

Next Gen Sequencing Tools to Derive Insights into
Protein Expression and Gene Function

by

Sena Bae

Department of Biomedical Engineering
Duke University

Date: _____

Approved:

Raphael H. Valdivia, Supervisor

Jingdong Tian

John F. Rawls

Brenton D. Hoffman, Chair

Charles A. Gersbach

Joseph Lucas

Dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in the Department of
Biomedical Engineering in the Graduate School of
Duke University

2017

ABSTRACT

Next Gen Sequencing Tools to Derive Insights into
Protein Expression and Gene Function

by

Sena Bae

Department of [Department]
Duke University

Date: _____

Approved:

Raphael H. Valdivia, Supervisor

Jingdong Tian

John F. Rawls

Brenton D. Hoffman, Chair

Charles A. Gersbach

Joseph Lucas

An abstract of a dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in the Department of
Biomedical Engineering in the Graduate School of
Duke University

2017

Copyright by
Sena Bae
2017

Abstract

Human physiology is heavily influenced by the colonization of microbes in the gastrointestinal tract. A major roadblock to understanding this process is our inability to genetically manipulate new bacterial species and experimentally assess the function of their genes. In order to map bacterial genes, we describe an application of chemical mutagenesis followed by population-based genomic sequencing. We chose to map genes responsible for motility in *Exiguobacterium acetylicum*, a representative intestinal Firmicutes bacterium that is intractable to molecular genetic manipulation. We derived strong associations between mutations in 57 *E. acetylicum* genes and impaired motility and also discovered new motility genes that were previously uncharacterized. We confirmed the genetic link between individual mutations and loss of motility for several of these genes by performing a large-scale analysis of spontaneous suppressor mutations. Furthermore, we generated isogenic strains that allowed us to establish that *Exiguobacterium* motility is important for the colonization of its vertebrate host.

This methodological advance in gene functional analysis of genetically intractable microbes has enabled us to identify 902 essential genes that are directly responsible for growth and survival. This is achieved by large-scale mutant sequencing analysis.

By curating the gene list, we assigned the essentiality of genes to uncharacterized genes as well. These results indicate that the genetic dissection of a complex trait, functional annotation of new genes, and the generation of mutant strains can all be accomplished in bacteria without the development of species-specific molecular genetic tools. Ultimately, this advance helps define the role of genes in complex environments.

To investigate the effect of silent mutations in a gene, we have designed and created thousands of AcGFP codon-variant libraries to determine the relationship between codon usage and protein expression. mRNA structures near the initial start codon regions are prominent factor for determining protein expression level, but variation in sequence beyond the start codon region also importantly modulates expression levels.

Contents

Abstract	iv
List of Tables	x
List of Figures	xi
Acknowledgements	xiv
1. Introduction	1
1.1 Overview	1
1.2 The role of the microbiome in vertebrate health.....	1
1.2.1 Firmicutes bacteria	4
1.3 The experimental challenges of performing genetic studies in gut microbes	6
1.4 Regulation of protein expression genetic manipulation.....	7
1.4.1 Redundancy of the genetic code	7
1.4.2 Codon usage and its impact on protein expression	8
2. Genomic-sequencing mutational enrichment analysis to identify motility genes in a genetically intractable microbe	11
2.1 Introduction.....	11
2.2 Material and Methods.....	13
2.2.1 De Novo assembly and Gene annotation of <i>E. acetylicum</i>	13
2.2.2 Prediction of Motility Genes.....	14
2.2.3 Chemical Mutagenesis experiment.....	14
2.2.4 Screening for Nonmotile <i>Exiguobacterium</i> Mutants.....	15
2.2.5 Genomic sequencing of <i>E. acetylicum</i> Mutants.....	15

2.2.6 Electron Microscopy Analysis of Flagellar	16
2.2.7 Isolation of Spontaneous Suppressor of Nonmotile <i>E. acetylicum</i> strains	16
2.2.8 Computational modeling	17
2.2.9 Gnotobiotic zebrafish colonization	18
2.2.10 Mutational Enrichment analysis	19
2.3 Results	22
2.3.1 Bioinformatic and Experimental approach of MEAPS	22
2.3.2 WGS analysis and mutational enrichment genes	28
2.3.3 Protein Functional verification of motile genes by spontaneous suppressor mutation.....	38
2.3.3 Discovery of <i>E. acetylicum</i> specific motility genes	45
2.4 Discussion.....	48
3. Genomic-sequencing mutational analysis to identify essential genes in a genetically intractable microbe	50
3.1 Introduction.....	50
3.2 Material and Methods.....	51
3.3 Results	51
3.3.1 Computational models to estimate the number of mutations needed to identify essential genes in chemically mutagenized <i>E. acetylicum</i>	51
3.3.2 Sequencing strategy to maximize the number of mutations identified in complex mutant pools.	56
3.3.3 A strategy to reduce the identification of false mutations.....	59
3.3.4 Approaches to identify putative <i>E. acetylicum</i> essential genes and conservation of these genes among <i>Exiguobacterium</i> sp.	60

3.3.5 Efforts at increasing the discovery rate of essential genes in the <i>E. acetylicum</i> .	66
3.3.6 False positive essential genes removal and Effects of mutation on protein function	70
3.3.7 Curation of essential genes by using information of the mutational impact of nonsynonymous amino acid substitutions	71
3.4 Discussion.....	73
4. Implementation of a high-throughput sequence analysis framework to design synthetic genes	74
4.1 Introduction.....	74
4.2 Material and Methods.....	74
4.2.1 Oligonucleotides synthesis to introduce random variations of synonymous mutations	74
4.2.2 Fragmented oligo assembly using the PCA method	75
4.2.3 Plasmid library construction using the CPEC method	76
4.2.4 High-throughput AcGFP protein screening using FACS.....	77
4.2.5 Quantitative real-time PCR (Q-PCR).....	78
4.3 Results	78
4.3.1 Combinatorial gene library construction and the codon usage calculation matrix	78
4.3.2 High-throughput screening and NGS data analysis approach	80
4.3.3 Regional effect of AcGFP protein expression level.....	87
5. Future directions	91
5.1 Development of a statistical framework to identify essential genes based on chemical mutagenesis	91

5.2 Establish a sequence analysis method to identify mutations that are depleted after purifying selection.....	92
Appendix A.....	93
Appendix B.....	97
Appendix C.....	102
Appendix D.....	107
Appendix E.....	108
Appendix F.....	109
References.....	111
Biography.....	124

List of Tables

Table 1: Category Subcategory Number of gene.....	67
Table 2: Distribution of the candidate essential gene production varied by number of GC bases/gene	70
Table 3: rpsQ nonsynonymous mutations analysis and its predicted functional effect ...	71

List of Figures

Figure 1: The genetic code and its redundancy	8
Figure 2: Schematic of the approach coupling chemical mutagenesis and DNA sequencing to perform a mutational enrichment analysis after phenotypic selection (MEAPS) and identify genes required for swarming motility	23
Figure 3: Transmission electron micrograph of <i>E. acetylicum</i> flagella (Bar: 0.5uM).....	24
Figure 4: Simulations of the rate of identification of motility genes based on the sequencing of non-motile mutants.	25
Figure 5: Isolation of non-motile <i>E. acetylicum</i> mutants.....	27
Figure 6: MEAPS of <i>E. acetylicum</i> strains defective for motility.	29
Figure 7: Comparative analysis of the distribution of synonymous mutations among <i>E. acetylicum</i> mutants.....	30
Figure 8: Genetic map of ORFs in motility Region I.	31
Figure 9: Cartoon schematic of the Gram positive flagellar apparatus displaying all components conserved in <i>E. acetylicum</i>	33
Figure 10: Overlapping set of putative <i>E. acetylicum</i> motility genes identified either by reciprocal BLAST homology searches, Pfam terms associated motility, and MEAPS.....	34
Figure 11: Comparative analysis of mutations in noncoding regions.....	35
Figure 12: Suppressors of nonsense mutations in putative structural flagellar genes confirm their role in motility.	36
Figure 13: Intragenic suppressor of a <i>hag1</i> nonsense allele restores the formation of wild type flagellar structures in <i>E. acetylicum</i>	37
Figure 14: Motility enhances <i>E. acetylicum</i> colonization of germ free zebrafish.....	39
Figure 15: Motility contributes to the efficient colonization of zebrafish	40
Figure 16: Motility Region II of <i>E. acetylicum</i> encodes for new motility genes.....	41

