative to *B. theta* sequence)	GGGCGTATTTCTTTTGCAGCT CCATCG
KL67	Sequence *chuR* ORF from CLGM chuR5, toward ORF start (nt 222 relative to *B. theta* sequence)	AAGCGGACGCATCAGCGTTTC TCCACC
KL68	Sequence *chuR* ORF from CLGM chuR5, toward ORF end (nt 1006 relative to *B. theta* sequence)	TCGGAACAATGAAATACCAAT CACTCC
KL69	Sequence *chuR* ORF from CLGM chuR5, toward ORF end (nt 1058 relative to *B. theta* sequence)	TCTATTTGCCTGCAACGGAGA ATGTCC
thrCIDMF (SalI)	F primer for amplifying *B. theta* ∼600-bp *thrC* fragment (BT_2401) with SalI adapter, designed by Eric Martens	GCGGTCGACGAGATTGCTTAT CGGGTAGCC
thrCIDMR (KpnI)	R primer for amplifying *B. theta* ∼600-bp *trpD* fragment (BT_2401) KpnI adapter, designed by Eric Martens	GCGGGTACCACACAAATCACG GCATTATCGG
trpDIDMF (SalI)	R primer for amplifying *B. theta* ∼350-bp *trpD* fragment (BT_0530) KpnI adapter, designed by Eric Martens	GCGGTCGACGGAAATGCGGGT TCCGGTTG

Continued on next page

Table 2.3 – *Continued from previous page*

Oligo	**Description**	**Sequence (5' to 3')**
trpDIDMR (KpnI)	R primer for amplifying *B. theta* ∼350-bp *trpD* fragment (BT_0530) KpnI adapter, designed by Eric Martens	GCGGGTACCGAATGTACGTAC CGCCAATCC
JC102	F sequencing primer for pJC8 [43]	TAACAATTTCACACAGGAAAC AGCTATGAC
JC103	R sequencing primer for pJC8 [43]	GCGATTAAGTTGGGTAACGCC AGGGTTTTC
KL-JC102	F sequencing primer for *B. theta* compatible fosmid; see Figure 6.4	TAACAATTTCACACAGGAAAC AGCTATGACG
KL-JC103	R sequencing primer for *B. theta* compatible fosmid; see Figure 6.4	GCGATTAAGTTGGGTAACGCC AGGGTTTTCG

2.3 Bacterial culture

2.3.1 Growth media

All recipes for media and solutions are provided in Appendix A. The following sections describe methods used for *E. coli* molecular biology work; for *B. theta* methods, see Section 5.6 of Chapter 5. *E. coli* was routinely grown at 37℃ using LB, with shaking at 200 rpm. For cultures to be used for alkaline lysis-based minipreps, *E. coli* was grown in either LB or TB media.

2.3.2 Antibiotics

Antiobiotics used in the culture of *E. coli* are summarized in Table 5.6. Concentrations for antibiotics are denoted using the abbreviation (see Table 2.4) followed by the concentration as a subscript; for example ampicillin at 100 µg/ml would be Ap_{100}. Note that antibiotic concentrations were halved when used in liquid media.

Table 2.4: Antibiotic concentrations used for *E. coli*.

Antibiotic	Abbrev.	Solvent	Final conc. agar
ampicillin	Ap	water	100 µg/ml
chloramphenicol	Cm	ethanol	10 µg/ml
gentamicin	Gm	water	25 µg/ml
kanamycin	Km	water	25 µg/ml
nalidixic acid	NA	water; add NaOH drops to dissolve	10 µg/ml
tetracycline	Tc	ethanol	10 µg/ml
trimethoprim	Tp	DMSO	400 µg/ml

2.4 DNA introduction and extraction methods

2.4.1 Calcium chloride-based competent cell preparation

Competent cell preparation was based on the protocol from Sambrook and Russell [251]. The desired strain was streaked from frozen stock onto LB agar with antibiotic selection, if possible (e.g., EPI300 was streaked onto LB Sm_{200}). A single colony was used to inoculate a liquid overnight culture, using the same antibiotic selection. The overnight culture was used to inoculate liquid LB media, without antibiotics, at a volume ratio of 1:200. The culture was grown to OD_{600} ∼0.9 [298], as measured on a Spectronic Spec 20D spectrophotometer (warmed up for at least 15 minutes). The culture flask was chilled on ice for ∼30 minutes to halt cell growth. All subsequent work was performed on ice to keep the cells cold at all times.

Cells were collected by centrifugation in polyethylene centrifuge bottles at 6,000×g at 4℃ for 10 minutes, using a rotor/adapter that was chilled at 4℃ for several hours. The supernatant was decanted and the cells were gently resuspended in 0.1 M $CaCl_2$ (chilled overnight at 4℃), at a ratio of approximately 1 volume per 2-3 volumes of overnight culture equivalent. The cells were again collected by centrifugation at 6,000×g at 4℃ for 10 minutes and the supernatant was decanted. The cells were then gently resuspended in the same volume of chilled 0.1 M $CaCl_2$ and incubated for several hours on ice or overnight on ice at 4℃. Cells were again pelleted at 6,000×g at 4℃ for 10 minutes, using a rotor/adapter that had been chilled at 4℃. The supernatant was decanted, the bottle was pop-spun, and all remaining supernatant was carefully removed. Cells were gently resuspended using 0.1 M $CaCl_2$ 15% glycerol (v/v; chilled overnight at 4℃) in a volume equal to 1.5% of the original culture volume. Cells were frozen at -80℃ in 0.2 or 1 ml aliquots.

2.4.2 Calcium chloride-based transformation

Calcium chloride-based transformation was based on the protocol from Sambrook and Russell [251]. Cells were thawed from -80℃ on ice, with periodic gentle flicking of the tube. DNA was mixed with cells in a microfuge tube, not exceeding a volume ratio of 1:10. The mixture was incubated on ice for 30 minutes, then heat-shocked at 42℃ for 90 seconds, followed by immediate transfer to ice for 1-2 minutes. 1 ml of LB was added, and cells were allowed to recover at 37℃ for 1 hour without shaking. Cells were pelleted by centrifugation at 8,000-13,000×g for 1 minute. The supernatant was decanted, leaving ∼100 µl to resuspend the cells for spreading onto selective agar plates.

2.4.3 Plasmid DNA miniprep

Home-made kit for routine plasmid preps

This protocol and the recipes for the solutions used in this protocol were obtained from the OpenWetWare version of the commercial Qiagen QIAprep Spin Miniprep Kit. Please see Section A.7 for the solution recipes.

Overnight cultures of *E. coli* were prepared using 3-5 ml LB or 2-3 ml TB with the appropriate antibiotics and supplementation. 2-5 ml of culture was pelleted in a 2-ml microfuge tube, and resuspended in 250 µl of Solution P1. 250 µl of the alkaline Solution P2 was added, and the tube was inverted ∼10 times to lyse the cells. 250 µl of Solution N3 was added and the tube was inverted ∼10 times to neutralize the mixture. Cell debris was pelleted by centrifugation at 21,000×g for 5-7 minutes. The supernatant containing the plasmid DNA was transferred to a silica spin column (BioBasic SD5005), the column was pop spun for ∼5 seconds at 13,000×g, and the flow-through was discarded. If the strain carrying the plasmid was not an *endA1* mutant, then

500 µl of PB wash solution was pop spun through the column to remove contaminating nucleases, and the flow-through was discarded. The column was then washed at least 2 times with 500-750 µl of PE wash solution by pop spinning and discarding the flow-through. As much ethanol wash as possible was removed by gentle tapping of the tube containing the flow-through onto a paper towel, and the column was spun for 2 minutes at 13,000×g. The spin column was transferred to a new microfuge tube, and 50 µl of $T_{10}E_{0.1}$ (pH 8.5) was added to the column. DNA was eluted by centrifugation at 10,000×g for 30 seconds. Miniprepped plasmid DNA was quantified using the Nanodrop ND-1000 Spectrophotometer.

Commercial kits for DNA sequencing

For samples intended for DNA sequencing, plasmid DNA was prepared using commercial miniprep kits according to the manufacturer's recommendations. Kits used were the EZ-10 Spin Column Plasmid DNA Mini-preps Kit (BioBasic BS614), the GeneJET Plasmid Miniprep Kit (Thermo-Fisher K0502), or the QIAprep spin miniprep kit (Qiagen 27106). Miniprepped plasmid DNA was quantified using the Nanodrop ND-1000 Spectrophotometer.

2.4.4 Plasmid DNA maxiprep

Large-scale preparations of plasmid DNA were based on the protocol from Charles, 1990 [35]; see Appendix A.9 for the solution recipes. All centrifugation steps were carried out at room temperature.

The desired strain were streaked from frozen stock onto LB agar with the appropriate antibiotics. A single colony was used to inoculate 5 ml liquid overnight culture, using the same antibiotic selection. The 5-ml overnight was then used to seed an

overnight 1 L culture, using the same antibiotic selection. The following day, the cells were pelleted by centrifuging at 7,000×g for 10 minutes, such that there were two cell pellets with the equivalent of 500 ml of culture each. Each pellet was resuspended in 10 ml TEG and pooled for 20 ml.

The cells were then lysed by the addition of 40 ml ALS followed by inversion ∼10 times. The mixture was neutralized with 30 ml HSS followed by inversion ∼10 times, and cooled at -70℃ for 20-30 minutes. The debris was pelleted by centrifuging at 10,000×g for 10-15 minutes, and the solution was decanted through cheesecloth into a fresh 250-ml centrifuge bottle. 90 ml of isopropanol was added to the solution to precipitate the DNA, followed by by centrifuging at 10,000×g for 10 minutes. The supernatant was discarded and the bottle was inverted on a paper towel to dry the pellet. The pellet was resuspended in 8 ml TE and the mixture was transferred to 40-ml centrifuge tube. 4 ml of 7.5 M NH_4Ac was added and mixed, and proteins were allowed to precipitate on ice for 15-30 minutes. Protein was pelleted by centrifuging at 10,000 × g for 10-15 minutes and the supernatant was transferred to a new tube. 12 ml isopropanol was added to the solution to precipitate the DNA, followed by centrifuging at 10,000×g for 10-15 minutes. The supernatant was discarded and the bottle was inverted on a paper towel to dry the pellet.

The pellet was resuspended in 800 µl TE and transferred to two microcentrifuge tubes, with 400 µl per tube. To each tube, 4 µl of 5 M NaCl and 5 µl of 10 mg/ml RNase A was added, followed by incubation at 37℃ for 30 minutes. 2.5 µl of 20% SDS and and 5 µl of 19.2 mg/ml Proteinase K was added, followed by incubation at 37℃ for 30 minutes. The mixture was then extracted with an equal volume of phenol-chloroform (1:1) and then extracted with an equal volume of only chloroform. To precipitate the DNA, 25 µl of 5 M NaCl and 500 µl isopropanol were added. The precipitated DNA was carefully removed with a pipette and dipped into 70% ethanol to wash and placed

into a new tube, with the precipitate from both tubes being combined. The DNA was allowed to dry, and then resuspended in 1 ml TE, and dissolved overnight at 4℃. To quantify, the DNA was diluted 1-in-10 and 1-in-100; 25 µl of these dilutions was quantified using the Nanodrop ND-1000 Spectrophotometer as well as run on a gel to confirm the concentrations. Typically, plasmid maxipreps can be obtained with concentrations ∼1 µg/µl.

2.4.5 HMW DNA extraction from fecal samples

Prior to DNA extraction, fecal samples were pre-processed based on the method described by Lee and Hallam [175], by placing 5 g of sample in a mortar with 1 ml of denaturing solution (4 M guanidine isothiocyanate, 10 mM Tris-HCl [pH 8.0], 1 mM EDTA, 0.5% beta-mercaptoethanol). The sample was frozen using liquid nitrogen, ground with a pestle to a homogeneous powder, then transferred to a conical tube for storage at -80°C.

DNA was extracted from soil or feces according to the method described by Zhou et al. [347]. Briefly, 5 g of soil or fecal sample were incubated in 13.5 ml of extraction buffer (100 mM Tris [pH 8.0], 100 mM EDTA, 100 mM sodium phosphate [pH 8.0], 1.5 M NaCl, 1% CTAB), with the addition of proteinase K (to 75 µg/ml), shaking at 37°C for 30 minutes. After adding SDS (to 2% w/v in 15 ml), the sample was incubated at 65°C for 2 h with gentle inversions every 15 minutes. After centrifugation at 6,000×g for 10 minutes at room temperature, the supernatant was collected, extracted with chloroform:isoamyl alcohol (24:1), and DNA was precipitated with 0.6 volumes of isopropanol at room temperature for 1 h. DNA was collected by centrifugation at 6,000× g for 20 minutes at room temperature, followed by a 70% ethanol wash. The DNA pellet was suspended overnight at 4°C in 0.5-3 ml of TE buffer (10 mM Tris-HCl

[pH 8.0] and 0.1 mM EDTA [pH 8.0]). The DNA was quantified by gel electrophoresis, using bacteriophage λ DNA as a standard (see Section 2.5.8).

2.4.6 HMW DNA extraction from pure cultures

DNA was isolated from liquid bacterial cultures based on a method described by Charles and Nester [36]. Briefly, cells were cultured in 50 ml of liquid media, and the cell pellets were recovered after centrifugation at 7000×g for 5 minutes at room temperature. Cells were washed with 8 ml of wash buffer (10 mM Tris [pH 8.0], 25 mM EDTA [pH 8.0], 150 mM NaCl), and resuspended in 4 ml of buffer (10 mM Tris [pH 8.0], 25 mM EDTA). The following were added, to a final volume of 5 ml: NaCl (to 0.5 M), proteinase K (to 0.5 mg/ml), and lysozyme (to 2.5 mg/ml). After incubation at 37°C for 30 minutes with shaking, 250 µl of 20% SDS were added, the mixture was incubated at 65°C for 60 minutes, then centrifuged at 6,000×g for 10 minutes at room temperature. The supernatant was collected, and protein was precipitated with 0.5 volumes of 7.5 M ammonium acetate on ice for 20 minutes. The mixture was centrifuged at 10,000×g for 15 minutes, the supernatant was collected and centrifuged at 8,500×g for 10 minutes to further clear the supernatant. The supernatant was decanted and the mixtured was extracted with chloroform in a 1:1 volume. The supernatant was collected and DNA was precipitated with 1 volume of isopropanol at room temperature for 30 minutes. DNA was spooled out, dipped in a 70% ethanol wash, and placed in a microfuge tube. The tube was centrifuged at 15,000×g for 1 minute, the supernatant was removed, and the pellet was allowed to dry. Finally, the pellet was allowed to dissolve in 2 ml of TE overnight at 4°C. The DNA was quantified by gel electrophoresis, using bacteriophage λ DNA as a standard (see Section 2.5.8).

2.5 DNA manipulation methods

2.5.1 Gel electrophoresis

Routine gel electrophoresis was carried out using TAE buffer; see Appendix A.6 for the 50× TAE stock recipe. The stock was diluted to 1× in 20-L working volumes and stored at room temperature for use. A concentration of 0.8% or 0.85% agarose was used to visualize bands greater than 10-20 kb, including genomic DNA preparations; 1.0% agarose was used for fragments ranging between 500 and 10,000 bp; and on the rare occasion, 2% agarose was used to visualize small bands, typically less than a few hundred basepairs. Gels were typically run using 5 V/cm. Commercial molecular ladders were used for size estimation: 25-50 ng of either the λ-HindIII Ladder or the 1-kb DNA Ladder (Thermo-Fisher FERSM0101 and FERSM0311, respectively). For visualization on the UV transilluminator, Gel Red stain was used; contrary to the manufacturer's recommendations, the stain was diluted 50,000× rather than 10,000×.

2.5.2 Ethanol precipitation

Ethanol precipitation was used to concentrate DNA or to change the buffer in which the DNA was dissolved. Ions were added in the form of either 1/10 volume of 3 M sodium acetate (pH 5.2), 1/50 volume of 5 M sodium chloride, or 1/2 volume of 7.5 M ammonium acetate. The solution was mixed, and alcohol was added in the form of either 3 volumes of ethanol or 1 volume of isopropanol. DNA was chilled either on ice or at -20℃ for 10-60 minutes, and centrifuged at 21,000×g for 10-30 minutes. The superntant was removed, the tube was pop spun, and the remaining supernatant was carefully removed. 100 µl of 70% ethanol was washed over the pellet and immediately removed. The pellet was allowed to dry with the tube inverted on a Kim Wipe for a

few minutes until the edges of the pellet began to become translucent. The DNA was dissolved in a small volume of TE buffer, typically 10-20 µl.

2.5.3 Gel extraction

This protocol is based on the Qiagen QIAquick Gel Extraction Kit, using a home-made binding buffer recipe [149]. Please see Appendix A.8 for the solution recipes.

The sample of DNA was run on an 1× TAE agarose gel, using the appropriate agarose concentration and 1 mM guanosine [111]. The desired fragment was excised, placed in a microfuge tube and weighed on an analytical balance. Binding buffer was added to the fragment, using 3 or 4 µl per mg of gel; for example, 300-400 µl for a 100-mg gel fragment. The gel was dissolved by incubating at 65℃ with frequent inverting and vortexing. After dissolution, the mixture was transferred to a silica spin column (BioBasic SD5005), the column was pop spun for ∼5 seconds at 13,000×g, and the flow-through was discarded. The column was then washed at least 2 times with 500 to 750 µl of PE wash solution by pop spinning and discarding the flow-through. As much ethanol wash as possible was removed by gentle tapping of the tube containing the flow-through onto a paper towel, and the column was spun for 2 minutes at 13,000×g. The spin column was transferred to a new microfuge tube, and 30-50 µl of $T_{10}E_{0.1}$ (pH 8.5) was added to the column. DNA was eluted by centrifugation at 10,000×g for 30 seconds. Extracted DNA was quantified using the Nanodrop ND-1000 Spectrophotometer.

2.5.4 Restriction enzyme digestion

Routine restriction enzyme digestion was carried out using the FastDigest line of enzymes from Thermo-Fisher Scientific, using the FastDigest universal Green Buffer with loading dye included. Digestion conditions were generally modified from the manufac-

turer's recommendations, herein described. Restriction digestion was either carried out on a larger scale to prepare DNA for cloning (Table 2.5) or on a smaller scale to confirm the results of cloning (Table 2.6). Enzyme volumes were not allowed to exceed 10% of the total reaction volume. Digests were either used directly for cloning after heat inactivation, or were purified by silica column using the protocol for gel extraction (see Section 2.5.3) with a 3-4:1 volume ratio of binding buffer to digest.

Table 2.5: General digestion recipe for cloning purposes.

DNA	∼1-3 µg
FastDigest enzyme (1U/µł)	1-3 µl
10× FastDigest Green Buffer	3-6 µl
sterile dH_2O	top up
Total	30-60 µl

Table 2.6: General digestion recipe for diagnostic purposes.

DNA	∼50-100 ng
FastDigest enzyme (1U/µł)	0.5 µl
10× FastDigest Green Buffer	1 µl
sterile dH_2O	top up
Total	10 µl

2.5.5 Ligation

Routine ligations were carried out in 10-15 µl volumes, using T4 DNA Ligase (ThermoFisher L0014) or Fast-Link DNA Ligase (Epicentre LK0750H) according to the manufacturer's recommendations. Sticky-end ligations were incubated for 1-3 hours at room temperature whereas blunt-end ligations were incubated overnight either at 16℃ or room temperature.

2.5.6 Estimation of digestion and dephosphorylation efficiency

The following outlines how to estimate the digestion and dephosphorylation efficiency for a large-scale preparation of vector for library construction. It is recommended that this be performed after purification of the backbone from the stuffer (by either gel extraction or electroelution) to test the integrity of the DNA for ligation, that is, ensuring that the ends of the DNA are ligatable.

First, the large-scale digestion and dephosphorylation was set up as in Table 2.7, using non-FastDigest Eco72I and FastAP (Themo-Fisher R0361 and F0651, respectively). The reaction was incubated for 3.5 hours at 37℃, heat-inactivated for 30 minutes at 80℃, and stored at -20℃.

Table 2.7: Recipe for large-scale digest and desphosphorylation.

vector DNA	100 µg
10× Tango Buffer	100 µl
Eco72I	30 µl
FastAP	30 µl
sterile dH_2O	top up
Total	1000 µl

After digestion and dephosphorylation, the mixture was assessed for cutting and dephosphorylation efficiency; reactions were set up as summarized in Table 2.8 using T4 polynucleotide kinase (Thermo-Fisher EF0651) and typically reactions were set up in duplicate. Reactions were incubated for 45 minutes at room temperature (not 37℃ specifically), followed by addition of 0.25 µl Fast-Link ligase (Epicentre LK0750H), and overnight incubation at 16℃. The mixtures were then used to transform home-made EPI300 competent cells.

Table 2.8: Recipes for assessment of digestion and dephosphorylation efficiency

	$-PNK + ligase$	$+PNK - ligase$	$+PNK + ligase$
DNA, dig. and dephos.	1	1	1
10× FL biffer	1	1	1
ATP, 10 mM	0.5	0.5	0.5
T4 PNK	0	0.5	0.5
H_2O	7.5	7	7
Total	10 µl	10 µl	10 µl
No. transformants	x	y	z

After transformation, colonies were counted (Table 2.8) and the efficiency of digestion and dephosphorylation were estimated using the two equations below. Typically, digestion efficiency was 97% and desphosphorylation efficiency was 99%.

$$\% \text{ of vector DNA that is cut} = \left(1 - \frac{y}{z}\right) \times 100$$

$$\% \text{ of cut vector that is desphosphorylated} = \left(1 - \frac{x - y}{z}\right) \times 100$$

2.5.7 Sanger DNA sequencing

For routine Sanger sequencing, samples were typically submitted to The Centre for Applied Genomics (Toronto) or BioBasic Inc. (Markham).

2.5.8 Gel quantification of genomic and metagenomic DNA

Both genomic and metagenomic DNA were quantified by agarose gel electrophoresis against a dilution series of commercial λ DNA (Thermo-Fisher FERSD0011; 300 ng/µł). For high-molecular-weight DNA species that may form a somewhat heterogeneous mixture, the Nanodrop ND-1000 Spectrophotometer may not be as inaccurate as quantification on an agarose gel.

A series of λ DNA dilutions was prepared to use as standards: 0, 5, 10, 25, 50, 75, and 100 ng. The standards were run on a 0.8% or 0.85% agarose gel pre-strained with Gel Red (Section 2.5.8), along with varying volumes of the sample(s) to be quantified, e.g., 0.1 and 0.9 µl (Figure 2.1A). Using the free software ImageJ [257], pixel intensity was quantified for the standards and samples (Figure 2.1B). A line of best fit was generated for the data points from the λ DNA standard, which was then used to estimate the concentration of DNA for the experimental sample(s) (Figure 2.1C).

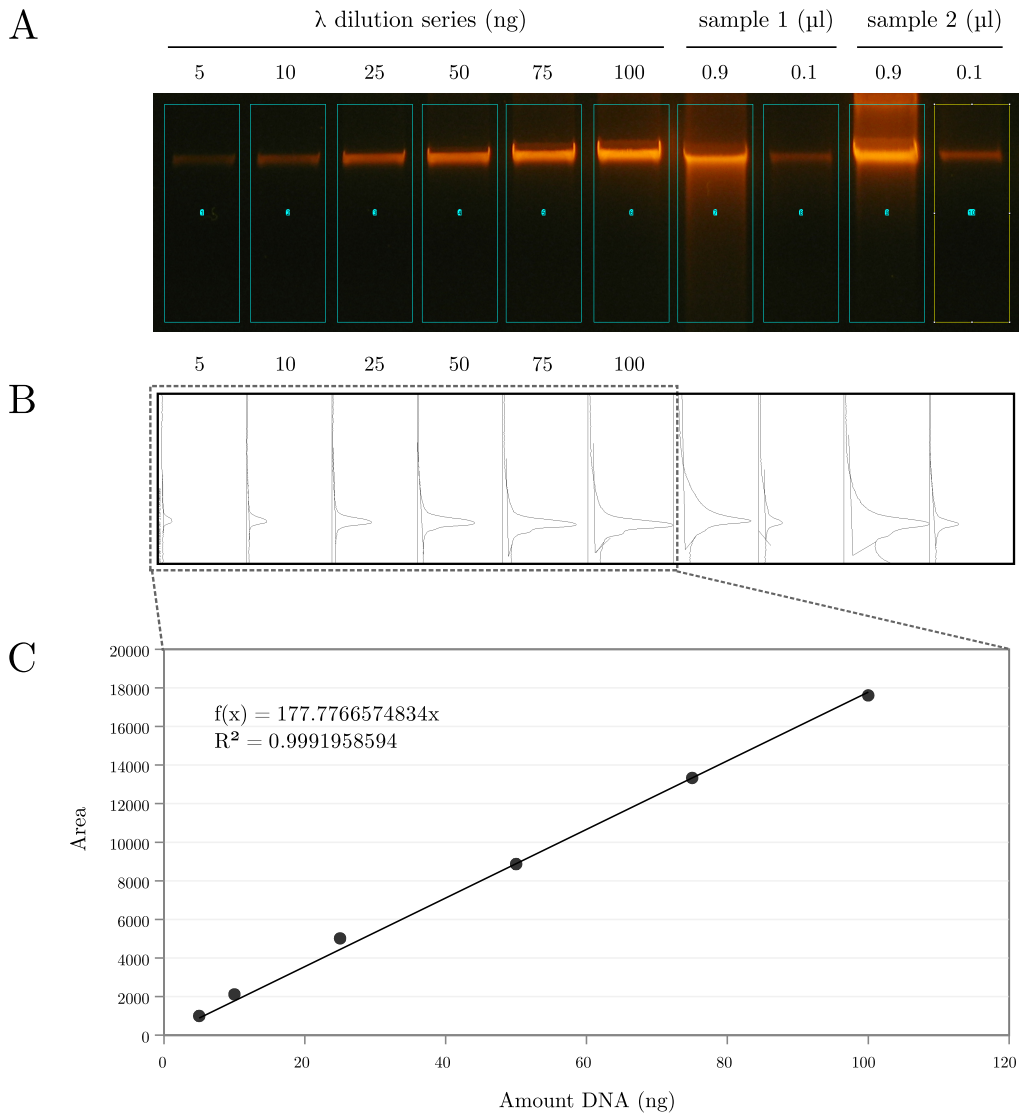


Figure 2.1: Gel quantification of high-molecular-weight DNA samples using λ DNA dilution standards. (A) Samples of unknown concentration are run on a gel against the λ standard. (B) ImageJ used to quantify pixel intensity in the selected lanes. (C) Pixel intensity for the λ standard is plotted and a line of best fit is generated.

2.5.9 Pulsed field gel electrophoresis

Pulsed-field gel electrophoresis was used to visualize/separate high-molecular-weight DNA fragments. The following section describes the protocol and parameters for electrophoresis as well as the preparation of λ DNA-based ladders.

Pulsed-field gel electrophoresis using Bio-Rad CHEF MAPPER

Gels were prepared using pulsed-field certified agarose (Bio-Rad 1620137) at 1% agarose in 100 ml 1× TAE buffer. The gel rig was filled with 1× TAE, the parameters on the Bio-Rad CHEF Mapper were set (Table 2.9), and the buffer was circulated to cool to 14℃. The cooling was stopped, the circulation was paused, the gel was was placed in the rig, and samples were loaded; DNA extracts were either run for diagnostics (500 ng) or for size-selection by excision (30 µg). The circulation was resumed followed by the cooling, and the run was allowed to proceed overnight (Table 2.9).

The next day, the gel was post-stained. For diagnostic gels, post-staining was done in 200 ml of 1× TAE buffer supplemented with 20-25 µl of Gel Red stain diluted 1-in-5 in dH_2O, shaking gently at room temperature for 1-2 hours; the gel was then rinsed in buffer, destained in 200 ml of buffer for 15-60 minutes, and visualized on a UV transilluminator. For excision gels, only the edges of the gel were stained and the fragment was excised without exposure to either Gel Red stain or UV/blue light (see Figure 2.2).

Table 2.9: Settings for pulsed-field gel electrophoresis on Bio-Rad CHEF Mapper.

Parameter	Diagnostic gel	Excision gel
input DNA range	10-100 kb	10-100 kb
calibration factor	1.0	1.0
buffer	0.5× TBE*	0.5× TBE*
temperature	14℃	14℃
agarose	1%	1%
voltage	6 V/cm	5 V/cm
pulse	1-10 s	0.5-8.5 s
ramping factor	linear	linear
runtime	16 h	14 h

Preparation of λ DNA molecular markers for pulsed-field electrophoresis

Commercial λ DNA (Thermo-Fisher FERSD0011; 300 ng/µł) was used to prepare home-made molecular weight markers for use in pulsed-field gel electrophoresis. The size of the λ genome is 48.5 kb. λ DNA was self-ligated using T4 DNA ligase (Thermo-Fisher FEREL0014) to generate concatemers appropriate for assessing the size range of crude DNA extracts: ∼50 kb, ∼100 kb, ∼150 kb, etc. The recipe for the self-ligation reaction is provided in Table 2.10. To generate a marker at ∼25 kb, λ DNA was digested with XbaI, which halves the 48.5-kb genome. The recipe for the digestion reaction is provided in Table 2.11.

*setting used although buffer was 1× TAE

The ligation and digestion mixtures were used to make a combined working ladder. λ-ligated and λ-XbaI were diluted to 5 ng/µl and 2.5 ng/µl, respectively, with loading dye added. For electrophoresis, 75-100 ng of the combined ladder was used; Figure 2.2 depicts the use of this combined ladder as a guide to excise a gel fragment, particularly in comparison with a commercial ladder whose largest marker is 40 kb (Invitrogen 10511-012).

Table 2.10: Ligation recipe for self-ligated λ DNA.

λ DNA (300 ng/µl)	33.3 µl
T4 DNA ligase	3 µl
10× T4 DNA Ligase Buffer	10 µl
sterile dH_2O	53.7 µl
Total	100 µl (100 ng/µl)

Table 2.11: Digestion recipe for XbaI-digested λ DNA.

λ DNA (300 ng/µł)	33.3 µl
FastDigest XbaI	10 µl
10× FastDigest Green Buffer	10 µl
sterile dH_2O	46.7 µl
Total	100 µl (100 ng/µl)

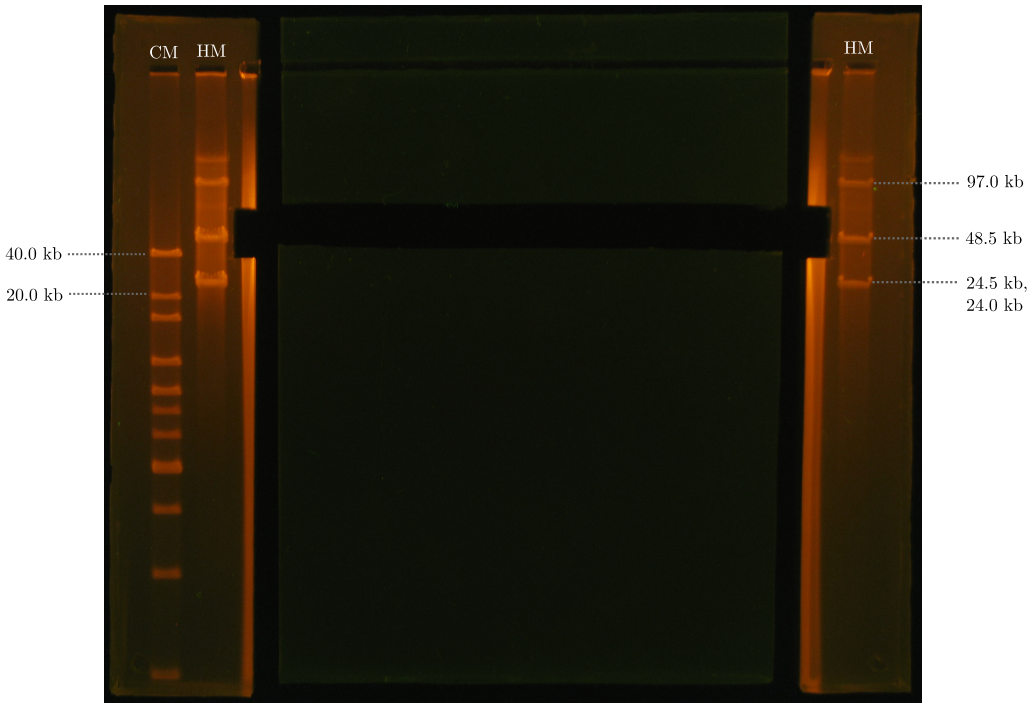


Figure 2.2: Pulsed-field gel electrophoresis using home-made λ DNA markers. CM: commercial marker, 1 kb Extension Ladder (Invitrogen 10511-012); HM: home-made λ marker, containing XbaI-digested λ and ligated λ DNA.

2.5.10 Electroelution

Preparation of dialysis tubing

Dialysis tubing (Sigma D-9652) was cut in forearm-length segments and immersed in 2% sodium bicarbonate, 1 mM EDTA. The tubing was boiled for 10 minutes, taking care to keep the tubing submerged. The tubing was then removed and thoroughly rinsed with distilled water straight from the tap, using three rinses outside and three inside. The tubing was immersed in 1 mM EDTA, boiled for another 10 minutes, and then transferred to 1 mM EDTA, 20% ethanol. All air trapped air bubbles were removed and the tubing was stored at 4℃ . Typically, ∼10 segments of tubing were prepared at a time; the tubing will keep for years in the storage solution.

Electroelution

The DNA to be electroeluted was run on either a typical agarose gel (for example, 100 µg of digested vector DNA) or a pulsed-field agarose gel (for example, 30 µg of crude extract DNA from feces), and the desired fragment was excised from the gel. The fragment was placed inside a segment of dialysis tubing that was previously thoroughly rinsed with distilled water and equilibrated to room temperature in 1× TAE. One end of the tubing was clamped, the same buffer was used to fill the tubing, and the other end was clamped (Figure 2.3A). The tubing was submerged in 1× TAE in the gel rig, and the DNA was eluted using ∼3 V/cm for 3 hours (Figure 2.3B). The buffer inside the tubing was then decanted into a sterile conical tube; the bag was rinsed twice with 2-3 ml of 1× TAE, and that buffer also retained, for a total volume of less than 50 ml. The mixture of DNA was subsequently concentrated using a 30 kDa Amicon centrifugal filter (Millipore UFC903024), followed by a standard ethanol precipitation (see Section 2.5.2).

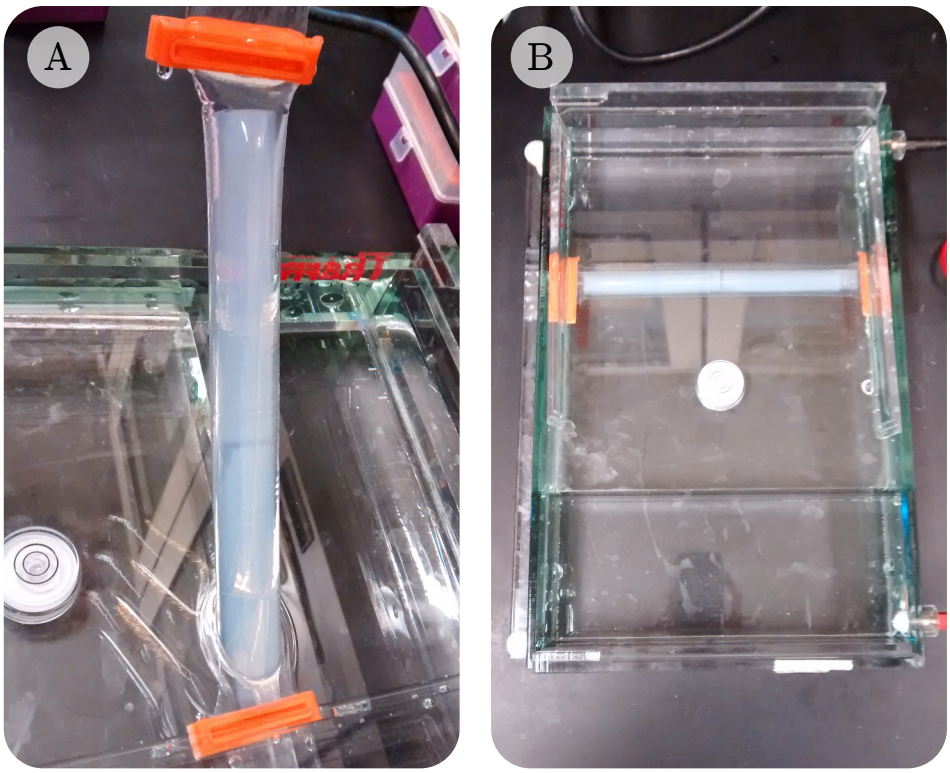


Figure 2.3: Setup of apparatus for electroelution. (A) Gel fragments containing desired the DNA are excised and placed in dialysis tubing with buffer. (B) The fragment is subjected to an electric field and the DNA migrates into the buffer contained in the dialysis tubing.

2.6 Summary of constructed libraries

Several genomic and metagenomic libraries were constructed in this study; protocols for library construction are provided in the specific materials and methods section of each chapter. Table 2.12 summarizes the details for each library: the library name, the source of the DNA, the vector used, the *E. coli* library host used for transduction, the approximate number of unique clones, and the estimated average insert size.

Table 2.12: Genomic and metagenomic libraries constructed in this study.

Library name	**DNA source**	**Vector**	**Host**	**No. clones**	**Estimated avg. insert size**
BT1	*B. theta* genomic DNA	pJC8	HB101	8,000	27 ±8 kb (n=17)
BF1	*B. frag* genomic DNA	pJC8	HB101	18,000	30 ±7 kb (n=18)
CLGM1	pooled human feces	pJC8	HB101	42,000	28 ±9 kb (n=36)
BT2	*B. theta* genomic DNA	pKL3	HB101	15,000	nd
CLGM2	pooled human feces	pKL3	HB101	65,000	nd
BT3	*B. theta* genomic DNA	pKL13[†]	EPI300	36,000	nd
CLGM3	pooled human feces	pKL13[†]	EPI300	115,000	26 ±10 kb (n=19)

[†]Eco72I stuffer fragment not purified from backbone prior to ligation; see Section 5.6.9

Chapter 3

Evaluation of pooled Illumina sequencing for metagenomic clones

3.1 Acknowledgements and declarations

The work presented in this chapter was published as a Research Article in the journal **PLOS ONE**. I was the primary author of this article. The citation for the article is:

> Lam KN, Hall MW, Engel K, Vey G, Cheng J, Neufeld JD, Charles TC (2014) Evaluation of a pooled strategy for high-throughput sequencing of cosmid clones from metagenomic libraries. *PLOS ONE* 9:e98968. doi:10.1371/journal.pone.0098968

I managed and performed all experiments/analyses described in this chapter with the exception of the following:

- In Section 3.4.1, I outlined and oversaw analyses carried out by **Michael Hall** to calculate sequencing read depth and extent of *E. coli* contamination for the samples. Mike Hall generated the read depth images in Appendix B.1.
- In Section 3.4.2, I and **Katja Engel** outlined and oversaw analyses carried out by **Greg Vey** and **Michael Hall** to estimate coverage from pooled sequencing.
- In Section 3.6.5 and Section 3.6.6, the management of samples for sequencing was organized by **Katja Engel**, who was then Project Manager for CM^{2}BL-related projects.
- In Table 3.7, the majority of the 92 DNA samples was prepared by **Jiujun Cheng**. **Cveta Manassieva** and **Tanya Romantsov** also contributed samples for sequencing.
- In Table 3.6, in addition to the CLGM1 human gut library I constructed, other libraries used were constructed by **Jiujun Cheng** as well as a previous lab member **Chunxia Wang**.

I also acknowledge the following contributions:

- The criticism of one **anonymous reviewer** led me to perform an all-by-all clone sequence similarity analysis in Section 3.4.3 that revealed an important caveat of our pooled sequencing approach.
- The text of the PLOS ONE manuscript, largely duplicated here, was proofread and edited by **Katja Engel**, **Josh Neufeld**, and **Trevor Charles**.

3.2 Abstract

High-throughput sequencing methods have been instrumental in the growing field of metagenomics, with technological improvements enabling greater throughput at decreased costs. Nonetheless, the economy of high-throughput sequencing cannot be fully leveraged in the sub-discipline of functional metagenomics. In this area of research, environmental DNA is typically cloned to generate large-insert libraries from which individual clones are isolated, based on specific activities of interest. Sequence data are required for complete characterization of such clones, but the sequencing of a large set of clones requires individual barcode-based sample preparation; this can become costly, as the cost of clone barcoding scales linearly with the number of clones processed, and thus sequencing a large number of metagenomic clones often remains cost-prohibitive.

This chapter investigates a hybrid Sanger/Illumina pooled sequencing strategy that omits barcoding altogether, and evaluates the strategy by comparing the pooled sequencing results to reference sequence data obtained from traditional barcode-based sequencing of the same set of clones. Using identity and coverage metrics, the results show that pooled sequencing can generate high-quality sequence data, without producing problematic chimeras. Though caveats of a pooled strategy exist and further optimization of the method is required to improve recovery of complete clone sequences and to avoid circumstances that generate unrecoverable clone sequences, our results demonstrate that pooled sequencing represents an effective and low-cost alternative for sequencing large sets of metagenomic clones.

3.3 Introduction

With the advent of high-throughput sequencing, metagenomics has emerged as a powerful way to explore DNA recovered from terrestrial, aquatic, and host-associated microbial communities. Sequence-based metagenomics involves bulk sequencing of environmental DNA and has generated a wealth of genome information from myriad environmental samples. With this wealth of sequence data serving as a foundational resource, the stage is set for function-based metagenomics, or functional metagenomics, which is arguably essential for the recovery and annotation of hypothetical proteins with as-yet-unknown functions [117, 242].

3.3.1 Sanger-based sequencing of metagenomic clones

Functional metagenomics allows exploration of the densely populated microbial habitats that are rich resources for the discovery of novel enzymes. Applying this approach, the genetic material of the microbial community is extracted from an environmental sample, and the DNA is cloned into appropriate vectors to generate metagenomic libraries that are maintained using *E. coli* as a surrogate host. These libraries may then be subjected to function-based activity screens, either in *E. coli* or various other surrogate hosts, after which positive clones are isolated for analysis.

A critical step in functional metagenomic studies is obtaining DNA sequence for the isolated clones in order to identify the gene(s) responsible for the function(s) of interest, particularly if the goal is to identify novel enzymes. Prior to the existence of high-throughput sequencing, it was, and still is, common to use other methods to identify the gene or operon carried on the insert DNA. One strategy is to Sanger-sequence the clone to obtain a sequence fragment, by primer-walking along the insert [86, 136, 270, 307] or first subcloning smaller fragments of the insert that carry the

activity of interest [20,85,105,135,186,190,230,236,237,259]. A variant of this strategy is to use transposon mutagenesis, which may be followed by screening for loss of activity [3, 52, 73, 119, 164, 169, 254, 282, 318, 329]. Regardless of the specific strategy, multiple steps are usually required to obtain sequence data for large-insert clones.

3.3.2 High-throughput sequencing of clones using barcodes

Although current high-throughput sequencing methods are an appropriate scale for sequencing of microbial genomes, the throughput is typically far greater than required for coverage of single clones. This has led to the practice of "multiplexing", which involves combining multiple clones for sequencing, using DNA barcodes (or indexes) to track sequence reads from individual clones within the larger set (Figure 3.1, Barcoded Sequencing). Examples of this strategy include the sequencing of large-insert clones identified from screens for enzymes involved in dietary fibre catabolism [299], prebiotic breakdown [34], and cellulosic biomass conversion [108]. Barcoded sequencing enables sequence data recovery from many clones simultaneously, yet the cost of barcoding every clone can be several-fold higher than the cost of the sequencing itself. This sample preparation cost can be a bottleneck for the smaller molecular microbiology lab, where isolating clones is relatively easy, but sequence analysis of the clones becomes cost-prohibitive.

3.3.3 Aims of this work

Our lab investigated the possibility of circumventing the barcoding step by testing a clone pooling and sequencing approach (Figure 3.1, Pooled Sequencing). As part of this sequencing strategy, end sequences for every clone are generated by Sanger-sequencing; these sequences are called "end-tags" to describe their role in the downstream sequence

retrieval process in which we match clones to next-generation sequence data assemblies. In a pooled method, clones are sequenced together and users rely on the post-sequencing assembly process to generate contigs that represent individual clones. After assembly, contigs exist in a pool; to retrieve a specific clone's contig, the clone's end-tags are used to query the pool.

A set of 92 large-insert clones was chosen for this analysis; cosmid clones were isolated previously by different members of the lab from various functional screens. End-tags were obtained from Sanger sequencing each clone and, concurrently, the clones were pooled for sequencing and assembly. Though the reduced cost of pooled sequencing is very attractive, the data obtained could be of poorer quality; while some compromise is of course made in a strategy that seeks economy, our lab was uncertain about the extent of the trade-off. Therefore, to evaluate the results of the pooled sequencing strategy, we had the same set of 92 large-insert clones sequenced using barcodes, generating sequences to which the pooled sequencing results could be compared. The aim was not to do a comparison of the two methods to show that the pooled method is superior; rather, the aim was to examine the results of the pooled sequencing approach, using high-quality reference sequences from traditional barcoded sequencing. Although a similar pooled clone sequencing method has recently been described by others for metagenome-derived medium-insert plasmids [69] and large-insert fosmids [321], this is the first report of using a pooled strategy for sequencing large-insert metagenomic clones while also critically evaluating the performance of this pooled strategy by comparing the results to barcoded reference sequences of the same clones.

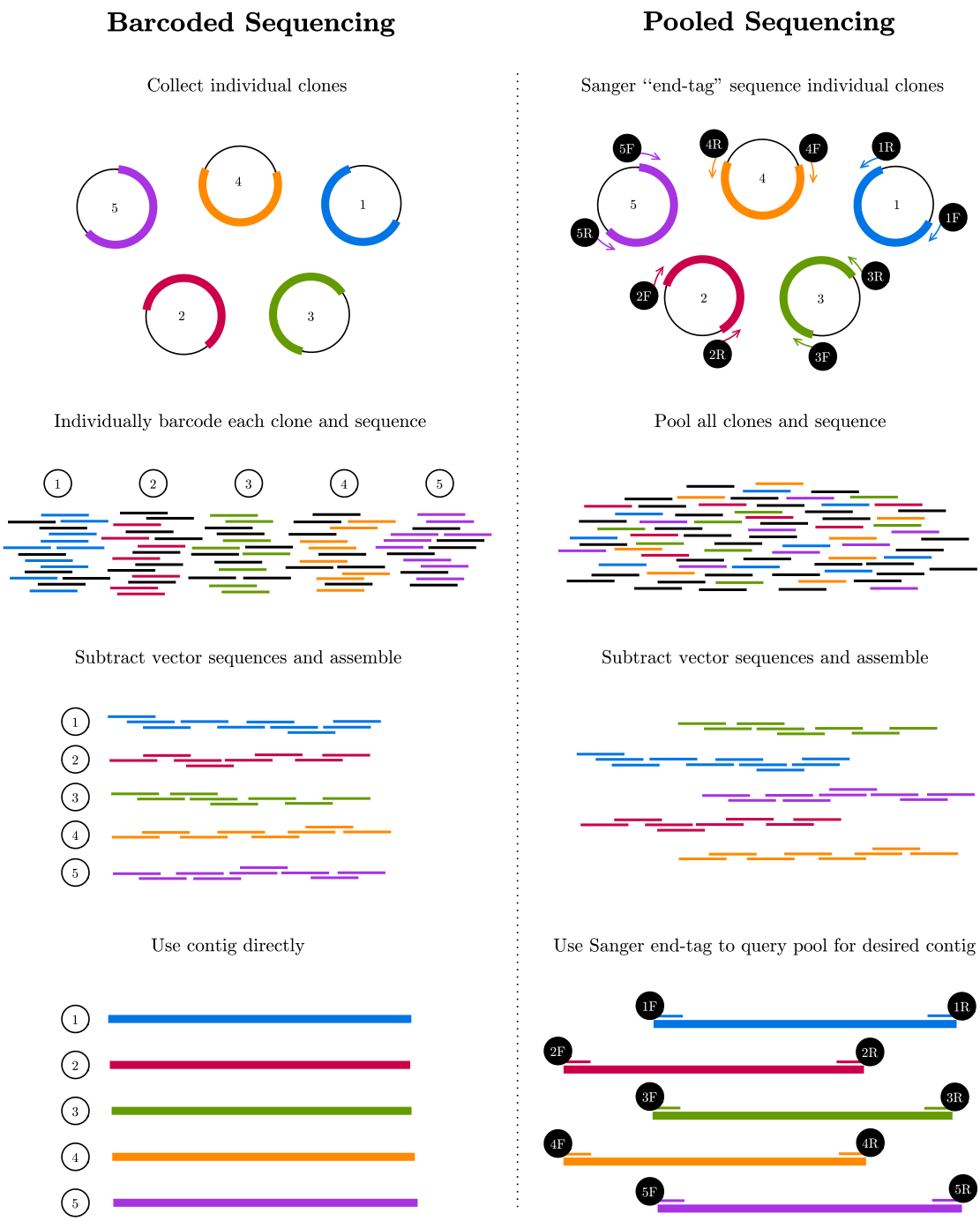


Figure 3.1: Overview of the two methods used in this study for sequencing of large-insert cosmid clones, barcoded sequencing and pooled sequencing. Traditional barcoded sequencing (left) uses DNA barcodes to keep clones as separate samples throughout the sequencing and assembly process. Pooled sequencing (right) involves combining clones into one sample for sequencing and assembly, and subsequently using previously obtained Sanger "end-tags" to retrieve specific clone sequences. [167]

3.4 Results and discussion

3.4.1 Pooled and barcoded sequencing results

A total of 92 cosmid clones were subjected to both pooled sequencing and barcoded sequencing. Of the 92 large-insert cosmid clones, I excluded 19 from subsequent analyses due to incomplete sequencing data. Of the excluded clones, 15 clones had insufficient barcoded sequence data for successful assembly. These samples appeared to have high contamination of *E. coli* genomic DNA and/or mobilizer plasmid DNA. Under my direction, Mike Hall examined the effect of contamination on clone assembly. The estimated percent *E. coli* contamination in each of the 92 samples ranged from 1% to nearly 50%, and, not surprisingly, the higher the contamination, the less likely a successful assembly (Figure 3.2).

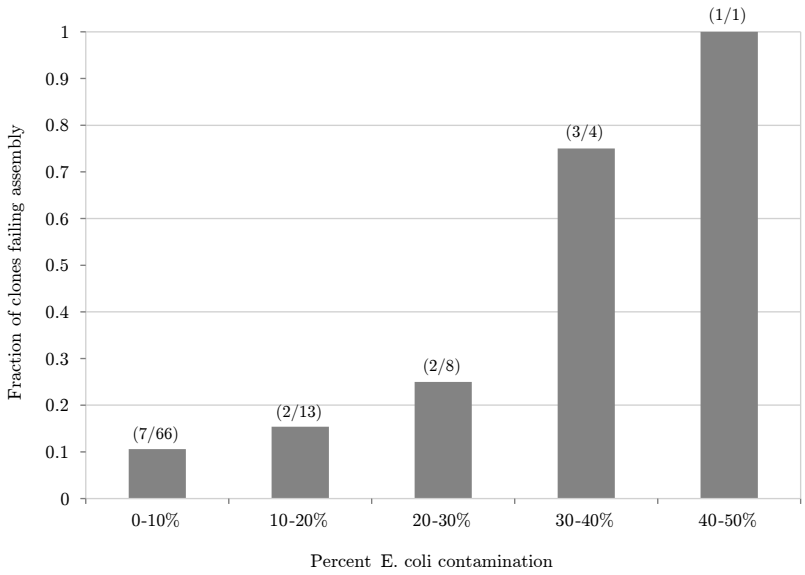


Figure 3.2: Fraction of clones failing assembly, binned by estimated percent *E. coli* contamination. Raw sequence data from barcoded sequencing of 92 clones were examined for *E. coli* contamination. [167]

The remaining 4 of the 19 clones repeatedly failed Sanger end sequencing reactions, possibly due to secondary structure associated with the insert DNA. In our lab's experience, it is occasionally difficult to obtain Sanger reads for certain clones, which we speculate may be caused by such secondary structure effects. In total, 73 clones yielded sufficient data for evaluation of the pooled sequencing results, using the barcoded sequencing results as a reference.

As a result of using different providers for the pooled and barcoded sequencing (see Section 3.6.6 and Section 3.6.5 for details), there was unequal depth of sequencing between the two sequencing approaches (Figure 3.3; see Table 3.3 for individual clone depth); however, it was the barcoded strategy that had the greater depth, which was ideal for its use as the reference data set.

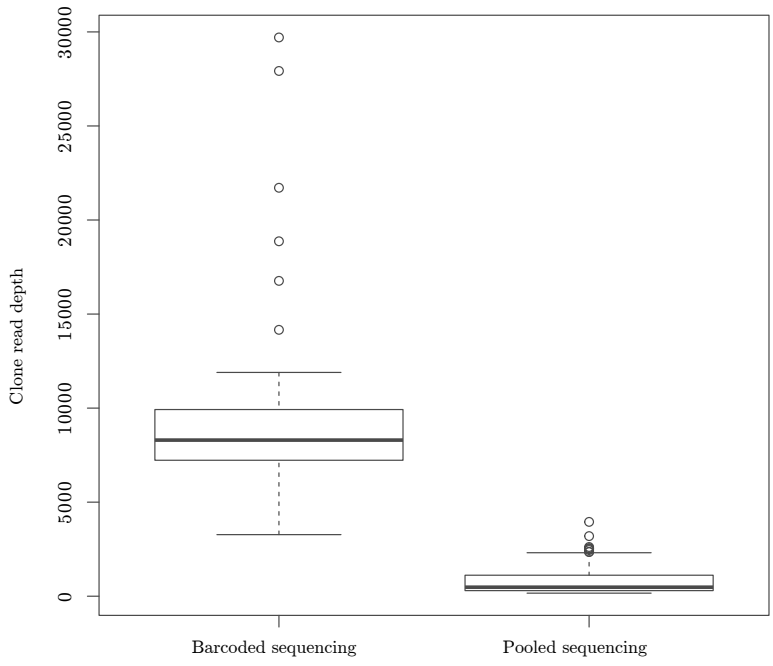


Figure 3.3: Clone sequencing read depth in barcoded sequencing versus pooled sequencing. Values from Table 3.3 were used to compare overall read depth for barcoded versus pooled sequencing strategies. [167]

3.4.2 Evaluation of pooled sequencing results

Using the set of 73 clones, the accuracy and completeness of the pooled sequencing approach was evaluated. First, contigs for each clone were retrieved from the pooled sequencing results using that clone's end tags (see Section 3.6.6 for details; retrieved contigs for all clones are provided in Table 3.9). Then, for each clone, the barcoded sequencing result (i.e., the "barcoded contig") was the reference to which the pooled sequencing result (i.e., the retrieved "pooled contig") was compared. Specifically, the retrieved pooled contig was aligned to its respective barcoded contig, using NCBI nucleotide BLAST [4] running the Megablast algorithm. By aligning the pooled contig to the barcoded contig for each clone, it was possible to quantitatively assess the pooled sequencing approach, by obtaining values for percent identity (i.e., did pooled sequencing return the expected sequence for the clone?) and percent coverage (i.e., did pooled sequencing return the expected length for the clone?). Katja Engel and Greg Vey assisted me in these analyses.

Our initial reservations about a pooled sequencing strategy centred on one major issue, which was that assembly of reads generated from a pooled sample may result in chimeric assemblies – that is, assemblies that are derived from more than one clone. However, when retrieved pooled contigs were aligned to barcoded contigs for each clone, the majority of clones showed alignments of greater than 99.9% identity, with identity values ranging from 99.4-100.0% (Figure 3.4).

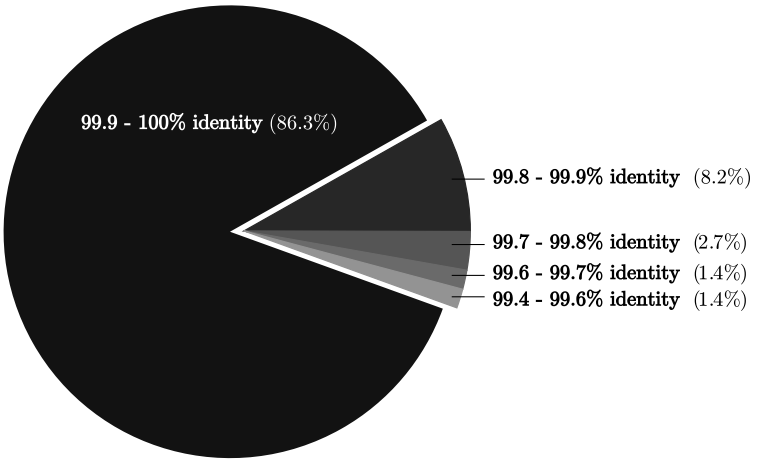


Figure 3.4: Alignment identity between pooled sequencing result and barcoded sequencing result. For all 73 clones, end-tags were used to retrieve contigs from pooled sequencing results; retrieved contigs were aligned to the reference barcoded sequencing result, and clones were binned by percent identity. [167]

Identity values showed high accuracy and little variability, indicating that the pooled sequencing strategy is capable of generating consistently accurate sequence data. Contrary to our concerns, the alignments showed no problems with chimeric sequences, and that most sequences had an error rate of less than one base per thousand. Indeed, this might be an overestimation of the error because the pooled sequencing and assembly method may mask the presence of single nucleotide polymorphisms (discussed further in Section 3.4.4).

The same alignments were used to determine clone coverage obtained by the pooled method and, in contrast to identity, the sequence coverage of pooled clones varied widely. To assess clone coverage, I first categorized the 73 clones into Clone Types (Type A, B, C, or D) based on whether one or both end-tags were obtained, whether the end-tags were able to retrieve a pooled contig, and whether one or two pooled contigs were retrieved (Figure 3.5; designations for each clone are provided in Table 3.1).

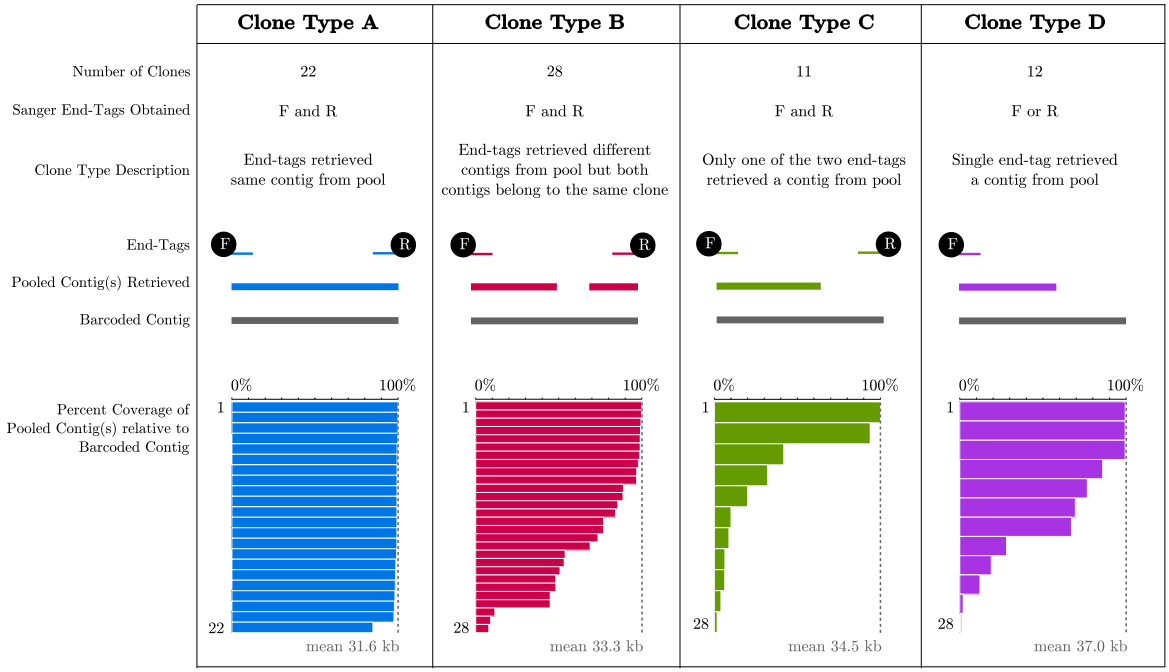


Figure 3.5: Percent coverage of pooled sequencing result relative to barcoded sequencing result. Each of the 73 clones was categorized into Clone Types A, B, C, or D by the number of end-tags obtained (one or two), whether the end-tag retrieved a contig from the pool, and the completeness of the retrieved pooled sequencing result relative to the reference barcoded sequencing result (full or partial coverage). Clone Type descriptions are given above. [167]

Table 3.1: Clone type classification for 73 clones. [167]

Count	Clone	Clone Type
1	BF4	B
2	BT2	A
3	Cel-1	B
4	Cel-32-1	B
5	Cel-3-22-2	B
6	Cel-60-1	B
7	CM-111	D
8	CM-123	A
9	CM-129	A
10	CM-130	A
11	CM-136	A
12	cm18	C
13	CM-18	A
14	CM-19	D
15	CM-2	C
16	Cm26	B
17	Cm3	D
18	Cm30	B
19	CM-31	D
20	CM-4	D
21	cm42	B
22	CM-69	C
23	CM-92	A
24	CX4s17	D
25	CX4s8	B
26	CX6-4	C
27	CX9-10	B
28	CX9s4	B
29	Km-1	C
30	lac-ec1	C
31	lac-ec104	D
32	lac111	C
33	lac121	B
34	lac-ec123	C
35	lac127	B
36	lac13	A
37	lac146	B
38	lac153	B
39	lac16	A
40	lac160	A
41	lac161	A
42	lac170	B
43	lac193	A
44	lac20	B
45	lac24B	C
46	lac27B	C
47	lac35B	C
48	lac36W	A
49	lac55	A
50	lac71	B
51	lac82	D
52	lac84	A
53	MEL125	B
54	MEL126	B
55	PO3	A
56	RCX18	B
57	RCX2	B
58	RCX24	A
59	RCX25	D
60	RCX28	B
61	RCX31	A
62	RCX32	B
63	RCX6	D
64	RCX7	D
65	RCX8	A
66	RCX9	D
67	RCX92	A
68	PCX9M1	A
69	PCX9M3	B
70	PCX9M5	B
71	Xyl 2	B
72	Xyl 3	B
73	Xyl 4	A

Type A represents the ideal outcome, in which the two end-tags retrieved the same contig from the pool; in this case, pooled sequencing resulted in ∼100% coverage for the clone. Type B represents a scenario in which end-tags retrieved different contigs due to a gap in coverage in the middle of the clone. Types C and D represent cases in which coverage was variable and likely underestimated, given that one of the two end-tags either failed to retrieve a contig or was simply missing, respectively. Coverage was highly variable, ranging from 0.4-100.0% over the 73 clones analyzed (Figure 3.5; percent coverage for all clones is provided in Table 3.2).

To determine how well the pooled sequencing strategy worked overall, I used the same coverage data (from Figure 3.5) to bin the 73 clones by coverage (Figure 3.6B). About one-half of the clones showed a retrieved coverage of 90-100%, with an overall average coverage of 71%. I next asked whether the retrieved coverage was an underestimation of the actual coverage achieved by pooled sequencing. To obtain an estimate of the actual coverage, it was necessary to account for unretrieved clone sequences in the pooled sequencing results, which would have occurred due to sequencing gaps, resulting in multiple contigs for a single clone. A comparison of the retrieved coverage to the actual coverage may help to determine whether increasing sequencing depth could increase clone coverage.

Mike Hall assisted me in recovering unretrieved sequences for each clone, using the reference barcoded sequencing result to query the pool (rather than using the end-tags). As an example of this difference, when the specific end-tags for Lactose clone 20 are used to retrieve its sequence from the pool, we obtained a retrieved coverage of 48% (Figure 3.6A); however, when the reference barcoded sequencing result is used instead to query the pooled sequencing results, the coverage improved to 95%. This latter value reflects the actual sequence coverage of the clone found in the pooled sequencing results.

This strategy was employed to correct for unretrieved sequences for all 73 clones, using a 250-base length cut-off and 99.6% identity cut-off; after this correction, coverage improved to an average of 85%, with over 80% of the clones showing 90-100% coverage (Figure 3.6C; retrieved versus estimated actual coverage for each clone is provided in Table 3.2).

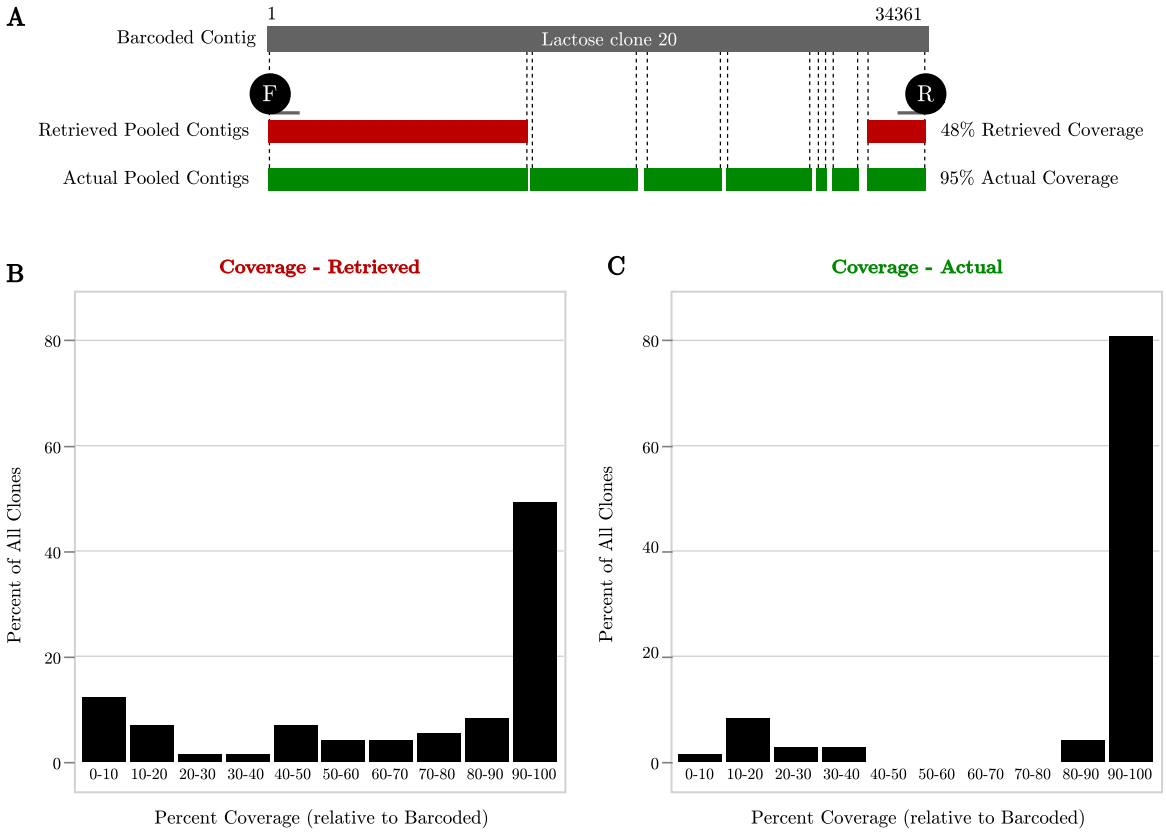


Figure 3.6: Retrieved coverage and estimated actual coverage of pooled sequencing relative to barcoded sequencing. (A) An example clone, Lactose clone 20, shows retrieved coverage at 48% (using end-tags as queries), but an actual coverage of 98% (using barcoded result as query). (B and C) Percent coverage for each of the 73 clones, binned in ten-percent increments. Retrieved coverage (B) is compared to estimated actual coverage (C). [167]

Table 3.2: Retrieved versus estimated actual coverage for 73 clones [167]

Count	Clone	Retrieved Coverage (Pooled relative to Barcoded)	Estimated Actual Coverage (Pooled relative to Barcoded)
1	BF4	0.7367	0.9854
2	BT2	0.9946	0.9946
3	Cel-1	0.5079	0.9289
4	Cel-32-1	0.4487	0.9404
5	Cel-3-22-2	0.4488	0.9404
6	Cel-60-1	0.4826	0.9459
7	CM-111	0.9941	0.9941
8	CM-123	0.9933	0.9933
9	CM-129	0.9917	0.9917
10	CM-130	0.9936	0.9936
11	CM-136	0.9934	0.9934
12	cm18	0.0394	0.1026
13	CM-18	0.9754	0.9754
14	CM-19	0.1214	0.9944
15	CM-2	0.4178	0.9931
16	Cm26	0.0900	0.1867
17	Cm3	0.0043	0.0911
18	Cm30	0.0789	0.1125
19	CM-31	0.9940	0.9940
20	CM-4	0.9939	0.9939
21	cm42	0.1159	0.2099
22	CM-69	0.1004	0.9856
23	CM-92	0.9947	0.9947
24	CX4s17	0.8590	0.9885
25	CX4s8	0.8417	0.8417
26	CX6-4	1.0000	1.0000
27	CX9-10	0.9908	0.9910
28	CX9s4	0.6895	0.9611
29	Km-1	0.9385	0.9619
30	lac-ec1	0.0878	0.3406
31	lac-ec104	0.1895	0.2589
32	lac111	0.1998	0.9667
33	lac121	0.8896	0.9806
34	lac-ec123	0.3211	0.3940
35	lac127	0.9922	0.9922
36	lac13	0.9937	0.9937
37	lac146	0.9794	0.9794
38	lac153	0.9875	0.9875
39	lac16	0.9865	0.9865
40	lac160	0.9941	0.9941
41	lac161	0.9941	0.9941
42	lac170	0.5329	0.9502
43	lac193	0.9940	0.9940
44	lac20	0.4826	0.9459
45	lac24B	0.0635	0.1732
46	lac27B	0.0624	0.1700
47	lac35B	0.0167	0.1278
48	lac36W	0.9938	0.9938
49	lac55	0.8491	0.8491
50	lac71	0.7695	0.9438
51	lac82	0.0204	0.8006
52	lac84	0.9817	0.9817
53	Mel-125	0.9905	0.9905
54	Mel-126	0.8557	0.9760
55	PO3	0.9782	0.9782
56	RCX18	0.9984	0.9984
57	RCX2	0.7704	0.9985
58	RCX24	0.9991	0.9991
59	RCX25	0.7666	1.0000
60	RCX28	0.9951	0.9951
61	RCX31	1.0000	1.0000
62	RCX32	0.5387	0.9968
63	RCX6	0.6946	0.9796
64	RCX7	0.2809	0.9970
65	RCX8	0.9836	0.9836
66	RCX9	0.6714	1.0000
67	RCX92	0.9853	0.9853
68	PCX9M1	1.0000	1.0000
69	PCX9M3	0.8872	0.9890
70	PCX9M5	1.0000	1.0000
71	Xyl 2	0.9692	0.9692
72	Xyl 3	0.9686	0.9686
73	Xyl 4	1.0000	1.0000

These data suggest that an increase in the sequencing depth of the pooled strategy may help to increase clone coverage, as this should reduce the occurrence of gaps that prevent retrieval of the full clone sequence. Indeed, others have shown full recovery of circular DNA molecules using a pooled sequencing approach in other applications. For example, bulk sequencing of the plasmid fraction of an activated sludge metagenome resulted in the complete assembly of forty plasmids, which were confirmed to be closed circular replicons by PCR [261], and pooled sequencing of mitochondrial genomes resulted in complete assembly of each, although the authors found that *de novo* transcriptome assemblers, designed for handling reads with differential coverage, provided much better assembly then assemblers meant for genomes [247]. Together, these results support our findings that a pooled strategy can be an effective alternative.

3.4.3 Clones with sequence similarity may have poor recovery

To determine if factors other than depth of sequencing affect clone coverage in a pooled approach, I first examined the sequence similarity between clones. To do this, I performed an all-by-all pair-wise BLAST comparison of clones, using their barcoded reference sequences (see Section 3.6.9 for details). I found that the majority of the 73 clones had little or no sequence similarity to any other clone in the pool (Figure 3.7A). However, some clones did have sequence similarity; furthermore, the clones that had sequence similarity were often the same clones that had poor retrieved coverage from pooled sequencing (Figure 3.7B). This was particularly striking when comparing to the actual coverage (Figure 3.7C), suggesting that increasing the depth of sequencing may improve clone coverage from pooled sequencing, but only for those clones that do not have sequence similarity to other clones present in the pool.

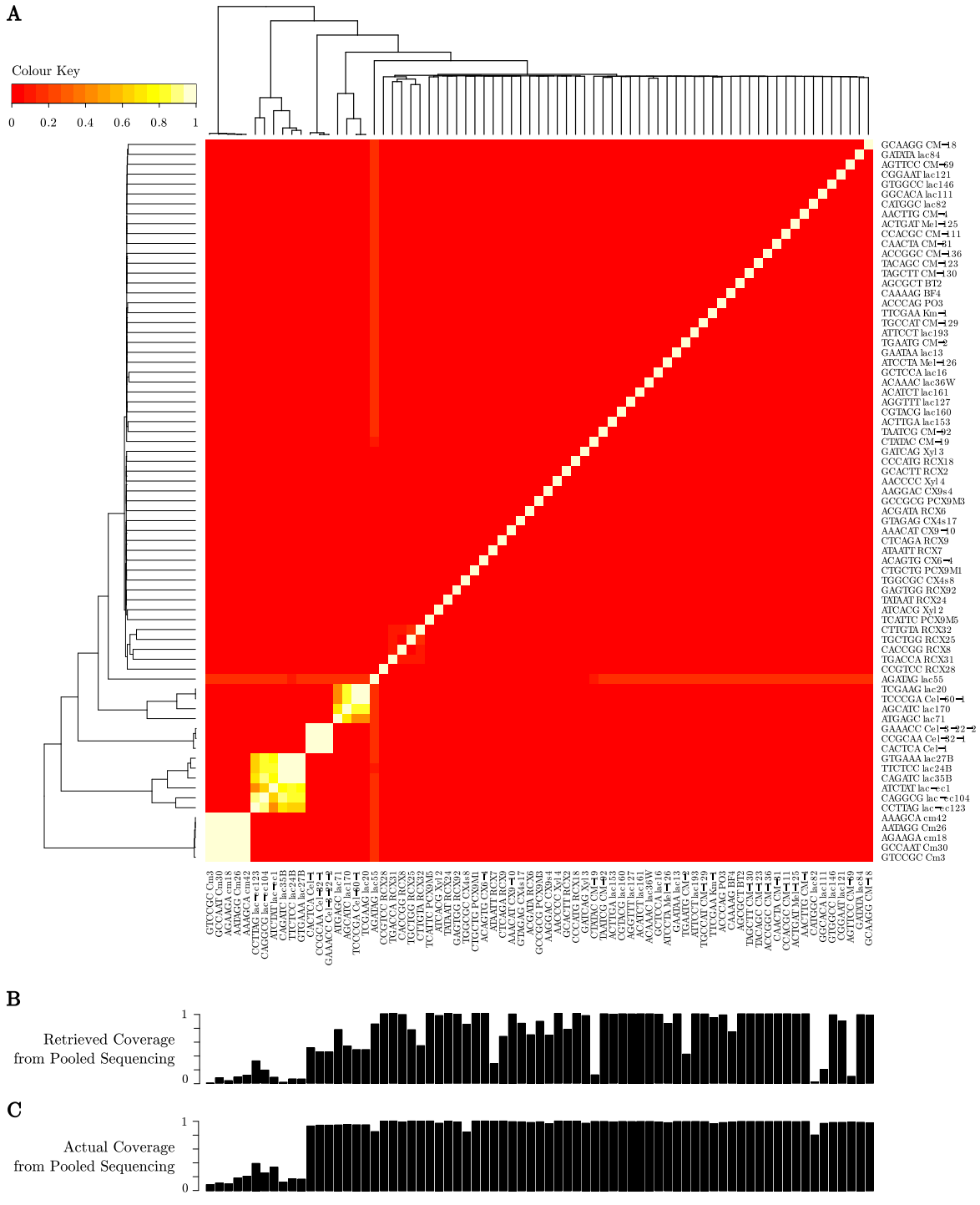


Figure 3.7: Heat map of clone sequence similarity and corresponding bar plots of clone coverage. Pair-wise sequence similarity is shown for all 73 clones (A), juxtaposed to their pooled sequencing coverage, showing both retrieved coverage (B) and actual coverage (C). [167]

I next asked what the sequencing read depth was for each clone to try to understand how the read depth and clone sequence similarity might be related. I asked Mike Hall to estimate the read depth of each of the 73 clones by aligning the raw reads to the assembled contig (see Section 3.6.8 for details; read depth for both pooled sequencing and barcoded sequencing for each clone is provided in Table 3.3). The idea that similar clones are problematic for a pooled sequencing strategy was corroborated using the data from Mike Hall's read depth analysis of each of the 73 clones. To examine the relationship between read depth and pooled sequencing coverage, I plotted the read depth of each clone against both its retrieved and actual coverage (Figure 3.8). I found that for a number of clones, the estimated read depth was particularly high and yet the coverage was unusually low; upon inspecting the identity of these clones, I found them to be the same clones that shared sequence similarity.

Perhaps not unexpectedly, these results suggest that when clones have sequence similarity, pooling and fragmenting the DNA for sequencing causes: (a) an overrepresentation of similar sequences in the pooled sequencing data, and (b) difficulty in assembling the sequences, leading to lack of coverage for the clones from which the sequences originate. There may be other factors that impact the success of pooled sequencing and assembly, such as the presence of repetitive sequences, but this work results suggest that sequencing depth and clone sequence similarity are two significant factors.

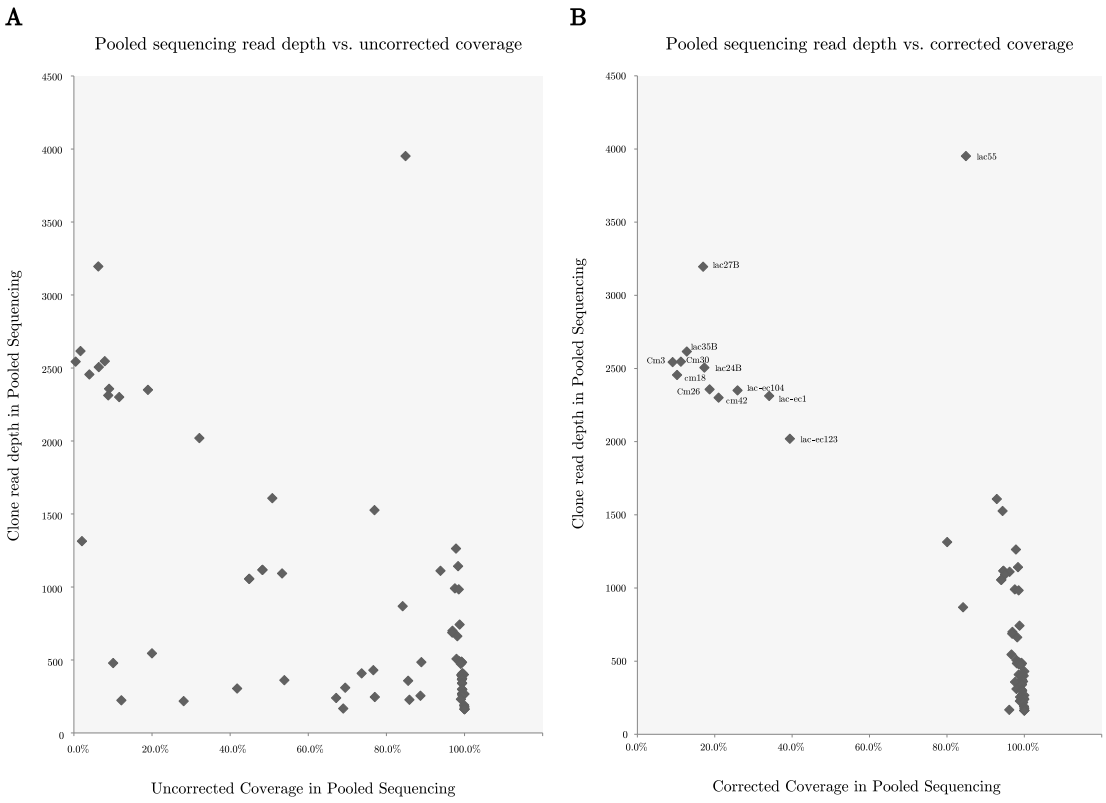


Figure 3.8: Clone read depth plotted against clone coverage in pooled sequencing. The overall read depth for each clone in the pooled sequencing strategy was estimated and plotted against either the uncorrected coverage (A) or corrected coverage (B). [167]

Table 3.3: Estimated read depth for both pooled and barcoded approaches, ranked by depth of pooled sequencing. [167]

Barcode	Clone Name	Pooled Sequencing Read Depth	Barcoded Sequencing Read Depth
AGATAG	lac55	3952	29704
GTGAAA	lac27B	3196	8300
CAGATC	lac35B	2616	8277
GCCAAT	Cm30	2547	7716
GTCCGC	Cm3	2544	9925
TTCTCC	lac24B	2506	8611
AGAAGA	cm18	2456	11288
AATAGG	Cm26	2357	8367
CAGGCG	lac-ec104	2350	10874
ATCTAT	lac-ec1	2312	8384
AAAGCA	cm42	2301	9909
CCTTAG	lac-ec123	2020	8490
CACTCA	Cel-1	1608	7606
ATGAGC	lac71	1526	8087
CATGGC	lac82	1314	11190
ACCCAG	PO3	1263	10374
CACCGG	RCX8	1142	7197
TCCCGA	Cel-60-1	1117	6917
TCGAAG	lac20	1117	5935
TTCGAA	Km-1	1111	8388
AGCATC	lac170	1093	3274
GAAACC	Cel-3-22-2	1056	11067
CCGCAA	Cel-32-1	1056	9250
GCAAGG	CM-18	990	18870
GAGTGG	RCX92	984	7952
TGGCGC	CX4s8	868	16764
ACTTGA	lac153	743	8880
ATCACG	Xyl 2	700	27923
GATCAG	Xyl 3	687	21714
GATATA	lac84	663	14161
GGCACA	lac111	545	7344
GTGGCC	lac146	507	7717
GCTCCA	lac16	491	9720
AGGTTT	lac127	489	8199
CGGAAT	lac121	485	10593
ACAAAC	lac36W	484	8909
AGTTCC	CM-69	479	10386
ACTGAT	Mel-125	474	9722
TGCTGG	RCX25	430	5919
CCGTCC	RCX28	409	7582
CAAAAG	BF4	408	11194
ATTCCT	lac193	402	7991
CCCATG	RCX18	400	7213
TGCCAT	CM-129	397	8858
ACCGGC	CM-136	391	9065
ACATCT	lac161	373	10605
TACAGC	CM-123	366	7313
CTTGTA	RCX32	361	5984
ATCCTA	Mel-126	358	10115
AACTTG	CM-4	342	8342
GAATAA	lac13	340	8453
ACGATA	RCX6	310	7541
TGAATG	CM-2	304	8761
TAGCTT	CM-130	300	8133
AGCGCT	BT2	297	11897
CGTACG	lac160	270	5424
CAACTA	CM-31	268	7539
CCACGC	CM-111	267	8209
ACAGTG	CX6-4	267	5986
TAATCG	CM-92	256	5029
GCCGCG	PCX9M3	255	5055
GCACTT	RCX2	246	6884
CTCAGA	RCX9	240	7227
AAACAT	CX9-10	231	8804
GTAGAG	CX4s17	227	5772
CTATAC	CM-19	224	7553
ATAATT	RCX7	218	6747
TATAAT	RCX24	191	6843
CTGCTG	PCX9M1	178	6976
AAGGAC	CX9s4	167	7853
AACCCC	Xyl 4	164	10947
TCATTC	PCX9M5	163	5026
TGACCA	RCX31	163	4348
Mean		881	9249

3.4.4 Consensus assemblies: a caveat of the pooled approach

Due to the nature of the pooled assembly, overlapping clones assemble into larger contigs. Indeed, three clones were determined to be overlapping by the barcoded sequence data, as well as the pooled sequence data (Figure 3.9).

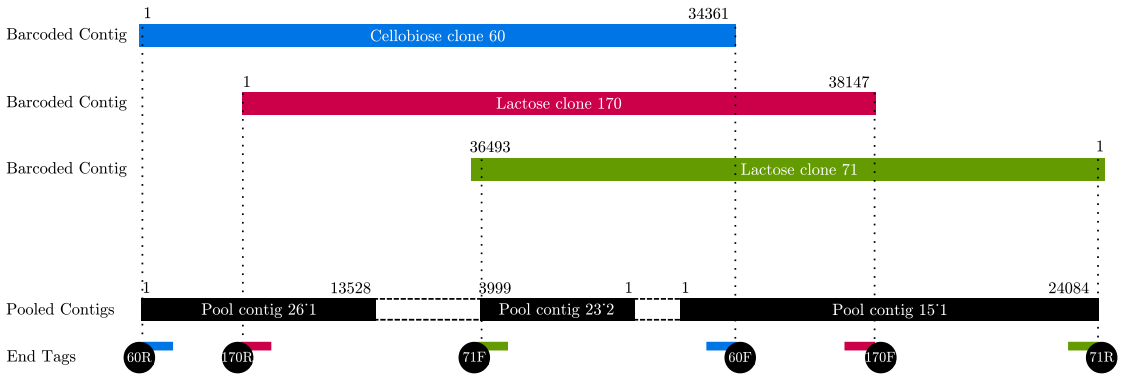


Figure 3.9: Overlapping clones assemble into one contig. Three overlapping clones as revealed by barcoded sequencing (above) and pooled sequencing (below). Locations of end-tags are indicated by vertical dashed lines. White dashed boxes indicate gaps in the pooled sequencing data; black boxes indicate a contig. Lengths of all contigs are given. [167]

In the latter, three contigs were retrieved from the pool using their six end-tags; more than one contig was retrieved due to incomplete sequencing and/or assembly by the pooled method, as discussed above (i.e., Figure 3.6 and Figure 3.7). Although this larger contig is derived from three clones, such a contig should not be classified as chimeric because it represents the metagenomic DNA as it would be found in nature. Furthermore, individual clone sequences can be easily delineated from the greater contig by alignment of clone end-tags to the contig (as illustrated in Figure 3.9).

This particular caveat of pooled sequencing can be viewed as a positive aspect rather than a negative one, because clones from different screens can be immediately

identified as overlapping simply from the clone sequence retrieval process. That being said, the assembly of a consensus sequence from overlapping clones may imply a loss of clone-specific information. It is possible that, in some cases, overlapping clones represent different strains of the same microorganism, or different alleles of the same genes(s). Through pooled assembly and depending on the assembler parameters, such clone-specific allelic information, in the form of single nucleotide polymorophisms (SNPs) or similar small sequence variations, may be lost – that is, the final consensus sequence may represent only the most frequent allele. If it should arise, the issue of information loss for allelic variations may be remedied by further analysis. For example, if clones were determined to be overlapping from the consensus contig obtained from pooled sequencing, it would be possible to examine the raw reads to determine if SNPs are present. If so, sequencing primers could be designed for the target loci to determine exactly which SNP(s) belong to which clones in the physical DNA collection.

3.4.5 Improvements and considerations

In this study, our lab investigated the quality of data obtained from pooled sequencing because this strategy offered an economical solution to the high cost of traditional barcoded sequencing. At the time this work began, there was a large cost difference in the two services that were available (Table 3.4 and Table 3.5). Since then, this difference has decreased, and it is likely that it will continue to do so with further developments in sequencing technology. At least for the time being, however, pooled sequencing remains a more affordable option for functional metagenomics research, particularly if a large number of clones must be sequenced.

Table 3.4: Cost of barcoded sequencing at the Genome Sciences Centre, BC Cancer Agency, Vancouver, Canada. [167]

Traditional Barcoded Illumina	
miniprep	$100.00
barcoded library construction	$8,700.00
sequencing	$1,300.00
assembly (in-house)	$0.00
total cost	$10,100.00
turnaround time	6 months
average coverage per clone	100% (reference)

Table 3.5: Cost of pooled sequencing at the Beijing Genomics Institute, Tai Po, Hong Kong. [167]

Sanger-Illumina Pooled Sequencing	
miniprep	$100.00
Sanger end-sequencing	$1,000.00
library construction	$400.00
pooled sequencing	$300.00
assembly and annotation	$400.00
total cost	$2,200.00
turnaround time	4 months
average coverage per clone, uncorrected	71%
average coverage per clone, corrected	85%

In our workflow, the lab concurrently had clones analyzed by pooled sequencing and by Sanger sequencing (for the generation of end-tags); this was done concurrently due to anticipation of a lengthy turnaround time for the Illumina sequencing results, which is typically (and was in fact) the case. However, given our experience, I recommend obtaining end sequences for all clones before carrying out pooled sequencing, due to the unexpected difficulty of Sanger-sequencing certain clones. Without two end-tags for each clone, it becomes difficult to retrieve the corresponding contig from the pool without further work, such as subcloning and sequencing fragments of the insert (which would negate the ease and economy of the pooled sequencing strategy).