Figure 17: GGDEF/EAL domain proteins in <i>E. acetylicum</i> . Only genes with more than 3 total non-synonymous mutations are shown.	43
Figure 18: Motility behavior and flagellar assembly of a FtsXQ82* nonsense <i>E. acetylicum</i> mutant and a spontaneous variant that regained motility (FtsXQ82W).....	44
Figure 19: Extragenic suppression analysis of non-motile <i>E. acetylicum</i> mutants identifies a role for cell wall modifications and c-di-GMP sensing in commensal Firmicutes motility.....	45
Figure 20: Schematic of suppressor mutations linking Ea2619 and Ea2862 to the regulation of swimming speed and direction.	47
Figure 21: Comparative analysis of the distribution of dN/dS ratio by different mutation load.....	54
Figure 22: Simulation of the rate identification of essential genes based on the sequencing of mutants with different mutational loads	55
Figure 23: Comparative analysis of the distribution of synonymous and nonsynonymous mutation across <i>E. acetylicum</i> genomes of 8500 mutants	61
Figure 24: The phylogenetic reconstruction of <i>Exiguobacterium</i> based on a concatenated matrix of 32 genes and <i>B. subtilis</i> genes	63
Figure 25: Comparison of number of GC bases per gene between all <i>E. acetylicum</i> genes and <i>B. subtilis</i> essential genes in <i>E. acetylicum</i> genome (upper panel). Comparison of number of GC bases per gene between candidate essential genes of <i>E. acetylicum</i> and <i>B. subtilis</i>	65
Figure 26: Comparison of number of GC bases per gene between candidate essential genes of <i>E. acetylicum</i> , <i>B. subtilis</i> essential genes in <i>E. acetylicum</i> genome and experimentally verified essential genes in <i>B. subtilis</i>	68
Figure 27: Overlapping set of putative <i>E. acetylicum</i> essential genes identified based on their homology to genes conserved among all <i>Exiguobacterium</i> , experimentally verified <i>B. subtilis</i> and candidate essential genes in <i>E. acetylicum</i> by analysis of mutational load	68
Figure 28: Combinatorial gene library cloning steps	79
Figure 29: PacBioRS sequence data filtering for high and low expression groups	83

Figure 30: Sliding window analysis of CAI, GC3, mRNA folding energy and tAI of high and low AcGFP expressed population.	86
Figure 31: Quantification of translation levels of the top five most abundant sequence reads from the high, low and WT AcGFP expression groups.	87
Figure 32: AcGFP protein expression levels of sequences with swapped regions between the low, high and WT AcGFP expression groups.	89

Acknowledgements

I thank to my advisor Dr. Raphael Valdivia and Dr. Jingdong Tian for guidance and help throughout my Ph.D years and am truly honored to complete my doctorate degree as Dr. Valdivia and Dr. Tian's student. With my advisor's supports and mentorship, I enjoyed research and could grow up as a scientist. All my research work could have better outcome and have more recognitions from people with their efforts as well. Again, I would like to express my deepest respect and gratitude to Dr. Valdivia and Dr. Tian.

Also, I would like to thank Dr. John Rawls for helpful suggestions, critical and mentorship. I have learned a lot collaborating research works with Dr. Rawls and his group. His great attitude as a scientist and thoughtful criticism has positively influence on my growth.

I thank you my lab members and friends who have helped and supported me with enormous kindness and supports. I feel truly lucky to study such a supportive environments and surrounded be great people.

Lastly, I thank to my family, especially my parents, Hyangsoon Park and Kijoon Bae. They have always supported me through my entire life and valued on education. I am always grateful to have wonderful parents and be loved by my family.

1. Introduction

1.1 Overview

The goal of this thesis is to develop bioinformatic and experimental approaches to understand the consequences of mutations on protein function. Chapter 1 describes the role of the microbiome in human health and introduces one intestinal microbiota component, a Firmicutes bacteria (Section 1.2), describes the experimental challenges of working with genetically intractable microbes (Section 1.3), and outlines the impact of synonymous mutations on protein expression level (Section 1.4). Chapter 2 presents a forward genetic approach to identify genes that are responsible for any chosen phenotype in the absence of molecular genetic tools. It is presented as a published manuscript titled “Genomic-sequencing mutational enrichment analysis to identify motility genes in a genetically intractable microbe”. Chapter 3 discusses the development of a methodology to identify essential genes in genetically intractable microbes. Chapter 4 establishes the guidelines to synthetic gene design by presenting a comprehensive understanding of the relationship between codon usage and protein expression level. Chapter 5 concludes with a discussion of the impact and significant of these approaches.

1.2 Microbiome vs microbiota

Microbiota refers to the collection of microorganisms in a specific niche, including bacteria, archaea, fungi and eukarya. The microbiome is defined as all of the collective genetic material within a microbiota. Bacteria and archaea can be classified by groups of related individuals based on 16S rRNA sequence identity. These sequences are highly conserved in different species of bacteria and archaea. For bacteria classified by 16S rRNA, a percent identity greater than 97% identifies an Operation Taxonomic Unit (OTU) and a percent identity greater than 99% identifies microorganisms of a species, though it can vary (1). In the human body, more than 10,000 different microbe species have identified. Individual humans share 99.9% sequence identity in terms of genome level. However, 80-90% of individuals are different from each other in terms of microbiome content (2). Even identical twin shares only 50% of their species level bacterial phylotypes (3). Characterization of the ecology of host associated microbial communities and distinctive microbial communities in individual would enhance our understanding of host-microbe interactions and its impact on health.

1.3 The role of the microbiome in vertebrate health

Hundreds of trillions of microbes inhabit and form complex communities in the vertebrate intestine (4). Over 1000 gut bacterial species (5) and 3.3 million non-redundant genes have been characterized (6) by analysis of fecal samples. Human

intestinal samples comprise of seven divisions of Bacteria including Firmicutes, Bacteroidetes, Actinobacteria, Fusobacteria, Proteobacteria, Verrucomicrobia, and Cyanobacteria. Firmicutes and Bacteroidetes constitute about 98% of all 16S rRNA sequences in mammals (7). The human gut microbiome has a relatively low number of bacterial divisions compared to other microbial habitats, but are diverse in lower phylogenetic levels (8). These microbial communities are increasingly recognized as environmental factors that influence host metabolism (9), nutrient absorption (10) and immune response (11). The composition of a microbial community in the gastrointestinal tract varies throughout different vertebrates and among individuals of the same species. Many factors constantly reshape microbial composition such as the health status of the host, environmental changes, and diet (12-15). The relative compositional changes in gut microbiome composition have been associated with diseases such as allergies (16-18), Celiac's disease (19), Gastric cancer (20), autism (21), obesity (22-24), Anorexia (25, 26), Crohn's disease (27) and Type II diabetes (28). An improved understanding of the function and structure of these complex microbial community would enhance human and animal health. Studying the interactions between host and their microbiota opens new avenues for understanding host health, diagnosis and treatment of diseases (29-32). The structure and role of microbiota are generally characterized by the abundance ratio of microbial communities. However, functional studies are largely lacking as to what

role individual genes of components of the microbe play in regulating the function of the community.

1.3.1 Firmicutes bacteria

Firmicutes bacteria is one of two major phyla that populate the vertebrate intestine. The proportion of Firmicutes significantly differs based on an individual's health conditions and nutrient intake. The composition of Firmicutes have dynamically evolved throughout life times (33), individual's health status independent from diet (22-24, 34). Also, the abundance of Firmicutes can vary in response to diet intake in humans (26, 35), mice (10) and zebrafish (15, 36). The diet-dependent population change of Firmicutes plays an important role in energy balance and nutrient absorption in the host (22, 26) . However, the genetic contribution of Firmicutes to nutrient absorption is poorly understood. One study uncovered the role of a specific Firmicutes isolated strain, *Exiguobacterium* strain ZWU0009, in promoting lipid absorption in the zebrafish gut in a diet-dependent manner (36). By taking advantage of the optical transparency of the zebrafish, *Exiguobacterium* strain ZWU0009 was shown to enrich and promote fatty acid uptake in the zebrafish gut in the presence of a high caloric diet (36).

1.3.2 Intestinal microbiome colonization factors

Composition and diversity of gut microbiome flexibly reshape bacterial communities, so an understanding of these colonization factors is important. Gut microbiome colonization requires adequate nutrient resources, chemical signals, and competition with other residing microbes. Several factors have been studied as prominent determinants of colonization including motility, chemotaxis, quorum sensing and commensal colonization factors. *Aeromonas veronii* Hm21 and *Vibrio* sp. strain ZWU0020 transposon mutants with defects in motility and chemotaxis showed reduction of colonization in zebrafish intestine (37). Motility enables microbes to move across a surface by the rotation of flagella, extension of pili, or facilitation of surfactants. The enterohemorrhagic gastrointestinal pathogen, *Escherichia coli*, unsuccessfully colonized in a gut due to disruption in the fructose sensing locus which modulates its pathogenicity (38). The colonization level of *Bacteroides fragilis* in the mice intestine is regulated by a genetic locus for commensal colonization factors which are conserved among intestinal *Bacteroides* (39). Since *B. fragilis* reaches saturated colonization, it effectively prevents others of the same species from colonizing the gut, without affecting colonization of relative species. Lastly, quorum sensing, a bacterial system for coordinating gene expression based on local bacterial population density, could be involved in the modulation of *E. coli* colonization in the intestine (40). From these factors,

it is clear that gut colonization by microbes is not only influenced by the genetics of individual cells, but also by species-species interactions.