Assembly for both the barcoded and the pooled sequencing strategies revealed contamination with *E. coli* genomic DNA sequences, indicating that minipreps of cosmid clones contained host DNA. Similar results were reported for genomic library BAC clones isolated for pooled sequencing [193]. Such contamination adds undesired DNA template to the sequencing reaction, affecting required-depth-of-coverage calculations, and possibly leading to insufficient sequencing and poor clone sequence recovery. This may have been a problem in our own incomplete recovery for the pooled strategy. We recommend removing contaminating genomic DNA by cesium chloride density purification or pre-treatment of samples with Plasmid-Safe DNase (Epicentre), which may help reduce genomic contamination up to ten-fold [16]. Clone sequence recovery was not problematic in the barcoded sequencing strategy because the sequencing depth was extremely high for the purpose of generating high-quality reference sequence data (Figure 3.3).

Another consideration for pooled sequencing relates to the problem of sequence similarity (Figure 3.7). These results indicate that clones that have sequence similarity are problematic in a pooled strategy, likely due to difficulties in assembling the similar reads and resulting in poor clone sequence recovery. The simple solution would be to

avoid pooling clones that share sequence similarity, but this remains a difficult, if not impossible, task without prior knowledge of the clone sequence. A possible way to reduce the potential for sequence similarity may be to assemble pools of clones such that the diversity of functional screens represented is maximized within a pool. In this way, the presence of homologous genes may be reduced.

One other consideration for the pooled sequencing strategy relates to the issue of consensus assemblies, which may occur for overlapping clones during assembly process (Figure 3.9). Since overlapping clones likely (though not always) result from the same functional screen, it is possible for the experimental biologist to minimize their presence by doing restriction profile comparisons prior to selecting clones for pooling and sequencing. It may also be possible to reduce loss of clone-specific sequence variation by using combinatorial or overlapping clone pooling approaches, which have been used by others for strategic sequencing of BAC clones from genomic libraries [30, 193] as well as plasmid-based oligonucleotide libraries [79]. In such an approach, a large set of clones is divided into subpools such that each clone is present in multiple subpools, but no two clones are in the same subpool more than once, which can help resolve ambiguity in the case that clones in one pool have sequence similarity. In the simplest approach for combining the barcoded and pooled sequencing strategies, a large pool of clones could be split into smaller subpools, each of which gets barcoded. By strategically using a mixture of barcoding, pooling, and/or duplicate sequencing, one can strike a balance between making use of sequencing power and being able to recover accurate and complete clone sequence information.

3.5 Conclusions

We explored a more economical sequencing strategy than barcoded sequencing by using a pooled sequencing method that successfully obtained sequence information for a set of large-insert clones. In particular, we validated this method by comparing the sequence data to reference data generated from barcoded sequencing of the same set of clones.

By observing identity and coverage between the two datasets for 73 clones, I have demonstrated high quality assemblies from the pooled sequencing dataset. Using the pooled strategy, retrieved clone sequences showed high accuracy, with identity at 99.9-100% for the majority of clones. The amount of sequence recovered for each clone, however, was variable; averaged across 73 clones, the retrieved coverage was 71%, with some clones showing full coverage, and others with minimal coverage. Correcting for sequencing gaps, the average coverage increased to 85%. These results suggest that increasing sequencing depth can improve clone coverage, but that clones that have sequence similarity are problematic in a pooled strategy regardless. Though pooled sequencing has generated promising results, refinement of the method is required: sequencing depth will need to be optimized to obtain maximum recovery of clone sequence, and the choice of clones to pool will also need consideration, to minimize the presence of clones with sequence similarity.

These results demonstrate that, with further optimization, a pooled sequencing approach could become the preferred method of generating clone sequence data, as its cost is a fraction of that of barcoded sequencing. It is important to note that clone sequence recovery may not be complete or even possible for all clones that have been pooled for sequencing; however, until the cost of barcoding many samples becomes affordable in the way that Sanger sequencing has become affordable, pooled sequencing of large sets of clones remains a relevant and reasonable strategy.

3.6 Specific materials and methods

3.6.1 Ethics Statement

Approval for the collection of human fecal samples was obtained from the Office of Research Ethics of the University of Waterloo in Waterloo, Canada, and written consent was obtained from the volunteers. No identification was attached to the collected samples and samples were pooled prior to use.

3.6.2 Isolation of HMW DNA

Soil samples were obtained from diverse environments across Canada [222]. Information regarding the metagenomic libraries constructed from Canadian soil samples is available online through the Canadian MetaMicrobiome Project website (http://www.cm2bl.org).

The isolation of high-molecular-weight DNA was previously described for fecal samples (Section 2.4.5) and for pure bacterial cultures (Section 2.4.6). Extracted DNA was either cloned directly or purified further by synchronous coefficient of drag alteration (SCODA) using the Aurora (Boreal Genomics) according to an established protocol [75]. Crude or SCODA-purified DNA was quantified by gel electrophoresis, using bacteriophage λ DNA as a standard.

3.6.3 Construction of large-insert metagenomic cosmid libraries

The cosmid vector pJC8 (Genbank accession KC149513; [43]) formed the backbone of all metagenomic libraries constructed in this study. In addition to constructing new libraries, existing metagenomic clones were used from previous libraries [320], constructed in the cosmid vector pRK7813 (Genbank accession KC442292; [139]). All libraries have entries in the NCBI BioSample database [13], and details regarding the libraries used in this study are summarized in Table 3.6.

Table 3.6: Metagenomic and genomic libraries screened.

Library name	**NCBI BioSample**	**DNA source**	**No. clones**	**Vector**	**Ref.**
12AC	SAMN02324088	soil (agricultural)	80,000	pJC8	[43]
BF1	SAMN02324093	*Bacteroides fragilis*	18,000	pJC8	this study
BT1	SAMN02324089	*Bacteroides thetaiotaomicron*	8,000	pJC8	this study
CLGM1	SAMN02324081	human feces	42,000	pJC8	this study
CX3	SAMN02324235	activated sludge (pulp and paper)	2,500	pRK7813	[320]
CX4	SAMN02393652	activated sludge (pulp and paper)	3,900	pRK7813	[320]
CX6	SAMN02393657	activated sludge (municipal)	3,300	pRK7813	[320]
CX9	SAMN02393684	soil (creek)	22,000	pRK7813	[320]
CX10	SAMN02393686	soil (creek)	8,700	pRK7813	[320]

Libraries were constructed as previously described [43]. Briefly, the vector pJC8 was digested with Eco72I/PmlI to produce blunt ends and then dephosphorylated. The backbone was purified from the 0.8 kb gentamicin resistance gene stuffer, either with an EZ-10 Spin Column DNA Gel Extraction Kit (BioBasic) or by electroelution. The high-molecular-weight DNA extracted from either environmental samples or pure culture (up to 25 µg of either crude or purified DNA) was size-selected by pulsed-field gel electrophoresis (PFGE) using a CHEF MAPPER Pulsed Field Gel Electrophoresis System (Bio-Rad). The gel fragment containing DNA of approximately 40-70 kb was excised, then electroeluted and concentrated using an Amicon Ultra Centrifugal Filter with 30 kDa MWCO (Millipore). Purified DNA (2.5 µg) was end-repaired using the End-It DNA End-Repair Kit (Epicentre). A phenol:chloroform extraction was performed to remove T4 polynucleotide kinase, and DNA was precipitated, resuspended in TE, and quantified by gel electrophoresis, using bacteriophage λ DNA as a standard. The purified and blunt-ended DNA was then ligated to the linearized cosmid vector. Ligations were carried out at 14℃ overnight with Fast-Link DNA Ligase (Epicentre), using 500 ng of end-repaired insert DNA and a vector-to-insert molar ratio of 10:1. Ligations were packaged into λ phage heads using Gigapack III XL Packaging Extract (Stratagene 200209) according to the manufacturer's instructions, and the final phage suspension was stored at 4℃.

To prepare cells for transduction, *E. coli* HB101 was streaked from frozen stock onto LB agar, and a single colony was then inoculated into 5 ml of LB. The culture was grown overnight at 37°C, and was subcultured 1:200 in 5 ml of LB supplemented with 0.2% maltose and 10 mM $MgSO_4$. The culture was grown to an OD_{600} of 0.8 (Spectronic Spec 20D). Cells were pelleted by centrifugation, resuspended in 2.5 ml of LB supplemented with 10 mM $MgSO_4$, and held on ice. For an estimate of phage concentration, 10 µl phage were mixed with 90 µl cells, and the mixture was incubated

at room temperature for 30 minutes, and moved to 37℃ for 30 minutes. Cells were pelleted by centrifugation and plated on LB with 20 µg/ml tetracycline to select for transductants. Plates were incubated overnight at 37℃ and colonies were counted to estimate phage concentration in the suspension. Finally, the transduction was scaled up to achieve approximately 1000 colonies per plate. Several plates were counted for an estimate of metagenomic library size, and then pooled and stored at -80℃. For regular use, libraries were propagated from the original frozen stock. For an estimate of average insert size, library stocks were streaked onto LB with 20 µg/ml tetracycline, and colonies were selected at random for restriction analysis.

3.6.4 Functional screens and positive clones

Various function-based screens were performed in our laboratory, including screens for antibiotic resistance genes, conjugation genes, and carbohydrate utilization genes. Tens to hundreds of positive clones were isolated from each screen although 92 distinct clones (based on restriction enzyme digestion patterns) were chosen for full sequencing. The list of clones and the screens from which they were isolated are provided (Table 3.7). Cosmid clone DNA was isolated from either *E. coli* HB101 or DH5α.

Table 3.7: Functional screens from which cosmid clones were isolated; bolded clone names indicate those excluded from analyses. [167]

Count	Clone Name	Functional Screen	Library Name	Vector Backbone
1	BF4	random clone Bacteroides fragilis cosmid library	BF1	pJC8
2	BT2	random clone Bacteroides theta cosmid library	BT1	pJC8
3	Cel-1	cellobiose utilization	12AC	pJC8
4	Cel-32-1	cellobiose utilization	12AC	pJC8
5	Cel-3-22-2	cellobiose utilization	12AC	pJC8
6	**Cel-3-24-2**	cellobiose utilization	12AC	pJC8
7	Cel-60-1	cellobiose utilization	12AC	pJC8
8	**CM-10**	conjugation	12AC	pJC8
9	**CM-110**	conjugation	12AC	pJC8
10	CM-111	conjugation	12AC	pJC8
11	CM-123	conjugation	12AC	pJC8
12	CM-129	conjugation	12AC	pJC8
13	CM-130	conjugation	12AC	pJC8
14	**CM-131**	conjugation	12AC	pJC8
15	**CM-135**	conjugation	12AC	pJC8
16	CM-136	conjugation	12AC	pJC8
17	**CM-15**	conjugation	12AC	pJC8
18	cm18	chloramphenicol resistance	12AC	pJC8
19	CM-18	conjugation	12AC	pJC8
20	CM-19	conjugation	12AC	pJC8
21	CM-2	conjugation	12AC	pJC8
22	**CM-20**	conjugation	12AC	pJC8
23	Cm26	chloramphenicol resistance	12AC	pJC8
24	Cm3	chloramphenicol resistance	12AC	pJC8
25	Cm30	chloramphenicol resistance	12AC	pJC8
26	CM-31	conjugation	12AC	pJC8
27	CM-4	conjugation	12AC	pJC8
28	cm42	chloramphenicol resistance	12AC	pJC8
29	**CM-45**	conjugation	12AC	pJC8
30	**CM-56**	conjugation	12AC	pJC8
31	**CM-64**	conjugation	12AC	pJC8
32	CM-69	conjugation	12AC	pJC8
33	CM-92	conjugation	12AC	pJC8
34	CX4s17	PHB synthesis	CX4	pRK7813
35	CX4s8	PHB synthesis	CX4	pRK7813
36	CX6-4	PHB synthesis	CX6	pRK7813
37	CX9-10	PHB synthesis	CX9	pRK7813
38	CX9s4	PHB synthesis	CX9	pRK7813
39	**lac97W**	lactose utilization	12AC	pJC8
40	Km-1	kanamycin resistance	12AC	pJC8
41	lac-ec1	lactose utilization	12AC	pJC8
42	**lac100B**	lactose utilization	12AC	pJC8
43	lac-ec104	lactose utilization	12AC	pJC8
44	lac111	lactose utilization	12AC	pJC8
45	lac121	lactose utilization	12AC	pJC8
46	**lac112W**	lactose utilization	12AC	pJC8
47	lac-ec123	lactose utilization	12AC	pJC8
48	lac127	lactose utilization	12AC	pJC8
49	lac13	lactose utilization	12AC	pJC8
50	lac146	lactose utilization	12AC	pJC8
51	lac153	lactose utilization	12AC	pJC8
52	lac16	lactose utilization	12AC	pJC8
53	lac160	lactose utilization	12AC	pJC8
54	lac161	lactose utilization	12AC	pJC8
55	lac170	lactose utilization	12AC	pJC8
56	lac193	lactose utilization	12AC	pJC8
57	lac20	lactose utilization	12AC	pJC8
58	**lac224**	lactose utilization	12AC	pJC8
59	lac24B	lactose utilization	12AC	pJC8
60	lac27B	lactose utilization	12AC	pJC8
61	lac35B	lactose utilization	12AC	pJC8
62	**lac36B**	lactose utilization	12AC	pJC8
63	lac36W	lactose utilization	12AC	pJC8
64	lac55	lactose utilization	12AC	pJC8
65	lac71	lactose utilization	12AC	pJC8
66	lac82	lactose utilization	12AC	pJC8
67	lac84	lactose utilization	12AC	pJC8
68	Mel-125	melibiose utilization	12AC	pJC8
69	Mel-126	melibiose utilization	12AC	pJC8
70	PO3	random clone human gut library	CLGM1	pJC8
71	**RCX11**	3-hydroxybutyrate utilization	CX4	pRK7813
72	**RCX12**	3-hydroxybutyrate utilization	CX4	pRK7813
73	**RCX13**	3-hydroxybutyrate utilization	CX4	pRK7813
74	**RCX15**	3-hydroxybutyrate utilization	CX4	pRK7813
75	RCX18	3-hydroxybutyrate utilization	CX9	pRK7813
76	RCX2	3-hydroxybutyrate utilization	CX3	pRK7813
77	RCX24	3-hydroxybutyrate utilization	CX9	pRK7813
78	RCX25	3-hydroxybutyrate utilization	CX9	pRK7813
79	RCX28	3-hydroxybutyrate utilization	CX9	pRK7813
80	RCX31	3-hydroxybutyrate utilization	CX9	pRK7813
81	RCX32	3-hydroxybutyrate utilization	CX10	pRK7813
82	RCX6	3-hydroxybutyrate utilization	CX4	pRK7813
83	RCX7	3-hydroxybutyrate utilization	CX9	pRK7813
84	RCX8	3-hydroxybutyrate utilization	CX9	pRK7813
85	RCX9	3-hydroxybutyrate utilization	CX10	pRK7813
86	RCX92	3-hydroxybutyrate utilization	CX9	pRK7813
87	PCX9M1	3-hydroxybutyrate utilization	CX9	pRK7813
88	PCX9M3	3-hydroxybutyrate utilization	CX9	pRK7813
89	PCX9M5	3-hydroxybutyrate utilization	CX9	pRK7813
90	Xyl 2	xylose utilization	CX9	pRK7813
91	Xyl 3	xylose utilization	CX9	pRK7813
92	Xyl 4	xylose utilization	CX9	pRK7813

3.6.5 Barcoded sequencing

Cosmid DNA was prepared from *E. coli* DH5α using a GeneJET Plasmid Miniprep Kit (Thermo Scientific), and 1-2 µg of DNA from each of the 92 samples was adjusted to ¿25 ng/µl. Samples were submitted to the BC Cancer Agency at the Michael Smith Genome Sciences Centre for individual barcoding and 75-base paired-end sequencing on the Illumina HiSeq 2000 platform, using in-house protocols and reagents for library construction. Clones were sequenced to a read depth of approximately 9000-fold, on average (Figure 3.3 and Table 3.3). This high coverage was ideal for a high-quality reference data set. Vector sequences were subtracted from the raw data by comparing all reads against the vector backbone using BLAST (with a requirement for 100% identity), and the data were assembled using ABySS version 1.3.2 [272]; default settings were used, with the exception of a k-mer length of 64. At the time of assembly, the complete sequence of the cosmid vector pJC8 was not yet available; as a result, vector subtraction used the closely related parent vector pRK404 (Genbank accession AY204475; [63]), and assemblies were checked subsequently for remaining vector sequences.

After assembly, the barcoded sequencing data were prepared in order to use as a reference for evaluation of the pooled sequencing data. For the majority of clones, assembly resulted in a single contig, usually exceeding 30 kb, as expected. For cases in which assembly resulted in more than one contig, contigs were manually checked for sequences from contaminating *E. coli* genomic DNA, helper plasmids, and cloning vectors, and those contigs were removed. For 3 clones, multiple contigs remained, indicating the samples may have been insufficiently sequenced, resulting in gaps. Accordingly, we concatenated the multiple large contigs and treated them as one contig. Using the described strategy, reference contigs were obtained for 77 out of 92 clones. The average contig length was 33.5 kb, with the largest being 47.2 kb and the smallest 1.8 kb. Though our cloning strategy enriches for high-insert clones, we have occasion-

ally observed smaller inserts after carrying out functional screening. These smaller inserts may have arisen from recombination and subsequent loss of cloned DNA after the library construction process. Sequence data have been made available for download (see below). Barcodes are provided in Table 3.8.

Table 3.8: Barcodes corresponding to each clone for Illumina sequencing. [167]

Count	Clone Name	Barcode
1	BF4	CAAAAG
2	BT2	AGCGCT
3	Cel-1	CACTCA
4	Cel-32-1	CCGCAA
5	Cel-3-22-2	GAAACC
6	Cel-3-24-2	GCCTTA
7	Cel-60-1	TCCCGA
8	CM-10	AAGACT
9	CM-110	CTAGCT
10	CM-111	CCACGC
11	CM-123	TACAGC
12	CM-129	TGCCAT
13	CM-130	TAGCTT
14	CM-131	GGCTAC
15	CM-135	TTAGGC
16	CM-136	ACCGGC
17	CM-15	CGATGT
18	cm18	AGAAGA
19	CM-18	GCAAGG
20	CM-19	CTATAC
21	CM-2	TGAATG
22	CM-20	ATGTCA
23	Cm26	AATAGG
24	Cm3	GTCCGC
25	Cm30	GCCAAT
26	CM-31	CAACTA
27	CM-4	AACTTG
28	cm42	AAAGCA
29	CM-45	CCAACA
30	CM-56	AGGCCG
31	CM-64	GATGCT
32	CM-69	AGTTCC
33	CM-92	TAATCG
34	CX4s17	GTAGAG
35	CX4s8	TGGCGC
36	CX6-4	ACAGTG
37	CX9-10	AAACAT
38	CX9s4	AAGGAC
39	jac97W	CGAGAA
40	Km-1	TTCGAA
41	lac-ec1	ATCTAT
42	lac100B	GACGGA
43	lac-ec104	CAGGCG
44	lac111	GGCACA
45	lac121	CGGAAT
46	lac112W	TCGGCA
47	lac-ec123	CCTTAG
48	lac127	AGGTTT
49	lac13	GAATAA
50	lac146	GTGGCC
51	lac153	ACTTGA
52	lac16	GCTCCA
53	lac160	CGTACG
54	lac161	ACATCT
55	lac170	AGCATC
56	lac193	ATTCCT
57	lac20	TCGAAG
58	lac224	CATTTT
59	lac24B	TTCTCC
60	lac27B	GTGAAA
61	lac35B	CAGATC
62	lac36B	AAATGC
63	lac36W	ACAAAC
64	lac55	AGATAG
65	lac71	ATGAGC
66	lac82	CATGGC
67	lacS4	GATATA
68	Mel-125	ACTGAT
69	Mel-126	ATCCTA
70	PO3	ACCCAG
71	RCX11	AAGCGA
72	RCX12	ACTCTC
73	RCX13	ATACGG
74	RCX15	CACGAT
75	RCX18	CCCATG
76	RCX2	GCACTT
77	RCX24	TATAAT
78	RCX25	TGCTGG
79	RCX28	CCGTCC
80	RCX31	TGACCA
81	RCX32	CTTGTA
82	RCX6	ACGATA
83	RCX7	ATAATT
84	RCX8	CACCGG
85	RCX9	CTCAGA
86	RCX92	GAGTGG
87	PCX9M1	CTGCTG
88	PCX9M3	GCCGCG
89	PCX9M5	TCATTC
90	Xyl 2	ATCACG
91	Xyl 3	GATCAG
92	Xyl 4	AACCCC

3.6.6 Sanger end-sequencing and pooled sequencing

Cosmid DNA was prepared from *E. coli* DH5α using a GeneJET Plasmid Miniprep Kit (Thermo Scientific). Aliquots of 100 ng from each of the 92 samples were pooled and concentrated to 125 ng/µl. The pooled samples were sequenced by the Beijing Genomics Institute (BGI) using 90-base paired-end sequencing on the Illumina HiSeq 2000 platform, using in-house protocols and reagents for library construction. Clones were sequenced to a read depth of approximately 900-fold on average (Figure 3.3), upon recommendation of >100-fold coverage. The service provider subtracted vector sequences using SOAPaligner version 2.21 [184] (again, using pRK404), and completed assembly using SOAPdenovo version 1.05 [185], using a k-mer size of 31, and BWA version 0.5.8 [181]. This resulted in 563 contigs ranging between 0.5 kb to 97.7 kb, with a mean contig length of 11.7 kb. Contigs exceeding the expected insert size were determined to be *E. coli* genomic DNA contamination, the presence of which did not interfere with clone sequence retrieval, as retrieval is done using clone end sequences.

Concurrent to pooled sequencing, samples were end-sequenced by Sanger sequencing at BioBasic Inc., Lucigen Corporation, or The Centre for Applied Genomics, to generate end-tags. One or both end sequences were obtained for 83 out of 92 clones. Sequencing primers used were standard M13 forward and M13 reverse from the sequencing facility, or custom primers JC102 (5'`TAACAATTTCACACAGGAAACAGCTATGAC`) and JC103 (5'`GCGATTAAGTTGGGTAACGCCAGGGTTTTC`). The obtained end-tags were then used to query the pooled sequencing results, using NCBI nucleotide BLAST [4] running the Megablast algorithm. In this manner, contigs were retrieved from the pool for each clone; see Table 3.9 for details. Pooled sequence data and end sequence data have been made available for download (see Section 3.6.10).

Table 3.9: Summary of retrieved contigs for the pooled sequencing approach. [167]

Count	Clone	Forward End-Tag (M13F/JC103)					Reverse End-Tag (M13R/JC102)				
		End Tag?	Tag Len	Retrieved Contig ID	Contig Len	Align. Identity	End Tag?	Tag Len	Retrieved Contig ID	Contig Len	Align. Identity
1	BF4	Y	661	scaffold196`1	9979	0.9847457627	Y	720	scaffold199`1	16126	0.9984802432
2	BT2	Y	517	scaffold258`1	39283	0.9712389381	Y	510	scaffold258`1	39283	0.9841986456
3	Cel-1	Y	1000	scaffold10`1	3594	0.9804347826	Y	563	scaffold7`1	17578	0.9936708861
4	Cel-32-1	Y	559	scaffold10`1	3594	0.9810526316	Y	561	scaffold7`1	17578	0.9977728285
5	Cel-3-22-2	Y	559	scaffold10`1	3594	0.9810526316	Y	561	scaffold7`1	17578	0.9977728285
6	Cel-60-1	Y	545	scaffold15`1	24084	0.9838337182	Y	1042	scaffold26`1	13528	0.9873817035
7	CM-111	Y	761	scaffold146`1	36324	0.9352226721	N				
8	CM-123	Y	805	scaffold155`1	32613	0.9955686854	Y	886	scaffold155`1	32613	0.982278481
9	CM-129	Y	633	scaffold213`1	25538	0.962745098	Y	759	scaffold213`1	25538	0.987933635
10	CM-130	Y	607	scaffold248`1	33740	0.987933635	Y	764	scaffold248`1	33740	1
11	CM-136	Y	602	scaffold63`1	31929	1	Y	402	scaffold63`1	31929	0.9765886288
12	cm18	Y	567	no hit			Y	844	scaffold151`1	3127	0.9943582511
13	CM-18	Y	995	scaffold126`1	8660	0.9836448598	Y	525	scaffold126`1	8660	1
14	CM-19	Y	564	scaffold260`2	4083	0.9953379953	N				
15	CM-2	Y	882	no hit			Y	519	scaffold185`1	13797	0.9976470588
16	Cm26	Y	747	scaffold151`1	3127	1	Y	1114	scaffold73`1	1676	0.9953488372
17	Cm3	Y	720	scaffold116`1	629	1	N				
18	Cm30	Y	1166	scaffold116`1	629	1	Y	763	scaffold127`1	2108	0.9906542056
19	CM-31	N					Y	343	scaffold246`1	34863	1
20	CM-4	Y	522	scaffold223`1	35472	0.9797979798	N				
21	cm42	Y	1039	scaffold151`1	3127	0.9855715871	Y	770	scaffold73`1	1676	0.9910979228
22	CM-69	Y	763	no hit			Y	921	scaffold110`1	2278	0.9875466999
23	CM-92	Y	1043	scaffold242`1	37050	0.9953271028	Y	641	scaffold242`1	37050	0.9811320755
24	CX4s17	N					Y	562	scaffold65`1	35609	0.9945454545
25	CX4s8	Y	766	scaffold35`2	3907	0.9957627119	Y	1121	scaffold54`1	2383	0.9465478842
26	CX6-4	Y	844	scaffold118`1	40976	0.9927971188	Y	688	no hit		
27	CX9-10	Y	927	scaffold13`1	7483	0.9812981298	Y	611	scaffold74`1	27257	0.995
28	CX9s4	Y	1184	scaffold234`1	12189	0.969273743	Y	681	scaffold210`1	12971	0.9910313901
29	Km-1	Y	687	scaffold56`1	32939	0.9963702359	Y	810	no hit		
30	lac-ec1	Y	743	no hit			Y	524	scaffold85`1	3220	0.9900497512
31	lac-ec104	N					Y	521	scaffold24`1	12023	0.9974811083
32	lac111	Y	561	scaffold45`5	6060	0.9946380697	Y	247	no hit		
33	lac121	Y	1166	scaffold128`1	22461	0.983463035	Y	598	scaffold134`1	3736	0.9940944882
34	lac-ec123	Y	813	scaffold24`1	12023	0.9897510981	Y	348	no hit		
35	lac127	Y	607	scaffold75`1	14099	0.9957983193	Y	841	scaffold109`1	17771	0.9919354839
36	lac13	Y	684	scaffold259`1	34172	0.9910394265	Y	762	scaffold259`1	34172	0.9864253394
37	lac146	Y	768	scaffold52`1	15507	0.9968652038	Y	727	scaffold249`1	10025	0.9920760697
38	lac153	Y	640	scaffold11`1	17026	0.9872881356	Y	1076	scaffold11`2	18778	0.9969325153
39	lac16	Y	920	scaffold84`2	32539	0.9849812265	Y	601	scaffold84`2	32539	0.9943019943
40	lac160	Y	603	scaffold243`1	36291	0.9978991597	Y	645	scaffold243`1	36291	0.9981751825
41	lac161	Y	641	scaffold77`1	35961	0.9902534113	Y	919	scaffold77`1	35961	0.9879951981
42	lac170	Y	445	scaffold15`1	24084	0.9860627178	Y	634	scaffold26`1	13528	0.9941634241
43	lac193	Y	601	scaffold135`1	35915	1	Y	679	scaffold135`1	35915	0.9982905983
44	lac20	Y	734	scaffold15`1	24084	0.9918032787	Y	962	scaffold26`1	13528	0.9844074844
45	lac24B	Y	653	no hit			Y	569	scaffold27`2	5893	0.9857397504
46	lac27B	Y	570	no hit			Y	809	scaffold27`2	5893	0.9937578027
47	lac35B	Y	855	scaffold97`1	1604	1	Y	686	no hit		
48	lac36W	Y	723	scaffold58`1	34351	0.9461325967	Y	467	scaffold58`1	34351	0.9892933619
49	lac55	Y	1096	scaffold150`1	1502	0.9657407407	Y	810	scaffold150`1	1502	0.9888888889
50	lac71	Y	501	scaffold23`2	3999	0.9893333333	Y	814	scaffold15`1	24084	0.9926199262
51	lac82	Y	493	scaffold71`1	701	0.9909090909	N				
52	lac84	Y	575	scaffold107`1	15739	0.9808362369	Y	406	scaffold107`1	15739	0.9971346705
53	Mel-125	Y	847	scaffold138`1	11638	0.9847645429	Y	436	scaffold138`3	21002	0.9794721408
54	Mel-126	Y	479	scaffold100`1	24125	0.9971590909	Y	516	scaffold205`1	7863	0.9926289926
55	PO3	Y	570	scaffold42`1	32512	0.994011976	Y	421	scaffold42`1	32512	1
56	RCX18	Y	961	scaffold6`1	5837	0.9853095488	Y	1047	scaffold39`1	28772	0.9912790698
57	RCX2	Y	1075	scaffold194`1	6764	0.9704433498	Y	574	scaffold194`4	19345	0.9891304348
58	RCX24	Y	811	scaffold108`1	43364	0.9899874844	Y	956	scaffold108`1	43364	0.9957805907
59	RCX25	N					Y	1039	scaffold25`1	31529	0.9889558233
60	RCX28	Y	561	scaffold43`1	35659	0.9746835443	Y	1076	scaffold43`2	3459	0.9885167464
61	RCX31	Y	840	scaffold32`1	39632	0.9401197605	Y	1213	scaffold32`1	39632	0.9634042553
62	RCX32	Y	538	scaffold206`1	3977	0.9961759082	Y	1045	scaffold88`1	15979	0.9884281581
63	RCX6	Y	602	scaffold31`2	26136	0.9956709957	N				
64	RCX7	Y	803	scaffold201`3	10607	0.9936788875	N				
65	RCX8	Y	801	scaffold30`1	33899	0.9885931559	Y	1005	scaffold30`1	33899	0.9867617108
66	RCX9	N					Y	643	scaffold46`1	23784	0.9919614148
67	RCX92	Y	1169	scaffold2`1	34796	0.9759572573	Y	1039	scaffold2`1	34796	0.982230997
68	PCX9M1	Y	526	scaffold181`1	37968	0.9633204633	Y	1228	scaffold181`1	37968	0.9885462555
69	PCX9M3	Y	524	scaffold55`2	7811	0.9595375723	Y	970	scaffold21`1	22427	0.9823651452
70	PCX9M5	Y	805	scaffold239`1	19889	0.9874529486	Y	1012	scaffold244`1	17171	0.9791459782
71	Xyl 2	Y	218	scaffold188`1	1491	0.9770642202	Y	209	scaffold61`1	1834	0.9959349593
72	Xyl 3	Y	683	scaffold87`1	4981	0.9939577039	Y	924	scaffold14`1	3665	0.9869423286
73	Xyl 4	Y	641	scaffold177`1	32570	0.9789644013	Y	180	scaffold177`1	32570	0.975

3.6.7 *E. coli* genomic DNA contamination analysis

Because contamination of samples with *E. coli* genomic DNA was found to affect downstream assembly of barcoded samples, raw data were used to estimate percent contamination. The genome of *E. coli* DH1 (Genbank accession CP001637) was used as a reference, being the parent of DH5α, the strain used in the lab for cosmid propagation. All sequence reads were examined for similarity to the DH1 genome, using a criterion of 100% identity. Contamination ranged from 1% to approximately 50% in the barcoded samples (Figure 3.2) and 5% in the pooled sample (data not shown).

3.6.8 Read depth analysis

Read depth was estimated for each clone, for both barcoded sequencing and pooled sequencing. In both cases, the barcoded clone sequence was used as the reference sequence; raw reads were aligned to the reference sequence using BWA version 0.7.6a [180] and depth at each base was counted using SAMtools version 0.1.18 [181]. Average read depth for each clone was calculated (Figure 3.3) as well as read depth at every base across each clone (Appendix B.1).

3.6.9 Clone sequence similarity analysis

Sequence similarity was estimated for all clones using BLAST [4] on the barcoded reference sequences, specifically blastn with an e-value cut-off of 0.001. In each pairwise comparison, the total alignment length was divided by the shorter clone length to obtain a similarity value between 0 and 1. Clones with no sequence similarity identifiable by BLAST were assigned a similarity value of 0.

3.6.10 Data availability

Raw sequence data are available at the NCBI Sequence Read Archive under Study SRP031898. Accession numbers for all SRA Experiments are provided (Table 3.10) as are Sanger end sequences for the pooled sequencing strategy (http://www.cm2bl.org/~data) and barcode information for the barcoded sequencing strategy (Table 3.8). In addition, raw data and relevant information for both barcoded and pooled sequencing may be accessed online: http://www.cm2bl.org/~data

Table 3.10: Accession numbers for datasets uploaded to NCBI SRA. [167]

NCBI Experiment Title	NCBI SRA Experiment Accession Number
Pooled sequencing of cosmid clones from metagenomic libraries	SRX367531
MetaMicrobiome-AAACAT	SRX375037
MetaMicrobiome-AAAGCA	SRX375038
MetaMicrobiome-AAATGC	SRX375039
MetaMicrobiome-AACCCC	SRX375040
MetaMicrobiome-AACTTG	SRX375041
MetaMicrobiome-AAGACT	SRX375042
MetaMicrobiome-AAGCGA	SRX375043
MetaMicrobiome-AAGGAC	SRX375044
MetaMicrobiome-AATAGG	SRX375045
MetaMicrobiome-ACAAAC	SRX375046
MetaMicrobiome-ACAGTG	SRX375047
MetaMicrobiome-ACATCT	SRX375048
MetaMicrobiome-ACCCAG	SRX375049
MetaMicrobiome-ACCGGC	SRX375050
MetaMicrobiome-ACGATA	SRX375051
MetaMicrobiome-ACTCTC	SRX375052
MetaMicrobiome-ACTGAT	SRX375053
MetaMicrobiome-ACTTGA	SRX375054
MetaMicrobiome-AGAAGA	SRX375055
MetaMicrobiome-AGATAG	SRX375056
MetaMicrobiome-AGCATC	SRX375057
MetaMicrobiome-AGCGCT	SRX375058
MetaMicrobiome-AGGCCG	SRX375059
MetaMicrobiome-AGGTTT	SRX375060
MetaMicrobiome-AGTTCC	SRX375061
MetaMicrobiome-ATAATT	SRX375062
MetaMicrobiome-ATACGG	SRX375063
MetaMicrobiome-ATCACG	SRX375064
MetaMicrobiome-ATCCTA	SRX375065
MetaMicrobiome-ATCTAT	SRX375066
MetaMicrobiome-ATGAGC	SRX375067
MetaMicrobiome-ATGTCA	SRX375068
MetaMicrobiome-ATTCCT	SRX375069
MetaMicrobiome-CAAAAG	SRX375070
MetaMicrobiome-CAACTA	SRX375071
MetaMicrobiome-CACCGG	SRX375072
MetaMicrobiome-CACGAT	SRX375073
MetaMicrobiome-CACTCA	SRX375074
MetaMicrobiome-CAGATC	SRX375075
MetaMicrobiome-CAGGCG	SRX375076
MetaMicrobiome-CATGGC	SRX375077
MetaMicrobiome-CATTTT	SRX375078
MetaMicrobiome-CCAACA	SRX375079
MetaMicrobiome-CCACGC	SRX375080
MetaMicrobiome-CCCATG	SRX375081
MetaMicrobiome-CCGCAA	SRX375082
MetaMicrobiome-CCGTCC	SRX375083
MetaMicrobiome-CCTTAG	SRX375084
MetaMicrobiome-CGAGAA	SRX375085
MetaMicrobiome-CGATGT	SRX375086
MetaMicrobiome-CGGAAT	SRX375087
MetaMicrobiome-CGTACG	SRX375088
MetaMicrobiome-CTAGCT	SRX375089
MetaMicrobiome-CTATAC	SRX375090
MetaMicrobiome-CTCAGA	SRX375091
MetaMicrobiome-CTGCTG	SRX375092
MetaMicrobiome-CTTGTA	SRX375093
MetaMicrobiome-GAAACC	SRX375094
MetaMicrobiome-GAATAA	SRX375095
MetaMicrobiome-GACGGA	SRX375096
MetaMicrobiome-GAGTGG	SRX375097
MetaMicrobiome-GATATA	SRX375098
MetaMicrobiome-GATCAG	SRX375099
MetaMicrobiome-GATGCT	SRX375100
MetaMicrobiome-GCAAGG	SRX375101
MetaMicrobiome-GCACTT	SRX375102
MetaMicrobiome-GCCAAT	SRX375103
MetaMicrobiome-GCCGCG	SRX375104
MetaMicrobiome-GCCTTA	SRX375105
MetaMicrobiome-GCTCCA	SRX375106
MetaMicrobiome-GGCACA	SRX375107
MetaMicrobiome-GGCTAC	SRX375108
MetaMicrobiome-GTAGAG	SRX375109
MetaMicrobiome-GTCCGC	SRX375110
MetaMicrobiome-GTGAAA	SRX375111
MetaMicrobiome-GTGGCC	SRX375112
MetaMicrobiome-TAATCG	SRX375113
MetaMicrobiome-TACAGC	SRX375114
MetaMicrobiome-TAGCTT	SRX375115
MetaMicrobiome-TATAAT	SRX375116
MetaMicrobiome-TCATTC	SRX375117
MetaMicrobiome-TCCCGA	SRX375118
MetaMicrobiome-TCGAAG	SRX375119
MetaMicrobiome-TCGGCA	SRX375120
MetaMicrobiome-TGAATG	SRX375121
MetaMicrobiome-TGACCA	SRX375122
MetaMicrobiome-TGCCAT	SRX375123
MetaMicrobiome-TGCTGG	SRX375124
MetaMicrobiome-TGGCGC	SRX375125
MetaMicrobiome-TTAGGC	SRX375126
MetaMicrobiome-TTCGAA	SRX375127
MetaMicrobiome-TTCTCC	SRX375128

Chapter 4

Analysis of cloning bias in metagenomic libraries

4.1 Acknowledgements and declarations

The work presented in this chapter was published as a Research article in the journal **Microbiome**. The citation for the article is:

> Lam KN, Charles TC (2015) Strong spurious transcription likely contributes to DNA insert bias in typical metagenomic clone libraries. *Microbiome* 3:22. doi: 10.1186/s40168-015-0086-5

Before publication, the content was also made publicly available as New Results on the pre-print server **bioRxiv**. The citation for the pre-print is:

> Lam KN, Charles TC (2015) Strong spurious transcription likely a cause of DNA insert bias in typical metagenomic clone libraries. *bioRxiv* doi: 10.1101/013763

Section 4.4.5 in the results of this chapter was published as part of a Perspective article in **Frontiers in Microbiology**. I was the primary author of this article. The citation for the article is:

> Lam KN, Cheng J, Engel K, Neufeld JD, Charles TC (2015) Current and future resources for functional metagenomics. *Frontiers in Microbiology* 6:1196. doi:10.3389/fmicb.2015.01196

I managed and performed all experiments and analyses described in this chapter with the following exceptions:

- In Section 4.4.5, **Jiujun Cheng** prepared DNA from the 12AC original soil sample and the corresponding metagenomic library.
- Also in Section 4.4.5, **Katja Engel** carried out V3 region PCR on these two samples and managed sequencing sample submission.

I also acknowledge the following contributions:

- The text of the *Microbiome* manuscript, largely duplicated here, was proofread and edited by my supervisor **Trevor Charles**.
- The text of the section from the *Frontiers in Microbiology* manuscript, largely duplicated here, was proofread and edited by **Katja Engel**, **Josh Neufeld**, **Trevor Charles**, and **Jiujun Cheng**.
- In the 16S rRNA gene analysis I carried out for the *Frontiers in Microbiology manuscript*, **Brent Seuradge** provided advice on using the AXIOME2 pipeline, **Michael J. Lynch** answered technical questions, and **Michael W. Hall** assisted in trouble-shoooting AXIOME2- and BIOM-related issues.

4.2 Abstract

Background: Clone libraries provide researchers with a powerful resource to study nucleic acid from diverse sources. Metagenomic clone libraries in particular have aided in studies of microbial biodiversity and function, and allowed the mining of novel enzymes. Libraries are often constructed by cloning large inserts into cosmid or fosmid vectors. Recently, there have been reports of GC bias in fosmid metagenomic libraries, and it was speculated to be a result of fragmentation and loss of AT-rich sequences during cloning. However, evidence in the literature suggests that transcriptional activity or gene product toxicity may play a role.

Results: To explore possible mechanisms responsible for sequence bias in clone libraries, I constructed a cosmid library from a human microbiome sample and sequenced DNA from different steps during library construction: crude extract DNA, size-selected DNA, and cosmid library DNA. I confirmed a GC bias in the final cosmid library, and provide evidence that the bias is not due to fragmentation and loss of AT-rich sequences but is likely occurring after DNA is introduced into *E. coli*. To investigate the influence of strong constitutive transcription, I searched the sequence data for consensus promoter sequences and found that *rpoD*/σ^{70} promoter sequences were underrepresented in the cosmid library. Furthermore, when I examined the genomes of taxa that were differentially abundant in the cosmid library relative to the original sample, I found the bias to be more correlated with the number of *rpoD*/σ^{70} consensus sequences in the genome than with simple GC content.

Conclusions: The GC bias of metagenomic libraries does not appear to be due to DNA fragmentation. Rather, analysis of promoter sequences provides support for the hypothesis that strong constitutive transcription from sequences recognized as *rpoD*/σ^{70} consensus-like in *E. coli* may lead to instability, causing loss of the plasmid or loss of the insert DNA that gives rise to the transcription. Despite widespread use of *E. coli* to propagate foreign DNA in metagenomic libraries, the effects of in vivo transcriptional activity on clone stability are not well understood. Further work is required to tease apart the effects of transcription from those of gene product toxicity.

4.3 Introduction

Clone libraries can be generated using a range of source material, from the DNA of a single organism to the DNA from environmental sources representing often complex microbial communities. Libraries generated from microbial communities are called metagenomic libraries, and they have been central to a powerful methodology contributing to understanding the diversity of microbial communities, expanding the knowledge of gene function, and mining for novel sequences encoding functions of interest. These activities all fall under the umbrella of functional metagenomics and require cloning the DNA, typically using low-copy vectors such as cosmids or fosmids. Cloned DNA is typically propagated in *E. coli*, and if the vector host range allows, the DNA can subsequently be transferred to other surrogate hosts that may be more suitable for heterologous expression.

4.3.1 Possible causes of sequence bias in metagenomic libraries

The general assumption in cloning-based metagenomic approaches is that foreign DNA can be stably maintained in *E. coli* and that the cloned DNA is a fair representation of the original sample. However, it has been previously observed that fosmid libraries exhibit a GC bias [54, 101, 304]. In general, such cloning biases may affect conclusions derived from analysis of the clone libraries. The observed GC bias of fosmid libraries was suggested to be due to fragmentation and subsequent loss of AT-rich sequences during the cloning process, purportedly because AT-rich sequences have fewer hydrogen bonds which makes them more vulnerable to non-perpendicular shear forces [304]. Other possible reasons for the bias in libraries include transcriptional activity of the cloned DNA [41] as well as toxicity from expressed genes [84, 287]. Though the exact mechanism(s) by which GC bias occurs has not yet been fully elucidated, the fragmenta-

tion explanation has been echoed by others [110, 192] despite being purely speculative and lacking experimental support. Indeed, in my own experience, extracting high-molecular-weight genomic DNA from low-GC organisms is no more difficult than from *E. coli*. I have previously constructed genomic libraries in cosmid vectors using DNA from *Bacteroides thetaiotaomicron* and *Bacteroides fragilis* (Table 2.12; both ∼43% GC) with no difficulties obtaining high-quality DNA (Figure 4.1) [167].

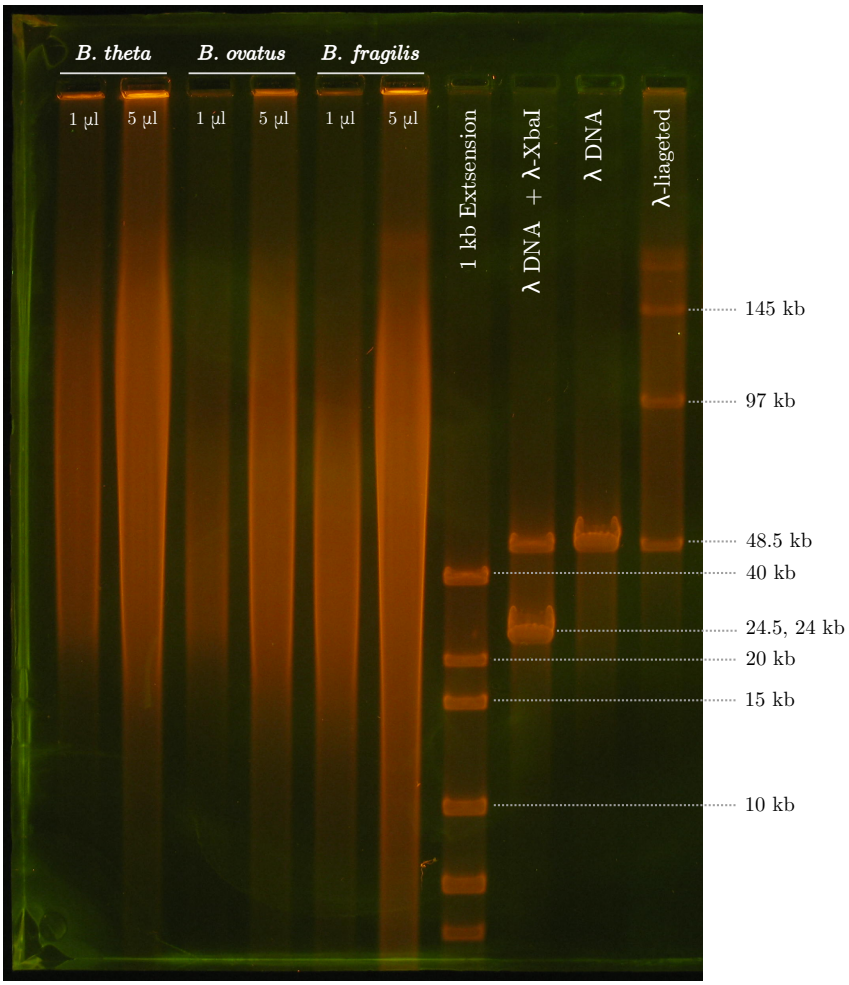


Figure 4.1: Pulsed-field gel electrophoresis of extracted *Bacteroides* genomic DNA. Genomic DNA extracted from *Bacteroides thetaiotaomicron*, *Bacteroides fragilis*, and *Bacteroides ovatus* was found to be high-molecular-weight by pulsed-field gel electrophoresis. For more details on the molecular markers, see Section 2.5.9.

Furthermore, in the Charles laboratory, we have observed that on occasion, cosmid clones from metagenomic libraries appear to have suffered insert loss, which is discussed in greater detail in the "Results and discussion" section below. Therefore, it seemed that the suggestion by Temperton et al. [304] that the GC bias in cosmid/fosmid libraries might be due to fragmentation of AT-rich sequences was unlikely to be true; rather, events occurring in vivo may be contributing substantially to the sequence bias of libraries.

4.3.2 Aims of this work

I investigated the nature of this GC bias, to characterize whether, and by what mechanism, biases may be introduced into the lab's own cosmid libraries. In particular, I wished to determine if fragmentation was a major cause of bias, or if there is evidence that the bias was indeed occurring in vivo. To answer this question, I constructed a cosmid library using DNA isolated from pooled human fecal samples, saving a portion of the DNA from three steps of the library construction process: (1) the crude extract DNA, (2) the size-selected DNA, and (3) the cloned DNA from the constructed cosmid library (Figure 4.2). The DNA samples were sequenced and the resulting datasets were analyzed to investigate if, where, and how any bias may have been introduced. Consistent with the aforementioned studies, I observed GC bias in the constructed cosmid library. However, the results indicate that fragmentation of DNA does not cause any significant bias; rather, the results are consistent with the hypothesis that the bias occurs after DNA is introduced into the *E. coli* host.

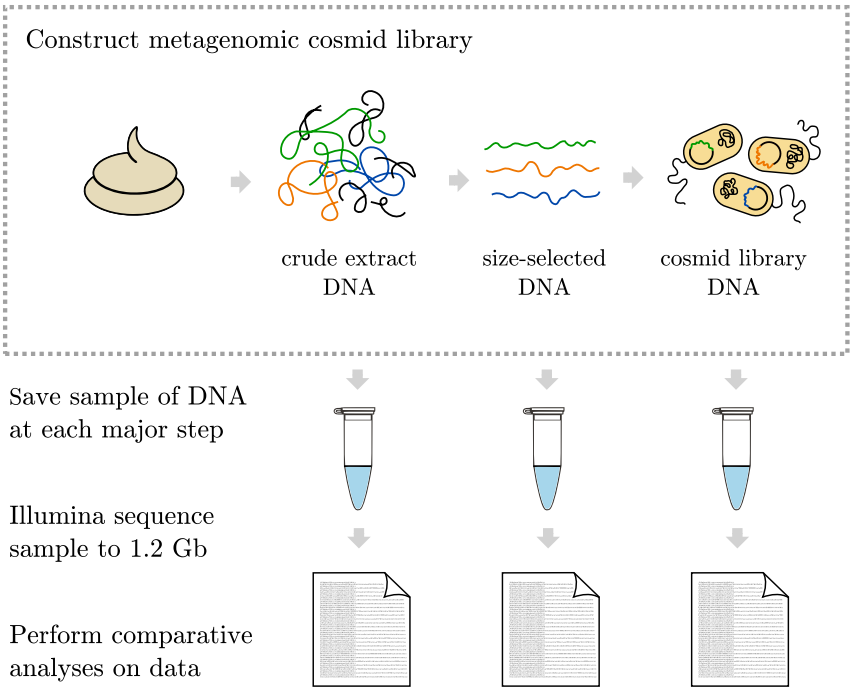


Figure 4.2: Overview of the experimental design for this library bias study. A pooled human fecal sample was used to construct a metagenomic cosmid library, during which DNA from three distinct steps was collected and sequenced in order to investigate possible sequence biases and at what steps the biases were introduced. [165]

4.4 Results and discussion

4.4.1 DNA sampling and sequencing results

I collected DNA at the three main steps of cosmid library construction: the crude extract DNA, the size-selected DNA, and the final cosmid library DNA (Figure 4.2). Before sequencing, I first checked the quality of each sample by gel electrophoresis (Figure 4.3). As expected, the crude extract was the only sample that contained a heavy smear of fragmented DNA; the selection for high-molecular-weight DNA greatly reduced fragmented DNA, as evidenced by its absence from the size-selected sample. The cosmid library sample exhibited the characteristic multiple banding pattern representing the various possible conformations of uncut circular DNA.

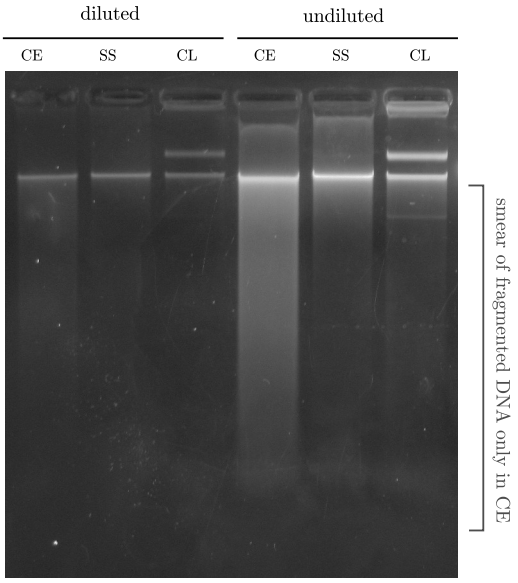


Figure 4.3: Gel electrophoresis of crude extract, size-selected, and cosmid library DNA samples. Diluted and undiluted amounts of each sample were gel electrophoresed for quality control check of DNA prior to Illumina sequencing. [165]

After confirming DNA quality, the samples were paired-end sequenced on an Illumina HiSeq 2000 platform, generating ∼1.2 Gb of DNA sequence per sample. It was expected that the cosmid library would be contaminated with *E. coli* genomic DNA and cosmid vector DNA as a result of (1) isolating cosmid DNA from *E. coli* cells and (2) the fact that each and every cosmid clone sequenced included its vector backbone. Thus, for fair treatment, I subtracted *E. coli* and pJC8 sequences from all samples (see "Methods" section). For *E. coli* and pJC8, respectively, 6701 and 164 reads were removed from crude extract data (∼0.05% of all reads); 9273 and 2410 from size-selected data (∼0.09%); and 851,410 and 2,130,004 from the cosmid library DNA (∼23%). As expected, the dataset originating from the cosmid library sample had the highest number of reads subtracted. Though the crude extract and size-selected samples contained a small amount, these likely represent true environmental sequences; however, their subtraction was necessary for equal treatment of all samples, and the small fraction removed should not affect overall conclusions from the data.

After host and vector sequence subtraction, I used Nonpareil [243] to estimate the overall sequencing coverage of the samples, which was ∼85% for the crude extract and size-selected samples and ∼95% for the cosmid library sample (Figure 4.4). Interestingly, despite one-quarter of reads in the cosmid library sample being from *E. coli* or pJC8, this sample appeared to have the best sequencing coverage, suggesting that the cosmid library suffered a decrease in diversity as a result of cloning bias. Overall, the relatively high sequencing coverage for all three samples was sufficient for the downstream comparative sequence analyses; for all subsequent results discussed here, the forward and reverse sequencing reads for the three samples were analyzed separately.

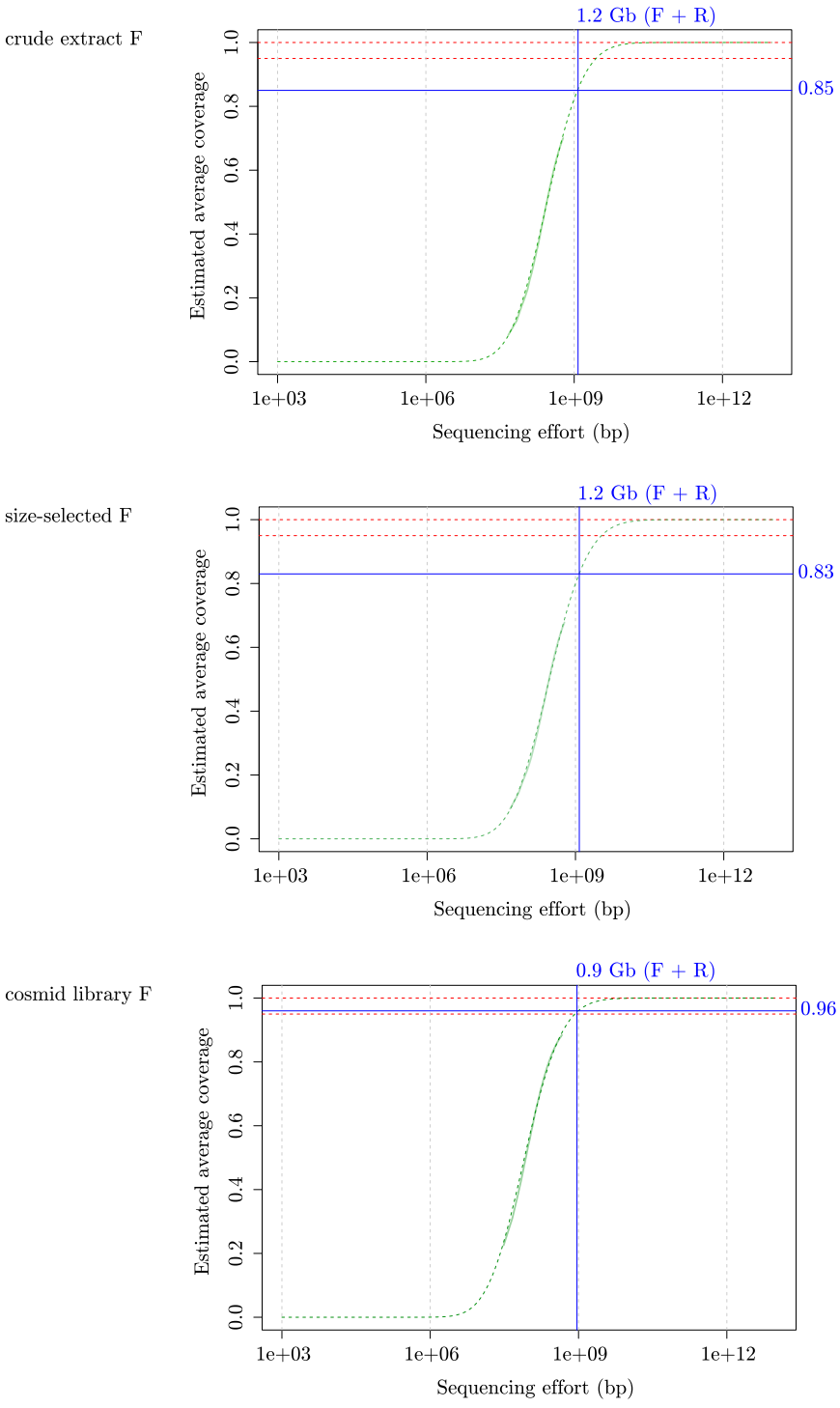


Figure 4.4: Estimate of sample sequencing coverage using Nonpareil. The software Nonpareil was used to estimate sequencing coverage for each of the three samples. The software takes a sequence data file as input and, based on the redundancy of the reads, calculates curves of coverage versus sequencing effort. [165]

4.4.2 GC bias is not caused by fragmentation of AT-rich DNA

The experimental design (Figure 4.2) was such that I could address whether the bias in the metagenomic library was due to fragmentation of DNA during cloning. Because both crude extract and size-selected samples were sequenced, I could determine whether the removed fragmented DNA from the crude extract (visible in Figure 4.3) led to a bias in the size-selected DNA sample. I calculated the percent GC in each of the three datasets and found that the GC bias was only present in the final cosmid library and not the size-selected sample (Table 4.1), effectively ruling out fragmentation as the mechanism for cosmid library bias.

Table 4.1: Percent GC of crude extract, size-selected, and cosmid library datasets. GC content was calculated after subtraction of *E. coli* and vector DNA from all samples. [165]

Sample/dataset	No. reads	No. Mb	%GC
Crude extract F	6,654,484	599	47.7
Crude extract R	6,654,567	599	47.8
Size-selected F	6,645,306	598	46.9
Size-selected R	6,645,817	598	46.9
Cosmid library F	5,134,020	462	53.0
Cosmid library R	5,191,538	467	53.1

After confirming that the bias occurs post size selection, I next asked if certain taxa were differentially represented across the samples to see if this would point to a possible reason for library sequence bias. I used Taxy [209] as well as Taxy-Pro [154] as part of the CoMet web server [189] to do a fast preliminary comparison of taxa abundance across the three different samples. Taxy calculates k-mer frequencies for the

dataset and then uses mixture modeling of k-mer frequencies of sequenced genomes to obtain a profile similar to that of the sample, whereas Taxy-Pro has a similar modeling approach but uses protein domains rather than k-mer frequencies. Both tools generated very similar profiles for the crude extract and the size-selected DNA but a very different profile for the cosmid library DNA (see Figure 4.5 for Taxy results), supporting the percent GC results.

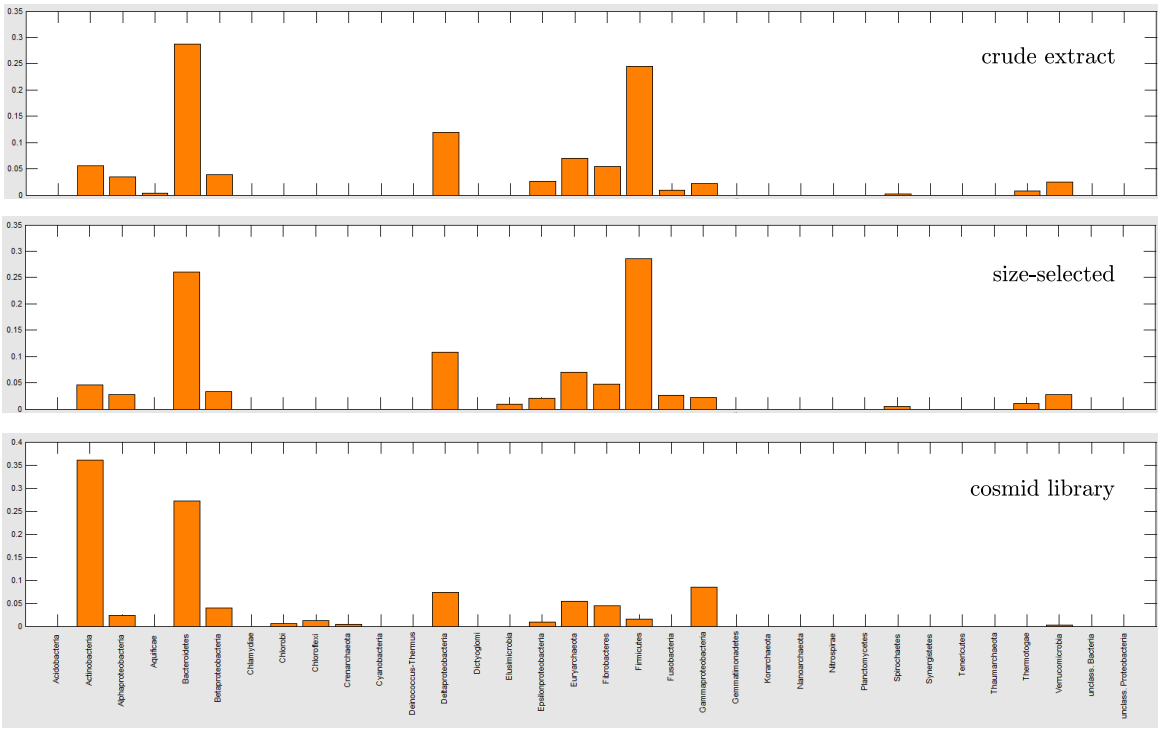


Figure 4.5: Distribution of bacterial phyla predicted by Taxy. The software Taxy was used to estimate the distribution of bacterial phyla in each of the three samples, using a k-mer length of 7.

With positive results from this preliminary work, I then performed more thorough taxonomic analyses using two different approaches; in the first, all sequencing reads were used, and in the second, only the 16S rRNA gene-containing reads were used.

In the first approach, I used the Metagenome Phylogenetic Analysis (MetaPhlAn) tool, a profiling tool that maps reads against clade-specific marker sequences [258] to estimate sample composition down to the species level (see Appendix C.1 for summary table of MetaPhlAn output). I examined the abundance of the top four most common phyla in human gut metagenomes to see whether there were large overall changes in taxa abundance across the samples (Figure 4.6). The crude extract and size-selected samples showed high Firmicutes and Bacteroidetes content with lower levels of Actinobacteria and Proteobacteria, compositions that are typical of gut-derived samples [72, 197, 288]. Notably, these results indicated that that DNA from the Firmicutes was nearly absent in the cosmid library sample, accompanied by an equivalent increase in the Actinobacteria. These results were consistent with the percent GC analysis, as members of the Firmicutes phylum are generally known to be low-GC, and those of the Actinobacteria, high-GC [97, 188].

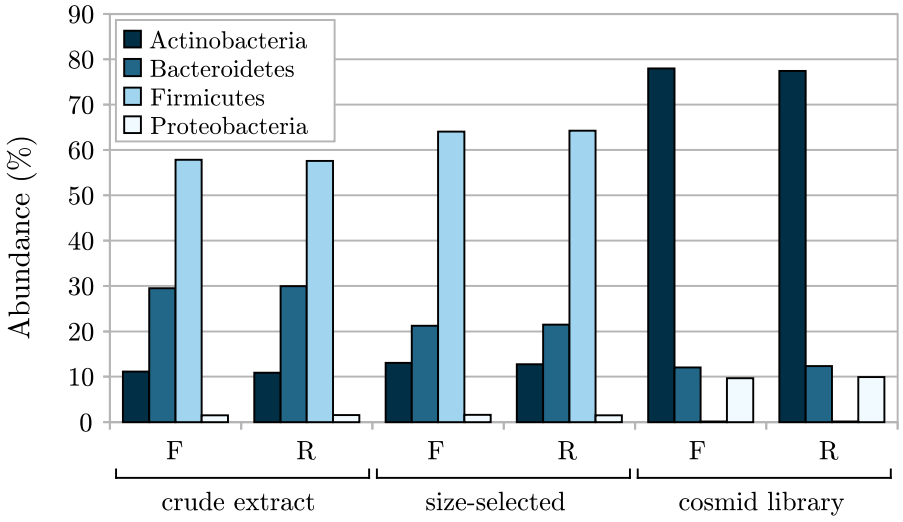


Figure 4.6: Histogram of abundance of the top four phyla in crude extract, size-selected, and cosmid library samples. Abundance of the Actinobacteria, Bacteroidetes, Firmicutes, and Proteobacteria phyla in each sample, as determined using MetaPhlAn. [165]

I also examined the MetaPhlAn results at the species level to see which genomes may be under- or overrepresented in the cosmid library, choosing to examine the top 50 most differentially abundant species (Figure 4.7). Several members of the Bifidobacterium genus were substantially overrepresented in the cosmid library while many members of the Firmicutes were completely or very nearly lost; for example, *Eubacterium rectale*, *Ruminococcus bromii*, and *Faecalibacterium prausnitzii* were all highly abundant in the original sample.

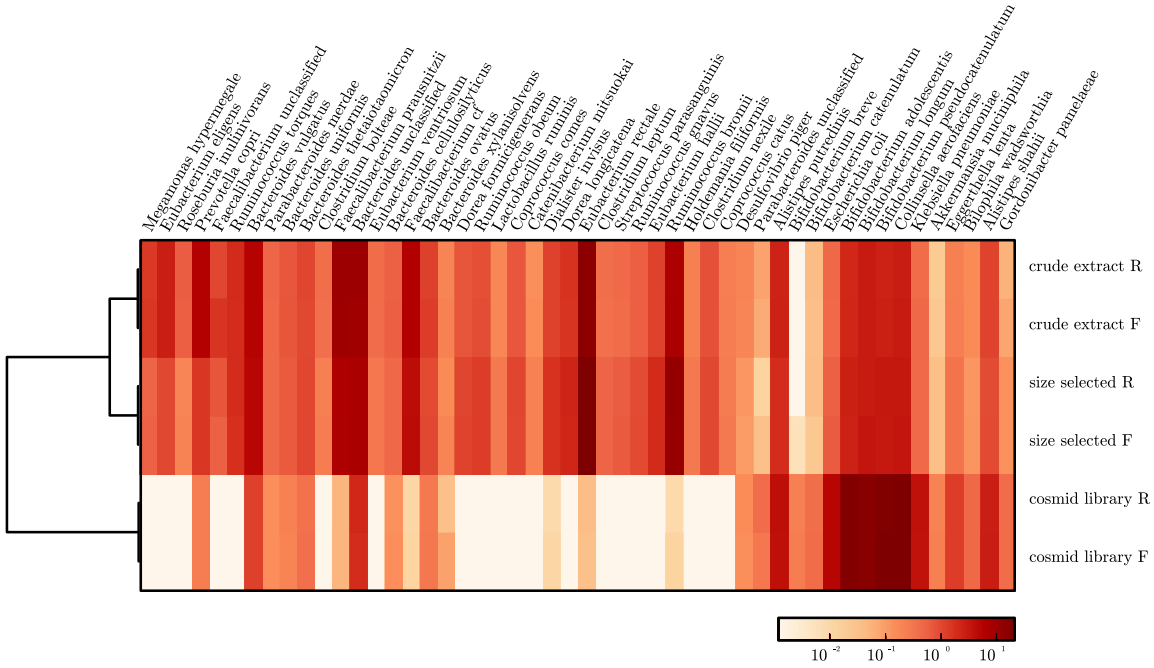


Figure 4.7: Heatmap of 50 species with differential abundance across crude extract, size-selected, and cosmid library samples. Abundance in each sample of the top 50 species determined to be differentially abundant using MetaPhlAn. Abundance is depicted on a log scale. [165]

In the second approach, I identified reads in the datasets that were from the 16S rRNA gene, and used the RDP classifier to classify these to the genus level (Figure 4.8). I found that analyses using only 16S rRNA gene-containing reads showed high agreement with analyses carried out using all reads (i.e., Figure 4.6), indicating that 16S rRNA gene content tracks well with genomic content in large-insert cosmid libraries. Both approaches – using all reads or only reads from the 16S rRNA gene – provided similar results, and both were in agreement with percent GC, Taxy, and Taxy-Pro results, all of which provide compelling evidence that cosmid library biases are not due to fragmentation of AT-rich sequences during the cloning process.

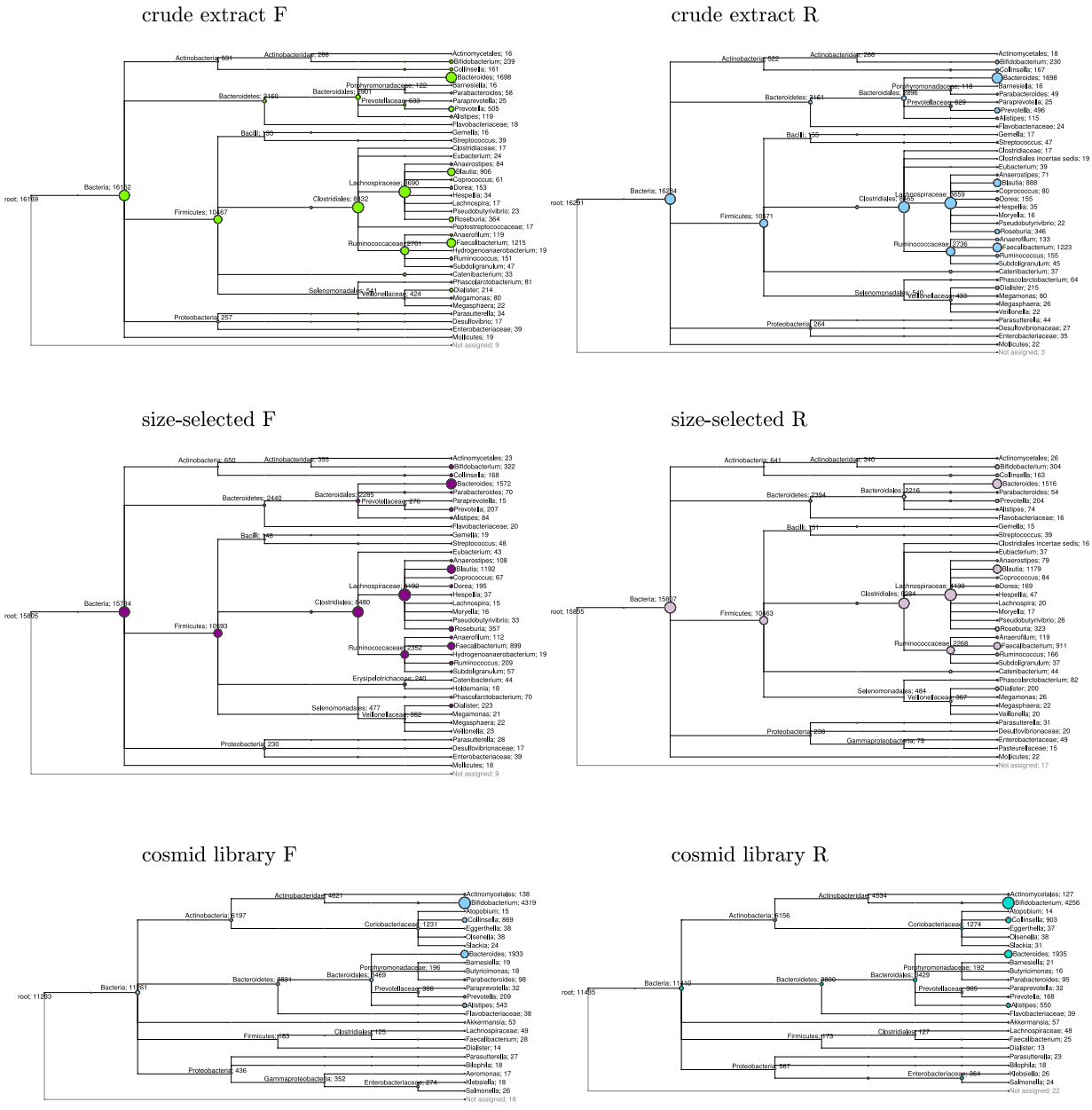


Figure 4.8: 16S rRNA gene analysis results using Infernal for identification of 16S-containing reads, RDP classifier to classify reads, and MEGAN for visualization of results. 16S rRNA sequences from forward and reverse datasets were classified for all three samples. [165]

4.4.3 GC content may be a proxy for *E. coli* σ^{70} promoter content

From these results, our laboratory's own experiences, and what was previously known in the literature, there was reason to suspect that the cause of the bias occurred in vivo. I wondered whether these AT-rich sequences might have a regulatory role in vivo and noticed that they may resemble the constitutive *E. coli* promoter, and in fact, I am not the first to suggest this resemblance [64, 218], particularly of the −10 Pribnow box (Figure 4.9).

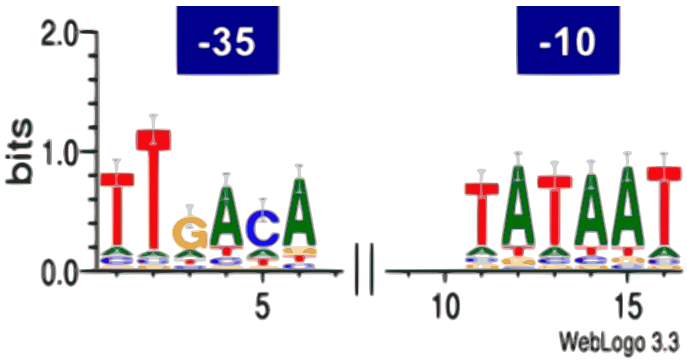


Figure 4.9: Sequence logo of *rpoD*/σ^{70} promoter consensus. The consensus sequence for *rpoD*/σ^{70} promoters is AT-rich. Adapted from [262]

To investigate whether transcription of the insert may be having a negative effect on its maintenance by the host cell, I analyzed the sequence data from the three samples for *E. coli* consensus promoter sequences; in particular, I was interested in examining the data for differential abundance of the *rpoD*/σ^{70} consensus sequence, as σ^{70} is the "house-keeping" sigma factor whose promoters are constitutive.

In my analysis, I used the known promoter consensus sequence for *rpoD*/σ^{70} [262], and, as negative controls, I used the consensus sequence for: *rpoE*/σ^{24} [241]; *rpoH*/σ^{32} [224]; *rpoN*/σ^{54}, which has a GC-rich consensus [346]; as well as the primary sigma factor of *Bacteroides*, σ^{ABfr} [15], because the *Bacteroides* genus had comparable abundance across the three samples (Figure 4.8) and because *Bacteroides* constitutive promoters are not recognized by *E. coli* [206]. I examined each of the three samples for relative abundance of these five consensus sequences; consensus sequences are provided in Table 4.2.

Table 4.2: Consensus promoter sequences for selected sigma factors.

Sigma factor	**Consensus sequence**	**Ref.**
rpoD (σ^{70})	TTGACAN$_{15-19}$TATAAT	[262]
rpoE (σ^{24})	GGAACTTN$_{15-19}$TCAAA	[241]
rpoH (σ^{32})	TTG[A/T][A/T][A/T]N$_{13-14}$CCCCAT[A/T]T	[224]
rpoN (σ^{54})	TGGCAN$_7$TGC	[346]
Bacteroides (σ^{ABfr})	TTTGN$_{19-21}$TAN$_2$TTTG	[15]

The results showed that while the crude extract and size-selected samples had similar promoter content profiles, the cosmid library exhibited a deviation (Figure 4.10). Supporting the hypothesis, only the *rpoD* consensus content was considerably different in abundance, by about an order of magnitude when compared to either the crude extract or size-selected sample.

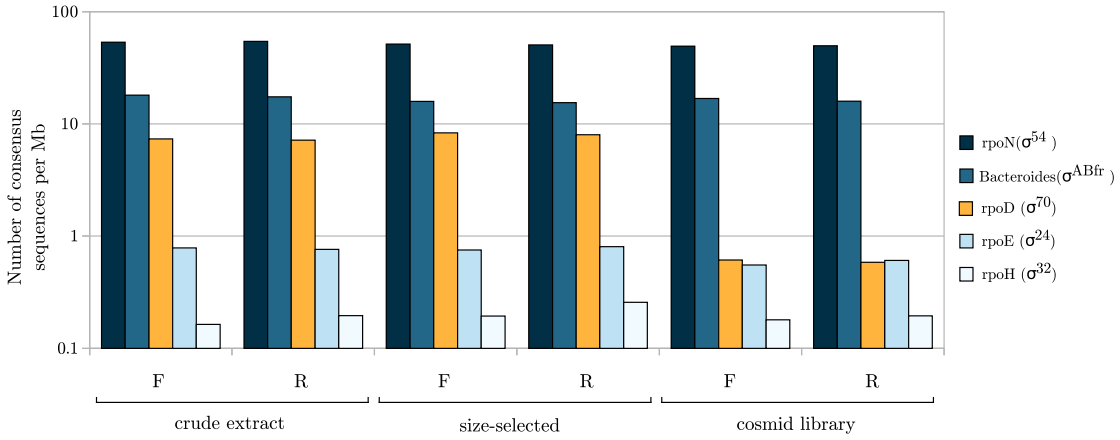


Figure 4.10: Histogram of sigma factor consensus sequence content in crude extract, size-selected, and cosmid library samples. Bars indicate the number of consensus sequences in each sample, for select *E. coli* sigma factors and the Bacteroides primary sigma factor, normalized to the amount of sequence data for that sample. Consensus content is depicted on a log scale. [165]

The loss of these specific sequences from the cosmid library suggests that the widely used cloning host *E. coli* may be problematic for cosmid-cloned fragments of DNA that incidentally contain constitutively active *rpoD* promoter sequences; indeed, these findings are supported by previous reports in the literature, which is discussed in more detail in the following section. If *E. coli* does in fact exclude constitutively active *rpoD* promoter-containing sequences, simply switching to a different cloning/library host (even if it were possible) would likely alleviate one problem only to introduce another, as all organisms have sequences from which constitutive transcription arises. It may be that multiple backgrounds, with different constitutively active sequences, are required for the maintenance of metagenomic libraries in an effort to increase sample representativeness.

Given that *rpoD* promoter sequences were underrepresented in the cosmid library and that certain species appear to be over- or underrepresented, I next asked whether a species' abundance in the cosmid library could be predicted from the *rpoD* consensus content of its genome. And in particular, is *rpoD* consensus content more predictive of library abundance than simple GC content?

To answer these questions, I turned to the results of the MetaPhlAn analysis, which gave me a list of the top 50 most differentially abundant species (Figure 4.7). To analyze the genomes of these species for possible sequence determinants of library abundance, I used the NCBI Genome database to find sequenced representatives of each species where possible and was able to retrieve 46 genomes (complete, draft, or whole genome shotgun sequences; see Section 4.6.6 for details); for each genome, I calculated the percent GC as well as the number of *rpoD* consensus promoter sequences present (Table 4.3).

Table 4.3: Length, percent GC, and *rpoD* consensus content of the 46 genomes. [165]

Filename	Length	GC%	rpod
Akkermansia˙muciniphila˙ATCC˙BAA-835.fasta	2664102	55.76	1
Alistipes˙putredinis˙DSM˙17216˙Scfld.fasta	2550678	53.25	2
Alistipes˙shahii˙WAL˙8301˙draft.fasta	3763317	55.82	3
Bacteroides˙cellulosilyticus˙DSM˙14838˙genomic˙scaffold.fasta	6870144	41.81	12
Bacteroides˙ovatus˙SD˙CMC˙3f˙contig.fasta	6775279	41.94	11
Bacteroides˙thetaiotaomicron˙VPI-5482.fasta	6260361	42.84	8
Bacteroides˙uniformis˙ATCC˙8492˙Scfld.fasta	4719097	46.43	5
Bacteroides˙vulgatus˙ATCC˙8482.fasta	5163189	42.2	15
Bacteroides˙xylanisolvens˙XB1A˙draft.fasta	5976145	40.67	13
Bifidobacterium˙adolescentis˙ATCC˙15703.fasta	2089645	59.18	0
Bifidobacterium˙breve˙ACS-071-V-Sch8b.fasta	2327492	58.73	1
Bifidobacterium˙catenulatum˙DSM 16992˙B˙catenulatum-1.0˙Cont.fasta	2058429	56.1	0
Bifidobacterium˙longum˙NCC2705.fasta	2256640	60.12	0
Bifidobacterium˙pseudocatenulatum˙DSM˙20438˙B˙pseudocatenulatum-1.0.1˙Cont.fasta	2304808	56.28	2
Bilophila˙wadsworthia˙3˙1˙6˙genomic˙scaffold.fasta	4391194	59.31	3
Catenibacterium˙mitsuokai˙DSM˙15897˙C˙mitsuokai-1.0˙Cont.fasta	2671313	36.82	30
Clostridium˙bolteae˙ATCC˙BAA-613˙Scfld.fasta	6557988	49.05	87
Clostridium˙leptum˙DSM˙753˙Scfld.fasta	3270209	50.18	63
Clostridium˙nexile˙DSM˙1787˙Scfld.fasta	3995628	38.74	53
Collinsella˙aerofaciens˙ATCC˙25986˙C˙aerofaciens-2.0˙Cont.fasta	2439869	60.55	1
Coprococcus˙catus˙GD˙7˙draft.fasta	3522704	42.43	67
Coprococcus˙comes˙ATCC˙27758˙genomic˙scaffold.fasta	3242215	42.45	50
Desulfovibrio˙piger˙ATCC˙29098˙Scfld.fasta	2867216	62.15	4
Dialister˙invisus˙DSM˙15470˙genomic scaffold.fasta	1895960	45.49	23
Dorea˙formicigenerans˙ATCC˙27755 D˙formicigenerans-3.0.1˙Cont.fasta	3186031	40.97	51
Dorea˙longicatena˙DSM˙13814˙Scfld.fasta	2915433	41.42	37
Eggerthella˙lenta˙DSM˙2243.fasta	3632260	64.2	1
Escherichia˙coli˙str˙K-12 substr˙MG1655.fasta	4641652	50.79	1
Eubacterium˙eligens˙ATCC˙27750.fasta	2144190	37.71	32
Eubacterium˙hallii˙DSM˙3353˙E˙hallii-1.0˙Cont.fasta	3290996	38.19	56
Eubacterium˙rectale˙ATCC˙33656.fasta	3449685	41.48	43
Eubacterium˙ventriosum˙ATCC˙27560˙Scfld.fasta	2870795	34.9	38
Faecalibacterium˙prausnitzii˙L2-6.fasta	3321367	55.57	48
Gordonibacter˙pamelaeae˙7-10-1-b˙draft.fasta	3608022	60.43	1
Holdemania˙filiformis˙DSM˙12042˙genomic˙scaffold.fasta	3932923	48.54	31
Klebsiella˙pneumoniae˙subsp˙pneumoniae˙HS11286.fasta	5333942	57.48	0
Lactobacillus˙ruminis˙ATCC˙27782.fasta	2066652	43.47	16
Megamonas˙hypermegale˙ART12˙1˙draft.fasta	2209938	30.51	55
Parabacteroides˙merdae˙ATCC˙43184˙Scfld.fasta	4434377	45.28	12
Prevotella˙copri˙DSM˙18205˙genomic˙scaffold.fasta	3512473	44.8	4
Roseburia˙inulinivorans˙DSM˙16841˙R˙inulinivorans-1.0.1˙Cont.fasta	4048462	41.93	54
Ruminococcus˙bromii˙L2-63˙draft.fasta	2249085	41.05	77
Ruminococcus˙gnavus˙ATCC˙29149˙R˙gnavus-1.0.1˙Cont.fasta	3501911	42.88	56
Ruminococcus˙obeum˙A2-162˙draft.fasta	3757491	41.75	58
Ruminococcus˙torques˙L2-14˙draft.fasta	3341681	40.14	34
Streptococcus˙parasanguinis˙ATCC˙15912.fasta	2153652	41.72	16

Next, to quantify bias in the cosmid library relative to the original sample (the crude extract), I calculated the change in abundance of the 46 species (using the average abundance of the forward and reverse datasets). I then plotted the change in abundance first against genome percent GC (Figure 4.10A) and second against *rpoD* consensus content, normalizing to genome size (Figure 4.10B). The results show that while library bias only generally correlates with GC content, library bias correlates surprisingly well with the *rpoD* consensus content of the genome.

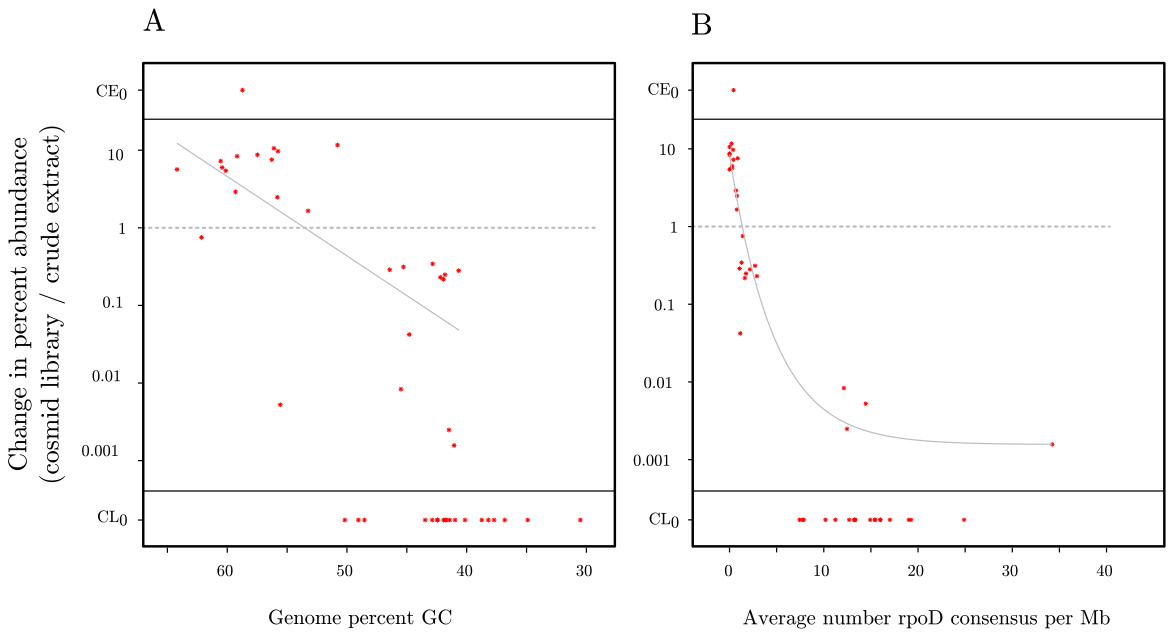


Figure 4.11: Bias in cosmid library relative to crude extract, against GC content or *rpoD* consensus content. Species abundance was obtained from MetaPhlAn analysis of the crude extract and cosmid library samples. Bias is calculated as change in percent abundance (cosmid library abundance / crude extract abundance) plotted against GC content (a) or *rpoD* consensus content (b). Change in abundance is depicted on a log scale; CE_0 values indicate zero abundance in the crude extract sample and CL_0 values indicate zero abundance in the cosmid library sample, as predicted by MetaPhlAn. [165]

These results suggest that GC content may be only a rough proxy for *rpoD* consensus content (as *rpoD* consensus sequences are AT-rich), but GC content itself may not be an accurate predictor of library presence/abundance; indeed, in some cases, a genome may have a moderate or relatively high percent GC but also possess an unusually high *rpoD* consensus content, leading to an underrepresentation in the cosmid library that could not have been predicted from GC content alone (Figure 4.10). These results are also consistent with the previous observation that library bias was more obvious among organisms with low GC content [54] because AT-rich genomes would have an increased number of *rpoD* promoter-like sequences simply by chance [219].

4.4.4 Examining the published literature: evidence for transcriptional activity of cloned AT-rich DNA interfering with stability of circular vectors

This chapter describes analyses concerning metagenomic DNA. However, if there are *rpoD* consensus-like sequences that are interfering with the maintenance of foreign DNA in *E. coli*, then the scope of the problem extends beyond metagenomics applications. Curious about the extent of the problem, I performed literature searches to find reports of experienced difficulties cloning AT-rich DNA and/or investigations of possible mechanisms for those difficulties. My search was fruitful, leading us to literature that spans the past three decades.