1.4 The experimental challenges of performing genetic studies in gut microbes

A limited number of “model” bacterial species which are amenable to molecular genetic manipulation have framed the bulk of our understanding of microbial biology. This is because in these bacteria one can readily disrupt genes and monitor the phenotypic consequences of these alterations. Insertional inactivation (41) and allelic replacement by homologous recombination (42) are popularly used tools to generate functional genetic study. Development of Next Generation DNA sequencing technology and bioinformatics combined with transposon mutagenesis lead to development of new tools to identify the function of genes in a high-throughput manner including transposon sequencing (Tn-seq) (43), high-throughput insertion tracking by deep sequencing (HITS) (44), insertion sequencing (INSeq)(45) and transposon-directed insertion site sequencing (TraDIS) (46).

However, most microbes are not amenable to routine molecular genetic manipulation, leaving very few options to experimentally determine the function of their genes. Nevertheless, it is possible to derive phenotype- genotype associations without molecular genetic tools, as has been shown in the obligate intracellular pathogen *Chlamydia trachomatis* (47), *E. coli* (48) and T7-like virus of *Vibriio cholerae* (49),

where whole genome sequencing (WGS) was used to identify chemically induced genetic variants. This approach has not been tested in Firmicutes bacteria or other gut microbes. Rapidly expanding knowledge of the gut microbiome and its contribution to host health requires the development of experimental methods to define gene function independently of molecular genetic tools.

1.5 Regulation of protein expression genetic manipulation

1.5.1 Redundancy of the genetic code

The genetic code uses 61 possible codons to code for 20 amino acids, with each amino acid encoded by an average of three possible codons. The DNA sequence of codons that encode the same amino acids generally differ by only one base. However, the codon sequence of three amino acids, Arginine, Leucine and Serine, are not similar to each other because these codons are coded by six independent codons (Figure 1).

		2nd base				
		U	C	A	G	
5' base	U	Phe	Ser	Tyr	Cys	U C A G
		Phe	Ser	Tyr	Cys	
		Leu	Ser	Stop	Stop	
		Leu	Ser	Stop	Trp	
	C	Leu	Pro	His	Arg	U C A G
		Leu	Pro	His	Arg	
		Leu	Pro	Gln	Arg	
		Leu	Pro	Gln	Arg	
	A	Ile	Thr	Asn	Ser	U C A G
		Ile	Thr	Asn	Ser	
		Ile	Thr	Lys	Arg	
		Met	Thr	Lys	Arg	
	G	Val	Ala	Asp	Gly	U C A G
		Val	Ala	Asp	Gly	
		Val	Ala	Glu	Gly	
		Val	Ala	Glu	Gly	

Figure 1: The genetic code and its redundancy

Due to this redundancy of the genetic code, a protein can be coded by many different permutations of codons. On average, a gene can have 3^n different DNA sequence that encode the same protein. For example, 100 amino acids can be encoded by roughly 3^{100} different sequences. The 3^{100} different sequences encode for the same protein and has no effect on final protein product, but difference in mRNA sequence can alter the transcript's secondary structure or the speed at which the ribosomes translate mRNAs.

1.5.2 Codon usage and its impact on protein expression

Many studies have highlighted the difference in the preferred codon usages between highly and lowly expressed genes within the same organism (50-54) . This

difference in codon usage implies that synthetic protein expression levels could be modulated by the choice of codon. Also, because different species differ in the codon usage for their highly expressed genes (55); a bacterial cells are often unable efficiently express highly expressed human. Additionally, an increasing number of synonymous mutations have been highly correlated with genetic diseases (56). This suggests that the accumulated effect of synonymous mutations suppresses or enhances protein expression levels to influence cellular processes.

Protein expression comprises multiple processes including DNA transcription, RNA translation and post-translational processing. Synonymous mutations in a gene can alter mRNA structure and change usage of available tRNA; factors that influence protein translational speed (57, 58). Indeed, several studies have shown a profound impact on protein expression level through the use of alternative codons (59-61). Nevertheless, these DNA sequence changes are not the only factor that modulates protein expression. Environmental conditions, regulatory elements (61) , ribosomal binding sites (62, 63) and gene copy number are known to contribute to protein expression level.

Codon usage was known as significantly affects protein expression level (64). Codon usage bias is indicated by codon usage frequency table and it can be used to score genes with the Codon Adaptation Index (55) . However, Kudla et al. showed that protein expression level is highly associated with 5' mRNA secondary structure rather

than codon usage (52). In contrast, Supek et al. concluded that codon bias is still a significant influence on protein expression level, particularly if 5' mRNA secondary has weak structure (65). These research efforts have deepened our knowledge of synonymous codon usage and its impact on protein translation level.

2. Genomic-sequencing mutational enrichment analysis to identify motility genes in a genetically intractable microbe

Sena Bae^{1,2}, Olaf Mueller¹, Sandi Wong¹, John F. Rawls¹ and Raphael H. Valdivia^{1*}

¹Center for the Genomics of Microbial Systems and Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, NC 27710

²Department of Biomedical Engineering, Duke University, Durham NC 27708.

Published in PNAS on November, 2016

2.1 Introduction

The advent of DNA sequence-based approaches to survey microbial environments has led to a deepened appreciation for the diversity, ubiquity, and functions of microbial life. For instance, the gastrointestinal tract of humans and other vertebrates is colonized by complex microbial communities that promote gut development, nutrient metabolism, and immune homeostasis (66). Of particular importance to human health, gut microbes have emerged as major risk determinants for obesity and metabolic disorders in part because of their role in modulating accessibility and absorption of energy-rich dietary nutrients in vertebrates (67). For example, colonization of germ-free zebrafish with *Exiguobacterium* sp. ZWU0009, a Firmicutes bacterium originally isolated from the zebrafish intestine, enhanced the ability of

intestinal enterocytes to absorb dietary fat (36). Unfortunately, the molecular bases for how bacteria like *Exiguobacterium* sp. ZWU0009 colonize the intestine and influence host physiology are poorly understood. Indeed, most microbes are not amenable to genetic manipulation because methods for robust DNA transformation, insertional mutagenesis and *trans* expression of genes are largely lacking. For a select group of microbial species, including members of the *Bacteroides* genus, some strains are amenable to transposon mutagenesis and have been invaluable in helping decipher the requirement of individual genes in gut colonization and nutrient homeostasis (68, 69). However, genetic tools do not exist for the vast majority of intestinal microbes. As a result, the function of individual genes and their contribution to host-microbe and microbe-microbe interactions within the gut often relies on information inferred from homology to genes characterized in phylogenetically unrelated, but genetically tractable bacterial systems. This reliance on previously characterized genes has emerged as a major block in the functional annotation of novel genes emerging from metagenomic studies.

There is broad interest in the role microbial communities play in human health. While DNA sequencing technologies enabled a broad assessment of microbial diversity and genomic content, our understanding of the molecular mechanisms underlying microbe-microbe and microbe-host interactions has proceeded much more slowly because only a small fraction of microbes are amenable to molecular genetic manipulation. We describe a method, independent of recombinant DNA tools, to

perform genetic analysis in any cultivatable microbial species. We identified new determinants of motility in a member of the vertebrate microbiome, the Firmicutes *Exiguobacterium acetylicum*, and experimentally determined a role for motility in animal colonization by this previously uncharacterized commensal bacteria that is important for host nutrient homeostasis.

2.2 Material and Methods

2.2.1 De Novo assembly and Gene annotation of *E. acetylicum*

Exiguobacterium sp. strain ZWU0009 (also referred to as ZF1EB02)(15) was grown in Brain Heart Infusion (BHI) broth (BD Biosciences) overnight at 30°C under aerobic conditions, and genomic DNA was isolated with a DNeasy blood and tissue kit (Qiagen). PacBio RS (Pacific Bioscience Inc.) library preparation and sequencing was performed at the Duke Sequencing and Genomic Technologies Shared Resource. PacBio reads were quality filtered and assembled with PacBio SMRT analysis software, using the HGAP2 protocol. The assembly resulted in circular contigs for the bacterial chromosome and four plasmids, and overlapping ends were trimmed. Subsequent gene annotation was performed with PROKKA(70). A complete, annotated genome and its analysis will be published separately. The Genbank accession number is GCA_000798945.1. Preliminary phylogenetic analysis based on whole genome

sequences indicate that *Exiguobacterium* sp. strain ZWU0009 is a variant of *Exiguobacterium acetylicum* (71).