It was reported that there are difficulties associated with cosmid cloning of very AT-rich genomic DNA [99, 106], and even when genomic libraries can be constructed, cosmid clones may be unstable [27, 120, 240, 265], which simply means that foreign DNA fragments are not able to be maintained in the *E. coli* library host. Thus, if selection is applied for a marker present on the vector, then in vivo events may lead

to insert deletion, which has been observed by our lab as well as others, despite using a host that is a *recA* mutant [265]. This is particularly evident when the library is constructed using a high-copy number vector (e.g., one containing a ColE1-type origin of replication), which has been experienced by our lab (Figure 5.10) and others [40] and is in agreement with the observation that F-based, single-copy fosmids perform better than multi-copy cosmids at stably maintaining insert DNA [148]. Loss of cloned sequence is even more widespread for inserts that have repetitive DNA sequences [33], as such sequences may be conducive to recombination. One way to combat insert loss is by minimizing outgrowth of the library-containing cells as much as possible [265], though this is not always feasible for shared cosmid libraries such as the Canadian MetaMicroBiome Library collection [222], which require outgrowth to generate stocks for sharing with the scientific community.

But what is the mechanism for plasmid instability? It was previously shown that transcriptional activity from a cloned strong promoter could affect plasmid stability by (1) interfering with the origin of replication via transcriptional read-through into the vector as well as (2) changing the abundance of protein products involved in plasmid copy number. Furthermore, plasmid instability was alleviated by placing transcriptional terminator sequences that flank the multiple cloning site [291]. It was also observed that strong phage promoters could only be cloned into plasmids that possess a downstream termination signal [100, 162]. Similarly, AT-rich pneumococcal DNA was found to contain a high incidence of *E. coli* strong promoter sequences, and that cloning of the DNA was improved by using a vector with efficient transcriptional terminators [40, 41, 289], although analysis of a set of pneumococcal promoter-containing sequences indicated that transcription strong enough to interfere with plasmid stability may be relatively rare and that other factors could be contributing to cloning difficulty [61].

Another consideration is that efficient transcription of poly-dT (as well as poly-dG) DNA tracts may cause the DNA to form a stable complex with its own accumulated transcription products, leading to transcriptional stalling that may interfere with the replication fork [152, 153, 160]. One particularly interesting observation that has surprisingly not attracted more interest is that linear cloning vectors with transcriptional terminators provide even more stability than circular vectors with transcriptional terminators [106, 107]. The advantage of these vectors is increased stability due to their linear conformation, but intriguingly, the mechanism remains unclear, although DNA supercoiling of plasmids is thought to play a role (Ronald Godiska, personal communication).

Our findings along with the aforementioned facts suggest that multiple, distinct mechanisms may be at play to cause cloning bias in *E. coli*, but that there is evidence that transcriptional activity of cloned DNA may be contributing to the sequence bias observed in metagenomic libraries. It is often assumed that toxicity of gene products may influence the stable maintenance or "clonability" of DNA in *E. coli* [84, 287, 302], but it is currently unclear whether gene product toxicity is a major factor in the bias of typical clone libraries constructed using circular vectors. It is interesting to consider that cloning bias could be due primarily to purely transcriptional activity rather than the often-blamed protein toxicity.

4.4.5 Cloning bias in a soil metagenomic library

The previous sections discuss the results of using shotgun sequencing to examine bias in a human fecal library (CLGM1 library; NCBI BioSample SAMN02324081). This section also discusses the results of 16S rRNA gene sequencing to examine bias in a corn field soil library (12AC library; NCBI BioSample SAMN02324088) [43]. Both

libraries were constructed using the same vector, the RK2-based cosmid pJC8 (Genbank accession KC149513). To examine possible bias in the soil library, I compared the 16S rRNA gene sequences from the original DNA that was extracted from the sample to the 16S rRNA gene sequences from the final cloned library DNA isolated from *E. coli*. Figure 4.12A summarizes analysis at the phylum level for both the fecal and soil samples.

At the phylum level, the fecal library differs substantially in the relative abundance of phyla compared to its corresponding extract, as discussed in the previous section. On the contrary, the relative abundance of phyla in the corn field soil library seemed similar to its extract (Figure 4.12A), although some caution should be exercised in their interpretation. Unfortunately, the majority of 16S rRNA gene sequences from the library sample were *E. coli* contamination, despite treating the library cosmid DNA preparation with Plasmid-Safe DNase to remove host genomic DNA prior to PCR, as well as obtaining on the order of millions of sequences from Illumina sequencing; after subtracting *E. coli* host sequences, I was left with approximately 30,000 sequences to represent the metagenomic library (see Section 4.6.7 for details). This high level of host contamination could be due to preferential amplification of template during PCR based on differences in DNA conformation: though present in very small quantities, linear DNA may be more efficiently amplified over supercoiled or closed circular plasmid DNA [39]. The issue of *E. coli* host contamination in 16S rRNA gene analysis needs to be addressed for future examination of bias in metagenomic libraries.

When I examined the soil samples more closely, I found that the similarity of the library and extract at the phylum level does not extend to the "species" level: examination of the individual OTUs in each sample revealed that only a small fraction of OTUs are shared between the library and original sample (Figure 4.12B). Interestingly, this analysis indicated that there were a number of OTUs in the library that were not iden-

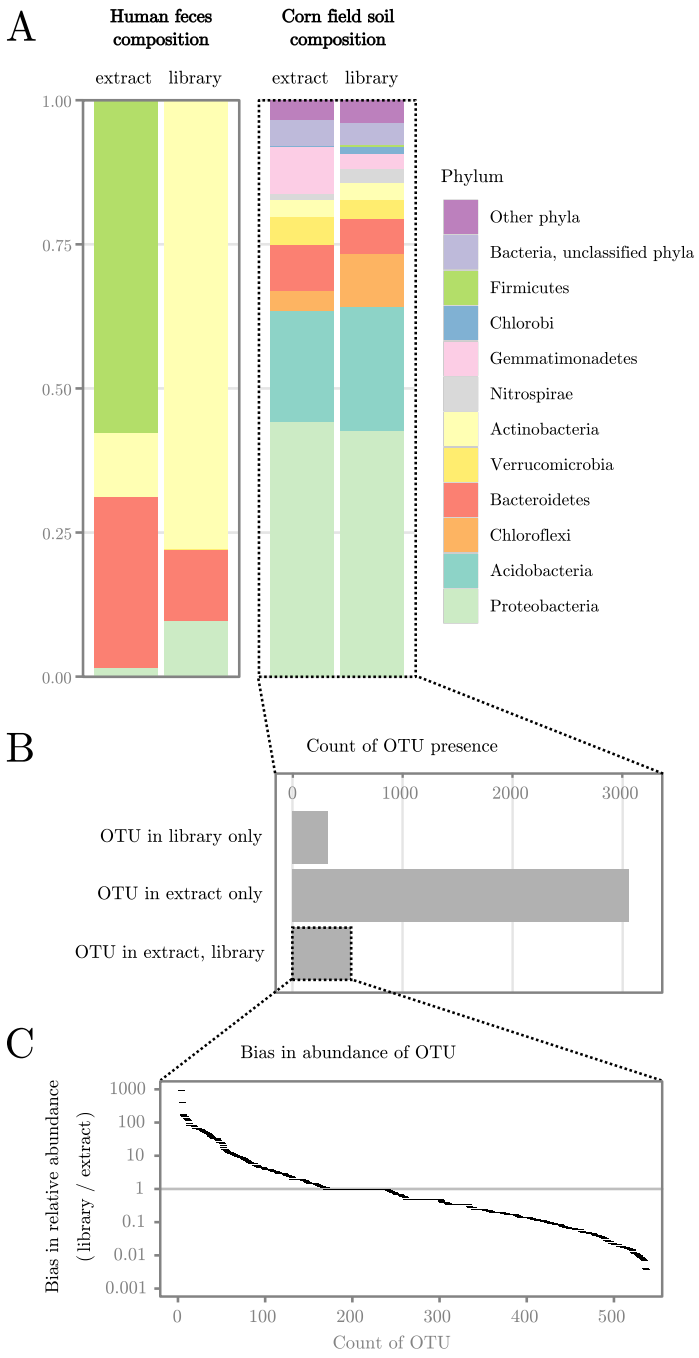


Figure 4.12: Metagenomic libraries exhibit cloning bias when compared to the original environmental sample. (A) Relative abundance of bacterial phyla from two previously constructed metagenomic libraries, a human fecal library [165] and a corn field soil library [43], compared to their original sample DNA extracts. (B) Number of OTUs identified from corn field soil DNA extract and library, and whether the OTUs were present in the library sample only, the extract sample only, or present in both. (C) Examination of cloning bias by comparing the relative abundance of OTUs that were present in both the DNA extract and the cosmid library, shown on a log scale; horizontal line at 1 denotes equal relative abundance in both samples.

tified in the extract sample (Figure 4.12B) and although this number is halved when the library data are compared to extract data that have not been rarefied (data not shown), they nevertheless remain, indicating that these OTUs are either extremely rare in the original sample and their DNA is preferentially cloned or that the identification of these OTUs is due to sequencing errors. A further analysis of the OTU fraction that is shared between extract and library samples shows a large range in the bias in relative abundance of each OTU, with some OTUs exhibiting a 1000-fold overrepresentation and others a 1000-fold underrepresentation in the library (Figure 4.12C). While there may be concern that 16S rRNA gene profiles of libraries compared to extracts may not provide an accurate comparison of cloned DNA content in general, I have shown in a previous section that for large-insert RK2 *oriV*-based cosmid libraries, 16S rRNA gene tracks well with genomic content (Figure 4.4.2). The analysis of the corn field DNA extract and corresponding metagenomic library suggests that though the overall relative abundance of phyla may remain similar, bias is occurring on the level of individual OTUs. This indicates that when trying to understand bias, using the popular representation of samples as barplots of bacterial phyla may be inappropriate; rather an OTU-level analysis may be required (Figure 4.12B versus Figure 4.12C). For mining purposes, the fact that certain taxa are under- or overrepresented might not pose a barrier to screening, but it may be useful to know from the beginning what sequences are not likely to be captured in libraries.

4.5 Conclusions

The results presented in this chapter and what was already known from the literature together support the hypothesis that GC bias in typical clone libraries (that is, using circular vectors) is related to constitutive promoter activity of the insert in *E. coli*,

although DNA topology as well as toxic protein effects may also influence insert and plasmid maintenance. In this analyses, I have focused only on would-be strong constitutive promoters in *E. coli* (*rpoD*/σ^{70} consensus sequences) because there is evidence that high-level transcription may have negative effects. It is important to acknowledge, however, that functional metagenomic approaches rely on *E. coli* (or other hosts) being able to transcribe and translate foreign DNA, in order to identify fragments encoding functions of interest. This ability of *E. coli* to initiate low-level transcription from diverse sources [214] and to be able to produce foreign proteins has been immensely advantageous for functional metagenomics and likely has contributed to the general assumption that *E. coli* is tolerant of foreign DNA, whether it expresses it or not. Our work, however, suggests that more careful consideration of cloning strategies may be required.

The stability of foreign DNA in *E. coli* is influenced by the copy number of its host plasmid and, as a result, single-copy fosmids may be ideal as the library backbone [148], although the success of some functional screens may be dependent on a higher gene dose. Possible alternatives to fosmid vectors include BACs [142] as well as linear vectors, which may provide exceptional stability [106]. However, *cos*-based vectors are likely to remain popular for their advantages: the availability of high-quality commercial packaging extracts, the efficiency of transduction over transformation, and the decreased probability of insert concatemers due to the phage head upper size limit. Though there exists variety in library cloning vectors, further work is required to understand how and to what extent cloning vector choice impacts library sequence bias.

Currently, there are three outstanding questions: (1) to what extent does transcription contribute to metagenomic library bias, (2) what factors affect whether transcription will be problematic, and (3) how can transcriptional effects be minimized so that DNA can be faithfully maintained in *E. coli*. An important consideration may

be the likelihood of an *rpoD* consensus sequence being cloned on any given fragment from a genome or metagenome. As an example, let us consider *Ruminococcus bromii*, which was one of the most highly abundant species in the original sample but became nearly absent in the cosmid library according to our analyses (∼7% versus ∼0.01%, respectively; see Section C.1). *R. bromii* has a genome size of 2.25 Mb; theoretically, its genome can be represented in ∼80 fragments if we consider that the average fragment in the particular cosmid library discussed here is ∼28 kb (data not shown). Given that there were 77 *rpoD* consensus sequences identified in its genome (Table 4.3), potentially many fragments could include a sequence that behaves as a strong, constitutive promoter in *E. coli*. I acknowledge that although this work supports the hypothesis that constitutive transcription contributes to library bias, more concrete evidence is required to confirm this hypothesis.

If strong transcription from the insert into the vector backbone contributes in part to the observed cloning bias—affecting the origin of replication, for example—it may be helpful to use vectors that include transcriptional terminators flanking the cloning site. Our lab is currently investigating the extent to which transcriptional terminators alleviate the cosmid library sequence bias, which may help tease apart the issue of transcription from that of gene product toxicity. While it is generally recognized that different expression hosts are needed for functional screening (discussed in Section 1.6.3), it is not as widely acknowledged that using *E. coli* as the sole cloning host for metagenomic DNA itself may be quite limiting due to the potential lack of sample representativeness from the outset. It is interesting that despite decades of using *E. coli* as "the workhorse of molecular biology," there is still much left to discover about how it tolerates exogenous DNA, which should serve as a reminder to us of how necessary it is to continually re-evaluate even our most basic methodological assumptions, particularly when they concern the inner workings of the cell.

4.6 Specific materials and methods

4.6.1 Sampling of DNA during fecal library construction

Methods for the construction of cosmid libraries, including the specific human gut metagenomic library discussed here (NCBI BioSample ID SAMN02324081), have been previously described in detail [167]. Briefly, DNA was extracted from pooled human fecal samples using freeze-grinding with liquid nitrogen followed by gentle lysis. Crude-extracted DNA was then size-selected by pulsed field gel electrophoresis using a CHEF Mapper Pulsed Field Gel Electrophoresis System (Bio-Rad), followed by electroelution, retaining fragments between approximately 40 and 70 kb. The size-selected DNA was end-repaired, purified, and ligated into the Eco72I site of linearized dephosphorylated pJC8 vector DNA (Genbank accession KC149513). The ligation product was packaged into λ phage heads using Gigapack III XL Packaging Extract (Stratagene 200209), followed by transduction of *E. coli* HB101. Transductants were recovered on LB agar supplemented with tetracycline (20 µg/ml) and incubated overnight at 37℃. Resulting colonies were enumerated to estimate library size (∼42,000 clones), and colonies were resuspended, pooled, and frozen at -80℃ to form the cosmid library stock.

During construction of the cosmid library, DNA was sampled from three steps: (1) the crude extract DNA, (2) the size-selected DNA, and (3) the final cosmid library DNA, prepared from the frozen stock using a GeneJET Plasmid Miniprep Kit (Thermo Scientific K0502).

4.6.2 Purification, quantification, and Illumina sequencing of DNA

Two of the three DNA samples, the cosmid library DNA and the size-selected DNA, were sufficiently pure for Illumina sequencing, as gauged by 260/280 and 260/230-nm ratios (Nanodrop ND-1000 Spectrophotometer); however, the crude extract DNA required further purification. Crude extract DNA concentration was estimated by gel electrophoresis, using bacteriophage λ DNA as a standard; ∼150 µg in 1 ml was purified and concentrated on the synchronous coefficient of drag alteration (SCODA) instrument (Boreal Genomics), using an established protocol [75].

All samples were re-quantified by gel electrophoresis, using bacteriophage λ DNA as a standard, and >2 µg of each sample was sent to the Beijing Genomics Institute (BGI, Hong Kong) for 90-base paired-end sequencing on the Illumina HiSeq 2000 platform, using their in-house protocols and reagents for 350-bp fragment library construction. Approximately 6.7 million reads were obtained in both the forward and the reverse direction, generating ∼1.2 Gb of sequence data per sample. All sequence data have been made publicly available (see "Data" section).

4.6.3 Subtraction of *E. coli* and vector DNA from fecal sequence data

The fecal cosmid library sequence data were expected to have substantial contamination with *E. coli* genomic DNA and pJC8 vector sequences. Sequence data were cleaned of contaminating *E. coli* genomic DNA and vector DNA, using BLAT [146] with a conservative criterion of 100% identity. To remove *E. coli* contamination, I used the genome of *E. coli* K12 MG1655 (Genbank accession U00096.3), which to our knowledge

is currently the closest sequenced relative of HB101, the library host strain. To remove vector contamination, I used the sequence of pJC8 (Genbank accession KC149513), formatted to simulate Eco72I-cut, cloning-ready vector by removing the 0.8-kb gentamicin resistance gene stuffer present between the two Eco72I sites.

4.6.4 Taxonomic analysis

To examine taxonomy based on only the 16S rRNA gene sequences present in the data, I identified 16S-containing reads using Infernal version 1.1 [220] and classified them using the RDP Classifier version 2.8 [323]. The classifier output was visualized using the MEtaGenome ANalyzer (MEGAN) version 5.6 [132]. To examine taxonomy using all sequence reads (i.e., not only those identified as 16S reads), I used the MetaPhlAn tool version 2.0, along with its built-in scripts for visualization [258].

4.6.5 Promoter analysis

To estimate promoter content in the data, I searched for known sigma factor consensus sequences for the *E. coli* sigma factors, $rpoD/\sigma^{70}$, $rpoE/\sigma^{24}$, $rpoH/\sigma^{32}$, $rpoN/\sigma^{54}$, as well as for the *Bacteroides* primary sigma factor, σ^{ABfr}. To do this, I used regular expression pattern matching with Python version 2.7.3; consensus promoter sequences and literature references were provided in Table 4.2 and regular expressions are provided in Table 4.4.

Table 4.4: Regular expressions used for selected promoter consensus sequences.

Sigma factor	Regular expression
rpoD (σ^{70})	`TTGACA.{15,19}TATAAT`
rpoE (σ^{24})	`GGAACTT.{15,19}TCAAA`
rpoH (σ^{32})	`TTG[AT][AT][AT].{13,14}CCCCAT[AT]T`
rpoN (σ^{54})	`TGGCA.{7}TGC`
Bacteroides (σ^{ABfr})	`TTTG.{19,21}TA.{2}TTTG`

4.6.6 Analysis of reference genomes

Genome sequences were downloaded from the NCBI Genbank database as complete genomes, draft genomes, or from whole genome shotgun sequencing projects. Organism names and accession numbers, as well as other relevant information, are provided (Table 4.5).

Table 4.5: NCBI accession numbers for genome sequences of the 46 species selected for percent GC and *rpoD* consensus content analysis. [165]

Species Name	Genome Status	NCBI Accession(s) downloaded 2014-10-17	NCBI Definition (abbreviated)	fasta seqs
Akkermansia muciniphila	comeplete	NC˙010655.1	Akkermansia muciniphila ATCC BAA-835 chromosome, complete genome	1
Alistipes putredinis	wgs	NZ˙DS499570:NZ˙DS499581[PACC]	Alistipes putredinis DSM 17216 Scfld	12
Alistipes shahii	draft	NC˙021030.1	Alistipes shahii WAL 8301 draft genome	1
Bacteroides˙cellulosilyticus	wgs	NZ˙EQ973486:NZ˙EQ973551[PACC]	Bacteroides cellulosilyticus DSM 14838 genomic scaffold	66
Bacteroides˙ovatus	wgs	NZ˙ADMO01000001:NZ˙ADMO01000156[PACC]	Bacteroides ovatus SD CMC 3f contig	156
Bacteroides˙thetaiotaomicron	comeplete	AE015928.1	Bacteroides thetaiotaomicron VPI-5482, complete genome	1
Bacteroides˙unclassified	n/a	n/a	n/a	n/a
Bacteroides˙uniformis	wgs	NZ˙DS362217:NZ˙DS362249[PACC]	Bacteroides uniformis ATCC 8492 Scfld	33
Bacteroides˙vulgatus	comeplete	NC˙009614.1	Bacteroides vulgatus ATCC 8482 chromosome, complete genome	1
Bacteroides˙xylanisolvens	draft	NC˙021017.1	Bacteroides vulgatus ATCC 8482 chromosome, complete genome	1
Bifidobacterium adolescentis	comeplete	NC˙008618.1	Bifidobacterium adolescentis ATCC 15703 chromosome, complete genome	1
Bifidobacterium breve	comeplete	NC˙017218.1	Bifidobacterium breve ACS-071-V-Sch8b chromosome, complete genome	1
Bifidobacterium catenulatum	wgs	NZ˙ABXY01000001:NZ˙ABXY01000031[PACC]	Bifidobacterium catenulatum DSM 16992 B˙catenulatum-1.0˙Cont	31
Bifidobacterium longum	comeplete	NC˙004307.2	Bifidobacterium longum NCC2705 chromosome, complete genome	1
Bifidobacterium pseudocatenulatum	wgs	NZ˙ABXX02000001:NZ˙ABXX02000036[PACC]	Bifidobacterium pseudocatenulatum DSM 20438 B˙pseudocatenulatum-1.0.1˙Cont	36
Bilophila wadsworthia	wgs	NZ˙KE150238:NZ˙KE150241[PACC]	Bilophila wadsworthia 3˙1˙6 genomic scaffold	4
Catenibacterium mitsuokai	wgs	NZ˙ACCK01000001:NZ˙ACCK01000475[PACC]	Catenibacterium mitsuokai DSM 15897 C˙mitsuokai-1.0˙Cont	475
Clostridium leptum	wgs	NZ˙DS480331:NZ˙DS480351[PACC]	Clostridium leptum DSM 753 Scfld	21
Clostridium nexile	wgs	NZ˙DS995337:NZ˙DS995353[PACC], NZ˙DS995602:NZ˙DS995683[PACC]	Clostridium nexile DSM 1787 Scfld	99
Clostridium˙bolteae	wgs	NZ˙DS480659:NZ˙DS480726[PACC]	Clostridium bolteae ATCC BAA-613 Scfld	68
Collinsella aerofaciens	wgs	NZ˙AAVN02000001:NZ˙AAVN02000025[PACC]	Collinsella aerofaciens ATCC 25986 C˙aerofaciens-2.0˙Cont	25
Coprococcus catus	draft	NC˙021009.1	Coprococcus catus GD/7 draft genome	1
Coprococcus comes	wgs	NZ˙GG662005:NZ˙GG662017[PACC]	Coprococcus comes ATCC 27758 genomic scaffold	13
Desulfovibrio piger	wgs	NZ˙DS996351:NZ˙DS996397[PACC]	Desulfovibrio piger ATCC 29098 Scfld	47
Dialister invisus	wgs	NZ˙GG698602.1	Dialister invisus DSM 15470 genomic scaffold Scfld0, whole genome shotgun sequence	1
Dorea formicigenerans	wgs	NZ˙AAXA02000001:NZ˙AAXA02000016[PACC]	Dorea formicigenerans ATCC 27755 D˙formicigenerans-3.0.1˙Cont	16
Dorea longicatena	wgs	NZ˙DS264384:NZ˙DS264419[PACC]	Dorea longicatena DSM 13814 Scfld	36
Eggerthella lenta	comeplete	NC˙013204.1	Eggerthella lenta DSM 2243 chromosome, complete genome	1
Escherichia coli	comeplete	NC˙000913.3	Escherichia coli str. K-12 substr. MG1655, complete genome	1
Eubacterium eligens	comeplete	NC˙012778.1	Eubacterium eligens ATCC 27750 chromosome, complete genome	1
Eubacterium hallii	wgs	NZ˙ACEP01000001:NZ˙ACEP01000175[PACC]	Eubacterium hallii DSM 3353 E˙hallii-1.0˙Cont	175
Eubacterium rectale	comeplete	NC˙012781.1	Eubacterium rectale ATCC 33656, complete genome	1
Eubacterium˙ventriosum	wgs	NZ˙DS264262:NZ˙DS264288[PACC]	Eubacterium ventriosum ATCC 27560 Scfld	27
Faecalibacterium˙cf	n/a	n/a	n/a	n/a
Faecalibacterium˙prausnitzii	comeplete	NC˙021042.1	Faecalibacterium prausnitzii L2-6, complete genome	1
Faecalibacterium˙unclassified	n/a	n/a	n/a	n/a
Gordonibacter pamelaeae	draft	NC˙021021.1	Gordonibacter pamelaeae 7-10-1-b draft genome	1
Holdemania filiformis	wgs	NZ˙GG657551:NZ˙GG657585[PACC]	Holdemania filiformis DSM 12042 genomic scaffold	35
Klebsiella pneumoniae	comeplete	NC˙016845.1	Klebsiella pneumoniae subsp. pneumoniae HS11286 chromosome, complete genome	1
Lactobacillus ruminis	comeplete	NC˙015975.1	Lactobacillus ruminis ATCC 27782 chromosome, complete genome	1
Megamonas hypermegale	draft	NC˙021041.1	Megamonas hypermegale ART12/1 draft genome	1
Parabacteroides unclassified	n/a	n/a	n/a	n/a
Parabacteroides˙merdae	wgs	NZ˙DS264460:NZ˙DS264552[PACC]	Parabacteroides merdae ATCC 43184 Scfld	93
Prevotella˙copri	wgs	NZ˙GG703852:NZ˙GG703878[PACC]	Prevotella copri DSM 18205 genomic scaffold	27
Roseburia inulinivorans	wgs	NZ˙ACFY01000001:NZ˙ACFY01000179[PACC]	Roseburia inulinivorans DSM 16841 R˙inulinivorans-1.0.1˙Cont	179
Ruminococcus bromii	draft	NC˙021013.1	Ruminococcus bromii L2-63 draft genome	1
Ruminococcus [Blautia] gnavus	wgs	NZ˙AAYG02000001:NZ˙AAYG02000043[PACC]	Ruminococcus gnavus ATCC 29149 R˙gnavus-1.0.1˙Cont	43
Ruminococcus [Blautia] obeum	draft	NC˙021022.1	Ruminococcus obeum A2-162 draft genome	1
Ruminococcus [Blautia] torques	draft	NC˙021015.1	Ruminococcus torques L2-14 draft genome	1
Streptococcus parasanguinis	comeplete	NC˙015678.1	Streptococcus parasanguinis ATCC 15912 chromosome, complete genome	1

4.6.7 16S rRNA analysis for soil extract and library

Construction of the 12AC library was previously described [43]. Crude DNA extract of corn field soil was purified using the SCODA method [75]. Cosmid library DNA was miniprepped from *E. coli* HB101 using a GeneJet Plasmid Miniprep kit (Thermo Scientific K0502). Cosmid DNA was treated with Plasmid-Safe ATP-dependent DNase according to the supplier's recommendations (Epicentre Biotechnologies E3101K). PCR was carried out on the samples as previously described, using bacterial V3-specific primers 5'`CCTACGGGAGGCAGCAG` and 5'`ATTACCGCGGCTGCTGG` [14]. Amplicons were sequenced at the NRC-PBI Saskatoon Research Facility (Saskatoon, Canada) using the Illumina GAIIx platform. Paired-end sequences were assembled using PANDAseq version 2.8 [205] using default parameters; 1,823,112 and 1,886,370 sequences were assembled for the extract and cosmid library sample, from an input of 1,960,793 and 2,035,138 paired-end sequences, respectively. *E. coli* sequences were filtered out, using a criterion of 100% identity to *E. coli* MG1655 (the closest sequenced relative of HB101), resulting in 233 sequences removed from the extract sample and 1,453,806 sequences removed from the cosmid library sample. Sequences were subsequently processed via AXIOME2 [195] running QIIME version 1.9, specifying UPARSE (USEARCH version 7.0) to cluster the sequences using default parameters and the RDP classifier version 2.2 trained with the Greengenes database version 13.8 to classify defined OTUs. From the resulting OTU table, *E. coli* was filtered a second time by manually removing OTUs classified as *Enterobacteriaceae*, which consisted of 109 sequences from the extract sample and 335,994 sequences from cosmid library sample. The extract sample was then rarefied using QIIME to match the cosmid library, retaining ∼30,000 sequences for each sample, altogether comprising ∼4000 OTUs.

4.6.8 Data availability

Raw Illumina sequence data for the CLGM1 human gut cosmid library (NCBI BioSample SAMN02324081), size-selected, and crude extract DNA samples are available at the NCBI Sequence Read Archive under Study SRP031898. Accession numbers for SRA Experiments are: SRX683591 for the crude extract, SRX683589 for the size-selected, and SRX683586 for the cosmid library. Sequence data for the 12AC corn field DNA extract and corresponding metagenomic library (NCBI BioSample SAMN02324088) previously constructed [43] have been deposited at NCBI SRA; accession numbers are SRX1015944 and SRX1015946 for the extract and cosmid library, respectively. In addition, raw data and other relevant data for this study may be accessed online: http://www.cm2bl.org/~data

Chapter 5

Development of *Bacteroides thetaiotaomicron* as a screening host

5.1 Acknowledgements and declarations

I performed all experiments and analyses described in this chapter.

I acknowledge the following contributions:

- This chapter uses methods for the culture and genetic manipulation of *B. theta* adapted from protocols that were generously shared by **Nicole Koropatkin** and **Eric Martens** from the University of Michigan.
- The plasmid pAFD1 was generously shared by **Nadja Shoemaker** from the laboratory of **Abigail Salyers**.
- The primers used in the construction of the *B. theta thrC* and *trpD* single recombinant mutants described in Section 5.4.3 were designed by **Eric Martens**.
- The previously published *B. theta chuR* deletion mutant [17] used in functional complementation screens in Section 5.4.3 was shared with me by **Elizabeth Cameron**, then a Ph.D. student in the laboratory of **Eric Martens**.
- Parts of the introduction for this chapter were written as part of a review for my graduate specialized studies course, BIOL 681.
- This chapter was proofread by my supervisor **Trevor Charles**.

5.2 Abstract

Functional metagenomic approaches are becoming increasingly important in this age of relatively inexpensive high-throughput sequencing, in which obtaining sequence data from metagenomes is widely accessible but lack of knowledge of gene function makes annotation of those datasets incomplete. Function-based approaches can help to fill this gap in knowledge by providing information about gene function for as-yet uncharacterized sequences through the cloning, expression, and functional screening/selection of DNA from metagenomes. Importantly, this process is dependent on the ability to express the cloned DNA in a surrogate host; though *E. coli* is a popular host for screening of metagenomic libraries, it may not be ideal.

Regarding human gut metagenomic DNA in particular, the Gammaproteobacteria *E. coli* may be inadequate due to barriers in transcription and/or translation. The bacterial community that inhabits the human distal gut is composed predominantly of members of the Bacteroidetes and Firmicutes phyla; though there are Proteobacteria present, they are usually vastly outnumbered. For one dominant member of the Bacteroidetes phylum, *Bacteroides thetaiotaomicron* (*B. theta*), it has been shown that the *E. coli* σ^{70} sigma factor is unable to substitute the function of the *Bacteroides* sigma factor in vivo and is therefore unable to transcribe *Bacteroides* DNA, although spurious transcription is possible.

With growing interest in the human gut microbiome, *B. theta* is attracting the attention of researchers interested in understanding its dominance and stability in the gut environment as well as those interested in harnessing these properties for microbiome engineering. In this chapter, I discuss how *B. theta* might be useful for functional metagenomics as a screening host, to express DNA present in gut-derived metagenomic libraries. *B. theta* is a good candidate because it already has reasonably well-developed molecular genetic methods, including methods for conjugation and mutant construction. In addition, it has inherent advantages such as aerotolerance and the ability to degrade various complex polysaccharides, which make it relatively easy to manipulate in a typical laboratory setting and provides potential phenotypes for functional complementation, respectively. Here, I present the results of developing *B. theta* VPI-5482 as a surrogate host for functional screening, through the construction of *B. theta*-compatible *cos*-based cloning vectors, generation of human gut metagenomic libraries, and attempt to complement *B. theta* mutants. In my first unsuccessful attempt, I constructed a high-

copy cosmid vector called pKL3, based on an existing *E. coli-B. theta* shuttle vector, but found after generating libraries that metagenomic DNA inserts maintained at high copy number were unstable and led to difficulty conjugating into *B. theta*. In my second attempt, I constructed a fosmid vector called pKL13, based on the commerical pCC1FOS, and found that metagenomic libraries were both more stable and exhibited sufficient conjugation efficiency for attempting functional screens in *B. theta*.

For *B. theta* mutant strains to use in complementation screens, I constructed amino acid auxotrophs using single recombination of a suicide vector to disrupt genes in either the threonine or tryptophan biosynthesis operons. Unfortunately, complementation of single recombinants proved unsuccessful as the recombinants had a tendency to revert to wild-type. Instead, I tried to complement an existing *B. theta* deletion mutant, a mutant missing the *chuR* gene that is required for growth on chondroitin sulfate as sole carbon source. This screen was successful, leading to the isolation and analysis of several complementing clones from the human gut metagenomic library, including one *chuR* gene exhibiting 97% nucleotide sequence identity to the wild-type VPI-5482 sequence. Unfortunately, however, this analysis also led to the discovery that fosmid clone DNA appeared to have recombined into the *B. theta chuR* mutant host genome.

The inability to retrieve fosmid clone DNA poses a barrier to screening of pooled metagenomic libraries; to tackle this problem, it was necessary to track individual clones being conjugated into *B. theta*. In a proof-of-principle experiment, I generated an arrayed collection using a subset of clones from the pKL13-based metagenomic library, and performed a two-step screening strategy to identify which clones in the array led to complementation of the *chuR* phenotype. Results from this attempt show that the method is promising, although mating conditions need to be refined to achieve the high throughput required for screening hundreds of thousands of clones in this manner. Based on the results presented here, *B. theta* has potential for use as a host in functional screening of gut-derived metagenomic DNA.

5.3 Introduction

Bacteroides thetaiotaomicron, or *B. theta*, is a microbe that is frequently a dominant member of the human gut, specifically the distal intestine [12, 339]. It is a Gram-negative anaerobe whose genome sequence was made available in 2003 [339]. The sequenced representative is the type strain from the Virginia Polytechnic Institute, VPI-5482; an alternative name for the same strain from the American Type Culture Collection is ATCC 29148 [339]. The type strain has one 6.3-Mb chromosome and one 33-kb plasmid called p5482.

As research interest concerning the role of the human-associated microbiota in human health has grown, and particularly of the human gut microbiota, so too has the interest in *B. theta* grown. Its dominance in the gut, its ability to break down complex polysaccharides from both the host as well as the host dietary intake, and its tractability in bacterial genetics has brought it to the forefront of human microbiota studies. This introduction will discuss *B. theta*'s role and functions in its symbiosis with the host, give an overview of molecular genetic methods used to work with *B. theta* in the laboratory, and finally, touch on the reasons that *B. theta* would be a suitable expression host for functional metagenomics.

5.3.1 Mutualistic role and polysaccharide utilization abilities

The digestion of complex polysaccharides in the gut requires the action of glycoside hydrolases (GHs) and polysaccharide lyases (PLs), enzymes which are able to hydrolyze glycosidic bonds and cleave carbohydrates using an elimination mechanism, respectively [56]. Interestingly, compared to the microbes that reside in our gastrointestinal tract, humans have no PLs and only a relatively small number of GHs, with only a handful of these participating in digestion, specifically of starch, sucrose, and lactose (Figure 5.1).

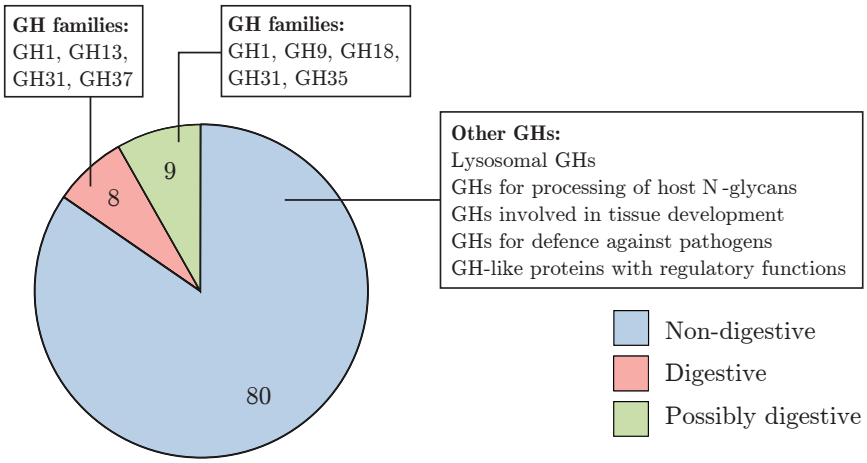


Figure 5.1: Classification of glycoside hydrolases encoded by the human genome. Adapted from [72].

In general, the gut microbiota allow energy to be harvested from many complex polysaccharides in the common human diet that would otherwise be undigestable, such as pectin, cellulose, and hemicellulose [12]. Our resident microbiota produce short chain fatty acids from fermentation of these polysaccharides, which are then taken up by our colonocytes [48], particularly butyrate [67]; in this manner, our microbiota have been estimated to produce between 5-10% of our energy requirements [207]. An assessment of a fraction of these bacteria whose genomes are sequenced reveals that many species possess GHs and PLs; in particular though, members of the Bacteroidetes have both a large number as well as diverse members of GHs and PLs, with *B. theta* close to the top (Figure 5.2).

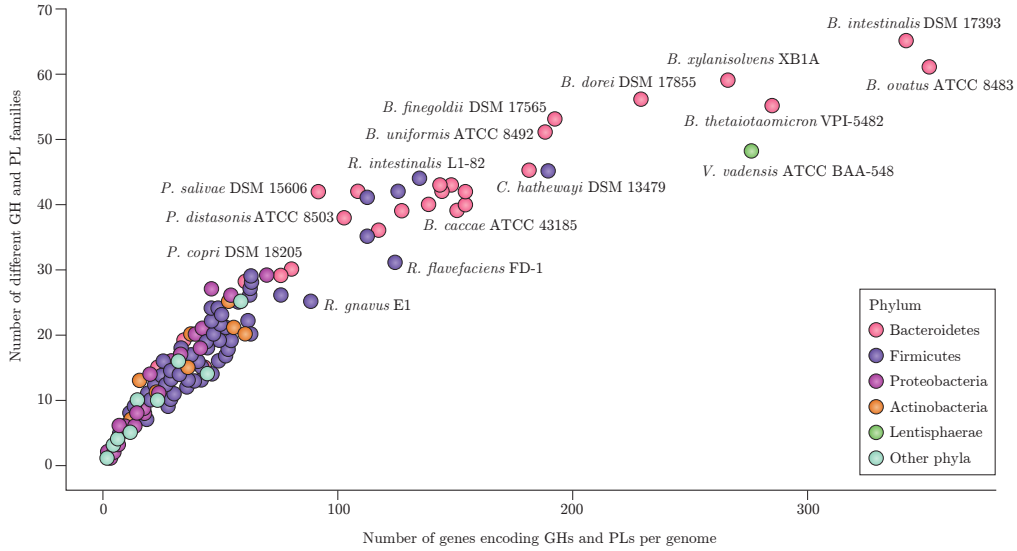


Figure 5.2: Total number and number of different GH and PL genes in gut bacterial genomes. Adapted from [72].

That *B. theta* would possess both a large number and diverse members of these enzymes is perhaps not surprising, as it has been characterized as a "generalist" with the ability to degrade a broad range of polysaccharides in the gut, in contrast to "specialists" that can only degrade one or a few polysaccharides [158]. Its relatively large genome size of ∼6.3 Mb has been attributed to this generalist lifestyle in the "use-it-or-lose-it" hypothesis of gene retention [221].

The Starch Utilization System (SUS) in *B. theta* is a canonical example of an operon devoted to the degradation of a particular polysaccharide (Figure 5.3). The system was first studied in the 1980s in the laboratory of the late Abigail Salyers. Using transposon mutagenesis, it was found that starch utilization mutants had insertions clustered within an 18-kb region of the chromosome [297]. Biochemical and genetic analyses of *B. theta* revealed that cells did not secrete extracellular enzymes, but instead bound starch for eventual degradation in the periplasm or cytoplasm [5, 6]. Later

work in the Salyers lab identified all 8 members of the *sus* locus, *susRABCDEFG* [57, 58, 238, 239]. Briefly, outer membrane proteins SusE, SusF, and SusD bind the starch molecule allowing it to be degraded into smaller oligosaccharides by the amylase SusG; SusE and SusF were shown to be not required for growth on starch [45] though they are involved in enhancing starch binding [32,263]. SusG-generated oligosaccharides are transported via the transporter SusC to the periplasm where SusA and SusB cleave them to form smaller mono- and disaccharides, which are finally transported into the cytoplasm for use by the cell. SusR is involved in activation of the locus and its expression is induced by maltose.

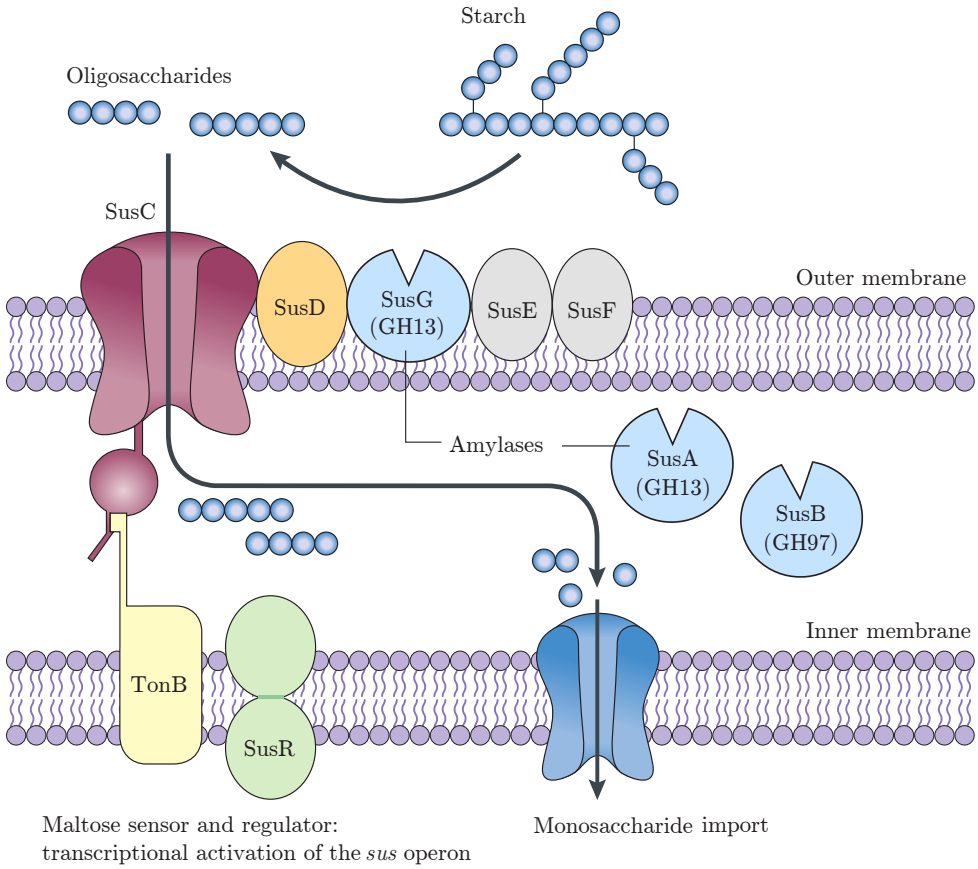


Figure 5.3: Overview of the canonical Starch Utilization System (SUS). Adapted from [158]

The *sus* locus is one example of an operon that encodes an entire membrane-associated multi-protein system for tackling the degradation of a specific polysaccharide, namely starch. *B. theta* uses similar operons to degrade other carbohydrates, called SUS-like systems or polysaccharide utilization loci (PULs) [202]. Remarkably, *B. theta* is estimated to have a total of 88 of these PULs, which comprise 18% of its genome and 866 of its genes, enabling it to degrade a wide range of glycans, from host-derived glycans such as mucin O-glycans and chondroitin sulfate to plant-derived glycans such as amylopectin and inulin [200, 202], although the majority of its 88 PULs are thought to be involved in the breakdown of plant polysaccharides [331].

These systems are interesting in that they afford members of the Bacteroidetes a competitive advantage, but each species may have its own micro-habitat or niche depending on the array of PULs its genome possesses. For example, *B. theta* seems well-suited for growth on host mucins while a related species, *Bacteroides ovatus*, may thrive on plant cell wall hemicellulose [158]. The SUS and SUS-like systems are of particular interest because *Bacteroides* mutants deficient in these systems may be good candidates for use as hosts in functional complementation screening.

5.3.2 Overview of molecular methods for *B. theta*

Over the past few decades, research interest in members of the *Bacteroides* has grown, leading to the development of molecular methods for use with these organisms, specifically with *B. theta*. Work in the Salyers' lab led to the development of *Bacteroides* as a genetic system, as well as the related *Porphyromonas* and *Prevotella*. Abigail Salyers was interested in the *Bacteroides* and related organisms for both their environmental and clinical significance. As mentioned, her lab did the initial studies on the *sus* locus in *B. theta*, but her lab also studied antibiotic resistance in *B. theta*, mediated by

mobile elements, which include conjugative transposons [250] and mobilizable insertion elements [182]. Their work developing genetic methods in *B. theta* culminated in the publication of two important reviews of genetic techniques in *B. theta*, one in 1999 called *"Genetic Methods for Bacteroides Species"* [249] and another in 2000, called *"Starting a new genetic system: lessons from Bacteroides"* [248]. In the following sections, I will attempt to summarize the microbiology and molecular genetic methods used to work with *B. theta*, both those that stem from early work and those that have been developed since then.

Vectors

All of the *B. theta*-compatible vectors in use today appear to use origins of replication that can be traced back to just a few native plasmids originally isolated from *Bacteroides* species. From the literature, the two most common originate from the 4.4-kb plasmid pB8-51 isolated from *Bacteroides eggerthii* B8-51 [268] and the 2.7-kb plasmid pBI143 isolated from *Bacteroides fragilis* IB143 [274]. Both plasmids have a copy number of approximately 10 to 20 in *Bacteroides* and, interestingly, the two origins have been shown to be compatible [290], although pB8-51 appears to have a broader host range and can replicate in *Prevotella* and *Porphyromonas* species in addition to *Bacteroides* species. The plasmid pBI143 was sequenced in 1995 [278], about one decade after its isolation.

A range of *B. theta* vectors have been developed: shuttle vectors and suicide vectors, many of which have been previously reviewed by Salyers *et al.* [249], and even expression vectors are available for use in *Bacteroides* [277]. Nucleotide sequence data available for some *B. theta* plasmids (native plasmids or cloning vectors) are summarized in Table 5.1.

Table 5.1: Plasmids relevant for genetics in *B. theta*, with available sequence

Vector	Source	Ref.
pBI143	Genbank U30316 (1995)	[278]
pFD288	Genbank U30830 (1995)	[278]
pBA	Genbank AF203972 (2006)	[336]
pFD1146	Genbank JQ776640 (2012)	[228]
pBUN24	Genbank EU818711 (2013)	[264]
pVAL-1	Genbank AB775653 (2014)	[314]
pTIO-1	Genbank AB775804 (2014)	[296]
pKNOCK-*bla-ermG*	https://gordonlab.wustl.edu/plasmids/	[159]
pKNOCK-*bla-tetQ*	https://gordonlab.wustl.edu/plasmids/	[200]
pNBU2-*bla-ermG*	https://gordonlab.wustl.edu/plasmids/	[159]
pNBU2-*bla-tetQ*	https://gordonlab.wustl.edu/plasmids/	[200]

There have not been many cosmid vectors constructed for use in the *Bacteroides*, however, as searches of the literature have turned up only two cosmids, both constructed in the late 1980s:

- pNJR1/pNJR5 [265] was constructed in Abigail Salyers' lab and employs the *Bacteroides* pB8-51 origin and the *E. coli* RSF1010 origin (IncQ).
- pOA10 [112] was constructed at UCSD and uses the less popular *Bacteroides* pCP1/pBFTM10 origin and the *E. coli* pBR322 origin.

Selectable markers and reporters

There are two antibiotic selectable markers that appear to be favoured for use in *B. theta*, erythromycin and tetracycline. Other antibiotics have been used successfully in *B. theta* in the literature, however, and a summary of the possibilities is presented in Table 5.2.

Table 5.2: Antibiotic markers in *B. theta*.

Antibiotic	Concentration	Reference
erythromycin (*ermF, ermG*)	10-25 µg/ml	[32, 268]
tetracycline (*tetQ**)	3 µg/ml	[265]
clindamycin (*ermF, ermG*)	5-20 µg/ml	[268, 274]
ampicillin (*cfxA*)	25-50 µg/ml	[182]
chloramphenicol (*cat*)	10-15 µg/ml	[277, 290]

Additionally, reporter systems that have been used successfully in *B. theta* or closely related species include:

- β-glucoronidase (*uidA*) [249]
- β-xylosidase (*xyaA*) [249]
- chloramphenicol acetyl transferase (*cat*) [15]
- catechol 2,3-dioxygenase (*xylE*) [38]
- luciferase, including *lux* and [206] and Nanoluc [215]

*distinct from *E. coli* tetracycline resistance

Conjugation

The native plasmids isolated from *Bacteroides* species – pBI143, pB8-51, and pBFTM10 – all have *mob* regions and can be mobilized by R751 or RP4/RK2 [267, 276] though these helper plasmids cannot replicate in the recipient [250]. Interestingly, despite the fact that R751 does not recognize the RK2 *oriT*, most or all of the *B. theta* plasmids can be mobilized by both R751 and RK2 [248].

Conjugations from an *E. coli* donor into a *B. theta* recipient can be done anaerobically on nitrocellulose filters placed on TYG agar plates [268, 279] or aerobically as a lawn on brain-heart-infusion blood plates [159]; in the latter method, anaerobic incubation is not required likely because the initial growth of *E. coli* sets up a barrier to the oxygen, allowing the anaerobic *B. theta* to grow between the agar surface and the *E. coli* lawn. Conjugations from a *B. theta* donor into an *E. coli* recipient are possible but require that the the *Bacteroides* strain express transfer genes, such as those from a conjugative transposon, as it has been shown that R751 integrated into the genome of *B. theta* was not able to mobilize out on its own, likely because R751 transfer genes are not expressed in *Bacteroides* [250, 269].

Conjugations require counter-selection. For conjugations from *B. theta* into *E. coli*, selection via aerobic incubation of plates is obviously sufficient, although transconjugants must be streaked for purity because *B. theta* can co-culture with *E. coli* [267]. Conjugations from *E. coli* into *B. theta* on the other hand require the use of antibiotics against the *E. coli* donor because it is a facultative anaerobe that is able to grow in the absence of oxygen. The *B. theta* type strain VPI-5482 has been reported to be naturally resistant to all aminoglycosides [268], up to 1 mg/ml [266] as well as nalidixic acid [306]. The antibiotics that can be used and their concentrations are listed in Table 5.3.

Table 5.3: Counter-selection against *E. coli*.

Antibiotic	Concentration	Reference
gentamicin	200 µg/ml	[314]
geneticin (G418)	400 µg/ml	[267]
nalidixic acid	100 µg/ml	[268, 306]
cefoxitin	50 µg/ml	[133]
streptomycin	200 µg/ml	this study
kanamycin	200 µg/ml	this study

Transduction

There is currently no transducing phage for *Bacteroides* [248]. A transducing phage would provide a means to isolate the genetic background of mutant strains to ensure the absence of other mutations, or to combine two mutations into a single background. However, the search for a transducing phage can be difficult and time-consuming [248], which is probably why such a tool for the *Bacteroides* remains elusive.

Electroporation

Wild-type *Bacteroides* strains are typically recalcitrant to the introduction of heterologous DNA, possibly due to the presence of restriction-modification systems. However, it has been shown that *E. coli*-derived DNA can be electroporated into some *Bacteroides* species [275], with especially high efficiency into *B. fragilis* [133]. The same group has reported being able to successfully electroporate *E. coli* HB101-derived DNA into *B. theta* VPI-5482 [133]. Interestingly, the Salyers group was not able to achieve this, but they have published that *B. theta*-derived DNA can be re-introduced into

B. theta via electroporation at high frequencies [182], both observations that I can confirm (unpublished data).

Mutant construction

B. theta mutant construction is fairly straightforward as suicide vectors and conjugation strategies are available. Single recombinants can be made using a suicide vector, such as pKNOCK-*bla-ermG* (Figure 5.15A) [159], which carries the *ori* R6K origin of replication that requires the use of λ-*pir* strains. Conveniently, constructs can be mated from *E. coli* S17-1 λ-*pir* in biparental conjugations [201] that are more efficient that triparental conjugations using a mobilizer strain .

Double crossover-based methods allow for the construction of clean deletions (e.g., the removal of a specific open reading frame), and can be generated using the suicide vector pExchange-*tdk* [159], a derivative of pKNOCK-*bla-ermG* that carries the *B. theta tdk* gene. The *tdk* gene provides the counter-selection that is required to make a clean deletion and must be used in combination with a *B. theta tdk* deletion mutant. In the presence of Tdk, *B. theta* becomes sensitive to the nucleotide analog 5-fluoro-2-deoxyuridine (FUdR) [159]. Thus, mutant construction involves the following steps:

- cloning the ORF's upstream and downstream regions into the vector, generating the deletion construct
- conjugating the new construct into the *tdk* mutant, selecting with erythromycin for integration of the suicide plasmid into the genome at the location of the ORF
- selecting with FUdR for loss of the integrated plasmid, followed by screening FUdR-resistant, erythromycin-sensitive clones for loss of the ORF, using PCR

5.3.3 Use of *B. theta* in systems biology and synthetic biology

Since this project began, there have been studies in systems biology and synthetic biology making use of *B. theta*. In a recent study, a functional genomic approach was used to explore which *B. theta* genes contribute to fitness in the gut: small fragments of *B. theta* genomic DNA were cloned into an *E. coli* expression vector to drive expression of *B. theta* genes, forgoing the requirement for *E. coli* to recognize native *B. theta* elements for transcription and translation [341]. The researchers introduced this library into mice; then, by sampling mouse feces that was shed and sequencing the DNA present, they were able to identify which *B. theta* genes were carried by the clones that dominated the population in the mouse gut as time progressed. Perhaps unsurprisingly, the two genes that dominated by far (>90% by sequencing) were ones involved in carbohydrate utilization: BT_1759 encodes a periplasmic glycoside hydrolase involved in hydrolyzing fructo-oligosaccharides and sucrose [285] and the adjacent BT_1758 encodes a glucose/galactose transporter. This experiment illustrates the potential of using functional genomics to understand how specific genes might contribute to a microbe's fitness in the host gut. Although this experiment was done in *E. coli* and using only *B. theta* genomic DNA, the next step would be to use larger inserts for cloning, metagenomic DNA from the whole gut community, or even a different surrogate host [81].

In another study, *B. theta* was engineered to respond to environmental cues present in the mouse gut by expressing a luciferase reporter gene as well as recording this encounter through the modification of its own DNA [215]; this is often described as equipping the organism with "synthetic genetic memory". First, as a foundation for their work, the researchers developed a repertoire of genetic parts to use in *B. theta*, including promoters and RBSs that together allow gene expression to be controlled over a 10^4-fold range. They also develop inducible systems based on *E. coli*'s IPTG-inducible

lac system as well as on *B. theta*'s previously characterized natural polysaccharide utilization systems, which encode hybrid two-component transcriptional regulators that sense and respond to the presence of carbohydrates such as rhamnose, chondroitin sulfate, and arabinogalactan [200,203,231]. Next, they design the responsive genetic memory by coupling the rhamnose utilization regulator to expression of serine integrases for unidirectional inversion of DNA at a designed "memory array" located on the chromosome. Another important contribution by the authors to the *B. theta* genetics toolbox is the development an inducible system for knocking down gene expression in *B. theta* by using CRISPR interference (CRISPRi) and they demonstrate that CRISPRi can be used to down-regulate gene expression in *B. theta* cells colonizing the mouse gut. These exciting developments in synthetic biology will hopefully spur efforts in microbiome engineering that may be important for the development of therapeutics to treat gastrointestinal diseases [286].

These examples in the recent literature illustrate *B. theta*'s potential in both pure and applied research and its utility as a model for both studying the adaptive functions of the microbiota in the gut as well as for manipulating the microbiota for the benefit of the host.

5.3.4 Suitability as a host for screening human gut metagenomic DNA

Functional metagenomics is dependent on the ability to effectively screen libraries for gene function, therefore requiring that the cloned fragments be expressed in the surrogate host. The human gut microbial community is dominated by members of the Bacteroidetes phylum, suggesting that human gut-derived libraries contain a large portion of Bacteroidetes genes. However, previous studies suggest a barrier to the expression

of *Bacteroides*-derived genes in the popular Proteobacteria host *E. coli* at the level of transcription due to lack of promoter recognition [206]. *B. theta*'s primary sigma factor recognizes a consensus sequence markedly different from *E. coli*'s σ^{70} (Figure 4.9); the consensus has been identified and comprises two elements situated at -33 and -7 from the start of transcription, separated by 19-21 nucleotides: $\texttt{TTTGN}_{19\text{-}21}\texttt{TAN}_2\texttt{TTTG}$ [15,317].

Most interestingly, though this would appear to be a contradiction of the above facts, there are at least several examples in the literature where functional screens of metagenomic libraries in an *E. coli* surrogate host have turned up positive clones carrying DNA that appears to be from *Bacteroides*:

- A metagenomic fosmid libary constructed from the fecal samples of patients with Crohn's Disease was screened for ability to modify NF-κB expression in human intestinal epithelial cells using a reporter system. NF-κB is a transcription factor that is involved in immune and inflammatory responses in the gut. This led to the identification of a clone whose insert's closest match was *Bacteroides vulgatus* [164, 196].

- A metagenomic fosmid libary constructed from the fecal sample of a healthy pescatarian was screened for carbohydrate-active enzymes able to degrade resistant substrates and/or able to withstand high temperature or extreme pH. Of the 26 clones sequenced, 9 were taxonomically assigned to members of the Bacteroidetes with 7 in the *Bacteroides* genus, on the basis of sequence similarity of predicted ORFs to known protein sequences [299].

- Fosmid libraries were generated from the foregut contents of Tammar wallabies and screened for ability to degrade cellulose or xylan. Sequencing and assembly of 33 fosmids resulted in contigs for which the majority were assigned to the order Bacteroidiales and half possessed homologs of genes present in *Bacteroides* PULs, including *susC* and *susD* [235].

- A BAC library constructed from a dairy cow rumen sample was screened for hydrolase activity. Subcloning and sequencing of positive clones revealed that that the endoglucanase genes from two of the clones had blastx best hits to *Bacteroides* species [108].

- A BAC library was constructed using whole intestinal samples from mice, and the library was screened for enhanced adherence to surfaces via biofilm. The two clones isolated were additionally tested for increased intestinal colonization in vivo in the mouse gut. The clones were sequenced and both blastn analysis and tetranucleotide frequency analysis revealed best hits to *Bacteroides* species [342].

I think that the most likely explanation for the successful isolation of Bacteroidetes-derived DNA from screening in *E. coli* is that the expression was due to spurious transcription at incidental *E. coli* σ^{70} consensus promoter-like sequences. Spurious transcription has been discussed in detail in Chapter 4 and simply means that transcription begins at a place on the DNA that is not at the native promoter of a gene. Bacteroidetes DNA could be expressed if transcription were to initiate spuriously and if *E. coli* were able to translate ORFs by recognizing RBSs present on the transcript. This scenario is plausible as *E. coli* has been demonstrated to recognize the RBS of the *B. theta* 16S rRNA operon despite not recognizing its promoter [206].

It is important to note that though this spurious transcription may have facilitated functional screening in the above cases, it cannot be relied on in general because stretches of cloned DNA may lack the sequences that give rise to such transcription in *E. coli*. There is currently a lack of suitable surrogate hosts for systematic functional screening of Bacteroidetes-derived DNA from the human gut metagenome. Given that *Bacteroides* are dominant members of the gut microbial community and some species are well-developed as genetic models, the development of a *Bacteroides* species as a host is a natural choice. In particular, the described genetic tools available for *B. theta* and its genetic tractability make it an ideal candidate for development as a surrogate host for functional metagenomics. This section further discusses the practical and technical aspects of this proposed development.

To use *B. theta* as a host for screening requires constructing a library using a cloning vector that is capable of replicating in both *E. coli* and *B. theta*. The library is constructed and maintained in *E. coli* as usual and subsequently transferred into a recipient *B. theta* strain in a triparental conjugation with the help of a mobilizer strain (Figure 5.4A). The *B. theta* transconjugants can then be plated on media selecting for functional complementation, that is, colonies of *B. theta* carrying cloned environmental DNA able to confer the desired phenotype upon the recipient; for example, wild-type *B. theta* can be selected on media containing an antibiotic to isolate library clones harbouring resistance genes (Figure 5.4B).

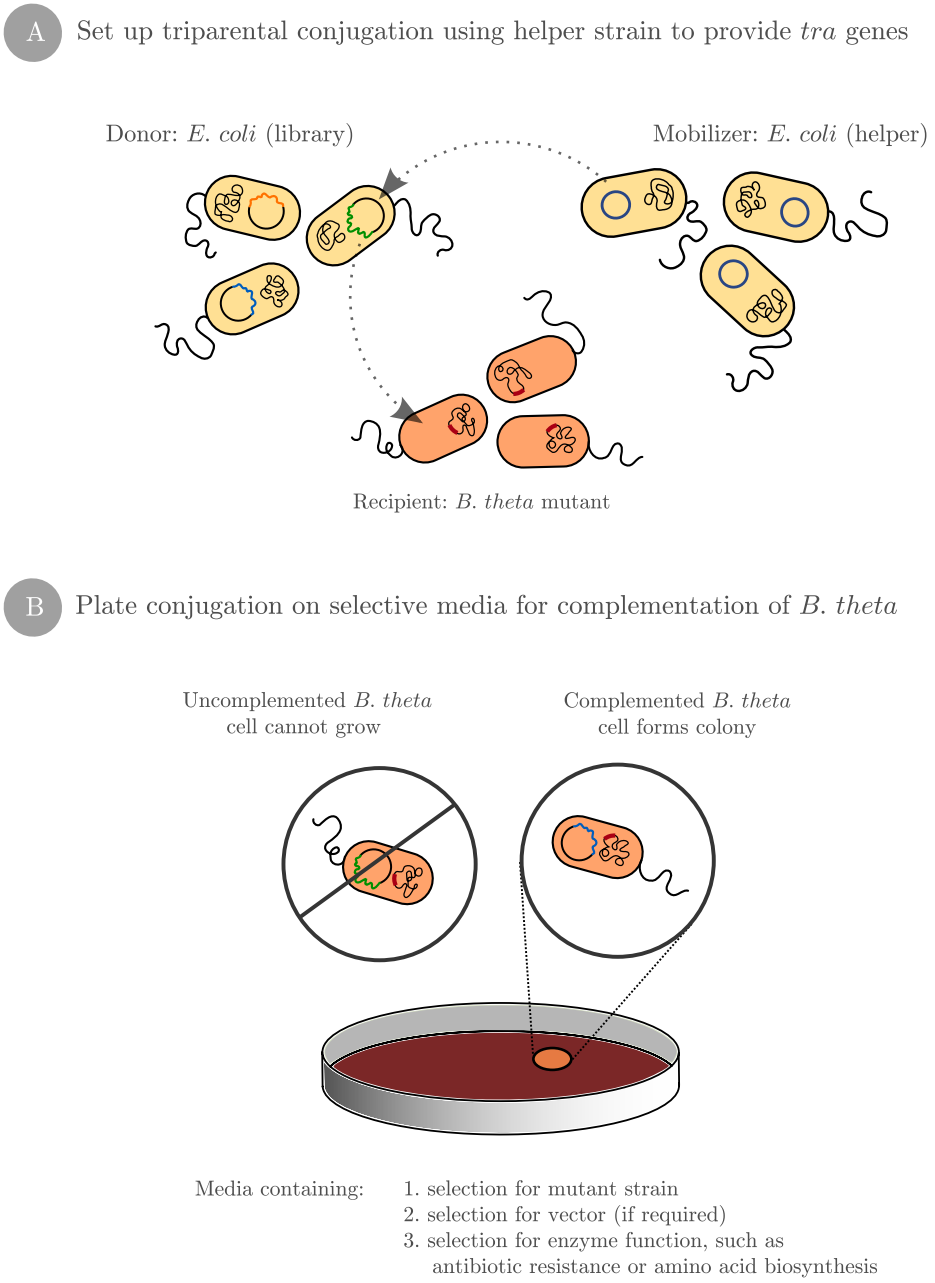


Figure 5.4: Overview of using *B. theta* as a host for functional metagenomics (A) Libraries from *E. coli* are conjugated into a *B. theta* mutant strain using a triparental mating and (B) functionally complemented *B. theta* transconjugants are grown on selective media.

Oxygen tolerance and laboratory culture

The culture of an obligate anaerobe requires growth in the absence of oxygen. *B. theta* is an obligate anaerobe but unlike other organisms that are highly sensitive to the presence of oxygen, it is able to survive for a limited time upon exposure to oxygen, making it convenient to work with in a laboratory setting. *B. theta* possesses enzymes that protect it from both superoxide- and hydrogen peroxide-induced damage to biological molecules, such as superoxide dismutase (SOD) [49], and catalase and other scavenging enzymes [216], respectively. Being an anaerobe, *B. theta* has a central metabolism that is blocked in the presence of oxygen. Its central metabolism has two iron-sulphur cluster enzymes that are sensitive to superoxide or molecular oxygen, which render them inactive; however, both can be repaired rapidly upon return to anaerobic conditions without new protein synthesis, explaining how *B. theta* can recover quickly after exposure to oxygen in the lab [227]. Outside of its central metabolism, *B. theta* has other iron-sulphur proteins that may also be affected by oxygen.

This ability to rapidly repair oxygen-induced damage makes it possible to culture *B. theta* without the use of an expensive anaerobic chamber. *B. theta* can be cultured in liquid using the pyrogallol method to create anerobic conditions inside a typical culture and the indicator dye resazurin can be used to to determine whether this has been done successfully (Figure 5.5 and Section 5.6.2). Culture on solid media in the absence of an anaerobic chamber can be done with the aid of a GasPak jar used in conjunction with one-time-use GasPak sachets that deplete oxygen inside the jar (Figure 5.6A); an even more cost-effective solution is to use inexpensive air-tight containers that can effectively replace GasPak jars (Figure 5.6B and C; Section 5.6.2).

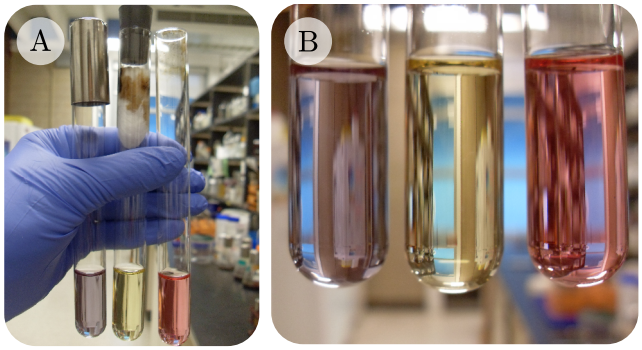


Figure 5.5: Resazurin as an indicator dye for oxidizing/reducing environments The dye resazurin is initially blue-purple in oxidizing conditions (left-most tube), turns irreversibly pink in reducing conditions (right-most tube), and reversibly colourless in anaerobic conditions (centre tube).

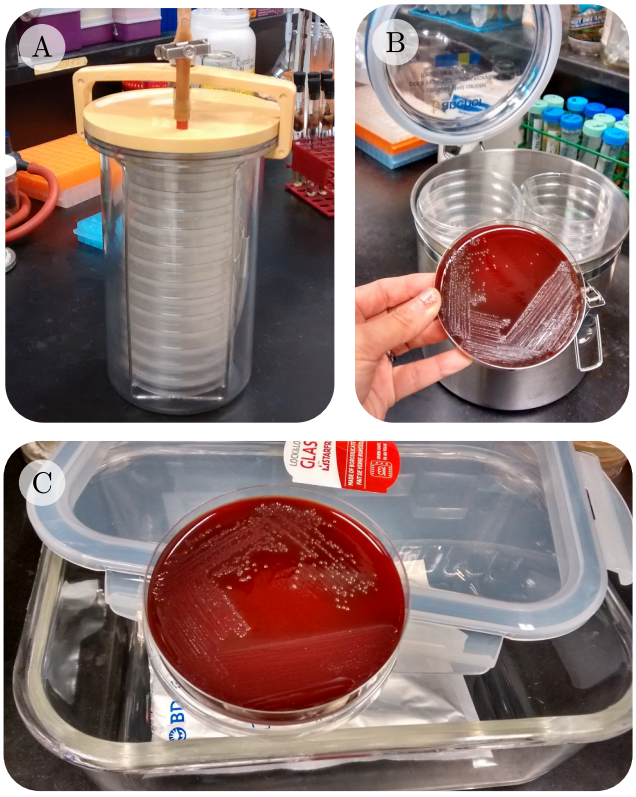


Figure 5.6: Anaerobic jars used in the culture of *B. theta*. (A) GasPak 100 System anaerobic jar, ∼$500; (B) Anchor Hocking stainless steel canister, $20; (C) Lock & Lock glass container, $7.

Stability of cloned *Bacteroides* DNA in *E. coli*

Although I have shown in Chapter 4 that major cloning bias can occur when constructing human gut metagenomic libraries, likely as a result of selection against AT-rich, *rpoD* consensus-containing sequence *in vivo* by the *E. coli* host, this appears to affect members of the Firmicutes to a much greater extent than members of the Bacteroidetes (Figure 4.5 and Figure 4.6). Although a previous study found large segments of *Bacteroides* DNA to be unstable in *E. coli* [265], I have found that using the low-copy cosmid vector pJC8, *Bacteroides*-derived content appears to be similar between the crude extracted DNA and the final cosmid library (Figure 4.7).

Again, the factors affecting the stability of cloned DNA are not well understood; however, my own observations support the notion that there is good representation of metagenomic DNA from the human gut that is likely to be expressed in *B. theta*. It is anticipated that *Bacteroides* DNA will be relatively stable in a low-copy IncP cosmid vector or single-copy fosmid vector, thereby facilitating functional screens in a *B. theta* host.

Functional complementation of *Bacteroides* mutant phenotypes

Though *E. coli* does not recognize *B. theta* promoters, it does recognize *B. theta* RBSs. One might be inclined to suggest that functional screening in *E. coli* could be improved by heterologous expression of the *B. theta* housekeeping sigma factor in *E. coli*; however, although the *B. theta* sigma factor has been shown to be able to interact with the *E. coli* RNA polymerase in vitro, the complex is unable to initiate transcription [317]. But even if this were possible, there is another reason why screening in *B. theta* would be more advantageous: *B. theta*'s various polysaccharide degradation abilities provide a range of phenotypes that can potentially be complemented on selective media, if

the appropriate *B. theta* mutant strains were available. Cosmid or fosmid libraries in particular may be very powerful for functional screening as the large DNA inserts of these libraries would capture the large operons that encode multi-protein systems characteristic of PULs (Figure 5.3).

5.3.5 Aims of this work

The objective of this work was to develop *B. theta* VPI-5482 as a surrogate host to use in functional screening of human gut metagenomic libraries. This required the construction of a library cloning vector with an origin of replication for *B. theta*, and generation of a metagenomic clone library using this vector. The library was used to attempt functional complementation of *B. theta* mutants possessing a suitable and relevant phenotype such as deficiency in the utilization of a particular polysaccharide as compared to wild-type. The goal was to isolate and sequence complementing clones with the hope of finding either novel complementing genes or at least genes different in sequence from the wild-type, thereby demonstrating the effectiveness and potential of using *B. theta* as a host.

5.4 Results and discussion

5.4.1 Problems arising from pUC-based cosmid libraries

Construction of a *B. theta*-compatible pUC-based cosmid pKL3

To be able to screen a library in a *B. theta* host, the library must be constructed using a vector that is able to replicate in *B. theta*. To generate a suitable cloning vector, I first started with the *E. coli*-*B. theta* shuttle vector pAFD1 (Figure 5.7).

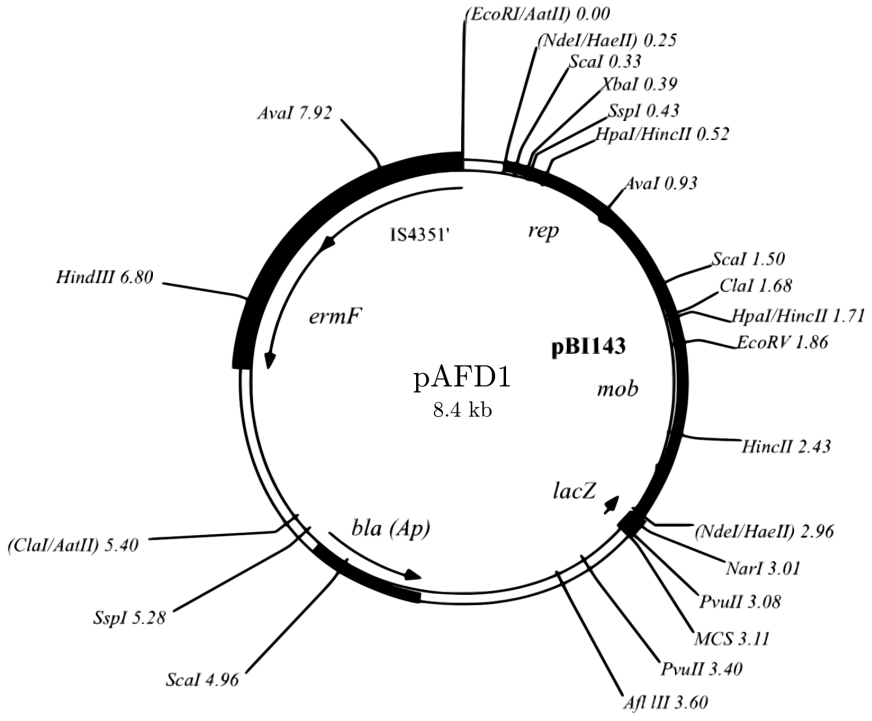


Figure 5.7: *Bacteroides* shuttle vector, pAFD1. Constucted in Abigail Salyers lab [249], this vector was generously shared by Nadja Shoemaker. Unique restriction sites in MCS: EcoRI, SstI, KpnI, SmaI/XmaI, BamHI, SalI, AccI, BspMI, PstI, SphI.

pAFD1 was constructed by ligating the native *Bacteroides* plasmid pBI143 [278] to the *E. coli* vector pUC19 [340], followed by introducing the *ermF* gene for erythromycin resistance in *B. theta*. To this base vector, I added the following elements, which are also summarized in Figure 5.8:

- A *cos* site, by cloning in the BglII fragment from the cosmid pHC79 into the compatible BamHI site of pAFD1 (Figure 5.8A), generating pKL1. The *cos* site enables packaging of DNA into λ phage heads.
- A polylinker (or multiple cloning site) to introduce the Eco72I restriction site (Figure 5.8B), generating pKL2. The Eco72I site was desired because this particular blunt-end restriction site has been used to successfully generate *cos*-based libraries and the preparation of digested, desphosphorylated vector DNA has become routine. The polylinker fragment was generated by phosphorylating and annealing two complementary oligos, KL10 and KL11 (see Section 5.6.4).
- The gentamicin resistance stuffer, as an Eco72I fragment from pJC8 into the Eco72I site of pKL2 (Figure 5.8C), generating pKL3. The stuffer is routinely included in vectors constructed in our laboratory to aid in restriction enzyme cleavage because we find that without a stuffer, digestion does not progress to completion or near-completion.

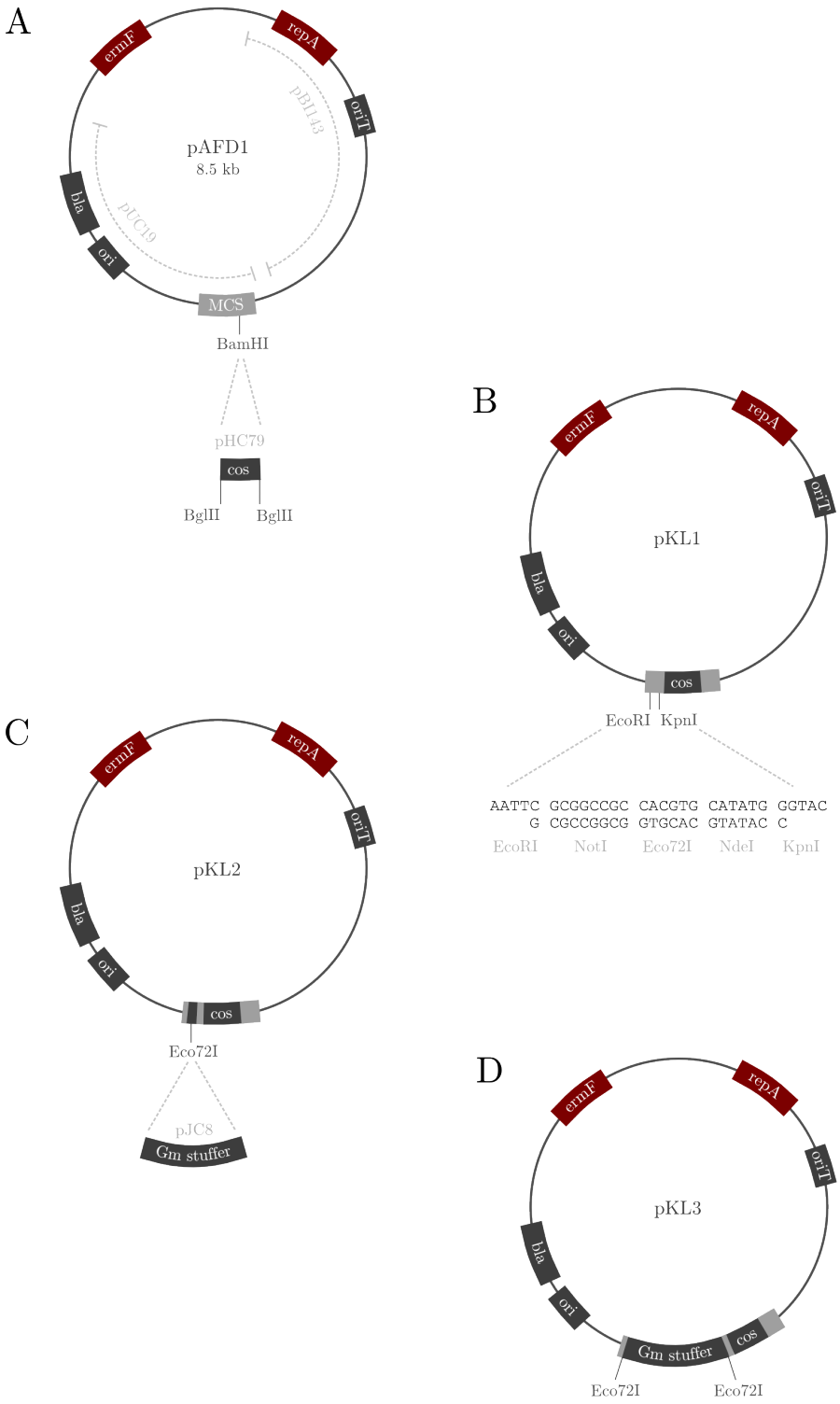


Figure 5.8: Construction of pUC-based *B. theta*-compatible cosmid vector pKL3. The shuttle vector pAFD1 (A) was modified by adding: the *cos* site from pHC79 as a BglII fragment, generating pKL1 (B); a polylinker carrying Eco72I, generating pKL2 (C); the gentamicin resistance stuffer from pJC8, generating pKL3 (D). Note that these are stylized diagrams and are not to scale.

Confirmation of pKL3 functionality; generation of clone libraries using pKL3

After constructing pKL3 (Figure 5.8D) from pAFD1, I then checked that the addition of the *cos* site and polylinker did not interfere with the vector's ability to replicate in *B. theta*. To do this, pKL2 was conjugated from *E. coli* S17-1 into *B. theta*, while pAFD1 was also conjugated as a positive control (Figure 5.9); note that pKL3 was not used because the presence of the gentamicin resistance gene stuffer would have interfered with the gentamicin used as *E. coli* counter-selection in this experiment. The results indicated that the constructed derivative was still functional in *B. theta* and that pKL3 could be used as a library backbone.

I then used this new pUC-based cosmid to construct a metagenomic library from a human fecal sample for screening in *B. theta*. The library was constructed in *E. coli* HB101 and named Charles Lab Gut Microbiome 2 (CLGM2; Figure 5.10A) because it was the second library to be constructed from the pooled stool samples of anonymous donors of the Charles Lab. I also constructed a library using *B. theta* genomic DNA for use as a control in selection experiments (Table 2.12).

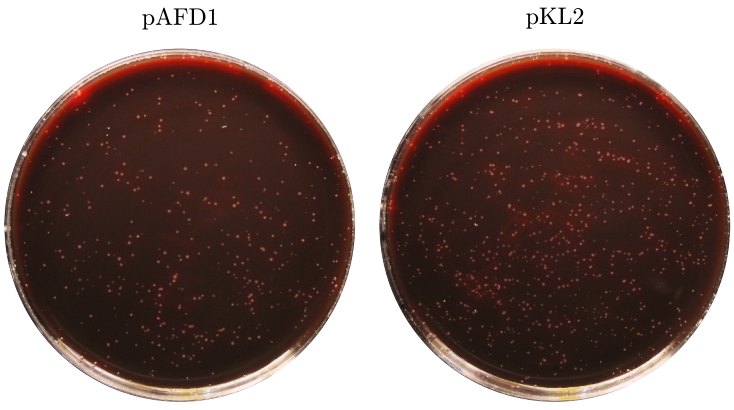


Figure 5.9: Conjugation of positive control pAFD1 and constructed derivative pKL2 into *B. theta*. pAFD1 and pKL2 were separately conjugated into *B. theta* to determine functionality of pKL2. Growth media: BHIH Em_{10} Gm_{200}

Instability of metagenomic insert DNA in high-copy vector

After the library was constructed, colonies were pooled from all the plates (Figure 5.10A) and frozen in aliquots as libraries typically are in the Charles Lab. One aliquot was used to plate isolated colonies from which random clones were selected for examination of insert size: cosmid DNA was miniprepped and subjected to an EcoR1-KpnI double digest to simultaneously release and digest the cloned insert DNA (Figure 5.10B).

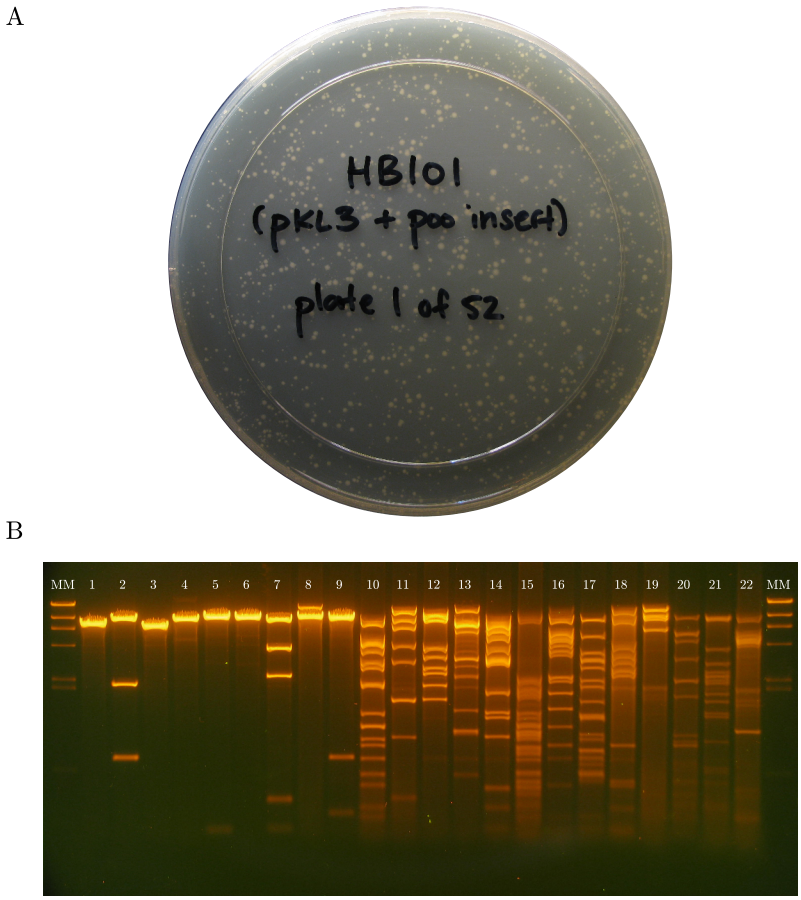


Figure 5.10: Random clones from CLGM2 library exhibit insert loss Randomly selected clones from CLGM2 library were miniprepped, ordered by DNA concentration, and subjected to EcoRI-KpnI double digest, revealing that nearly half have insert sizes much smaller than expected.

The diagnostic digest of 22 random clones yielded an unexpected result: while clones #10 to #22 exhibited restriction patterns typical of large cosmid DNA inserts, clones #1 to #9 had noticeably smaller or even non-existent DNA inserts (Figure 5.10B). This result suggested that a sizeable portion of the library was unstable; the possible causes of this instability that lead to cloning bias were previously discussed in Chapter 4 (see Section 4.4.4). Despite the observed instability, I decided to try to use this library due to time constraints.

Difficulty conjugating CLGM2 metagenomic library

To use the library and attempt to carry out functional screening in a *B. theta* host, the library requires transfer from *E. coli* to *B. theta* via conjugation. To do this, I carried out a triparental conjugation using the library strain HB101(CLGM2) as donor, *B. theta* as recipient, and J53(R751) as helper (Figure 5.11A); I also simultaneously conjugated the empty vector from HB101(pKL2) into *B. theta* as a control. It was necessary to use R751 as the helper plasmid instead of the commonly used pRK600 or pRK2013 to avoid plasmid incompatibility issues as pKL2/pKL3 and pRK600/pRK2013 are all ColE1-related plasmids.

The conjugation was plated on media selecting for the transconjugant, *B. theta* carrying the conjugated cosmids; recall that *B. theta* has natural resistance to nalidixic acid and aminoglycosides, such as kanamycin. While the empty vector showed an acceptable conjugation efficiency, the efficiency of the CLGM2 library was poor (Figure 5.11B). This poor transfer of the library was not specific to the *B. theta* recipient, as conjugation was also poor for an *E. coli* recipient when tested (data not shown). The reason for the library's poor transfer is not clear, although it maybe be related to the high-copy number of the vector backbone in combination with maintaining large DNA inserts that may be transcriptionally active.

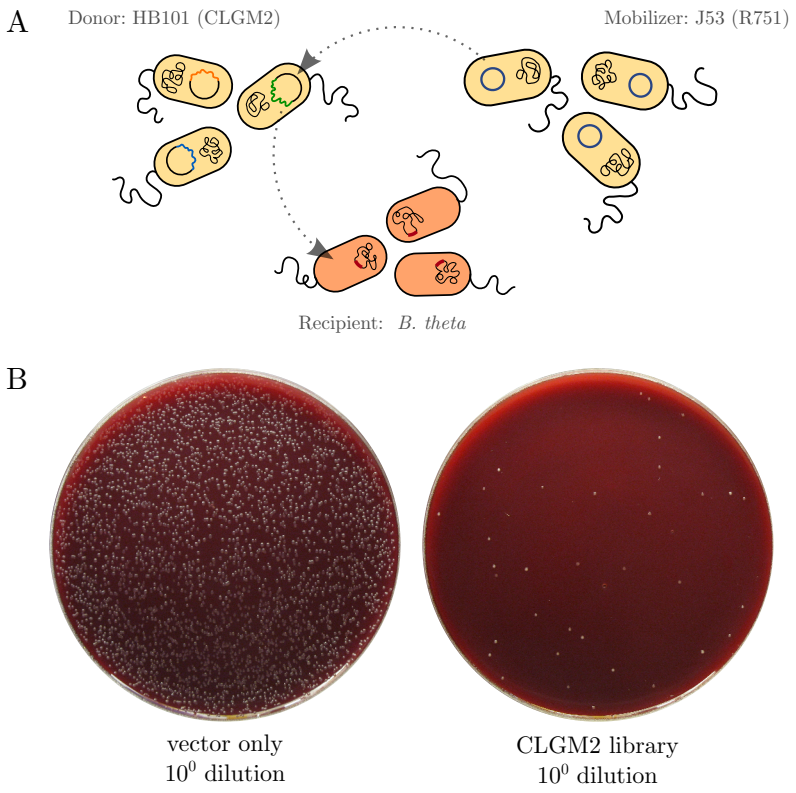


Figure 5.11: Triparental conjugation of CLGM2 library into *B. theta*. (A) Overview of triparental conjugation experiment for transfer of CLGM2 library from *E. coli* HB101 donor to *B. theta* recipient. (B) Result of conjugation into *B. theta* of vector alone (left) or CLGM2 library (right). Growth media: BHIH Em_{25} NA_{25} Km_{200}

A poor efficiency of conjugation into *B. theta* severely hinders the success of functional screens because library clone DNA cannot be transferred to the recipient in order to undergo selection. In combination, the instability of insert DNA in the library and the poor transfer of the library into recipient cells rendered the CLGM2 library effectively unuseable. Therefore, I decided to re-build the system, using a single-copy vector backbone to avoid possible high copy number-related problems.

5.4.2 Efficient conjugation of fosmid-based libraries into *B. theta*

Construction of a *B. theta*-compatible fosmid pKL13

For the backbone of the new library cloning vector, I decided to use the commercial vector pCC1FOS (Figure 5.12A). The properties, advantages, and disadvantages of this vector are discussed in greater detail in Section 6.3.2 of the following Chapter 6.

Briefly, pCC1FOS replicates as a single-copy fosmid in *E. coli* strains as it carries the F plasmid origin of replication. In addition, it carries the RK2 origin of replication which, combined with the *trfA* gene product, increases copy number in members of the Proteobacteria. For example, the commercial strain *E. coli* EPI300 has been designed for use with pCC1FOS: EPI300 carries *trfA* under the control of an arabinose-inducible promoter, which allows the fosmid to be maintained at single-copy but induced to a higher copy number when desired. The vector also carries the chloramphenicol resistance gene for selection in *E. coli*.

pCC1FOS is used widely for the construction of fosmid libraries; both the popularity and the properties of pCC1FOS made it an attractive choice for use as a base vector for construction of *B. theta*-compatible libraries. The following points below describe the step-by-step construction of the pCC1FOS *B. theta*-compatible derivative pKL13; the steps are also summarized graphically (Figure 5.12):

- The gentamicin resistance stuffer was added, as an Eco72I fragment from pJC8 into the Eco72I site of pCC1FOS, generating pKL4 (Figure 5.12B). As previously, the stuffer was added to aid digestion of the vector for library cloning.
- An *oriT* sequence was added to allow the vector to be conjugated between strains, particularly between *E. coli* and *B. theta*. The sequence was PCR-amplified as an ∼800-bp fragment from pJC8 using primers KL12 and KL13 with HindIII

adapters, and ligated into the unique HindIII site of pKL4, generating pKL5 (Figure 5.12C). Though the actual functional *oriT* sequence is only ∼100 bp, including the surrounding region reportedly improves transfer frequency by two orders of magnitude [113].

- A fragment from pAFD1 was added, which includes (a) the *ermF* gene encoding erythromycin resistance as a selectable marker for *B. theta* and (b) the *repA* gene and internal *ori* for replication in *B. theta*. The fragment was PCR-amplified as an ∼4-kb fragment from pAFD1 using primers KL14 and KL15 with EcoRI adapters, ligated into pJET1.2 forming pKL8, and subcloned as an EcoRI fragment from pKL8 into the unique EcoRI site of pKL5, generating pKL6 (Figure 5.12D). Note that because the sequence of pAFD1 was not known, I deduced the fragment's probable sequence and designed PCR primers based on related vectors that have been sequenced: the sequence of *repA* was determined from the native *B. fragilis* plasmid pBI143 [278]; the sequence of *ermF* was determined from the vectors pFD288 and pFD1146 [228, 278], which are related to pAFD1 through the shared *ermF* marker that was originally from pBF4 [326]. I was uncertain about the sequence for the portion between the *ermF* and *repA* elements, so to obtain the complete sequence, I carried out primer walking (see Section 5.6.6).

- Deletion of the gentamicin resistance gene stuffer, generating pKL7 (Figure 5.12E). At this time, I was finishing my work on Chapter 4, and decided to include transcriptional terminators that flank the cloning site in my new vector (see next point), which required removing this stuffer.

- In place of the gentamicin resistance stuffer, I cloned in what I called the "transcriptional terminator" fragment. The elements of this fragment are discussed in detail in Section 6.4.1. The fragment includes: two unidirectional transcriptional terminators that stop potential insert-initiated transcription from going into the vector backbone, and a stuffer comprising a gentamicin resistance gene as well as a P_{tac} promoter for terminator testing purposes (see Section 6.4.3). The fragment was cloned as a blunt SwaI fragment from pKL9 into the blunt Eco72I site of pKL7, terminating the existing Eco72I sites but reintroducing new Eco72I sites, which flank the stuffer (Figure 5.12F).

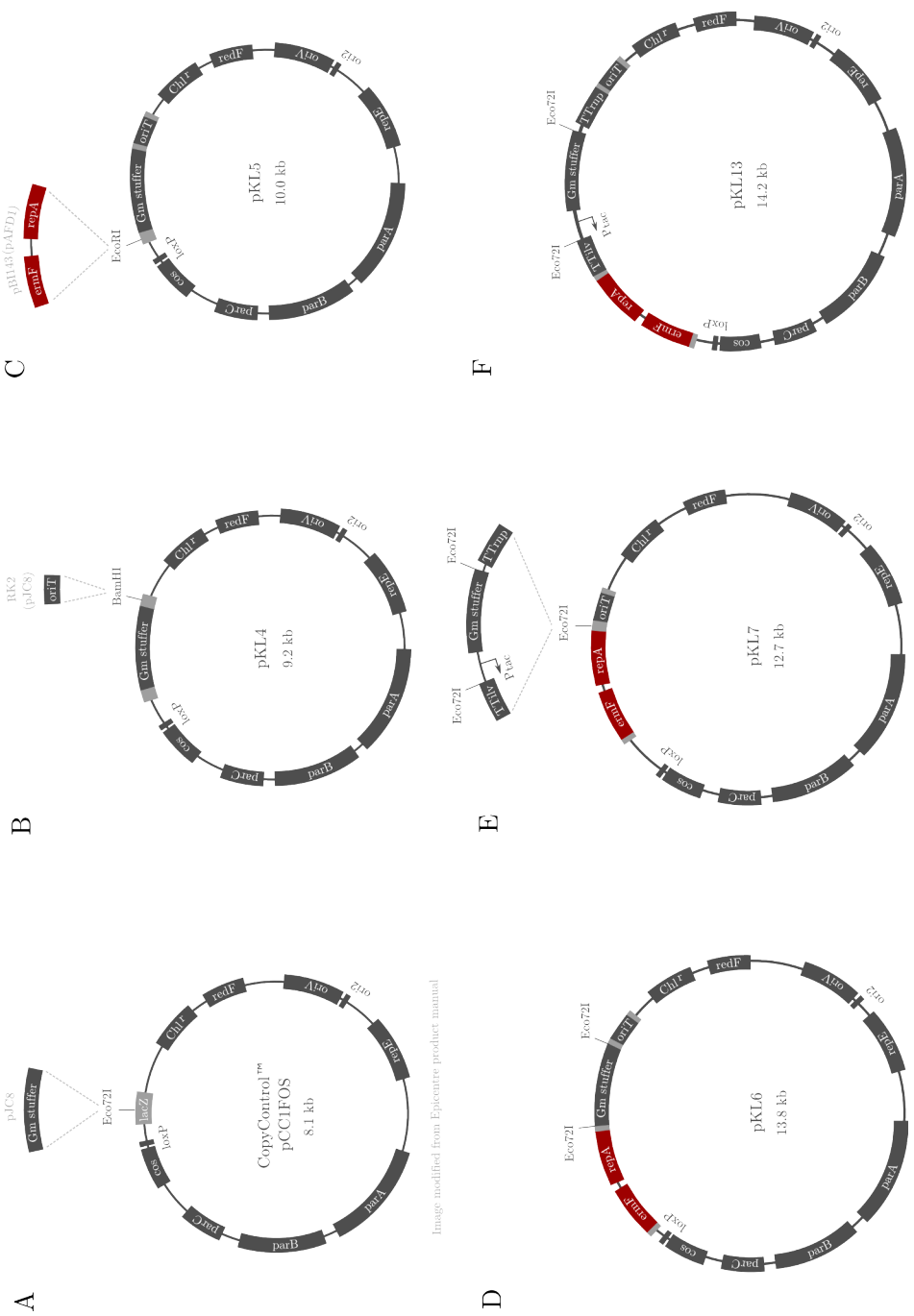


Figure 5.12: Construction of *B. theta*-compatible fosmid vector pKL13. The commerical vector pCC1FOS (A) was modified by adding the gentamicin resistance stuffer from pJC8, generating pKL4 (B); the fragment carrying the *oriT* from pJC8 with BamHI adapters, generating pKL5 (C); the fragment from pAFD1 carrying *ermF* and *repA*-ori with EcoRI adapters, generating pKL6 (D); deleting the gentamicin resistance stuffer, generating pKL7 (E); adding the transcriptional terminator fragment, generating pKL13 (F). Note that these are stylized diagrams and are not to scale.

Analysis of new vector passaged through *B. theta*; generation of clone libraries using pKL13

After constructing the new vector, I performed a check to see that the vector was behaving as expected. Because the pCC1FOS backbone is not a vector that is normally used in the *Bacteroides*, the check was important to make sure that the new vector is stable in *B. theta* and was therefore appropriate to use as a library cloning vector.

To perform the check, I used pK11; note that pKL11 is identical to pKL13 except for a point mutation in one of the transcriptional terminators and the removal of the stuffer between the Eco72I sites (see Table 2.2). I carried out a triparental mating to conjugate pKL11 from *E. coli* HB101 to *B. theta*, using DH5α(pRK600) as helper; following this, six clones of *B. theta* carrying pKL11 were selected and streak-purified, fosmid DNA was isolated from the clones, and the DNA was re-introduced into *E. coli* for subsequent isolation and restriction analysis (Figure 5.13A). Note that plasmid miniprepped DNA from *B. theta* cannot be analyzed directly because it contains DNA from *B. theta*'s own native plasmid (see Section 5.3), which complicates restriction digest analyses.

The *B. theta*-passaged fosmid DNA isolated from *E. coli* was digested and compared to digested pKL11 from *E. coli* that had not been passaged through *B. theta* (Figure 5.13B). From the results, it can be seen that the passaged vector DNA is the same size as the original vector, meaning undesired recombination events that may have increased or decreased the vector size did not occur. Importantly, this experiment demonstrates that the vector is stable and can be isolated intact by plasmid miniprep from *B. theta*; this point will returned later in Section 5.4.4 where I encounter difficulties isolating plasmid DNA.

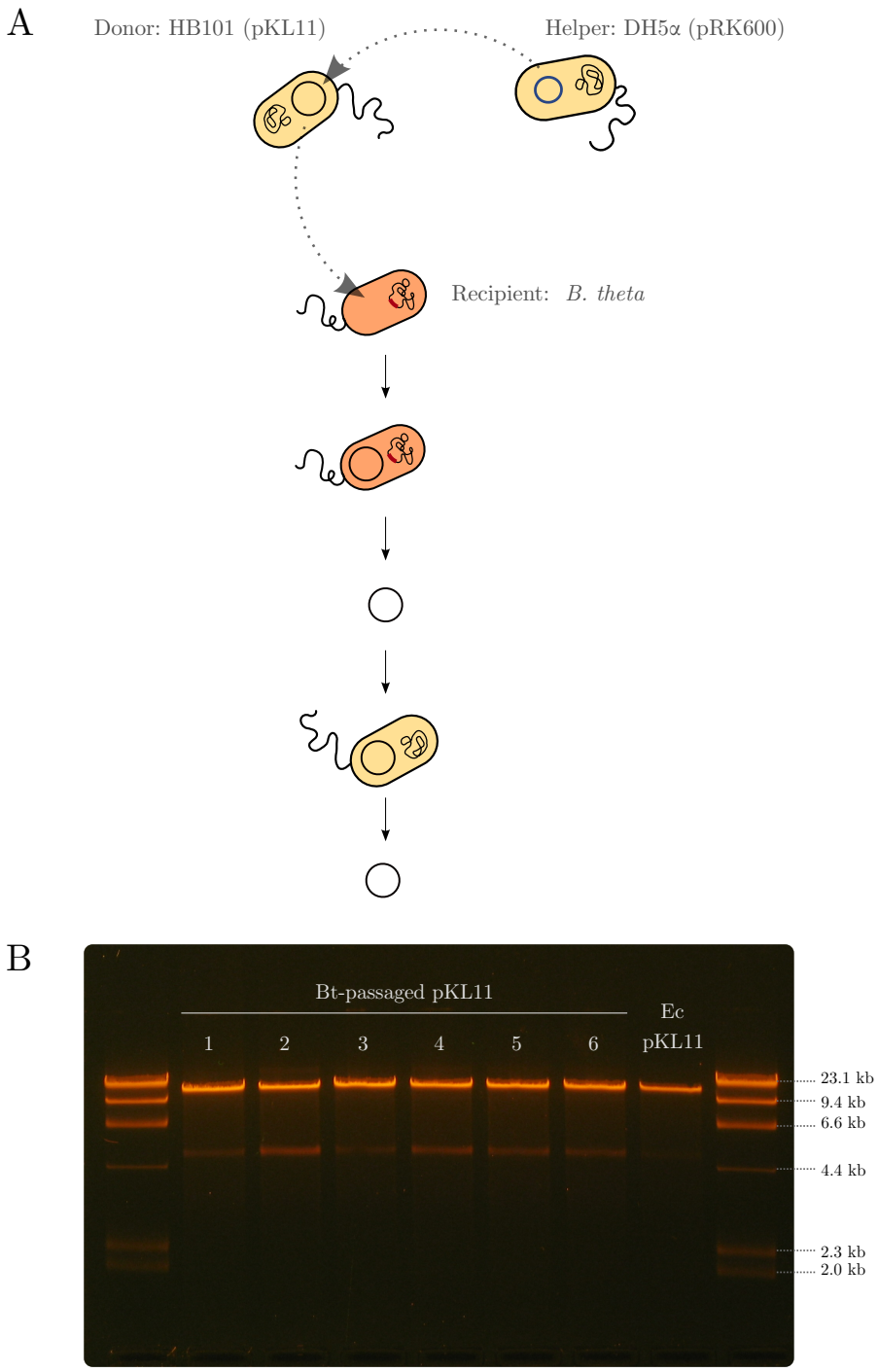


Figure 5.13: Analysis of fosmid vector DNA passaged through *B. theta* and re-introduced into *E. coli*. (A) pKL11 was conjugated from *E. coli* to *B. theta* in a triparental mating; plasmid DNA was isolated from six *B. theta* clones carrying pKL11, re-introduced into *E. coli*, and isolated from *E. coli* for analysis (B) Gel electrophoresis of Eco72I-digested *B. theta*-passaged pKL11, against a control preparation of pKL11 from *E. coli*.

After making sure the fosmid vector was stable in *B. theta*, I used pKL13 to generate clone libraries. Library construction was carried out using a protocol as described earlier with the exception that the Eco72I stuffer was not separated from the vector preparation prior to ligation to the genomic/metagenomic DNA (see Section 5.6.9 for technical details). As before, I generated two libraries to use in selection experiments: a *B. theta* genomic library named BT3, and human gut metagenomic library named CLGM3 (see Table 2.12). Both libraries were constructed in an EPI300 background, because EPI300 offers copy-number inducibility and I found that it transduces at least as well as HB101 (Table 5.4).

Table 5.4: Transduction efficiency using HB101, S17-1, or EPI300.

Strain used	Number of transductants	
	Trial 1 count	Trial 2 count
HB101	162	413
S17-1	34	61
EPI300	592	430

Conjugation of CLGM3 metagenomic library into *B. theta* host

Hoping that using new single-copy vector backbone would resolve the conjugation problems encountered, I performed a triparental mating to transfer the library from EPI300 to *B. theta*, using HB101(pRK2013) as helper (Figure 5.14A). A similar mating using the pKL13 vector alone was done alongside as a control. Note that the pRK2013 helper is a ColE1 plasmid, and is compatible with pKL13, which carries the F and RK2 origins.

The mating was plated on media selecting for *B. theta* transconjugants (Figure 5.14B). Comparing the dilution plate giving rise to colonies between Figure 5.11B and Figure 5.14B, it can be seen that the conjugation efficiency of the vector alone is improved using the single-copy fosmid, but more importantly, the efficiency of CLGM3 is showing an improvement of easily one thousand-fold. The marked improvement in transfer of the library meant that it was well-suited for functional screening in *B. theta*. Before proceeding to a screen, however, I first wanted to more quantitatively assess the conjugation efficiencies.

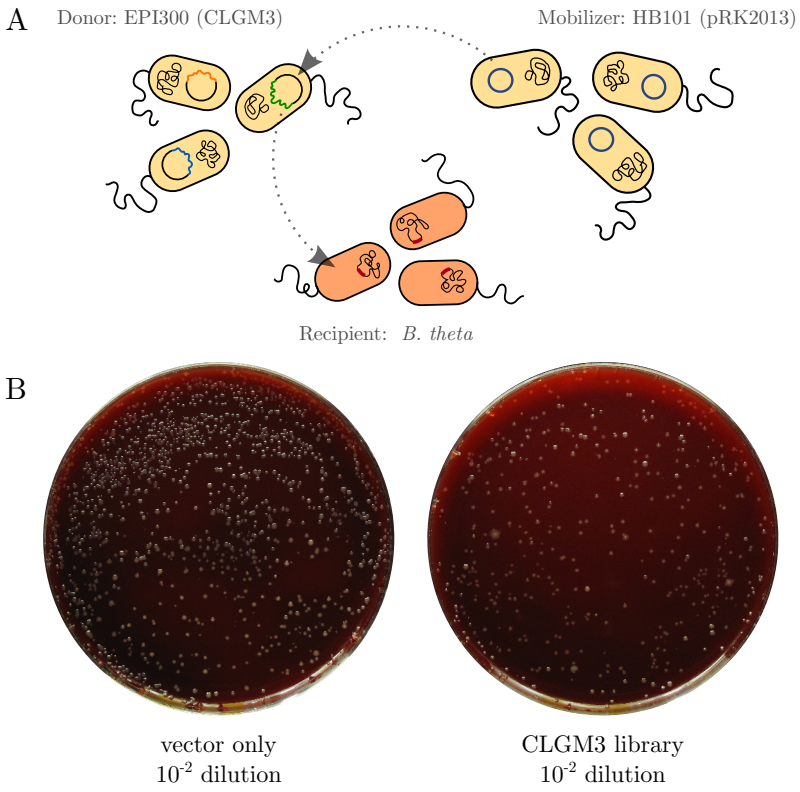


Figure 5.14: Triparental conjugation of CLGM3 library into *B. theta*. (A) Overview of triparental conjugation experiment for transfer of CLGM3 library from *E. coli* EPI300 donor to *B. theta* recipient. (B) Result of conjugation into *B. theta* of vector alone (left) or CLGM3 library (right). Growth media: BHIH Em_{25} NA_{25} Km_{200}

Conjugation efficiencies

To calculate the efficiency of conjugation of both empty pKL13 and the CLGM3 library into *B. theta*, I repeated the triparental conjugations as depicted in Figure 5.14A. The matings were serially diluted and plated on media with different antibiotics to select for the donor, recipient, or transconjugant:

- Donor: *E. coli* EPI300 (pKL13/CLGM3), on LB Cm_{10}
- Recipient: *B. theta*, on BHIH NA_{25} Km_{200}
- Transconjugant: *B. theta* (pKL13/CLGM3), on BHIH Em_{25} NA_{25} Km_{200}

From counting the number of colonies arising on the plates for each of the donor, recipient, and transconjugant dilutions, it was possible to determine the conjugation efficiency with respect to the donor as well as the recipient, which is simply the number of transconjugants divided by the number of donors or recipients, respectively (Table 5.5).

Table 5.5: Conjugation efficiency of pKL13 vector and CLGM3 library into *B. theta*.

	pKL13 vector only	CLGM3 library
relative to donor	2.1×10^{-5}	8.2×10^{-6}
relative to recipient	2.6×10^{-2}	1.1×10^{-2}

For matings in which *B. theta* is the recipient, it would be most useful to refer to the conjugation efficiency with respect to the recipient as this is the limiting factor; this is because conjugations are performed aerobically where *B. theta* growth can only occur after the *E. coli* cells have formed a lawn, thereby protecting *B. theta* from atmospheric oxygen (see Section 5.6.8 for details on methods); hence, the recipient cell count is much lower than the donor cell count.

The conjugation efficiency was calculated to be 2.6×10^{-2} for pKL13 and 1.1×10^{-2} for the CLGM3 library (Table 5.5). This means that 2-3% of *B. theta* cells present in the pKL13 conjugation will receive the vector; for the library, this number is closer to 1%. Though the fraction of transconjugants obtained from a mating is not as high as, for example, matings involving *Sinorhizobium meliloti* as recipient [94], the frequency of transfer was sufficiently high to move forward and try functional screening using *B. theta* as an expression host.

5.4.3 Functional complementation using a *B. theta* host

Construction of *B. theta* single recombinant amino acid auxotrophs and attempt at complementation

To execute a functional screen as described in Figure 5.4, a prerequisite is having a *B. theta* mutant whose phenotype can be complemented and, ideally, the complemented mutant can be selected rather than screened for. During my visit to laboratory of Eric Martens at the University of Michigan, I constructed two mutants for this purpose; both were mutants in amino acid biosynthesis: the first was a threonine auxotroph and the second, a tryptophan auxotroph.

For a quick construction, rather than making clean deletions, I settled for generating single recombinant mutants by disrupting the *thrC* (BT_2401) and *trpD* (BT_0530) genes. To do this, I PCR-amplified and cloned an internal fragment from either the *thrC* or *trpD* gene into *B. theta* suicide vector pKNOCK-*bla-tetQ* (Figure 5.15A), generating pKL21 and pKL22, respectively. The constructed plasmids were then mated into wild-type *B. theta*; pKNOCK-*bla-tetQ* is unable to replicate in *B. theta* and thus selection for tetracycline resistance allows isolation of single recombinants in which the plasmid has integrated into the genome at the locus specified by the cloned fragment. I isolated

threonine and tryptophan auxotrophs and checked their phenotype on minimal media; as expected, the *thr* mutant could not grow unless threonine was supplemented and the *trp* mutant could not grow unless tryptophan was supplemented (Figure 5.15B).

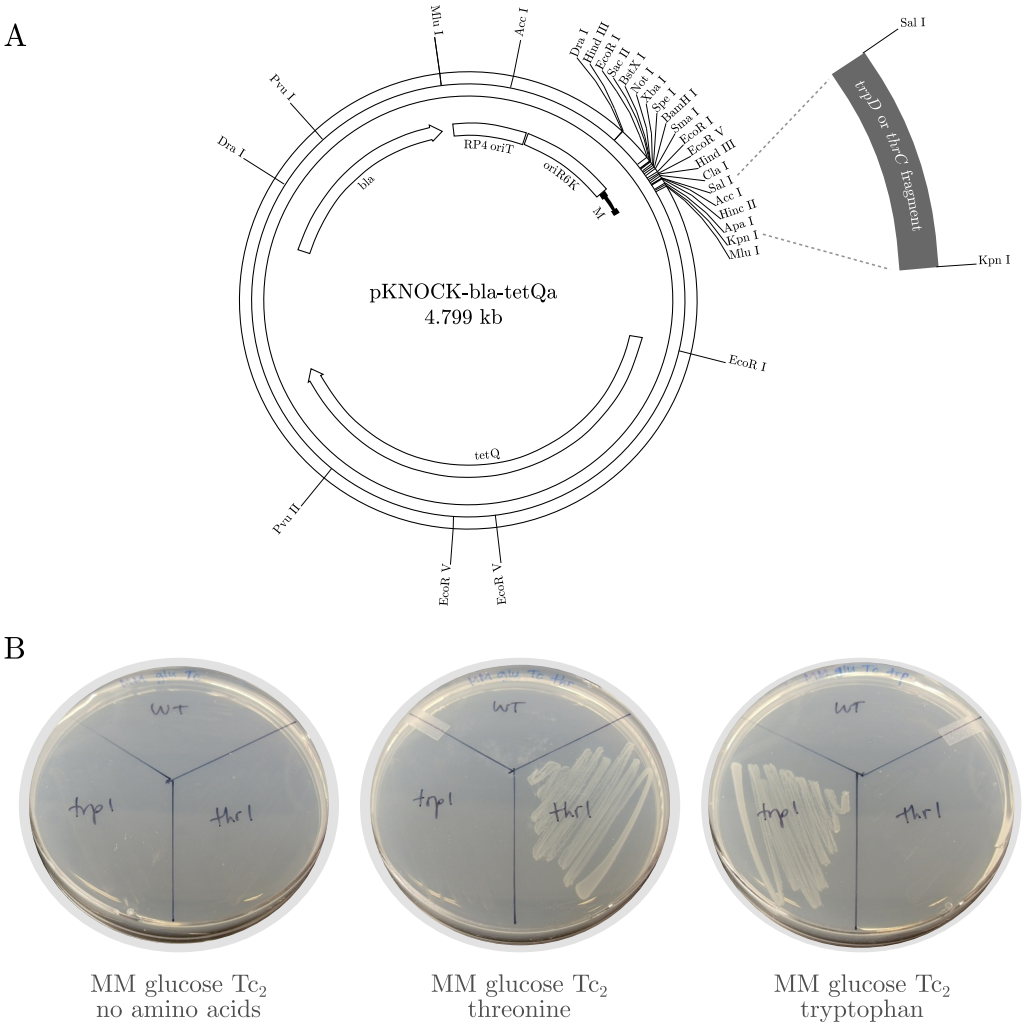
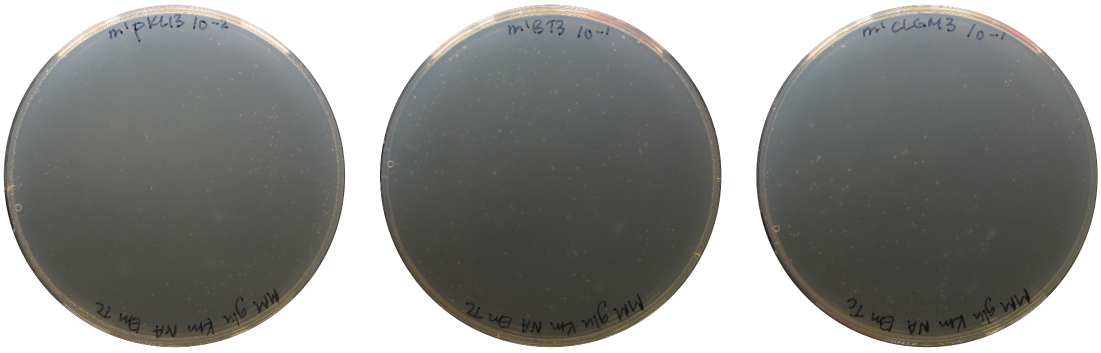


Figure 5.15: Construction of *B. theta* single recombinant amino acid auxotrophs. (A) A fragment of either *thrC* or *trpD* was PCR-amplified and cloned into the *B. theta* suicide vector pKNOCK-*bla*-*tetQ*; adapted from [200] (B) Phenotypic check of constructed mutants on minimal media; WT: wild-type, trp1: tryptophan auxotroph, thr1: threonine auxotroph.

Out of the two mutants, I decided to use the *B. theta* tryptophan auxotroph in the first functional screen of the CLGM3 library. I mated the CLGM3 library from *E. coli* EPI300 into the *B. theta* tryptophan auxotroph, and selected for complemented transconjugants on minimal media with no supplemented amino acids; as negative and positive controls, I also mated the vector, pKL13, as well as the *B. theta* genomic library, BT3, respectively (Figure 5.16). Unfortunately, though the CLGM3 metagenomic library and BT3 genomic library matings gave rise to colonies on the selective media, the vector-only control did as well – at an even greater frequency. It was most likely that the single recombinant mutant was unstable and the vector was recombining out of the chromosome, despite the inclusion of tetracycline as selection; that is, the mutant was reverting to wild-type phenotype under the selection for functional tryptophan biosynthesis genes. The greater frequency of reversion seen for the vector over the two libraries can likely be attributed to a greater efficiency of conjugation for smaller plasmids; this was also evident in Figure 5.14B.



Vector only
10^{-2} dilution

B. theta genomic library
10^{-1} dilution

Gut metagenomic library
10^{-1} dilution

Figure 5.16: Results of functional screen for tryptophan biosynthesis genes in *B. theta* single recombinant. The vector-only control, pKL13 (left), the BT3 genomic library (centre), and the CLGM3 metagenomic library (right) were mated into the *B. theta* tryptophan auxotroph and conjugations were plated on media selecting for complementation. Growth media: MM glucose Tc_2 Em_{25} NA_{25} Km_{200}

It was most regrettable that I did not construct deletion mutants instead of single recombinant mutants: if the *trpD* gene were deleted instead of simply interrupted, there would be no possibility of reversion to wild-type phenotype. Given the time constraints, however, it was not feasible to begin the construction of clean deletions of the *thrC* or *trpD* genes; rather, as Eric Martens suggested, I made use of a *B. theta* deletion mutant that had been previously constructed and characterized.

Successful complementation of the *B. theta chuR / anSME* mutant

The mutant chosen for the next attempt at functional complementation was *B. theta* Δ*chuR*, also called Δ*anSME* [17]. The *chuR/anSME* gene (BT_0238) was first identified by Abigail Salyers' group through transposon mutagenesis as a ***r***egulator of ***c***hondroitin sulfate and ***h***eparin ***u***tilization [44]. Knocking out this single gene renders *B. theta* unable to grow on chondroitin sulfate or heparin as sole carbon source, as shown in Figure 5.17 [†].

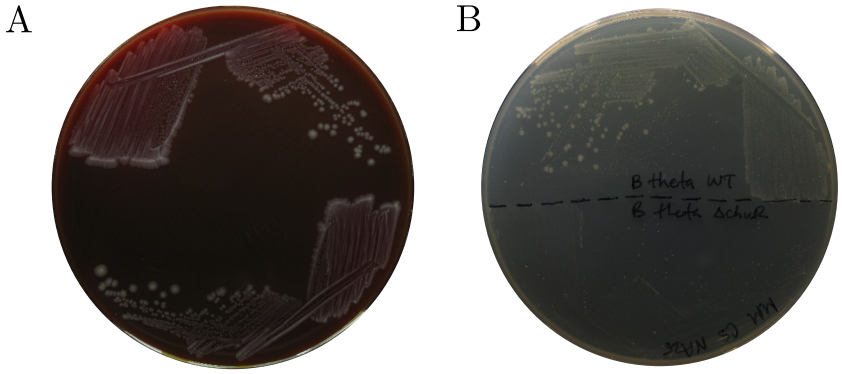


Figure 5.17: Phenotype of *B. theta* wild-type and Δ*chuR* mutant. Phenotype of the *B. theta* wild-type (top half) and Δ*chuR* mutant (bottom half) on BHIH complex media (A) or minimal media with chondroitin sulfate as sole carbon source (B).

[†]Note that the strain isogenic to Δ*chuR* is Δ*tdk*, which is in turn isogenic to the wild-type. The wild-type and Δ*tdk* exhibit comparable growth on chondroitin sulfate; see Appendix D.1

Chondroitin sulfate is a polysaccharide that is composed of alternating N-acetyl-galactosamine and glucuronic acid residues, with the sugar residues carrying sulfate groups at certain positions [80]. The breakdown of this polysaccharide requires the action of sulfatase enzymes, of which *B. theta* may encode up to 28 [17]; however, the sulfatases must be modified post-translationally by the product of the *chuR/anSME* gene, an ***an***aerobic ***s***ulfatase ***m***aturase ***e***nzyme [18]; without the post-translational modification, the sulfatases are not active. The 1.2-kb *chuR/anSME* gene is part of a three-gene operon but is currently the only characterized member (Figure 5.18). The phenotype being dependent on the single *chuR* gene, as well as the clean phenotype of the *B. theta* Δ*chuR* mutant on chondroitin sulfate as sole carbon source (Figure 5.17B), make it a very good candidate for functional complementation.

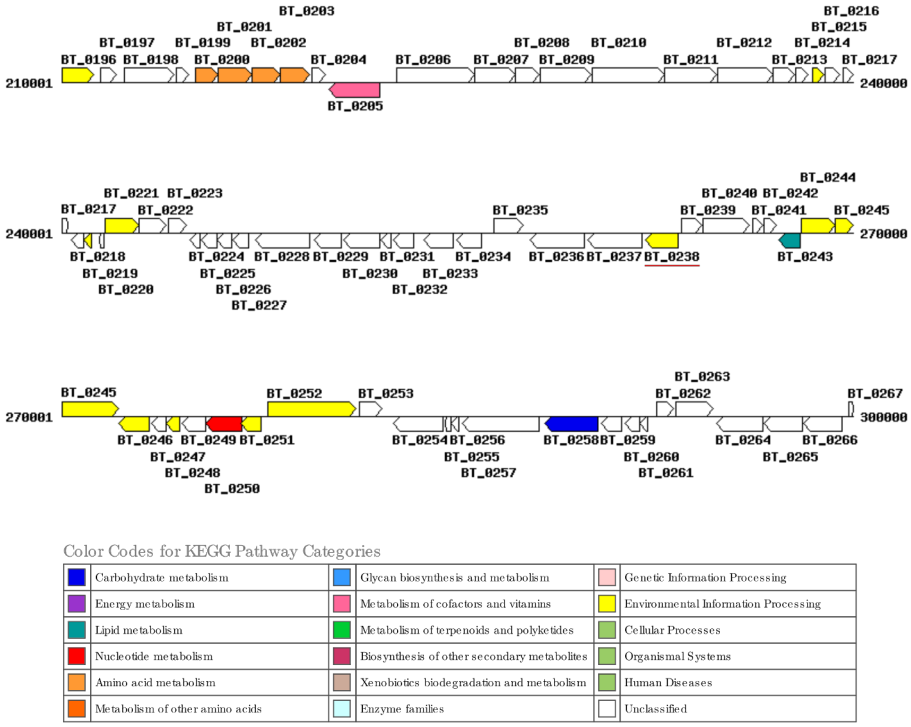


Figure 5.18: Genomic region of the *B. theta chuR* (*anSME*; BT_0238) gene. The 1.2-kb *chuR* gene (underlined in red) of the *B. theta* genome and its surrounding region. Adapted from the KEGG Genome Database [143]

To screen the CLGM3 library for *chuR/anSME* genes, I once again performed a triparental conjugation, mating the CLGM3 library from *E. coli* EPI300 into the *B. theta* Δ*chuR* strain, selecting on minimal media with chondroitin sulfate. Also as before, for negative and positive controls, respectively, I performed matings of the vector, pKL13, as well as the *B. theta* genomic library, BT3. Each of the three conjugations was plated on multiple plates to select for transconjugants with ability to use chondroitin sulfate as sole carbon source; one of each is shown in Figure 5.19.

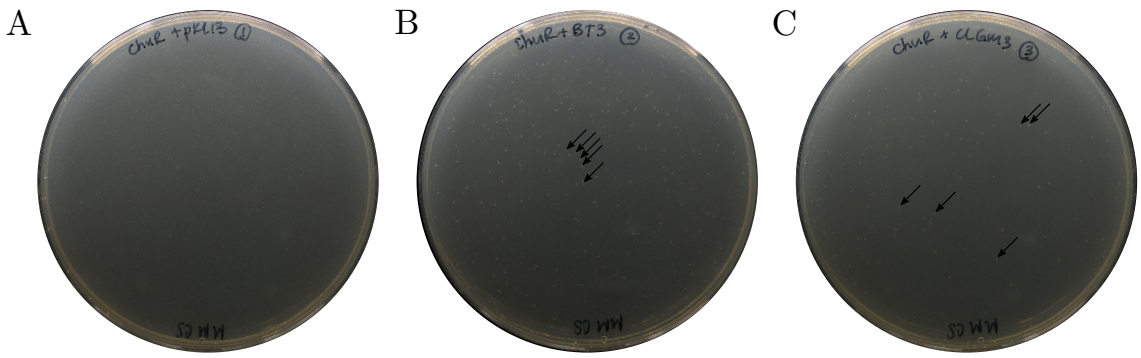


Figure 5.19: Results of functional screen for *chuR/anSME* genes using *B. theta* Δ*chuR* background. Selection plates onto which conjugations were spread, using as donor: pKL13 vector only (A), BT3 genomic library (B), and CLGM3 metagenomic library (C). Black arrows indicate several examples of isolated colonies. Growth media: MM chondroitin sulfate

Unlike my first attempt at complementation, the negative control had no colonies (Figure 5.19A). The positive control, using *B. theta*'s own genomic DNA to complement the mutant, resulted in colonies, as was expected (Figure 5.19B). Most importantly, the experimental mating using the CLGM3 metagenomic library also yielded colonies (Figure 5.19C). This result indicates that the *B. theta* Δ*chuR* mutant can be complemented using cloned metagenomic DNA from the human gut, although the phylogenetic origin of the complementing DNA remained to be determined. From the BT3 and CLGM3

plates, I streak-purified colonies to confirm the restored phenotype and to purify the clone in the case that one colony arose from more than one complemented cell. The positive clones from the streak-purification provide clear evidence that the mutant's ability to grow on chondroitin sulfate has been restored (Figure 5.20). After the difficulties that I encountered, that the functional screen seemed to be working well was promising. The next step was to isolate the complementing fosmids from *B. theta* for eventual restriction analyses and DNA sequencing.

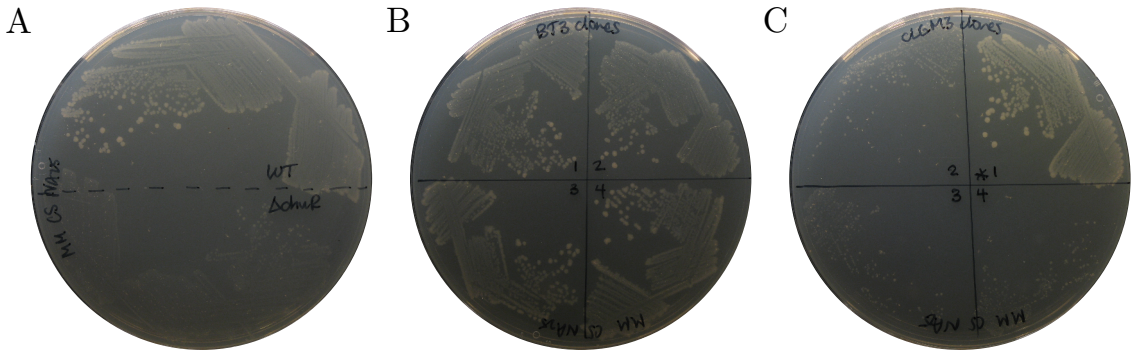


Figure 5.20: Streak purification of complementing *chuR/anSME* clones. Control streaks of wild-type and Δ*chuR* (A), four complementing clones from the BT3 library (B), and one complementing clone from the CLGM3 library (C). Growth media: MM chondroitin sulfate

5.4.4 DNA of positive clones appears to be integrated into the host genome

In Section 5.4.2, I showed that the fosmid vector could be isolated from *B. theta* and re-introduced into *E. coli*. Now, with streak-purified complementing clones from the successful *chuR/anSME* screen of both the BT3 library and the CLGM3 library, I needed to employ the same method to isolate the clone DNA from the *B. theta* Δ*chuR* host. I inoculated the clones in liquid media for a plasmid miniprep, and included the antibiotic erythromycin in the media to ensure that the fosmid backbone was present. The first clue that something was amiss was when only about half of the clones grew up in the liquid media containing the antibiotic. I proceeded to do the plasmid miniprep for those clones that grew; when I attempted to transform *E. coli* with the preparation, however, I did not obtain transformants for any of the samples, which indicated that there was no fosmid DNA isolated from *B. theta*.

At this point, I hypothesized that the fosmid DNA may have integrated into the host genome. If the DNA were in fact integrated into the genome, this would be unfortunate as the functional metagenomic method employed in our lab hinges on being able to retrieve the DNA for sequence analysis. With this hypothesis in mind, I isolated genomic DNA from the same clones to analyze, that is, from the clones that did grow in the presence of erythromycin. Genomic DNA was prepared from the following strains for analysis:

- BT3 library: *chuR* clones #2, 5, 6, 9, 10, in *B. theta* Δ*chuR* background
- CLGM3 library: *chuR* clones #1, 2, 3, 4, 5, 8, 9, in *B. theta* Δ*chuR* background
- *B. theta* Δ*chuR*, as a control
- wild-type *B. theta*, as a control

To establish whether the genomic DNA contained integrated fosmid DNA, I performed a PCR to test for the presence of the fosmid's *oriT* sequence, and I included pKL13 as a positive control (Figure 5.21A). As suspected, all of the clones from the BT3 and CLGM3 library were positive for the *oriT* while the wild-type and Δ*chuR* controls were negative. This suggested that the fosmid DNA was integrated into the genome of the Δ*chuR* background; the location of integration is uncertain but recombination would theoretically be possible anywhere along shared homologous tracts of DNA, which would likely be present on the complementing *chuR* fosmid clone.

Following that line of thought, if the fosmid DNA had recombined into the genome for so many clones, could it be that most or even all of the fosmid clones were carrying DNA from *B. theta* strains (rather than other species) present in the pooled fecal samples? This scenario could explain the clones' propensity for homologous recombination. To see if this was the case, I designed PCR primers for the ORF of the *B. theta* *chuR* gene; these primers are likely to amplify only exact or very close matches to the *B. theta* VPI-5482 wild-type sequence (primers KL61 and KL62 were 35 and 40 bases in length, respectively; see Table 2.3). I carried out this PCR, using the pKL13 plasmid DNA and Δ*chuR* genomic DNA as negative controls (Figure 5.21B). As expected, all of the clones from the BT3 library were positive; and from the CLGM3 library, all but one clone (*chuR* clone #2) showed amplification using primers based on the *B. theta* *chuR* sequence. I tried reducing the annealing temperature of the PCR in an attempt to amplify the *chuR* ORF from CLGM3 clone #2, but a PCR product was not obtained even when using an annealing temperature as low as 45℃. This suggests that this clone may be carrying a copy of *chuR* that is quite different in sequence from *B. theta*; unfortunately, such sequences are the ones desired in a functional metagenomics approach and the problem of recombination prevented the retrieval of the clone's *chuR*-complementing sequence.

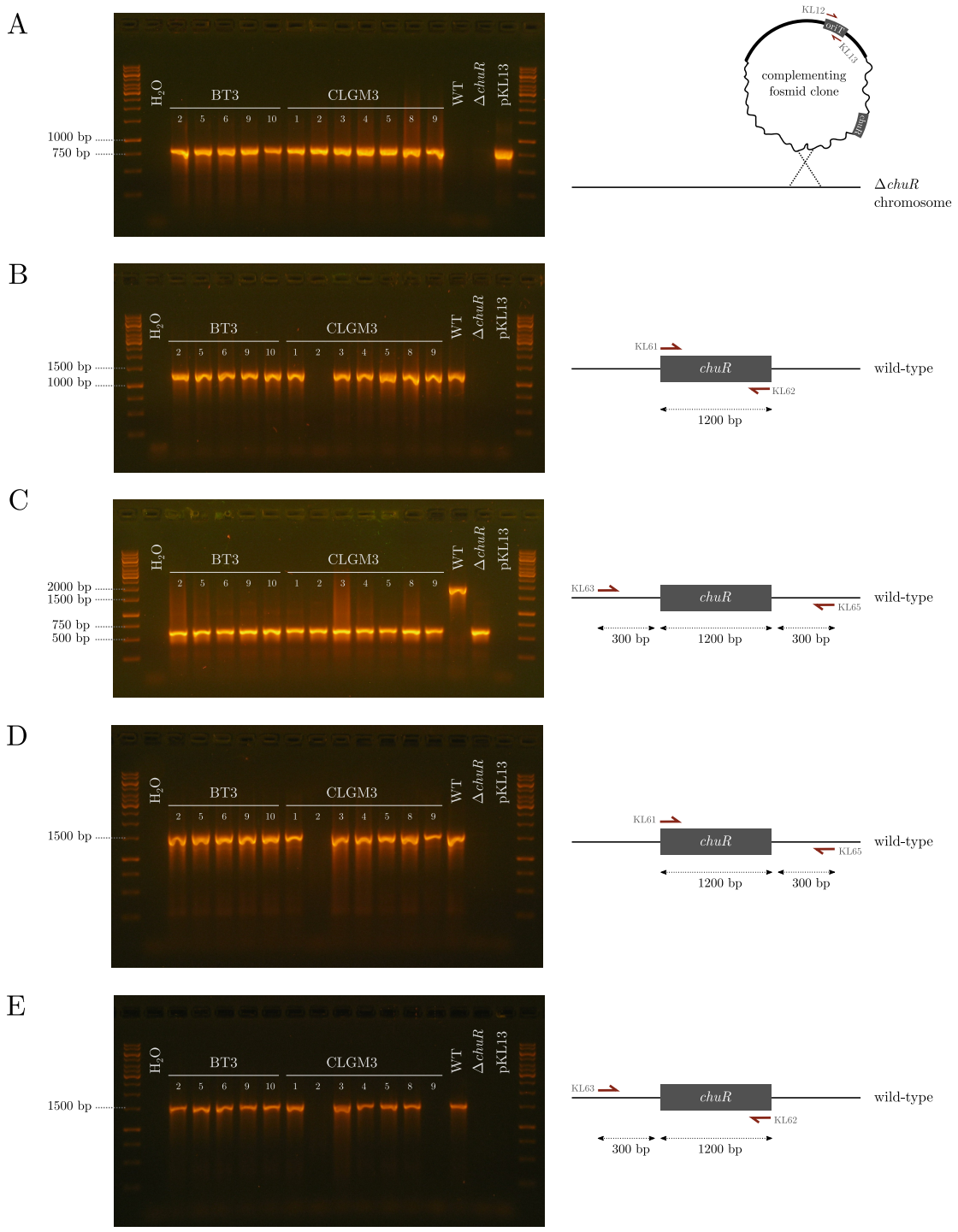


Figure 5.21: PCR analysis supporting the hypothesis that complementing fosmid DNA is integrated into the genome of *B. theta* Δ*chuR* host. PCR for: (A) the *oriT* sequence on the pKL13 vector backbone; (B) *chuR* ORF based on *B. theta* wild-type sequence; (C) fragment 300 bp upstream to 300 bp downstream of the *chuR* ORF; (D) *chuR* ORF plus −300 bp downstream; (E) *chuR* ORF plus 300 bp upstream.

From Figure 5.21B, it appeared that all but one clone from the CLGM3 library had a *chuR* gene exactly or very similar to the *B. theta* VPI-5482 wild-type, because PCR using *B. theta*-specific primers was successful. However, before I proceeded to analyze the sequences for these amplified ORFs, I first wanted to perform another check to support the hypothesis that the fosmid clones had integrated into the genome of the Δ*chuR* background. This deletion strain carries a clean removal of the 1,200-bp *chuR* ORF, and primers designed to +300 bp upstream and −300 bp downstream of the ORF would amplify only 600 bp from the mutant versus 1,800 bp from the wild-type. I used such primers to confirm that indeed the *chuR* 600-bp deletion fragment in the host genome was still intact for all BT3 and CLGM3 library clones (Figure 5.21C).

The result of this last PCR was somewhat surprising, however, for another reason. I had expected the BT3 library clones (and perhaps some of the CLGM3 clones as well) to exhibit both the 600-bp and 1800-bp bands – the prior from the *B. theta* Δ*chuR* background and the latter from the complementing fosmid DNA carrying the *B. theta* *chuR* gene. That all of the BT3 clones from Figure 5.21C were exhibiting just the 600-bp band suggested that the smaller product may be preferred in the PCR. To determine if this was the case, I used primer combinations such that the smaller PCR product was not a possibility: amplifying either the *chuR* ORF plus 300 bp downstream or amplifying the *chuR* ORF plus 300 bp upstream (Figure 5.21D and E, respectively). The results of this PCR confirmed that indeed the smaller PCR product was preferred and that the wild-type complementing DNA was present in the clones originating from the BT3 genomic library. Interestingly, 6 of the 7 clones from the CLGM3 human gut library also showed amplification (Figure 5.21D), supporting my hypothesis that these gut clones likely carried *B. theta* DNA – although CLGM3 clone #9 did not produce a PCR product in the amplification that included the 300-bp upstream of the ORF (Figure 5.21E), a result that suggests this particular complementing fosmid may

simply not be carrying a fragment that includes this 300-bp upstream region.

Consistent with a lack of amplification of the *chuR* ORF for CLGM3 clone #2 in Figure 5.21B, this clone did not produce PCR products in either Figure 5.21D or Figure 5.21E. For the 6 other clones isolated from the CLGM3 human gut library, however, the successful amplification of the *chuR* ORF (Figure 5.21B) meant that sequence analysis of the complementing ORF on the metagenomic DNA was possible.

5.4.5 Sequence analysis of positive clones isolated from complementation of *B. theta* reveals a *chuR* variant

Of the 6 metagenomic *chuR* ORFs that were amplified (Figure 5.21B), I suspected that all or most of them would be near or exact matches to the *B. theta* VPI-5482 *chuR* ORF. To analyze the sequence of these ORFs, the PCR products from CLGM3 *chuR* clones #1, 3, 4, 5, 8, and 9 were purified and submitted for Sanger sequencing. As a control, I also sequenced a PCR product originating from the *B. theta* genomic libary, BT3 *chuR* clone #2; this sequence should be the wild-type *B. theta* sequence, consistent with the source DNA used to make the BT3 library.

After Sanger sequencing, the single BT3 and 6 CLGM3 *chuR* sequences were aligned (Figure 5.22). All but one of the metagenomic *chuR* sequences were an exact match to the *B. theta* wild-type *chuR* sequence. To reiterate, this result was not surprising if homologous recombination occurred for all of these clones, suggesting that there was significant sequence similarity between the host genome and the DNA carried on the fosmid clones.

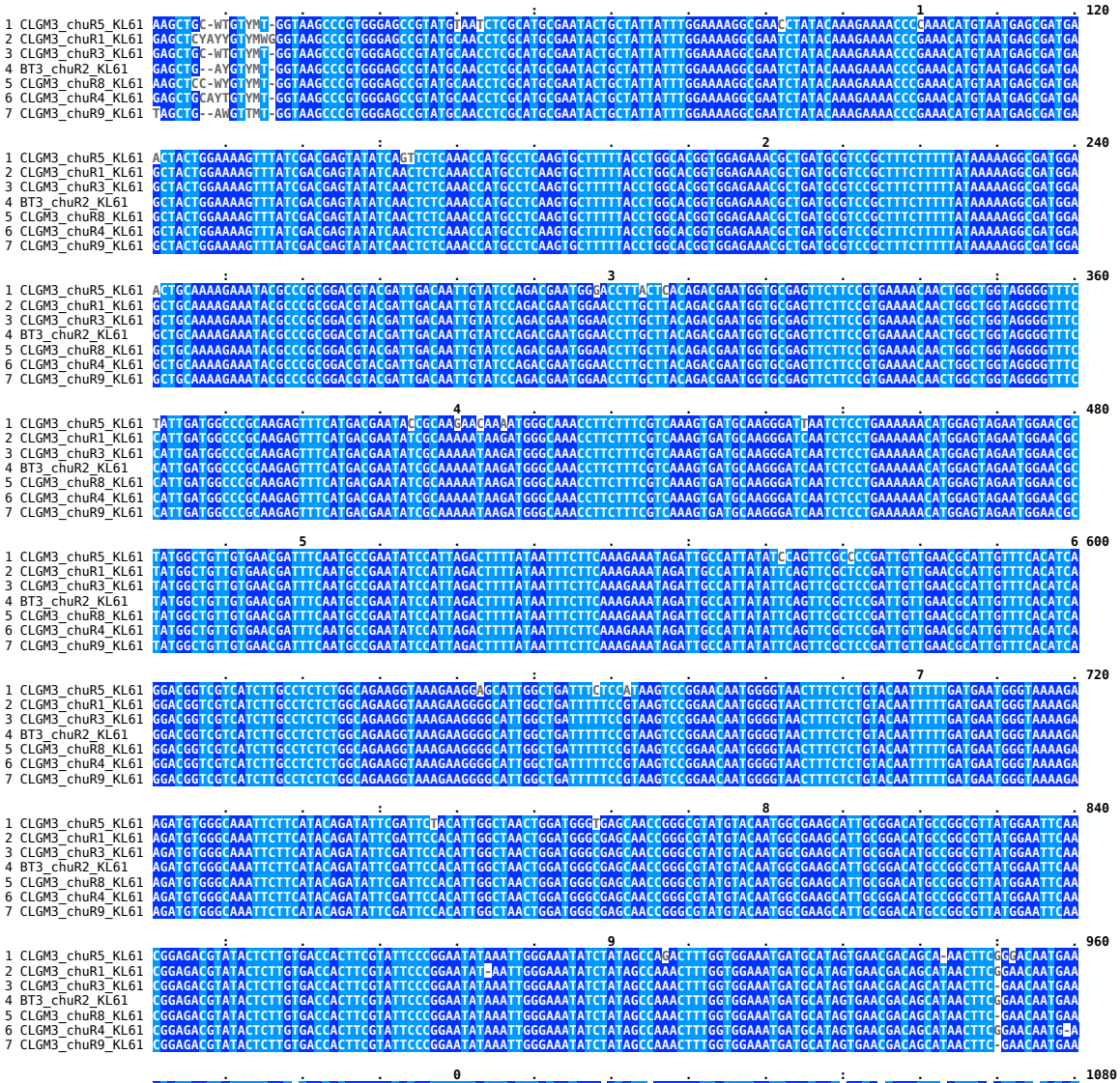


Figure 5.22: Sequence anlaysis of *chuR* ORFs PCR-amplified from positive clones isolated from BT3 and CLGM3 libraries. Alignment of sequences from the *chuR* ORFs from one clone from the BT3 library (BT3_chuR2) and six clones from the CLGM3 library (CLGM3_chuR1, chuR3, chuR4, chuR5, chuR8, and chuR9). Alignment generated using MUSCLE [70] and the alignment visualized using MView [28] on the EMBL-EBI web server [208], with colouring of purines/pyrimidines and mismatches.

Though it was not surprising that nearly all sequenced *chuR* ORFs from the integrated CLGM3 library clones were exact matches to wild-type *B. theta*, this outcome was interesting in a different light: it meant that nearly all positive clones isolated from the human gut metagenomic library in the *chuR/anSME* screen were of *B. theta* origin, albeit of "wild" *B. theta* from the feces of the volunteers who contributed to the library. Should we be surprised that nearly all *chuR* sequences recovered are from *B. theta*, rather than from other species? Perhaps no, considering that *Bacteroides* is the most common genus in human fecal samples [8] and that *B. theta* is often a dominating species in the distal gut [339]. To see if this ORF was present in public metagenomes, I performed a BLAST analysis, using the *B. theta chuR* sequence to query the NCBI database of assembled metagenomic contigs, and found exact or near identical full-length sequences in over a dozen assembled gut metagenomes (Table D.1 in Appendix D.3), suggesting that this particular *chuR* sequence may be relatively widespread, as would be expected for a gene from a common gut microbe. However, I was also interested in whether non-identical *chuR/anSME* genes have been annotated in metagenomes; a BLAST search using blastx against the NCBI env_nr database suggests that indeed there may be many proteins of varying sequence similarity that can potentially complement the Δ*chuR* mutant (Table D.2 in Appendix D.3)

From the alignment of the *chuR* sequences, one metagenomic *chuR* sequence was not identical to the *B. theta* wild-type – CLGM3 *chuR* clone #5 (Figure 5.22). The full ORF was obtained for this clone by Sanger sequencing (see Section 5.6.13 for primer and sequence data details). It shared ∼97% nucleotide identity with the wild-type using blastn, and its best hit in the NCBI nr database was *B. theta* VPI-5482 using megablast. Comparing its translated sequence to the *B. theta chuR* 415-residue protein sequence revealed three changes at the amino acid level: Asn62Ser, Val232Ile, and His325Gln (Figure 5.23).

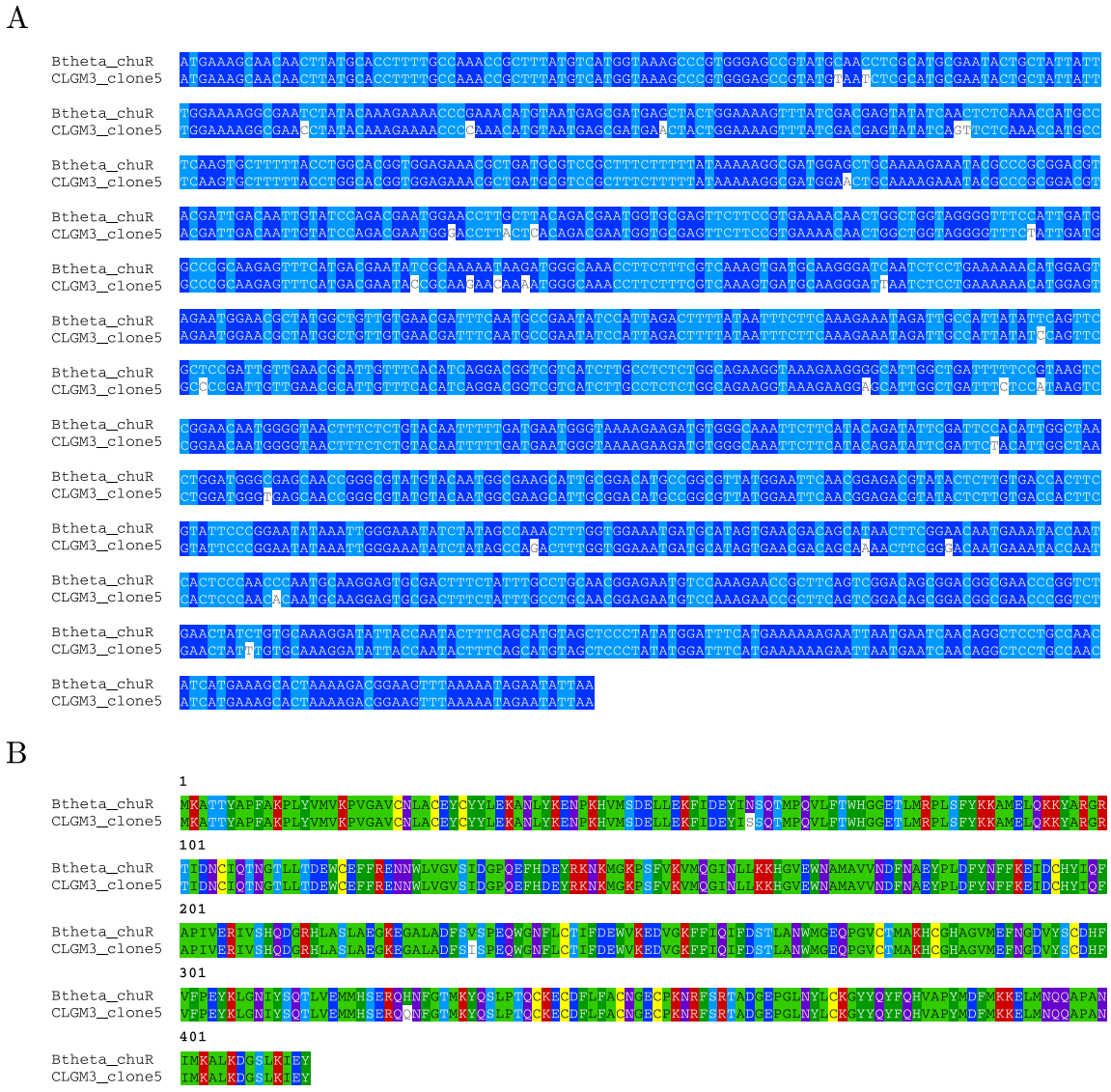


Figure 5.23: Alignment of the *chuR* sequence of CLGM3 *chuR* clone #5 to *B. theta* VPI-5482 *chuR* (BT_0238). Sanger sequencing reads were obtained from CLGM3 chuR clone #5 and the reads were assembled using Geneious version 6.0. The assembly was aligned to the wild-type sequence using MUSCLE [70] and the alignment visualized using MView [28] on the EMBL-EBI web server [208]. Alignments were generated for the ORF nucleotide sequence (A) and the translated ORF sequence (B). Residues differing from *B. theta* wild-type are indicated in white.

The three amino acid changes observed for this clone were in not in the three conserved cysteine clusters thought to be involved in the ability of the *chuR* enzyme to mature sulfatase enzymes [18].

The level of sequence similarity of this clone to wild-type *B. theta* suggests that this particular *chuR* gene carried by clone #5 may belong to an as-yet unsequenced species in the *Bacteroides* genus or perhaps another strain of *B. theta*, based on blastx results from querying the *B. theta chuR* sequence against the NCBI Refseq protein database (Table D.3 in Appendix D.3). The identification of a *chuR* gene from a human gut metagenomic library that is different in sequence from the *B. theta* VPI-5482 host is a clear indication that functional screening of metagenomic libraries using *B. theta* is a viable strategy.

5.4.6 Attempt to use arrayed libraries to track individual donor fosmids in complementation screens

The unanticipated problem of presumed homologous recombination in *B. theta* was an obstacle to screening using the lab's usual strategy, which requires retrieving the complementing fosmid from the transconjugant after the functional complementation screen. Using a *recA* mutant of *B. theta* as a host was one possibility that may have reduced the probability of recombination; however, a constructed *recA* mutant of *B. theta* was reported to have the unexpected phenotype of sensitivity to oxygen [49]. This increased sensitivity would make *B. theta* less versatile to work with in a laboratory setting and therefore the use of a *B. theta recA* mutant did not seem suitable.

Another solution to tackle the problem of unintended recombination was to modify the screening strategy so that I could track the fosmid clones being conjugated into the *B. theta* recipient. By tracking the clones in individual conjugations, any positive result

can be traced back to the specific *E. coli* clone used as donor, so that there is no need to retrieve DNA from *B. theta* at all, and the risk of not being able to retrieve the clone is obviated. Unfortunately, to track the clones for conjugations into *B. theta* required essentially "de-pooling" the fosmid libraries to obtain individual clones for tracking. To test this strategy with a subset of the libraries, I arrayed ∼600 clones from the BT3 genomic library and ∼1000 clones from the CLGM3 metagenomic library, making an arrayed collection of individual clone stocks in 96-well format (Figure 5.24).

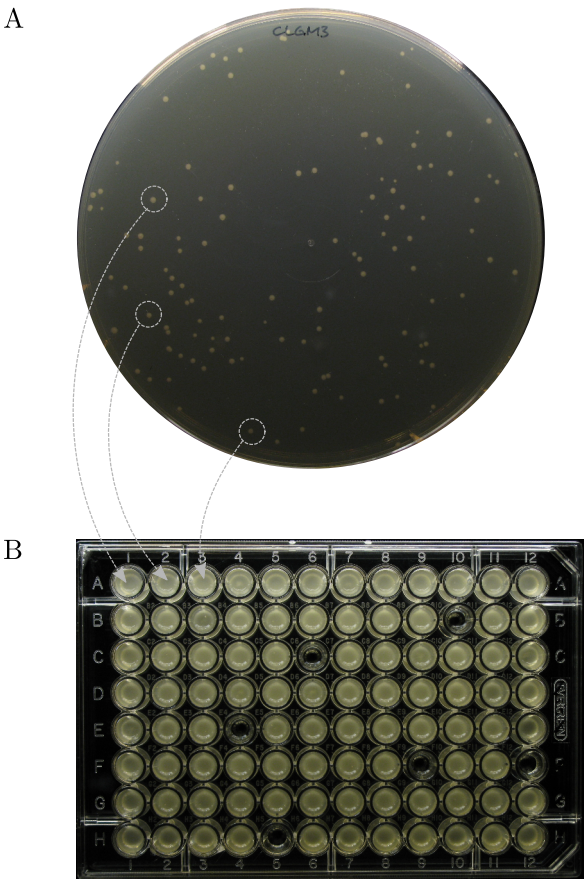


Figure 5.24: Arraying ∼1000 clones from the CLGM3 fosmid library (A) A frozen aliquot of the pooled CLGM3 library was diluted and plated for isolated colonies; (B) colonies were picked, inoculated, and saved in 96-well format. Six blank wells were included on each of the 12 plates as negative inoculation controls.

Though the libraries were arrayed to isolate individual clones, it was not feasible to carry out a separate mating for each clone; considering that the full CLGM3 metagenomic library contains ∼115 000 clones, this would not be a viable future strategy – without prior development of small-scale, high-throughput *E. coli*-*B. theta* conjugations and likely investing in and optimizing a robotic liquid handling system. Rather than carry out conjugations using single clones as the *E. coli* donor in matings, I instead used a pooled-clone mating system in which two rounds of conjugation were required, using a spot-conjugation method devised for moderately increased throughput (see Section 5.6.8 for description of two *E. coli*-*B. theta* conjugation methods used in this study):

Round 1: Pooled conjugations. In the first round, the 12 clones in each row of every plate of the arrayed collection were pooled (Figure 5.25A) and the pool was used as the donor in a mating with the *B. theta* Δ*chuR* recipient (Figure 5.25B). The conjugation spot was resuspended, washed, and streaked out on selective media to isolate complemented transconjugants – that is, those *B. theta* recipient cells that received a library fosmid carrying a gene that could provide the missing *chuR* function (Figure 5.25C).

Round 2: Resolution conjugations. Any positive clone arising from the first round was double checked by streak purification on the same selective media (Figure 5.25D). Then, to resolve which clone in that particular pool was responsible for the complementation, a second round of conjugation was carried out using individual clones as donor. Though it is possible that more than one of the 12 clones led to the positive result, the likely scenario is that just one of the clones was responsible for the complementation.

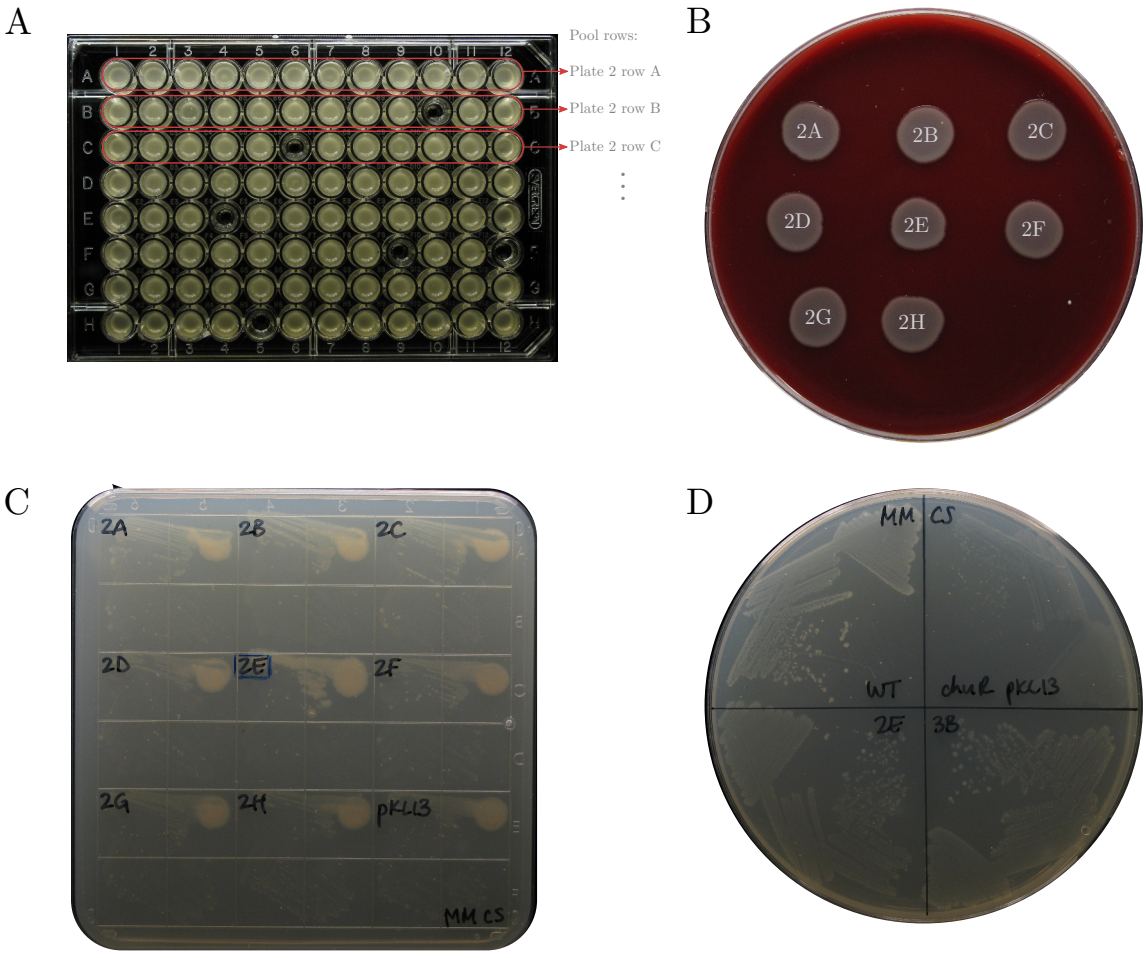


Figure 5.25: Functional complementation of *B. theta* Δ*chuR* using pooled *E. coli* donors from arrayed CLGM3 library. (A) Clones from each row were pooled for every row of each of the 12 96-well plates; rows were tracked by plate and row, e.g., the clone pool from Plate 2 Row A was labeled Pool 2A. (B) Pools from each plate were mated into the *B. theta* Δ*chuR* deletion using the spot conjugation method. (C) Spots were resuspended, washed, and streaked on minimal media with chondroitin sulfate as sole carbon source; positive pools were identified, e.g. Pool 2E. (D) Putative-positive complemented transconjugants were re-streaked on the same media for confirmation of phenotype.

This strategy was applied to screen the arrayed CLGM3 metagenomic library for *chuR*-complementing clones and in the pooled-conjugation round, a number of rows from various 96-well plates were identified as having a positive clone(s). However, the spot-mating strategy requires optimization because it is difficult to select the complemented transconjugants from the heavy background of *E. coli*; put another way, the mating spot contains high background making it difficult to both obtain and gauge a positive (Figure 5.25C). Though the natural inclination may be to perform the matings anaerobically to favour the recipient growth, conjugations using IncP systems have been documented to require oxygen for high-frequency transfer and may not work well anaerobically [249].

With putative positives from the pooled conjugations, I then performed resolution-round conjugations to identify single clones in the pool that were responsible for the complemented phenotype. Due to the described difficulties in this strategy and time constraints, I was only able to identify two putative positive clones that restored the ability to use chondroitin sulfate to the *B. theta* Δ*chuR* recipient: from the 5B pool that gave a positive in the first round (pooled clones from Plate 5, Row B), clone #5B2 was identified as the putative clone responsible for the complementation (Well 2). Interestingly, clone #5B9 was also identified as having an intermediate phenotype, between that of the wild-type and the deletion mutant; I streak-purified both of the 5B2 and 5B9 clones to confirm their phenotype on minimal media with chondroitin sulfate as sole carbon source (Figure 5.26).

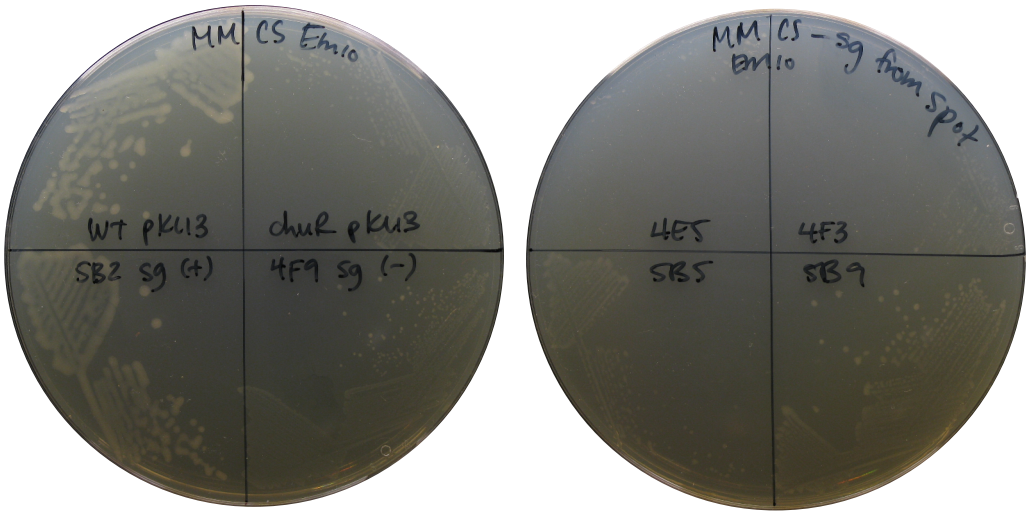


Figure 5.26: Streak purification of *B. theta chuR* carrying CLGM3 fosmid clone 5B2 or 5B9, for confirmation of phenotype. *B. theta chuR* carrying CLGM3 fosmid 5B2 (left plate, bottom-left quadrant) exhibits functional complementation when compared to the positive control wild-type (left plate, top-left quadrant) and negative control vector-only streak (left plate, top-right quadrant). *B. theta chuR* carrying CLGM3 fosmid 5B9 shows an intermediate phenotype between wild-type and mutant (right plate, bottom-right quadrant).

After identifying the specific wells of the arrayed collection with the putative clone carrying a *chuR*-complementing gene (Plate 5, Row B, Wells 2 and 9), I was then able to go back to the collection and examine the DNA from *E. coli* that had never been passaged through *B. theta*. Diagnostic digests of these clones showed a high-molecular-weight insert for both clones, although interestingly, copy number induction of these clones led to loss of the insert (Figure D.4 in Appendix D.1). BLAST analysis, using the megablast algorithm, detected no sequence similarity between the *ilvGEDA* and *rnpB* terminators, so it is unclear how the insert could have recombined out.

With only two complementing clones identified in the resolution mating round, this two-step strategy to screen the arrayed collection will have to be repeated to identify more putative individual complementing clones. Further analysis of the two complementing clones is also required, to determine the origin of insert DNA carried

by the clones and whether the DNA encodes a *chuR* ORF that is novel in sequence. Though this two-step method appears to be a viable strategy for screening human gut metagenomic libraries in a *B. theta* host, optimization of the method will be required to reduce *E. coli* background, raise the frequency of obtaining transconjugants, and increase throughput for *E. coli*-*B. theta* conjugations as well as selection for phenotypic complementation.

5.5 Conclusions

B. theta is becoming more widely used in both pure and applied research. Its important role in degrading polysaccharides in the host gut and its dominance in the microbiota community make it an ideal candidate for study and manipulation. In this Chapter, *B. theta* was chosen to be developed as a host to screen gut-derived metagenomic DNA because it would likely be able to express a greater fraction of the cloned DNA than would *E. coli*. Unexpectedly, the complementation of a *B. theta chuR* mutant suggested that *B. theta* is prone to homologous recombination, which presents difficulties for the screening of pooled metagenomic libraries. Screening of arrayed clone libraries is possible and is presented here, but the strategy is labour-intensive and likely requires a semi-automated high-throughput approach; with such an approach, the conditions for *E. coli*-*B. theta* conjugations will also require optimization. Though difficulties were encountered in using a *B. theta* host to screen a human gut library, the identification of a *chuR* gene different in sequence from the *B. theta* wild-type demonstrates that *B. theta* shows some promise as screening host.

5.6 Specific materials and methods

5.6.1 Strains and plasmids

The *E. coli* and *B. theta* strains and plasmids used were described in Chapter 2, specifically Table 2.1 for strains and Table 2.2 for plasmids.

5.6.2 Growth media and anaerobic culture

Methods for the culture of *B. theta* were based on those generously shared by Nicole Koropatkin and Eric C. Martens of the University of Michigan.

Culture in liquid media

B. theta was routinely cultured in liquid broth using brain heart infusion broth (BD Biosciences B237200), supplemented with 1.2 µM histidine, 1.9 µM hematin, 1 µg/ml menadione, and 0.5 µg/ml cysteine. I called this media BHI+; see Appendix A.4 for the recipe. Before discovering that *B. theta* grows very well in BHI+, I also used TYG for liquid culture; see Appendix A.3 for the recipe.

Cultures of *B. theta* were started by inoculation either from a single colony or from frozen stock, using the pyrogallol method [128]: after inoculation, two cotton balls were inserted into the mouth of the culture tube using sterile forceps, with the second cotton ball not fully inserted. The cotton was lit using the flame of a Bunsen burner to purge the culture tube of oxygen; after the flame extinguished, the cotton ball was pushed about an inch further into the culture tube, and overtop of the cotton ball was added 200 µl of 20% $NaCO_3$ (w/v) and 200 µl of 35% pyrogallol (w/v), and the tube was then immediately plugged with a rubber stopper. Pyrogallol is activated in the

presence of alkalinity to react with oxygen creating a reducing environment.

Cultures of *B. theta* were incubated at 37℃ without shaking. Typically, resazurin was added to the liquid media as an indicator dye (1 µg/ml): it is blue in an oxidizing environment, turns irreversibly pink in a reducing environment, and reversibly colourless in the absence of oxygen (Figure 5.5).

Culture on solid media

B. theta was routinely cultured on agar using brain heart infusion broth (BD Biosciences B237200), supplemented with 10% defibrinated horse blood (Bio-media Unlimited MOHD500); see Appendix A.4 for the recipe. *B. theta* was also cultured on solid minimal media; see Appendix A.5 for the recipe.

Agar plates were incubated in air-tight jars with GasPak EZ Anaerobe sachets (BD Biosciences B260678) to deplete oxygen. Originally, the air-tight container used was the GasPak 100 System 13 × 23 cm polycarbonate jar; however, inexpensive air-tight containers purchased from local stores demonstrated comparable results, including Anchor Hocking stainless steel canisters and Lock & Lock glass containers (Figure 5.6). Lubricating grease was applied to the gaskets of air-tight containers to ensure a good seal.

5.6.3 Antibiotics

Antiobiotics used in the culture of *B. theta* are summarized in Table 5.6. Concentrations for antibiotics are denoted using the abbreviation (see Table 5.6) followed by the concentration as a subscript; for example erythromycin at 10 µg/ml would be Em_{10}. Note that antibiotic concentrations were halved when used in liquid media.

Table 5.6: Antibiotic concentrations used for *B. theta*

Antibiotic	**Abbrev.**	**Solvent**	**Final conc.**
erythromycin	Em	ethanol	10-25 µg/ml
gentamicin	Gm	dH_2O	200 µg/ml
kanamycin	Km	dH_2O	200 µg/ml
nalidixic acid	NA	dH_2O	25 µg/ml
tetracycline	Tc	ethanol	2 µg/ml

5.6.4 Preparation of DNA polylinker/MCS from complementary oligos

The following protocol was used to phosphorylate and anneal oligos KL10 and KL11 to form a polylinker. See Table 2.3 for DNA sequences. The protocol for annealing complementary oligos is based on the protocol from OpenWetWare (http://openwetware.org/wiki/Endy:Annealing_complementary_primers).

Phosphorylation of oligos

Oligos KL10 (30 bases) and KL11 (22 bases) were each diluted to 100 pmol/µł and 40 µl of each were used in separate phosphorylation reactions, using T4 polynucleotide kinase (Thermo-Fisher K0031) according to the recipe in Table 5.7. This volume corresponded to 36 µg and 27 µg for KL10 and KL11, respectively. The reactions were incubated at 37℃ for 1.5 hours, followed by heat inactivation at 80-85℃ for 20 minutes and cooling on ice.

Table 5.7: Recipe for phosphorylating oligos.

oligo DNA (100 pmol/µł)	40 µl
10× T4 DNA Ligase Buffer	5 µl
T4 PNK (10 units; in excess)	1 µl
sterile dH_2O	4 µl
Total	50 µl (80 pmol/µł)

Annealing complementary oligos

Phosphorylated KL10 and KL11 were combined in an annealing reaction mix (Table 5.8). The tube was placed in a floating rack and incubated in a beaker of boiling water for 5 minutes. The beaker was then removed from the heat and allowed to cool to room temperature slowly over ∼20 minutes, with later cooling sped up by placing the beaker on ice. As a check, 0.5 µl of the annealed KL10/KL11 reaction was run on a 2% agarose gel, against 0.5 µl and 1 µl of the phosphorylated KL10 and KL11 as controls (Figure D.1 in Appendix D.1). The generated polylinker was stored at -20℃ until ready to be used for ligating to the vector, EcoRI- and KpnI-digested pKL1.

Table 5.8: Recipe for annealing complementary oligos.

phosphorylated KL10	20 µl (14.6 µg)
phosphorylated KL11	20 µl (10.8 µg)
0.85% NaCl	10 µl (14 mM final)
Total	50 µl (508 ng/µł)

5.6.5 PCR of *ermF-repA* and *oriT*

The *oriT* fragment was amplified from pJC8 (10 ng) using primers KL12/KL13 (possessing HindIII adapters) and the *ermF-repA* fragment was amplified from pAFD1 (10 ng) using primers KL14/KL15 (possessing EcoRI adapters). KOD Hot Start DNA Polymerase (Novagen 71086) was used according to the manufacturer's recommendations. The touchdown PCR protocol used for both fragments is summarized in Table 5.9. To prepare for cloning, the PCR products were gel extracted, digested with the appropriate restriction enzyme, and column-purified, using routine protocols previously described in Chapter 2.

Table 5.9: Touchdown PCR protocol for *ermF-repA* and *oriT*.

Temperature	Duration	
94℃	2 min	
98℃	10 sec	×6 cycles; ↓1℃/cycle
65 → 59℃	30 sec	
68℃	1 min/kb; round up nearest min	
98℃	10 sec	×25 cycles
58℃	30 sec	
68℃	1 min/kb; round up to nearest min	
68℃	5 min	
20℃	hold	

5.6.6 Primer walking to sequence the *ermF-repA* fragment

The *ermF-repA* fragment from pAFD1 was sequenced in order to compile the complete sequence for the constructed vector pKL13. The ∼4-kb fragment was sequenced by primer walking using oligos KL14, KL16, KL33, KL42, KL43, KL45, and KL46 (Table 2.3). Multiple templates were sequenced from which the consensus was taken, using different combinations of the following for each round of primer walking: pAFD1, pKL6, pKL7, and pKL8 (Table 2.2). The consensus sequence for the *ermF-repA* fragment is included in Appendix D.2. See Section 5.6.13 for information on sequence data availability.

5.6.7 Miniprep of plasmid DNA from *B. theta*

Plasmid DNA was isolated from liquid *B. theta* cultures using the QIAprep spin miniprep kit (Qiagen 27106), according to the manufacturer's recommendations, including optional washes to reduce nuclease contamination. Typically, 5 ml of culture was used for plasmid minipreps.

5.6.8 Conjugation from *E. coli* donor to *B. theta* recipient

Lawn conjugations

The following protocol was based on one shared with me by Nicole Koropatkin and Eric Martens from the University of Michigan. Matings were carried out using 5 ml of each of the donor, mobilizer, and recipient strains.

The *E. coli* donor and mobilizer were cultured in 5-ml LB supplemented with the appropriate antibiotics and grown to OD_{600} of ∼0.4; *B. theta* recipient cultures were cultured in 5 ml BHI+ and grown to OD_{600} of ∼0.3-0.4 (Spectronic Spec 20D spectrophotometer). Cultures were placed on ice to halt cell growth. Cultures were transferred to 15-ml conical tubes and cells were pelleted by centrifuging at 7,000×g at room temperature for 5 minutes. The supernatant was removed and the cells were resuspended in either BHI+ or 1× Bt salts (see Appendix A.5). Donor, mobilizer, and recipient were mixed in a final volume of 1 ml, and the mixture was swirled evenly over the surface of a BHIH agar plate. The plate was dried for several minutes in a laminar flow hood and then incubated aerobically overnight with the agar side down.

Overnight mating lawns were scraped off the agar plate with a wooden stick and resuspended in 2 ml BHI+ or 1× Bt salts. Typically, serial ten-fold dilutions were made from 10^{-1} to 10^{-3}, and 100 µl of each dilution was plated on the BHIH supplemented with appropriate antibiotics to select for transconjugants – typically, Km_{200} and NA_{25} to select against *E. coli* and $Em_{10\text{-}25}$ to select for the vector. If the mating lawn was plated on minimal media, then the initial resuspension of the mating lawn was washed to remove complex media components; this was accomplished by at least three repititions of centifugation and resuspension in 1 ml 1× Bt salts.

Spot conjugations for increased throughput

I modified the preceding conjugation protocol shared by Nicole Koropatkin and Eric Martens to achieve a higher throughput for mating library clones into *B. theta*. Using spots rather than lawns, up to 10 or 12 matings can be performed per agar plate (Figure 5.25B). Matings were carried out using 5 ml-equivalents of each of the donor and mobilizer spotted onto 10-15 ml-equivalent of spread-plated recipient.

Cultures of *E. coli* and *B. theta* were grown as for the lawn conjugations. 10-15 ml of the *B. theta* recipient culture was centrifuged; the supernatant was removed and the cells were resuspended in 100 µl 1× Bt salts (see Appendix A.5) and spread on a BHIH plate, and the plate was dried for several minutes in the laminar flow hood. 5 ml of each of the *E. coli* donor and mobilizer were centrifuged; the supernatant was removed from both, the mobilizer cell pellet was resuspended in 20 µl 1× Bt salts, the resuspension was transferred to the donor cell pellet for resuspension, and then the mixture was spotted onto the plate overlaying the *B. theta* cells (Figure 5.25B). The mating spots were dried for several minutes in the laminar flow hood and then incubated aerobically overnight with the agar side down.

Overnight spot matings were processed exactly as lawn matings, with the only difference being that the volume used for resuspension was smaller: 500 µl 1× Bt salts or BHI+ was used instead of 2 ml.

5.6.9 Genomic and metagenomic library construction

The libraries constructed in this chapter using either pKL3 or pKL13 are summarized in Table 2.12. Libraries were constructed as described previously in Section 3.6.3, with some exceptions that are detailed below.

pKL3-based libraries

The CLGM2 metagenomic library and BT2 genomic library (see Section 2.6) were both constructed using pKL3. Library construction was carried out as previously described in Section 3.6.3, with the minor exception that transductants were selected on ampicillin instead of tetracycline. This was due to the resistance marker present on the base vector, pAFD1, which was used to construct pKL3 (Figure 5.8).

pKL13-based libraries

The CLGM3 metagenomic library and BT3 genomic library (see Section 2.6) were both constructed using pKL13. Library construction was carried out as previously described in Section 3.6.3, with a few exceptions. First, transductants were selected on chloramphenicol instead of tetracycline; this was due to the resistance marker present on the base vector, pCC1FOS, which was used to construct pKL13 (Figure 5.12). Second, EPI300 was used for the library host instead of HB101, due to its advantageous copy number control feature when used in conjunction with pCC1FOS.

The third and last exception to library construction is a highly unusual and therefore notable one: the pKL13 vector backbone was not purified away from the stuffer between the Eco72I sites; that is, the vector was simply digested to release the stuffer, and the mixture was used for ligation to high-molecular weight metagenomic or ge-

nomic DNA. This means that some subset of clones in the CLGM3 and BT3 libraries may be "contaminated" with the pKL13 stuffer. The reason for not removing the stuffer was that I had technical difficulties doing so: after purifying the digested and dephosphorylated backbone by electroelution and achieving a concentrated preparation of ∼350 ng/µl, I was no longer able to ligate the vector, which I discovered in carrying out calculations for digestion and dephosphorylation efficiency using T4 PNK and ligase (as described in Section 2.5.6). To ensure that the preparation had not been contaminated with nucleases, I ran the purified DNA on an agarose gel and saw that it was indeed intact (Figure D.2 in Appendix D.1). It is still unclear why the vector was no longer ligatable, but it is possible that after digestion and dephosporylation, the vector ends may be sensitive to disruption when subjected to an electric field. In any case, I was forced to make a preparation of the vector without stuffer purification to use in library construction. After constructing the CLGM3 and BT3 libraries, I estimated using the CLGM3 library that the percent of gentamicin-resistant clones is 1-2%, which provides an estimate of the upper limit for stuffer contamination (the stuffer carries a gentamicin resistance gene; see Figure 5.12).

5.6.10 Construction of *thrC* and *trpD* single recombinants

The ∼600-bp *thrC* (BT_2401) and ∼350-bp *trpD* (BT_0530) fragments were amplified from *B. theta* genomic DNA (55 ng) using primers thrCIDMF-SalI/thrCIDMR-KpnI and trpDIDMF-SalI/trpDIDMR-KpnI, respectively (see Table 2.3), with restriction enzyme adapters as indicated by the primer names. Pfx DNA Polymerase (Invitrogen 11708-013) was used according to the manufacturer's recommendations. The PCR protocol used for both fragments is summarized in Table 5.10.

Table 5.10: PCR protocol for *thrC* and *trpD* fragments.