2.2.2 Prediction of Motility Genes

To identify genes with a potential role in motility, we performed a keyword-search of the NCBI refseq protein database (<http://www.ncbi.nlm.nih.gov/refseq/>) using the terms 'firmicutes' and 'motility'. Redundancies in the retrieved set of 36769 protein sequences were removed by clustering with USEARCH8 (72) (<http://drive5.com/usearch/>), and output centroids were used as a database for reciprocal BLAST queries with predicted *Exiguobacterium* sp. ZWU0009 protein models. This analysis led to the identification of 126 potential motility genes (Appendix A). A separate bioinformatics search based on Pfam domains in the EMBL-EBI database (<http://pfam.xfam.org/>) using the keywords 'motility', 'flagellar' and 'chemotaxis' yielded 709 unique domains, which matched to 653 predicted *E. acetylicum* genes. Approximately 9% of these genes (56/653) were among the 102 putative motility genes identified by MEAPS (Appendix B).

2.2.3 Chemical Mutagenesis experiment

Overnight cultures of *E. acetylicum* were exposed to 2.5-20mg/mL of ethyl methanesulfonate (EMS) or *N*-Ethyl-*N*-Nitrosourea (ENU) (Sigma-Aldrich) in phosphate-buffered saline (PBS) for 1 h. The mutagenized bacteria were rinsed three times in PBS and incubated in BHI broth for 6-8 h to allow recovery from mutagen exposure. Mutagenized bacterial pools were stored in BHI/15% glycerol at -80°C.

2.2.4 Screening for Nonmotile *Exiguobacterium* Mutants

To select for non-motile mutants, approximately 10^7 EMS or ENU-treated *E. acetylicum* strains were streaked on the center of BHI plates made with 0.3% agar. After overnight incubation at 30°C, bacteria at the initial site of inoculation were collected and re-inoculated in the center of another 0.3% agar plate to enrich for non-swarming mutants (73). After three rounds of enrichment, individual bacterial colonies were tested in a 96-well plating assay (74) to identify mutants that failed to swarm in low percentage agar. Transparent wells contained non-motile mutants whose growth was restricted to the site of inoculation, whereas uniformly cloudy wells contained motile mutants that swarmed across the entire well (**Figure 5**). For the 440 non-motile mutant collection, 108 independent chemically mutagenized pools were used.

2.2.5 Genomic sequencing of *E. acetylicum* Mutants

E. acetylicum strains were cultured in BHI broth overnight at 30°C, and genomic DNA isolated with a DNeasy blood and tissue kit (Qiagen). Pools were assembled consisting of 2.5 ng of total DNA isolated from each of 20 individual strains and a pre-sequenced mutant strain. Sequencing libraries were prepared and barcoded with a Nextera DNA library preparation kit (Illumina) or NEBNext ultra DNA library prep kit (New England Biolabsenses), as recommended by the manufacturers. Five barcoded pools totaling 100 *E. acetylicum* strains were sequenced as single 50 base pairs reads on either a HiSeq2000 or HiSeq2500 sequencing platform (Illumina). Duplicated reads from

raw sequence reads were removed using VSEARCH (<https://github.com/torognes/vsearch>). Unique sequence reads derived from the mutagenized strains were mapped against the reference genome with Bowtie2 (75). To identify single nucleotide variants (SNVs) among complex pools of mutants we used SNVer (single nucleotide variant caller) (76). We omitted the Bonferroni method that SNVer uses for multiple-comparison corrections, and instead used a false discovery rate estimation to adjust raw *p*-value outputs from SNVer. The false discovery rate was set to 5%.

2.2.6 Electron Microscopy Analysis of Flagellar

E. acetylicum wild-type, non-motile strains, and its suppressor strains were grown in brain heart infusion (BHI) medium at 30°C. Midlog bacteria were sedimented at 2000 rpm in a microcentrifuge and the bacterial pellet washed with PBS three times. The pellet was resuspended in PBS and applied onto a EM grid for one minute followed rinsed with water three times. Bacterial cells on the grid were fixed stained with 2% uranyl acetate and imaged using a Philips CM10 transmission electron microscope.

2.2.7 Isolation of Spontaneous Suppressor of Nonmotile *E. acetylicum* strains

To identify spontaneous suppressors of non-motile mutant strains, each non-motile mutant strain was grown independently in BHI broth overnight, and 10 µL of the resulting culture was streaked on the center of 0.3% agar BHI plates and incubated overnight. Bacteria at the edge of the growth zone were collected and re-inoculated in

the center of another 0.3% agar plate. The cycle was repeated until a clear enrichment for motile variants was observed. We isolated rifampin (Rif) resistant variants of these non-motile mutants, and their suppressors, by plating saturated cultures on BHI plates supplemented with 10g/mL rifampin. The selection for Rif^R variants did not affect the mutants or their suppressor's motility status.

2.2.8 Computational modeling

We developed a computational model to predict what fraction of motility genes we would identify as a function of the number of non-motile mutants we examined. Our model simulated the genomic distribution of point mutations in non-motile mutant strains, with simulated loads of 5, 10, and 20 mutations per mutant strain (**Figure 4**). Genes were mutated at random, with the likelihood of mutation proportional to the number of G:C base pairs (the target of EMS and ENU) in each gene. Mutations were randomly generated as nonsynonymous, synonymous, or non-coding regions with a probability of 0.6, 0.3, and 0.1 respectively, based on our observation of rates in pilot experiments. We made two simplifying assumptions in developing the model: i) there are 100 genes in *E. acetylicum* that are essential for motility, which is consistent with what is found in other bacterial species (77); and ii) for each non-motile strain a single mutation is assumed to be responsible for the loss-of-motility and all other mutations are assumed to have no phenotypic consequence. Genes were ranked by the number of nonsynonymous mutations observed in the simulation run, and a motility gene

discovery rate was calculated as the fraction of genes labeled as motility genes that ranked in the top 3%. A final average motility gene discovery rate (Fig S1) is the result from 1,000 simulations. Simulations were performed with MATLAB R2015b (Mathworks). The Pseudo-code for the simulation is provided below.

2.2.9 Gnotobiotic zebrafish colonization

All zebrafish experiments were conducted in conformity with the Public Health Service Policy on Humane Care and Use of Laboratory Animals using protocols approved by the Institutional Animal Care and Use Committee of Duke University. Derivation and colonization of germ-free zebrafish (Tübingen strain) was performed as described (78) with the following exceptions. Germ-free zebrafish were reared in sterile tissue culture flasks in 30mL Gnotobiotic Zebrafish Medium (GZM) and received daily 83.3% media changes starting at 6 days post-fertilization (dpf). Starting at 5 dpf, larvae were fed approximately 2.5mg/day with a custom-formulated diet consisting of, by mass, 45% protein, 15.07% fat, 12.09% carbohydrate, 0.97% fiber, 4x vitamin supplement, 19.57% ash (Zeigler Brothers, Inc.), pelleted and sterilized by irradiation (absorbed dose range 106.5-135.2 kGy; Neutron Products, Inc.). At 6 dpf, overnight shaking cultures of *E. acetylicum* motile and non-motile rifampin-resistant and rifampin-sensitive strains were mixed in 1:500 ratios. Each strain mixture was pelleted and resuspended in 450 μ l of sterile PBS for every 1 ml of strain mixture that was pelleted. Subsequently, zebrafish larvae were colonized by immersion as follows: 83.3% of the media was removed from

each flask, leaving 5ml GZM, and 400 μ l of one of the bacterial mixture suspensions was added to each flask, resulting in an inoculum concentration of $1-6 \times 10^8$ CFU/ml. After 30 minutes of immersion in the *E. acetylicum* strain mixtures, 25 ml GZM was added to each flask. Each flask subsequently underwent three consecutive 83.3% media changes. To assess whole larvae bacterial loads immediately after inoculation at 0 days post-inoculation (dpi) or at 3 days post-inoculation, larvae were euthanized via tricaine overdose (sterile-filtered buffered tricaine at 0.83mg/ml) and individual larvae were homogenized in 500 μ L PBS using a Tissue-Tearor (Biospec) for 1 min at maximum speed. Multiple dilutions of each larval homogenate from the respective flask were plated on BHI with or without 1 μ g/ml rifampin. The number of colony forming units were assessed after overnight growth at 28°C.

2.2.10 Mutational Enrichment analysis

The number of synonymous, nonsynonymous and nonsense mutations were determined for each gene for both the selected and unselected mutant pools. The unselected pool was used as background correction for biases in mutational frequency associated with the particulars of the *E. acetylicum* genome (e.g. GC content, mutational hotspots) by subtracting the number of mutations per gene found in the unselected pool from the selected pool. The number of nonsynonymous mutations per gene after correction was used to generate a rank order list of genes potentially required for motility. A cutoff point for putative motility genes was determined by the number of independent

mutational events expected to have led to a loss of motility (n=440 for the total number of non motile mutants isolated and the assumption that there is one causal mutation per mutant). Genes (n=37) with a net overrepresentation of 4 or more nonsynonymous mutations in the non-motile group accounted for ~60% of the expected mutational events. If a cutoff point of 3 or more nonsynonymous mutations is used, the list of genes increased to 88 and accounted for 92% of expected mutational events.