Temperature	Duration	
94℃	5 min	
94℃	15 sec	×30 cycles
58℃	30 sec	
68℃	60 sec	
68℃	5 min	
10℃	hold	

To prepare for cloning, the PCR products were purified using a QIAquick PCR Purification Kit (Qiagen 28104) and digested using NEB enzymes according to the manufacturer's recommendations in a sequential double digest: purified PCR products were digested with SalI in NEB Buffer 3 (NEB R0138), the sample was ethanol precipitated, and the DNA was resuspended in NEB Buffer 1 for digest with KpnI (NEB R0142). The digested fragments were purified by gel extraction using a QIAquick Gel Extraction Kit (Qiagen 28704), and ligated to similarly cut and purified pKNOCK-*bla*-

tetQb. Ligations were microdialyzed against water using DNA filter paper (Millipore VCWP09025), and then used to electroporate S17-1 λ-*pir*. Clones were streak-purified, then screened and verified by restriction digest. Clones of pKNOCK-*bla*-*tetQb*carrying the *thrC* and *trpD* fragment were named pKL21 and pKL22, respectively.

pKL21 and pKL22 were conjugated from S17-1 λ-pir into wild-type *B. theta* in a biparental mating using the lawn conjugation method (Section 5.6.8). Mating lawns were resuspended and diluted, and transconjugants carrying the integrated plasmid were selected on BHIH Gm_{200} Tc_2. Transconjugants were streak-purified and inoculated into 5 ml TYG Tc_2 for generation of frozen stocks; the *B. theta* BtUW1 and BtUW2 strains were added to the Charles lab strain collection (see Table 2.1). The phenotype of the strains were also checked on minimal media with and without the appropriate amino acid supplementation (Figure 5.15B).

5.6.11 Genomic DNA miniprep of *B. theta*

This protocol is a scaled-down version of the one described in Section 2.4.6, which is based on the method described by Charles and Nester [36]. Briefly, *B. theta* was cultured in 10 ml of liquid media with the appropriate antibiotics, and the cell pellets were recovered after centrifugation at 7000×g for 5 minutes at room temperature. Cells were resuspended in 400 µl buffer (10 mM Tris [pH 8.0], 25 mM EDTA). The following were added: 50 µl 5 M NaCl, 10 µl 10 mg/ml RNase A, 5 µl 19.2 mg/ml proteinase K (optional), and the tube was inverted several times. 25 µl 20% SDS was added and the sample was incubated at 65℃ for 30-60 minutes. 260 µl 7.5 M ammonium acetate was added and the sample was incubated on ice for 20 minutes. The mixture was centrifuged at 21,000×g for 15 minutes, the supernatant was decanted carefully, and the mixture was extracted with chloroform in a 1:1 volume. The DNA was precipitated with 800 µl isopropanol, and pelleted by centrifuging at 21,000×g for 3 minutes. The pellet was washed with 100 µl 70% ethanol, centrifuged at 21,000×g for 1 minute, the supernatant was removed, and the pellet was allowed to dry. Finally, the pellet was allowed to dissolve in 50 µl of TE overnight at 4℃. The DNA was quantified by gel electrophoresis, using bacteriophage λ DNA as a standard (see Section 2.5.8).

5.6.12 Analysis of genomic DNA for fosmid clone recombination using PCR

Genomic DNA was isolated from the *B. theta* clones carrying *chuR*-complementing fosmid DNA, and used as template in the PCR. *Taq*-based 2X PCR Master Mix (Thermo Scientific K0171) was used according to the manufacturer's recommendations, with the exception that RNAseA was typically added to the reaction in small amounts to remove RNA contamination from the genomic DNA prep. The general touchdown PCR protocol used is summarized in Table 5.11. Target PCR products and their corresponding primer sets were (see Table 2.3 for primer details):

- RK2 *oriT*(∼800 bp): KL12, KL13
- *chuR* ORF (∼1200 bp): KL61, KL62
- *chuR* ORF + 300-bp upstream and downstream (∼1800 bp): KL63, KL65
- *chuR* ORF + 300-bp downstream (∼1500 bp): KL61, KL65
- *chuR* ORF + 300-bp upstream (∼1500 bp): KL63, KL62

Table 5.11: Touchdown PCR protocol for analysis of genomic DNA.

Temperature	Duration	
95℃	3 min	
95℃	30 sec	×11 cycles; ↓1℃/cycle
60 → 50℃	30 sec	
72℃	1 min/kb	
95℃	30 sec	×20 cycles
50℃	30 sec	
72℃	1 min/kb	
72℃	5 min	
20℃	hold	

5.6.13 Data availability

The expected sequence for the *B. theta*-compatible pKL13 fosmid is provided in Appendix D.2 and has been submitted to NCBI Genbank (NCBI accession KU746975). The sequence of the *ermF-repA* fragment from pAFD1 is provided in Appendix D.2. Sanger sequencing reads for *chuR* clone #5 are provided in Appendix D.2. Additionally, raw sequencing data in ABI (.ab1) format can be accessed online: `https://github.com/itskathylam/phd`

Chapter 6

Inclusion of transcriptional terminators in cloning vectors

6.1 Acknowledgements and declarations

Part of the introduction of this chapter was published as part of a Perspective article in the journal **Frontiers in Microbiology**. I was the primary author of this article. The citation for the article is:

> Lam KN, Cheng J, Engel K, Neufeld JD, Charles TC (2015) Current and future resources for functional metagenomics. *Frontiers in Microbiology* 6:1196. doi:10.3389/fmicb.2015.01196

I performed all experiments and analyses described in this chapter and I acknowledge the following contributions:

- The introduction of the *Frontiers in Microbiology* manuscript, largely duplicated here, was proofread and edited by **Katja Engel**, **Josh Neufeld**, **Trevor Charles**, and **Jiujun Cheng**.
- This remainder of this chapter was proofread by my supervisor **Trevor Charles**.

6.2 Abstract

Functional metagenomics is a powerful experimental approach for studying gene function, starting from the extracted DNA of mixed microbial populations. A functional approach relies on the construction and screening of metagenomic libraries – physical libraries that contain DNA cloned from environmental metagenomes. Library construction is often a technically challenging and laborious endeavour, thus necessitating the careful design of library cloning vectors to ensure the presence of elements that aid in the library's downstream applications.

The commercial fosmid vector pCC1FOS is widely used for the construction of metagenomic libraries. As I described in Chapter 5, I used pCC1FOS as the base plasmid to construct the *B. theta*-compatible library vector pKL13, introducing various additional elements, including two transcriptional terminators that flank the cloning site, which were anticipated to reduce insert-borne transcription into the vector backbone should such transcription be problematic for clone stability. The two terminators are taken from the *ilvGEDA* and *rnpB* genes of *E. coli* MG1655, which were documented to be strong terminators. Here, I provide the rationale for the design of the transcriptional terminator (TT) fragment encoding the terminators, describe its synthesis and cloning, and most importantly, present the results of testing the functionality of the two terminators using the fluorescent reporter GFPuv. With the use of a simple testing scheme, both terminators appear to be reducing transcription in vivo, justifying their inclusion in the pKL13 fosmid.

Finally, in the last results section of this chapter, I discuss how the TT fragment may be taken advantage of in future experiments to test whether the transcriptional terminators help protect against or alleviate the observed cloning bias of metagenomic libraries. Several constructs have been built for this purpose and though such experiments are outside the scope of this work, it will be important for the functional metagenomics approach to understand the factors that affect DNA representation in clone libraries.

6.3 Introduction

6.3.1 The challenges of constructing large-insert metagenomic libraries

The functional metagenomic approach and the steps involved in constructing libraries using a *cos*-based vector were previously described (Section 1.6.1 and Figure 1.1). With the number of steps involved, the construction of a metagenomic library can be a laborious and time-consuming procedure, requiring a high level of skill at the laboratory bench. There are several technically challenging steps in the process of metagenomic library construction. First, the DNA extracted from the environmental sample must be of sufficient length for efficient packaging into lambda phage heads, which have a lower size limit for packaging [229]. Extraction usually employs gentle lysis to avoid shearing the DNA [347] but even so it may be difficult to achieve large fragment sizes [141]. I find that starting with crude DNA extracts containing at least ∼75 kb fragments leads to high-quality *cos*-based libraries, and it is crucial to check the fragment size range of crude extracts by pulsed-field electrophoresis before proceeding. In my experience, a particularly useful and affordable molecular ladder to use for pulsed-field gels is self-ligated lambda DNA, which can be easily prepared in-house and results in bands at ∼50, ∼100, and ∼150 kb. A freeze-grinding step prior to extraction [175] can substantially improve cell lysis. Although this additional step might also fragment DNA [26], I find that it does not hinder library construction, consistent with previous work showing that freeze-grinding results in minimal shearing [347].

Extracts are often contaminated with compounds that co-purify with DNA, requiring additional purification steps that may lead to loss of DNA. Common contaminants in soil-derived DNA extracts are humic acids, visible as a brown coloring of the extract. Such contaminants may interfere with enzymatic reactions [303]. Non-linear electrophoresis is effective for contaminant removal [232] and generates purified and highly concentrated DNA suitable for PCR or metagenomic analysis [75], yet requires access to specialized equipment. I have found that for library construction, humic acids can simply be allowed to run off the gel during pulsed-field electrophoresis of crude extract for size-selection because humic acids travel much faster than large DNA fragments when subjected to an electric field. Alternatively, to avoid contaminating the circulating buffer, electrophoresis can be paused after contaminants have formed a front, the part of the gel containing humic acids excised, and then this region replaced with fresh gel [43].

After the DNA has been size-selected and electroeluted from a pulsed-field gel, it must be end-repaired and then ligated to a desphosphorylated and blunt-ended vector. To ensure a proper size range of DNA (∼25 to 40 kb) before ligation, the DNA can be checked for co-migration with the largest band of a lambda-HindIII ladder on a typical agarose gel [26] or, as I prefer, running the sample on a pulsed-field gel for a more accurate size assessment. The end-repair is a particularly challenging step in library construction because there is no simple way to confirm that ends are indeed blunt following the end repair step. My current strategy is to use a small amount of the ligation to transform *E. coli* prior to the costly packaging step; resulting transformants indicate the presence of circular DNA molecules arising from ligation of successfully blunt-ended fragments. Though the ligation conditions may not favour the formation of circular molecules, this is currently the best proxy for successful end-repair.

Other challenges include the sensitivity of packaging extracts as well as the preparation of purified digested dephosphorylated vector DNA for ligation. Although excellent commercial products are available for both reagents, in-house vector preparation may still be required when specific expression hosts are to be used in functional screening that are outside the host range of available commercial vectors [43,50,308,330]. The culminating step of library construction is the transduction of *E. coli*. Although it is possible to generate many thousands of clones with the first attempt, troubleshooting may be required to increase library size in some cases. When the transduction results in a disappointingly small number of transductants (zero in the worst case!), it is not easy to determine the cause.

Indeed, metagenomic library construction is in many ways a craft that takes time and practice to master. Given that there are substantial challenges and costs associated with library construction, as well as possible difficulties in obtaining rare environmental samples, a clear corollary is that researchers active in this field ought to find ways to maximize these valuable resources for shared benefit. In particular, collections of metagenomic libraries that can be used in a variety of hosts would be extremely valuable if able to be accessed by the wider scientific community. Our lab and collaborators have previously made metagenomic libraries publicly available [222] and continue to advocate for increased sharing and strategizing [37]. Though there are obvious administrative obstacles, services such as Addgene [125] may facilitate these efforts.

6.3.2 Properties of pCC1FOS, a popular vector for library construction

Due to the difficulties of library construction, commercial products that aid in generation of cosmid or fosmid libraries are popular. Indeed, one widely used cloning-ready commercial vector is pCC1FOS (Genbank accession EU140751; available from Epicentre Biotechnologies), shown in Figure 6.1. In recent years, as functional metagenomics has gained traction, a number of metagenomic libraries from remarkably diverse environments have been constructed using pCC1FOS, some of which are listed in Table 6.1.

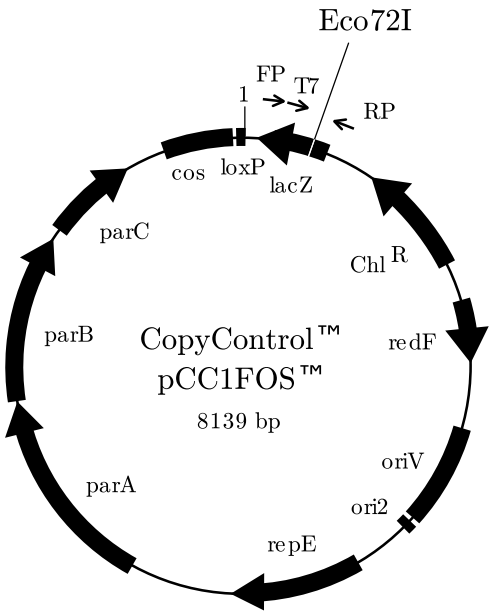


Figure 6.1: Commercial fosmid vector, pCC1FOS. pCC1FOS is avilable from Epicentre Biotechnologies. Notable elements include: chloramphenicol resistance, an F origin of replication for *E. coli*, and an RK2 origin of replication for Proteobacteria compatible with strains carrying *trfA*.

Table 6.1: Examples of metagenomic libraries constructed from diverse environmental samples using cloning vector pCC1FOS or derivatives. Libraries that are based on the commercial pCC1FOS or pCC2FOS vector can be screened in any RK2-compatible host that expresses the *trfA* gene product required for the broad-host-range RK2 *oriV* origin of replication.

Sampled environment	Vector; screening host(s)	Ref.
Host-associated environments		
bovine rumen	pCC1FOS; *E. coli*	[321]
elephant feces	pCC1FOS; *E. coli*	[237]
human distal ileum	pCC1FOS; *E. coli*	[34]
human feces	pCC1FOS; *E. coli*	[138]
human feces (pescatarian)	pCC1FOS; *E. coli*	[299]
marine sponge	pCC1FOS	[343]
termite gut	pCC1FOS, pCC2FOS; *E. coli*	[191, 324]
Extreme environments		
Alaskan soil	pCC1FOS; *E. coli*	[3]
Alaskan floodplain soil	pCC1FOS; *E. coli*	[335]
Antarctic Pennisula meltwater	pCC1FOS; *E. coli*	[87]
glacial ice	pCC1FOS; *E. coli*	[270]
hot spring sediment/biofilm	pCT3FK; *E. coli*, *T. thermophilus*[*]	[177]
hydrothermal fluids	pCC1FOS; *E. coli*	[24]
Marine or freshwater environments		
bog	pCC1FOS; *E. coli*	[282]
marine sediment	pRS44; *P. fluorescens*, *X. campestris*[†]	[1]
ocean tidal flat sediment	pCC1FOS; *E. coli*	[173, 174]
ocean water column	pCC1FOS	[59]
river sediment	pCC1FOS; *E. coli*	[237]

Continued on next page

[*] *Thermus thermophilus*
[†] *Pseudomonas fluorescens*, *Xanthomonas campestris*

Table 6.1 – *Continued from previous page*

Sampled environment	**Vector; screening host(s)**	**Ref.**
Polluted environments		
crude oil-contaminated shore	pMPO579; *E. coli*‡	[305]
polluted river	pCC1FOS; *E. coli*	[316]
Agricultural, engineered, or other environments		
activated sludge	pCC1FOS, pCC2FOS; *E. coli*	[294, 345]
compost: leaf branch	pCC1FOS; *E. coli*	[295]
compost: lumber waste	pCT3FK; *E. coli*, *T. thermophilus**	[177]
compost: wood, manure, plant debris	pCC1FOS; *E. coli*	[226]
decomposing leaf litter	pCC1FOS; *E. coli*	[225]
orchard soil	pCC1FOS; *E. coli*	[65]
sugarcane bagasse	pCC1FOS	[213]

The pCC1FOS cloning vector has several advantages over other commercial options. It carries a chloramphenicol resistance (*cat*) marker that is superior to the common ampicillin resistance (*bla*) marker; because beta-lactamases that break down ampicillin are secreted into the media, satellite colony formation sometimes arises on ampicillin selection plates, and this background growth can be particularly problematic for the dense platings that are often required for library construction. In addition to an F plasmid origin of replication for single-copy maintenance, the pCC1FOS vector also carries an *oriV* origin of replication from the RK2 plasmid. The *oriV* is a broad-host-range origin, conferring the ability to replicate in diverse members of the *Proteobacteria* [10], but requires the *trfA* gene product for replication and results in an

‡derivatives of *E. coli* EPI300 to increase transcription

estimated 15 copies per cell [68]. Though *trfA* is not carried by the fosmid, it can be provided in trans; notably, the commercial *E. coli* strain EPI300 (also available from Epicentre Biotechnologies) carries *trfA* under the control of an inducible promoter that is advertised to increase copy number from 1 copy per cell to 10-200 copies. The strain likely possesses a *trfA* copy-up mutant allele under control of *araC*-P_{BAD}, which is induced by L-arabinose [333]. In the past, our lab has preferred HB101 as a library host due to its receptiveness to transduction, but I have found that EPI300 appears to transduce at least as well as, if not better than, HB101 (Table 5.4). It also has the advantages of being an *endA1* mutant and supporting copy-number inducibility, allowing for less-degraded and higher-yield plasmid DNA preparations, respectively.

pCC1FOS lacks an origin of transfer

Despite its popularity, pCC1FOS has some disadvantages that make resulting libraries less versatile than they could be. First, pCC1FOS does not possess an origin of transfer (*oriT*) that would allow the fosmid to be efficiently transferred by conjugation, mediated by a helper plasmid, to other species that may be more suitable for heterologous expression or even to different strains of *E. coli*. Others have achieved conjugation capabilities by adding the RK2 *oriT* to pCC1FOS [1, 29, 305]. To enable conjugation after library construction has already taken place, still others have retrofitted individual pCC1FOS-based clones with an *oriT* [29, 179]. This retrofitting strategy has also been used for the cosmid vector SuperCos-1 [115], which is an alternative *cos*-based cloning vector (Stratagene, Agilent Technologies). These modifications illustrate the need for fosmid and cosmid vector design to include the *oriT* so that duplication of work can be avoided. It is possible that transformation can be used to transfer libraries to other hosts, but only for recipients that are amenable to those techniques and that will not reject DNA that has been synthesized in *E. coli* due to the presence of host

restriction-modification systems. There is little leeway here though if desired hosts are isolates that have not yet been adapted for routine laboratory techniques.

pCC1FOS is not inherently broad-host-range

Given that the broad-host-range *oriV* is used to achieve a higher copy number in conjunction with EPI300 expressing the *trfA* gene, another disadvantage of pCC1FOS is that *trfA* is not included on the vector. The consequence is that species that would otherwise be able to use the *oriV* cannot replicate pCC1FOS. Perhaps it is not surprising then that for the vast majority of studies highlighted here (Table 6.1), *E. coli* was used as the screening host. This is an enormous disadvantage for functional metagenomics because different clones can be isolated from the same metagenomic library when different screening hosts are used [50, 204]. Our lab has found that using the legume-symbiont *Sinorhizobium meliloti* as a host results in a much greater diversity of clones than *E. coli* when screening a corn field soil metagenomic library for beta-galactosidase activity, though this greater diversity does not appear to be related to phylogenetic distance of the origin of the cloned DNA to the surrogate host [Cheng et al., in preparation]. The importance of devising systems that allow for functional screening in diverse expression hosts has been reviewed by others [71, 187, 302, 312], but what of the large number of libraries that have already been constructed? Can we make use of them for screening in non-*E. coli* hosts? The libraries listed in Table 6.1, as well as potentially many other metagenomic libraries constructed using pCC1FOS or derivatives, would be accessible to any RK2-compatible host if a copy of the *trfA* gene were also made available. This solution has already been applied by others: one group inserted a wild-type *trfA* gene into the chromosome of the Gammaproteobacteria species *Pseudomonas fluorescens* and *Xanthomonas campestris* for screening of libraries constructed using a pCC1FOS derivative [1]. Another group inserted *araC*-

P_{BAD}-*trfA* into the chromosome of *E. coli* to give copy number inducibility to the lambda Red recombineering strain EL350 [327]. The introduction of *trfA* into RK2-compatible species is a straightforward way to expand the range of expression hosts for existing pCC1FOS-based libraries.

An alternative to inserting the *trfA* gene into desired expression hosts for maintaining metagenomic clones is to modify the vector for integration into the host genome, which bypasses the requirement for *trfA*. This strategy has already been employed to integrate clones into a locus in the genome of the thermophile *Thermus thermophilus* for functional screening: pCC1FOS was first modified to include a *T. thermophilus* selectable marker as well as regions for homologous recombination at the target locus [7]. In our lab, pCC1FOS has also been modified by John Heil to carry ΦC31 *att* sites [122] for integrase-mediated site-specific recombination of cloned insert DNA into the genomes of landing pad strains, including *Sinorhizobium meliloti*, *Ochrobactrum anthropi*, and *Agrobacterium tumefaciens* [123]. As a general strategy, however, chromosomal insertion is potentially less useful than recombinant clone maintenance due to the difficulty in retrieving the integrated insert DNA for manipulation, including DNA sequence analysis, when non-arrayed (i.e., pooled) libraries have been screened.

6.3.3 Inclusion of transcriptional terminators in cloning vectors

In addition to an *oriT* and broad-host range *oriV*, pCC1FOS may also be improved by the addition of transcriptional terminators that flank the fosmid's Eco72I cloning site (Figure 6.1). The benefits of using terminators for cloning have previously been discussed (Section 4.4.4); briefly, transcriptional terminators may help alleviate cloning bias in some cases where DNA, particularly AT-rich DNA, may contain sequences that resemble the σ^{70} promoter consensus sequence. Spurious transcription initiating from efficient promoters near the vector-insert junction can interfere with the plasmid's origin of replication or can lead to overproduction of proteins involved in control of plasmid copy number, affecting plasmid maintenance [291]. For cloning metagenomic DNA, it may be a good precaution to include terminators that prevent transcription into the vector backbone and indeed, commercial vectors are available that make use of transcriptional terminators to combat this problem.

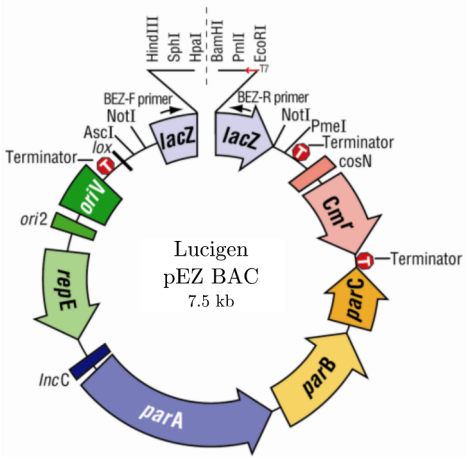


Figure 6.2: Lucigen pEZ BAC cloning vector includes transcriptional terminators. Transcriptional terminators indicated by red stop signs. Two terminators flank the cloning site to reduce isert-driven transcription and one terminator follows the *parC* gene to reduce vector-driven transcription into the insert. Adapted from Lucigen BAC Cloning Kits product manual.

For example, the pEZ BAC cloning vector from Lucigen Corporation (Figure 6.2) has two terminators that flank the cloning site to reduce insert-driven transcription. Interestingly, it also has another terminator to reduce vector-driven transcription into the insert. In one particular line of vectors available from Lucigen, the linear pJAZZ vectors, the two terminators flanking the cloning site were disclosed as the phage T7 terminator and the *E. coli rrnB* terminator [106], both of which have been documented as relatively strong terminators in standardized tests of terminator efficiency [31].

6.3.4 Testing the efficiency of transcriptional terminators

The characterization of transcriptional terminator strength has been of recent interest as more parts are needed to build complex systems in synthetic biology endeavours. This has led to the standardized testing of hundreds of natural and synthetic transcriptional terminators to both understand their sequence determinants to aid in prediction and modelling of terminators, as well as to find strong terminators for use in designed biological systems that require a tight control of transcription by RNA polymerase [31, 42].

Efforts to characterize terminators so far have focussed on only intrinsic termination, also called Rho-independent termination, which is one of two ways that transcription can be terminated in *E. coli* and accounts for ∼80% of terminators in its genome [252]. Intrinsic terminators consist of a GC-rich hairpin-forming sequence followed by a run of Ts in the DNA, which is called the oligoT tract in the DNA [157] or the U-tract [42] or poly-U tail [31] in the corresponding RNA. The folding of the nascent RNA into a hairpin disrupts the RNA:DNA hybrid that stabilizes the transcription elongation complex, leading to its dissociation; the stretch of Ts downstream from the hairpin sequence is important in this process because it contributes to pausing

of the elongation complex, allowing time for hairpin formation [114, 157].

In contrast to intrinsic termination, Rho-dependent termination is more complex: Rho binds to sites on the RNA in a sequence-specific manner, and can traverse the transcript to catch up with the elongation complex to cause RNA release; however, the binding sites can be separated from the site of transcriptional termination by hundreds of nucleotides, and the factors leading transcriptional termination by Rho are currently not well understood [252]. This complexity makes Rho-dependent termination difficult to predict on the basis of sequence and thus efforts to characterize terminators have concentrated on the more straightforward intrinsic terminators.

The strength (or efficiency) of intrinsic terminators has been measured using devices designed for standardized testing with fluorescent reporter proteins: briefly, flow cytometry is used to compare the level of expression of a reporter downstream of the transcriptional terminator to the level of expression of an upstream reporter, normalizing to measurements obtained from a control construct lacking a terminator (Figure 6.3).

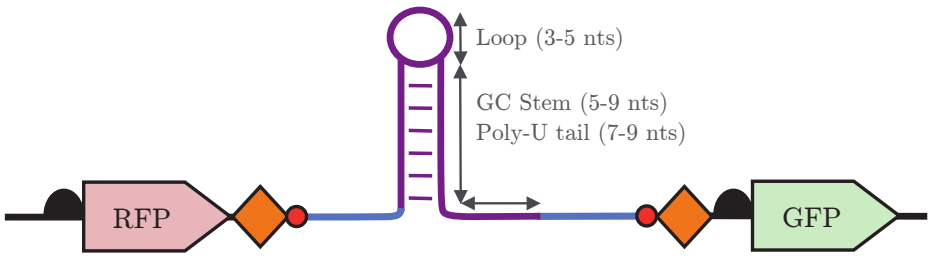


Figure 6.3: Device for standardized testing of transcriptional terminators. Flourescent protein reporter genes are used to measure the efficiency of transcriptional terminators. Adapted from [31].

To be able to compare measurements of terminator strength requires careful design of the testing device, including the choice of upstream and downstream reporters, as well as understanding possible effects of neighbouring sequence context, which can influence terminator strength [31]. In any case, the use of fluorescent reporters is a convenient way to gauge whether transcriptional terminators are functioning in vivo.

6.3.5 Aims of this work

In Chapter 5, I used the commercial vector pCC1FOS to construct the *B. theta*-compatible fosmid vector pKL13. I included the *oriT* for conjugation ability and two unidirectional, Rho-independent transcriptional terminators that flank the cloning site to reduce potential transcription into the vector. The latter were introduced by the cloning of a synthesized fragment. This chapter elaborates on how the transcriptional terminator (TT) fragment was designed, providing rationale for each element, particularly those required for terminator testing. The main objective was to test the functionality of the terminators, that is, to determine whether each of the two terminators was indeed able to reduce transcription in the fosmid context. Such confirmation would justify their inclusion in the pKL13 library cloning vector. Given this objective, the testing was intended to be a crude check and involved just one reporter protein, GFPuv. This reporter was measured in the presence versus absence of each of the two transcriptional terminators, demonstrating that each terminator behaved as expected in reducing transcription.

6.4 Results and discussion

6.4.1 Design of a transcriptional terminator fragment

After deciding to introduce terminators to reduce potential transcription into the pCC1FOS vector backbone, I considered two options for the synthesis of the fragment encoding transcriptional terminators, using Integrated DNA Technologies (IDT) as the manufacturer: custom gene synthesis or gBlocks gene fragments. The difference between the two is that custom gene synthesis delivers the desired fragment cloned into a plasmid, whereas gBlock fragments arrive as uncloned double-stranded fragments. Because a gBlock fragment is not cloned, the product will contain a small proportion of incorrect sequences, such as insertions or deletions, although the product is accompanied by an estimated purity and a recommendation from IDT regarding the probability of obtaining the correct clone. For example, for a ∼1,500-bp fragment with approximately 85% purity, the IDT technical support team suggests that users screen about 6 colonies for >95% chance of the correct clone.

The price for gene synthesis was estimated to be nearly three times that for gBlock synthesis. The transcriptional terminator (TT) fragment was synthesized as a gBlock, and because the lab was offered a free trial, I also designed the TT fragment to include the ∼1-kb gentamicin resistance stuffer as well as all the elements required for testing both the transcriptional terminators, short of a reporter gene. The final design came to 1,500 bp (Figure 6.4; DNA sequences provided in Table 6.2).

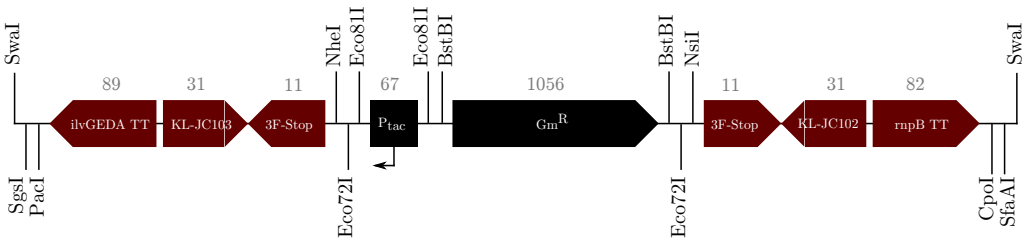


Figure 6.4: Transcriptional terminator (TT) fragment design. The length of each element is indicated by the number above the element. Note that this diagram is stylized and is therefore not to scale.

More specifically, the following elements were included in the TT fragment:

- A stuffer fragment, flanked by Eco72I sites (a.k.a PmlI; `CAC^GTG`). As described in Section 5.4.2, the presence of a stuffer aids in complete digestion of the vector; the vector is typically purified to remove the stuffer prior to ligation.
- A gentamicin resistance gene within the stuffer, identical to the sequence from pJC8. The resistance gene is flanked by BstBI sites (a.k.a. Bsp119I; `TT^CGAA`) for optional removal or swap of the resistance gene. The resistance gene was included because (1) there was no cost to synthesis, (2) it would reduce required cloning downstream, and (3) the resistance gene confers antibiotic resistance that would make cloning the synthesized fragment more straightforward.
- An inducible P_{tac} promoter, also within the stuffer. P_{tac} is a strong promoter, possessing the consensus sequence for $rpoD/\sigma^{70}$ (Figure 4.9). The promoter is flanked by Eco81I sites (a.k.a. Bsu36I; `CC^TNAGG`). This restriction enzyme was specifically chosen for being a 7-cutter and lacking specificity at the centre base; though Eco81I will cut on either side of the P_{tac}, the two sites are actually different in sequence: `CC^TAAGG` on the NheI side versus `CC^TCAGG` on the NsiI side.

This design allows for the future swapping-in of different promoters, if desired: complementary oligos can be synthesized, annealed to form the double-stranded promoter sequence, and then cloned in directionally.

- Two transcriptional terminators positioned to reduce transcription outward – that is, into the vector backbone. Both the *ilvGEDA* and *rnpB* T1 transcriptional terminators are from *E. coli* MG1655; sequences were taken from the comprehensive study on terminator efficiency by Cambray et al. [31]. These were not the strongest terminators reported in that study because the strong stem-loop structures associated with very strong terminators were incompatible with gBlock synthesis; however, both of the chosen terminators were reported by Cambray et al. to reduce expression 64- to 128-fold, which still make them very good transcriptional terminators.

- A 3-frame translational stop upstream of each of the two transcriptional terminators. These two translational stops differ in sequence to avoid potential problems with homologous recombination (Table 6.2). They were designed upstream of the transcriptional stops to ensure that the latter are effective: if perchance ribosomes were actively translating the nascent mRNA, transcriptional termination may be abolished due to interference with the formation of the stem-loop structure in Rho-independent termination [337].

- Two primer-binding sites for Sanger end-sequencing of cloned inserts, KL-JC102 and KL-JC103. Other than the addition of an extra base, these sequences are identical to the sequencing primer sites for pJC8 (the extra base was included to ensure the 3' end of the primer ends with two bases that are either C or G). These sites are internal to the transcriptional terminators because the stem-loop structures may hinder Sanger sequencing. Furthermore, each primer-binding

site was positioned between the translational stop and the transcriptional stop because there is evidence to suggest that transcriptional termination is effective when the spacing is ∼20-60 bases [337].

- A pair of unique restriction sites downstream of each of the two transcriptional terminators, for directional cloning of a downstream reporter gene to test terminator functionality. The two pairs were: SgsI (a.k.a. AscI; `GG^CGCGCC`) and PacI (`TTAAT^TAA`) on the *ilvGEDA* side, and CpoI (a.k.a. RsrII; `CG^GWCCG`) and SfaAI (a.k.a. SgfI, AsiSI; `GCGAT^CGC`) on the *rnpB* side.

- Single restriction sites upstream of each translational stop: NheI (`G^CTAGC`) on the *ilvGEDA* side and NsiI (a.k.a. Mph1103I; `ATGCA^T`) on the *rnpB* side. These were included for two reasons, with only the first being relevant to this chapter: (1) in the case that an additional upstream reporter gene had to be cloned for testing transcriptional termination (as in Figure 6.3) – ideally two sites would have been included on each side for directional cloning but unique restriction sites were limited for a vector of this size (Figure 5.12); (2) so that the cloned insert DNA can be released from the vector for restriction digest analysis of clones from metagenomic libraries.

- A SwaI site (`ATTT^AAAT`) on both ends of the fragment so that the entire TT fragment can be subcloned from one vector to another.

Table 6.2: DNA sequences for elements of the TT fragment.

Element	Length	Sequence (5' to 3')
excess bases	5	GCATA
SwaI	8	ATTTAAAT
SfaAI	8	GCGATCGC
spacer	8	GACCTGCT
CpoI	7	CGGACCG
rnpB T1 TT	82	GACAGTCATTCATCTTTCTGCCCCTCCAAAAGCAAAAACCCGCCGAAGCGGGTTTTTACGTAAATCAGGTGA AACTGACCGA[§]
KL-JC102 F	31	TAACAATTTCACACAGGAAACAGCTATGACG
3F stop 1	11	TCACCTAGTTA[§]
NsiI	6	ATGCAT
Eco72I	6	CACGTG
BstBI	6	TTCGAA
Gm resistance	1056	CGTGTTGCCCCAGCAATCAGCGCGACCTTGCCCCTCCAACGTCATCTCGTTCTCCGCTCATGAGCTCAGCCA ATCGACTGGCGAGCGGCATCGCATTCTTCGCATCCCGCCCTCTGGCGGATGCAGGAAGATCAACGGATCTCG GCCCAGTTGACCCAGGGCTGTCGCCACAATGTCGCGGGAGCGGATCAACCGAGCAAAGGCATGACCGACTGG ACCTTCCTTCTGAAGGCTCTTCTCCTTGAGCCACCTGTCCGCCAAGGCAAAGCGCTCACAGCAGTGGTCATT CTCGAGATAATCGACGCGTACCAACTTGCCATCCTGAAGAATGGTGCAGTGTCTCGGCACCCCATAGGGAAC CTTTGCCATCAACTCGGCAAGATGCAGCGTCGTGTTGGCATCGTGTCCCACGCCGAGGAGAAGTACCTGCCC ATCGAGTTCATGGACACGGGCGACCGGGCTTGCAGGCGAGTGAGGTGGCAGGGGCAATGGATCAGAGATGAT CTGCTCTGCCTGTGGCCCCGCTGCCGCAAAGGCAAATGGATGGGCGCTGCGCTTTACATTTGGCAGGCGCCA GAATGTGTCAGAGACAACTCCAAGGTCCGGTGTAACGGGCGACGTGGCAGGATCGAACGGCTCGTCGTCCAG ACCTGACCACGAGGGCATGACGAGCGTCCCTCCCGGACCCAGCGCAGCACGCAGGGCCTCGATCAGTCCAAG TGGCCCATCTTCGAGGGGCCGGACGCTACGGAAGGAGCTGTGGACCAGCAGCACACCGCCGGGGGTAACCCC AAGGTTGAGAAGCTGACCGATGAGCTCGGCTTTTCGCCATTCGTATTGCACGACATTGCACTCCACCGCTGA TGACATCAGTCGATCATAGCACGATCAACGGCACTGTTGCAAATAGTCGGTGGTGATAAACTTATCATCCCC TTTTGCTGATGGAGCTGCACATGAACCCATTCAAAGGCCGGCATTTTCAGCGTGACATCATTCTGTGGGCCG TACGCTGGTACTGCAAATACGGCATCAGTTACCGTGAGCCGGAGGATC[¶]
BstBI	6	TTCGAA
Eco81I	7	CCTCAGG

Continued on next page

[§]sequence shown has been reverse-complemented for continuity; see Figure 6.4
[¶]synthesized gBlock fragment differs from this sequence by a point mutation; see Appendix E.2

Table 6.2 – *Continued from previous page*

Element	Length	Sequence (5' to 3')
P_{tac}	67	GAGCTGTTGACAATTAATCATCGGCTCGTATAATGTGTGGAATTGTGAGCGGATAACAATTTCACAC
Eco81I	7	CCTAAGG
Eco72I	6	CACGTG
NheI	6	GCTAGC
3F stop 2	11	TGACTAAGTGA
KL-JC103 R	31	CGAAAACCCTGGCGTTACCCAACTTAATCGC‖
ilvGEDA TT	89	TAGAGATCAAGCCTTAACGAACTAAGACCCCCGCACCGAAAGGTCCGGGGGGTTTTTTTTGACCTTAAAAACA TAACCGAGGAGCAGACA
PacI	8	TTAATTAA
spacer	8	ATCCAGCC
SgsI	8	GGCGCGCC
SwaI	8	ATTTAAAT
excess bases	5	TTGAC

6.4.2 Synthesis and cloning of terminator fragment

The TT fragment was synthesized by IDT as a gBlock gene fragment in the form of 200 fmol (200 ng) of the product – blunt DNA fragments with phosphorylated 5' ends. I attempted to clone the TT fragment into two different vectors concurrently: first, directly into pKL7 (Figure 5.12) to generate the desired *B. theta*-compatible fosmid vector; second, in case the first attempt did not work, the TT fragment was also cloned into the intermediate vector pJET1.2 for eventual transfer to pKL7. Both cloning attempts were successful and a couple of clones from each were chosen for screening by sequencing to find the correct clone; the pJET1.2-based clones were called pKL9 and the fosmid-based clones were called pKL10 (Table 2.2).

‖sequence shown has been reverse-complemented for continuity; see Figure 6.4

Both clones in the fosmid backbone contained a deletion of a critical base in the *ilvGEDA* terminator (Figure 6.5A and B) and in fact, all four clones contained a deletion in the gentamicin resistance gene fragment (Figure 6.5A and C), although this did not affect the gentamicin resistance phenotype. Not surprisingly, the error in the terminator was in a run of As (corresponding to the U-tract) near the core stem-loop structure. From this experience, it is probably advisable to use gBlocks gene fragments for sequences without known strong secondary structure and for fragments of relatively small size to minimize the cost of screening by sequencing.

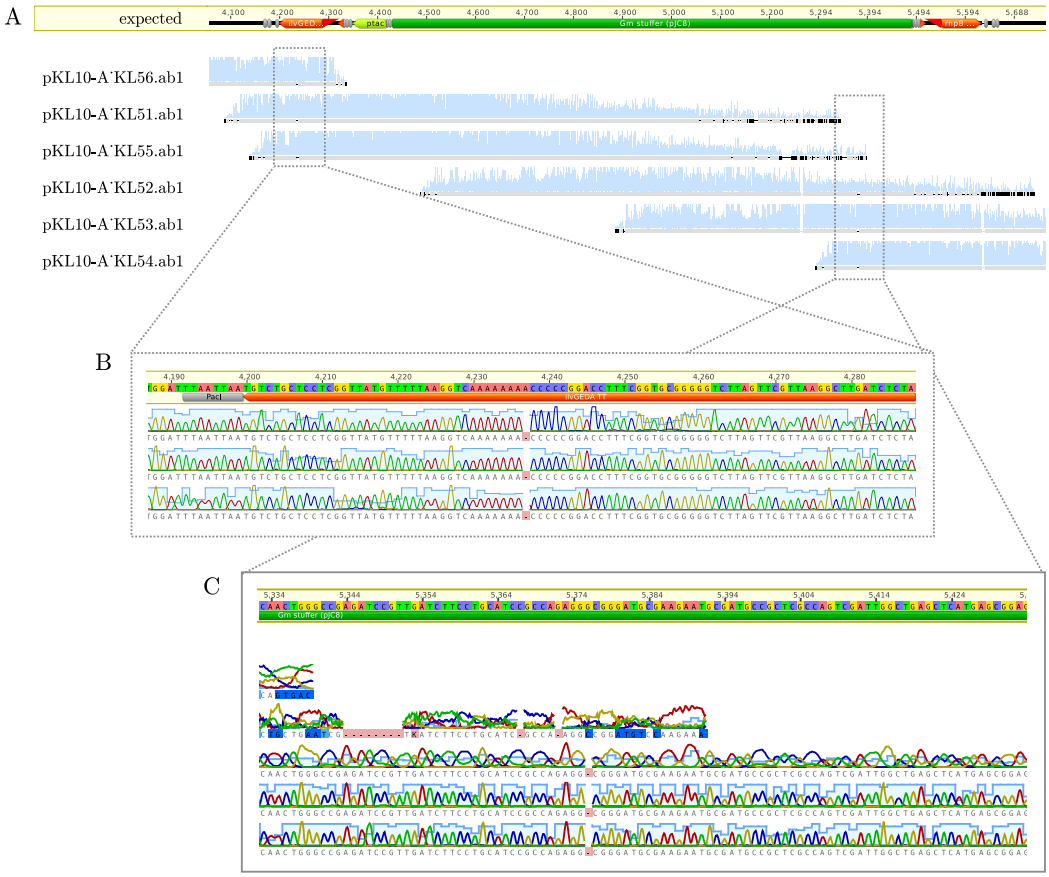


Figure 6.5: Screening by Sanger sequencing for correct TT fragment sequence. Six Sanger reads were obtained for pKL10 and aligned to the expected sequence (A) revealing two errors: one in the *ilvGEDA* terminator (B) and the other in the gentamicin resistance gene fragment (C). Adapted from images generated by Geneious version 6.0 created by BioMatters.

Because pKL9 carried only the inconsequential error in the gentamicin resistance gene fragment, the TT fragment was usable (see Appendix E.2 for sequence). Accordingly, the TT fragment was subcloned from pKL9 as a blunt-ended SwaI-fragment into the blunt Eco72I site of pKL7, generating pKL13 (Figure 5.12), which was the final library vector that I used for constructing *B. theta*-compatible libraries in Chapter 5. After constructing pKL13, the TT fragment was double-checked by restriction digest, using all of the enzymes whose sites were designed into the fragment (Figure 6.6).

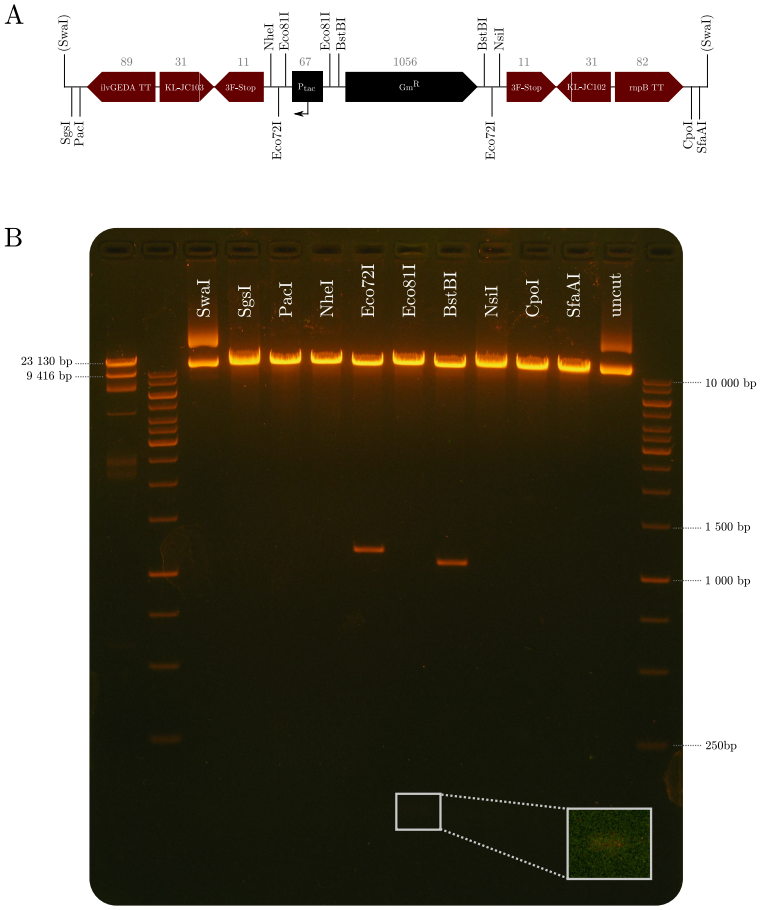


Figure 6.6: Restriction digest check of TT fragment cloned in pKL13. Restriction enzyme sites and sizes of the elements in the TT fragment (A) were checked by restriction digest alongside uncut pKL13 as control (B); pullout shows agarose gel section under adjusted brightness and contrast to increase DNA fragment visibility.

6.4.3 Testing functionality of transcriptional terminators

Though the two chosen transcriptional terminators, the *ilvGEDA* and *rnpB* T1 terminators, had been previously characterized [31], I wanted to confirm that the terminators were indeed functional in their new context to justify their inclusion in the *B. theta*-compatible library vector, pKL13 (Figure 5.12). This confirmation was not meant to be a precise quantification of terminator efficiency; rather, it was intended to be a crude check of function. To do a simple check of transcriptional terminator functionality, a reporter gene can be cloned downstream of the terminator, and the level of expression of that reporter gene can be compared in the presence versus absence of the terminator. For a reporter, I chose green fluorescent protein (GFP), specifically the GFPuv variant isolated by molecular evolution and determined to be 16-18 times brighter than wild-type GFP [51]. Though the fluoresecence of another variant called enhanced GFP (EGFP) has been reported to be even higher – about 35 times brighter than wild-type GFP [344], EGFP may be more appropriate for eukaryotic rather than bacterial systems [144].

Having chosen the reporter gene, the vector was then prepared for the introduction of the GFPuv reporter: to test each terminator, the P_{tac} promoter must be upstream of that terminator and the GFPuv reporter must be downstream. pKL13 had the P_{tac} promoter oriented toward the *ilvGEDA* terminator, and thus I cloned GFPuv downstream of the terminator to generate pKL15, and then deleted the terminator to generate pKL16 (Figure 6.7, left). To generate constructs for testing the *rnpB* terminator, I first reversed the orientation of the Eco72I stuffer (see Section 6.6.2) to generate in pKL17, which placed the P_{tac} upstream of the *rnpB* terminator; analogous to the first set of constructs, I cloned GFPuv downstream of the terminator to generate pKL18, and then deleted the terminator to generate pKL19 (Figure 6.7, right).

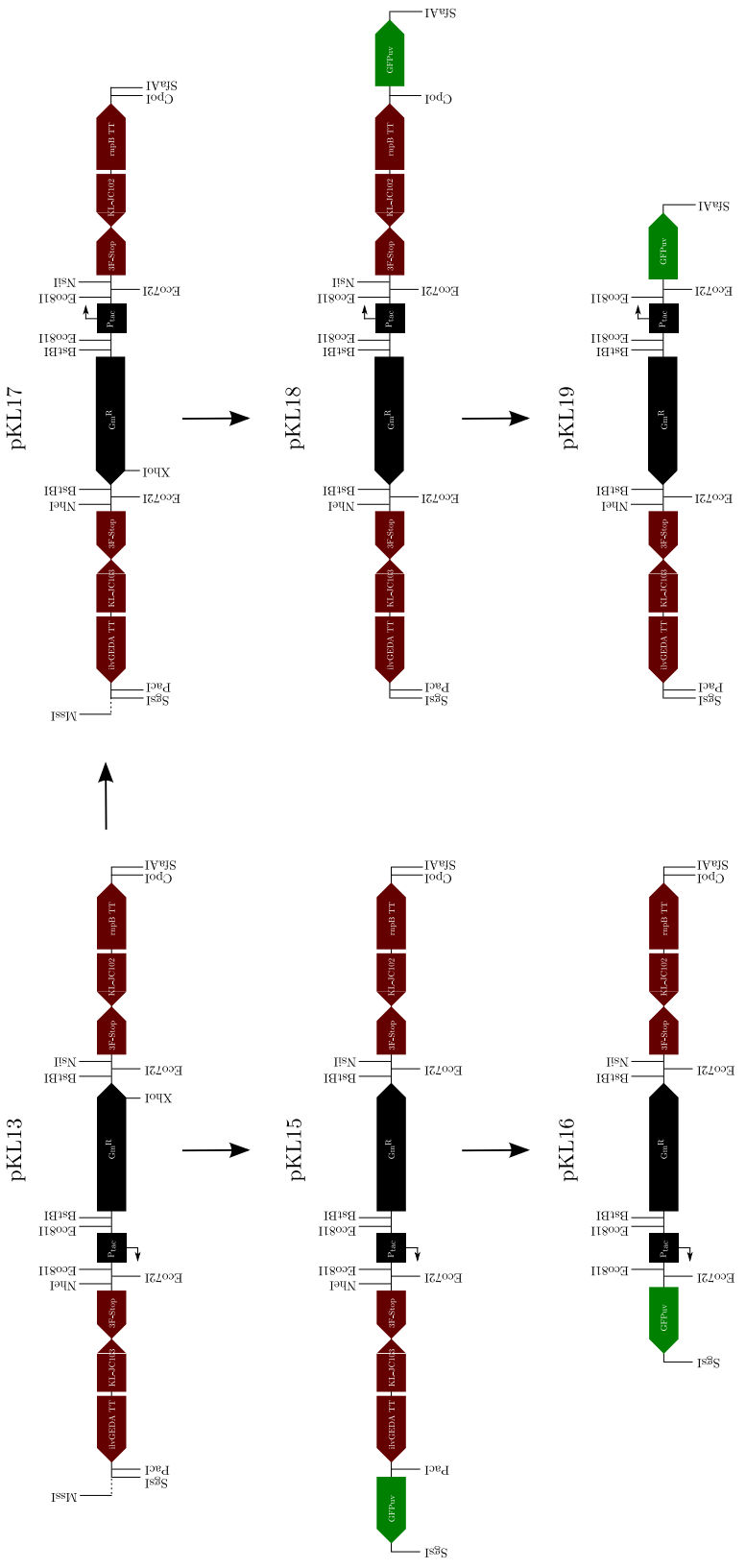


Figure 6.7: Overview of constructed plasmids for testing of transcriptional terminators using GFPuv reporter gene. To test functionality of the *ilvGEDA* and *rnpB* transcriptional terminator sequences, respectively: the P_{tac} promoter orientation was manipulated (pKL13 and pKL17), the GFPuv reporter gene was cloned (pKL15 and pKL18), and the terminators were deleted (pKL16 and pKL19).

In each of the two plasmids that now lacked either the *ilvGEDA* or *rnpB* terminator– pKL16 or pKL19, respectively – DNA was deleted starting from the translational stop to the transcriptional stop, inclusive (Figure 6.7). I decided to delete the entire segment instead of simply deleting the stem-loop-containing sequences because the segment was designed to work as a unit for the termination of transcription. To see if the deleted sequences were conferring transcriptional termination in the fosmid context, the fluorescence from expressed GFPuv was compared in EPI300 cells carrying constructs with versus without the terminator unit – that is, pKL15 was compared to pKL16 while pKL18 was compared to pKL19 – under two different conditions (Figure 6.8).

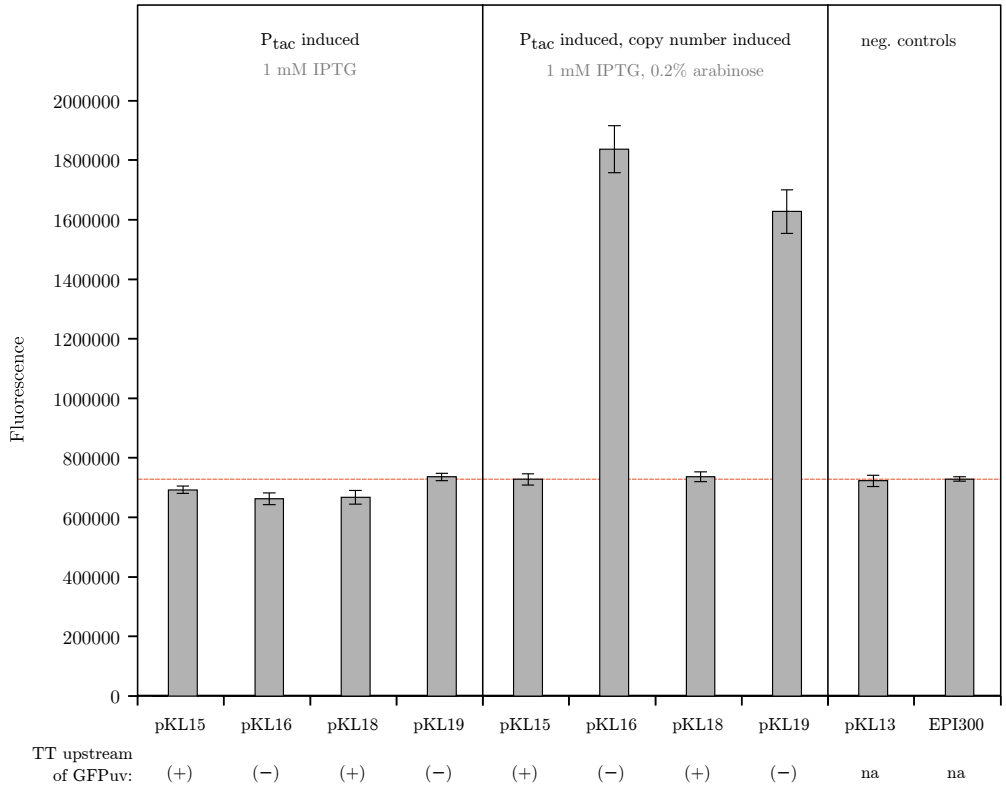


Figure 6.8: Fluorescence from EPI300 cells expressing GFPuv with or without transcriptional terminators. Strains were grown under two different conditions to assay GFPuv fluorescence from constructs with and without the *ilvGEDA* or *rnpB* terminating sequences; red line indicates background fluorescence of empty cells.

The two different conditions used to test the presence versus absence of the terminator units were: (1) P_{tac} promoter induction alone using IPTG, or (2) P_{tac} promoter induction in combination with copy number induction using arabinose. In the latter condition, the presence of the terminators resulted in cells displaying a level of fluorescence that was comparable to negative-control cells that lacked GFPuv; conversely, the absence of either the *ilvGEDA* terminator unit (pKL16) or the *rnpB* terminator unit (pKL19) led to an increase in fluorescence as a result of higher GFPuv transcription (Figure 6.8, centre and right panels). This result confirmed that the two unidirectional terminators are functional in the pKL13 context. Interestingly, this difference was only observed when plasmid copy number was induced (Figure 6.8, left versus centre panel), indicating that there is a limit of detection with the current experimental set-up (see Section 6.6.5). It would be interesting to know what the exact copy number is for these plasmid constructs that were compared, as Epicentre provides a rather large range for copy number (from 10 to 200 copies per cell) without explanation of the influencing factors [77].

In considering copy number for these constructs, it is conceivable that the copy number of plasmids with the terminator may be different from the copy number of the those lacking the terminator, as copy number can be affected by various factors, such as growth media composition and nutrient limitation [95] or, in this case, the presence/size of cloned DNA [280,333]. In the case of these GFPuv testing constructs, however, the difference of ∼150 bases between constructs being compared is unlikely to lead to very large differences in copy number, although if there were a difference, it is more likely that increased transcription would lead to decreased copy number, meaning that the difference observed in GFPuv expression would be even greater if plasmid copy number were controlled for. It would be interesting to see how strong transcription affects plasmid copy number for this particular vector. To control for

differences in copy number, plasmid copy number can be estimated for each plasmid in the particular strain under the specific growth condition [172]; alternatively, differences in plasmid copy number can be accounted for by simply using testing constructs that make use of an upstream reporter gene in addition to a downstream reporter gene so that transcription can be normalized to variability in reporter gene expression owing to factors other than transcription termination (Figure 6.3). That being said, the precise quantification of terminator efficiency is beyond the scope of this thesis, although I did design the TT fragment to allow for upstream reporter gene cloning (Figure 6.4).

6.4.4 Constructs for testing the effect of transcription on cloning bias

The TT fragment was designed with two intentions: (1) to include terminators in the *B. theta*-comptaible vector where they may help alleviate cloning bias (as discussed in Section 4.5), and (2) to use in further experiments to test the extent to which transcriptional terminators protect against cloning bias. For the latter, one future goal is to compare the cloning bias between two metagenomic libraries that have been constructed in a vector with transcriptional terminators versus one without. To prepare vectors for this purpose, I deleted the P_{tac}-gentamicin stuffer in pKL13 (Figure 5.12) and replaced it with only the gentamicin stuffer gene from pJC8, generating pKL20, although the orientation of the gentamicin stuffer gene in pKL20 is currently uncertain (Figure 6.9A, B, and C). The next step would be to delete the two transcriptional terminator units to obtain a vector identical to pKL20 but for the missing terminators (Figure 6.9C and D).

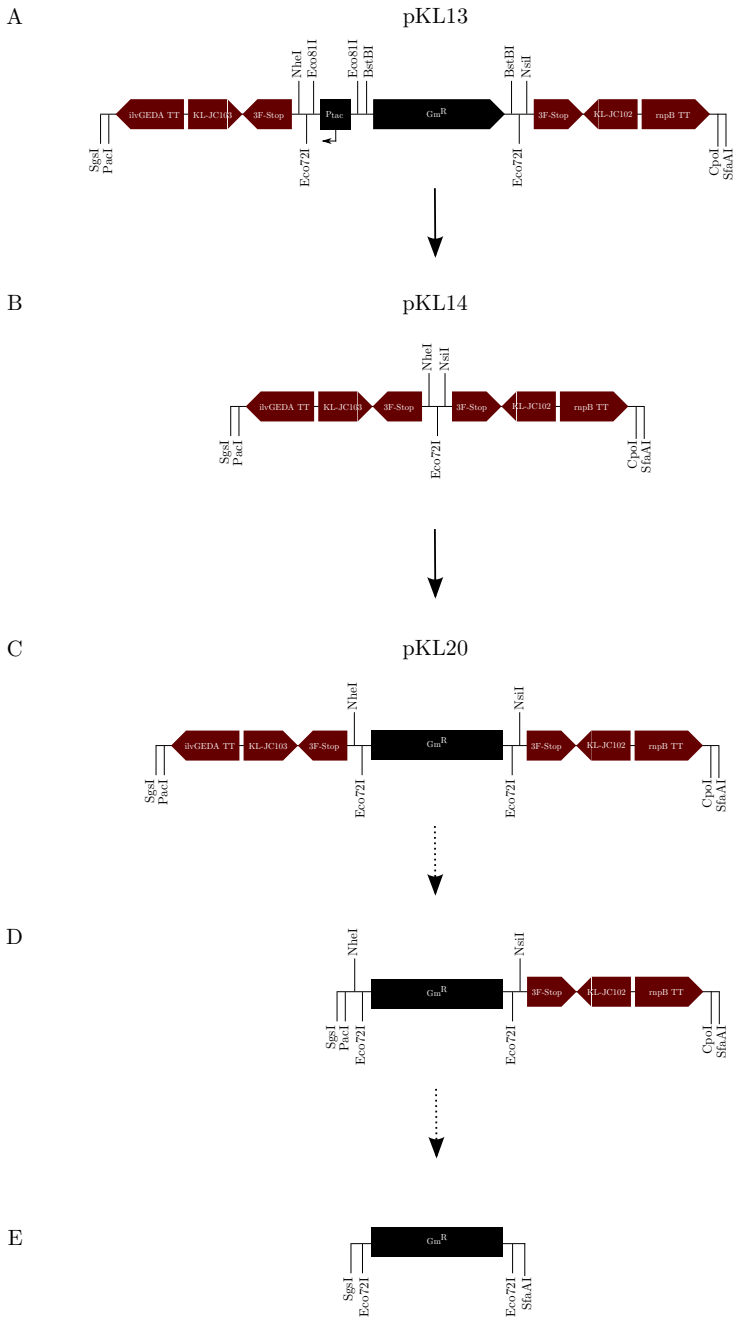


Figure 6.9: Vectors for future work to test the effect of transcription terminators on cloning bias. (A) pKL13, containing the TT fragment; (B) pKL14, in which the P_{tac}-gentamicin stuffer was removed as a Eco72I frament; (C) pKL20, which contains just the gentamicin resistance gene stuffer from pJC8; (D) and (E) show suggested next steps for removal of the two transcriptional terminators to generate a vector that can be used for direct comparison to pKL20.

Cloning bias could then be compared between the two vectors, pKL20 and that of Figure 6.9E, after using them to construct metagenomic libraries from the same DNA sample. For cloning bias experiments, it may be desirable to first delete the *ermF-repA* fragment from pKL20, which would reduce cloning vector size by ∼4 kb (see Figure 5.12C and D).

A advantage of using the pCC1FOS backbone is that plasmid copy number can be induced, allowing comparison of cloning bias not only between presence versus absence of terminators, but also between single-copy versus multi-copy maintenance of metagenomic libraries. With carefully designed experiments, it may be possible to tease apart the factors that affect library representativeness – distinguishing transcriptional effects from copy-number effects, though it may be easier to do so with cloned fragments smaller in size than typical fosmid inserts.

Although the pCC1FOS backbone allows for copy number induction, cloning bias could be observed under even greater copy number, as would be the case for pUC-based vectors [340], which have been reported in the literature at up to 500-700 copies per cell [210]. It may be interesting to determine whether transcriptional terminators alleviate cloning bias under these conditions; in fact, the pUC19-based, high-copy pKL3 cosmid that I constructed is one vector that could be used for this purpose (Figure 5.8). To transfer the TT fragment – and any derivatives constructed from it – to a different vector, the fragment can simply be subcloned as an blunted SgsI-SfaAI fragment into the destination plasmid (Figure 6.9).

6.5 Conclusions

In the previous Chapter 5, pCC1FOS was modified to include an *oriT* to allow the vector to be conjugated between strains, as well as a TT fragment carrying transcriptional terminators that flank the cloning site to block transcription into the vector backbone. This chapter described the design, synthesis, and characterization of the TT fragment. Using GFPuv, the *ilvGEDA* and *rnpB* transcriptional terminator units were determined to be functional in the pKL13 fosmid context. This chapter also described the construction of plasmids and fragments that may be used to test the effect of transcription on the observed cloning bias of metagenomic libraries, although the various factors that lead to cloning bias and their relative contributions remain to be elucidated.

6.6 Specific materials and methods

6.6.1 Preparation of pCC1FOS-based vectors using arabinose induction

Plasmid minipreps of pCC1FOS-based vectors were prepared from cultures that had been induced with either 1× commerical autoinduction solution (Epicentre AIS107F) or 0.2% arabinose. EpiCentre sells the solution without details about composition, but based on the literature, it is clear that the inducer of plasmid copy number is arabinose. I induced using a final concentration of 0.2% arabinose (see Appendix E.1) although it might be useful to drop concentration to 0.02% [147]. I did not test varying concentrations of arabinose for optimal yields of plasmid DNA.

6.6.2 Reversing orientation of stuffer fragment

The construction of pKL17 from pKL13 required reversing the orientation of the stuffer fragment, so that the P_{tac} would be oriented in the opposite direction (see Figure 6.7 for construct diagrams). To release the stuffer, 1 µg of pKL13 was digested with Eco72I (Thermo-Fisher FD0364) in 20 µl and heat-inactivated at 80℃. For ligation, 1 µl of the digest was used in a 10 µl ligation using T4 DNA ligase (Thermo-Fisher EL0014). Ligations were used to transform EPI300 and clones were streak-purified and screened by restriction enzyme digest (see Appendix E.1 for agarose gel image).

6.6.3 Cloning of GPFuv

The GFPuv ORF and RBS were amplified from pGFPuv (1 ng; Genbank accession U62636) using primers KL47/KL48 (with PacI and SgsI adapters) or KL49/KL50 (with CpoI and SfaAI adapters) for cloning into pKL13 or pKL17, respectively (see Figure 6.7 for construct diagrams).

High-fidelity Phusion DNA polymerase (Thermo-Fisher F-530L) was used according to the manufacturer's recommendations. The two-step PCR protocol used is summarized in Table 5.9. To prepare for cloning, the PCR products were gel extracted, digested with the appropriate restriction enzymes, and column-purified, using routine protocols previously described in Chapter 2.

Table 6.3: PCR protocol for GFPuv.

Temperature	Duration	
98℃	30 sec	
98℃	10 sec	×30 cycles
72℃	30 sec	
72℃	5 min	
22℃	hold	

6.6.4 Deletion of transcriptional terminators

Plasmid DNA was prepared for pKL15 and pKL18; 3 µg was used for PacI-NheI and NsiI-CpoI double digestion, respectively, to release the transcriptional terminators (see Figure 6.7 for construct diagrams). Digestions were incubated at 37℃ for 2.5 hours, and the vector backbone was gel extracted and purified. 200 ng of each sample was used in end-repair reaction using the End-It DNA End-Repair Kit (Epicentre ER81050) in a volume of 20 µl, according to the manufacturer's instructions. The reaction volume was then doubled by the addition of water to achieve 0.5 mM ATP concentration, and Fast-Link buffer and ligase were added (Epicentre LK0750H), according to the manufacturer's instructions. The ligation was incubated overnight at room temperature.

After the end-repair and ligation, the two desired constructs – with the transcriptional terminator deleted – no longer had the restriction sites that flanked the deleted terminator sequence. To effectively remove those DNA molecules that still had these sites due to possible incomplete digestion, the ligations were subjected to another double digest with the corresponding enzymes; this step digests undesired molecules, enriching for the correct ones. The digest was then used to transform EPI300 and clones were streak-purified and screened by restriction enzyme digest (see Appendix E.1 for agarose gel images).

6.6.5 Fluorescence assay for GFPuv expression

Strains were streaked from frozen stock onto solid media with the appropriate antibiotics. For each strain, an isolated colony was inoculated in triplicate into 5 ml of liquid media, using experimental and control conditions (Table 6.4). After overnight culture, 500 µl was transferred to 4.5 ml of saline and used to take an OD_{600} reading (Spectronic 20 spectrophotometer). The remaining 4.5 ml of culture was centrifuged at 8,000×g for 1 minute and resuspended in 1 ml saline.

The sample were effectively standardized by OD in the following manner: using the OD values, a standardization factor for each sample was calculated by dividing the OD of the sample with the lowest OD, by the OD of that sample; a fraction of each 1 ml sample was taken corresponding to the calculated dilution factor. Cells were pelleted by centrifugation at 8,000×g for 1 minute and resuspended in 300 µl saline. Samples were transferred to a black opaque microtiter plate for fluoresence assay on the FilterMax F5 Multi-Mode Microplate Reader (Molecular Devices) using the Softmax Pro software (version 6.2.2); the filters used were 360/35 nm for excitation and 535/25 nm for emission.

Table 6.4: Strains used in fluorescence assay to test transcriptional terminators.

Media	Strains
LB Cm_5, 1 mM IPTG, 0.2% arabinose	pKL15, pKL16, pKL17, and pKL18; all in EPI300
LB Cm_5, 1 mM IPTG	pKL15, pKL16, pKL17, and pKL18; all in EPI300
LB Cm_5	pKL13 in EPI300
LB Sm_{100}	EPI300

Chapter 7

Summary, future directions, and concluding remarks

7.1 Acknowledgements and declarations

Part of the discussion of this chapter was published as part of a Perspective article in the journal **Frontiers in Microbiology**. I was the primary author of this article. The citation for the article is:

> Lam KN, Cheng J, Engel K, Neufeld JD, Charles TC (2015) Current and future resources for functional metagenomics. *Frontiers in Microbiology* 6:1196. doi:10.3389/fmicb.2015.01196

I also acknowledge the following contributions:

- The text of the *Frontiers in Microbiology* manuscript was proofread and edited by **Katja Engel**, **Josh Neufeld**, **Trevor Charles**, and **Jiujun Cheng**.
- This remainder of this chapter was proofread by my supervisor **Trevor Charles**.

7.2 Abstract

Method development will be crucial to the continuing success of functional metagenomics for elucidating and understanding microbial gene function. This thesis has focused on development and analysis of methods for functional metagenomics, including devising strategies for large-insert clone sequencing (Chapter 3), understanding sequence bias in metagenomic libraries (Chapter 4), expanding screening host range for gut-derived libraries (Chapter 5), and exploring the importance of transcriptional terminators in cloning vectors (Chapter 6). The results presented in this thesis contribute towards method advancement, but also suggest new avenues for further investigation.

The near future may bring changes to the functional metagenomics field. For example, improvements in long-read sequencing technology, making it possible to obtain on the order of thousands of bases of accurate DNA sequence, will undoubtedly change clone sequencing strategies. On the other hand, expression host development will likely advance at a slower pace, with steady and likely labour-intensive work to generate or modify organisms to make them suitable for heterologous screening. Another issue to be addressed is that of sequence bias in clone libraries, particularly for libraries constructed using gut-derived DNA; factors contributing to library bias need to be better understood to inform strategies to address such bias. These methodological improvements will complement sequence-based metagenomics methods, providing basic knowledge about gene function as well as supporting applied work aimed at mining novel enzymes and engineering or modifying microbiomes.

7.3 Summary and claims of contributions to knowledge

This section briefly summarizes the broad goals of each of the four data chapters in this thesis, as well as lists my claims of contributions to scientific knowledge based on the results of each chapter.

Chapter 3: Evaluation of pooled sequencing for metagenomic clones

In Chapter 3, I presented the results of using a pooled method for sequencing large-insert cosmid clones isolated from functional screens of metagenomic libraries. Illumina sequence data from the pooled approach were evaluated against reference data obtained from barcoded sequencing of the same clones. The objective was to determine the extent to which the more cost-effective pooled sequencing strategy was capable of generating accurate and near-complete assemblies for the metagenomic inserts. My specific claims of contributions to knowledge are:

1. By comparison to the barcoded reference data, I showed that DNA sequence for large-insert metagenomic clones can be effectively recovered from a pooled short-read (75-base) sequencing approach.

2. I showed that two major factors affecting clone sequence recovery are sequencing depth and clone sequence similarity. In the first case, coverage of the clones can be improved by increasing the depth of sequencing to close any potential gaps; however, in the latter case, coverage may not improve for those clones in the pool that have high sequence similarity (but not identical) due to problems assembling the short reads.

Chapter 4: Analysis of cloning bias in metagenomic libraries

In Chapter 4, I presented the results of analyzing sequence bias in metagenomic libraries and exploring the possible causes of this bias during library construction. I did this by analyzing data obtained from sequencing the DNA at various points in the construction of a human fecal metagenomic library. The objective was to determine if DNA fragmentation was a major cause of cloning bias or alternatively, if events occurring in vivo in *E. coli* were a more important factor. My specific claims of contributions to knowledge are:

3. I showed that the low-copy cosmid-based human gut metagenomic library did suffer from cloning bias but that DNA fragmentation/size selection was not a major cause of this bias; rather, the bias appears to occur after introduction of the cloned DNA into *E. coli*.

4. By analyzing the sequence data for promoter consensus sequences, I provided support for the hypothesis that spurious transcription in *E. coli* may be a major cause of bias. I emphasized how this finding is in agreement with older published results which I found by careful examination of the scientific literature.

Chapter 5: Development of *B. theta* as a screening host

In Chapter 5, I presented the results of efforts to develop *Bacteroides thetaiotaomicron* as a host for screening human gut metagenomic libraries. Arguably the most important chapter of this thesis, it was also the most challenging. The objective was to construct a cloning vector able to replicate in *B. theta*, generate *B. theta*-compatible clone libraries using such a vector, and finally to demonstrate that constructed libraries can be successfully screened in a *B. theta* host. Positive clones isolated from a proof-

of-principle functional screen would support the notion of using *B. theta* as a host for screening gut-derived DNA. My specific claims of contributions to knowledge are:

5. I constructed a mobilizable *B. theta*-compatible fosmid vector, pKL13, and used this vector to construct a *B. theta* genomic library as well as a human gut metagenomic library. Both the vector and the libraries are resources that may be useful in future functional metagenomics work.

6. By introducing both libraries into a *B. theta* deletion mutant unable to grow on chondroitin sulfate as sole carbon source, I achieved complementation thereby demonstrating that it was possible to carry out functional screening in *B. theta*, particularly of a metagenomic library.

7. Although I found that fosmid clone DNA appeared to be integrated into the genome of *B. theta*, I was able to obtain and analyze partial DNA sequence data from the metagenomic clones that were able to complement the *B. theta* *chuR* mutant. Through this, I identified a *chuR* ORF that showed high sequence similarity to the VPI-5482 strain but was not found in the NCBI nr database, indicating that this is a novel *chuR* ORF.

Chapter 6: Inclusion of transcriptional terminators in cloning vectors

In Chapter 6, I presented the results of designing, cloning, and testing transcriptional terminators for a fosmid vector. The objective was to introduce elements to reduce insert-driven transcription into the vector backbone, as well as to make a terminator-containing construct general enough for introduction into other cloning vectors. My specific claims of contributions to knowledge are:

8. I incorporated two transcriptional terminators into the *B. theta*-compatible fosmid pKL13 that flank the site of large-insert cloning, and demonstrated their functionality in that context.

9. I generated constructs that will be useful for future experiments to examine whether the presence of transcriptional terminators will alleviate the cloning bias observed for metagenomic libraries.

7.4 Future directions and perspective

Function-based approaches are likely to be increasingly important as the fields of microbial ecology and metagenomics advance. The development and refinement of methods for functional metagenomics will be instrumental in this advancement [74]. The work described in this thesis was carried out towards this goal, although further work needs to be done to expand on the findings presented. Accordingly, there are several broad considerations discussed below that are relevant to method development for functional metagenomics.

Sequencing clones from metagenomic libraries

Although Chapter 3 was focused on a pooled-clone Illumina sequencing strategy and discussed the limitation of pooling clones for short-read sequencing, it is possible that short-read technologies will soon be obsolete. Within the last decade, there has been marked increase even in the length of reads obtained by Illumina (Solexa) instruments, from less than 50 bases on the Illumina GA II ten years ago to 2×300 bases on the Illumina MiSeq today. Although Illumina offers the lowest error rate among sequencing technologies currently in popular use, at $\leq$1% [245, 255], other sequencing technologies that are able to offer much longer read lengths may soon gain the advantage as they improve their error rates. For example, Pacific Biosciences sequencing can generate reads several-kb long on average, although the throughput and $\sim$15% error rate need to be improved for it to gain more widespread usage [76].

A particularly exciting long-read sequencing technology that is being developed comes from Oxford Nanopore Technologies, with a median length in the thousands of bases and upper-limit length of tens of thousands of bases [9]; the length obtained, however, depends on the quality of the input DNA, which offers the prospect of obtaining the entire DNA sequence of a typical fosmid insert in just a single read! Like PacBio sequencing, this technology is also limited by a high error rate, which is close to $\sim$30% [9] although a rate of 4% has been reported by the company [121]. The refinement and availability of affordable long-read sequencing technologies may soon obviate the need for the more difficult methods involved in short-read sequencing and assembly, particularly for clone pools.

Representativeness of metagenomic libraries

Though not so much a concern for functional screens, it is interesting to consider the factors that influence library representativeness; elucidating these factors may lead to the development of better strategies for accessing the full potential of environmental metagenomes. If spurious transcription does indeed contribute substantially to cloning bias, it would be worth investigating strategies to alleviate such transcription. For example, the use of transcriptional terminators has already been discussed in detail in Chapter 4 and Chapter 6; in the latter, I generated constructs containing terminators that flank the site of cloning, which can be introduced into different vectors for library construction and examination of bias (Section 6.4.4). It is important to note that for fosmid vectors, inserts may be very large and events occurring at the vector-insert junction may contribute to only a small fraction of the observed bias; on the other hand, these events may cause the whole insert to be lost.

For tackling potential transcription more globally, that is, across the entire cloned fragment, another possibility is based on the observation that *E. coli* H-NS (histone-like nucleoid structuring) protein binds AT-rich DNA, including sequences that may be recognized by the *E. coli* housekeeping sigma factor σ^{70} [168], silencing spurious transcription by RNA polymerase [273]. It is possible that increasing the cellular concentration of H-NS will suppress transcription from σ^{70} promoter-like sequences in cloned metagenomic DNA, thereby reducing transcriptional effects that may potentially lead to insert exclusion. The caveat of using H-NS, however, is that suppression of transcription may be undesirable if the host used for library construction is to be used directly for functional screening.

Appropriate hosts for functional screening

Depending on the target activity, functional screens can exhibit a low hit rate [312] the reasons for which might include barriers at the level of both transcription and translation. Improving *E. coli* as a screening host to address these problems will likely improve future hit rates. Examples include introducing heterologous sigma factors to guide RNA polymerase to otherwise untranscribed regions [98], employing T7 RNA polymerase to help drive transcription [305], as well as forming hybrid ribosomes [151] that may influence expression.

Nevertheless, it will be important to move beyond *E. coli* into different screening hosts, particularly for the complementation of mutant phenotypes not possible in *E. coli*, such as those of *B. theta* and other members of the Bacteroidetes described in Chapter 5. The future of functional metagenomics will likely see the development of a greater variety of alternative hosts for functional screening, which will not only likely lead to an increase in the hit rates of functional screens but also make available a broader range of phenotypes for functional complementation.

Functional metagenomics using a mouse model

An exciting avenue of research involves performing functional screens in vivo, that is, in a germ-free (gnotobiotic) mouse model, to explore how particular genes contribute to fitness in terms of host colonization or other effects on the host organism. This has already been demonstrated in principle using *E. coli* to screen a *B. theta* genomic library (in an expression vector) for fitness determinants in a mouse model [341]. Moving to a metagenomic library is the obvious next step [81]. A further exciting step would be to carry out functional screening in *B. theta* or another closely related host, should the development of such organisms for functional metagenomics be successful.

7.5 Concluding remarks

Method development is and will continue to be important in the functional metagenomics field, particularly as (1) interest in the human microbiome drives research into characterizing microbial gene function and understanding the mechanisms that lead to effects on the host organism, and (2) knowledge of gene function is required to complement sequence-based metagenomics research. The identification of obstacles to cloning and screening will aid in the development of new tools and technologies for functional metagenomics, providing us with greater reach in terms of what we are able to gather from functional screens. Refining function-based methods will be crucial for the bioprospecting of novel enzymes and compounds, for the determination of gene function to guide the development of reliable models of microbial ecosystem functioning, and to support efforts in microbiome engineering and development of therapeutics.

Bibliography

[1] Aakvik, T., Degnes, K. F., Dahlsrud, R., Schmidt, F., Dam, R., Yu, L., Völker, U., Ellingsen, T. E., and Valla, S. A plasmid RK2-based broad-host-range cloning vector useful for transfer of metagenomic libraries to a variety of bacterial species. *FEMS Microbiology Letters 296* (2009), 149–158. [Page 21], [Page 222], [Page 224], [Page 225]

[2] Alivisatos, A. P., Blaser, M. J., Brodie, E. L., Chun, M., Dangl, J. L., Donohue, T. J., Dorrestein, P. C., Gilbert, J. A., Green, J. L., Jansson, J. K., Knight, R., Maxon, M. E., McFall-Ngai, M. J., Miller, J. F., Pollard, K. S., Ruby, E. G., Taha, S. A., and Consortium, U. M. I. A unified initiative to harness Earth's microbiomes. *Science 350* (2015), 507–508. [Page 12], [Page 13], [Page 14]

[3] Allen, H. K., Moe, L. A., Rodbumrer, J., Gaarder, A., and Handelsman, J. Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil. *The ISME Journal 3* (2009), 243–251. [Page 62], [Page 222]

[4] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology 215* (1990), 403–410. [Page 67], [Page 93], [Page 95]

[5] Anderson, K. L., and Salyers, A. A. Biochemical evidence that starch breakdown by *Bacteroides thetaiotaomicron* involves outer membrane starch-binding sites and periplasmic starch-degrading enzymes. *Journal of Bacteriology 171* (1989), 3192–3198. [Page 142]

[6] Anderson, K. L., and Salyers, A. A. Genetic evidence that outer membrane binding of starch is required for starch utilization by *Bacteroides thetaiotaomicron*. *Journal of Bacteriology 171* (1989), 3199–3204. [Page 142]

[7] Angelov, A., Mientus, M., Liebl, S., and Liebl, W. A two-host fosmid system for functional screening of (meta)genomic libraries from extreme thermophiles. *Systematic and Applied Microbiology 32* (2009), 177–185. [Page 226]

[8] Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H. B., Nielsen, T., Pons, N., Poulain,

J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E. G., Wang, J., Guarner, F., Pedersen, O., de Vos, W. M., Brunak, S., Doré, J., Consortium, M., Weissenbach, J., Ehrlich, S. D., Bork, P., Antolín, M., Artiguenave, F., Blottiere, H. M., Almeida, M., Brechot, C., Cara, C., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Foerstner, K. U., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Huber, W., van Hylckama-Vlieg, J., Jamet, A., Juste, C., Kaci, G., Knol, J., Lakhdari, O., Layec, S., Le Roux, K., Maguin, E., Mérieux, A., Melo Minardi, R., M'rini, C., Muller, J., Oozeer, R., Parkhill, J., Renault, P., Rescigno, M., Sanchez, N., Sunagawa, S., Torrejon, A., Turner, K., Vandemeulebrouck, G., Varela, E., Winogradsky, Y., and Zeller, G. Enterotypes of the human gut microbiome. *Nature 473* (2011), 174–180. [Page 7], [Page 190]

[9] Ashton, P. M., Nair, S., Dallman, T., Rubino, S., Rabsch, W., Mwaigwisya, S., Wain, J., and O'Grady, J. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature Biotechnology 33* (2014), 296–300. [Page 258]

[10] Ayres, E. K., Thomson, V. J., Merino, G., Balderes, D., and Figurski, D. H. Precise deletions in large bacterial genomes by vector-mediated excision (VEX): the *trfA* gene of promiscuous plasmid RK2 is essential for replication in several Gram-negative hosts. *Journal of Molecular Biology 230* (1993), 174–185. [Page 223]

[11] Bäckhed, F., Ding, H., Wang, T., Hooper, L. V., Koh, G. Y., Nagy, A., Semenkovich, C. F., and Gordon, J. I. The gut microbiota as an environmental factor that regulates fat storage. *Proceedings of the National Academy of Sciences of the United States of America 101* (2004), 15718–15723. [Page 8]

[12] Backhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A., and Gordon, J. I. Host-bacterial mutualism in the human intestine. *Science 307* (2005), 1915–1920. [Page 5], [Page 6], [Page 8], [Page 140], [Page 141]

[13] Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., Kimelman, M., Pruitt, K. D., Resenchuk, S., Tatusova, T., Yaschenko, E., and Ostell, J. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Research 40* (2012), D57–D63. [Page 86]

[14] Bartram, A. K., Lynch, M. D. J., Stearns, J. C., Moreno-Hagelsieb, G., and Neufeld, J. D. Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. *Applied and Environmental Microbiology 77* (2011), 3846–3852. [Page 134]

[15] Bayley, D. P., Rocha, E. R., and Smith, C. J. Analysis of *cepA* and other *Bacteroides fragilis* genes reveals a unique promoter structure. *FEMS Microbiology Letters 193* (2000), 149–154. [Page 116], [Page 147], [Page 153]

[16] Bazzani, R. P., Cai, Y., Hebel, H. L., Hyde, S. C., and Gill, D. R. The significance of plasmid DNA preparations contaminated with bacterial genomic DNA on

inflammatory responses following delivery of lipoplexes to the murine lung. *Biomaterials 32* (2011), 9854–9865. [Page 82]

[17] Benjdia, A., Martens, E. C., Gordon, J. I., and Berteau, O. Sulfatases and a radical S-adenosyl-L-methionine (AdoMet) enzyme are key for mucosal foraging and fitness of the prominent human gut symbiont, *Bacteroides thetaiotaomicron*. *Journal of Biological Chemistry 286* (2011), 25973–25982. [Page 27], [Page 137], [Page 180], [Page 181]

[18] Benjdia, A., Subramanian, S., Leprince, J., Vaudry, H., Johnson, M. K., and Berteau, O. Anaerobic sulfatase-maturating enzymes, first dual substrate radical S-adenosylmethionine enzymes. *Journal of Biological Chemistry 283* (2008), 17815–17826. [Page 181], [Page 192]

[19] Berer, K., Mues, M., Koutrolos, M., Rasbi, Z. A., Boziki, M., Johner, C., Wekerle, H., and Krishnamoorthy, G. Commensal microbiota and myelin autoantigen cooperate to trigger autoimmune demyelination. *Nature 479* (2011), 538–541. [Page 11]

[20] Berlemont, R., Delsaute, M., Pipers, D., D'Amico, S., Feller, G., Galleni, M., and Power, P. Insights into bacterial cellulose biosynthesis by functional metagenomics on Antarctic soil samples. *The ISME Journal 3* (2009), 1070–1081. [Page 62]

[21] Bethesda Research Laboratories. BRL pUC host: *E. coli* DH5α competent cells. *Focus 8* (1986), 8. [Page 27]

[22] Bik, E. M., Eckburg, P. B., Gill, S. R., Nelson, K. E., Purdom, E. A., Francois, F., Perez-Perez, G., Blaser, M. J., and Relman, D. A. Molecular analysis of the bacterial microbiota in the human stomach. *Proceedings of the National Academy of Sciences of the United States of America 103* (2006), 732–737. [Page 4]

[23] Blaser, M. J. *Missing microbes: how the overuse of antibiotics is fueling our modern plagues*. HarperCollins Publishers Ltd, 2014. [Page 12]

[24] Böhnke, S., and Perner, M. A function-based screen for seeking RubisCO active clones from metagenomes: novel enzymes influencing RubisCO activity. *The ISME Journal 9* (2015), 735–745. [Page 222]

[25] Boyer, H. W., and Roulland-Dussoix, D. A complementation analysis of the restriction and modification of DNA in *Escherichia coli*. *Journal of Molecular Biology 41* (1969), 459–472. [Page 27]

[26] Brady, S. F. Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. *Nature Protocols 2* (2007), 1297–1305. [Page 218], [Page 219]

[27] Brayton, K., De Villiers, E., Fehrsen, J., Nxomani, C., Collins, N., and Allsopp, B. Cowdria ruminantium DNA is unstable in a SuperCos1 library. *Onderstepoort Journal of Veterinary Research 117* (1999), 111–117. [Page 121]

[28] Brown, N. P., Leroy, C., and Sander, C. MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics 14* (1998), 380–381. [Page 189], [Page 191]

[29] Buck, J. D. *Physiological effects of heterologous expression of proteorhodopsin photosystems*. PhD thesis, Massachusetts Institute of Technology, 2012. [Page 224]

[30] Cai, W. W., Chen, R., Gibbs, R. A., and Bradley, A. A clone-array pooled shotgun strategy for sequencing large genomes. *Genome Research 11* (2001), 1619–1623. [Page 83]

[31] Cambray, G., Guimaraes, J. C., Mutalik, V. K., Lam, C., Mai, Q.-A., Thimmaiah, T., Carothers, J. M., Arkin, A. P., and Endy, D. Measurement and modeling of intrinsic transcription terminators. *Nucleic Acids Research 41* (2013), 5139–5148. [Page 228], [Page 229], [Page 230], [Page 233], [Page 239]

[32] Cameron, E. A., Maynard, M. A., Smith, C. J., Smith, T. J., Koropatkin, N. M., and Martens, E. C. Multidomain carbohydrate-binding proteins involved in *Bacteroides thetaiotaomicron* starch metabolism. *Journal of Biological Chemistry 287* (2012), 34614–34625. [Page 143], [Page 147]

[33] Case, S. T. Selective deletion of large segments of Balbiani ring DNA during molecular cloning. *Gene 20* (1982), 169–176. [Page 122]

[34] Cecchini, D. A., Laville, E., Laguerre, S., Robe, P., Leclerc, M., Doré, J., Henrissat, B., Remaud-Siméon, M., Monsan, P., and Potocki-Véronèse, G. Functional metagenomics reveals novel pathways of prebiotic breakdown by human gut bacteria. *PLOS ONE 8* (2013), e72766. [Page 62]

[35] Charles, T. C. *Construction of a genetic linkage map of the Rhizobium meliloti 1600 kilobase megaplasmid pRmeSU47b, generation of defined megaplasmid deletions, and study of megaplasmid-borne genes*. PhD thesis, McMaster University, 1990. [Page 40]

[36] Charles, T. C., and Nester, E. W. A chromosomally encoded two-component sensory transduction system is required for virulence of *Agrobacterium tumefaciens*. *Journal of Bacteriology 175* (1993), 6614–6625. [Page 43], [Page 212]

[37] Charles, T. C., and Neufeld, J. D. Open resource metagenomics. In *Encyclopedia of Metagenomics*, K. E. Nelson, Ed. Springer New York, New York, 2015, pp. 573–575. [Page 220]

[38] Chatzidaki-Livanis, M., Coyne, M. J., Roche-Hakansson, H., and Comstock, L. E. Expression of a uniquely regulated extracellular polysaccharide confers a large-capsule phenotype to *Bacteroides fragilis*. *Journal of Bacteriology 190* (2008), 1020–1026. [Page 147]

[39] Chen, J., Kadlubar, F. F., and Chen, J. Z. DNA supercoiling suppresses real-time PCR: a new approach to the quantification of mitochondrial DNA damage and repair. *Nucleic Acids Research 35* (2007), 1377–1388. [Page 124]

[40] Chen, J.-D., and Morrison, D. A. Cloning of *Streptococcus pneumoniae* DNA fragments in Escherichia coli requires vectors protected by strong transcriptional terminators. *Gene 55* (1987), 179–187. [Page 122]

[41] Chen, J.-D., and Morrison, D. A. Construction and properties of a new insertion vector, pJDC9, that is protected by transcriptional terminators and useful for cloning of DNA from *Streptococcus pneumoniae*. *Gene 64* (1988), 155–164. [Page 102], [Page 122]

[42] Chen, Y.-J., Liu, P., Nielsen, A. A. K., Brophy, J. A. N., Clancy, K., Peterson, T., and Voigt, C. A. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nature Methods 10* (2013), 659–664. [Page 228]

[43] Cheng, J., Pinnell, L., Engel, K., Neufeld, J. D., and Charles, T. C. Versatile broad-host-range cosmids for construction of high quality metagenomic libraries. *Journal of Microbiological Methods 99* (2014), 27–34. [Page 30], [Page 36], [Page 86], [Page 87], [Page 123], [Page 125], [Page 134], [Page 135], [Page 219], [Page 220]

[44] Cheng, Q., Hwa, V., and Salyers, A. A. A locus that contributes to colonization of the intestinal tract by *Bacteroides thetaiotaomicron* contains a single regulatory gene (chuR) that links two polysaccharide utilization pathways. *Journal of Bacteriology 174* (1992), 7185–7193. [Page 180]

[45] Cho, K. H., and Salyers, A. A. Biochemical Analysis of Interactions between Outer Membrane Proteins That Contribute to Starch Utilization by *Bacteroides thetaiotaomicron*. *Journal of Bacteriology 183* (2001), 7224–7230. [Page 143]

[46] Clayton, T. A., Baker, D., Lindon, J. C., Everett, J. R., and Nicholson, J. K. Pharmacometabonomic identification of a significant host-microbiome metabolic interaction affecting human drug metabolism. *Proceedings of the National Academy of Sciences of the United States of America 106* (2009), 14728–14733. [Page 9]

[47] Colin, P.-Y., Kintses, B., Gielen, F., Miton, C. M., Fischer, G., Mohamed, M. F., Hyvönen, M., Morgavi, D. P., Janssen, D. B., and Hollfelder, F. Ultrahigh-throughput discovery of promiscuous enzymes by picodroplet functional metagenomics. *Nature Communications 6* (2015), 10008. [Page 21]

[48] Comstock, L. E. Importance of glycans to the host-*Bacteroides* mutualism in the mammalian intestine. *Cell Host and Microbe 5* (2009), 522–526. [Page 141]

[49] Cooper, A. J., Kalinowski, A. P., Shoemaker, N. B., and Salyers, A. A. Construction and characterization of a *Bacteroides thetaiotaomicron recA* mutant: transfer of *Bacteroides* integrated conjugative elements is RecA independent. *Journal of Bacteriology 179* (1997), 6221–6227. [Page 157], [Page 192]

[50] Craig, J. W., Chang, F.-Y., Kim, J. H., Obiajulu, S. C., and Brady, S. F. Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse Proteobacteria. *Applied and Environmental Microbiology 76* (2010), 1633–1641. [Page 21], [Page 220], [Page 225]

[51] Crameri, A., Whitehorn, E. A., Tate, E., and Stemmer, W. P. Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nature Biotechnology 14* (1996), 315–319. [Page 239]

[52] Culligan, E. P., Sleator, R. D., Marchesi, J. R., and Hill, C. Functional metagenomics reveals novel salt tolerance loci from the human gut microbiome. *The ISME Journal 6* (2012), 1916–1925. [Page 62]

[53] Curtis, M. M., Hu, Z., Klimko, C., Narayanan, S., Deberardinis, R., and Sperandio, V. The gut commensal *Bacteroides thetaiotaomicron* exacerbates enteric infection through modification of the metabolic landscape. *Cell Host & Microbe 16* (2014), 759–769. [Page 10]

[54] Danhorn, T., Young, C. R., and DeLong, E. F. Comparison of large-insert, small-insert and pyrosequencing libraries for metagenomic analysis. *The ISME Journal 6* (2012), 2056–2066. [Page 102], [Page 121]

[55] David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, a. S., Varma, Y., Fischbach, M. a., Biddinger, S. B., Dutton, R. J., and Turnbaugh, P. J. Diet rapidly and reproducibly alters the human gut microbiome. *Nature 505* (2013), 559–563. [Page 7]

[56] Davies, G. J., Gloster, T. M., and Henrissat, B. Recent structural insights into the expanding world of carbohydrate-active enzymes. *Current Opinion in Structural Biology 15* (2005), 637–645. [Page 140]

[57] D'Elia, J. N., and Salyers, A. A. Contribution of a neopullulanase, a pullulanase, and an α-glucosidase to growth of *Bacteroides thetaiotaomicron* on starch. *Journal of Bacteriology 178* (1996), 7173–7179. [Page 143]

[58] D'Elia, J. N., and Salyers, A. A. Effect of regulatory protein levels on utilization of starch by *Bacteroides thetaiotaomicron*. *Journal of Bacteriology 178* (1996), 7180–7186. [Page 143]

[59] DeLong, E. F., Preston, C. M., Mincer, T., Rich, V., Hallam, S. J., Frigaard, N.-U., Martinez, A., Sullivan, M. B., Edwards, R., Brito, B. R., Chisholm, S. W., and Karl, D. M. Community genomics among stratified microbial assemblages in the ocean's interior. *Science 311* (2006), 496–503. [Page 222]

[60] Dewhirst, F. E., Chen, T., Izard, J., Paster, B. J., Tanner, A. C. R., Yu, W.-h., Lakshmanan, A., and Wade, W. G. The human oral microbiome. *Journal of Bacteriology 192* (2010), 5002–5017. [Page 4]

[61] Dillard, J. P., and Yother, J. Analysis of *Streptococcus pneumoniae* sequences cloned into *Escherichia coli*: effect of promoter strength and transcription terminators. *Journal of Bacteriology 173* (1991), 5105–5109. [Page 122]

[62] Dinsdale, E. A., Edwards, R. A., Hall, D., Angly, F., Breitbart, M., Brulc, J. M., Furlan, M., Desnues, C., Haynes, M., Li, L., McDaniel, L., Moran, M. A., Nelson, K. E., Nilsson, C., Olson, R., Paul, J., Brito, B. R., Ruan,

Y., Swan, B. K., Stevens, R., Valentine, D. L., Thurber, R. V., Wegley, L., White, B. A., and Rohwer, F. Functional metagenomic profiling of nine biomes. *Nature 452* (2008), 629–632. [Page 15]

[63] Ditta, G., Schmidhauser, T., Yakobson, E., Lu, P., Liang, X. W., Finlay, D. R., Guiney, D., and Helinski, D. R. Plasmids related to the broad host range vector, pRK290, useful for gene cloning and for monitoring gene expression. *Plasmid 13* (1985), 149–153. [Page 90]

[64] Djordjevic, G., Bojovic, B., Banina, A., and Topisirovic, L. Cloning of promoter-like sequences from Lactobacillus paracasei subsp. paracasei CG11 and their expression in *Escherichia coli, Lactococcus lactis,* and *Lactobacillus reuteri*. *Canadian Journal of Microbiology 40* (1994), 1043–1050. [Page 115]

[65] Donato, J. J., Moe, L. A., Converse, B. J., Smart, K. D., Berklein, F. C., McManus, P. S., and Handelsman, J. Metagenomic analysis of apple orchard soil reveals antibiotic resistance genes encoding predicted bifunctional proteins. *Applied and Environmental Microbiology 76* (2010), 4396–4401. [Page 223]

[66] Donia, M. S., and Fischbach, M. A. Small molecules from the human microbiota. *Science 349* (2015), 1254766–1254766. [Page 9]

[67] Donohoe, D. R., Garge, N., Zhang, X., Sun, W., O'Connell, T. M., Bunger, M. K., and Bultman, S. J. The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon. *Cell Metabolism 13* (2011), 517–526. [Page 9], [Page 141]

[68] Durland, R. H., and Helinski, D. R. Replication of the broad-host-range plasmid RK2: Direct measurement of intracellular concentrations of the essential TrfA replication proteins and their effect on plasmid copy number. *Journal of Bacteriology 172* (1990), 3849–3858. [Page 224]

[69] Džunková, M., D'Auria, G., Pérez-Villarroya, D., and Moya, A. Hybrid sequencing approach applied to human fecal metagenomic clone libraries revealed clones with potential biotechnological applications. *PLOS ONE 7* (2012), e47654. [Page 63]

[70] Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics 5* (2004), 113. [Page 189], [Page 191]

[71] Ekkers, D. M., Cretoiu, M. S., Kielak, A. M., and van Elsas, J. D. The great screen anomaly—a new frontier in product discovery through functional metagenomics. *Applied Microbiology and Biotechnology 93* (2012), 1005–1020. [Page 21], [Page 225]

[72] El Kaoutari, A., Armougom, F., Gordon, J. I., Raoult, D., and Henrissat, B. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nature Reviews Microbiology 11* (2013), 497–504. [Page 8], [Page 111], [Page 141], [Page 142]

[73] Elend, C., Schmeisser, C., Leggewie, C., Babiak, P., Steele, H. L., Reymond, J., Jaeger, K., Streit, R., Carballeira, J. D., and Streit, W. R.

Isolation and biochemical characterization of two novel metagenome-derived esterases. *Applied and Environmental Microbiology 72* (2006), 3637–3645. [Page 62]

[74] Engel, K., Ashby, D., Brady, S. F., Cowan, D. A., Doemer, J., Edwards, E. A., Fiebig, K., Martens, E. C., McCormac, D., Mead, D. A., Miyazaki, K., Moreno-Hagelsieb, G., O'Gara, F., Reid, A., Rose, D. R., Simonet, P., Sjöling, S., Smalla, K., Streit, W. R., Tedman-Jones, J., Valla, S., Wellington, E. M. H., Wu, C.-C., Liles, M. R., Neufeld, J. D., Sessitsch, A., and Charles, T. C. Meeting report: 1st International Functional Metagenomics Workshop May 7-8, 2012, St. Jacobs, Ontario, Canada. *Standards in Genomic Sciences 8* (2013), 106–111. [Page 257]

[75] Engel, K., Pinnell, L., Cheng, J., Charles, T. C., and Neufeld, J. D. Nonlinear electrophoresis for purification of soil DNA for metagenomics. *Journal of Microbiological Methods 88* (2012), 35–40. [Page 85], [Page 130], [Page 134], [Page 219]

[76] English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D. M., Reid, J. G., Worley, K. C., and Gibbs, R. A. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLOS ONE 7* (2012), e47768. [Page 258]

[77] Epicentre. Product manual for CopyControl™ fosmid library production kit with pCC1FOS™ vector, 2015. [Page 242]

[78] Erickson, A. R., Cantarel, B. L., Lamendella, R., Darzi, Y., Mongodin, E. F., Pan, C., Shah, M., Halfvarson, J., Tysk, C., Henrissat, B., Raes, J., Verberkmoes, N. C., Fraser, C. M., Hettich, R. L., and Jansson, J. K. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLOS ONE 7* (2012), e49138. [Page 11]

[79] Erlich, Y., Chang, K., Gordon, A., Ronen, R., Navon, O., Rooks, M., and Hannon, G. J. DNA Sudoku–harnessing high-throughput sequencing for multiplexed specimen analysis. *Genome Research 19* (2009), 1243–1253. [Page 83]

[80] Esko, J. D., Kimata, K., and Lindahl, U. Proteoglycans and sulfated glycosaminoglycans. In *Essentials of Glycobiology*, A. Varki, R. D. Cummings, J. D. Esko, H. H. Freeze, P. Stanley, C. R. Bertozzi, G. W. Hart, and M. E. Etzler, Eds., 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2009, ch. 16. [Page 181]

[81] Faith, J. J. Bridging the knowledge gap: from microbiome composition to function. *Molecular Systems Biology 11* (2015), 793. [Page 151], [Page 260]

[82] Faith, J. J., Ahern, P. P., Ridaura, V. K., Cheng, J., and Gordon, J. I. Identifying gut microbe-host phenotype relationships using combinatorial communities in gnotobiotic mice. *Science Translational Medicine 6* (2014), 220ra11. [Page 13]

[83] Faith, J. J., Colombel, J.-f., and Gordon, J. I. Identifying strains that contribute to complex diseases through the study of microbial inheritance. *Proceedings of the National Academy of Sciences of the United States of America 112* (2015), 633–640. [Page 11]

[84] Feingersch, R., and Béjà, O. Bias in assessments of marine SAR11 biodiversity in environmental fosmid and BAC libraries? *The ISME Journal 3* (2009), 1117–1119. [Page 102], [Page 123]

[85] Feng, Y., Duan, C.-J., Pang, H., Mo, X.-C., Wu, C.-F., Yu, Y., Hu, Y.-L., Wei, J., Tang, J.-L., and Feng, J.-X. Cloning and identification of novel cellulase genes from uncultured microorganisms in rabbit cecum and characterization of the expressed cellulases. *Applied Microbiology and Biotechnology 75* (2007), 319–328. [Page 62]

[86] Ferrer, M., Golyshina, O. V., Chernikova, T. N., Khachane, A. N., Reyes-Duarte, D., Santos, V. a. P. M. D., Strompl, C., Elborough, K., Jarvis, G., Neef, A., Yakimov, M. M., Timmis, K. N., and Golyshin, P. N. Novel hydrolase diversity retrieved from a metagenome library of bovine rumen microflora. *Environmental Microbiology 7* (2005), 1996–2010. [Page 61]

[87] Ferrés, I., Amarelle, V., Noya, F., and Fabiano, E. Construction and screening of a functional metagenomic library to identify novel enzymes produced by Antarctic bacteria. *Advances in Polar Science 26* (2015), 96–101. [Page 222]

[88] Ferrieres, L., Hemery, G., Nham, T., Guerout, A.-M., Mazel, D., Beloin, C., and Ghigo, J.-M. Silent mischief: bacteriophage Mu insertions contaminate products of *Escherichia coli* random mutagenesis performed using suicidal transposon delivery plasmids mobilized by broad-host-range RP4 conjugative machinery. *Journal of Bacteriology 192* (2010), 6418–6427. [Page 27]

[89] Finan, T. M., Kunkel, B., De Vos, G. F., and Signer, E. R. Second symbiotic megaplasmid in *Rhizobium meliloti* carrying exopolysaccharide and thiamine synthesis genes. *Journal of bacteriology 167* (1986), 66–72. [Page 30]

[90] Forslund, K., Hildebrand, F., Nielsen, T., Falony, G., Le Chatelier, E., Sunagawa, S., Prifti, E., Vieira-Silva, S., Gudmundsdottir, V., Krogh Pedersen, H., Arumugam, M., Kristiansen, K., Yvonne Voigt, A., Vestergaard, H., Hercog, R., Igor Costea, P., Roat Kultima, J., Li, J., Jørgensen, T., Levenez, F., Dore, J., consortium, M., Bjørn Nielsen, H., Brunak, S., Raes, J., Hansen, T., Wang, J., Dusko Ehrlich, S., Bork, P., and Pedersen, O. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature 528* (2015), 262–266. [Page 11]

[91] Forsythe, P., Sudo, N., Dinan, T., Taylor, V. H., and Bienenstock, J. Mood and gut feelings. *Brain, Behavior, and Immunity 24* (2010), 9–16. [Page 11]

[92] Frank, D. N., St Amand, A. L., Feldman, R. A., Boedeker, E. C., Harpaz, N., and Pace, N. R. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences of the United States of America 104* (2007), 13780–13785. [Page 11]

[93] Franzosa, E. a., Huang, K., Meadow, J. F., Gevers, D., Lemon, K. P., Bohannan, B. J. M., and Huttenhower, C. Identifying personal microbiomes using metagenomic codes. *Proceedings of the National Academy of Sciences 112* (2015), E2930–E2938. [Page 7]

[94] Friedman, A. M., Long, S. R., Brown, S. E., Buikema, W. J., and Ausubel, F. M. Construction of a broad host range vector and its use in the genetic analysis of Rhizobium mutants. *Gene 18* (1982), 289–296. [Page 177]

[95] Friehs, K. Plasmid copy number and plasmid stability. *Advances in Biochemical Engineering / Biotechnology 86* (2004), 47–82. [Page 242]

[96] Fujimura, K. E., Demoor, T., Rauch, M., Faruqi, A. A., Jang, S., C. Johnson, C., Boushey, H. A., Zoratti, E., Ownby, D., Lukacs, N. W., and Lynch, S. V. House dust exposure mediates gut microbiome *Lactobacillus* enrichment and airway immune defense against allergens and virus infection. *Proceedings of the National Academy of Sciences 111* (2013), 805–810. [Page 11]

[97] Gabor, E. M., Alkema, W. B. L., and Janssen, D. B. Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environmental Microbiology 6* (2004), 879–886. [Page 18], [Page 111]

[98] Gaida, S. M., Sandoval, N. R., Nicolaou, S. A., Chen, Y., Venkataramanan, K. P., and Papoutsakis, E. T. Expression of heterologous sigma factors enables functional screening of metagenomic and heterologous genomic libraries. *Nature Communications 6* (2015), 7045. [Page 260]

[99] Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A., Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M.-S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M. W., Vaidya, A. B., Martin, D. M. A., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S. A., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Newbold, C., Davis, R. W., Fraser, C. M., and Barrell, B. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature 419* (2002), 498–511. [Page 121]

[100] Gentz, R., Langner, A., Chang, A. C. Y., Cohent, S. N., and Bujard, H. Cloning and analysis of strong promoters is made possible by the downstream placement of a RNA termination signal. *Proceedings of the National Academy of Sciences of the United States of America 78* (1981), 4936–4940. [Page 122]

[101] Ghai, R., Martin-Cuadrado, A.-B., Molto, A. G., Heredia, I. G., Cabrera, R., Martin, J., Verdú, M., Deschamps, P., Moreira, D., López-García, P., Mira, A., and Rodriguez-Valera, F. Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. *The ISME Journal 4* (2010), 1154–1166. [Page 102]

[102] Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., and Nelson, K. E. Metagenomic analysis of the human distal gut microbiome. *Science 312* (2006), 1355–1359. [Page 4], [Page 5]

[103] Gilmore, M. S., and Ferretti, J. J. The thin line between gut commensal and pathogen. *Science 299* (2003), 1999–2002. [Page 10]

[104] Giongo, A., Gano, K. A., Crabb, D. B., Mukherjee, N., Novelo, L. L., Casella, G., Drew, J. C., Ilonen, J., Knip, M., Hyöty, H., Veijola, R., Simell, T., Simell, O., Neu, J., Wasserfall, C. H., Schatz, D., Atkinson, M. A., and Triplett, E. W. Toward defining the autoimmune microbiome for type 1 diabetes. *The ISME Journal 5* (2011), 82–91. [Page 10]

[105] Gloux, K., Berteau, O., El Oumami, H., Béguet, F., Leclerc, M., and Doré, J. A metagenomic β-glucuronidase uncovers a core adaptive function of the human intestinal microbiome. *Proceedings of the National Academy of Sciences of the United States of America 108* (2010), 4539–4546. [Page 62]

[106] Godiska, R., Mead, D., Dhodda, V., Wu, C., Hochstein, R., Karsi, A., Usdin, K., Entezam, A., and Ravin, N. Linear plasmid vector for cloning of repetitive or unstable sequences in *Escherichia coli*. *Nucleic Acids Research 38* (2010), e88. [Page 121], [Page 123], [Page 127], [Page 228]

[107] Godiska, R., Patterson, M., Schoenfeld, T., and Mead, D. A. Beyond pUC: vectors for cloning unstable DNA. In *Optimization of the DNA Sequencing Process*, J. Kieleczawa, Ed. Jones and Bartlett Publishers, Sudbury, Massachusetts, 2005, pp. 55–75. [Page 123]

[108] Gong, X., Gruninger, R. J., Qi, M., Paterson, L., Forster, R. J., Teather, R. M., and McAllister, T. A. Cloning and identification of novel hydrolase genes from a dairy cow rumen metagenomic library and characterization of a cellulase gene. *BMC Research Notes 5* (2012), 566. [Page 21], [Page 62], [Page 154]

[109] Goodman, A. L., Kallstrom, G., Faith, J. J., Reyes, A., Moore, A., Dantas, G., and Gordon, J. I. Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proceedings of the National Academy of Sciences of the United States of America 108* (2011), 6252–6257. [Page 4], [Page 13]

[110] Gori, F., Tringe, S. G., Folino, G., van Hijum, S. A. F. T., Op den Camp, H. J. M., Jetten, M. S. M., and Marchiori, E. Differences in sequencing technologies improve the retrieval of anammox bacterial genome from metagenomes. *BMC Genomics 14* (2013), 7. [Page 103]

[111] Gründemann, D., and Schömig, E. Protection of DNA during preparative agarose gel electrophoresis against damage induced by ultraviolet light. *BioTechniques 21* (1996), 898–903. [Page 45]

[112] Guiney, D. G., Bouic, K., Hasegawa, P., and Matthews, B. Construction of shuttle cloning vectors for *Bacteroides fragilis* and use in assaying foreign tetracycline resistance gene expression. *Plasmid 20* (1988), 17–22. [Page 146]

[113] Guiney, D. G., and Yakobson, E. Location and nucleotide sequence of the transfer origin of the broad host range plasmid RK2. *Proceedings of the National Academy of Sciences of the United States of America 80* (1983), 3595–3598. [Page 169]

[114] Gusarov, I., and Nudler, E. The mechanism of intrinsic transcription termination. *Molecular Cell 3* (1999), 495–504. [Page 229]

[115] Gust, B., Challis, G. L., Fowler, K., Kieser, T., and Chater, K. F. PCR-targeted *Streptomyces* gene replacement identifies a protein domain needed for biosynthesis of the sesquiterpene soil odor geosmin. *Proceedings of the National Academy of Sciences of the United States of America 100* (2003), 1541–1546. [Page 224]

[116] Haiser, H. J., Gootenberg, D. B., Chatman, K., Sirasani, G., Balskus, E. P., and Turnbaugh, P. J. Predicting and manipulating cardiac drug inactivation by the human gut bacterium *Eggerthella lenta*. *Science 341* (2013), 295–298. [Page 10]

[117] Handelsman, J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews 68* (2004), 669–685. [Page 61]

[118] Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology 5* (1998), R245–R249. [Page 4]

[119] Hao, Y., Winans, S. C., Glick, B. R., and Charles, T. C. Identification and characterization of new LuxR/LuxI-type quorum sensing systems from metagenomic libraries. *Environmental Microbiology 12* (2010), 105–117. [Page 19], [Page 62]

[120] Hattori, Y., Omori, H., Hanyu, M., Kaseda, N., Mishima, E., Kaneko, T., Tabata, S., and Saeki, K. Ordered cosmid library of the *Mesorhizobium loti* MAFF303099 genome for systematic gene disruption and complementation analysis. *Plant and Cell Physiology 43* (2002), 1542–1557. [Page 121]

[121] Hayden, E. C. Nanopore genome sequencer makes its debut. *Nature* (2012). [Page 258]

[122] Heil, J. *Bacterial chromosome engineering for applications in metabolic engineering*. PhD thesis, University of Waterloo, 2015. [Page 226]

[123] Heil, J. R., Cheng, J., and Charles, T. C. Site-specific bacterial chromosome engineering: ΦC31 integrase mediated cassette exchange (IMCE). *Journal of Visualized Experiments* (2012), e3698. [Page 226]

[124] Helinski, D. R. Replication of an origin-containing derivative of plasmid RK2 dependent on a plasmid function provided in trans. *Proceedings of the National Academy of Sciences of the United States of America 76* (1979), 1648–1652. [Page 30]

[125] Herscovitch, M., Perkins, E., Baltus, A., and Fan, M. Addgene provides an open forum for plasmid sharing. *Nature Biotechnology 30* (2012), 316–317. [Page 220]

[126] Hofstad, T., and Kristoffersen, T. Chemical characteristics of endotoxin from *Bacteroides fragilis* NCTC 9343. *Journal of General Microbiology 61* (1970), 15–19. [Page 27]

[127] Hohn, B., and Collins, J. A small cosmid for efficient cloning of large DNA fragments. *Gene 11* (1980), 291–298. [Page 30]

[128] Holdeman, L., Cato, E., and Moore, W. *Anaerobe laboratory manual*. Virginia Polytechnic Institute and State University Anaerobe Laboratory, Blacksburg, Va, 1977. [Page 199]

[129] Holmes, E., Loo, R. L., Stamler, J., Bictash, M., Yap, I. K. S., Chan, Q., Ebbels, T., De Iorio, M., Brown, I. J., Veselkov, K. A., Daviglus, M. L., Kesteloot, H., Ueshima, H., Zhao, L., Nicholson, J. K., and Elliott, P. Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature 453* (2008), 396–400. [Page 11]

[130] Hooper, L. V., Midtvedt, T., and Gordon, J. I. How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annual Review of Nutrition 22* (2002), 283–307. [Page 9]

[131] Hsiao, E. Y., McBride, S. W., Hsien, S., Sharon, G., Hyde, E. R., McCue, T., Codelli, J. a., Chow, J., Reisman, S. E., Petrosino, J. F., Patterson, P. H., and Mazmanian, S. K. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell 155* (2013), 1451–1463. [Page 11]

[132] Huson, D. H., Mitra, S., Ruscheweyh, H.-J., Weber, N., and Schuster, S. C. Integrative analysis of environmental sequences using MEGAN4. *Genome Research 21* (2011), 1552–1560. [Page 131]

[133] Ichimura, M., Nakayama-Imaohji, H., Wakimoto, S., Morita, H., Hayashi, T., and Kuwahara, T. Efcient electrotransformation of *Bacteroides fragilis*. *Applied and Environmental Microbiology 76* (2010), 3325–3332. [Page 149]

[134] Jacobsen, U. P., Nielsen, H. B., Hildebrand, F., Raes, J., Sicheritz-Ponten, T., Kouskoumvekaki, I., and Panagiotou, G. The chemical interactome space between the human host and the genetically defined gut metabotypes. *The ISME Journal 7* (2012), 730–742. [Page 9]

[135] Jeon, J. H., Kim, S.-J., Lee, H. S., Cha, S.-S., Lee, J. H., Yoon, S.-H., Koo, B.-S., Lee, C.-M., Choi, S. H., Lee, S. H., Kang, S. G., and Lee, J.-H. Novel metagenome-derived carboxylesterase that hydrolyzes β-lactam antibiotics. *Applied and Environmental Microbiology 77* (2011), 7830–7836. [Page 62]

[136] Jiang, C., Ma, G., Li, S., Hu, T., Che, Z., Shen, P., Yan, B., and Wu, B. Characterization of a novel beta-glucosidase-like activity from a soil metagenome. *Journal of Microbiology 47* (2009), 542–548. [Page 61]

[137] Jobanputra, R. S., and Datta, N. Trimethoprim R factors in enterobacteria from clinical specimens. *Journal of Medical Microbiology 7* (1974), 169–177. [Page 30]

[138] Jones, B. V., Begley, M., Hill, C., Gahan, C. G. M., and Marchesi, J. R. Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. *Proceedings of the National Academy of Sciences of the United States of America 105* (2008), 13580–13585. [Page 222]

[139] Jones, J. D., and Gutterson, N. An efficient mobilizable cosmid vector, pRK7813, and its use in a rapid method for marker exchange in *Pseudomonas fluorescens* strain HV37a. *Gene 61* (1987), 299–306. [Page 86]

[140] Kaddurah-Daouk, R., Baillie, R. a., Zhu, H., Zeng, Z.-B., Wiest, M. M., Nguyen, U. T., Wojnoonski, K., Watkins, S. M., Trupp, M., and Krauss, R. M. Enteric microbiome metabolites correlate with response to Simvastatin treatment. *PLOS ONE 6* (2011), e25482. [Page 9]

[141] Kakirde, K. S., Parsley, L. C., and Liles, M. R. Size does matter: application-driven approaches for soil metagenomics. *Soil Biology & Biochemistry 42* (2010), 1911–1923. [Page 21], [Page 218]

[142] Kakirde, K. S., Wild, J., Godiska, R., Mead, D. A., Wiggins, A. G., Goodman, R. M., Szybalski, W., and Liles, M. R. Gram negative shuttle BAC vector for heterologous expression of metagenomic libraries. *Gene 475* (2011), 57–62. [Page 127]

[143] Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research 42* (2014), D199–D205. [Page 181]

[144] Kawate, T., and Gouaux, E. Fluorescence-detection size-exclusion chromatography for precrystallization screening of integral membrane proteins. *Structure 14* (2006), 673–681. [Page 239]

[145] Kelly, D., King, T., and Aminov, R. Importance of microbial colonization of the gut in early life to the development of immunity. *Mutation Research 622* (2007), 58–69. [Page 6]

[146] Kent, W. J. BLAT - the BLAST-like alignment tool. *Genome Research 12* (2002), 656–664. [Page 130]

[147] Khlebnikov, A., Risa, Ø., Skaug, T., Carrier, T. A., and Keasling, J. D. Regulatable arabinose-inducible gene expression system with consistent control in all cells of a culture. *Journal of Bacteriology 182* (2000), 7029–7034. [Page 247]

[148] Kim, U.-J., Shizuya, H., de Jong, P. J., Birren, B., and Simon, M. I. Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Research 20* (1992), 1083–1085. [Page 122], [Page 127]

[149] Kim, Y.-C., and Morrison, S. L. A rapid and economic in-house dna purification method using glass syringe filters. *PLOS ONE 4* (2009), e7750. [Page 45], [Page 301]

[150] Kirjavainen, P. V., and Gibson, G. R. Healthy gut microflora and allergy: factors influencing development of the microbiota. *Annals of Medicine 31* (1999), 288–292. [Page 6]

[151] Kitahara, K., Yasutake, Y., and Miyazaki, K. Mutational robustness of 16S ribosomal RNA, shown by experimental horizontal gene transfer in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America 109* (2012), 19220–19225. [Page 260]

[152] Kiyama, R., and Oishi, M. Instability of plasmid DNA maintenance caused by transcription of poly (dT)-containing sequences in *Escherichia coli*. *Gene 150* (dec 1994), 57–61. [Page 123]

[153] Kiyama, R., and Oishi, M. In vitro transcription of a poly (dA)- poly (dT)-containing sequence is inhibited by interaction between the template and its transcripts. *Nucleic Acids Research 24* (1996), 4577–4583. [Page 123]

[154] Klingenberg, H., Asshauer, K. P., Lingner, T., and Meinicke, P. Protein signature-based estimation of metagenomic abundances including all domains of life and viruses. *Bioinformatics 29* (2013), 973–980. [Page 109]

[155] Knights, D., Ward, T. L., Mckinlay, C. E., Miller, H., Gonzalez, A., and Mcdonald, D. Rethinking "enterotypes". *Cell Host and Microbe 16* (2014), 433–437. [Page 7]

[156] Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., Angenent, L. T., and Ley, R. E. Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences of the United States of America 108* (2011), 4578–4585. [Page 6]

[157] Komissarova, N., Becker, J., Solter, S., Kireeva, M., and Kashlev, M. Shortening of RNA:DNA hybrid in the elongation complex of RNA polymerase is a prerequisite for transcription termination shortening of RNA:DNA hybrid in the elongation complex of RNA polymerase is a prerequisite for transcription termination. *Molecular Cell 10* (2002), 1151–1162. [Page 228], [Page 229]

[158] Koropatkin, N. M., Cameron, E. A., and Martens, E. C. How glycan metabolism shapes the human gut microbiota. *Nature Reviews Microbiology 10* (2012), 323–335. [Page 142], [Page 143], [Page 144]

[159] Koropatkin, N. M., Martens, E. C., Gordon, J. I., and Smith, T. J. Starch catabolism by a prominent human gut symbiont is directed by the recognition of amylose helices. *Structure 16* (2008), 1105–1115. [Page 27], [Page 146], [Page 148], [Page 150]

[160] Krasilnikova, M. M., Samadashwily, G. M., Krasilnikov, A. S., and Mirkin, S. M. Transcription through a simple DNA repeat blocks replication elongation. *The EMBO Journal 17* (1998), 5095–5102. [Page 123]

[161] Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A., Takami, H., Morita, H., Sharma, V. K., Srivastava, T. P., Taylor, T. D., Noguchi, H., Mori, H., Ogura, Y., Ehrlich, D. S., Itoh, K., Takagi, T., Sakaki, Y., Hayashi, T., and Hattori, M. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Research 14* (2007), 169–181. [Page 14]

[162] Kwon, Y.-S., Kim, J., and Kang, C. Viability of *E. coli* cells containing phage RNA polymerase and promoter: interference of plasmid replication by transcription. *Genetic Analysis: Biomolecular Engineering 14* (1998), 133–139. [Page 122]

[163] Lagier, J.-C., Armougom, F., Million, M., Hugon, P., Pagnier, I., Robert, C., Bittar, F., Fournous, G., Gimenez, G., Maraninchi, M., Trape, J.-F., Koonin, E. V., La Scola, B., and Raoult, D. Microbial culturomics: paradigm shift in the human gut microbiome study. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases 18* (2012), 1185–93. [Page 4]

[164] Lakhdari, O., Cultrone, A., Tap, J., Gloux, K., Bernard, F., Ehrlich, S. D., Lefèvre, F., Doré, J., and Blottière, H. M. Functional metagenomics: a high throughput screening method to decipher microbiota-driven NF-κB modulation in the human gut. *PLOS ONE 5* (2010), e13092. [Page 7], [Page 20], [Page 62], [Page 153]

[165] Lam, K. N., and Charles, T. C. Strong spurious transcription likely contributes to DNA insert bias in typical metagenomic clone libraries. *Microbiome 3* (2015), 22. [Page 105], [Page 106], [Page 108], [Page 109], [Page 111], [Page 112], [Page 114], [Page 117], [Page 119], [Page 120], [Page 125], [Page 133]

[166] Lam, K. N., Cheng, J., Engel, K., Neufeld, J. D., and Charles, T. C. Current and future resources for functional metagenomics. *Frontiers in Microbiology 6* (2015), 1196. [Page 16]

[167] Lam, K. N., Hall, M. W., Engel, K., Vey, G., Cheng, J., Neufeld, J. D., and Charles, T. C. Evaluation of a pooled strategy for high-throughput sequencing of cosmid clones from metagenomic libraries. *PLOS ONE 9* (2014), e98968. [Page 64], [Page 65], [Page 66], [Page 68], [Page 69], [Page 70], [Page 72], [Page 73], [Page 75], [Page 77], [Page 78], [Page 79], [Page 81], [Page 89], [Page 92], [Page 94], [Page 97], [Page 103], [Page 129]

[168] Landick, R., Wade, J. T., and Grainger, D. C. H-NS and RNA polymerase: a love–hate relationship? *Current Opinion in Microbiology 24* (2015), 53–59. [Page 259]

[169] Lang, K. S., Anderson, J. M., Schwarz, S., Williamson, L., Handelsman, J., and Singer, R. S. Novel florfenicol and chloramphenicol resistance gene discovered in alaskan soil by using functional metagenomics. *Applied and Environmental Microbiology 76* (2010), 5321–5326. [Page 62]

[170] Lawrence, E. How salmonella survive the stomach. *Nature News* (1998). [Page 6]

[171] Lederberg, J., and Mccray, A. T. 'Ome sweet 'omics - a genealogical treasury of words. *The Scientist 15* (2001), 8. [Page 4]

[172] Lee, C., Kim, J., Shin, S. G., and Hwang, S. Absolute and relative QPCR quantification of plasmid copy number in *Escherichia coli*. *Journal of Biotechnology 123* (2006), 273–280. [Page 243]

[173] Lee, D.-H., Choi, S.-L., Rha, E., Kim, S. J., Yeom, S.-J., Moon, J.-H., and Lee, S.-G. A novel psychrophilic alkaline phosphatase from the metagenome of tidal flat sediments. *BMC Biotechnology 15* (2015), 1. [Page 222]

[174] Lee, M. H., Lee, C. H., Oh, T. K., Song, J. K., and Yoon, J. H. Isolation and characterization of a novel lipase from a metagenomic library of tidal flat sediments: evidence for a new family of bacterial lipases. *Applied and Environmental Microbiology 72* (2006), 7406–7409. [Page 222]

[175] Lee, S., and Hallam, S. J. Extraction of high molecular weight genomic DNA from soils and sediments. *Journal of Visualized Experiments 33* (2009), e1569. [Page 42], [Page 218]

[176] Lee, Y. K., and Mazmanian, S. K. Has the microbiota played a critical role in the evolution of the adaptive immune system? *Science 330* (2010), 1768–1773. [Page 5]

[177] Leis, B., Angelov, A., Mientus, M., Li, H., Pham, V. T. T., Lauinger, B., Bongen, P., Pietruszka, J., Gonçalves, L. G., Santos, H., and Liebl, W. Identification of novel esterase-active enzymes from hot environments by use of the host bacterium *Thermus thermophilus*. *Frontiers in Microbiology 6* (2015), 275. [Page 222], [Page 223]

[178] Lennox, E. S. Transduction of linked genetic characters of the host by bacteriophage P1. *Virology 1*, 2 (1955), 190–206. [Page 295]

[179] Li, C., Zhang, F., and Kelly, W. L. Heterologous production of thiostrepton A and biosynthetic engineering of thiostrepton analogs. *Molecular BioSystems 7* (2011), 82–90. [Page 224]

[180] Li, H., and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics 25* (2009), 1754–1760. [Page 95]

[181] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. The sequence alignment/map format and SAMtools. *Bioinformatics 25* (2009), 2078–2079. [Page 93], [Page 95]

[182] Li, L. Y., Shoemaker, N. B., and Salyers, A. A. Characterization of the mobilization region of a *Bacteroides* insertion element (NBU1) that is excised and transferred by *Bacteroides* conjugative transposons. *Journal of Bacteriology 175* (1993), 6588–6598. [Page 145], [Page 147], [Page 150]

[183] Li, M., Wang, B., Zhang, M., Rantalainen, M., Wang, S., Zhou, H., Zhang, Y., Shen, J., Pang, X., Zhang, M., Wei, H., Chen, Y., Lu, H., Zuo, J., Su, M., Qiu, Y., Jia, W., Xiao, C., Smith, L. M., Yang, S., Holmes, E., Tang, H., Zhao, G., Nicholson, J. K., Li, L., and Zhao, L. Symbiotic gut microbes modulate human metabolic phenotypes. *Proceedings of the National Academy of Sciences of the United States of America 105* (2008), 2117–2122. [Page 15]

[184] Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., and Wang, J. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics 25* (2009), 1966–1967. [Page 93]

[185] Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J., and Wang, J. De novo assembly

of human genomes with massively parallel short read sequencing. *Genome Research 20* (2010), 265–272. [Page 93]

[186] Li, Y., Wexler, M., Richardson, D. J., Bond, P. L., and Johnston, A. W. B. Screening a wide host-range, waste-water metagenomic library in tryptophan auxotrophs of *Rhizobium leguminosarum* and of *Escherichia coli* reveals different classes of cloned trp genes. *Environmental Microbiology 7* (2005), 1927–1936. [Page 21], [Page 62]

[187] Liebl, W., Angelov, A., Juergensen, J., Chow, J., Loeschcke, A., Drepper, T., Classen, T., Pietruzska, J., Ehrenreich, A., Streit, W. R., and Jaeger, K.-E. Alternative hosts for functional (meta)genome analysis. *Applied Microbiology and Biotechnology 98* (2014), 8099–8109. [Page 225]

[188] Lightfield, J., Fram, N. R., and Ely, B. Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. *PLOS ONE 6* (2011), e17677. [Page 111]

[189] Lingner, T., Asshauer, K. P., Schreiber, F., and Meinicke, P. CoMet — a web server for comparative functional profiling of metagenomes. *Nucleic Acids Research 39* (2011), W518–W523. [Page 12], [Page 109]

[190] Liu, J., Liu, W.-d., Zhao, X.-l., Shen, W.-j., Cao, H., and Cui, Z.-l. Cloning and functional characterization of a novel endo-β-1,4-glucanase gene from a soil-derived metagenomic library. *Applied Microbiology and Biotechnology 89* (2010), 1083–1092. [Page 62]

[191] Liu, N., Yan, X., Zhang, M., Xie, L., Wang, Q., Huang, Y., Zhou, X., Wang, S., and Zhou, Z. Microbiome of fungus-growing termites: a new reservoir for lignocellulase genes. *Applied and Environmental Microbiology 77* (2011), 48–56. [Page 222]

[192] Lombard, N., Prestat, E., van Elsas, J. D., and Simonet, P. Soil-specific limitations for access and analysis of soil microbial communities by metagenomics. *FEMS Microbiology Ecology 78* (2011), 31–49. [Page 103]

[193] Lonardi, S., Duma, D., Alpert, M., Cordero, F., Beccuti, M., Bhat, P. R., Wu, Y., Ciardo, G., Alsaihati, B., Ma, Y., Wanamaker, S., Resnik, J., Bozdag, S., Luo, M.-C., and Close, T. J. Combinatorial pooling enables selective sequencing of the barley gene space. *PLoS Computational Biology 9* (2013), e1003010. [Page 82], [Page 83]

[194] Lubys, A. Vectors comprising toxic genes for cloning and expression, 2008. [Page 30]

[195] Lynch, M. D. J., Masella, A. P., Hall, M. W., Bartram, A. K., and Neufeld, J. D. AXIOME: automated exploration of microbial diversity. *GigaScience 2* (2013), 3. [Page 134]

[196] Manichanh, C., Rigottier-Gois, L., Bonnaud, E., Gloux, K., Pelletier, E., Frangeul, L., Nalin, R., Jarrin, C., Chardon, P., Marteau, P., Roca, J.,

and Dore, J. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut 55* (2006), 205–211. [Page 11], [Page 153]

[197] Marchesi, J. R. Human distal gut microbiome. *Environmental Microbiology 13* (2011), 3088–3102. [Page 111]

[198] Marchesi, J. R., Dutilh, B. E., Hall, N., Peters, W. H. M., Roelofs, R., Boleij, A., and Tjalsma, H. Towards the human colorectal cancer microbiome. *PLOS ONE 6* (2011), e20447. [Page 11]

[199] Marchesi, J. R., and Ravel, J. The vocabulary of microbiome research: a proposal. *Microbiome 3* (2015), 31. [Page 4]

[200] Martens, E. C., Chiang, H. C., and Gordon, J. I. Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host and Microbe 4* (2008), 447–457. [Page 30], [Page 144], [Page 146], [Page 152], [Page 178]

[201] Martens, E. C., Heungens, K., and Goodrich-Blair, H. Early colonization events in the mutualistic association between *Steinernema carpocapsae* nematodes and *Xenorhabdus nematophila* bacteria. *Journal of Bacteriology 185* (2003), 3147–3154. [Page 150]

[202] Martens, E. C., Koropatkin, N. M., Smith, T. J., and Gordon, J. I. Complex glycan catabolism by the human gut microbiota: the Bacteroidetes Sus-like paradigm. *The Journal of Biological Chemistry 284* (2009), 24673–24677. [Page 144]

[203] Martens, E. C., Lowe, E. C., Chiang, H., Pudlo, N. a., Wu, M., McNulty, N. P., Abbott, D. W., Henrissat, B., Gilbert, H. J., Bolam, D. N., and Gordon, J. I. Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS Biology 9* (2011), e1001221. [Page 152]

[204] Martinez, A., Kolvek, S. J., Yip, C. L. T., Hopke, J., Brown, K. A., MacNeil, I. A., and Osburne, M. S. Genetically modified bacterial strains and novel bacterial artificial chromosome shuttle vectors for constructing environmental libraries and detecting heterologous natural products in multiple expression hosts. *Applied and Environmental Microbiology 70* (2004), 2452–2463. [Page 225]

[205] Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G., and Neufeld, J. D. PANDAseq: paired-end assembler for Illumina sequences. *BMC Bioinformatics 13* (2012), 31. [Page 134]

[206] Mastropaolo, M. D., Thorson, M. L., and Stevens, A. M. Comparison of *Bacteroides thetaiotaomicron* and *Escherichia coli* 16S rRNA gene expression signals. *Microbiology 155* (2009), 2683–2693. [Page 116], [Page 147], [Page 153], [Page 154]

[207] McNeil, N. I. The contribution of the large-intestine to energy supplies in man. *American Journal of Clinical Nutrition 39* (1984), 338–342. [Page 141]

[208] McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., Cowley, A. P., and Lopez, R. Analysis tool web services from the EMBL-EBI. *Nucleic acids research 41* (2013), W597–W600. [Page 189], [Page 191]

[209] Meinicke, P., Asshauer, K. P., and Lingner, T. Mixture models for analysis of the taxonomic composition of metagenomes. *Bioinformatics 27* (2011), 1618–1624. [Page 109]

[210] Mergulhão, F. J. M., Monteiro, G. a., Larsson, G., Sandén, a. M., Farewell, A., Nystrom, T., Cabral, J. M. S., and Taipa, M. a. Medium and copy number effects on the secretion of human proinsulin in *Escherichia coli* using the universal stress promoters uspA and uspB. *Applied Microbiology and Biotechnology 61* (2003), 495–501. [Page 245]

[211] Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., and Edwards, R. A. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics 9* (2008), 386. [Page 14]

[212] Meyer, R. J., and Shapiro, J. A. Genetic organization of the broad-host-range IncP-1 plasmid R751. *Journal of Bacteriology 143* (1980), 1362–1373. [Page 30]

[213] Mhuantong, W., Charoensawan, V., Kanokratana, P., Tangphatsornruang, S., and Champreda, V. Comparative analysis of sugarcane bagasse metagenome reveals unique and conserved biomass-degrading enzymes among lignocellulolytic microbial communities. *Biotechnology for Biofuels 8* (2015), 16. [Page 223]

[214] Miller, W. G., and Simons, R. W. DNA from diverse sources manifests cryptic low-level transcription in *Escherichia coli*. *Molecular Microbiology 4* (1990), 881–893. [Page 127]

[215] Mimee, M., Tucker, A. C., Voigt, C. A., and Lu, T. K. Programming a human commensal bacterium, *Bacteroides thetaiotaomicron*, to sense and respond to stimuli in the murine gut microbiota. *Cell Systems 1* (2015), 62–71. [Page 147], [Page 151]

[216] Mishra, S., and Imlay, J. A. An anaerobic bacterium, *Bacteroides thetaiotaomicron*, uses a consortium of enzymes to scavenge hydrogen peroxide. *Molecular Microbiology 90* (2013), 1356–1371. [Page 157]

[217] Moreno-hagelsieb, G. The power of operon rearrangements for predicting functional associations. *Computational and Structural Biotechnology Journal 13 13* (2015), 402–406. [Page 14]

[218] Morrison, D. A., and Jaurin, B. Streptococcus pneumoniae possesses canonical *Escherichia coli* (sigma 70) promoters. *Molecular Microbiology 4* (1990), 1143–1152. [Page 115]

[219] Mulligan, M. E., and McClure, W. R. Analysis of the occurrence of promoter-sites in DNA. *Nucleic Acids Research 14* (1986), 109–126. [Page 121]

[220] Nawrocki, E. P., and Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics 29* (2013), 2933–2935. [Page 131]

[221] Nayfach, S., and Pollard, K. S. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biology 16* (2015), 51. [Page 142]

[222] Neufeld, J., Engel, K., Cheng, J., Moreno-Hagelsieb, G., Rose, D., and Charles, T. Open resource metagenomics: a model for sharing metagenomic libraries. *Standards in Genomic Sciences 5* (2011), 203–210. [Page 85], [Page 122], [Page 220]

[223] Neufeld, J. D., Chen, Y., Dumont, M. G., and Murrell, J. C. Marine methylotrophs revealed by stable-isotope probing, multiple displacement amplification and metagenomics. *Environmental Microbiology 10* (2008), 1526–1535. [Page 13]

[224] Nonaka, G., Blankschien, M., Herman, C., Gross, C. A., and Rhodius, V. A. Regulon and promoter analysis of the *E. coli* heat-shock factor, $\sigma 32$, reveals a multifaceted cellular response to heat stress. *Genes & Development 20* (2006), 1776–1789. [Page 116]

[225] Nyyssönen, M., Tran, H. M., Karaoz, U., Weihe, C., Hadi, M. Z., Martiny, J. B. H., Martiny, A. C., and Brodie, E. L. Coupled high-throughput functional screening and next generation sequencing for identification of plant polymer decomposing enzymes in metagenomic libraries. *Frontiers in Microbiology 4* (2013), 282. [Page 223]

[226] Ohlhoff, C. W., Kirby, B. M., Van Zyl, L., Mutepfa, D. L., Casanueva, A., Huddy, R. J., Bauer, R., Cowan, D. A., and Tuffin, M. An unusual feruloyl esterase belonging to family VIII esterases and displaying a broad substrate range. *Journal of Molecular Catalysis B: Enzymatic 118* (2015), 79–88. [Page 223]

[227] Pan, N., and Imlay, J. A. How does oxygen inhibit central metabolism in the obligate anaerobe *Bacteroides thetaiotaomicron*? *Molecular Microbiology 39* (2001), 1562–1571. [Page 157]

[228] Parker, A. C., and Jeffrey Smith, C. Development of an IPTG inducible expression vector adapted for *Bacteroides fragilis*. *Plasmid 68* (2012), 86–92. [Page 146], [Page 169]

[229] Parks, R. J., and Graham, F. L. A helper-dependent system for adenovirus vector production helps define a lower limit for efficient DNA packaging. *Journal of Virology 71* (1997), 3293–3298. [Page 218]

[230] Parsley, L. C., Consuegra, E. J., Kakirde, K. S., Land, A. M., Harper, W. F., and Liles, M. R. Identification of diverse antimicrobial resistance determinants carried on bacterial, plasmid, or viral metagenomes from an activated sludge microbial assemblage. *Applied and Environmental Microbiology 76* (2010), 3753–3757. [Page 62]

[231] Patel, E. H., Paul, L. V., Patrick, S., and Abratt, V. R. Rhamnose catabolism in *Bacteroides thetaiotaomicron* is controlled by the positive transcriptional regulator RhaR. *Research in Microbiology 159* (2008), 678–684. [Page 152]

[232] Pel, J., Broemeling, D., Mai, L., Poon, H.-L., Tropini, G., Warren, R. L., Holt, R. A., and Marziali, A. Nonlinear electrophoretic response yields a unique parameter for separation of biomolecules. *Proceedings of the National Academy of Sciences of the United States of America 106* (2009), 14796–14801. [Page 219]

[233] Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., Bonazzi, V., McEwen, J. E., Wetterstrand, K. a., Deal, C., Baker, C. C., Di Francesco, V., Howcroft, T. K., Karp, R. W., Lunsford, R. D., Wellington, C. R., Belachew, T., Wright, M., Giblin, C., David, H., Mills, M., Salomon, R., Mullins, C., Akolkar, B., Begg, L., Davis, C., Grandison, L., Humble, M., Khalsa, J., Little, a. R., Peavy, H., Pontzer, C., Portnoy, M., Sayre, M. H., Starke-Reed, P., Zakhari, S., Read, J., Watson, B., and Guyer, M. The NIH human microbiome project. *Genome Research 19* (2009), 2317–2323. [Page 4]

[234] Petrenko, P., Kurtz, D., Lobb, B., Neufeld, J., and Doxey, A. MetAnnotate: Function-specific taxonomic profiling and comparison of metagenomes. *BMC Bioinformatics 13* (2015), 92. [Page 12]

[235] Pope, P. B., Denman, S. E., Jones, M., Tringe, S. G., Barry, K., Malfatti, S. A., McHardy, A. C., Cheng, J.-F., Hugenholtz, P., McSweeney, C. S., and Morrison, M. Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase profiles different from other herbivores. *Proceedings of the National Academy of Sciences of the United States of America 107* (2010), 14793–14798. [Page 154]

[236] Pottkämper, J., Barthen, P., Ilmberger, N., Schwaneberg, U., Schenk, A., Schulte, M., Ignatiev, N., and Streit, W. R. Applying metagenomics for the identification of bacterial cellulases that are stable in ionic liquids. *Green Chemistry 11* (2009), 957–965. [Page 62]

[237] Rabausch, U., Juergensen, J., Ilmberger, N., Böhnke, S., Fischer, S., Schubach, B., Schulte, M., and Streit, W. R. Functional screening of metagenome and genome libraries for detection of novel flavonoid-modifying enzymes. *Applied and Environmental Microbiology 79* (2013), 4551–4563. [Page 62], [Page 222]

[238] Reeves, A. R., D'Elia, J. N., Frias, J., and Salyers, A. A. A *Bacteroides thetaiotaomicron* outer membrane protein that is essential for utilization of maltooligosaccharides and starch. *Journal of Bacteriology 178* (1996), 823–830. [Page 143]

[239] Reeves, A. R., Wang, G.-r., and Salyers, A. A. Characterization of four outer membrane proteins that play a role in utilization of starch by *Bacteroides thetaiotaomicron*. *Journal of Bacteriology 179* (1997), 643–649. [Page 143]

[240] Rhee, D.-K. Instability of pneumococcus library in pHC79 and pACYC184. *Archives of Pharmacal Research 18* (1995), 31–37. [Page 121]

[241] Rhodius, V. A., Suh, W. C., Nonaka, G., West, J., and Gross, C. A. Conserved and variable functions of the σE stress response in related genomes. *PLoS Biology 4* (2006), e2. [Page 116]

[242] Riesenfeld, C. S., Schloss, P. D., and Handelsman, J. Metagenomics: genomic analysis of microbial communities. *Annual Review of Genetics 38* (2004), 525–552. [Page 4], [Page 61]

[243] Rodriguez-R, L. M., and Konstantinidis, K. T. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics 30* (2014), 629–635. [Page 107]

[244] Roller, M., Lucic, V., Nagy, I., Perica, T., and Vlahovicek, K. Environmental shaping of codon usage and functional adaptation across microbial communities. *Nucleic Acids Research 41* (2013), 8842–8852. [Page 15]

[245] Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., and Jaffe, D. B. Characterizing and measuring bias in sequence data. *Genome Biology 14* (2013), R51. [Page 258]

[246] Round, J. L., and Mazmanian, S. K. The gut microbiota shapes intestinal immune responses during health and disease. *Nature Reviews Immunology 9* (2009), 313–323. [Page 5], [Page 11]

[247] Rubinstein, N. D., Feldstein, T., Shenkar, N., Botero-Castro, F., Griggio, F., Mastrototaro, F., Delsuc, F., Douzery, E. J. P., Gissi, C., and Huchon, D. Deep sequencing of mixed total DNA without barcodes allows efficient assembly of highly plastic ascidian mitochondrial genomes. *Genome Biology and Evolution 5* (2013), 1185–1199. [Page 74]

[248] Salyers, A. A., Bonheyo, G., and Shoemaker, N. B. Starting a new genetic system: lessons from *Bacteroides*. *Methods 20* (2000), 35–46. [Page 145], [Page 148], [Page 149]

[249] Salyers, A. A., Shoemaker, N., Cooper, A., Elia, J. D., and Shipman, J. A. Genetic methods for *Bacteroides* species. *Methods in Microbiology 29* (1999), 230–249. [Page 30], [Page 145], [Page 147], [Page 161], [Page 196]

[250] Salyers, A. A., Shoemaker, N. B., Stevens, A. M., and Li, L.-Y. Conjugative transposons: an unusual and diverse set of integrated gene transfer elements. *Microbiological Reviews 59* (1995), 579–590. [Page 145], [Page 148]

[251] Sambrook, J., and Russell, D. W. *Molecular cloning: a laboratory manual*, 3rd ed. Cold Spring Harbour Press, Cold Spring Harbour, New York, 2001. [Page 38], [Page 39]

[252] Santangelo, T. J., and Artsimovitch, I. Termination and antitermination: RNA polymerase runs a stop sign. *Nature Reviews Microbiology 9* (2011), 319–329. [Page 228], [Page 229]

[253] Savage, D. C. Microbial ecology of the gastrointestinal tract. *Annual Review of Microbiology 31* (1977), 107–133. [Page 5]

[254] Schallmey, M., Ly, A., Wang, C., Meglei, G., Voget, S., Streit, W. R., Driscoll, B. T., and Charles, T. C. Harvesting of novel polyhydroxyalkanaote (PHA) synthase encoding genes from a soil metagenome library using phenotypic screening. *FEMS Microbiology Letters 321* (2011), 150–156. [Page 21], [Page 62]

[255] Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., and Quince, C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research 43* (2015), e37. [Page 258]

[256] Schloss, P. D. Nurturing the microbiome field. *Science 350* (2015), 1044. [Page 13]

[257] Schneider, C. a., Rasband, W. S., and Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods 9*, 7 (2012), 671–675. [Page 49]

[258] Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods 9* (2012), 811–814. [Page 111], [Page 131]

[259] Sen, D., Van der Auwera, G., Rogers, L. M., Thomas, C. M., Brown, C. J., and Top, E. M. Broad-host-range plasmids from agricultural soils have IncP-1 backbones with diverse accessory genes. *Applied and Environmental Microbiology 77* (2011), 7975–7983. [Page 62]

[260] Sender, R., Fuchs, S., and Milo, R. Revised estimates for the number of human and bacteria cells in the body. *bioRxiv* (2016). [Page 5]

[261] Sentchilo, V., Mayer, A. P., Guy, L., Miyazaki, R., Green Tringe, S., Barry, K., Malfatti, S., Goessmann, A., Robinson-Rechavi, M., and van der Meer, J. R. Community-wide plasmid gene mobilization and selection. *The ISME Journal 7* (2013), 1173–1186. [Page 74]

[262] Shimada, T., Yamazaki, Y., Tanaka, K., and Ishihama, A. The whole set of constitutive promoters recognized by RNA polymerase RpoD holoenzyme of *Escherichia coli*. *PLOS ONE 9* (2014), e90447. [Page 115], [Page 116]

[263] Shipman, J. a., Berleman, J. E., and Salyers, A. A. Characterization of four outer membrane proteins involved in binding starch to the cell surface of *Bacteroides thetaiotaomicron*. *Journal of Bacteriology 182* (2000), 5365–5372. [Page 143]

[264] Shkoporov, A. N., Khokhlova, E. V., Kulagina, E. V., Smeianov, V. V., Kuchmiy, A. A., Kafarskaya, L. I., and Efimov, B. A. Analysis of a novel 8.9 kb cryptic plasmid from *Bacteroides uniformis*, its long-term stability and spread within human microbiota. *Plasmid 69* (2013), 146–159. [Page 146]

[265] Shoemaker, N. B., Barber, R. D., and Salyers, A. A. Cloning and characterization of a *Bacteroides* conjugal tetracycline-erythromycin resistance element by using a shuttle cosmid vector. *Journal of Bacteriology 171* (1989), 1294–1302. [Page 121], [Page 122], [Page 146], [Page 147], [Page 159]

[266] Shoemaker, N. B., Getty, C., Gardner, J. F., and Salyers, A. A. Tn4351 transposes in *Bacteroides* spp. and mediates the integration of plasmid R751 into the *Bacteroides* chromosome. *Journal of Bacteriology 165* (1986), 929–936. [Page 148]

[267] Shoemaker, N. B., Getty, C., Guthrie, E. P., and Salyers, A. A. Regions in *Bacteroides* plasmids pBFTM10 and pB8-51 that allow Escherichia coli-Bacteroides shuttle vectors to be mobilized by IncP plasmids and by a conjugative Bacteroides tetracycline resistance element. *Journal of Bacteriology 166* (1986), 959–965. [Page 148], [Page 149]

[268] Shoemaker, N. B., Guthrie, E. P., Salyers, a. a., and Gardner, J. F. Evidence that the clindamycin-erythromycin resistance gene of *Bacteroides* plasmid pBF4 is on a transposable element. *Journal of Bacteriology 162* (1985), 626–32. [Page 145], [Page 147], [Page 148], [Page 149]

[269] Shoemaker, N. B., Wang, G.-r., and Salyers, A. A. The *Bacteroides* mobilizable insertion element , NBU1, integrates into the 3' end of a Leu-tRNA gene and has an integrase that is a member of the lambda integrase family. *Microbiology 178* (1996), 3594–3600. [Page 148]

[270] Simon, C., Herath, J., Rockstroh, S., and Daniel, R. Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice. *Applied and Environmental Microbiology 75* (2009), 2964–2968. [Page 61], [Page 222]

[271] Simon, R., Priefer, U., and Puhler, A. A broad host range mobilization system for in vivo genetic engineering: transposon mutagenesis in gram negative bacteria. *Biotechnology 1* (1983), 784–791. [Page 27]

[272] Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., and Birol, I. ABySS: a parallel assembler for short read sequence data. *Genome Research 19* (2009), 1117–1123. [Page 90]

[273] Singh, S. S., Singh, N., Bonocora, R. P., Fitzgerald, D. M., Wade, J. T., and Grainger, D. C. Widespread suppression of intragenic transcription initiation by H-NS. *Genes & Development 28* (2014), 214–219. [Page 259]

[274] Smith, C. J. Development and use of cloning systems for *Bacteroides fragilis*: cloning of a plasmid-encoded clindamycin resistance determinant. *Journal of Bacteriology 164* (1985), 294–301. [Page 145], [Page 147]

[275] Smith, C. J., Parker, A., and Rogers, M. B. Plasmid transformation of *Bacteroides* spp. by electroporation. *Plasmid 24* (1990), 100–109. [Page 149]

[276] Smith, C. J., and Parker, A. C. A gene product related to Tral is required for the mobilization of *Bacteroides* mobilizable transposons and plasmids. *Molecular Microbiology 20* (1996), 741–750. [Page 148]

[277] Smith, C. J., Rogers, M. B., and McKee, M. L. Heterologous gene expression in *Bacteroides fragilis*. *Plasmid 27* (1992), 141–154. [Page 145], [Page 147]

[278] Smith, C. J., Rollins, L. A., and Parker, A. C. Nucleotide sequence determination and genetic analysis of the *Bacteroides* plasmid, pBI143. *Plasmid 34* (1995), 211–222. [Page 145], [Page 146], [Page 162], [Page 169]

[279] Smith, C. J., Welch, R. A., and Macrina, F. L. Two independent conjugal transfer systems operating in *Bacteroides fragilis* V479-1. *Journal of Bacteriology 151* (1982), 281–287. [Page 148]

[280] Smith, M. A., and Bidochka, M. J. Bacterial fitness and plasmid loss: the importance of culture conditions and plasmid size. *Canadian Journal of Microbiology 44* (1998), 351–355. [Page 242]

[281] Sobhani, I., Tap, J., Roudot-Thoraval, F., Roperch, J. P., Letulle, S., Langella, P., Corthier, G., Van Nhieu, J. T., and Furet, J. P. Microbial dysbiosis in colorectal cancer (CRC) patients. *PLOS ONE 6* (2011), e16393. [Page 11]

[282] Sommer, M. O., Church, G. M., and Dantas, G. A functional metagenomic approach for expanding the synthetic biology toolbox for biomass conversion. *Molecular Systems Biology 6* (2010), 360. [Page 62], [Page 222]

[283] Sommer, M. O. A., Dantas, G., and Church, G. M. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science 325* (2009), 1128–1131. [Page 10], [Page 17]

[284] Sonnenburg, E. D., Smits, S. A., Tikhonov, M., Higginbottom, S. K., Wingreen, N. S., and Sonnenburg, J. L. Diet-induced extinctions in the gut microbiota compound over generations. *Nature 529* (2016), 212–215. [Page 7]

[285] Sonnenburg, E. D., Zheng, H., Joglekar, P., Higginbottom, S. K., Firbank, S. J., Bolam, D. N., and Sonnenburg, J. L. Specificity of polysaccharide use in intestinal Bacteroides species determines diet-induced microbiota alterations. *Cell 141* (2010), 1241–1252. [Page 151]

[286] Sonnenburg, J. L. Microbiome engineering. *Nature 518* (2015), S10–S10. [Page 152]

[287] Sorek, R., Zhu, Y., Creevey, C. J., Francino, M. P., Bork, P., and Rubin, E. M. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science 318* (2007), 1449–1452. [Page 102], [Page 123]

[288] Spor, A., Koren, O., and Ley, R. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Reviews Microbiology 9* (2011), 279–290. [Page 7], [Page 8], [Page 111]

[289] Stassi, D. L., and Lacks, S. A. Effect of strong promoters on the cloning in *Escherichia coli* of DNA fragments from *Streptococcus pneumoniae*. *Gene 18* (1982), 319–328. [Page 122]

[290] Stevens, A. M., Shoemaker, N. B., Li, L. Y., and Salyers, A. A. Tetracycline regulation of genes on *Bacteroides* conjugative transposons. *Journal of Bacteriology 175* (1993), 6134–6141. [Page 145], [Page 147]

[291] Stueber, D., and Bujard, H. Transcription from efficient promoters can interfere with plasmid replication and diminish expression of plasmid specified genes. *The EMBO Journal 1* (1982), 1399–1404. [Page 122], [Page 227]

[292] Stulberg, E., Fravel, D., Proctor, L. M., Murray, D. M., LoTempio, J., Chrisey, L., Garland, J., Goodwin, K., Graber, J., Harris, M. C., Jackson, S., Mishkind, M., Porterfield, D. M., and Records, A. An assessment of US microbiome research. *Nature Microbiology 1* (2016), 15015. [Page 13]

[293] Suenaga, H. Targeted metagenomics: a high-resolution metagenomics approach for specific gene clusters in complex microbial communities. *Environmental Microbiology 14* (2011), 13–22. [Page 13]

[294] Suenaga, H., Ohnuki, T., and Miyazaki, K. Functional screening of a metagenomic library for genes involved in microbial degradation of aromatic compounds. *Environmental Microbiology 9* (2007), 2289–2297. [Page 223]

[295] Sulaiman, S., Yamato, S., Kanaya, E., Kim, J.-J., Koga, Y., Takano, K., and Kanaya, S. Isolation of a novel cutinase homolog with polyethylene terephthalate-degrading activity from leaf-branch compost by using a metagenomic approach. *Applied and Environmental Microbiology 78* (2012), 1556–1562. [Page 223]

[296] Tagawa, J., Inoue, T., Naito, M., Sato, K., Kuwahara, T., Nakayama, M., Nakayama, K., Yamashiro, T., and Ohara, N. Development of a novel plasmid vector pTIO-1 adapted for electrotransformation of *Porphyromonas gingivalis*. *Journal of Microbiological Methods 105* (2014), 174–179. [Page 146]

[297] Tancula, E., Feldhaus, M. J., Bedzyk, L. A., and Salyers, A. A. Location and characterization of genes involved in binding of starch to the surface of *Bacteroides thetaiotaomicron*. *Journal of Bacteriology 174* (1992), 5609–5616. [Page 142]

[298] Tang, X., Nakata, Y., Li, H.-O., Zhang, M., Gao, H., Fujita, A., Sakatsume, O., Ohta, T., and Yokoyama, K. The optimization of preparations of competent cells for transformation of *E. coli*. *Nucleic Acids Research 22* (1994), 2857–2858. [Page 38]

[299] Tasse, L., Bercovici, J., Pizzut-Serin, S., Robe, P., Tap, J., Klopp, C., Cantarel, B. L., Coutinho, P. M., Henrissat, B., Leclerc, M., Doré, J., Monsan, P., Remaud-Simeon, M., and Potocki-Veronese, G. Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. *Genome Research 11* (2010), 1605–1612. [Page 20], [Page 62], [Page 153], [Page 222]

[300] Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., and Natale, D. A. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics 4* (2003), 41. [Page 14]

[301] Tatusov, R. L., Koonin, E. V., and Lipman, D. J. A genomic perspective on protein families. *Science 278* (1997), 631–637. [Page 14]

[302] Taupp, M., Mewis, K., and Hallam, S. J. The art and design of functional metagenomic screens. *Current Opinion in Biotechnology 22* (2011), 465–472. [Page 21], [Page 123], [Page 225]

[303] Tebbe, C. C., and Vahjen, W. Interference of humic acids and DNA extracted directly from soil in detection and transformation of recombinant DNA from bacteria and a yeast. *Applied and Environmental Microbiology 59* (1993), 2657–2665. [Page 219]

[304] Temperton, B., Field, D., Oliver, A., Tiwari, B., Mühling, M., Joint, I., and Gilbert, J. A. Bias in assessments of marine microbial biodiversity in fosmid libraries as evaluated by pyrosequencing. *The ISME Journal 3* (2009), 792–796. [Page 102], [Page 104]

[305] Terrón-González, L., Medina, C., Limón-Mortés, M. C., and Santero, E. Heterologous viral expression systems in fosmid vectors increase the functional analysis potential of metagenomic libraries. *Scientific Reports 3* (2013), 1107. [Page 223], [Page 224], [Page 260]

[306] Thomson, A. M., Flint, H. J., Béchet, M., Martin, J., and Dubourguier, H.-C. A new *Escherichia coli:Bacteroides* shuttle vector, pRRI207, based on the Bacteroides ruminicola plasmid replicon pRRI2. *Current Microbiology 24* (1992), 49–54. [Page 148], [Page 149]

[307] Torres-Cortés, G., Millán, V., Ramírez-Saad, H. C., Nisa-Martínez, R., Toro, N., and Martínez-Abarca, F. Characterization of novel antibiotic resistance genes identified by functional metagenomics on soil samples. *Environmental Microbiology 13* (2011), 1101–1114. [Page 19], [Page 61]

[308] Troeschel, S. C., Drepper, T., Leggewie, C., Streit, W. R., and Jaeger, K.-E. Novel tools for the functional expression of metagenomic DNA. In *Metagenomics: Methods and Protocols*, W. R. Streit and R. Daniel, Eds., Methods in Molecular Biology. Humana Press, New York, 2010, ch. 8, pp. 117–139. [Page 21], [Page 220]

[309] Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., Egholm, M., Henrissat, B., Heath, A. C., Knight, R., and Gordon, J. I. A core gut microbiome in obese and lean twins. *Nature 457* (2009), 480–484. [Page 8]

[310] Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. The human microbiome project. *Nature 449* (2007), 804–810. [Page 7]

[311] Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature 444* (2006), 1027–1031. [Page 8], [Page 10]

[312] Uchiyama, T., and Miyazaki, K. Functional metagenomics for enzyme discovery: challenges to efficient screening. *Current Opinion in Biotechnology 20* (2009), 616–622. [Page 21], [Page 225], [Page 260]

[313] Ufarté, L., Potocki-Veronese, G., and Laville, É. Discovery of new protein families and functions: new challenges in functional metagenomics for biotechnologies and microbial ecology. *Frontiers in Microbiology 6* (2015), 563. [Page 21]

[314] Valentine, P. J., Shoemaker, N. B., and Salyers, A. A. Mobilization of *Bacteroides* plasmids by *Bacteroides* conjugal elements. *Journal of Bacteriology 170* (1988), 1319–1324. [Page 146], [Page 149]

[315] Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H., and Smith, H. O. Environmental genome shotgun sequencing of the Sargasso Sea. *Science 304* (2004), 66–74. [Page 13]

[316] Vercammen, K., Garcia-Armisen, T., Goeders, N., Van Melderen, L., Bodilis, J., and Cornelis, P. Identification of a metagenomic gene cluster containing a new class A beta-lactamase and toxin-antitoxin systems. *MicrobiologyOpen 2* (2013), 674–683. [Page 223]

[317] Vingadassalom, D., Kolb, A., Mayer, C., Rybkine, T., Collatz, E., and Podglajen, I. An unusual primary sigma factor in the Bacteroidetes phylum. *Molecular Microbiology 56* (2005), 888–902. [Page 153], [Page 159]

[318] Voget, S., Steele, H. L., and Streit, W. R. Characterization of a metagenome-derived halotolerant cellulase. *Journal of Biotechnology 126* (2006), 26–36. [Page 62]

[319] Walker, A. W., Duncan, S. H., Louis, P., and Flint, H. J. Phylogeny, culturing, and metagenomics of the human gut microbiota. *Trends in Microbiology 22* (2014), 267–274. [Page 4]

[320] Wang, C., Meek, D. J., Panchal, P., Archibald, F. S., Driscoll, B. T., Charles, T. C., and Boruvka, N. Isolation of poly-3-hydroxybutyrate metabolism genes from complex microbial communities by phenotypic complementation of bacterial mutants. *Applied and Environmental Microbiology 72* (2006), 384–391. [Page 21], [Page 86]

[321] Wang, L., Hatem, A., Catalyurek, U. V., Morrison, M., and Yu, Z. Metagenomic insights into the carbohydrate-active enzymes carried by the microorganisms adhering to solid digesta in the rumen of cows. *PLOS ONE 8* (2013), e78507. [Page 63], [Page 222]

[322] Wang, M., Doak, T. G., and Ye, Y. Subtractive assembly for comparative metagenomics, and its application to type 2 diabetes metagenomes. *Genome Biology 16* (2015), 243. [Page 11]

[323] Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology 73* (2007), 5261–5267. [Page 131]

[324] Warnecke, F., Luginbühl, P., Ivanova, N., Ghassemian, M., Richardson, T. H., Stege, J. T., Cayouette, M., McHardy, A. C., Djordjevic, G., Aboushadi, N., Sorek, R., Tringe, S. G., Podar, M., Martin, H. G., Kunin, V., Dalevi, D., Madejska, J., Kirton, E., Platt, D., Szeto, E., Salamov,

A., Barry, K., Mikhailova, N., Kyrpides, N. C., Matson, E. G., Ottesen, E. A., Zhang, X., Hernández, M., Murillo, C., Acosta, L. G., Rigoutsos, I., Tamayo, G., Green, B. D., Chang, C., Rubin, E. M., Mathur, E. J., Robertson, D. E., Hugenholtz, P., and Leadbetter, J. R. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature 450* (2007), 560–565. [Page 222]

[325] Wassenaar, T. M. *Bacteria: the benign, the bad, and the beautiful*. Wiley, 2012. [Page 6]

[326] Welch, R. A., and Macrina, F. L. Physical characterization of *Bacteroides fragilis* R plasmid pBF4. *Journal of bacteriology 145* (1981), 867–872. [Page 169]

[327] Westenberg, M., Bamps, S., Soedling, H., Hope, I. A., and Dolphin, C. T. *Escherichia coli* MW005: lambda Red-mediated recombineering and copy-number induction of oriV-equipped constructs in a single host. *BMC Biotechnology 10* (2010), 27. [Page 226]

[328] Wexler, H. M. *Bacteroides*: the good, the bad, and the nitty-gritty. *Clinical Microbiology Reviews 20* (2007), 593–621. [Page 10]

[329] Wexler, M., Bond, P. L., Richardson, D. J., and Johnston, A. W. B. A wide host-range metagenomic library from a waste water treatment plant yields a novel alcohol/aldehyde dehydrogenase. *Environmental Microbiology 7* (2005), 1917–1926. [Page 62]

[330] Wexler, M., and Johnston, A. W. B. Wide host-range cloning for functional metagenomics. In *Metagenomics: Methods and Protocols*, W. R. Streit and R. Daniel, Eds., vol. 668 of *Methods in Molecular Biology*. Humana Press, New York, 2010, ch. 5, pp. 77–96. [Page 220]

[331] White, B. a., Lamed, R., Bayer, E. a., and Flint, H. J. Biomass utilization by gut microbiomes. *Annual review of microbiology* (2014), 279–296. [Page 144]

[332] Wikoff, W. R., Anfora, A. T., Liu, J., Schultz, P. G., Lesley, S. A., Peters, E. C., and Siuzdak, G. Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proceedings of the National Academy of Sciences of the United States of America 106* (2009), 3698–3703. [Page 9]

[333] Wild, J., Hradecna, Z., and Szygbalski, W. Conditionally amplifiable BACs: switching from single-copy to high-copy vectors and genomic clones. *Genome Research 12* (2002), 1434–1444. [Page 224], [Page 242]

[334] Wilke, A., Bischof, J., Gerlach, W., Glass, E., Harrison, T., Keegan, K. P., Paczian, T., Trimble, W. L., Bagchi, S., Grama, A., Chaterji, S., and Meyer, F. The MG-RAST metagenomics database and portal in 2015. *Nucleic Acids Research 44* (2015), D590–D594. [Page 14]

[335] Williamson, L. L., Borlee, B. R., Schloss, P. D., Guan, C., Allen, H. K., and Handelsman, J. Intracellular screen to identify metagenomic clones that induce

or inhibit a quorum-sensing biosensor. *Applied and Environmental Microbiology 71* (2005), 6335–6344. [Page 222]

[336] Wong, C., Klieve, A., Hamdorf, B., Schafer, D., Bräu, L., Seet, S., and Gregg, K. Family of shuttle vectors for ruminal *Bacteroides*. *Journal of Molecular Microbiology and Biotechnology* (2003), 123–132. [Page 146]

[337] Wright, J. J., and Hayward, R. S. Transcriptional termination at a fully Rho-independent site in *Escherichia coli* is prevented by uninterrupted translation of the nascent RNA. *The EMBO Journal 6* (1987), 1115–1119. [Page 233], [Page 234]

[338] Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F. D., and Lewis, J. D. Linking long-term dietary patterns with gut microbial enterotypes. *Science 334* (2011), 105–108. [Page 7]

[339] Xu, J., Bjursell, M. K., Himrod, J., Deng, S., Carmichael, L. K., Chiang, H. C., Hooper, L. V., and Gordon, J. I. A genomic view of the human-*Bacteroides* thetaiotaomicron symbiosis. *Science 299* (2003), 2074–2076. [Page 27], [Page 140], [Page 190]

[340] Yanisch-Perron, C., Vieira, J., and Messing, J. Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene 33* (1985), 103–119. [Page 162], [Page 245]

[341] Yaung, S. J., Deng, L., Li, N., Braff, J. L., Church, G. M., Bry, L., Wang, H., and Gerber, G. K. Improving microbial fitness in the mammalian gut by in vivo temporal functional metagenomics. *Molecular Systems Biology 11* (2015), 788. [Page 151], [Page 260]

[342] Yoon, M. Y., Lee, K.-M., Yoon, Y., Go, J., Park, Y., Cho, Y.-J., Tannock, G. W., and Yoon, S. S. Functional screening of a metagenomic library reveals operons responsible for enhanced intestinal colonization by gut commensal microbes. *Applied and Environmental Microbiology 79* (2013), 3829–3938. [Page 154]

[343] Yung, P. Y., Burke, C., Lewis, M., Egan, S., Kjelleberg, S., and Thomas, T. Phylogenetic screening of a bacterial, metagenomic library using homing endonuclease restriction and marker insertion. *Nucleic Acids Research 37* (2009), e144. [Page 222]

[344] Zhang, G., Gurtu, V., and Kain, S. R. An enhanced green fluorescent protein allows sensitive detection of gene transfer in mammalian cells. *Biochemical and Biophysical Research Communications 227* (1996), 707–711. [Page 239]

[345] Zhang, T., Han, W.-J., and Liu, Z.-P. Gene cloning and characterization of a novel esterase from activated sludge metagenome. *Microbial Cell Factories 8* (2009), 67. [Page 223]

[346] Zhao, K., Liu, M., and Burgess, R. R. Promoter and regulon analysis of nitrogen assimilation factor, σ54, reveal alternative strategy for *E. coli* MG1655 flagellar biosynthesis. *Nucleic Acids Research 38* (2010), 1273–1283. [Page 116]

[347] Zhou, J., Bruns, M. A., and Tiedje, J. M. DNA recovery from soils of diverse composition. *Applied and Environmental Microbiology 62* (1996), 316–322. [Page 42], [Page 218]

Appendix A

Recipes for media and solutions

A.1 LB: lysogeny broth (or Luria-Bertani media)

This recipe is for the Lennox variety of LB [178].

Prepare:

- 10 g tryptone
- 5 g yeast extract
- 5 g sodium chloride
- top to 1000 ml with distilled water

Aliquot 200 ml per bottle. If preparing solid media, add:

- 3-4 g agar per bottle

Autoclave. Store at room temperature and steam agar media prior to use.

A.2 TB: terrific broth media

This protocol was adapted from Cold Spring Harbor Protocols: http://cshprotocols.cshlp.org/content/2006/1/pdb.rec8620

Prepare:

- 12 g tryptone
- 24 g yeast extract
- 8 ml 50% glycerol
- top to 900 ml with distilled water

Aliquot 90 ml per bottle and autoclave. Prior to use, add per bottle:

- 10 ml 0.17 M KH_2PO_4, 0.72 M K_2HPO_4

A.3 TYG: tryptone yeast glucose media

This recipe was adapted from one shared with me by Nicole Koropatkin and Eric Martens from the University of Michigan.

Prepare and autoclave:

- 10 g tryptone
- 5 g yeast extract
- 2 g glucose
- top to 860 ml with distilled water

Add per 172 ml of media:

- 20 ml potassium phosphate buffer, pH 7.2
- 8 ml TYG salts (per litre: 0.5 g $MgSO_4-7\,H_2O$, 10 g $NaHCO_3$, 2 g NaCl)
- 50 µl 0.8% $CaCl_2$
- 50 µl 0.4 mg/ml $FeSO_4{\cdot}_7H_2O$

Store at room temperature. Prior to use, add per 5 ml of broth:

- 5 µl 1.2 mM histidine-1.9 mM hematin solution (hematin dissolved in 1 M NaOH, neutralized with equivalent volume 1 M HCl, and histidine solution added)
- 5 µl 1 mg/ml menadione (Vitamin K; dissolved in ethanol)
- 20 µl 0.25 mg/ml resazurin indicator
- 50 µl 50 mg/ml cysteine, thawed from -20℃

A.4 BHI: brain heart infusion media

BHI blood agar (BHIH)

This recipe was shared with me by Nicole Koropatkin and Eric Martens from the University of Michigan.

Prepare:

- 37 g brain heart infusion power (BD cat. no. B211059)
- top to 900 ml with distilled water

Aliquot 450 ml per bottle and add:

- 10 g agar per bottle; include a stir bar

Autoclave. Store at room temperature. Steam prior to use, cool agar on stir plate, and add:

- 50 ml defibrinated horse blood, equilibrated to room temperature

BHI broth with supplementation (BHI+)

This recipe was adapted from the TYG recipe shared with me by Nicole Koropatkin and Eric Martens from the University of Michigan.

Prepare:

- 37 g brain heart infusion powder (BD cat. no. B211059)
- top to 1 L with distilled water

Aliquot 200 ml per bottle and autoclave. Before use, add per 5 ml of broth:

- 5 µl 1.2 mM histidine-1.9 mM hematin solution (hematin dissolved in 1 M NaOH, neutralized with equivalent volume 1 M HCl, and histidine solution added)
- 5 µl 1 mg/ml menadione (Vitamin K; dissolved in ethanol)
- 20 µl 0.25 mg/ml resazurin indicator
- 50 µl 50 mg/ml cysteine, thawed from -20℃

A.5 Bt MM: *B.theta* minimal media

This recipe was adapted from one shared with me by Nicole Koropatkin and Eric Martens from the University of Michigan, specifically by addition of trace elements.

Prepare and autoclave to store at room temperature or use directly:

- 2.5 g carbon source (e.g. glucose or chondroitin sulfate)
- 10 g agar, for solid media; include a stir bar
- top to 440 ml with distilled water

Use stir plate to mix while adding:

- 50 ml 10× Bt salts (per litre: 136 g KH_2PO_4, 8.75 g NaCl, 11.25 g $(NH_4)_2SO_4$)
- 5 ml 50 mg/ml cysteine, thawed from -20℃
- 500 µl 1.2 mM histidine-1.9 mM hematin solution (hematin dissolved in 1 M NaOH, neutralized with equivalent volume 1 M HCl, and histidine solution added)
- 500 µl trace elements (per litre: 0.247 g H_3BO_3, 0.1 g $CuSO_4{\cdot}_5H_2O$, 0.338 g $MnSO_4{\cdot}H_2O$, 0.282 g $ZnSO_4{\cdot}_7H_2O$, 0.056 g $CoSO_4{\cdot}_7H_2O$, 0.048 g $Na_2MoO_4{\cdot}_2H_2O$)
- 500 µl 0.8% $CaCl_2$
- 500 µl 0.4 mg/ml $FeSO_4{\cdot}_7H_2O$
- 500 µl 1 mg/ml menadione (vitamin K; dissolved in ethanol)
- 500 µl 0.1 M $MgCl_2$
- 500 µl 0.01 mg/ml vitamin B12 (dissolved in ethanol)

Note that *B. theta* minimal media plates should not be stored at all; they should be prepared fresh on the day they are required.

A.6 TAE: tris acetic acid EDTA electrophoresis buffer

This protocol was adapted from OpenWetWare: http://openwetware.org/wiki/1X_TAE

For 50× TAE stock, prepare in a starting volume of ∼600-700 ml:

- 242 g Tris free base
- 18.6 g disodium EDTA (add before glacial acetic acid)
- 57.1 ml glacial acetic acid
- pH to 8.0 (optional; should be about 8)
- top to 1000 ml

Dilute 1 in 50 with distilled water. The 1× dilution can be stored at room temperature for weeks in a large carboy.

A.7 Plasmid miniprep solutions

Recipes for the following solutions were obtained from OpenWetWare and were based on buffers from the Qiagen QIAprep Spin Miniprep Kit. The recipes are reproduced here but can be found at: http://openwetware.org/wiki/Qiagen_Buffers

P1: resuspension solution

Prepare and autoclave:

- 50 mM Tris-HCl pH 8.0
- 10 mM EDTA

Add:

- RNaseA to 100 µg/ml

Store at 4℃ .

P2: lysis solution

Prepare non-sterile:

- 200 mM NaOH, from 2 M stock
- 1% SDS, from 20% stock (may require heating if precipitated; do not steam)

N3: neutralization solution

Prepare non-sterile:

- 4.2 M guanidine hydrchloride (or guanidine isothiocyanate)
- 0.9 M potassium acetate
- pH to 4.8

PB: optional wash solution

Prepare non-sterile:

- 5 M guanidine hydrchloride (or guanidine isothiocyanate)
- 30% isopropanol

PE: ethanol wash solution

Prepare:

- 10 mM Tris-HCl pH 7.5, sterile
- 80% ethanol

A.8 Gel extraction solutions

The recipe for the binding buffer was found in the literature [149]. The recipe for the ethanol wash was obtained from OpenWetWare and is based on buffers from the Qiagen QIAquick Gel Extraction Kit; see http://openwetware.org/wiki/Qiagen_Buffers

Binding buffer

Prepare non-sterile:

- 140 mM MES-NaOH (pH 7.0)
- 20 mM EDTA
- 5.5 M guanidine isothiocyanate

PE: ethanol wash solution

Same as for plasmid miniprep.

A.9 Plasmid maxiprep solutions

The recipes for these alkaline lysis solutions were provided by my supervisor, Trevor Charles.

TEG: resuspension

Dilute from concentrated stocks using dH_2O to make:

- 50 mM Tris-Cl pH 8.0 (1/20 of 1 M stock)
- 20 mM Na-EDTA pH 8.0 (1/10 of 0.2 stock)
- 1% glucose

ALS: alkaline lysis

Dilute from concentrated stocks using sterile dH_2O to make:

- 0.2 M NaOH (1/10 of 2 M stock)
- 1% SDS (1/20 of 20% stock)

Store in plastic bottle; do not store in glass.