Two additional criteria were used to generate a “confidence score” that further sub classified these 88 putative motility genes. First, we used the distribution of synonymous and nonsynonymous mutations per gene to perform a Fisher’s exact test (see below) to assess the probability that the spectrum of synonymous vs. nonsynonymous mutations deviates from a random distribution. Prior to performing Fisher’s exact tests, the data was pre-filtered to remove genes that have zero or 1 nonsynonymous mutation from the unselected group. The Fisher’s exact test was applied to the number of synonymous and nonsynonymous mutations identified for each gene in the selected and the unselected groups, and the number of non mutagenized GC bases within each gene. The low frequency of mutational events per gene results in very few genes achieving significant *p*-values, but nonetheless provides a standardized means of generating a likelihood rank.

As a complementary approach, we monitored only the frequency of nonsense mutations, as they are the most likely ones to lead to a loss of function. We determined

the number of nonsense mutations per gene among non-motile mutants, divided them by the potential number of codons in that gene that could be switched to a nonsense codon by a single EMS mutational event, and used that number to rank these genes. This added an additional 14 genes that had been excluded from the preliminary list based on Fisher's exact test. Some of these include small genes with high homology to motility genes, but which had been excluded because of they had very few nonsynonymous mutations.

Next, we ranked the combined list of potential motility genes ($n=102$) by their Fisher's exact test p -value and their normalized frequency of nonsense mutations, and assigned each gene a relative score (scale: 5 to 1) to reflect their relative rank. For instance, for nonsense mutations, arbitrary scores were given to genes in the top 10, 20, 30 or 40 percentile of the normalized frequency of nonsense mutation (only 40 genes out of 102 genes had nonsense mutations). For genes ranked by their Fisher's exact test p -values, arbitrary scores of 5-1 were given to genes in the top 10, 25, 45, or 75 percentile. Finally, an overall score was then determined by multiplying both values. These values were re-ranked and genes were categorized as highly likely (for the top 103 mutational events - 10 genes), very likely (103 mutational events/18 genes), likely (102 mutational events/31 genes), and possible (124 mutational events/43 genes) motility genes. We eliminated from further consideration 2 genes where the number of synonymous mutations was far

greater in the unselected than in the selected group, as this difference is not expected to contribute to motility.

2.3 Results

2.3.1 Bioinformatic and Experimental approach of MEAPS

We sought to develop broad methods for genetic analysis of “genetically intractable” microbes, by using *Exiguobacterium sp* as a representative gut microbe. We first generated a draft genome sequence of strain ZWU0009 (Taxonomy ID: 1224749) (15, 36). The ZWU0009 genome is ~3.2 Mb and includes 3289 CDS, 30 rRNA operons and 76 tRNAs. A comparative genome analysis indicated a close relationship with other previously characterized *Exiguobacterium* and that this isolate is a new variant of *E. acetylicum* (71). We reasoned that one could apply whole genome sequencing to monitor experimentally induced genetic variations in *E. acetylicum* ZWU0009 and then derive associations between gene variants and phenotypically selected traits, a process we term Mutational Enrichment Analysis after Phenotypic Selection (MEAPS) (**Figure 2**).

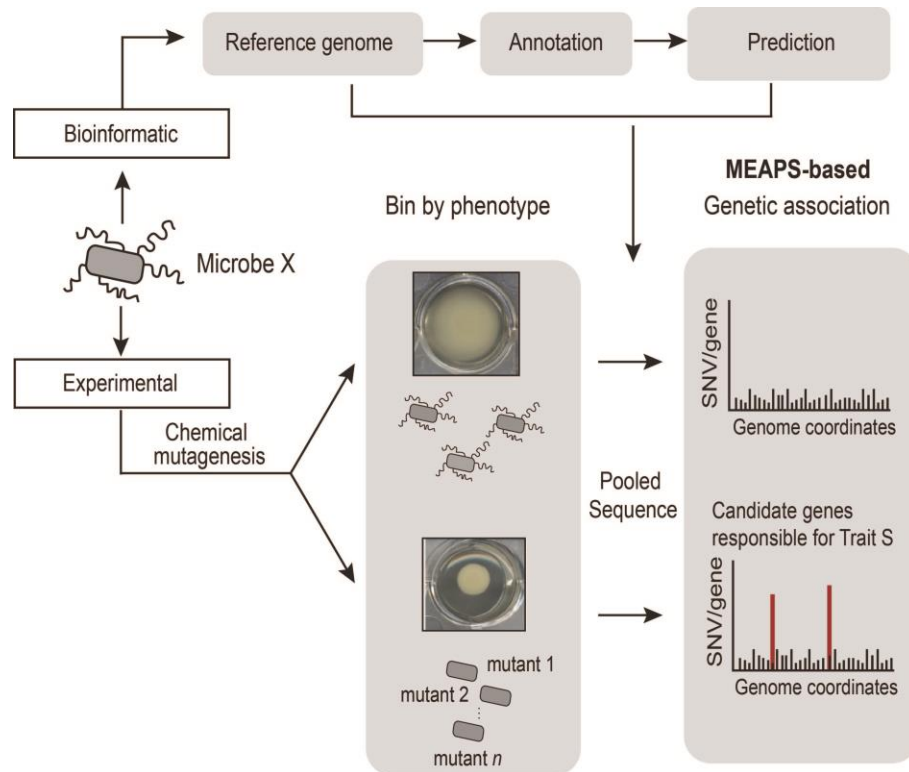


Figure 2: Schematic of the approach coupling chemical mutagenesis and DNA sequencing to perform a mutational enrichment analysis after phenotypic selection (MEAPS) and identify genes required for swarming motility

We chose to perform a MEAPS dissection of *Exiguobacterium* motility, a multigenic complex trait that is important for some bacterial pathogens to colonize the vertebrate gut (79, 80). In contrast, the role of motility in colonization by commensal bacteria is less clear. Metatranscriptomic analysis of healthy gut microbiotas indicates a broad dampening of the expression of motility genes in Firmicutes and Proteobacteria as a result of innate and adaptive immune responses (81, 82), suggesting that the expression of flagella may confer a disadvantage to commensal bacteria. *E. acetylicum*

expresses peritrichous flagella, as assessed by transmission electron microscopy (TEM), when grown in rich media (**Figure 3**).

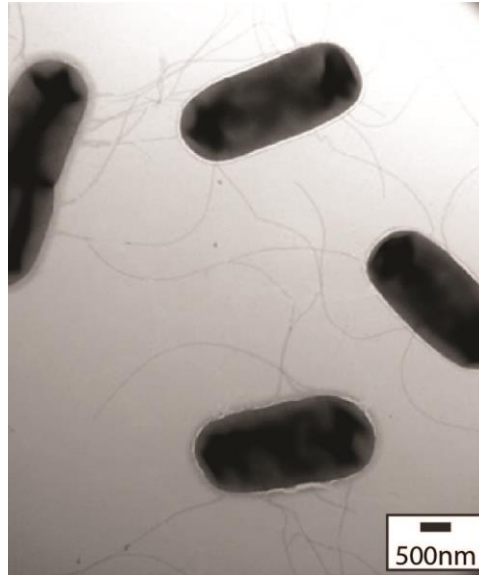


Figure 3: Transmission electron micrograph of *E. acetylicum* flagella (Bar: 0.5uM).

Based on keyword searches and reciprocal BLAST queries we predicted that 126 *E. acetylicum* genes are potentially involved in motility (**Appendix A**). To determine the number of non-motile *E. acetylicum* mutants we needed to sequence to identify mutations overrepresented in motility loci we first modeled MEAPS experiments with the assumption that 100 genes are required for motility and each non-motile mutant had one motility-disabling mutation. These simulations indicated that by sequencing the genomes of 400-500 non-motile mutants with an average of 10 mutations/genome we can capture ~70% of all motility genes (**Figure 4**). Further changes in the number of

mutants sequenced or mutagenesis rates only led to marginal increases to the number of new motility genes that could be identified.

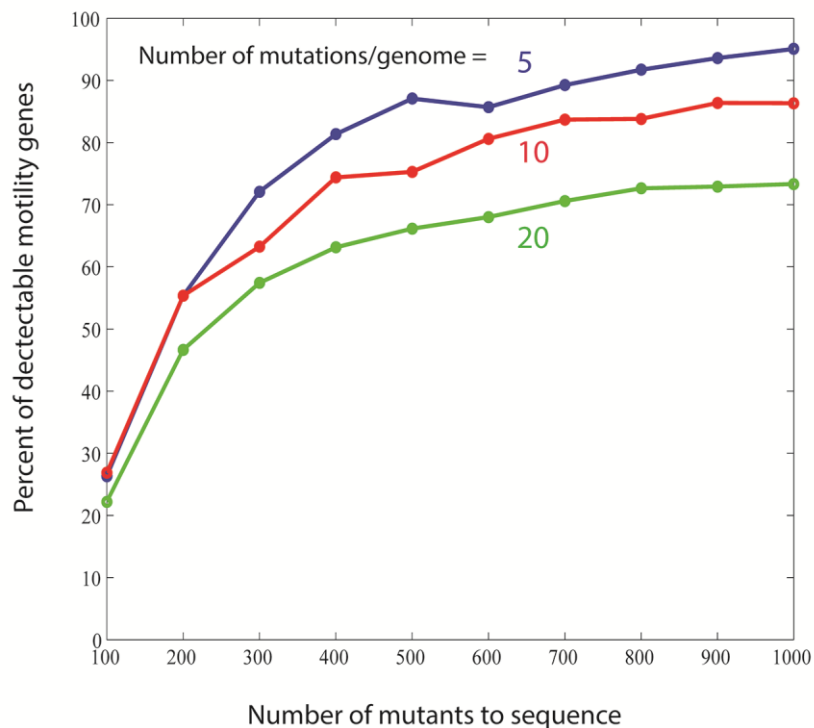


Figure 4: Simulations of the rate of identification of motility genes based on the sequencing of non-motile mutants.