HSS: neutralization

Dissolve in 60 ml:

- 147 g of K-Ac

Then add:

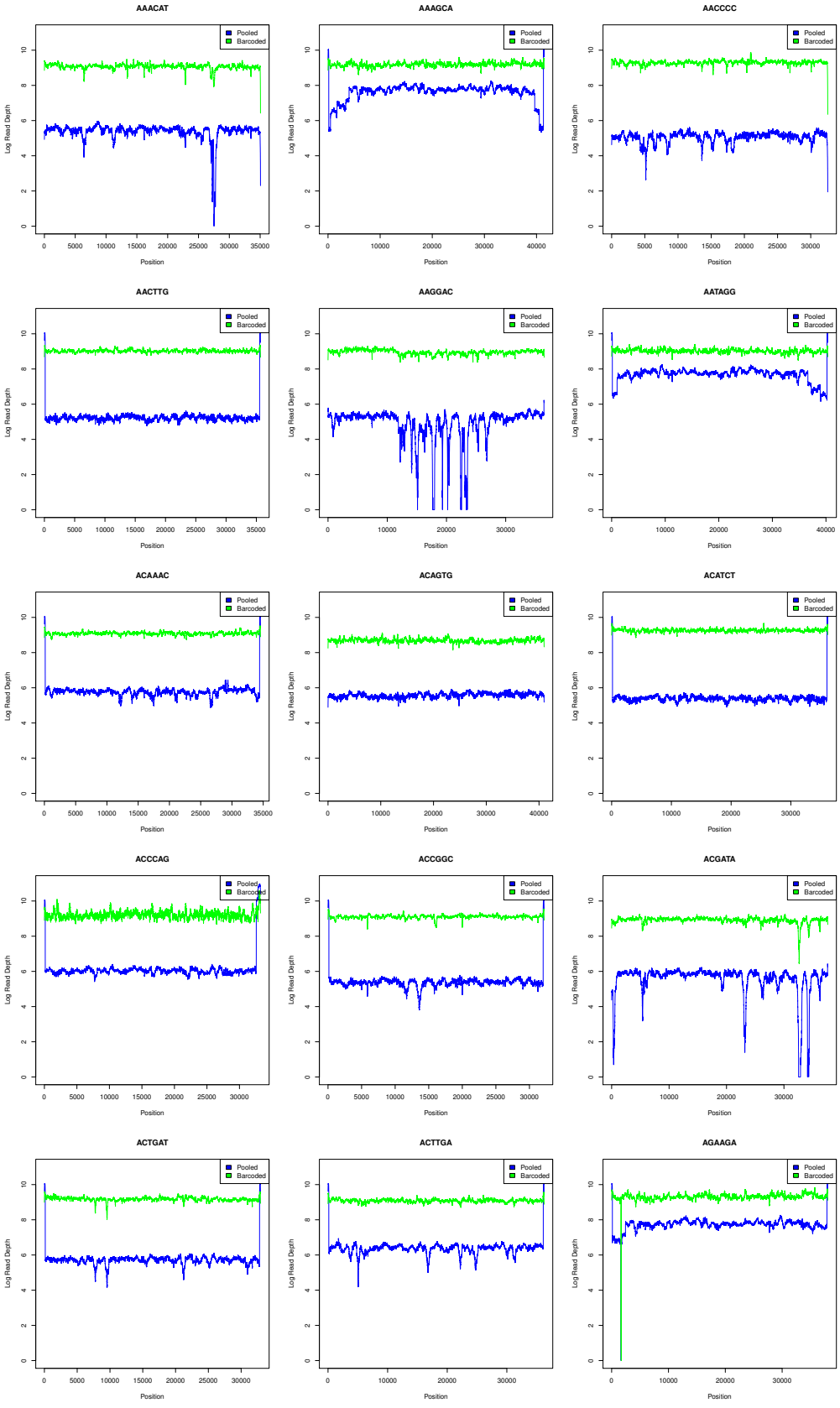
- 282 ml glacial acetic acid
- top up to 500 ml with dH2O

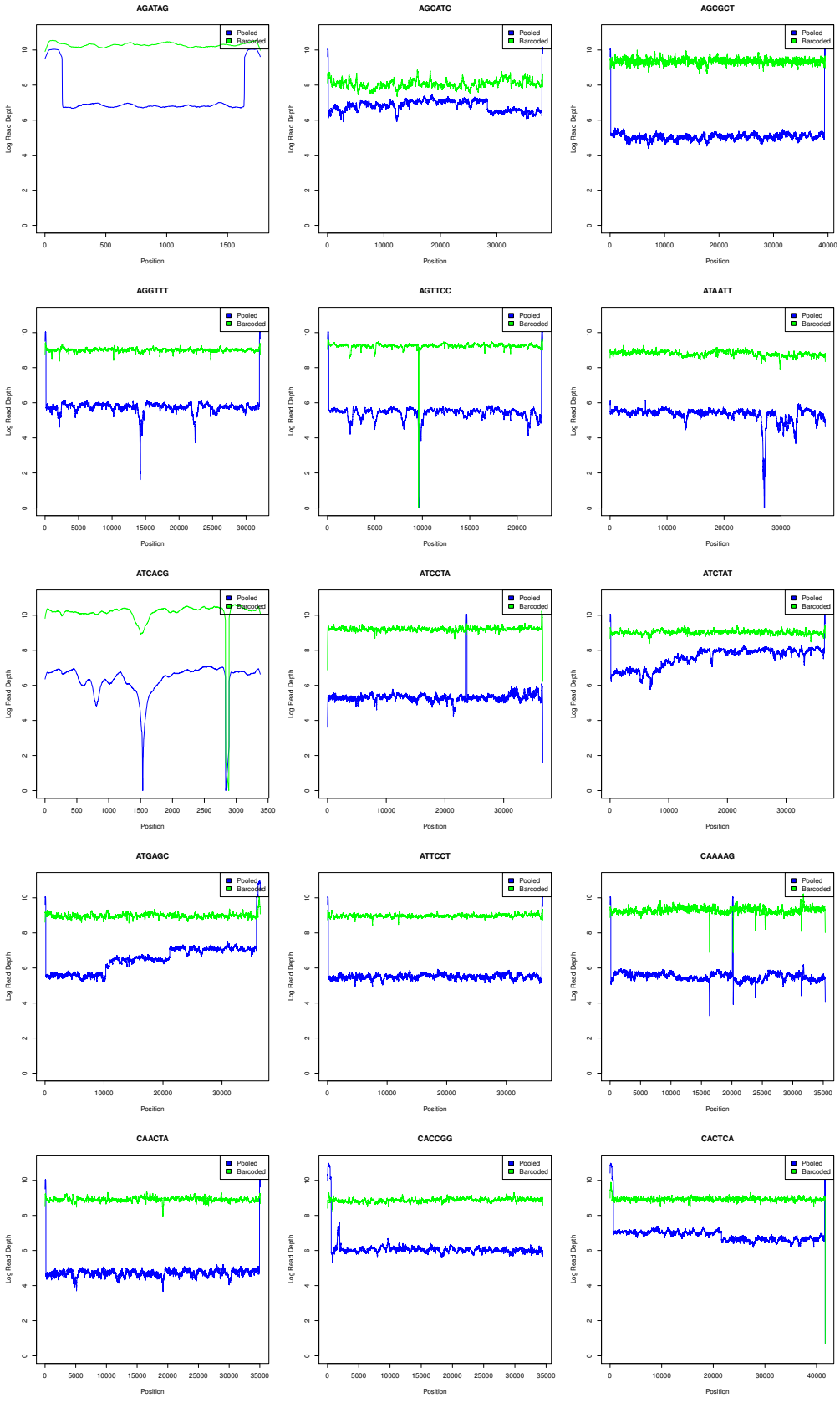
Appendix B

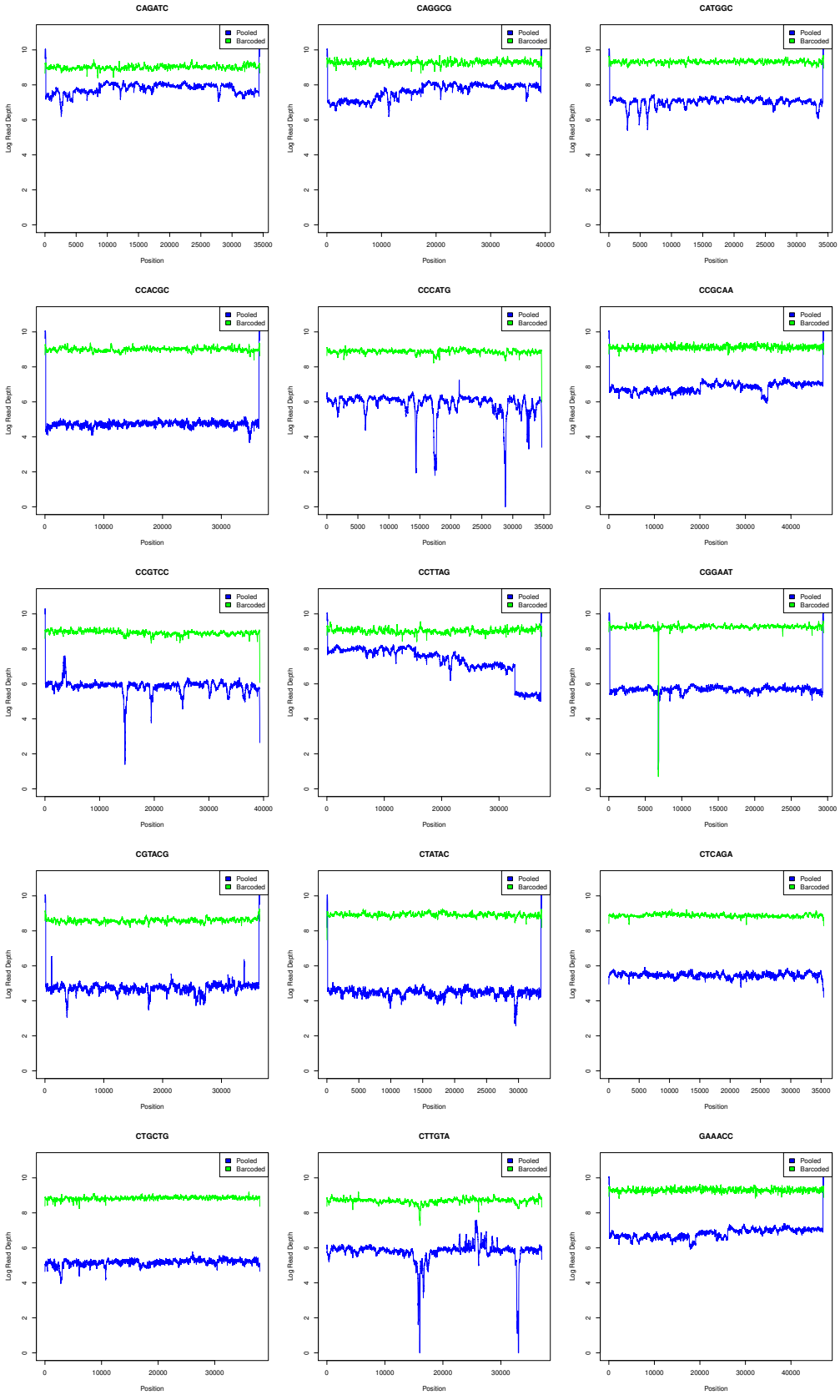
Supplementary information for Chapter 3

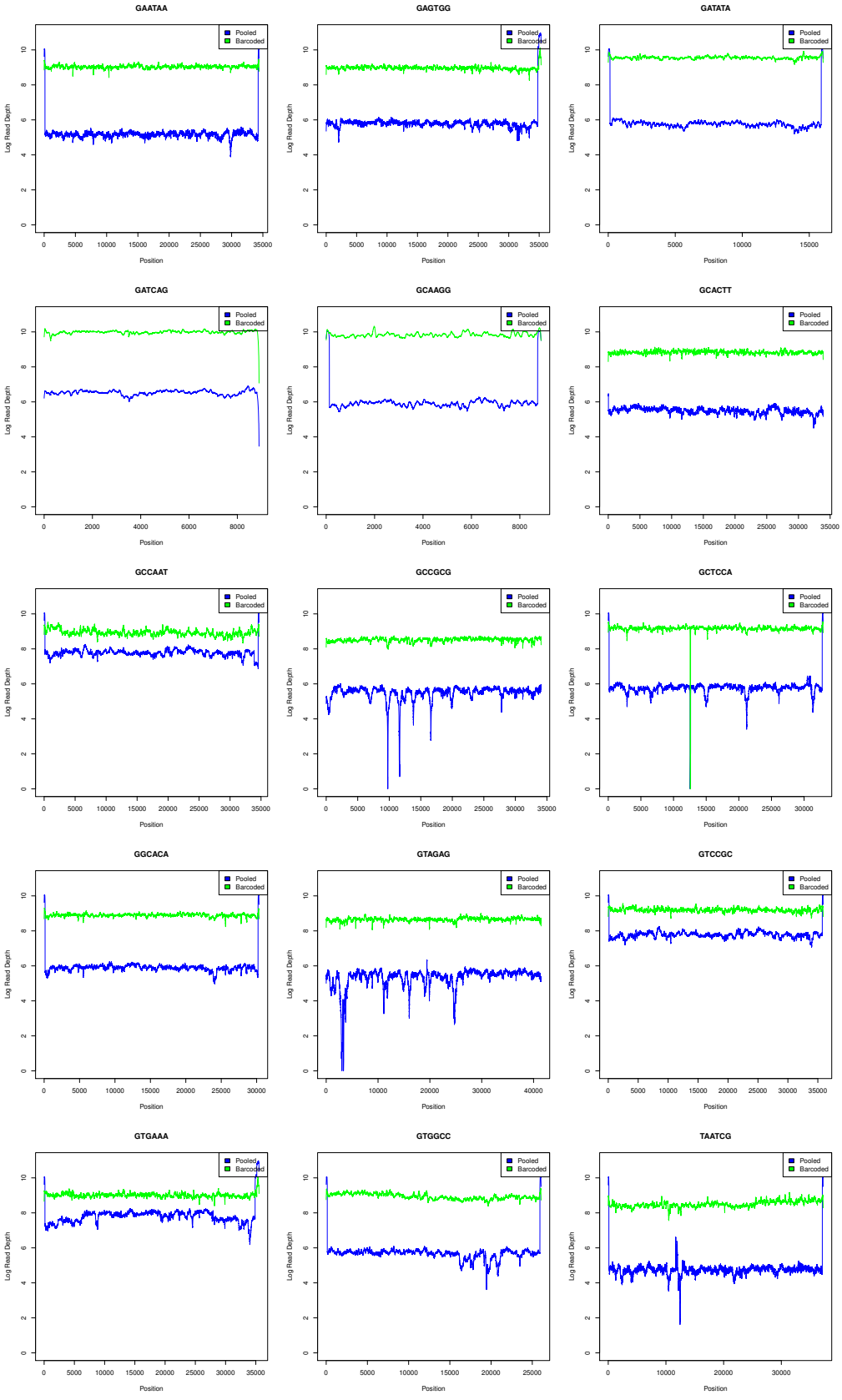
B.1 Clone sequencing read depth

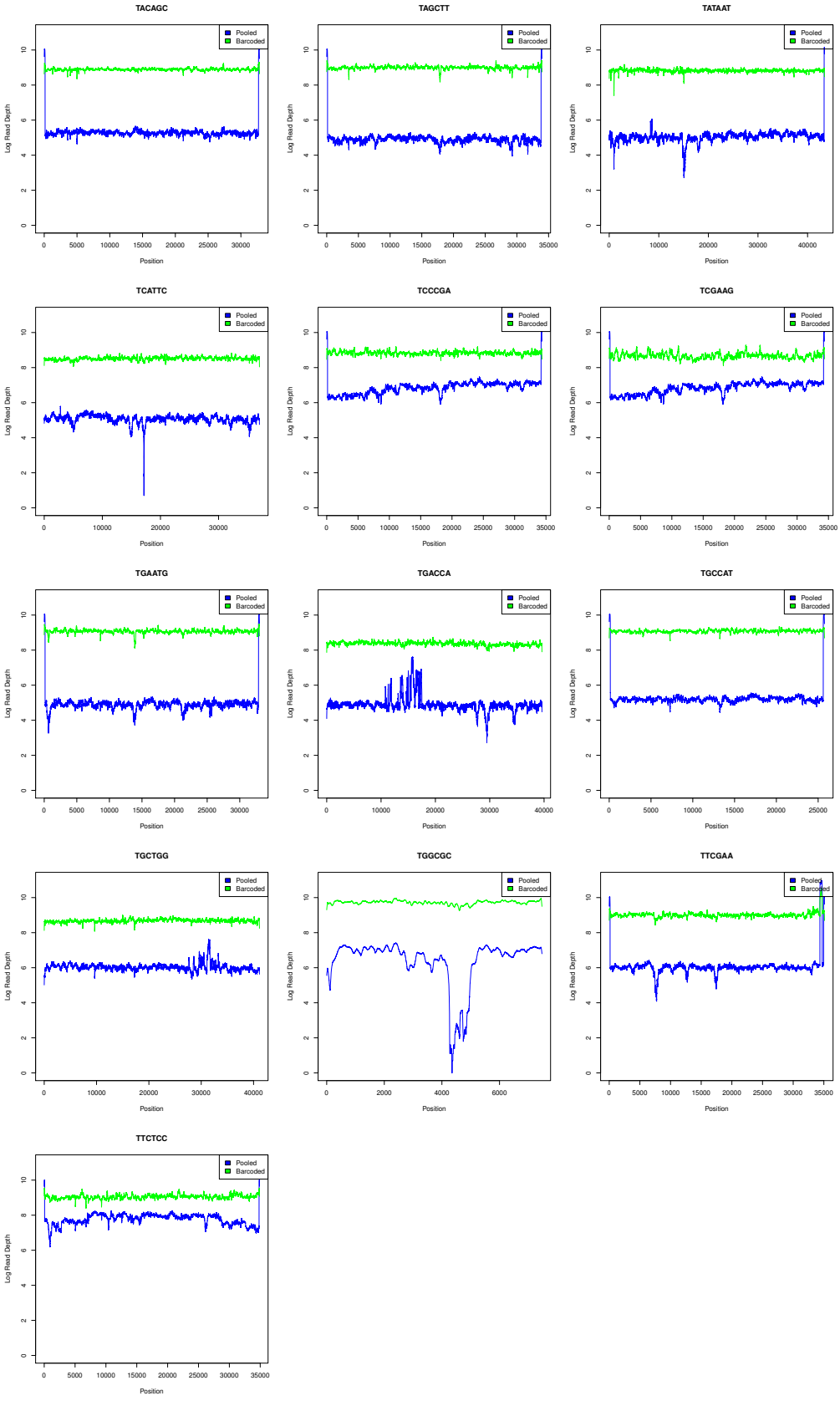
Graphs of clone sequencing read depth for each of the 73 clones are presented on the following pages. For each clone, the barcoded contig was used as a reference, to which raw reads from pooled sequencing or barcoded sequencing were aligned.


AAACAT
AAAGCA
AACCCC
AACTTG
AAGGAC
AATAGG
ACAAAC
ACAGTG
ACATCT
ACCCAG
ACCGGC
ACGATA
ACTGAT
ACTTGA
AGAAGA
Pooled
Barcoded
Log Read Depth
Position


AGATAG
AGCATC
AGCGCT
AGGTTT
AGTTCC
ATAATT
ATCACG
ATCCTA
ATCTAT
ATGAGC
ATTCCT
CAAAAG
CAACTA
CACCGG
CACTCA
Pooled
Barcoded
Log Read Depth
Position


CAGATC
CAGGCG
CATGGC
CCACGC
CCCATG
CCGCAA
CCGTCC
CCTTAG
CGGAAT
CGTACG
CTATAC
CTCAGA
CTGCTG
CTTGTA
GAAACC
Pooled
Barcoded
Log Read Depth
Position


GAATAA
GAGTGG
GATATA
GATCAG
GCAAGG
GCACTT
GCCAAT
GCCGCG
GCTCCA
GGCACA
GTAGAG
GTCCGC
GTGAAA
GTGGCC
TAATCG
Pooled
Barcoded
Log Read Depth
Position


TACAGC
TAGCTT
TATAAT
TCATTC
TCCCGA
TCGAAG
TGAATG
TGACCA
TGCCAT
TGCTGG
TGGCGC
TTCGAA
TTCTCC
Pooled
Barcoded
Log Read Depth
Position

B.2 Python scripts

Parse all-by-all BLAST results

```
#!/usr/bin/python

from Bio.Blast import NCBIXML
'''
#From command line, execute all-by-all blastn to generate results.xml:
#blastn -query contigs-5.fa -subject contigs-5.fa -evalue .001 -out results.xml
 ↪  -outfmt 5
'''

from interval import Interval, IntervalSet

file = open("results.xml")
blast_records = NCBIXML.parse(file)

##accumulate distance between contig pairs in dictionary (where 1 = identical)
distance = {}

##for each queried sequence
for blast_record in blast_records:
    #print "\n" + blast_record.query
    #print str(blast_record.query_letters)

    ##for each subject sequence
    for alignment in blast_record.alignments:

        ##accumulate hsp intervals for each subject sequence, by iterating
 ↪  through each hsp
        hsp_interval_list = []
        for hsp in alignment.hsps:

            ##if alignment was on subject complement, subtract alignment length
 ↪  from start to get interval
            if hsp.frame == (1,-1):
                hsp_interval = IntervalSet([Interval(hsp.sbjct_start,
 ↪  hsp.sbjct_start - hsp.align_length)])
                hsp_interval_list.append(hsp_interval)

            ##otherwise, alignment was on subject given strand, add alignment
 ↪  length to start to get interval
            else:
                hsp_interval = IntervalSet([Interval(hsp.sbjct_start,
 ↪  hsp.sbjct_start + hsp.align_length)])
                hsp_interval_list.append(hsp_interval)

```

```
39
40          ##use interval addition to remove overlapping regions over hsps
41          new_intervalset = IntervalSet()
42          for interval in hsp_interval_list:
43                  new_intervalset = new_intervalset + interval
44
45          ##calculate length of the subject sequence that was involved in the
    ↪   alignment = [aligned length]
46          range_list =[]
47          for interval in new_intervalset:
48              start = interval.lower_bound
49              end = interval.upper_bound
50              for i in range(start, end):
51                  range_list.append(i)
52
53          ##check which of query/subject is shorter; then divide the [aligned
    ↪   length] by length of the shorter one
54          ##note: blast_record.query_letters = query length; alignment.length =
    ↪   subject length
55          ##keep track of the fraction and query/subject names for putting in dict
56          fraction = 0
57          if blast_record.query_letters <= alignment.length:
58              fraction = float(len(range_list))/blast_record.query_letters
59          else:
60              fraction = float(len(range_list))/alignment.length
61
62          ##save the fraction (distance), which represents the homology between
    ↪   the query and subject
63          ##put the names into a list to sort; this overwrites duplicate key-value
    ↪   pairs in the dictionary
64          name_pair = [str(blast_record.query), str(alignment.hit_def)]
65          name_pair = sorted(name_pair)
66          new_name_pair = ":".join(name_pair)
67          distance[str(new_name_pair)] = fraction
68
69  ##write distances to file
70  out = open("out.txt", "w")
71  for item in distance:
72      #print item + "\t\t\t" + str(distance[item])
73      names = item.split(":")
74      row = names[0] + "," + names[1] + "," + str(distance[item]) + "\n"
75      out.write(row)
76
77
```

Appendix C

Supplementary information for Chapter 4

C.1 MetaPhlAn output of taxa abundance

The output of MetaPhlAn is presented on the following pages (Table C.1); results for both the forward and reverse reads are included for all three samples: the crude extract, the size-selected, and the cosmid library.

Table C.1: Summary of Metaphlan output.

ID	cosmid'library'F	cosmid'library'R	size'selected'F	size'selected'R	crude'extract'F	crude'extract'R
k"Archaea	0	0	0.00435	0.00178	0	0
k"Archaea—p"Euryarchaeota	0	0	0.00435	0.00178	0	0
k"Archaea—p"Euryarchaeota—c"Methanobacteria	0	0	0.00435	0.00178	0	0
k"Archaea—p"Euryarchaeota—c"Methanobacteria—o"Methanobacteriales	0	0	0.00435	0.00178	0	0
k"Archaea—p"Euryarchaeota—c"Methanobacteria—o"Methanobacteriales—f"Methanobacteriaceae	0	0	0.00435	0.00178	0	0
k"Archaea—p"Euryarchaeota—c"Methanobacteria—o"Methanobacteriales—f"Methanobacteriaceae—g"Methanosphaera	0	0	0.00435	0.00178	0	0
k"Archaea—p"Euryarchaeota—c"Methanobacteria—o"Methanobacteriales—f"Methanobacteriaceae—g"Methanosphaera—s"Methanosphaera'stadtmanae	0	0	0.00435	0.00178	0	0
k"Bacteria	100	100	99.99565	99.99822	100	100
k"Bacteria—p"Actinobacteria	77.97123	77.44348	13.0776	12.73214	11.14398	10.89272
k"Bacteria—p"Actinobacteria—c"Actinobacteria	77.97123	77.44348	13.0776	12.73214	11.14398	10.89272
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Actinomycetales	0.05618	0.05321	0	0	0.00107	0
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Actinomycetales—f"Actinomycetaceae	0.01117	0.01193	0	0	0.00107	0
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Actinomycetales—f"Actinomycetaceae—g"Actinomyces	0.01117	0.01193	0	0	0.00107	0
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Actinomycetales—f"Actinomycetaceae—g"Actinomyces—s"Actinomyces'odontolyticus	0.01117	0.01193	0	0	0.00107	0
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Actinomycetales—f"Micrococcaceae	0.045	0.04128	0	0	0	0
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Actinomycetales—f"Micrococcaceae—g"Rothia	0.045	0.04128	0	0	0	0
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Actinomycetales—f"Micrococcaceae—g"Rothia—s"Rothia'mucilaginosa	0.045	0.04128	0	0	0	0
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Bifidobacteriales	54.98204	54.69089	9.20296	8.88183	7.87557	7.72522
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Bifidobacteriales—f"Bifidobacteriaceae	54.98204	54.69089	9.20296	8.88183	7.87557	7.72522
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Bifidobacteriales—f"Bifidobacteriaceae—g"Bifidobacterium	54.98204	54.69089	9.20296	8.88183	7.87557	7.72522
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Bifidobacteriales—f"Bifidobacteriaceae—g"Bifidobacterium—s"Bifidobacterium'adolescentis	18.44349	18.6723	2.55627	2.60934	2.23682	2.20449
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Bifidobacteriales—f"Bifidobacteriaceae—g"Bifidobacterium—s"Bifidobacterium'breve	0.24638	0.22541	0.00578	0.00119	0	0
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Bifidobacteriales—f"Bifidobacteriaceae—g"Bifidobacterium—s"Bifidobacterium'catenulatum	0.41224	0.39189	0.02275	0.03965	0.03462	0.04135
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Bifidobacteriales—f"Bifidobacteriaceae—g"Bifidobacterium—s"Bifidobacterium'longum	16.0106	16.00286	3.40731	2.95913	2.92716	2.95269
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Bifidobacteriales—f"Bifidobacteriaceae—g"Bifidobacterium—s"Bifidobacterium'pseudocatenulatum	19.86934	19.39843	3.21084	3.27252	2.67697	2.52669
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Coriobacteriales	22.93302	22.69937	3.87464	3.85031	3.26734	3.1675
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Coriobacteriales—f"Coriobacteriaceae	22.93302	22.69937	3.87464	3.85031	3.26734	3.1675
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Coriobacteriales—f"Coriobacteriaceae—g"Collinsella	21.21125	20.97976	3.40455	3.37866	2.95099	2.88259
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Coriobacteriales—f"Coriobacteriaceae—g"Collinsella—s"Collinsella'aerofaciens	21.21125	20.97976	3.40455	3.37866	2.95099	2.88259
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Coriobacteriales—f"Coriobacteriaceae—g"Eggerthella	1.32363	1.30562	0.33117	0.33524	0.23774	0.22797
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Coriobacteriales—f"Coriobacteriaceae—g"Eggerthella—s"Eggerthella'lenta	1.32363	1.30562	0.33117	0.33524	0.23774	0.22797
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Coriobacteriales—f"Coriobacteriaceae—g"Gordoniibacter	0.39814	0.414	0.13892	0.13641	0.07862	0.05694
k"Bacteria—p"Actinobacteria—c"Actinobacteria—o"Coriobacteriales—f"Coriobacteriaceae—g"Gordoniibacter—s"Gordoniibacter'pamelaeae	0.39814	0.414	0.13892	0.13641	0.07862	0.05694
k"Bacteria—p"Bacteroidetes	12.02304	12.32339	21.24834	21.47013	29.50008	29.97218
k"Bacteria—p"Bacteroidetes—c"Bacteroidia	12.02304	12.32339	21.24834	21.47013	29.50008	29.97218
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales	12.02304	12.32339	21.24834	21.47013	29.50008	29.97218
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Bacteroidaceae	4.38677	4.55248	16.24903	16.56666	19.28831	19.85961
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Bacteroidaceae—g"Bacteroides	4.38677	4.55248	16.24903	16.56666	19.28831	19.85961
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Bacteroidaceae—g"Bacteroides—s"Bacteroides'caccae	0.02079	0.02183	0.09667	0.09511	0.11722	0.11225
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Bacteroidaceae—g"Bacteroides—s"Bacteroides'cellulosilyticus	0.14931	0.13825	0.4679	0.46279	0.58924	0.56722
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Bacteroidaceae—g"Bacteroides—s"Bacteroides'coprocola	0	0	0.00149	0	0.02463	0.02992
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Bacteroidaceae—g"Bacteroides—s"Bacteroides'dorei	0.04273	0.04379	0.13965	0.17031	0.16946	0.18184
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Bacteroidaceae—g"Bacteroides—s"Bacteroides'eggerthii	0	0	0.01255	0.00905	0.02386	0.01879
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Bacteroidaceae—g"Bacteroides—s"Bacteroides'finegoldii	0.01691	0.01896	0.09499	0.12025	0.16818	0.16985
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Bacteroidaceae—g"Bacteroides—s"Bacteroides'fragilis	0.00975	0.00487	0.16213	0.16516	0.15296	0.10438
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Bacteroidaceae—g"Bacteroides—s"Bacteroides'intestinalis	0	0	0.01787	0.01626	0.02399	0.02366
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Bacteroidaceae—g"Bacteroides—s"Bacteroides'ovatus	0.2891	0.2753	1.00239	0.96548	1.35037	1.24824
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Bacteroidaceae—g"Bacteroides—s"Bacteroides'stercoris	0	0	0.00534	0	0.04992	0.0393
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Bacteroidaceae—g"Bacteroides—s"Bacteroides'thetaiotaomicron	0.37345	0.34607	1.03798	1.0592	1.00328	1.09393
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Bacteroidaceae—g"Bacteroides—s"Bacteroides'unclassified	1.94406	2.20922	7.37793	7.50927	9.23866	9.92434
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Bacteroidaceae—g"Bacteroides—s"Bacteroides'uniformis	0.20634	0.22197	0.66519	0.67599	0.76074	0.72604
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Bacteroidaceae—g"Bacteroides—s"Bacteroides'vulgatus	1.24732	1.23935	4.96679	5.15465	5.39491	5.41351
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Bacteroidaceae—g"Bacteroides—s"Bacteroides'xylanisolvens	0.087	0.03287	0.20016	0.16315	0.22089	0.20633
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Porphyromonadaceae	0.45788	0.57906	0.53245	0.51046	0.64013	0.66208
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Porphyromonadaceae—g"Odoribacter	0	0	0.00323	0.00411	0.02077	0.01833
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Porphyromonadaceae—g"Odoribacter—s"Odoribacter'splanchnicus	0	0	0.00323	0.00411	0.02077	0.01833
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Porphyromonadaceae—g"Parabacteroides	0.45788	0.57906	0.52921	0.50635	0.61936	0.64376
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Porphyromonadaceae—g"Parabacteroides—s"Parabacteroides'distasonis	0	0	0.01105	0.00648	0.01848	0.02003
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Porphyromonadaceae—g"Parabacteroides—s"Parabacteroides'johnsonii	0.00907	0.00771	0.03531	0.03845	0.06263	0.0511
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Porphyromonadaceae—g"Parabacteroides—s"Parabacteroides'merdae	0.14744	0.15162	0.4486	0.44856	0.46743	0.48847
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Porphyromonadaceae—g"Parabacteroides—s"Parabacteroides'unclassified	0.30136	0.41973	0.03425	0.01287	0.07082	0.08416
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Prevotellaceae	0.24934	0.24731	1.55259	1.55722	5.90395	5.86084
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Prevotellaceae—g"Prevotella	0.24934	0.24731	1.55259	1.55722	5.90395	5.86084
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Prevotellaceae—g"Prevotella—s"Prevotella'copri	0.24934	0.24731	1.55259	1.55722	5.90395	5.86084
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Rikenellaceae	6.92906	6.94454	2.91427	2.83578	3.66768	3.58964
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Rikenellaceae—g"Alistipes	6.92906	6.94454	2.91427	2.83578	3.66768	3.58964
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Rikenellaceae—g"Alistipes—s"Alistipes'putredinis	4.07745	4.16533	1.93769	1.90829	2.52303	2.4627
k"Bacteria—p"Bacteroidetes—c"Bacteroidia—o"Bacteroidales—f"Rikenellaceae—g"Alistipes—s"Alistipes'shahii	2.85161	2.77921	0.97658	0.92749	1.14465	1.12694
k"Bacteria—p"Firmicutes	0.12246	0.1209	64.05676	64.28564	57.83246	57.5862
k"Bacteria—p"Firmicutes—c"Bacilli	0	0	1.1275	1.12086	0.77354	0.81422
k"Bacteria—p"Firmicutes—c"Bacilli—o"Lactobacillales	0	0	1.1275	1.12086	0.77354	0.81422
k"Bacteria—p"Firmicutes—c"Bacilli—o"Lactobacillales—f"Lactobacillaceae	0	0	0.2723	0.27887	0.18615	0.18228
k"Bacteria—p"Firmicutes—c"Bacilli—o"Lactobacillales—f"Lactobacillaceae—g"Lactobacillus	0	0	0.2723	0.27887	0.18615	0.18228
k"Bacteria—p"Firmicutes—c"Bacilli—o"Lactobacillales—f"Lactobacillaceae—g"Lactobacillus—s"Lactobacillus'ruminis	0	0	0.2723	0.27887	0.18615	0.18228
k"Bacteria—p"Firmicutes—c"Bacilli—o"Lactobacillales—f"Streptococcaceae	0	0	0.8552	0.84199	0.58739	0.63194
k"Bacteria—p"Firmicutes—c"Bacilli—o"Lactobacillales—f"Streptococcaceae—g"Streptococcus	0	0	0.8552	0.84199	0.58739	0.63194
k"Bacteria—p"Firmicutes—c"Bacilli—o"Lactobacillales—f"Streptococcaceae—g"Streptococcus—s"Streptococcus'australis	0	0	0	0	0.00098	0
k"Bacteria—p"Firmicutes—c"Bacilli—o"Lactobacillales—f"Streptococcaceae—g"Streptococcus—s"Streptococcus'parasanguinis	0	0	0.65839	0.62618	0.40991	0.44214
k"Bacteria—p"Firmicutes—c"Bacilli—o"Lactobacillales—f"Streptococcaceae—g"Streptococcus—s"Streptococcus'salivarius	0	0	0.19291	0.20442	0.15672	0.16197
k"Bacteria—p"Firmicutes—c"Bacilli—o"Lactobacillales—f"Streptococcaceae—g"Streptococcus—s"Streptococcus'thermophilus	0	0	0.0039	0.01139	0.01977	0.02783
k"Bacteria—p"Firmicutes—c"Clostridia	0.11169	0.112	59.90936	60.12072	53.74329	53.51115
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales	0.11169	0.112	59.90936	60.12072	53.74329	53.51115
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Clostridiaceae	0	0	2.14418	2.13152	2.01952	2.05992
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Clostridiaceae—g"Clostridium	0	0	2.14418	2.13152	2.01952	2.05992
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Clostridiaceae—g"Clostridium—s"Clostridium'asparagiforme	0	0	0	0	0.092	0.09041
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Clostridiaceae—g"Clostridium—s"Clostridium'bartlettii	0	0	0.09874	0.09638	0.13193	0.13979
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Clostridiaceae—g"Clostridium—s"Clostridium'boltease	0	0	0.24807	0.23127	0.37966	0.3917
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Clostridiaceae—g"Clostridium—s"Clostridium'cf	0	0	0.0322	0.03402	0.02676	0.03709
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Clostridiaceae—g"Clostridium—s"Clostridium'hathewayi	0	0	0.04425	0.04094	0.03153	0.02295
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Clostridiaceae—g"Clostridium—s"Clostridium'leptum	0	0	0.53848	0.53178	0.36475	0.39324
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Clostridiaceae—g"Clostridium—s"Clostridium'nexile	0	0	1.00147	1.01887	0.834	0.83276
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Clostridiaceae—g"Clostridium—s"Clostridium'scindens	0	0	0.10052	0.1023	0.05908	0.0583

k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Clostridiaceae—g"Clostridium—s"Clostridium'symbiosum	0	0	0.08044	0.07595	0.09982	0.09368
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Clostridiales'Family XI Incertae Sedis	0	0	0	0.01519	0	0
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Clostridiales'Family XI Incertae Sedis—g"Clostridiales'Family XI Incertae Sedis'unclassified	0	0	0	0.01519	0	0
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Clostridiales'uncl	0	0	0.17969	0.14638	0.17607	0.12738
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Clostridiales'uncl—g"Blautia	0	0	0.17969	0.14638	0.17607	0.12738
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Clostridiales'uncl—g"Blautia—s"Blautia'hydrogenotrophica	0	0	0.01461	0.01872	0.01027	0.00784
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Clostridiales'uncl—g"Blautia—s"Blautia'unclassified	0	0	0.16508	0.12766	0.16581	0.11954
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Eubacteriaceae	0.03335	0.04003	24.42598	24.53932	19.17275	19.06187
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Eubacteriaceae—g"Eubacterium	0.03335	0.04003	24.42598	24.53932	19.17275	19.06187
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Eubacteriaceae—g"Eubacterium—s"Eubacterium'eligens	0	0	1.01733	1.00474	2.74208	2.69392
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Eubacteriaceae—g"Eubacterium—s"Eubacterium'hallii	0	0	2.03682	1.99023	1.17307	1.17818
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Eubacteriaceae—g"Eubacterium—s"Eubacterium'limosum	0	0	0.00953	0.00807	0	0.0017
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Eubacteriaceae—g"Eubacterium—s"Eubacterium'rectale	0.03335	0.04003	21.04378	21.22138	14.77741	14.73719
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Eubacteriaceae—g"Eubacterium—s"Eubacterium'siraeum	0	0	0.02265	0.02074	0.05463	0.05181
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Eubacteriaceae—g"Eubacterium—s"Eubacterium'ventriosum	0	0	0.29587	0.29417	0.42556	0.39907
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Lachnospiraceae	0	0	5.0478	5.04547	4.05289	4.08144
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Lachnospiraceae—g"Coprococcus	0	0	1.31986	1.35062	0.96688	0.94963
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Lachnospiraceae—g"Coprococcus—s"Coprococcus'catus	0	0	0.28149	0.27011	0.24307	0.2411
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Lachnospiraceae—g"Coprococcus—s"Coprococcus'comes	0	0	1.03837	1.08051	0.72381	0.70853
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Lachnospiraceae—g"Dorea	0	0	3.42776	3.43285	2.36608	2.37726
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Lachnospiraceae—g"Dorea—s"Dorea'formicigenerans	0	0	1.12478	1.10088	0.73896	0.73629
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Lachnospiraceae—g"Dorea—s"Dorea'longicatena	0	0	2.30298	2.33198	1.62712	1.64097
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Lachnospiraceae—g"Roseburia	0	0	0.30018	0.262	0.71993	0.75456
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Lachnospiraceae—g"Roseburia—s"Roseburia'intestinalis	0	0	0.08844	0.06943	0.1149	0.12747
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Lachnospiraceae—g"Roseburia—s"Roseburia'inulinivorans	0	0	0.21175	0.19257	0.60503	0.62708
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Ruminococcaceae	0.07834	0.07197	28.11171	28.24284	28.32206	28.18054
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Ruminococcaceae—g"Anaerotruncus	0	0	0.07834	0.08228	0.0747	0.07836
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Ruminococcaceae—g"Anaerotruncus—s"Anaerotruncus'colihominis	0	0	0.07834	0.08228	0.0747	0.07836
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Ruminococcaceae—g"Faecalibacterium	0.06524	0.06307	11.96027	11.82916	17.6229	17.44306
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Ruminococcaceae—g"Faecalibacterium—s"Faecalibacterium'cf	0.01223	0.01123	4.52787	4.39973	6.195	6.21572
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Ruminococcaceae—g"Faecalibacterium—s"Faecalibacterium'prausnitzii	0.05301	0.05183	6.87077	6.68096	9.82464	10.16358
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Ruminococcaceae—g"Faecalibacterium—s"Faecalibacterium'unclassified	0	0	0.56163	0.74848	1.60325	1.06376
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Ruminococcaceae—g"Ruminococcus	0.0131	0.00891	16.05678	16.32131	10.61971	10.64383
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Ruminococcaceae—g"Ruminococcus—s"Ruminococcus'bromii	0.0131	0.00891	11.80023	11.97115	7.02601	6.99219
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Ruminococcaceae—g"Ruminococcus—s"Ruminococcus'gnavus	0	0	0.96046	0.92369	0.60017	0.60577
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Ruminococcaceae—g"Ruminococcus—s"Ruminococcus'lactaris	0	0	0.1605	0.15906	0.14736	0.13247
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Ruminococcaceae—g"Ruminococcus—s"Ruminococcus'obeum	0	0	1.31916	1.32292	0.88979	0.87322
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Ruminococcaceae—g"Ruminococcus—s"Ruminococcus'torques	0	0	1.81643	1.9445	1.95638	2.04018
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Ruminococcaceae—g"Subdoligranulum	0	0	0.01632	0.01009	0.00475	0.01528
k"Bacteria—p"Firmicutes—c"Clostridia—o"Clostridiales—f"Ruminococcaceae—g"Subdoligranulum—s"Subdoligranulum'variabile	0	0	0.01632	0.01009	0.00475	0.01528
k"Bacteria—p"Firmicutes—c"Erysipelotrichi	0	0	0.67104	0.67619	0.49115	0.47528
k"Bacteria—p"Firmicutes—c"Erysipelotrichi—o"Erysipelotrichales	0	0	0.67104	0.67619	0.49115	0.47528
k"Bacteria—p"Firmicutes—c"Erysipelotrichi—o"Erysipelotrichales—f"Erysipelotrichaceae	0	0	0.67104	0.67619	0.49115	0.47528
k"Bacteria—p"Firmicutes—c"Erysipelotrichi—o"Erysipelotrichales—f"Erysipelotrichaceae—g"Catenibacterium	0	0	0.21822	0.21838	0.15408	0.14172
k"Bacteria—p"Firmicutes—c"Erysipelotrichi—o"Erysipelotrichales—f"Erysipelotrichaceae—g"Catenibacterium—s"Catenibacterium'mitsuokai	0	0	0.21822	0.21838	0.15408	0.14172
k"Bacteria—p"Firmicutes—c"Erysipelotrichi—o"Erysipelotrichales—f"Erysipelotrichaceae—g"Coprobacillus	0	0	0.17556	0.1848	0.10687	0.10931
k"Bacteria—p"Firmicutes—c"Erysipelotrichi—o"Erysipelotrichales—f"Erysipelotrichaceae—g"Coprobacillus—s"Coprobacillus'bacterium	0	0	0.17556	0.1848	0.10687	0.10931
k"Bacteria—p"Firmicutes—c"Erysipelotrichi—o"Erysipelotrichales—f"Erysipelotrichaceae—g"Holdemania	0	0	0.27726	0.27302	0.23019	0.22424
k"Bacteria—p"Firmicutes—c"Erysipelotrichi—o"Erysipelotrichales—f"Erysipelotrichaceae—g"Holdemania—s"Holdemania'filiformis	0	0	0.27726	0.27302	0.23019	0.22424
k"Bacteria—p"Firmicutes—c"Negativicutes	0.01077	0.0089	2.34886	2.36787	2.82449	2.78555
k"Bacteria—p"Firmicutes—c"Negativicutes—o"Selenomonadales	0.01077	0.0089	2.34886	2.36787	2.82449	2.78555
k"Bacteria—p"Firmicutes—c"Negativicutes—o"Selenomonadales—f"Acidaminococcaceae	0	0	0.14545	0.16028	0.19088	0.15305
k"Bacteria—p"Firmicutes—c"Negativicutes—o"Selenomonadales—f"Acidaminococcaceae—g"Acidaminococcaceae'unclassified	0	0	0.14545	0.16028	0.19088	0.15305
k"Bacteria—p"Firmicutes—c"Negativicutes—o"Selenomonadales—f"Veillonellaceae	0.01077	0.0089	2.20341	2.20758	2.63361	2.6325
k"Bacteria—p"Firmicutes—c"Negativicutes—o"Selenomonadales—f"Veillonellaceae—g"Dialister	0.01077	0.0089	1.6548	1.67255	1.16702	1.19352
k"Bacteria—p"Firmicutes—c"Negativicutes—o"Selenomonadales—f"Veillonellaceae—g"Dialister—s"Dialister'invisus	0.01077	0.0089	1.6548	1.67255	1.16702	1.19352
k"Bacteria—p"Firmicutes—c"Negativicutes—o"Selenomonadales—f"Veillonellaceae—g"Megamonas	0	0	0.54684	0.53234	1.46205	1.42808
k"Bacteria—p"Firmicutes—c"Negativicutes—o"Selenomonadales—f"Veillonellaceae—g"Megamonas—s"Megamonas'hypermegale	0	0	0.54684	0.53234	1.46205	1.42808
k"Bacteria—p"Firmicutes—c"Negativicutes—o"Selenomonadales—f"Veillonellaceae—g"Veillonella	0	0	0.00176	0.00269	0.00454	0.01091
k"Bacteria—p"Firmicutes—c"Negativicutes—o"Selenomonadales—f"Veillonellaceae—g"Veillonella—s"Veillonella'unclassified	0	0	0.00176	0.00269	0.00454	0.01091
k"Bacteria—p"Proteobacteria	9.65819	9.90084	1.57642	1.47372	1.49978	1.52766
k"Bacteria—p"Proteobacteria—c"Deltaproteobacteria	0.68372	0.77154	0.30892	0.28465	0.44194	0.51069
k"Bacteria—p"Proteobacteria—c"Deltaproteobacteria—o"Desulfovibrionales	0.68372	0.77154	0.30892	0.28465	0.44194	0.51069
k"Bacteria—p"Proteobacteria—c"Deltaproteobacteria—o"Desulfovibrionales—f"Desulfovibrionaceae	0.68372	0.77154	0.30892	0.28465	0.44194	0.51069
k"Bacteria—p"Proteobacteria—c"Deltaproteobacteria—o"Desulfovibrionales—f"Desulfovibrionaceae—g"Bilophila	0.4798	0.44882	0.12539	0.1176	0.16021	0.15952
k"Bacteria—p"Proteobacteria—c"Deltaproteobacteria—o"Desulfovibrionales—f"Desulfovibrionaceae—g"Bilophila—s"Bilophila'wadsworthia	0.4798	0.44882	0.12539	0.1176	0.16021	0.15952
k"Bacteria—p"Proteobacteria—c"Deltaproteobacteria—o"Desulfovibrionales—f"Desulfovibrionaceae—g"Desulfovibrio	0.20392	0.32272	0.18353	0.16705	0.28173	0.35117
k"Bacteria—p"Proteobacteria—c"Deltaproteobacteria—o"Desulfovibrionales—f"Desulfovibrionaceae—g"Desulfovibrio—s"Desulfovibrio'desulfuricans	0.06007	0.16577	0.07934	0.04913	0.08314	0.1506
k"Bacteria—p"Proteobacteria—c"Deltaproteobacteria—o"Desulfovibrionales—f"Desulfovibrionaceae—g"Desulfovibrio—s"Desulfovibrio'piger	0.14385	0.15694	0.10419	0.11792	0.19858	0.20057
k"Bacteria—p"Proteobacteria—c"Gammaproteobacteria	8.97448	9.1293	1.2675	1.18907	1.05785	1.01697
k"Bacteria—p"Proteobacteria—c"Gammaproteobacteria—o"Alteromonadales	0.01216	0	0	0	0.03468	0.0038
k"Bacteria—p"Proteobacteria—c"Gammaproteobacteria—o"Alteromonadales—f"Shewanellaceae	0.01216	0	0	0	0.03468	0.0038
k"Bacteria—p"Proteobacteria—c"Gammaproteobacteria—o"Alteromonadales—f"Shewanellaceae—g"Shewanella	0.01216	0	0	0	0.03468	0.0038
k"Bacteria—p"Proteobacteria—c"Gammaproteobacteria—o"Alteromonadales—f"Shewanellaceae—g"Shewanella—s"Shewanella'oneidensis	0.01216	0	0	0	0.03468	0.0038
k"Bacteria—p"Proteobacteria—c"Gammaproteobacteria—o"Enterobacteriales	8.95048	9.12138	1.09247	1.00767	0.94133	0.91524
k"Bacteria—p"Proteobacteria—c"Gammaproteobacteria—o"Enterobacteriales—f"Enterobacteriaceae	8.95048	9.12138	1.09247	1.00767	0.94133	0.91524
k"Bacteria—p"Proteobacteria—c"Gammaproteobacteria—o"Enterobacteriales—f"Enterobacteriaceae—g"Enterobacter	0.08686	0.07763	0	0	0	0
k"Bacteria—p"Proteobacteria—c"Gammaproteobacteria—o"Enterobacteriales—f"Enterobacteriaceae—g"Enterobacter—s"Enterobacter'cloacae	0.08686	0.07763	0	0	0	0
k"Bacteria—p"Proteobacteria—c"Gammaproteobacteria—o"Enterobacteriales—f"Enterobacteriaceae—g"Escherichia	5.16069	5.22912	0.65059	0.56245	0.43353	0.45782
k"Bacteria—p"Proteobacteria—c"Gammaproteobacteria—o"Enterobacteriales—f"Enterobacteriaceae—g"Escherichia—s"Escherichia'coli	5.16069	5.22912	0.65059	0.56245	0.43353	0.45782
k"Bacteria—p"Proteobacteria—c"Gammaproteobacteria—o"Enterobacteriales—f"Enterobacteriaceae—g"Klebsiella	3.70294	3.81462	0.44188	0.44522	0.5078	0.45742
k"Bacteria—p"Proteobacteria—c"Gammaproteobacteria—o"Enterobacteriales—f"Enterobacteriaceae—g"Klebsiella—s"Klebsiella'pneumoniae	3.70294	3.81462	0.42971	0.44389	0.45181	0.40999
k"Bacteria—p"Proteobacteria—c"Gammaproteobacteria—o"Enterobacteriales—f"Enterobacteriaceae—g"Klebsiella—s"Klebsiella'unclassified	0	0	0.01218	0.00133	0.05598	0.04744
k"Bacteria—p"Proteobacteria—c"Gammaproteobacteria—o"Pasteurellales	0.01184	0.00793	0.17502	0.1814	0.08184	0.09792
k"Bacteria—p"Proteobacteria—c"Gammaproteobacteria—o"Pasteurellales—f"Pasteurellaceae	0.01184	0.00793	0.17502	0.1814	0.08184	0.09792
k"Bacteria—p"Proteobacteria—c"Gammaproteobacteria—o"Pasteurellales—f"Pasteurellaceae—g"Haemophilus	0.01184	0.00793	0.17502	0.1814	0.08184	0.09792
k"Bacteria—p"Proteobacteria—c"Gammaproteobacteria—o"Pasteurellales—f"Pasteurellaceae—g"Haemophilus—s"Haemophilus'parainfluenzae	0.01184	0.00793	0.17502	0.1814	0.08184	0.09792
k"Bacteria—p"Verrucomicrobia	0.22507	0.2114	0.03653	0.03659	0.02369	0.02124
k"Bacteria—p"Verrucomicrobia—c"Verrucomicrobiae	0.22507	0.2114	0.03653	0.03659	0.02369	0.02124
k"Bacteria—p"Verrucomicrobia—c"Verrucomicrobiae—o"Verrucomicrobiales	0.22507	0.2114	0.03653	0.03659	0.02369	0.02124
k"Bacteria—p"Verrucomicrobia—c"Verrucomicrobiae—o"Verrucomicrobiales—f"Verrucomicrobiaceae	0.22507	0.2114	0.03653	0.03659	0.02369	0.02124
k"Bacteria—p"Verrucomicrobia—c"Verrucomicrobiae—o"Verrucomicrobiales—f"Verrucomicrobiaceae—g"Akkermansia	0.22507	0.2114	0.03653	0.03659	0.02369	0.02124
k"Bacteria—p"Verrucomicrobia—c"Verrucomicrobiae—o"Verrucomicrobiales—f"Verrucomicrobiaceae—g"Akkermansia—s"Akkermansia'muciniphila	0.22507	0.2114	0.03653	0.03659	0.02369	0.02124

C.2 Python scripts

Filter *E. coli* or vector reads using BLAT, in batch

```
#!/usr/bin/python

from Bio import SeqIO
import sys
import os
import time

#FUNCTIONS

#run blat and parse results; return a set of unique read names that are hits to the subject
def run_blat(files_dir, reads_filename, subject_filename):

    #run blat in the shell
    results_filename = reads_filename + "_BLAT_" + subject_filename + ".psl"
    os.system("blat " + subject_filename + " " + files_dir + reads_filename + " " + files_dir + results_filename)

    #open results
    results_file = open(files_dir + results_filename)

    #clear the header lines
    for i in range(0,5):
        results_file.readline()

    #track the names of reads that are 100% identical to E. coli (90 base identity)
    match_names = set()
    for line in results_file:

        #parse the line
        line = line.split('\t')
        match = line[0]
        mismatch = line[1]
        gaps = line[6]
        query_name = line[9]

        #if the match was 100% identical (90 bases), accumulate the name
        if match == '90' and mismatch == '0' and gaps == '0':
            match_names.add(query_name)

    #delete psl files
    os.system("rm " + files_dir + "*.psl")

```

```
42      return match_names
43
44  #INPUT FILES
45
46  filenames_dir = sys.argv[1]
47  vector_filename = sys.argv[2]
48  ec_filename = sys.argv[3]
49
50  #get list of filenames into array to process
51  filenames = os.listdir(filenames_dir)
52  filenames.sort()
53
54  #RUN BLAT AND PARSE RESULTS FOR EACH FILE
55
56  #write summary file of results
57  summary_file = open(filenames_dir + "summary.txt", "w")
58  summary_file.write("filename \ttotal reads \ttotal dirty \tec \tvector \n")
59
60  #process files
61  for filename in filenames:
62
63      #get sets of read names that are hits
64      ec_hits = run_blat(filenames_dir, filename, ec_filename)
65      vector_hits = run_blat(filenames_dir, filename, vector_filename)
66
67      #track for summary file
68      total_count = 0
69      total_dirty_count = 0
70      vector_count = 0
71      ec_count = 0
72
73      #write clean and dirty reads to  new files; also summary file
74      clean_file = open(filenames_dir + filename + "_clean_chked.fa", "w")
75      dirty_file = open(filenames_dir + filename + "_dirty_chked.fa", "w")
76
77      #open the reads file; for each FASTA sequence read
78      for seq_record in SeqIO.parse(filenames_dir + filename, "fasta"):
79          total_count = total_count + 1
80
81          if (seq_record.id in ec_hits):
82              SeqIO.write(seq_record, dirty_file, "fasta")
83              ec_hits.remove(seq_record.id) #remove id from set to make following
    ↪  searches faster
84              ec_count = ec_count + 1
85              total_dirty_count = total_dirty_count + 1
86
87          elif (seq_record.id in vector_hits):
88              SeqIO.write(seq_record, dirty_file, "fasta")
89              vector_hits.remove(seq_record.id) #remove id from set to make
    ↪  following searches faster
90              vector_count = vector_count + 1
91              total_dirty_count = total_dirty_count + 1
```

```
92
93          #if not in list of read names, it's a clean read
94          else:
95
96              #write to clean file
97              SeqIO.write(seq_record, clean_file, "fasta")
98
99      #write to summary
100     output = filename + "\t" + str(total_count) + "\t" + str(total_dirty_count)
    ↪  + "\t" + str(ec_count) + "\t" + str(vector_count) + "\n"
101     summary_file.write(output)
```

Check filtering of *E. coli* and vector reads, in batch

```
#!/usr/bin/python

from Bio import SeqIO
import sys
import os
import time

#input: directory of files to process; fasta Ec file; fasta vector file
filenames_dir = sys.argv[1]
vector_filename = sys.argv[2]
ec_filename = sys.argv[3]

#get list of filenames into array to process
filenames = os.listdir(filenames_dir)
filenames.sort()

#for ec, vector: get the sequence, rev comp of the sequence, in preparation for checking
ec = SeqIO.read(ec_filename, "fasta")
ec_rc = ec.reverse_complement()
vector = SeqIO.read(vector_filename, "fasta")
vector_rc = vector.reverse_complement()

#prep output file
outfile = open(filenames_dir + "results_Ec_or_pJC8.txt", "w")
outfile.write("filename \ttotal \tboth \tEc \tvector \tunaccounted \n")

#process each file
for filename in filenames:

    #check whether each read in the file is from pJC8 or Ec or both; should not be any unaccounted, but track in case
    both_count = 0
    ec_count = 0
    vector_count = 0
    unaccounted = 0
    total = 0
    unaccounted_file = open(filenames_dir + filename + "_unaccounted_reads", "w")

    for seq_record in SeqIO.parse(filenames_dir + filename, "fasta"):
        total = total + 1

        #if seq in both
        if (seq_record.seq in ec.seq or seq_record.seq in ec_rc.seq):
            ec_count = ec_count + 1
            if (seq_record.seq in vector.seq or seq_record.seq in vector_rc.seq):
```

```
                vector_count = vector_count + 1
                both_count = both_count + 1

        elif (seq_record.seq in vector.seq or seq_record.seq in vector_rc.seq):
            vector_count = vector_count + 1

        #this shouldn't happen
        else:
            unaccounted = unaccounted + 1
            SeqIO.write(seq_record, unaccounted_file, "fasta")

    #write to output file: filename, total num reads, num Ec reads, num pjc8 reads
    output_line = filename + "\t" + str(total) + "\t" + str(both_count) + "\t" + str(ec_count) + "\t" + str(vector_count) + "\t" + str(unaccounted) + "\n"
    outfile.write(output_line)
```

Calculate percent GC, in batch

```
#!/usr/bin/python
from Bio import SeqIO
import sys
import os

#function to calc percent gc from all seqs in a fasta file
def get_gc(files_dir, filename):

    #track number of each base
    bases = {'A':0, 'C':0, 'G':0, 'T':0}

    #open the reads file; for each FASTA sequence, track bases in seq
    for seq_record in SeqIO.parse(files_dir + filename, "fasta"):
        for base in seq_record.seq:
            if base == 'A':
                bases['A'] = bases['A'] + 1
            elif base == 'C':
                bases['C'] = bases['C'] + 1
            elif base == 'G':
                bases['G'] = bases['G'] + 1
            else:
                bases['T'] = bases['T'] + 1

    #do the stats
    total_bases = float(sum(bases.values()))
    gc = (bases['G'] + bases['C']) / total_bases * 100
    return gc

#input file in fasta
filenames_dir = sys.argv[1]
filenames = os.listdir(filenames_dir)
filenames.sort()

#summary file
results_file = open(filenames_dir + "summary.txt", "w")
results_file.write("filename \t%GC \n")

#process each file
for filename in filenames:
    gc = get_gc(filenames_dir, filename)
    output = filename + "\t" + str(gc) + "\n"
    results_file.write(output)
```

Find consensus promoter sequences, in batch

```
#!/usr/bin/python

from Bio import SeqIO
import sys
import os
import re

#FUNCTIONS

#look for consensus sequences 1 promoter; return count
def find_one_consensus(sequence, filename):

    #compile regex
    p = re.compile(sequence)
    count = 0

    #iterate through each fasta sequence
    for seq_record in SeqIO.parse(filename, "fasta"):

        #check the sequence
        for match in p.finditer(str(seq_record.seq)):
            count = count + 1

        #check the reverse complement
        for match in p.finditer(str(seq_record.reverse_complement().seq)):
            count = count + 1

    return count

#look for consensus sequences for 5 promoters; return a string to be printed to file
def find_all_consensus(files_dir, reads_filename):

    #file location
    location = files_dir + reads_filename

    #rpoD sigma 70
    rpod_count = find_one_consensus("TTGACA.{15,19}TATAAT", location)

    #rpoE sigma 24
    rpoe_count = find_one_consensus("GGAACTT.{15,19}TCAAA", location)

    #rpoH sigma 32
    rpoh_count = find_one_consensus("TTG[AT][AT][AT].{13,14}CCCCAT[AT]T", location)

    #rpoN sigma 54
    rpon_count = find_one_consensus("TGGCA.{7}TGC", location)
```

```
47
48      #Bacteroides sigma AB
49      bacteroides_count = find_one_consensus("TTTG.{19,21}TA.{2}TTTG", location)
50
51      output = filename + "\t" + str(rpod_count) + "\t" + str(rpoe_count) + "\t" +
        ↪  str(rpoh_count) + "\t" + str(rpon_count)+ "\t" + str(bacteroides_count) +
        ↪  "\n"
52      return output
53
54  #INPUT FILES
55
56  filenames_dir = sys.argv[1]
57  filenames = os.listdir(filenames_dir)
58  filenames.sort()
59
60  #PROCESS ALL FILES
61
62  #write summary file of results
63  summary_file = open(filenames_dir + "summary.txt", "w")
64  summary_file.write("filename \trpoD reads \trpoE \trpoH \trpoN \tBacteroides
    ↪  \n")
65
66  #process files
67  for filename in filenames:
68
69      #get sets of read names that are hits
70      output = find_all_consensus(filenames_dir, filename)
71
72      #write to summary
73      summary_file.write(output)
```

Calculate phyla percentages from OTU table

```
import sys
import os

otu_filename = sys.argv[1]

#prep outfile
phyla_filename = os.path.splitext(otu_filename)[0] + "_phyla_percent.txt"
phyla_file = open(phyla_filename, "w")

#get otu table
otu_file = open(otu_filename, "r")

#discard first header line
otu_file.readline()

#start dict to keep phyla counts
cosmid = {}
bulk = {}

#process each line, adding to both dicts
for line in otu_file:
    line = line.split(",")
    bulk_count = int(line[1])
    cosmid_count = int(line[2])
    phylum = line[4]

    #check if phylum in either dict and add accordingly
    if phylum in cosmid:
        cosmid[phylum] = cosmid[phylum] + cosmid_count
        bulk[phylum] = bulk[phylum] + bulk_count
    else:
        cosmid[phylum] = cosmid_count
        bulk[phylum] = bulk_count

#given a dictionary of phyla counts, return dict of phyla fractions
def get_phyla_fractions(phyla_dict):

    #get total member count
    total = 0
    for phylum in phyla_dict:
        total = total + phyla_dict[phylum]
    total = float(total)

    #make new dict of fractions
    new_dict = {}
    for phylum in phyla_dict:
        new_dict[phylum] = phyla_dict[phylum]/total

```

```
    return new_dict

cosmid_fraction = get_phyla_fractions(cosmid)
bulk_fraction = get_phyla_fractions(bulk)

#write phyla fractions to new file
for item in cosmid_fraction:
    phyla_file.write(item)
    phyla_file.write("\t")
    phyla_file.write(str(format(cosmid_fraction[item], '.9f')))
    phyla_file.write("\t")
    phyla_file.write(str(format(bulk_fraction[item], '.9f')))
    phyla_file.write("\n")

phyla_file.close()
```

Appendix D

Supplementary information for Chapter 5

D.1 Images

Annealing of oligos KL10 and KL11

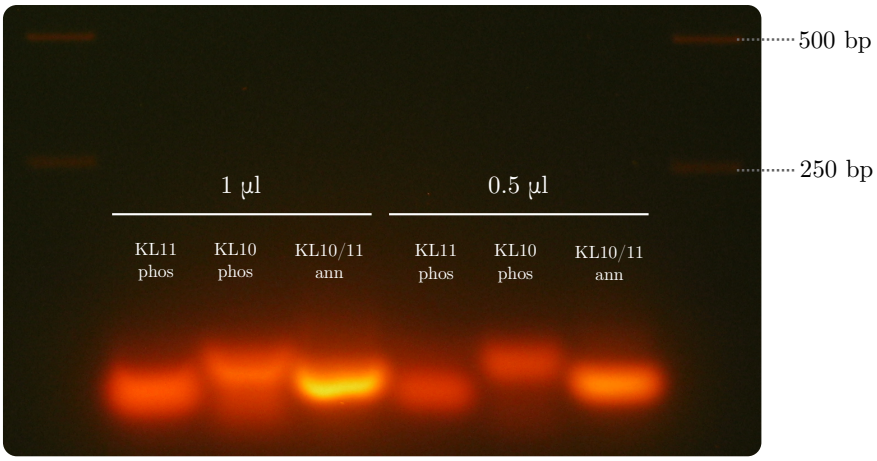


Figure D.1: Agarose gel of annealed complementary oligos. 1 µl and 0.5 µl of phosphorylated and annealed KL10/KL11 were run against unannealed phosphorylated controls of each individually.

pKL13 preparation post-stuffer removal

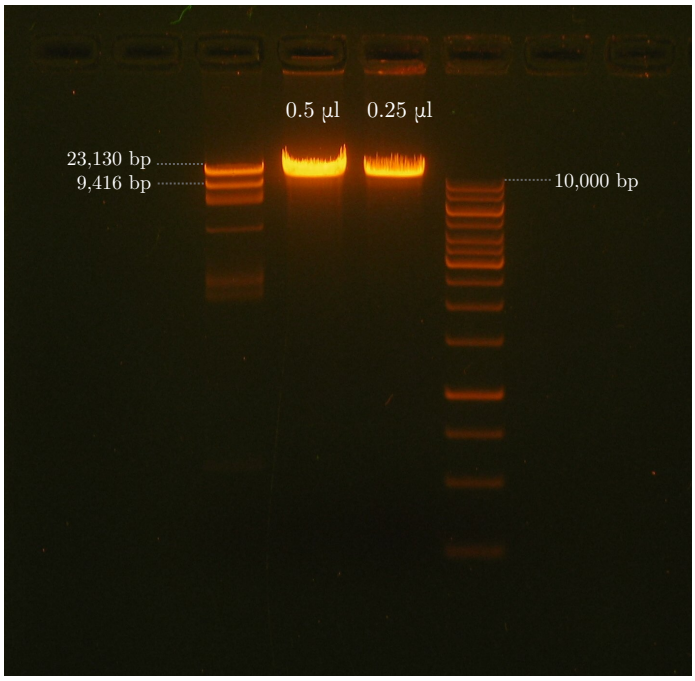


Figure D.2: Agarose gel of unligatable pKL13 vector prep after removal of stuffer. Agarose gel showed that the vector preparation was nuclease-free and highly concentrated, but ligation attempts with this vector were unsuccessful.

Phenotype of *B. theta* VPI-5482 wild-type versus Δ*tdk* on chondroitin sulfate

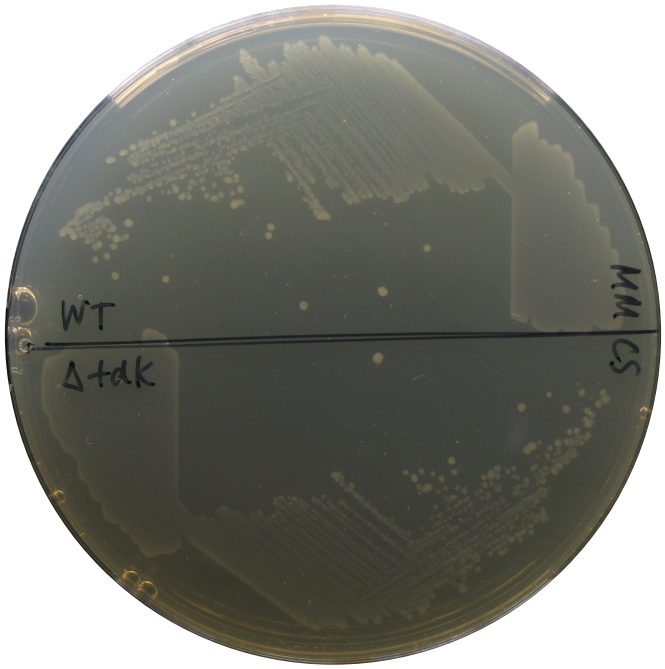


Figure D.3: Comparable phenotype of *B. theta* VPI-5482 wild-type versus Δ*tdk* on chondroitin sulfate as sole carbon source Note that Δ*tdk* is isogenic to the Δ*chuR* mutant used for functional screening. In the Charles lab collection, they have been designated *B. theta* BtUW24 and BtUW25, respectively; see Table 2.1.

Agarose gel of CLGM3 *chuR* complementing clones with versus without arabinose induction

The following gel images show the unexpected negative effect of using copy-number induction on insert stability. Fosmid DNA of *chuR* complementing CLGM3 clones was miniprepped and digested from cultures that were either copy-number induced (Figure D.4A) or cultures in which fosmid DNA was present in single copy (Figure D.4B). The insert was observed to be lost in the former case.

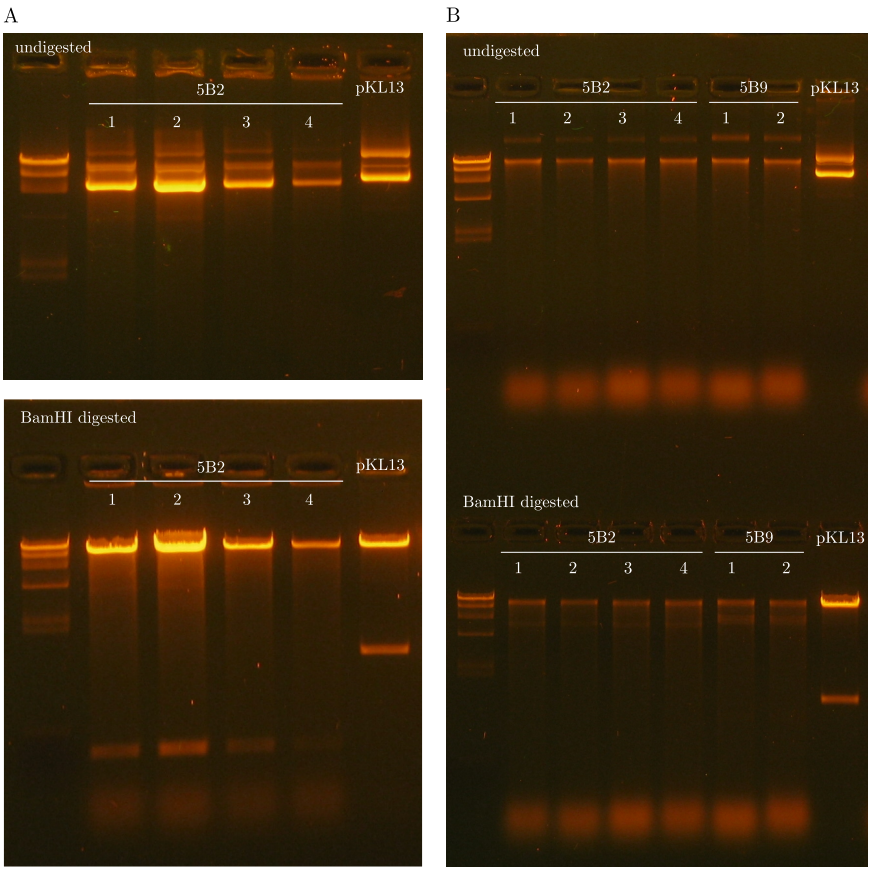


Figure D.4: Agarose gel of CLGM3 *chuR* complementing clones with versus without copy number induction. (A) CLGM3 clone 5B2 was miniprepped and digested from cultures grown in the presence of 0.2% arabinose. (B) CLGM3 clones 5B2 and 5B9 were miniprepped and digested from cultures grown in the absence of arabinose.

D.2 Sequence data

ermF-repA fragment (pAFD1)

>*ermF-repA* fragment (pAFD1)
ATAACAGCCGGTGACAGCCGGCTGACAGGGGGTTAAGGGGGCTTGTCCCCTTACACACGCACTCTTTAGGGTGCTAG
TGTGCTATCACCATACTGCATAGGTGCGAAGTTAGTGAATGTTTTGTAAATGCACAAATAAAGGGAAAAACATTTGG
ATTTGCGATAATAAAGTACTACCTTTGTTGCTGACCAAACGGTAGCTGACCGATACGGGAGAGTTACCAAAATACAA
GCCGCTGGAGTTAATTGACGGACATCCGACATCTCCAGCGGCTTTATTTTTGCCTATCTGCTTCGCCTAGGCACACC
AGTACCTCTACTAAAAATGTACTTCAAAGATACTTATTTTCTACCGACTTGATAGTTTTTACCCCATATTCTTGGAC
ATTTTTCCCCCATGAGGTTATCTTTGTAGGGTGAAAGAGAAACCCATAAACGGGGATAGATTGAATGCTGGGAAGCA
TAAACAATCGGGGTAAGGTTAGCGAACCTTGCCTTTCATCCCCCATTATAACTTTACATAGAGGAACTTTATCTATC
CCCCCCCGCCCCCAAAGGGGGAGCGACCAAACGGCAGCTTCACTCAATGGAGTGTTACTGTTCATCAAAGCCAAGTG
ATAATTGTCGTTTCTCTGCTTCTTCTTTCTTTTGGGCAGCTAAAGTCTTTTTCCGAACGTATGTTTTAGCAAATGTC
ACTCGGTCACCATTGAATACTATCAGAGGATTAATAAACCAAAGATTATCGGCTGGTCCTCGGGCTATGATTTCAGC
TTTTACAAGTTCTGCAAGTCCTTTATAAACGGCTTTGTCTGTTTTGTATTTGGTATATTCTAGGCATTTTTTTCTAT
TGAAAATGATTAAATCATTTTTGGGTTTCATGCAGGTCATAAAGTAACCAAAAACCCGAATAGCTGCTTGTGATAGG
TCAAAGAATGCAGCAAAGTTAGAAAGATACAATTTAGTGAATTGTTCTTCATCTACTTCTATTTGACGGATAAACGA
AGTCTTAAACACTTCTCCAGTTTCAGTGTCGGCTAAAGCTACTACAGCTCTCTTATCGCCACCACTATTACTCTTAT
ACTTTTTAACAACATGATTTTCAATACCTTCTATAGCTTGTTTCATAAAAGGATTTTCTTCGTTCTTTTGAAAATCG
GTTAACTTAACTGCTTTTTTATTTTCCATTTTGATATGTTTTTGGGAAATATTATTCTCCACAAAGTAAACTATTAT
TTTCCATAAAAACAATATTAAGGGAAATATTATTTTCCTATTTAGTATCATATTAGGAAATCGGTATTTTCTAGATT
GGAAAATGAGAATTTCCAATATGGAAAATGCCCTATATTGTGTATCAAGTACTTAACTTATTCTATTTCTTTTATTC
TTAATATACCCCCAAAACAGCACAAAATCAGTCACTTAAAAATCATCGGTCGGGGAATGGTGCACTCTCAGTACAAT
CTGCTCTGATGCCGCATAGTTAAGCCAGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGC
TCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCG
AAACGCGCGAGACGAAAGGGCCTCGTGATACGCCTATTTTTATAGGTTAATGTCATGATAATAATGGTTTCTTAAAT
TCTGATTAATAATTTGTTTAAATTTTTCGTTTGGCGTGAGGTATCCAAGTCTTTTACGAGGTCGATTATTGAGTTTA
TTTTCAATCCACTTAATCTGTTTGTTGGTTACTTCACTAAAGTCCTTACCCTTTGGGATATACTGCCTGATAAGCCC
GTTGGTGTTTTCATTGGCACCACGTTCCCATGAGTGGTATGGTTTGCAAAAATAGAATTTTATTTCCAATTTTTGCG
CAATTTCCTCGTGCTTTGCAAACTCCTTTCCATTGTCAGCCGTAATTGTGTGTATTAAGTTTTTCACTTTCCGCAGT
GCCCATACTGCAATCTTAGCTACCGGGATGGCTTCTTTTCCCGACAACTTGCGTATCCAGACCCTGCTTGTTGCTCT
GTCGTTAATGGTAAGAATGGCACCTTTGTGGTTCTTACCAATAATTGTATCTATCTCTAAATCACCAAATCTCTCCT
TCAGTTCCACTATCTCGGGACGCTCATCAATATCCACCCTGCCTGGGATAAATCCTCGCCCTGCATTTTTAGAACCA
CGTTTGGCATACCTGCGACCTTGTCTGCGAAGATATTTGTGCAGTTTGCCACCCCGCCGCTTATCCTCCCAAATCCA
GCGATATATCGTTTCGTGAGATACCATCGCAATTCCCTCCAAGCGGCTCCTGCCGACAATCTGCTCCGGGCTGAATC
CTTTCTTCAACAGCTTTATTATCCGTTTTCTCATTGCCGGTGTAAGCACTTCCTTGCGATGTTTTTGCTGCTTGCGC
CTGTCTGCTTTTCGCTGGGCAAGCTCCATGCTATAGCTACCACTTCGGGCGTCGCAATTGCGCTTTATCTCCCTGTA
AACAGTGCTTTTATCTACTCCGATAGCTTCCGCTATTGCTTTTTTGCTCATCGGTATTTGCAACATCATAGAAATTG
CATACCTTTGTTCCTCGGTTATATGTTTGCTCATCTGCAACTTTTTTTTCTTTGGACGGACAATTAAAGCAAAGATA
GCAAACTTTATCCATTCAGAGTGAGAGAAAGGGGGACATTGTCTCTCTTTCCTCTCTGAAAAATAAATGTTTTTATT
GCTTATTATCCGCACCCAAAAAGTTGCATTTATAAGTTGAACTCAAGAAGTATTCACCTGTAAGAAGTTACTAATGA
CAAAAAAGAAATTGCCCGTTCGTTTTACGGGTCAGCACTTTACTATTGATAAAGTGCTAATAAAAGATGCAATAAGA
CAAGCAAATATAAGTAATCAGGATACGGTTTTAGATATTGGGGCAGGCAAGGGGTTTCTTACTGTTCATTTATTAAA

AATCGCCAACAATGTTGTTGCTATTGAAAACGACACAGCTTTGGTTGAACATTTACGAAAATTATTTTCTGATGCCC
GAAATGTTCAAGTTGTCGGTTGTGATTTTAGGAATTTTGCAGTTCCGAAATTTCCTTTCAAAGTGGTGTCAAATATT
CCTTATGGCATTACTTCCGATATTTTCAAAATCCTGATGTTTGAGAGTCTTGGAAATTTTCTGGGAGGTTCCATTGT
CCTTCAATTAGAACCTACACAAAAGTTATTTTCGAGGAAGCTTTACAATCCATATACCGTTTTCTATCATACTTTTT
TTGATTTGAAACTTGTCTATGAGGTAGGTCCTGAAAGTTTCTTGCCACCGCCAACTGTCAAATCAGCCCTGTTAAAC
ATTAAAAGAAAACACTTATTTTTTGATTTTAAGTTTAAAGCCAAATACTTAGCATTTATTTCCTGTCTGTTAGAGAA
ACCTGATTTATCTGTAAAAACAGCTTTAAAGTCGATTTTCAGGAAAAGTCAGGTCAGGTCAATTTCGGAAAAATTCG
GTTTAAACCTTAATGCTCAAATTGTTTGTTTGTCTCCAAGTCAATGGTTAAACTGTTTTTTGGAAATGCTGGAAGTT
GTCCCTGAAAAATTTCATCCTTCGTAGTTCAAAGTCGGGTGGTTGTCAAGATGATTTTTTTGGTTTGGTGTCGTCTT
TTTTTAAGCTGCCGCATAACGGCTGGCAAATTGGCGATGGAGCCGACTTTTAGCACAAATGTTGAATAGAATTACTA
ATCTTCAACATTGCACAAAAGT

pKL13

Below is the theoretical sequence for pKL13.

>pKL13_expected
GCGGCCGCAAGGGGTTCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCTGGCTTAACTATGCGGCATCAGAGCAGATT
GTACTGAGAGTGCACCATATGCGGTGTGAAATACCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGCGCCATTC
GCCATTCAGCTGCGCAACTGTTGGGAAGGGCGATCGGTGCGGGCCTCTTCGCTATTACGCCAGCTGGCGAAAGGGGG
ATGTGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGTTGTAAAACGACGGCCAGTGAAT
TGTAATACGACTCACTATAGGGCGAATTCATAACAGCCGGTGACAGCCGGCTGACAGGGGGTTAAGGGGGCTTGTCC
CCTTACACACGCACTCTTTAGGGTGCTAGTGTGCTATCACCATACTGCATAGGTGCGAAGTTAGTGAATGTTTTGTA
AATGCACAAATAAAGGGAAAAACATTTGGATTTGCGATAATAAAGTACTACCTTTGTTGCTGACCAAACGGTAGCTG
ACCGATACGGGAGAGTTACCAAAATACAAGCCGCTGGAGTTAATTGACGGACATCCGACATCTCCAGCGGCTTTATT
TTTGCCTATCTGCTTCGCCTAGGCACACCAGTACCTCTACTAAAAATGTACTTCAAAGATACTTATTTTCTACCGAC
TTGATAGTTTTTACCCCATATTCTTGGACATTTTTCCCCCATGAGGTTATCTTTGTAGGGTGAAAGAGAAACCCATA
AACGGGGATAGATTGAATGCTGGGAAGCATAAACAATCGGGGTAAGGTTAGCGAACCTTGCCTTTCATCCCCCATTA
TAACTTTACATAGAGGAACTTTATCTATCCCCCCCCGCCCCCAAAGGGGGAGCGACCAAACGGCAGCTTCACTCAAT
GGAGTGTTACTGTTCATCAAAGCCAAGTGATAATTGTCGTTTCTCTGCTTCTTCTTTCTTTTGGGCAGCTAAAGTCT
TTTTCCGAACGTATGTTTTAGCAAATGTCACTCGGTCACCATTGAATACTATCAGAGGATTAATAAACCAAAGATTA
TCGGCTGGTCCTCGGGCTATGATTTCAGCTTTTACAAGTTCTGCAAGTCCTTTATAAACGGCTTTGTCTGTTTTGTA
TTTGGTATATTCTAGGCATTTTTTTCTATTGAAAATGATTAAATCATTTTTGGGTTTCATGCAGGTCATAAAGTAAC
CAAAAACCCGAATAGCTGCTTGTGATAGGTCAAAGAATGCAGCAAAGTTAGAAAGATACAATTTAGTGAATTGTTCT
TCATCTACTTCTATTTGACGGATAAACGAAGTCTTAAACACTTCTCCAGTTTCAGTGTCGGCTAAAGCTACTACAGC
TCTCTTATCGCCACCACTATTACTCTTATACTTTTTAACAACATGATTTTCAATACCTTCTATAGCTTGTTTCATAA
AAGGATTTTCTTCGTTCTTTTGAAAATCGGTTAACTTAACTGCTTTTTTATTTTCCATTTTGATATGTTTTTGGGAA
ATATTATTCTCCACAAAGTAAACTATTATTTTCCATAAAAACAATATTAAGGGAAATATTATTTTCCTATTTAGTAT
CATATTAGGAAATCGGTATTTTCTAGATTGGAAAATGAGAATTTCCAATATGGAAAATGCCCTATATTGTGTATCAA
GTACTTAACTTATTCTATTTCTTTTATTCTTAATATACCCCCAAAACAGCACAAAATCAGTCACTTAAAAATCATCG
GTCGGGGAATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGCCCCGACACCCGCCAACACC
CGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCA
TGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGACGAAAGGGCCTCGTGATACGCCTATTTTTATAGGTT
AATGTCATGATAATAATGGTTTCTTAAATTCTGATTAATAATTTGTTTAAATTTTTCGTTTGGCGTGAGGTATCCAA
GTCTTTTACGAGGTCGATTATTGAGTTTATTTTCAATCCACTTAATCTGTTTGTTGGTTACTTCACTAAAGTCCTTA
CCCTTTGGGATATACTGCCTGATAAGCCCGTTGGTGTTTTCATTGGCACCACGTTCCCATGAGTGGTATGGTTTGCA
AAAATAGAATTTTATTTCCAATTTTTGCGCAATTTCCTCGTGCTTTGCAAACTCCTTTCCATTGTCAGCCGTAATTG
TGTGTATTAAGTTTTTCACTTTCCGCAGTGCCCATACTGCAATCTTAGCTACCGGGATGGCTTCTTTTCCCGACAAC
TTGCGTATCCAGACCCTGCTTGTTGCTCTGTCGTTAATGGTAAGAATGGCACCTTTGTGGTTCTTACCAATAATTGT
ATCTATCTCTAAATCACCAAATCTCTCCTTCAGTTCCACTATCTCGGGACGCTCATCAATATCCACCCTGCCTGGGA
TAAATCCTCGCCCTGCATTTTTAGAACCACGTTTGGCATACCTGCGACCTTGTCTGCGAAGATATTTGTGCAGTTTG
CCACCCCGCCGCTTATCCTCCCAAATCCAGCGATATATCGTTTCGTGAGATACCATCGCAATTCCCTCCAAGCGGCT
CCTGCCGACAATCTGCTCCGGGCTGAATCCTTTCTTCAACAGCTTTATTATCCGTTTTCTCATTGCCGGTGTAAGCA
CTTCCTTGCGATGTTTTTGCTGCTTGCGCCTGTCTGCTTTTCGCTGGGCAAGCTCCATGCTATAGCTACCACTTCGG
GCGTCGCAATTGCGCTTTATCTCCCTGTAAACAGTGCTTTTATCTACTCCGATAGCTTCCGCTATTGCTTTTTTGCT
CATCGGTATTTGCAACATCATAGAAATTGCATACCTTTGTTCCTCGGTTATATGTTTGCTCATCTGCAACTTTTTTT
TCTTTGGACGGACAATTAAAGCAAAGATAGCAAACTTTATCCATTCAGAGTGAGAGAAAGGGGGACATTGTCTCTCT

TTCCTCTCTGAAAAATAAATGTTTTTATTGCTTATTATCCGCACCCAAAAAGTTGCATTTATAAGTTGAACTCAAGA
AGTATTCACCTGTAAGAAGTTACTAATGACAAAAAAGAAATTGCCCGTTCGTTTTACGGGTCAGCACTTTACTATTG
ATAAAGTGCTAATAAAAGATGCAATAAGACAAGCAAATATAAGTAATCAGGATACGGTTTTAGATATTGGGGCAGGC
AAGGGGTTTCTTACTGTTCATTTATTAAAAATCGCCAACAATGTTGTTGCTATTGAAAACGACACAGCTTTGGTTGA
ACATTTACGAAAATTATTTTCTGATGCCCGAAATGTTCAAGTTGTCGGTTGTGATTTTAGGAATTTTGCAGTTCCGA
AATTTCCTTTCAAAGTGGTGTCAAATATTCCTTATGGCATTACTTCCGATATTTTCAAAATCCTGATGTTTGAGAGT
CTTGGAAATTTTCTGGGAGGTTCCATTGTCCTTCAATTAGAACCTACACAAAAGTTATTTTCGAGGAAGCTTTACAA
TCCATATACCGTTTTCTATCATACTTTTTTTGATTTGAAACTTGTCTATGAGGTAGGTCCTGAAAGTTTCTTGCCAC
CGCCAACTGTCAAATCAGCCCTGTTAAACATTAAAAGAAAACACTTATTTTTTGATTTTAAGTTTAAAGCCAAATAC
TTAGCATTTATTTCCTGTCTGTTAGAGAAACCTGATTTATCTGTAAAAACAGCTTTAAAGTCGATTTTCAGGAAAAG
TCAGGTCAGGTCAATTTCGGAAAAATTCGGTTTAAACCTTAATGCTCAAATTGTTTGTTTGTCTCCAAGTCAATGGT
TAAACTGTTTTTTGGAAATGCTGGAAGTTGTCCCTGAAAAATTTCATCCTTCGTAGTTCAAAGTCGGGTGGTTGTCA
AGATGATTTTTTTGGTTTGGTGTCGTCTTTTTTTAAGCTGCCGCATAACGGCTGGCAAATTGGCGATGGAGCCGACT
TTTAGCACAAATGTTGAATAGAATTACTAATCTTCAACATTGCACAAAAGTGAATTCGAGCTCGGTACCCGGGGATC
CCACAAATGGCGCGCCGGCTGGATTTAATTAATGTCTGCTCCTCGGTTATGTTTTTAAGGTCAAAAAAAACCCCCGG
ACCTTTCGGTGCGGGGGTCTTAGTTCGTTAAGGCTTGATCTCTAGCGATTAAGTTGGGTAACGCCAGGGTTTTCGTC
ACTTAGTCAGCTAGCCACGTGCCTTAGGGTGTGAAATTGTTATCCGCTCACAATTCCACACATTATACGAGCCGATG
ATTAATTGTCAACAGCTCCCTGAGGTTCGAAGATCCTCCGGCTCACGGTAACTGATGCCGTATTTGCAGTACCAGCG
TACGGCCCACAGAATGATGTCACGCTGAAAATGCCGGCCTTTGAATGGGTTCATGTGCAGCTCCATCAGCAAAAGGG
GATGATAAGTTTATCACCACCGACTATTTGCAACAGTGCCGTTGATCGTGCTATGATCGACTGATGTCATCAGCGGT
GGAGTGCAATGTCGTGCAATACGAATGGCGAAAAGCCGAGCTCATCGGTCAGCTTCTCAACCTTGGGGTTACCCCCG
GCGGTGTGCTGCTGGTCCACAGCTCCTTCCGTAGCGTCCGGCCCCTCGAAGATGGGCCACTTGGACTGATCGAGGCC
CTGCGTGCTGCGCTGGGTCCGGGAGGGACGCTCGTCATGCCCTCGTGGTCAGGTCTGGACGACGAGCCGTTCGATCC
TGCCACGTCGCCCGTTACACCGGACCTTGGAGTTGTCTCTGACACATTCTGGCGCCTGCCAAATGTAAAGCGCAGCG
CCCATCCATTTGCCTTTGCGGCAGCGGGGCCACAGGCAGAGCAGATCATCTCTGATCCATTGCCCCTGCCACCTCAC
TCGCCTGCAAGCCCGGTCGCCCGTGTCCATGAACTCGATGGGCAGGTACTTCTCCTCGGCGTGGGACACGATGCCAA
CACGACGCTGCATCTTGCCGAGTTGATGGCAAAGGTTCCCTATGGGGTGCCGAGACACTGCACCATTCTTCAGGATG
GCAAGTTGGTACGCGTCGATTATCTCGAGAATGACCACTGCTGTGAGCGCTTTGCCTTGGCGGACAGGTGGCTCAAG
GAGAAGAGCCTTCAGAAGGAAGGTCCAGTCGGTCATGCCTTTGCTCGGTTGATCCGCTCCCGCGACATTGTGGCGAC
AGCCCTGGGTCAACTGGGCCGAGATCCGTTGATCTTCCTGCATCCGCCAGAGGCGGGATGCGAAGAATGCGATGCCG
CTCGCCAGTCGATTGGCTGAGCTCATGAGCGGAGAACGAGATGACGTTGGAGGGGCAAGGTCGCGCTGATTGCTGGG
GCAACACGTTCGAACACGTGATGCATTAACTAGGTGACGTCATAGCTGTTTCCTGTGTGAAATTGTTATCGGTCAGT
TTCACCTGATTTACGTAAAAACCCGCTTCGGCGGGTTTTTGCTTTTGGAGGGGCAGAAAGATGAATGACTGTCCGGT
CCGAGCAGGTCGCGATCGCATTTGTGGGATCCTCTAGAGTCGACCTGCAGGCATGCAAGCTTGCGCTTTTCCGCTGC
ATAACCCTGCTTCGGGGTCATTATAGCGATTTTTTCGGTATATCCATCCTTTTTCGCACGATATACAGGATTTTGCC
AAAGGGTTCGTGTAGACTTTCCTTGGTGTATCCAACGGCGTCAGCCGGGCAGGATAGGTGAAGTAGGCCCACCCGCG
AGCGGGTGTTCCTTCTTCACTGTCCCTTATTCGCACCTGGCGGTGCTCAACGGGAATCCTGCTCTGCGAGGCTGGCC
GGCTACCGCCGGCGTAACAGATGAGGGCAAGCGGATGGCTGATGAAACCAAGCCAACCAGGAAGGGCAGCCCACCTA
TCAAGGTGTACTGCCTTCCAGACGAACGAAGAGCGATTGAGGAAAAGGCGGCGGCGGCCGGCATGAGCCTGTCGGCC
TACCTGCTGGCCGTCGGCCAGGGCTACAAAATCACGGGCGTCGTGGACTATGAGCACGTCCGCGAGCTGGCCCGCAT
CAATGGCGACCTGGGCCGCCTGGGCGGCCTGCTGAAACTCTGGCTCACCGACGACCCGCGCACGGCGCGGTTCGGTG
ATGCCACGATCCTCGCCCTGCTGGCGAAGATCGAAGAGAAGCAGGACGAGCTTGGCAAGGTCATGATGGGCGTGGTC
CGCCCGAGGGCAGAGCCATGACTTTTTTAGCCGCTAAAACGGCCGGGGGGTGCGCGTGATTGCCAAGCACGTCCCCA
TGCGCTCCATCAAGAAGAGCGACTTCGCGGAGCTGGTGAAGTACATCACCGACGAGCAAGGCAAGACCGAAAGCTTG
AGTATTCTATAGTCTCACCTAAATAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCT
CACAATTCCACACAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACAT

TAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGC
GAACCCCTTGCGGCCGCCCGGGCCGTCGACCAATTCTCATGTTTGACAGCTTATCATCGAATTTCTGCCATTCATCC
GCTTATTATCACTTATTCAGGCGTAGCAACCAGGCGTTTAAGGGCACCAATAACTGCCTTAAAAAAATTACGCCCCG
CCCTGCCACTCATCGCAGTACTGTTGTAATTCATTAAGCATTCTGCCGACATGGAAGCCATCACAAACGGCATGATG
AACCTGAATCGCCAGCGGCATCAGCACCTTGTCGCCTTGCGTATAATATTTGCCCATGGTGAAAACGGGGGCGAAGA
AGTTGTCCATATTGGCCACGTTTAAATCAAAACTGGTGAAACTCACCCAGGGATTGGCTGAGACGAAAAACATATTC
TCAATAAACCCTTTAGGGAAATAGGCCAGGTTTTCACCGTAACACGCCACATCTTGCGAATATATGTGTAGAAACTG
CCGGAAATCGTCGTGGTATTCACTCCAGAGCGATGAAAACGTTTCAGTTTGCTCATGGAAAACGGTGTAACAAGGGT
GAACACTATCCCATATCACCAGCTCACCGTCTTTCATTGCCATACGAAATTCCGGATGAGCATTCATCAGGCGGGCA
AGAATGTGAATAAAGGCCGGATAAAACTTGTGCTTATTTTTCTTTACGGTCTTTAAAAAGGCCGTAATATCCAGCTG
AACGGTCTGGTTATAGGTACATTGAGCAACTGACTGAAATGCCTCAAAATGTTCTTTACGATGCCATTGGGATATAT
CAACGGTGGTATATCCAGTGATTTTTTTCTCCATTTTAGCTTCCTTAGCTCCTGAAAATCTCGATAACTCAAAAAAT
ACGCCCGGTAGTGATCTTATTTCATTATGGTGAAAGTTGGAACCTCTTACGTGCCGATCAACGTCTCATTTTCGCCA
AAAGTTGGCCCAGGGCTTCCCGGTATCAACAGGGACACCAGGATTTATTTATTCTGCGAAGTGATCTTCCGTCACAG
GTATTTATTCGCGATAAGCTCATGGAGCGGCGTAACCGTCGCACAGGAAGGACAGAGAAAGCGCGGATCTGGGAAGT
GACGGACAGAACGGTCAGGACCTGGATTGGGGAGGCGGTTGCCGCCGCTGCTGCTGACGGTGTGACGTTCTCTGTTC
CGGTCACACCACATACGTTCCGCCATTCCTATGCGATGCACATGCTGTATGCCGGTATACCGCTGAAAGTTCTGCAA
AGCCTGATGGGACATAAGTCCATCAGTTCAACGGAAGTCTACACGAAGGTTTTTGCGCTGGATGTGGCTGCCCGGCA
CCGGGTGCAGTTTGCGATGCCGGAGTCTGATGCGGTTGCGATGCTGAAACAATTATCCTGAGAATAAATGCCTTGGC
CTTTATATGGAAATGTGGAACTGAGTGGATATGCTGTTTTTGTCTGTTAAACAGAGAAGCTGGCTGTTATCCACTGA
GAAGCGAACGAAACAGTCGGGAAAATCTCCCATTATCGTAGAGATCCGCATTATTAATCTCAGGAGCCTGTGTAGCG
TTTATAGGAAGTAGTGTTCTGTCATGATGCCTGCAAGCGGTAACGAAAACGATTTGAATATGCCTTCAGGAACAATA
GAAATCTTCGTGCGGTGTTACGTTGAAGTGGAGCGGATTATGTCAGCAATGGACAGAACAACCTAATGAACACAGAA
CCATGATGTGGTCTGTCCTTTTACAGCCAGTAGTGCTCGCCGCAGTCGAGCGACAGGGCGAAGCCCTCGGCTGGTTG
CCCTCGCCGCTGGGCTGGCGGCCGTCTATGGCCCTGCAAACGCGCCAGAAACGCCGTCGAAGCCGTGTGCGAGACAC
CGCGGCCGGCCGCCGGCGTTGTGGATACCTCGCGGAAAACTTGGCCCTCACTGACAGATGAGGGGCGGACGTTGACA
CTTGAGGGGCCGACTCACCCGGCGCGGCGTTGACAGATGAGGGGCAGGCTCGATTTCGGCCGGCGACGTGGAGCTGG
CCAGCCTCGCAAATCGGCGAAAACGCCTGATTTTACGCGAGTTTCCCACAGATGATGTGGACAAGCCTGGGGATAAG
TGCCCTGCGGTATTGACACTTGAGGGGCGCGACTACTGACAGATGAGGGGCGCGATCCTTGACACTTGAGGGGCAGA
GTGCTGACAGATGAGGGGCGCACCTATTGACATTTGAGGGGCTGTCCACAGGCAGAAAATCCAGCATTTGCAAGGGT
TTCCGCCCGTTTTTCGGCCACCGCTAACCTGTCTTTTAACCTGCTTTTAAACCAATATTTATAAACCTTGTTTTTAA
CCAGGGCTGCGCCCTGTGCGCGTGACCGCGCACGCCGAAGGGGGGTGCCCCCCCTTCTCGAACCCTCCCGGTCGAGT
GAGCGAGGAAGCACCAGGGAACAGCACTTATATATTCTGCTTACACACGATGCCTGAAAAAACTTCCCTTGGGGTTA
TCCACTTATCCACGGGGATATTTTTATAATTATTTTTTTTATAGTTTTTAGATCTTCTTTTTTAGAGCGCCTTGTAG
GCCTTTATCCATGCTGGTTCTAGAGAAGGTGTTGTGACAAATTGCCCTTTCAGTGTGACAAATCACCCTCAAATGAC
AGTCCTGTCTGTGACAAATTGCCCTTAACCCTGTGACAAATTGCCCTCAGAAGAAGCTGTTTTTTCACAAAGTTATC
CCTGCTTATTGACTCTTTTTTATTTAGTGTGACAATCTAAAAACTTGTCACACTTCACATGGATCTGTCATGGCGGA
AACAGCGGTTATCAATCACAAGAAACGTAAAAATAGCCCGCGAATCGTCCAGTCAAACGACCTCACTGAGGCGGCAT
ATAGTCTCTCCCGGGATCAAAAACGTATGCTGTATCTGTTCGTTGACCAGATCAGAAAATCTGATGGCACCCTACAG
GAACATGACGGTATCTGCGAGATCCATGTTGCTAAATATGCTGAAATATTCGGATTGACCTCTGCGGAAGCCAGTAA
GGATATACGGCAGGCATTGAAGAGTTTCGCGGGGAAGGAAGTGGTTTTTTATCGCCCTGAAGAGGATGCCGGCGATG
AAAAAGGCTATGAATCTTTTCCTTGGTTTATCAAACGTGCGCACAGTCCATCCAGAGGGCTTTACAGTGTACATATC
AACCCATATCTCATTCCCTTCTTTATCGGGTTACAGAACCGGTTTACGCAGTTTCGGCTTAGTGAAACAAAAGAAAT
CACCAATCCGTATGCCATGCGTTTATACGAATCCCTGTGTCAGTATCGTAAGCCGGATGGCTCAGGCATCGTCTCTC
TGAAAATCGACTGGATCATAGAGCGTTACCAGCTGCCTCAAAGTTACCAGCGTATGCCTGACTTCCGCCGCCGCTTC
CTGCAGGTCTGTGTTAATGAGATCAACAGCAGAACTCCAATGCGCCTCTCATACATTGAGAAAAAGAAAGGCCGCCA

GACGACTCATATCGTATTTTCCTTCCGCGATATCACTTCCATGACGACAGGATAGTCTGAGGGTTATCTGTCACAGA
TTTGAGGGTGGTTCGTCACATTTGTTCTGACCTACTGAGGGTAATTTGTCACAGTTTTGCTGTTTCCTTCAGCCTGC
ATGGATTTTCTCATACTTTTTGAACTGTAATTTTTAAGGAAGCCAAATTTGAGGGCAGTTTGTCACAGTTGATTTCC
TTCTCTTTCCCTTCGTCATGTGACCTGATATCGGGGGTTAGTTCGTCATCATTGATGAGGGTTGATTATCACAGTTT
ATTACTCTGAATTGGCTATCCGCGTGTGTACCTCTACCTGGAGTTTTTCCCACGGTGGATATTTCTTCTTGCGCTGA
GCGTAAGAGCTATCTGACAGAACAGTTCTTCTTTGCTTCCTCGCCAGTTCGCTCGCTATGCTCGGTTACACGGCTGC
GGCGAGCGCTAGTGATAATAAGTGACTGAGGTATGTGCTCTTCTTATCTCCTTTTGTAGTGTTGCTCTTATTTTAAA
CAACTTTGCGGTTTTTTGATGACTTTGCGATTTTGTTGTTGCTTTGCAGTAAATTGCAAGATTTAATAAAAAAACGC
AAAGCAATGATTAAAGGATGTTCAGAATGAAACTCATGGAAACACTTAACCAGTGCATAAACGCTGGTCATGAAATG
ACGAAGGCTATCGCCATTGCACAGTTTAATGATGACAGCCCGGAAGCGAGGAAAATAACCCGGCGCTGGAGAATAGG
TGAAGCAGCGGATTTAGTTGGGGTTTCTTCTCAGGCTATCAGAGATGCCGAGAAAGCAGGGCGACTACCGCACCCGG
ATATGGAAATTCGAGGACGGGTTGAGCAACGTGTTGGTTATACAATTGAACAAATTAATCATATGCGTGATGTGTTT
GGTACGCGATTGCGACGTGCTGAAGACGTATTTCCACCGGTGATCGGGGTTGCTGCCCATAAAGGTGGCGTTTACAA
AACCTCAGTTTCTGTTCATCTTGCTCAGGATCTGGCTCTGAAGGGGCTACGTGTTTTGCTCGTGGAAGGTAACGACC
CCCAGGGAACAGCCTCAATGTATCACGGATGGGTACCAGATCTTCATATTCATGCAGAAGACACTCTCCTGCCTTTC
TATCTTGGGGAAAAGGACGATGTCACTTATGCAATAAAGCCCACTTGCTGGCCGGGGCTTGACATTATTCCTTCCTG
TCTGGCTCTGCACCGTATTGAAACTGAGTTAATGGGCAAATTTGATGAAGGTAAACTGCCCACCGATCCACACCTGA
TGCTCCGACTGGCCATTGAAACTGTTGCTCATGACTATGATGTCATAGTTATTGACAGCGCGCCTAACCTGGGTATC
GGCACGATTAATGTCGTATGTGCTGCTGATGTGCTGATTGTTCCCACGCCTGCTGAGTTGTTTGACTACACCTCCGC
ACTGCAGTTTTTCGATATGCTTCGTGATCTGCTCAAGAACGTTGATCTTAAAGGGTTCGAGCCTGATGTACGTATTT
TGCTTACCAAATACAGCAATAGTAATGGCTCTCAGTCCCCGTGGATGGAGGAGCAAATTCGGGATGCCTGGGGAAGC
ATGGTTCTAAAAAATGTTGTACGTGAAACGGATGAAGTTGGTAAAGGTCAGATCCGGATGAGAACTGTTTTTGAACA
GGCCATTGATCAACGCTCTTCAACTGGTGCCTGGAGAAATGCTCTTTCTATTTGGGAACCTGTCTGCAATGAAATTT
TCGATCGTCTGATTAAACCACGCTGGGAGATTAGATAATGAAGCGTGCGCCTGTTATTCCAAAACATACGCTCAATA
CTCAACCGGTTGAAGATACTTCGTTATCGACACCAGCTGCCCCGATGGTGGATTCGTTAATTGCGCGCGTAGGAGTA
ATGGCTCGCGGTAATGCCATTACTTTGCCTGTATGTGGTCGGGATGTGAAGTTTACTCTTGAAGTGCTCCGGGGTGA
TAGTGTTGAGAAGACCTCTCGGGTATGGTCAGGTAATGAACGTGACCAGGAGCTGCTTACTGAGGACGCACTGGATG
ATCTCATCCCTTCTTTTCTACTGACTGGTCAACAGACACCGGCGTTCGGTCGAAGAGTATCTGGTGTCATAGAAATT
GCCGATGGGAGTCGCCGTCGTAAAGCTGCTGCACTTACCGAAAGTGATTATCGTGTTCTGGTTGGCGAGCTGGATGA
TGAGCAGATGGCTGCATTATCCAGATTGGGTAACGATTATCGCCCAACAAGTGCTTATGAACGTGGTCAGCGTTATG
CAAGCCGATTGCAGAATGAATTTGCTGGAAATATTTCTGCGCTGGCTGATGCGGAAAATATTTCACGTAAGATTATT
ACCCGCTGTATCAACACCGCCAAATTGCCTAAATCAGTTGTTGCTCTTTTTTCTCACCCCGGTGAACTATCTGCCCG
GTCAGGTGATGCACTTCAAAAAGCCTTTACAGATAAAGAGGAATTACTTAAGCAGCAGGCATCTAACCTTCATGAGC
AGAAAAAAGCTGGGGTGATATTTGAAGCTGAAGAAGTTATCACTCTTTTAACTTCTGTGCTTAAAACGTCATCTGCA
TCAAGAACTAGTTTAAGCTCACGACATCAGTTTGCTCCTGGAGCGACAGTATTGTATAAGGGCGATAAAATGGTGCT
TAACCTGGACAGGTCTCGTGTTCCAACTGAGTGTATAGAGAAAATTGAGGCCATTCTTAAGGAACTTGAAAAGCCAG
CACCCTGATGCGACCACGTTTTAGTCTACGTTTATCTGTCTTTACTTAATGTCCTTTGTTACAGGCCAGAAAGCATA
ACTGGCCTGAATATTCTCTCTGGGCCCACTGTTCCACTTGTATCGTCGGTCTGATAATCAGACTGGGACCACGGTCC
CACTCGTATCGTCGGTCTGATTATTAGTCTGGGACCACGGTCCCACTCGTATCGTCGGTCTGATTATTAGTCTGGGA
CCACGGTCCCACTCGTATCGTCGGTCTGATAATCAGACTGGGACCACGGTCCCACTCGTATCGTCGGTCTGATTATT
AGTCTGGGACCATGGTCCCACTCGTATCGTCGGTCTGATTATTAGTCTGGGACCACGGTCCCACTCGTATCGTCGGT
CTGATTATTAGTCTGGAACCACGGTCCCACTCGTATCGTCGGTCTGATTATTAGTCTGGGACCACGGTCCCACTCGT
ATCGTCGGTCTGATTATTAGTCTGGGACCACGATCCCACTCGTGTTGTCGGTCTGATTATCGGTCTGGGACCACGGT
CCCACTTGTATTGTCGATCAGACTATCAGCGTGAGACTACGATTCCATCAATGCCTGTCAAGGGCAAGTATTGACAT
GTCGTCGTAACCTGTAGAACGGAGTAACCTCGGTGTGCGGTTGTATGCCTGCTGTGGATTGCTGCTGTGTCCTGCTT
ATCCACAACATTTTGCGCACGGTTATGTGGACAAAATACCTGGTTACCCAGGCCGTGCCGGCACGTTAACCGGGCTG

CATCCGATGCAAGTGTGTCGCTGTCGACGAGCTCGCGAGCTCGGACATGAGGTTGCCCCGTATTCAGTGTCGCTGAT
TTGTATTGTCTGAAGTTGTTTTTACGTTAAGTTGATGCAGATCAATTAATACGATACCTGCGTCATAATTGATTATT
TGACGTGGTTTGATGGCCTCCACGCACGTTGTGATATGTAGATGATAATCATTATCACTTTACGGGTCCTTTCCGGT
GATCCGACAGGTTACGGGGCGGCGACCTCGCGGGTTTTCGCTATTTATGAAAATTTTCCGGTTTAAGGCGTTTCCGT
TCTTCTTCGTCATAACTTAATGTTTTTATTTAAAATACCCTCTGAAAAGAAAGGAAACGACAGGTGCTGAAAGCGAG
CTTTTTGGCCTCTGTCGTTTCCTTTCTCTGTTTTTGTCCGTGGAATGAACAATGGAAGTCCGAGCTCATCGCTAATA
ACTTCGTATAGCATACATTATACGAAGTTATATTCGAT

Sanger sequencing reads of CLGM3 *chuR* clones #5

Primer KL61

>CLGM3 chuR clone 5 primer KL61
AAGCTGCWTGTYMTGGTAAGCCCGTGGGAGCCGTATGTAATCTCGCATGCGAATACTGCTATTATTTGGAAAAGGCG
AACCTATACAAAGAAAACCCCAAACATGTAATGAGCGATGAACTACTGGAAAAGTTTATCGACGAGTATATCAGTTC
TCAAACCATGCCTCAAGTGCTTTTTACCTGGCACGGTGGAGAAACGCTGATGCGTCCGCTTTCTTTTTATAAAAAGG
CGATGGAACTGCAAAAGAAATACGCCCGCGGACGTACGATTGACAATTGTATCCAGACGAATGGGACCTTACTCACA
GACGAATGGTGCGAGTTCTTCCGTGAAAACAACTGGCTGGTAGGGGTTTCTATTGATGGCCCGCAAGAGTTTCATGA
CGAATACCGCAAGAACAAAATGGGCAAACCTTCTTTCGTCAAAGTGATGCAAGGGATTAATCTCCTGAAAAAACATG
GAGTAGAATGGAACGCTATGGCTGTTGTGAACGATTTCAATGCCGAATATCCATTAGACTTTTATAATTTCTTCAAA
GAAATAGATTGCCATTATATCCAGTTCGCCCCGATTGTTGAACGCATTGTTTCACATCAGGACGGTCGTCATCTTGC
CTCTCTGGCAGAAGGTAAAGAAGGAGCATTGGCTGATTTCTCCATAAGTCCGGAACAATGGGGTAACTTTCTCTGTA
CAATTTTTGATGAATGGGTAAAAGAAGATGTGGGCAAATTCTTCATACAGATATTCGATTCTACATTGGCTAACTGG
ATGGGTGAGCAACCGGGCGTATGTACAATGGCGAAGCATTGCGGACATGCCGGCGTTATGGAATTCAACGGAGACGT
ATACTCTTGTGACCACTTCGTATTCCCGGAATATAAATTGGGAAATATCTATAGCCAGACTTTGGTGGAAATGATGC
ATAGTGAACGACAGCAAACTTCGGGACAATGAAATACCAATCACTCCCAACACAATGCAAGGAGTGCGACTTTCTAT
TTGCCTGCAACGGARATGTCCAAAGAACCGCTTCAGTCGGACAGCGGACGGCGAACCCGGTCTGACTATTTGTGCAA
AGGATATTACCAATACTTTCASMWGTAGCYTCCTATWWTGGATTYMTGAAAAARRATTAATGAATCAMCA

Primer KL62

>CLGM3 chuR clone 5 primer KL62
GTAATGATGTTYGGCAGGAGCCTGTTGATTCATTAATTCTTTTTTCATGAAATCCATATAGGGAGCTACATGCTGAA
AGTATYGGTAATATCCTTTGCACAAATAGTTCAGACCGGGTTCGCCGTCCGCTGTCCGACTGAAGCGGTTCTTTGGA
CATTCTCCGTTGCAGGCAAATAGAAAGTCGCACTCCTTGCATTGTGTTGGGAGTGATTGGTATTTCATTGTCCCGAA
GTTTTGCTGTCGTTCACTATGCATCATTTCCACCAAAGTCTGGCTATAGATATTTCCCAATTTATATTCCGGGAATA
CGAAGTGGTCACAAGAGTATACGTCTCCGTTGAATTCCATAACGCCGGCATGTCCGCAATGCTTCGCCATTGTACAT
ACGCCCGGTTGCTCACCCATCCAGTTAGCCAATGTAGAATCGAATATCTGTATGAAGAATTTGCCCACATCTTCTTT
TACCCATTCATCAAAAATTGTACAGAGAAAGTTACCCCATTGTTCCGGACTTATGGAGAAATCAGCCAATGCTCCTT
CTTTACCTTCTGCCAGAGAGGCAAGATGACGACCGTCCTGATGTGAAACAATGCGTTCAACAATCGGGGCGAACTGG
ATATAATGGCAATCTATTTCTTTGAAGAAATTATAAAAGTCTAATGGATATTCGGCATTGAAATCGTTCACAACAGC
CATAGCGTTCCATTCTACTCCATGTTTTTTCAGGAGATTAATCCCTTGCATCACTTTGACGAAAGAAGGTTTGCCCA
TTTTGTTCTTGCGGTATTCGTCATGAAACTCTTGCGGGCCATCAATAGAAACCCCTACCAGCCAGTTGTTTTCACGG
AAGAACTCGCACCATTCGTCTGTGAGTAAGTCCCATTCGTCTGGATACAATTGTCAATCGTACGTCCGCGGGCGTAT
TTCTTTTGCAGTTCCATCGCGTTTTATAAAAGAAAGCGGACGCATCAGCGTTTCTCACGTGCCAGTAAAAGCACTTG
AGCATGGTTTGAGACTGATATACTCGTCGATAACTTTTCCAGTAGTCATCGCTCATACATGTTTGGGGTTTTCTTTT
GTATAGTCGCTTTCAAWAATAGCAGTATCSCATGCGAGAATACATACGCTCACGGGCTTACATGACTAGCRGTTKGC
TAAGTGAG

Primer KL66

>CLGM3 chuR clone 5 primer KL66
CCAATCTTYCAGGAGCGGACGCATCAGCGTTTCTCCACCGTGCCAGGTAAAAAGCACTTGAGGCATGGTTTGAGAAC
TGATATACTCGTCGATAAACTTTTCCAGTAGTTCATCGCTCATTACATGTTTGGGGTTTTCTTTGTATAGGTTCGCC
TTTTCCAAATAATAGCAGTATTCGCATGCGAGATTACATACGGCTCCCACGGGCTTTACCATGACATAAAGCGGTTT
GGCAAAAGGTGCATAAGTTGGTGCTTTCATCATACTGATGCGCCTGCGTGAGCGAGGTTTCCGGCGAGAGGGGGGTA
AACAGTTCASCCKGYGCCTGCTCCGGCTTCARCCSCAGAGGAGGSSAGCAGAAGAAGARAGGACGGGRGGAGGAGTC
AGAAKCTTATGTTGTTTATTCGWGGGAAGGCCATGTCGGGKGCGCCGATCATKASTGGGATMAGCWASTTTCCSAAG
CCRCCRATTATRATAGGKATRACTATAAAGAAAATTMTAACGAARGCATGGKCGGTAACRATAACATTGTAAATCTG
GACATCACCCATTASGGSCCYGGGTTGGWTTAGCTCGGMTYRRGMAGKGMSGCTTAWKGCCGYGCCRCCTATYCCRG
CTCRGGCACCWAACACTATATRRGGGTGCCKATATCTTRRRWW

Primer KL67

>CLGM3 chuR clone 5 primer KL67
ATGCGTCCGTAAGCACTTGAGGCATGGTTTGAGAACTGATATACTCGTCGATAAACTTTTCCAGTAGTTCATCGCTC
ATTACATGTTTGGGGTTTTCTTTGTATAGGTTCGCCTTTTCCAAATAATAGCAGTATTCGCATGCGAGATTACATAC
GGCTCCCACGGGCTTTACCATGACATAAAGCGGTTTGGCAAAAGGTGCATAAGTTGTTGCTTTCATSATGGGCCTTC
CCCCCCCGGSKGGGGGGCGCTYCYGGSYCCCCCCCCCTMCCTSTKCCTGTCCGCKRGCSYCCKCGGGGGARGMWSSK
SYAAAAKGMWYYMGCTGRCCTCSGWTCCKCCCTCACACCKGARASRKSGWCAKYAKAGGRSCRTMASWWAACYTAMS
AGTTKCYWCCWCCCCYCTTGGWTGGGGGCSCGYCSYTRSTCGGSSYCRCTARTWKTKWWMGAACAARWTSRAACWCG
AAACKCKWAGCTTGWTACTGTTATCMAAACTWCTAARGAGTSWGAWCCGCCMGRMKAAKAYTKGTTCTCCTTCTTCY
CCCCCYCCCMSCKTTGGCAGRTCTWAGTWCTACSCMRWWAWAGMTCRASGACSRRWTTARGMMGTKCYTSTARTGCG
GAYCASGAGYYAACGMWRTRWRMKGTTAKSTRCTGTTTGGAARTTWGCAAAA

Primer KL68

>CLGM3 chuR clone 5 primer KL68
CCCACGCATGAGTGCGACTTTCTATTTGCCTGCAACGGAGAATGTCCAAAGAACCGCTTCAGTCGGACAGCGGACGG
CGAACCCGGTCTGAACTATTTGTGCAAAGGATATTACCAATACTTTCAGCATGTAGCTCCCTATATGGATTTCATGA
AAAAAGAATTAATGAATCAACAGGCTCCTGCCAACATCATGAAAGCACTAAAAGACGGAAGTTTAAAAATAGAATAT
TAAASGCMGSCCGGMGGGACCAARRCTMCSCCCCCCATSKGTTCTTCTGSTGGCCGSCTRAYGAGGKGKWGAAGCCM
SWKMAGAWRTTGWGMKTACTCCTGATCCACRTCTCGTWAGAAACGGGKMCAMGAAMAGAGARAARCRWARCWAWYSA
TTCCTTCCGTCCSCTTAGGTTYCGTACSCSRGMKGRGTGGGWCACWCMAMMSKKTTKGAAYMATWAKWAAMCTGCSA
ARCCCGGCCATAWKACYATAWCCCAAAMYAWTAWWTATWWGGRCSGMSCSYGYAGRACTSTWCYWCTRCYMCTATCC
KYYCWWCCCYRRAWSCCKKARTTSCCMWMCTACTRYMAGATRGAKATKMRTWAGAMGTKCYGGGCAGYGCGGGYGAM
GCGGYWAACRGKKYCTSMGGSTGMTMTRTMTTGMYWYGKGGRAYMCAGYCMKGKARRKWRTTAWAAKGRCAAYWTYT
TTMSAR

Primer KL69

>CLGM3 chuR clone 5 primer KL69
TAAGAMCGCTTMGTCGGACGCGGACGGCGAACCCGGTCTGAACTATTTGTGCAAAGGATATTACCAATACTTTCAGC
ATGTAGCTCCCTATATGGATTTCATGAAAAAAGAATTAATGAATCAACAGGCTCCTGCCAACATCATGAAAGCACTA
AAAGACGGAAGTTTAAAAATAGAATATTAACGCKTTGGTGYCTTTTGKKCGGATKGKSTTGCYGSATMAMCTTACGA
GSRCGGKCARGTGKAGGWAAAGAAMCKCCMCCCAACYTTCWKSCSWYGCYWGCGAGGGCGGTWGCSGGRGCSAAAAA
RAAAARAGGTTYGTTKTKYCTGCCTTKYTTCWASGSCAGMKGAGRARAGSAAGASARGAARGGCTGRTGARAGTYKC
CRYAAMMTSTGGACGYGGRCTAMKCRMGAGGKGGSSCCKCSARCCRTKATGGAWAWAGGASRKACTCRGATGCCGAA
CCMTGGTACTATWATCCAWYMRACYAMWTAKMAAMGSRGRKCCMAGGGAMARRAAWTCACTKTCATTSTCGSKAYMA
MSCCCTSGCGSCACKAGATCYCTCCGCTGRKMGAAGATKAASTAGGATAAARGMWGTTCCYRYCCAGTSCGGCTCAY
KMAKGMAASRGRTCAAMKGSTGYGASYAGCSTGCKMTASCTGSSGSYWRCCKGTMWKSRTAWTCTCTCMKKWWKATR
TAGAGAGCA

D.3 BLAST analyses

BLAST of *B. theta chuR* against NCBI WGS metagenome contigs

The following page summarizes the results of BLAST analysis using the Megablast algorithm, querying the *B. theta* VPI-5482 *chuR/anSME* gene (BT_0238) against the NCBI WGS database, specifying tax_id 408169 for assembled metagenomic contigs.

Table D.1: BLAST (megablast) results for *B. theta chuR* against metagenomic contigs.

Description	Max score	Total score	Query cover	E value	Ident	Accession
gut metagenome genome assembly P2E7-k21-2014-09-20, contig contig-179000065, whole genome shotgun sequence	2300	2300	100.00%	0	100.00%	CEBV01025663.1
gut metagenome genome assembly P2E0-k21-2014-09-20, contig contig-4470000086, whole genome shotgun sequence	2300	2300	100.00%	0	100.00%	CEAB01052623.1
gut metagenome genome assembly P22E0-k21-2014-09-20, contig contig-576000124, whole genome shotgun sequence	2300	2300	100.00%	0	100.00%	CDZR01059274.1
gut metagenome genome assembly P22E7-k21-2014-09-20, contig contig-4000110, whole genome shotgun sequence	2300	2300	100.00%	0	100.00%	CDZN01021567.1
gut metagenome genome assembly P15E7-k21-2014-09-20, contig contig-58, whole genome shotgun sequence	2300	2300	100.00%	0	100.00%	CDYN01010010.1
gut metagenome genome assembly P3E7-k21-2014-09-20, contig contig-32000034, whole genome shotgun sequence	2289	2289	100.00%	0	99.00%	CEAK01009572.1
gut metagenome genome assembly P23C90-k21-2014-09-20, contig contig-79000054, whole genome shotgun sequence	2274	2274	98.00%	0	100.00%	CDZU01019025.1
gut metagenome genome assembly P20E0-k21-2014-09-20, contig contig-328000126, whole genome shotgun sequence	2139	2139	100.00%	0	98.00%	CDZL01023776.1
gut metagenome genome assembly P20E90-k21-2014-09-20, contig contig-1784000118, whole genome shotgun sequence	2139	2139	100.00%	0	98.00%	CDZJ01030116.1
gut metagenome genome assembly P11E7-k21-2014-09-20, contig contig-475000077, whole genome shotgun sequence	2139	2139	100.00%	0	98.00%	CDYJ01032401.1
gut metagenome genome assembly P11E7-k21-2014-09-20, contig contig-64000044, whole genome shotgun sequence	2139	2139	100.00%	0	98.00%	CDYJ01018206.1
gut metagenome genome assembly P13E90-k21-2014-09-20, contig contig-4649000084, whole genome shotgun sequence	2134	2134	100.00%	0	98.00%	CDYU01038733.1
gut metagenome genome assembly P13E7-k21-2014-09-20, contig contig-2656000067, whole genome shotgun sequence	2134	2134	100.00%	0	98.00%	CDYM01021406.1
gut metagenome genome assembly P22E90-k21-2014-09-20, contig contig-158000023, whole genome shotgun sequence	1796	1796	78.00%	0	100.00%	CDZS01010650.1
gut metagenome genome assembly P17E90-k21-2014-09-20, contig contig-625000026, whole genome shotgun sequence	1679	1679	78.00%	0	98.00%	CDZK01010597.1
gut metagenome genome assembly P9E7-k21-2014-09-20, contig contig-114000075, whole genome shotgun sequence	1546	1546	100.00%	0	89.00%	CDZX01019418.1
gut metagenome genome assembly P14E90-k21-2014-09-20, contig contig-1458000087, whole genome shotgun sequence	1546	1546	100.00%	0	89.00%	CDZB01057382.1
gut metagenome genome assembly P14E7-k21-2014-09-20, contig contig-10000044, whole genome shotgun sequence	1546	1546	100.00%	0	89.00%	CDZA01027718.1
gut metagenome genome assembly P11E90-k21-2014-09-20, contig contig-3536000126, whole genome shotgun sequence	1546	1546	100.00%	0	89.00%	CDYR01053020.1
gut metagenome genome assembly P11E90-k21-2014-09-20, contig contig-76000025, whole genome shotgun sequence	1546	1546	100.00%	0	89.00%	CDYR01010643.1
gut metagenome genome assembly P10E90-k21-2014-09-20, contig contig-39000086, whole genome shotgun sequence	1546	1546	100.00%	0	89.00%	CDYK01030801.1
gut metagenome genome assembly P10E0-k21-2014-09-20, contig contig-19000019, whole genome shotgun sequence	1546	1546	100.00%	0	89.00%	CDYI01008742.1
gut metagenome genome assembly P9E90-k21-2014-09-20, contig contig-51000017, whole genome shotgun sequence	1546	1546	100.00%	0	89.00%	CDTY01007659.1
gut metagenome genome assembly P38C7-k21-2014-09-20, contig contig-101000018, whole genome shotgun sequence	1541	1541	100.00%	0	89.00%	CEAH01010496.1
gut metagenome genome assembly P38C90-k21-2014-09-20, contig contig-618000043, whole genome shotgun sequence	1541	1541	100.00%	0	89.00%	CEAG01020626.1
gut metagenome genome assembly P38C0-k21-2014-09-20, contig contig-657000020, whole genome shotgun sequence	1541	1541	100.00%	0	89.00%	CEAF01011572.1
gut metagenome genome assembly P25C90-k21-2014-09-20, contig contig-3488000001, whole genome shotgun sequence	1541	1541	100.00%	0	89.00%	CEAA01000885.1
gut metagenome genome assembly P25C7-k21-2014-09-20, contig contig-2000100, whole genome shotgun sequence	1541	1541	100.00%	0	89.00%	CDZY01046211.1
gut metagenome genome assembly P25C0-k21-2014-09-20, contig contig-3510000034, whole genome shotgun sequence	1541	1541	100.00%	0	89.00%	CDZW01013549.1
gut metagenome genome assembly P11E7-k21-2014-09-20, contig contig-403000088, whole genome shotgun sequence	1541	1541	99.00%	0	89.00%	CDYJ01036724.1
gut metagenome genome assembly P15E90-k21-2014-09-20, contig contig-135000065, whole genome shotgun sequence	1360	1360	59.00%	0	100.00%	CDYY01027497.1
Human gut metagenome DNA, contig sequence: F2-Y_034152, whole genome shotgun sequence	1229	1229	70.00%	0	92.00%	BABA01034152.1
Chicken gut metagenome c108720, whole genome shotgun sequence	1227	1227	56.00%	0	98.00%	JFBN01021268.1
gut metagenome genome assembly P20E7-k21-2014-09-20, contig contig-146000123, whole genome shotgun sequence	1175	1175	53.00%	0	98.00%	CDZM01024581.1
gut metagenome genome assembly P3E7-k21-2014-09-20, contig contig-211000094, whole genome shotgun sequence	1157	1157	93.00%	0	85.00%	CEAK01026614.1
gut metagenome genome assembly P6C0-k21-2014-09-20, contig contig-37000001, whole genome shotgun sequence	1153	1153	100.00%	0	83.00%	CEBY01000578.1
gut metagenome genome assembly P6C7-k21-2014-09-20, contig contig-846000044, whole genome shotgun sequence	1153	1153	100.00%	0	83.00%	CEAZ01023225.1
gut metagenome genome assembly P6C7-k21-2014-09-20, contig contig-75000008, whole genome shotgun sequence	1153	1153	100.00%	0	83.00%	CEAZ01004158.1
gut metagenome genome assembly P6C90-k21-2014-09-20, contig contig-95, whole genome shotgun sequence	1153	1153	100.00%	0	83.00%	CEAX01040480.1
gut metagenome genome assembly P13E7-k21-2014-09-20, contig contig-86000092, whole genome shotgun sequence	1153	1153	100.00%	0	83.00%	CDYM01029143.1
gut metagenome genome assembly P10E90-k21-2014-09-20, contig contig-3440000062, whole genome shotgun sequence	1153	1153	100.00%	0	83.00%	CDYK01022411.1
gut metagenome genome assembly P10E7-k21-2014-09-20, contig contig-820000014, whole genome shotgun sequence	1153	1153	100.00%	0	83.00%	CDYF01004576.1
gut metagenome genome assembly P8C0-k21-2014-09-20, contig contig-7000103, whole genome shotgun sequence	1134	1134	97.00%	0	84.00%	CECJ01021445.1
gut metagenome genome assembly P8C90-k21-2014-09-20, contig contig-56000001, whole genome shotgun sequence	1134	1134	97.00%	0	84.00%	CEAI01000265.1
gut metagenome genome assembly P8C7-k21-2014-09-20, contig contig-1902000025, whole genome shotgun sequence	1134	1134	97.00%	0	84.00%	CEAE01006842.1
gut metagenome genome assembly P12E90-k21-2014-09-20, contig contig-124000056, whole genome shotgun sequence	1134	1134	97.00%	0	84.00%	CDYL01015147.1
gut metagenome genome assembly P12E7-k21-2014-09-20, contig contig-1108000122, whole genome shotgun sequence	1134	1134	97.00%	0	84.00%	CDYE01010333.1
gut metagenome genome assembly P22E90-k21-2014-09-20, contig contig-1259000039, whole genome shotgun sequence	1120	1120	98.00%	0	83.00%	CDZS01017641.1
gut metagenome genome assembly P3E7-k21-2014-09-20, contig contig-13000052, whole genome shotgun sequence	1118	1118	97.00%	0	83.00%	CEAK01014454.1
gut metagenome genome assembly P11E90-k21-2014-09-20, contig contig-358000043, whole genome shotgun sequence	1118	1118	97.00%	0	83.00%	CDYR01018557.1
gut metagenome genome assembly P11E7-k21-2014-09-20, contig contig-484000011, whole genome shotgun sequence	1118	1118	97.00%	0	83.00%	CDYJ01004907.1
gut metagenome genome assembly P11E0-k21-2014-09-20, contig contig-427000011, whole genome shotgun sequence	1118	1118	97.00%	0	83.00%	CDYG01004237.1
Uncultured Bacteroides sp. TS29_contig120613, whole genome shotgun sequence	1118	1118	97.00%	0	83.00%	ADJT01001577.1
gut metagenome genome assembly P15E90-k21-2014-09-20, contig contig-3699000064, whole genome shotgun sequence	1114	1114	98.00%	0	83.00%	CDYY01027356.1
gut metagenome genome assembly P22E7-k21-2014-09-20, contig contig-580000085, whole genome shotgun sequence	1112	1112	97.00%	0	83.00%	CDZN01016749.1
gut metagenome genome assembly P1E90-k21-2014-09-20, contig contig-3956000077, whole genome shotgun sequence	1112	1112	97.00%	0	83.00%	CDZF01029773.1
gut metagenome genome assembly P17E90-k21-2014-09-20, contig contig-1000031, whole genome shotgun sequence	1107	1107	97.00%	0	83.00%	CDZK01012354.1
gut metagenome genome assembly P17E0-k21-2014-09-20, contig contig-2522000019, whole genome shotgun sequence	1107	1107	97.00%	0	83.00%	CDYT01005576.1
gut metagenome genome assembly P8C0-k21-2014-09-20, contig contig-1526000091, whole genome shotgun sequence	1098	1098	100.00%	0	83.00%	CECJ01019076.1
gut metagenome genome assembly P4E90-k21-2014-09-20, contig contig-81, whole genome shotgun sequence	1098	1098	100.00%	0	83.00%	CEAN01008867.1
gut metagenome genome assembly P4E0-k21-2014-09-20, contig contig-1344000072, whole genome shotgun sequence	1098	1098	100.00%	0	83.00%	CEAM01008655.1
gut metagenome genome assembly P2E90-k21-2014-09-20, contig contig-267000058, whole genome shotgun sequence	1098	1098	100.00%	0	83.00%	CEAC01018913.1
gut metagenome genome assembly P22E90-k21-2014-09-20, contig contig-174000057, whole genome shotgun sequence	1098	1098	100.00%	0	83.00%	CDZS01025645.1
gut metagenome genome assembly P22E0-k21-2014-09-20, contig contig-151000108, whole genome shotgun sequence	1098	1098	100.00%	0	83.00%	CDZR01051462.1
gut metagenome genome assembly P22E7-k21-2014-09-20, contig contig-115000011, whole genome shotgun sequence	1098	1098	100.00%	0	83.00%	CDZN01002285.1
gut metagenome genome assembly P20E0-k21-2014-09-20, contig contig-4, whole genome shotgun sequence	1098	1098	100.00%	0	83.00%	CDZL01000577.1
gut metagenome genome assembly P20E90-k21-2014-09-20, contig contig-1000054, whole genome shotgun sequence	1098	1098	100.00%	0	83.00%	CDZJ01013511.1
gut metagenome genome assembly P11E7-k21-2014-09-20, contig contig-82000075, whole genome shotgun sequence	1098	1098	100.00%	0	83.00%	CDYJ01031218.1
gut metagenome genome assembly P4E7-k21-2014-09-20, contig contig-7, whole genome shotgun sequence	1092	1092	100.00%	0	83.00%	CECB01000484.1
gut metagenome genome assembly P2E7-k21-2014-09-20, contig contig-13000091, whole genome shotgun sequence	1092	1092	100.00%	0	83.00%	CEBV01035922.1
gut metagenome genome assembly P2E0-k21-2014-09-20, contig contig-10000091, whole genome shotgun sequence	1092	1092	100.00%	0	83.00%	CEAB01054968.1
gut metagenome genome assembly P21E90-k21-2014-09-20, contig contig-39000037, whole genome shotgun sequence	1092	1092	100.00%	0	83.00%	CDZQ01017534.1
gut metagenome genome assembly P21E0-k21-2014-09-20, contig contig-3967000011, whole genome shotgun sequence	1092	1092	100.00%	0	83.00%	CDZP01004662.1
gut metagenome genome assembly P21E7-k21-2014-09-20, contig contig-67, whole genome shotgun sequence	1092	1092	100.00%	0	83.00%	CDZO01021369.1
gut metagenome genome assembly P17E7-k21-2014-09-20, contig contig-1796000017, whole genome shotgun sequence	1092	1092	100.00%	0	83.00%	CDYP01003565.1
gut metagenome genome assembly P9E0-k21-2014-09-20, contig contig-52000056, whole genome shotgun sequence	1086	1086	100.00%	0	82.00%	CEAD01017927.1
Human gut metagenome DNA, contig sequence: F2-Y_034151, whole genome shotgun sequence	1086	1086	63.00%	0	91.00%	BABA01034151.1
gut metagenome genome assembly P20E7-k21-2014-09-20, contig contig-52000037, whole genome shotgun sequence	1085	1085	97.00%	0	83.00%	CDZM01007421.1
gut metagenome genome assembly P5E90-k21-2014-09-20, contig contig-31, whole genome shotgun sequence	1081	1081	100.00%	0	82.00%	CEAQ01005343.1
gut metagenome genome assembly P5E7-k21-2014-09-20, contig contig-1000113, whole genome shotgun sequence	1081	1081	100.00%	0	82.00%	CEAP01016593.1
gut metagenome genome assembly P5E0-k21-2014-09-20, contig contig-1000008, whole genome shotgun sequence	1081	1081	100.00%	0	82.00%	CEAO01001616.1
gut metagenome genome assembly P25C90-k21-2014-09-20, contig contig-13000090, whole genome shotgun sequence	1081	1081	100.00%	0	82.00%	CEAA01038482.1
gut metagenome genome assembly P25C7-k21-2014-09-20, contig contig-69000065, whole genome shotgun sequence	1081	1081	100.00%	0	82.00%	CDZY01029940.1
gut metagenome genome assembly P25C0-k21-2014-09-20, contig contig-171000038, whole genome shotgun sequence	1081	1081	100.00%	0	82.00%	CDZW01014946.1
gut metagenome genome assembly P23C7-k21-2014-09-20, contig contig-4814000032, whole genome shotgun sequence	1081	1081	100.00%	0	82.00%	CDZV01016032.1
gut metagenome genome assembly P23C90-k21-2014-09-20, contig contig-5000026, whole genome shotgun sequence	1081	1081	100.00%	0	82.00%	CDZU01009275.1
gut metagenome genome assembly P23C0-k21-2014-09-20, contig contig-4924000103, whole genome shotgun sequence	1081	1081	100.00%	0	82.00%	CDZT01052918.1
gut metagenome genome assembly P11E90-k21-2014-09-20, contig contig-3000020, whole genome shotgun sequence	1081	1081	100.00%	0	82.00%	CDYR01008444.1
gut metagenome genome assembly P12E90-k21-2014-09-20, contig contig-6, whole genome shotgun sequence	1081	1081	100.00%	0	82.00%	CDYL01001587.1
gut metagenome genome assembly P12E0-k21-2014-09-20, contig contig-98, whole genome shotgun sequence	1081	1081	100.00%	0	82.00%	CDYH01018076.1
gut metagenome genome assembly P12E7-k21-2014-09-20, contig contig-13000083, whole genome shotgun sequence	1081	1081	100.00%	0	82.00%	CDYE01007002.1
gut metagenome genome assembly P9E90-k21-2014-09-20, contig contig-543000015, whole genome shotgun sequence	1081	1081	100.00%	0	82.00%	CDTY01007069.1
gut metagenome genome assembly P18E7-k21-2014-09-20, contig contig-118000077, whole genome shotgun sequence	1070	1070	100.00%	0	82.00%	CDZC01019985.1
gut metagenome genome assembly P17E90-k21-2014-09-20, contig contig-1230000094, whole genome shotgun sequence	1042	1042	96.00%	0	82.00%	CDZK01036991.1
gut metagenome genome assembly P5E90-k21-2014-09-20, contig contig-113000023, whole genome shotgun sequence	1033	1033	45.00%	0	99.00%	CEAQ01004113.1
gut metagenome genome assembly P10E7-k21-2014-09-20, contig contig-820000010, whole genome shotgun sequence	918	918	79.00%	0	83.00%	CDYF01003333.1
gut metagenome genome assembly P9E7-k21-2014-09-20, contig contig-850000067, whole genome shotgun sequence	891	891	77.00%	0	83.00%	CDZX01017508.1
gut metagenome genome assembly P3E0-k21-2014-09-20, contig contig-335000076, whole genome shotgun sequence	880	880	80.00%	0	82.00%	CEAJ01016762.1
gut metagenome genome assembly P3E90-k21-2014-09-20, contig contig-692000068, whole genome shotgun sequence	859	859	67.00%	0	85.00%	CEAL01020555.1
gut metagenome genome assembly P17E0-k21-2014-09-20, contig contig-1507000010, whole genome shotgun sequence	843	843	80.00%	0	82.00%	CDYT01003220.1

BLAST of *B. theta chuR* against NCBI metagenomic proteins

The following page summarizes the results of BLAST analysis using blastx, querying the translated *B. theta* VPI-5482 *chuR/anSME* gene (BT_0238) against the NCBI env_nr database for matches to metagenomic proteins. Bolded results indicate proteins described as regulators of sulfatases.

Table D.2: BLAST (blastx) results for *B. theta chuR* against metagenomic proteins.

Description	Max score	Total score	Query cover	E value	Identity	Accession
regulator of arylsulfatase activity [gut metagenome]	**789**	**789**	**99.00%**	**0**	**99.00%**	**EJX02180.1**
transcriptional regulator [gut metagenome]	746	746	97.00%	0	85.00%	EJX05687.1
regulator of arylsulfatase activity [gut metagenome]	**696**	**696**	**97.00%**	**0**	**79.00%**	**EJW97079.1**
regulator of arylsulfatase activity [gut metagenome]	**653**	**653**	**97.00%**	**0**	**73.00%**	**EJW94445.1**
regulator of arylsulfatase activity [gut metagenome]	**589**	**589**	**97.00%**	**0**	**65.00%**	**EJX08867.1**
regulator of arylsulfatase activity [mine drainage metagenome]	420	420	94.00%	9.00E-141	49.00%	CBI09006.1
hypothetical protein GOS 1705184 [marine metagenome]	408	408	99.00%	9.00E-135	46.00%	EDJ38325.1
Anaerobic sulfatase-maturating enzyme-like protein AslB [human gut metagenome]	**384**	**384**	**94.00%**	**1.00E-127**	**45.00%**	**ETJ17740.1**
hypothetical protein GOS 2771047 [marine metagenome]	381	381	93.00%	3.00E-126	46.00%	ECW15524.1
hypothetical protein GOS 9576743 [marine metagenome]	359	359	94.00%	3.00E-117	43.00%	EBF56645.1
hypothetical protein GOS 1912926 [marine metagenome]	342	342	90.00%	5.00E-111	43.00%	EDA92056.1
hypothetical protein GOS 2939750 [marine metagenome]	342	342	90.00%	8.00E-111	43.00%	ECV22775.1
hypothetical protein LCGC14 0569180 [marine sediment metagenome]	333	333	96.00%	1.00E-107	40.00%	KKN56736.1
unnamed protein product [marine sediment metagenome]	322	322	89.00%	2.00E-104	43.00%	GAF78354.1
hypothetical protein GOS 1100717 [marine metagenome]	321	321	92.00%	1.00E-102	41.00%	EDE61078.1
hypothetical protein GOS 9380034 [marine metagenome]	291	291	62.00%	1.00E-093	50.00%	EBG75305.1
hypothetical protein LCGC14 0093800 [marine sediment metagenome]	275	275	90.00%	2.00E-085	38.00%	KKO03456.1
unnamed protein product [marine sediment metagenome]	250	250	80.00%	8.00E-077	40.00%	GAF67375.1
hypothetical protein LCGC14 0644950 [marine sediment metagenome]	249	249	94.00%	1.00E-075	35.00%	KKN49236.1
unnamed protein product [marine sediment metagenome]	247	247	68.00%	3.00E-076	41.00%	GAF90965.1
hypothetical protein LCGC14 0691880 [marine sediment metagenome]	244	244	93.00%	1.00E-073	35.00%	KKN44555.1
hypothetical protein GOS 9597097 [marine metagenome]	232	232	88.00%	6.00E-069	34.00%	EBF44117.1
unnamed protein product [marine sediment metagenome]	230	230	61.00%	5.00E-070	41.00%	GAI42214.1
unnamed protein product [marine sediment metagenome]	216	216	55.00%	6.00E-065	47.00%	GAI16149.1
unnamed protein product [marine sediment metagenome]	213	213	50.00%	3.00E-064	49.00%	GAJ02304.1
hypothetical protein GOS 4005653 [marine metagenome]	185	185	39.00%	2.00E-053	50.00%	ECG00326.1
hypothetical protein LCGC14 3094110 [marine sediment metagenome]	181	181	69.00%	1.00E-050	34.00%	KKK53506.1
Radical SAM domain protein [mine drainage metagenome]	**179**	**179**	**52.00%**	**3.00E-061**	**40.00%**	**EQD53795.1**
unnamed protein product [marine sediment metagenome]	178	178	68.00%	7.00E-050	36.00%	GAI74417.1
unnamed protein product [marine sediment metagenome]	176	176	63.00%	1.00E-049	37.00%	GAG65792.1
Arylsulfatase regulator (Fe-S oxidoreductase) [human gut metagenome]	**168**	**168**	**34.00%**	**1.00E-047**	**53.00%**	**EKC66373.1**
regulator of arylsulfatase activity [human gut metagenome]	**164**	**164**	**33.00%**	**4.00E-046**	**54.00%**	**EKC67400.1**
hypothetical protein LCGC14 0496100 [marine sediment metagenome]	162	162	55.00%	1.00E-044	36.00%	KKN63973.1
unnamed protein product [marine sediment metagenome]	146	146	33.00%	1.00E-039	49.00%	GAG76514.1
anaerobic sulfatase-maturating enzyme [gut metagenome]	**144**	**144**	**64.00%**	**2.00E-037**	**31.00%**	**EJW98913.1**
unnamed protein product [marine sediment metagenome]	141	141	34.00%	1.00E-037	48.00%	GAG04863.1
hypothetical protein GOS 2819697 [marine metagenome]	136	136	62.00%	2.00E-033	32.00%	ECV87608.1
unnamed protein product [marine sediment metagenome]	135	135	27.00%	5.00E-036	53.00%	GAG98419.1
unnamed protein product [marine sediment metagenome]	130	130	39.00%	2.00E-033	40.00%	GAJ24261.1
unnamed protein product [marine sediment metagenome]	129	129	40.00%	6.00E-033	41.00%	GAJ09634.1
unnamed protein product [marine sediment metagenome]	127	127	38.00%	2.00E-032	41.00%	GAG08714.1
hypothetical protein GOS 1405925 [marine metagenome]	127	127	83.00%	2.00E-030	27.00%	EDC83749.1
unnamed protein product [marine sediment metagenome]	124	124	28.00%	1.00E-031	49.00%	GAI25361.1
unnamed protein product [marine sediment metagenome]	123	123	28.00%	2.00E-031	48.00%	GAH39779.1
unnamed protein product [marine sediment metagenome]	120	120	30.00%	4.00E-030	45.00%	GAH65020.1
hypothetical protein GOS 1097727 [marine metagenome]	116	116	46.00%	4.00E-027	34.00%	EDE62797.1
unnamed protein product [marine sediment metagenome]	115	115	29.00%	2.00E-028	46.00%	GAH01613.1
hypothetical protein GOS 2866977 [marine metagenome]	115	115	52.00%	3.00E-027	33.00%	ECV60952.1
unnamed protein product [marine sediment metagenome]	115	115	51.00%	2.00E-027	32.00%	GAF67662.1
unnamed protein product [marine sediment metagenome]	114	114	26.00%	1.00E-027	47.00%	GAI73590.1
hypothetical protein GOS 1921962 [marine metagenome]	113	113	83.00%	1.00E-025	25.00%	EDA87173.1
hypothetical protein GOS 9432695 [marine metagenome]	110	110	40.00%	4.00E-025	32.00%	EBG44141.1
hypothetical protein GOS 1740763 [marine metagenome]	108	108	24.00%	8.00E-026	49.00%	EDJ18224.1
unnamed protein product [marine sediment metagenome]	108	108	38.00%	9.00E-026	33.00%	GAH64000.1
unnamed protein product [marine sediment metagenome]	108	108	38.00%	4.00E-025	33.00%	GAG22266.1
hypothetical protein GOS 8408113 [marine metagenome]	104	104	40.00%	4.00E-022	33.00%	EBM44987.1
hypothetical protein LCGC14 2565200 [marine sediment metagenome]	102	102	75.00%	4.00E-022	26.00%	KKL09503.1
hypothetical protein GOS 4562577 [marine metagenome]	100	100	32.00%	6.00E-023	38.00%	ECJ37744.1
hypothetical protein GOS 9386950 [marine metagenome]	98.6	98.6	56.00%	2.00E-021	24.00%	EBG71229.1
sulfatase regulatory protein [mine drainage metagenome]	**95.5**	**95.5**	**18.00%**	**8.00E-022**	**57.00%**	**EQD66842.1**
unnamed protein product [marine sediment metagenome]	95.1	95.1	22.00%	2.00E-021	48.00%	GAG44443.1
hypothetical protein GOS 5191566 [marine metagenome]	95.1	95.1	35.00%	1.00E-020	33.00%	ECB22898.1
hypothetical protein GOS 5526512 [marine metagenome]	92.8	92.8	64.00%	6.00E-019	25.00%	ECT54974.1
unnamed protein product [marine sediment metagenome]	92	92	34.00%	2.00E-019	34.00%	GAG90103.1
unnamed protein product [marine sediment metagenome]	89	89	22.00%	3.00E-019	43.00%	GAH43108.1
unnamed protein product [marine sediment metagenome]	88.2	88.2	20.00%	5.00E-019	45.00%	GAH16442.1
hypothetical protein GOS 7334446 [marine metagenome]	87.8	87.8	78.00%	4.00E-017	25.00%	EBT06716.1
radical SAM domain-containing protein [mine drainage metagenome]	**87.4**	**87.4**	**88.00%**	**1.00E-016**	**22.00%**	**EQD53563.1**
hypothetical protein GOS 1503414 [marine metagenome]	87	87	89.00%	1.00E-016	24.00%	EDC28425.1
hypothetical protein GOS 1957776 [marine metagenome]	86.7	86.7	56.00%	2.00E-017	27.00%	EDA67724.1
hypothetical protein GOS 9494848 [marine metagenome]	85.9	85.9	22.00%	3.00E-017	43.00%	EBG06992.1
hypothetical protein OBE 11873 [human gut metagenome]	80.1	80.1	11.00%	2.00E-016	72.00%	EKC54820.1
hypothetical protein GOS 9515512 [marine metagenome]	79.3	79.3	18.00%	5.00E-015	49.00%	EBF94215.1
unnamed protein product [marine sediment metagenome]	77.4	77.4	19.00%	1.00E-014	47.00%	GAG97456.1
unnamed protein product [marine sediment metagenome]	74.7	74.7	30.00%	6.00E-014	32.00%	GAI92648.1
radical SAM domain-containing protein [human gut metagenome]	**74.7**	**74.7**	**43.00%**	**2.00E-012**	**30.00%**	**EKC58775.1**
hypothetical protein GOS 7012149 [marine metagenome]	74.7	74.7	36.00%	2.00E-013	28.00%	EBU74838.1
hypothetical protein GOS 9617987 [marine metagenome]	73.9	73.9	41.00%	1.00E-012	30.00%	EBF31477.1
unnamed protein product [marine sediment metagenome]	73.6	73.6	20.00%	1.00E-013	37.00%	GAJ17489.1
hypothetical protein LCGC14 1197950 [marine sediment metagenome]	73.6	73.6	79.00%	5.00E-012	23.00%	KKM94474.1
unnamed protein product [marine sediment metagenome]	73.2	73.2	26.00%	3.00E-013	34.00%	GAH46004.1
arylsulfatase regulator [human gut metagenome]	**73.2**	**73.2**	**37.00%**	**6.00E-013**	**26.00%**	**EKC74149.1**
unnamed protein product [marine sediment metagenome]	71.2	71.2	16.00%	6.00E-013	48.00%	GAJ20278.1
hypothetical protein LCGC14 1882520 [marine sediment metagenome]	70.9	70.9	13.00%	4.00E-011	55.00%	KKL92659.1
hypothetical protein LCGC14 1566370 [marine sediment metagenome]	70.1	70.1	49.00%	4.00E-011	29.00%	KKM29052.1
hypothetical protein LCGC14 1351350 [marine sediment metagenome]	68.6	68.6	18.00%	1.00E-011	43.00%	KKM79300.1
hypothetical protein GOS 2993326 [marine metagenome]	68.6	68.6	38.00%	2.00E-010	32.00%	ECU92664.1
Radical SAM domain protein [mine drainage metagenome]	**68.6**	**68.6**	**38.00%**	**5.00E-011**	**26.00%**	**EQD79036.1**
unnamed protein product [marine sediment metagenome]	68.2	68.2	32.00%	9.00E-011	29.00%	GAH57071.1
hypothetical protein Q604 UNBC13573G0001 [human gut metagenome]	66.2	66.2	15.00%	3.00E-011	45.00%	ETJ31946.1
unnamed protein product [marine sediment metagenome]	65.5	65.5	43.00%	7.00E-010	26.00%	GAF75495.1
hypothetical protein GOS 9625176 [marine metagenome]	63.9	63.9	81.00%	5.00E-009	24.00%	EBF27058.1
unnamed protein product [marine sediment metagenome]	63.2	63.2	29.00%	1.00E-009	35.00%	GAI24881.1
hypothetical protein GOS 9446156 [marine metagenome]	62.8	62.8	68.00%	1.00E-008	25.00%	EBG35936.1
hypothetical protein LCGC14 0223720 [marine sediment metagenome]	62.4	62.4	41.00%	2.00E-008	26.00%	KKN90789.1
hypothetical protein LCGC14 0491760 [marine sediment metagenome]	62	62	82.00%	2.00E-008	22.00%	KKN64433.1
unnamed protein product [marine sediment metagenome]	61.2	61.2	18.00%	2.00E-009	38.00%	GAG90102.1
hypothetical protein GOS 5487000 [marine metagenome]	60.8	60.8	57.00%	3.00E-008	22.00%	ECE73955.1
unnamed protein product [marine sediment metagenome]	60.5	60.5	25.00%	1.00E-008	35.00%	GAI37089.1
unnamed protein product [marine sediment metagenome]	60.5	60.5	37.00%	1.00E-008	31.00%	GAG31974.1

BLAST of *B. theta chuR* against NCBI Refseq proteins

The following page summarizes the results of BLAST analysis using blastx, querying the translated *B. theta* VPI-5482 *chuR/anSME* gene (BT_0238) against the NCBI Refseq protein database for matches to known proteins.

Table D.3: BLAST (blastx) results for *B. theta chuR* against Refseq proteins.

Description	Max score	Score	Query cover	E value	Identity	Accession
MULTISPECIES: anaerobic sulfatase maturase [Bacteroides]	874	874	99.00%	0	100.00%	WP 008766211.1
anaerobic sulfatase maturase [Bacteroides thetaiotaomicron]	872	872	99.00%	0	99.00%	WP 016267954.1
anaerobic sulfatase maturase [Bacteroides thetaiotaomicron]	871	871	99.00%	0	99.00%	WP 022471893.1
anaerobic sulfatase-maturase [Bacteroides thetaiotaomicron]	870	870	99.00%	0	99.00%	WP 048697144.1
anaerobic sulfatase maturase [Bacteroides thetaiotaomicron]	869	869	99.00%	0	99.00%	WP 054959252.1
anaerobic sulfatase maturase [Bacteroides faecis]	851	851	99.00%	0	98.00%	WP 010537511.1
anaerobic sulfatase maturase [Bacteroides thetaiotaomicron]	850	850	99.00%	0	97.00%	WP 022301748.1
anaerobic sulfatase maturase [Bacteroides caccae]	814	814	99.00%	0	91.00%	WP 005680548.1
MULTISPECIES: anaerobic sulfatase maturase [Bacteroides]	811	811	99.00%	0	91.00%	WP 004297342.1
anaerobic sulfatase maturase [Bacteroides faecichinchillae]	810	810	99.00%	0	91.00%	WP 025074644.1
anaerobic sulfatase maturase [Bacteroides finegoldii]	808	808	99.00%	0	91.00%	WP 007759188.1
MULTISPECIES: anaerobic sulfatase maturase [Bacteroides]	813	813	99.00%	0	90.00%	WP 008643298.1
MULTISPECIES: anaerobic sulfatase maturase [Bacteroides]	811	811	99.00%	0	90.00%	WP 008021790.1
MULTISPECIES: anaerobic sulfatase maturase [Bacteroides]	810	810	99.00%	0	90.00%	WP 004315949.1
anaerobic sulfatase maturase [Bacteroides ovatus]	809	809	99.00%	0	90.00%	WP 004306013.1
anaerobic sulfatase maturase [Bacteroides finegoldii]	808	808	99.00%	0	90.00%	WP 022276071.1
anaerobic sulfatase maturase [Bacteroides finegoldii]	807	807	99.00%	0	90.00%	WP 032839687.1
anaerobic sulfatase maturase [Bacteroides ovatus]	807	807	99.00%	0	90.00%	WP 004319514.1
anaerobic sulfatase maturase [Bacteroides acidifaciens]	802	802	99.00%	0	90.00%	WP 044656247.1
anaerobic sulfatase maturase [Bacteroides pyogenes]	796	796	99.00%	0	90.00%	WP 027326227.1
anaerobic sulfatase maturase [Bacteroides pyogenes]	793	793	99.00%	0	89.00%	WP 021646122.1
MULTISPECIES: anaerobic sulfatase maturase [Bacteroides]	772	772	98.00%	0	87.00%	WP 002561758.1
anaerobic sulfatase maturase [Bacteroides salyersiae]	768	768	98.00%	0	86.00%	WP 005923804.1
anaerobic sulfatase maturase [Bacteroides fragilis]	766	766	98.00%	0	86.00%	WP 042985698.1
MULTISPECIES: anaerobic sulfatase maturase [Bacteroides]	766	766	98.00%	0	86.00%	WP 005789094.1
anaerobic sulfatase maturase [Bacteroides fragilis]	765	765	98.00%	0	86.00%	WP 032570972.1
anaerobic sulfatase maturase [Bacteroides fragilis]	764	764	98.00%	0	86.00%	WP 014299157.1
anaerobic sulfatase maturase [Bacteroides fragilis]	763	763	98.00%	0	86.00%	WP 010993230.1
anaerobic sulfatase maturase [Bacteroides fragilis]	762	762	98.00%	0	86.00%	WP 032580200.1
anaerobic sulfatase maturase [Bacteroides fragilis]	762	762	98.00%	0	86.00%	WP 032528627.1
anaerobic sulfatase maturase [Bacteroidaceae bacterium MS4]	756	756	97.00%	0	86.00%	WP 042368025.1
anaerobic sulfatase maturase [Bacteroides fragilis]	768	768	99.00%	0	85.00%	WP 005807432.1
MULTISPECIES: anaerobic sulfatase maturase [Bacteroides]	766	766	99.00%	0	85.00%	WP 032530728.1
anaerobic sulfatase maturase [Bacteroides fragilis]	765	765	99.00%	0	85.00%	WP 005821013.1
anaerobic sulfatase maturase [Bacteroides fragilis]	764	764	99.00%	0	85.00%	WP 005780285.1
anaerobic sulfatase maturase [Bacteroides oleiciplenus]	751	751	97.00%	0	85.00%	WP 009131239.1
anaerobic sulfatase maturase [Bacteroides intestinalis]	749	749	97.00%	0	85.00%	WP 007664013.1
anaerobic sulfatase maturase [Candidatus Bacteroides timonensis]	747	747	97.00%	0	85.00%	WP 044264327.1
anaerobic sulfatase maturase [Bacteroides cellulosilyticus]	746	746	97.00%	0	85.00%	WP 029428463.1
MULTISPECIES: anaerobic sulfatase maturase [Bacteroides]	746	746	97.00%	0	85.00%	WP 007214474.1
anaerobic sulfatase maturase [Bacteroides cellulosilyticus]	745	745	97.00%	0	85.00%	WP 007217653.1
anaerobic sulfatase maturase [Bacteroides reticulotermitis]	753	753	99.00%	0	84.00%	WP 044161034.1
anaerobic sulfatase maturase [Bacteroides eggerthii]	747	747	98.00%	0	84.00%	WP 004292631.1
anaerobic sulfatase maturase [Bacteroides eggerthii]	745	745	98.00%	0	84.00%	WP 004290378.1
anaerobic sulfatase maturase [Bacteroides gallinarum]	745	745	98.00%	0	84.00%	WP 018668146.1
anaerobic sulfatase maturase [Bacteroides helcogenes]	744	744	97.00%	0	84.00%	WP 013548456.1
anaerobic sulfatase maturase [Bacteroides stercoris]	744	744	98.00%	0	84.00%	WP 005654844.1
anaerobic sulfatase maturase [Bacteroides stercoris]	744	744	98.00%	0	84.00%	WP 016661344.1
anaerobic sulfatase maturase [Bacteroides plebeius]	734	734	97.00%	0	84.00%	WP 007559240.1
MULTISPECIES: anaerobic sulfatase maturase [Bacteria]	745	745	98.00%	0	83.00%	WP 005835998.1
anaerobic sulfatase maturase [Bacteroides uniformis]	743	743	98.00%	0	83.00%	WP 057088086.1
anaerobic sulfatase maturase [Bacteroides fluxus]	742	742	98.00%	0	83.00%	WP 009124014.1
anaerobic sulfatase maturase [Bacteroides clarus]	742	742	98.00%	0	83.00%	WP 009120536.1
MULTISPECIES: anaerobic sulfatase maturase [Bacteroides]	741	741	98.00%	0	83.00%	WP 005826708.1
anaerobic sulfatase maturase [Bacteroides uniformis]	740	740	98.00%	0	83.00%	WP 044467894.1
anaerobic sulfatase maturase [Bacteroides uniformis]	736	736	98.00%	0	83.00%	WP 016273382.1
anaerobic sulfatase maturase [Bacteroides uniformis]	734	734	98.00%	0	82.00%	WP 035448004.1
anaerobic sulfatase maturase [Bacteroides coprosuis]	720	720	98.00%	0	81.00%	WP 006745530.1
anaerobic sulfatase maturase [Bacteroides plebeius]	692	692	96.00%	0	81.00%	WP 007558660.1
anaerobic sulfatase maturase [Bacteroides coprophilus]	711	711	97.00%	0	80.00%	WP 008140154.1
anaerobic sulfatase maturase [Bacteroides massiliensis]	704	704	98.00%	0	80.00%	WP 005941469.1
anaerobic sulfatase maturase [Bacteroides propionicifaciens]	723	723	98.00%	0	79.00%	WP 018108809.1
anaerobic sulfatase maturase [Bacteroides coprocola]	702	702	97.00%	0	78.00%	WP 007570292.1
anaerobic sulfatase maturase [Bacteroides barnesiae]	698	698	98.00%	0	77.00%	WP 018709694.1
MULTISPECIES: anaerobic sulfatase maturase [Bacteroides]	689	689	98.00%	0	77.00%	WP 007833026.1
anaerobic sulfatase maturase [Bacteroides vulgatus]	687	687	98.00%	0	77.00%	WP 005850852.1
anaerobic sulfatase maturase [Bacteroides vulgatus]	687	687	98.00%	0	77.00%	WP 005840257.1
anaerobic sulfatase maturase [Bacteroides vulgatus]	686	686	98.00%	0	77.00%	WP 032953086.1
anaerobic sulfatase maturase [Bacteroides vulgatus]	686	686	98.00%	0	77.00%	WP 016271815.1
MULTISPECIES: anaerobic sulfatase maturase [Bacteroides]	684	684	98.00%	0	77.00%	WP 008667464.1
MULTISPECIES: anaerobic sulfatase maturase [Bacteroides]	682	682	98.00%	0	77.00%	WP 016275423.1
anaerobic sulfatase maturase [Bacteroides uniformis]	678	678	97.00%	0	76.00%	WP 057253591.1
MULTISPECIES: anaerobic sulfatase maturase [Bacteroides]	670	670	98.00%	0	75.00%	WP 005829655.1
anaerobic sulfatase maturase [Prevotella pleuritidis]	675	675	97.00%	0	74.00%	WP 021584912.1
anaerobic sulfatase maturase [Prevotella pleuritidis]	672	672	97.00%	0	74.00%	WP 024991366.1
anaerobic sulfatase-maturase [Parabacteroides goldsteinii]	657	657	98.00%	0	74.00%	WP 048315582.1
MULTISPECIES: anaerobic sulfatase maturase [Parabacteroides]	656	656	98.00%	0	74.00%	WP 028729461.1
MULTISPECIES: anaerobic sulfatase maturase [Parabacteroides]	655	655	98.00%	0	74.00%	WP 010803049.1
anaerobic sulfatase maturase [Parabacteroides goldsteinii]	655	655	98.00%	0	74.00%	WP 007656924.1
anaerobic sulfatase maturase [Parabacteroides goldsteinii]	655	655	98.00%	0	74.00%	WP 046147140.1
anaerobic sulfatase maturase [Dysgonomonas capnocytophagoides]	656	656	96.00%	0	73.00%	WP 026626529.1
anaerobic sulfatase maturase [Parabacteroides johnsonii]	656	656	98.00%	0	73.00%	WP 008149604.1
anaerobic sulfatase maturase [Parabacteroides merdae]	649	649	97.00%	0	73.00%	WP 005649385.1
anaerobic sulfatase maturase [Parabacteroides merdae]	648	648	97.00%	0	73.00%	WP 005643440.1
MULTISPECIES: anaerobic sulfatase maturase [Bacteroidales]	645	645	96.00%	0	73.00%	WP 005862697.1
anaerobic sulfatase maturase [Prevotella enoeca]	657	657	97.00%	0	72.00%	WP 036888348.1
anaerobic sulfatase maturase [Prevotella bergensis]	657	657	98.00%	0	72.00%	WP 044123378.1
anaerobic sulfatase maturase [Barnesiella viscericola]	656	656	98.00%	0	72.00%	WP 025277267.1
anaerobic sulfatase maturase [Parabacteroides johnsonii]	654	654	99.00%	0	72.00%	WP 008157651.1
anaerobic sulfatase maturase [Bacteroides salanitronis]	651	651	97.00%	0	72.00%	WP 013617485.1
anaerobic sulfatase maturase [Barnesiella intestinihominis]	647	647	97.00%	0	72.00%	WP 008862184.1
anaerobic sulfatase maturase [Bacteroides sp. 3`1`19]	645	645	96.00%	0	72.00%	WP 008779794.1
anaerobic sulfatase maturase [Parabacteroides distasonis]	644	644	96.00%	0	72.00%	WP 057327522.1
MULTISPECIES: anaerobic sulfatase maturase [Parabacteroides]	644	644	96.00%	0	72.00%	WP 005857302.1
anaerobic sulfatase maturase [Parabacteroides distasonis]	642	642	96.00%	0	72.00%	WP 036611496.1
anaerobic sulfatase maturase [Parabacteroides distasonis]	642	642	96.00%	0	72.00%	WP 011966246.1
anaerobic sulfatase maturase [Bacteroides paurosaccharolyticus]	650	650	98.00%	0	71.00%	WP 024993888.1
anaerobic sulfatase maturase [Prevotella buccalis]	644	644	97.00%	0	70.00%	WP 036873332.1
anaerobic sulfatase maturase [Prevotella buccalis]	642	642	97.00%	0	70.00%	WP 004350830.1
MULTISPECIES: anaerobic sulfatase maturase [Prevotella]	639	639	97.00%	0	70.00%	WP 023056581.1

Appendix E

Supplementary information for Chapter 6

E.1 Images

Arabinose induction

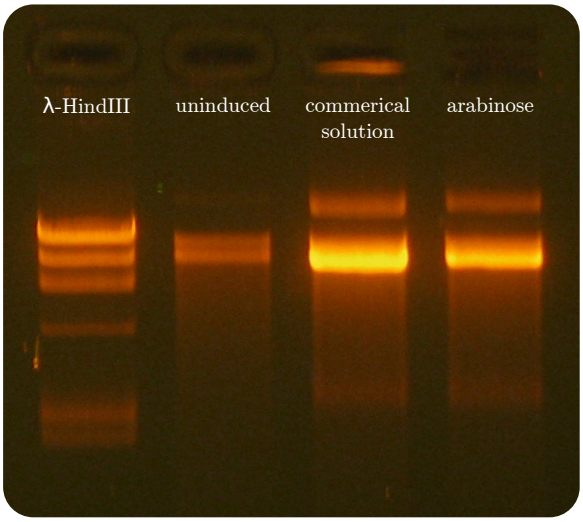


Figure E.1: Agarose gel of miniprepped DNA following induction using arabinose versus commerical solution. Plasmid minipreps of pKL13 were compared from three cultures: an uninduced negative control, induction using 1× autoinduction solution (Epicentre), or induction using 0.2% arabinose.

Confirmation of pKL17

Six putative pKL17 clones were screened using two restriction digests; see Figure 6.7 for construct diagrams.

1. SfaAI-SgsI double digest, to check fragment still present (Figure E.2A)
    - Expected for pKL17: 1470 bp
2. MssI-XhoI double digest, for orientation of stuffer (Figure E.2B)
    - Expected for pKL13: 1300 bp (control)
    - Expected for pKL17: 745 bp

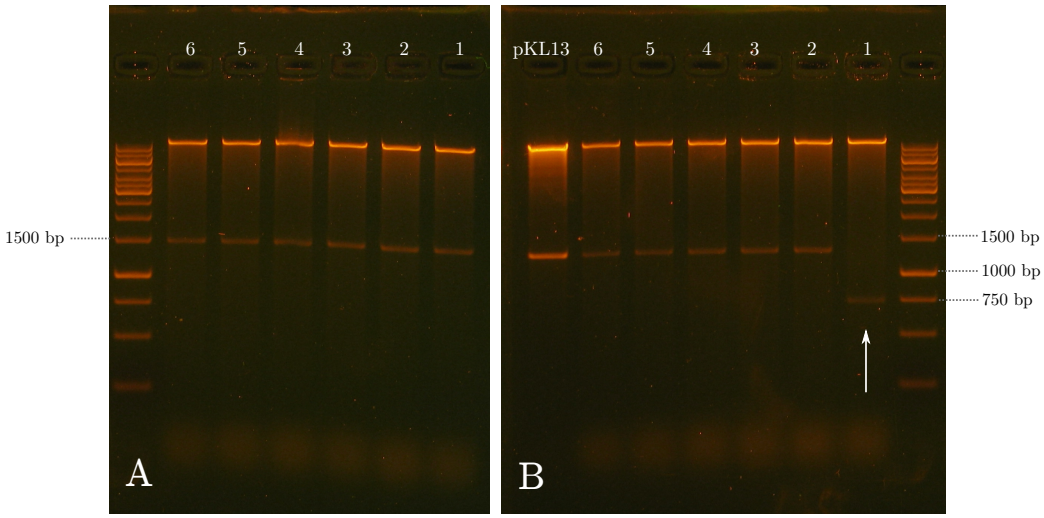


Figure E.2: Agarose gel of putative pKL17 clones. Putative clones of pKL17 were digested with: (A) SfaAI and SgsI, and (B) MssI and XhoI for orientation; white arrow indicates clone with desired restriction pattern.

Confirmation of pKL16

Six putative pKL16 clones were screened using two restriction digests; see Figure 6.7 for construct diagrams.

1. NheI and PacI single digests, along with uncut control (Figure E.3A)
    - Expected for pKL18: both PacI and NheI cut once (positive control)
    - Expected for pKL16: neither PacI nor NheI will cut
2. Eco72I-SgsI double digest (Figure E.3B)
    - Expected for pKL15: 1155 bp, 1021 bp (control)
    - Expected for pKL16: 1155 bp, 886 bp

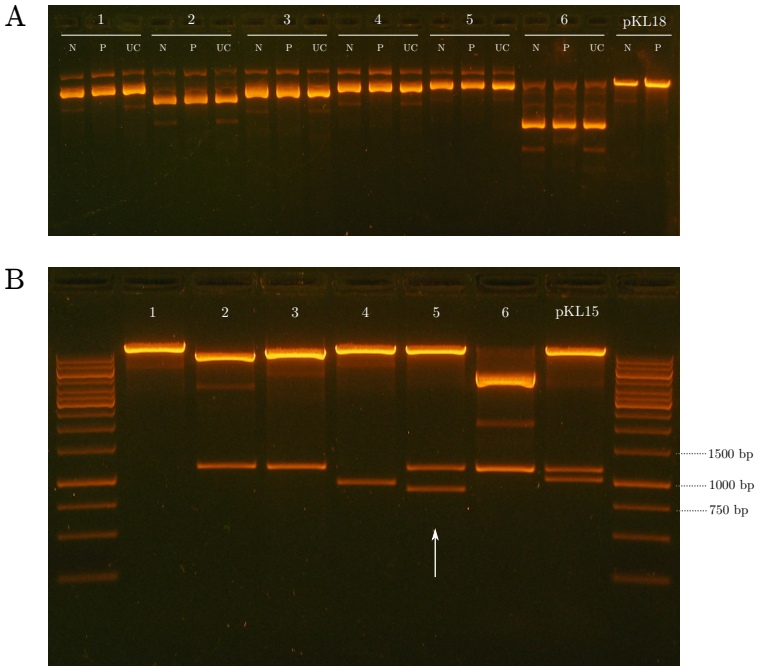


Figure E.3: Agarose gel of putative pKL16 clones. Putative clones of pKL16 were digested with: **(A)** NheI (N) and PacI (P) individually to check for loss of sites, using uncut DNA as a control (UC); **(B)** and Eco72I and SgsI doubly to confirm deletion of the terminator; white arrow indicates clone with desired restriction pattern.

Confirmation of pKL19

Six putative pKl19 clones were screened using two restriction digests; see Figure 6.7 for construct diagrams.

1. NsiI and CpoI single digests, along with uncut control (Figure E.4A)
    - Expected for pKL18: both NsiI and CpoI cut once (positive control)
    - Expected for pKL19: neither NsiI nor CpoI will cut
2. Eco72I-SfaAI double digest (Figure E.4B)
    - Expected for pKL18: 1155 bp, 1012 bp (control)
    - Expected for pKL19: 1155 bp, 885 bp

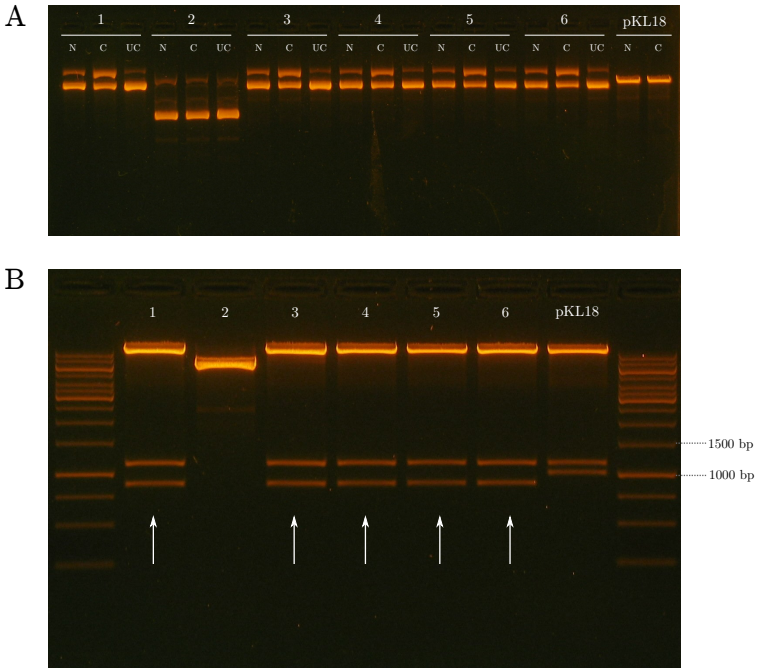


Figure E.4: Agarose gel of putative pKL19 clones. Putative clones of pKL19 were digested with: **(A)** NsiI (N) and CpoI (C) individually to check for loss of sites, using uncut DNA as a control (UC); **(B)** and Eco72I and SfaAI doubly to confirm deletion of the terminator; white arrow indicates clones with desired restriction pattern.

E.2 Sequence data

TT fragment cloned in pKL13

Note that the sequence of the actual TT fragment in pKL13 differs from the designed sequence by one base, due to a point mutation in the synthesis of the fragment.

```
>TT fragment cloned in pKL13
AAATGGCGCGCCGGCTGGATTTAATTAATGTCTGCTCCTCGGTTATGTTTTTAAGGTCAAAAAAAACCCCCGGACCT
TTCGGTGCGGGGGTCTTAGTTCGTTAAGGCTTGATCTCTAGCGATTAAGTTGGGTAACGCCAGGGTTTTCGTCACTT
AGTCAGCTAGCCACGTGCCTTAGGGTGTGAAATTGTTATCCGCTCACAATTCCACACATTATACGAGCCGATGATTA
ATTGTCAACAGCTCCCTGAGGTTCGAAGATCCTCCGGCTCACGGTAACTGATGCCGTATTTGCAGTACCAGCGTACG
GCCCACAGAATGATGTCACGCTGAAAATGCCGGCCTTTGAATGGGTTCATGTGCAGCTCCATCAGCAAAAGGGGATG
ATAAGTTTATCACCACCGACTATTTGCAACAGTGCCGTTGATCGTGCTATGATCGACTGATGTCATCAGCGGTGGAG
TGCAATGTCGTGCAATACGAATGGCGAAAAGCCGAGCTCATCGGTCAGCTTCTCAACCTTGGGGTTACCCCCGGCGG
TGTGCTGCTGGTCCACAGCTCCTTCCGTAGCGTCCGGCCCCTCGAAGATGGGCCACTTGGACTGATCGAGGCCCTGC
GTGCTGCGCTGGGTCCGGGAGGGACGCTCGTCATGCCCTCGTGGTCAGGTCTGGACGACGAGCCGTTCGATCCTGCC
ACGTCGCCCGTTACACCGGACCTTGGAGTTGTCTCTGACACATTCTGGCGCCTGCCAAATGTAAAGCGCAGCGCCCA
TCCATTTGCCTTTGCGGCAGCGGGGCCACAGGCAGAGCAGATCATCTCTGATCCATTGCCCCTGCCACCTCACTCGC
CTGCAAGCCCGGTCGCCCGTGTCCATGAACTCGATGGGCAGGTACTTCTCCTCGGCGTGGGACACGATGCCAACACG
ACGCTGCATCTTGCCGAGTTGATGGCAAAGGTTCCCTATGGGGTGCCGAGACACTGCACCATTCTTCAGGATGGCAA
GTTGGTACGCGTCGATTATCTCGAGAATGACCACTGCTGTGAGCGCTTTGCCTTGGCGGACAGGTGGCTCAAGGAGA
AGAGCCTTCAGAAGGAAGGTCCAGTCGGTCATGCCTTTGCTCGGTTGATCCGCTCCCGCGACATTGTGGCGACAGCC
CTGGGTCAACTGGGCCGAGATCCGTTGATCTTCCTGCATCCGCCAGAGGCGGGATGCGAAGAATGCGATGCCGCTCG
CCAGTCGATTGGCTGAGCTCATGAGCGGAGAACGAGATGACGTTGGAGGGGCAAGGTCGCGCTGATTGCTGGGGCAA
CACGTTCGAACACGTGATGCATTAACTAGGTGACGTCATAGCTGTTTCCTGTGTGAAATTGTTATCGGTCAGTTTCA
CCTGATTTACGTAAAAACCCGCTTCGGCGGGTTTTTGCTTTTGGAGGGGCAGAAAGATGAATGACTGTCCGGTCCGA
GCAGGTCGCGATCGCATTT
```