Mutations in the *E. acetylicum* genome were generated *in silico* at random GC base pairs. The mutagenesis rates were altered to generate 5, 10 or 20 mutations per genome, with the assumption that only one mutation was responsible for the loss of motility. The number of “motility” genes was fixed at 100. The simulations indicate that at an average of 10 mutation/genome, sequencing of 400-500 non-motile mutants will lead to the identification of 70-75% of motility genes based on their high frequency of accumulation of nonsynonymous mutations.

We generated *E. acetylicum* mutants by treating bacteria with ethyl methyl sulfonate (EMS) and used the ability of *E. acetylicum* to swarm in soft agar plates to enrich for non-motile mutants. *E. acetylicum* inoculated in the center of a 0.3% agar plate will spread as large halos of turbidity emerging from the inoculation site. After serial passages of pools of mutagenized bacteria in soft agar with repeated collection of bacteria from the site of inoculation, we tested bacterial clones for defects in swarming motility (**Figure 5**) and stored them as individual clones.

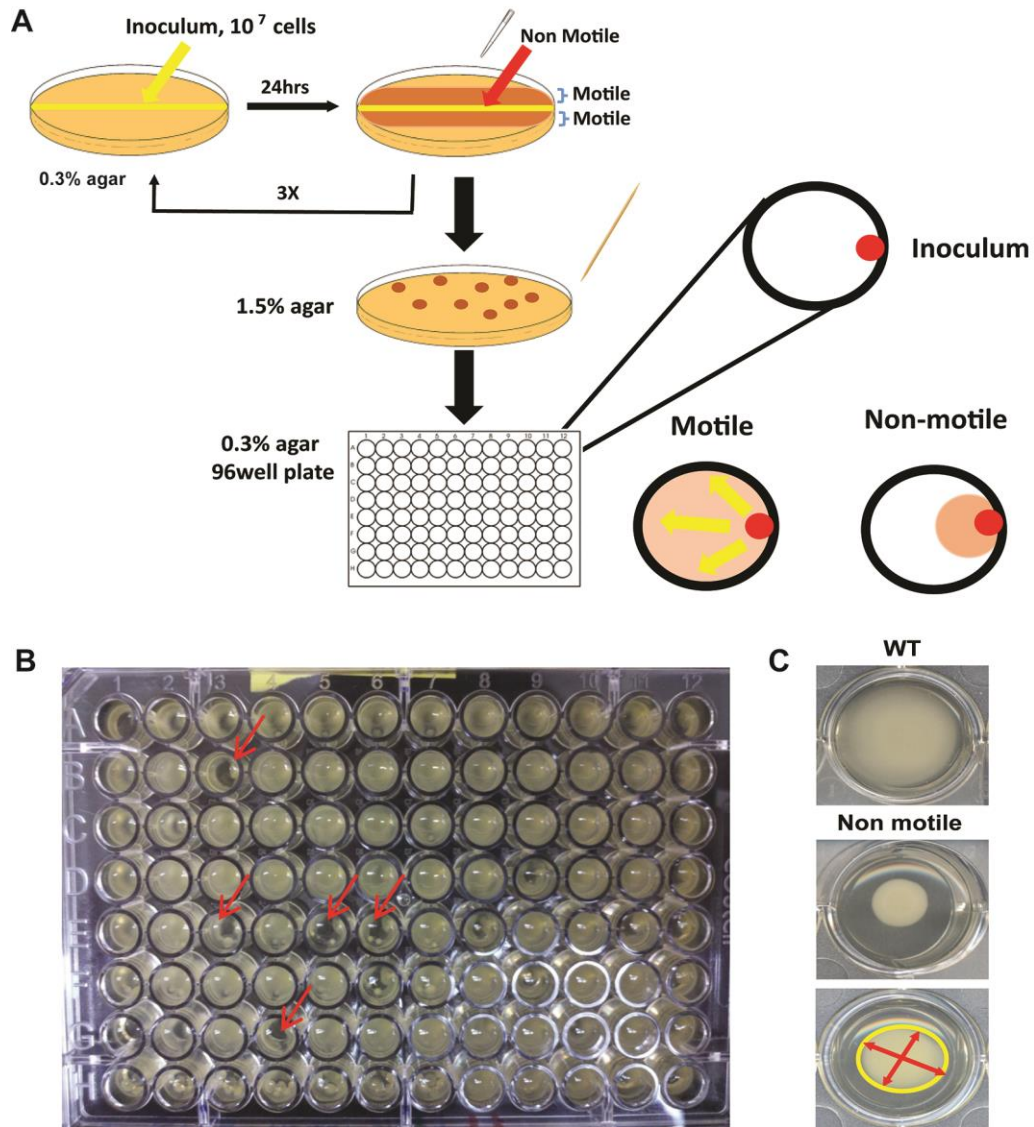


Figure 5: Isolation of non-motile *E. acetylicum* mutants.

A. Chemically mutagenized bacteria were inoculated on 0.3% agar plates to enrich for non-motile strains based on their inability to migrate away from the inoculation site. After 3-4 passages, individual clones were derived from bacteria that remained near the inoculation site and individually tested in a 96 well plate format. Each clone was inoculated at the edge of the well and incubated for 24h. Motile strains covered the entire well and appeared turbid. Strains impaired for motility remained near the edge.

B. Representative image of a 96 well plate used to screen for motility defects. Candidate non-motile mutants are indicated with red arrows.

C. Secondary confirmation tests for the loss of motility. Saturated bacterial cultures were inoculated in the middle of 0.3%

agar in 24 well plates and incubated for 24h. The relative diameter of the bacterial colony was used as criteria to identify mutants with strong defects in swarming motility.

2.3.2 WGS analysis and mutational enrichment genes

We next applied a pooling strategy to sequence 440 non-motile *E. acetylicum* mutants and identified all the mutagen-induced single nucleotide variants (SNV) by mapping unique sequence reads to the reference genome. To compensate for mutational biases due to gene length and relative %GC content, we also sequenced 700 EMS-generated mutants that had not been selected for the loss of motility. Overall, we identified 4009 and 5013 SNVs among the selected (non-motile) and unselected strains, respectively. Next, we compared mutation frequencies between these sets of mutant strains and identified two regions in the *E. acetylicum* genome that displayed a marked accumulation of nonsynonymous, but not of synonymous, mutations (**Figure 6** and **Figure 7**).

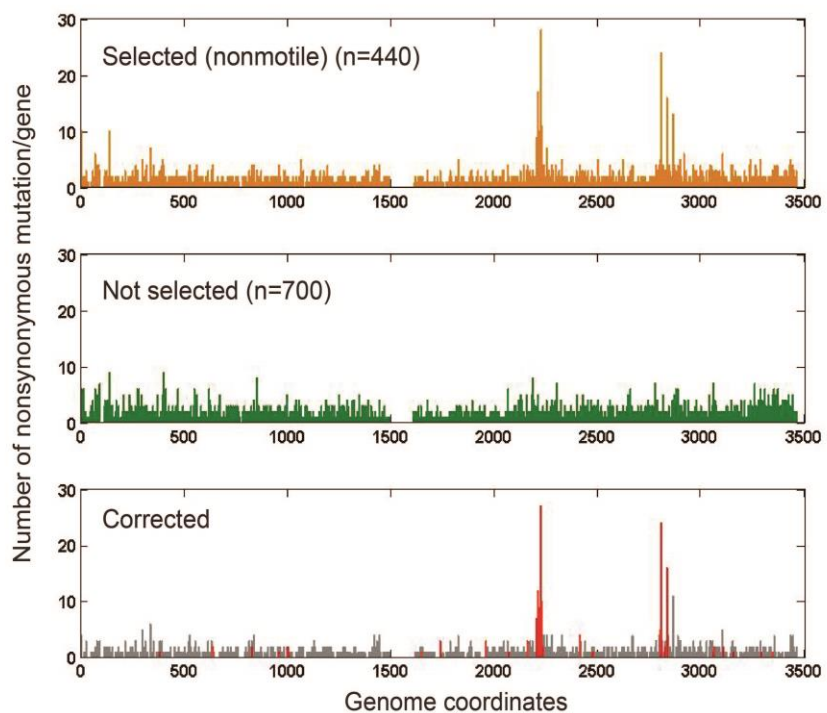


Figure 6: MEAPS of *E. acetylicum* strains defective for motility.

E. acetylicum mutants were selected based on the loss of swarming ($n=440$) or not selected ($n=700$) and their genomes sequenced. The normalized frequency of nonsynonymous mutations reveals two chromosomal regions (RI and RII) that preferentially accumulate mutations in non-motile *E. acetylicum* strains. Red bars highlight predicted motility genes based on their similarities to motility genes in other bacteria.

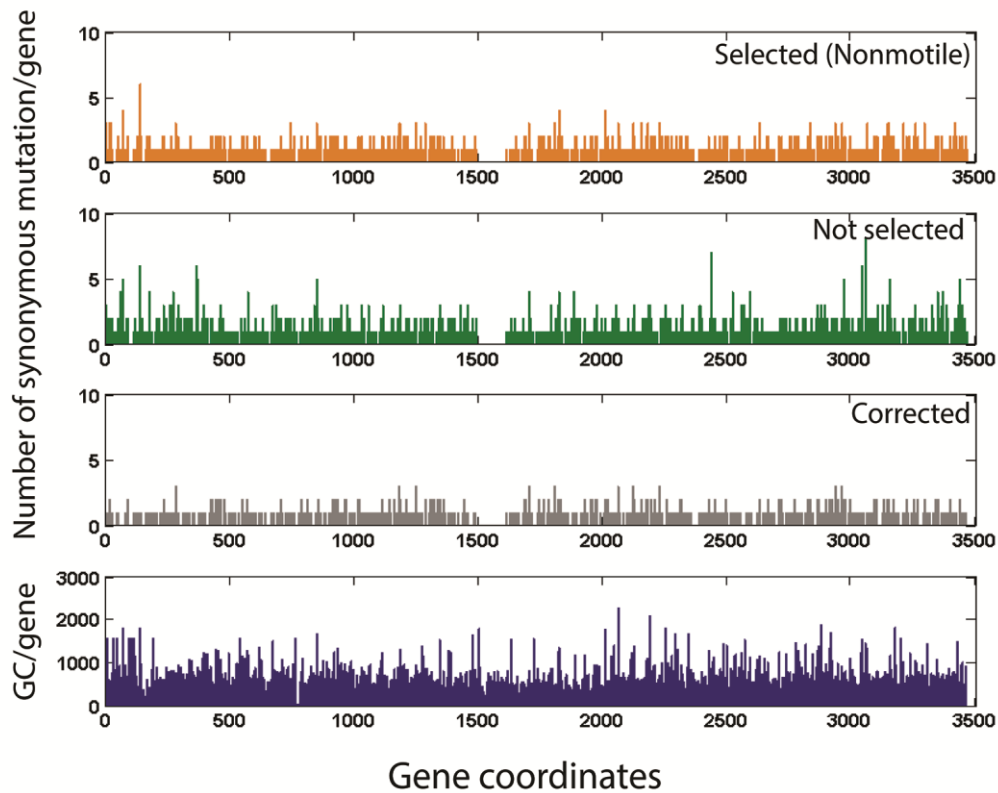


Figure 7: Comparative analysis of the distribution of synonymous mutations among *E. acetylicum* mutants.

The location and abundance of synonymous mutations in *E. acetylicum* mutants that were selected or not selected based on motility was determined by sequencing the genomes of pools of mutants. The corrected frequency of synonymous mutation in non-motile *E. acetylicum* mutants reveals no clear bias in mutations in any particular locus. Number of GC bases per gene (bottom panel) indicates the likelihood of mutational biases based solely on targets for EMS mutagenesis. Note: The regions between 500-1600 were omitted from analysis as this region contains rDNA and other repetitive regions that confound alignment of sequencing reads to a unique locus.

Region I encompasses most predicted flagellar structural genes and chemotaxis genes (Figure 8).

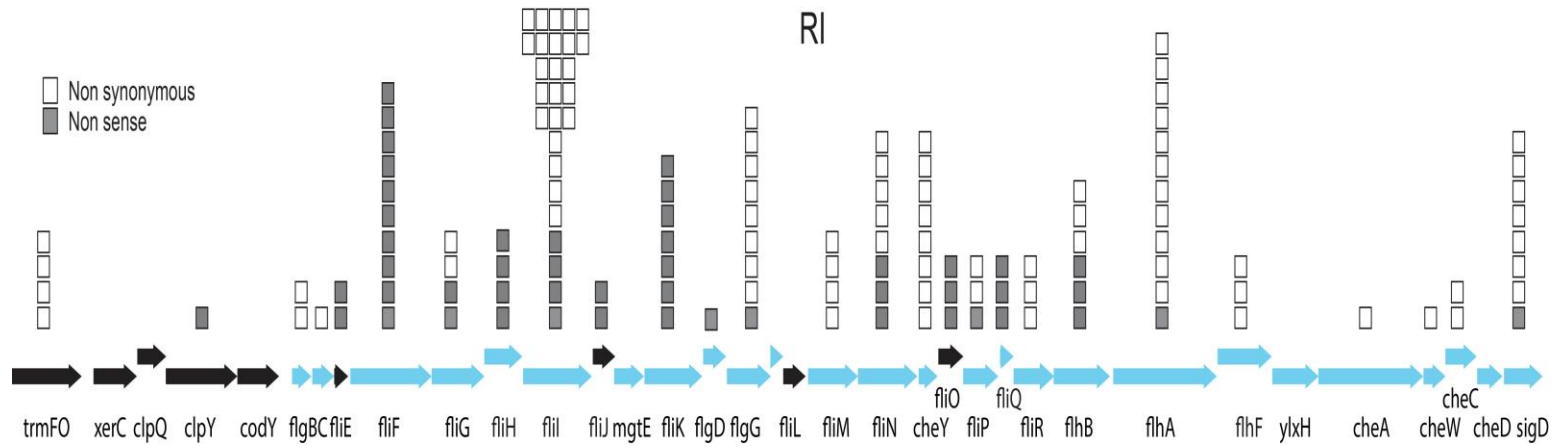


Figure 8: Genetic map of ORFs in motility Region I.

Genes with homology to predicted motility genes (Appendix A) are shown as blue arrows and the number of nonsense and non-synonymous mutations identified, after correction, is represented by gray and white squares, respectively

Region II encoded additional predicted flagellar structural components, including two flagellin (*hag*) genes, and the two component regulatory system, DegS/DegU (77). Using more stringent criteria (see Methods), we defined 57 genes as most likely required for motility, including 27 genes homologous to genes not commonly associated with motility in Gram positive bacteria and 7 genes encoding proteins of unknown function.

In general, the highest confidence 21 genes identified by MEAPS (**Figure 9** and **Appendix B**) are homologous to genes previously associated with the assembly of flagella or the regulation of motility in bacteria (77). Overall, 102 genes were overrepresented among non-motile strains including 19% of genes identified as potential motility genes by key word searches (**Appendix B**).

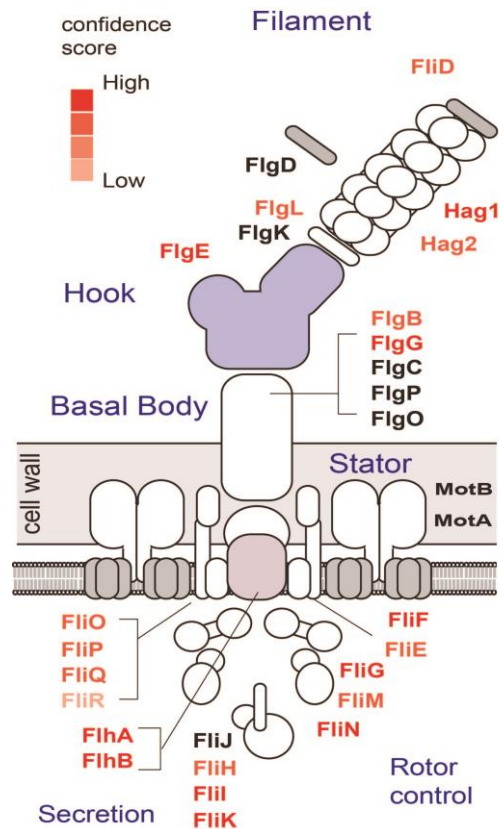


Figure 9: Cartoon schematic of the Gram positive flagellar apparatus displaying all components conserved in *E. acetylicum*.

Components identified by MEAPS are shown in different font colors to reflect confidence of their relative association with motility (**Appendix B**).

Strikingly, only 23% (N=24) and 55% (N=56) of these 102 genes were identified as potential motility genes by reciprocal BLAST homology and Pfam terms, respectively (**Figure 10**). These results imply that significant fraction of genes predicted by bioinformatics to participate in *E. acetylicum* motility do not play a role in motility under the conditions tested.

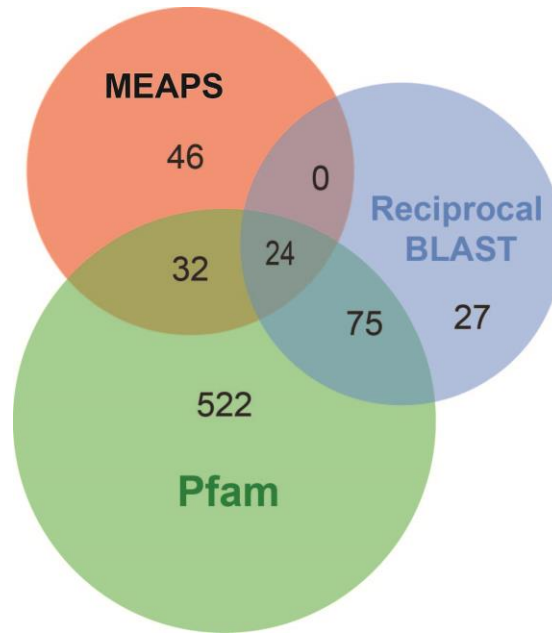


Figure 10: Overlapping set of putative *E. acetylicum* motility genes identified either by reciprocal BLAST homology searches, Pfam terms associated motility, and MEAPS.

We also monitored intergenic regions (**Figure 11**) and identified mutations in three loci that were overrepresented in non-motile mutants, including SNVs mapping to the predicted ribosome-binding site of *cheY* and *flhA*, encoding putative chemotaxis and flagellar structural proteins, respectively.

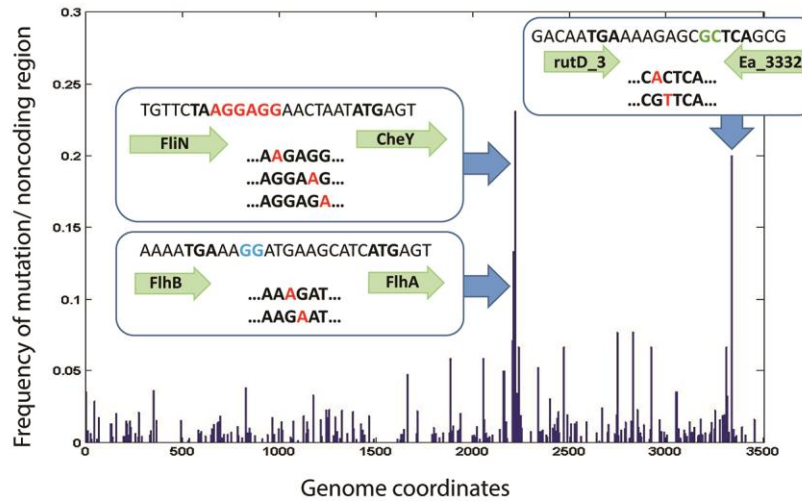


Figure 11: Comparative analysis of mutations in noncoding regions.

Distribution of single nucleotide variants (SNVs) in non-coding regions of *E. acetylicum* mutants that had been selected based on their motility. The frequency of mutation was normalized to the number of nucleotides in the non-coding region after correcting for mutation biases. In the regions exhibiting abnormally high frequencies of mutations, the nucleotide variants mapped to predicted ribosome sites of *cheY* and *flhA* respectively. The consequence of mutations in the intergenic region between *rutD3* and *ea3382* on motility is currently unknown since these genes do not appear to be directly involved in motility.

Although MEAPS led to strong associations between mutations in specific genes and the loss of motility, it is difficult to unequivocally assign causality to any one mutation, especially since the average number of SNVs per mutant strains is 9.3. We applied another basic tool in microbial genetics, the isolation of spontaneous genetic suppressors mutations, to determine if the genes identified were responsible for the loss of motility. For this analysis we isolated strains from among our mutant collection with nonsense mutations in *flhA*, required for flagellar biosynthesis (83); *fliE*, *fliF*, *fliK* and

flgG1, encoding core components of the basal structure (84); *fliM*, encoding the flagellar M-ring and switch component (85); *flgN*, encoding a secretion chaperone (86); and *hag1*, encoding one of the two predicted flagellin subunits. Mutants were passaged on soft agar plates and bacteria were collected from the leading edge of the inoculation spot. After 3-4 passages, clonal populations of strains that had regained motility were isolated and the DNA region spanning the predicted motility gene was sequenced. In all instances, the suppressor mutations either reversed the original nonsense mutation or changed adjacent nucleotides to generate a reading codon (**Figure 12**), indicating that for loss-of-function mutations in predicted structural flagellar genes and accessory factor(s), the only path to restore motility was to repair the nonsense mutation.

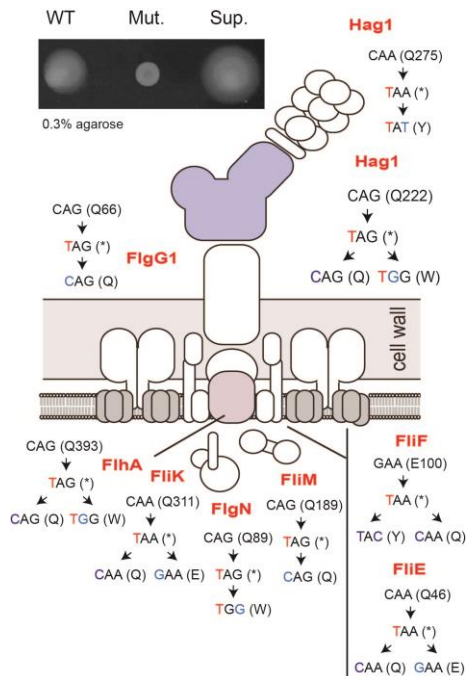


Figure 12: Suppressors of nonsense mutations in putative structural flagellar genes confirm their role in motility.

Strains with nonsense mutations in flagellar components were passaged in soft agar to enrich for spontaneous motile variants (Inset: Hag^{Q222*} and its Hag^{Q222W} suppressor). Sequence analysis indicated the presence of reversions and intragenic suppressor mutations that restored the reading frame. The relevant sequence of mutated and suppressed codons is shown.

In a representative *hag1* nonsense mutant, flagella were observed by TEM only after intragenic repair of the nonsense lesion to generate a Q to W codon switch (**Figure 13**). These findings provide compelling genetic evidence that the mutations identified by MEAPS were indeed responsible for the loss of motility.

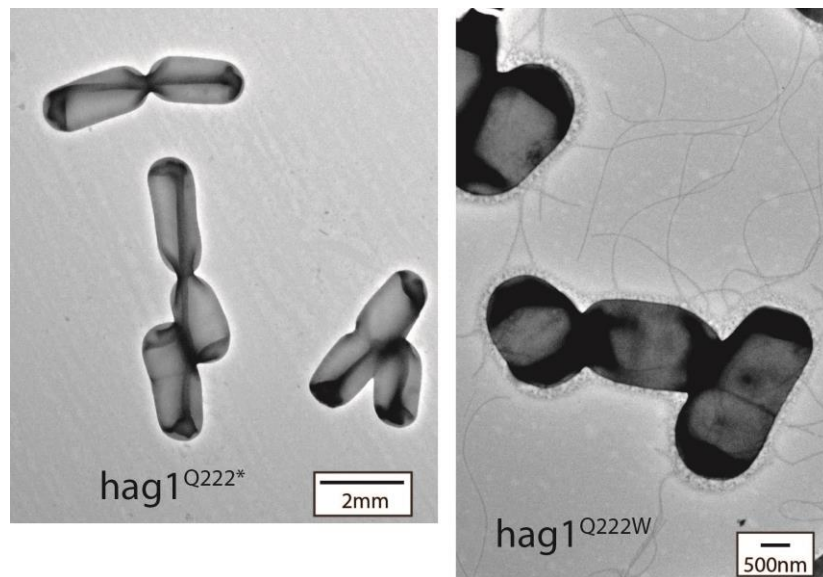


Figure 13: Intragenic suppressor of a *hag1* nonsense allele restores the formation of wild type flagellar structures in *E. acetylicum*.

2.3.3 Protein Functional verification of motile genes by spontaneous suppressor mutation

The availability of genetically defined non-motile and co-isogenic suppressor strains allowed us to test if flagellar motility is required for *E. acetylicum* colonization of zebrafish hosts. We isolated rifampin resistant variants of a strain bearing a *hag1* (*ea2793*) nonsense mutation (Hag1^{Q222*}) and its suppressed sister strain (Hag1^{Q222W}) and performed competitive colonization experiments in germ-free zebrafish larvae. The relative enrichment of one strain over the other at 0 and 3 days post inoculation was then assessed by plating larvae-associated bacteria and enumerating rifampin resistant colonies. The non-motile *hag1* mutants were rapidly outcompeted by their motile suppressors even when the majority of the starting inoculum consisted of non-motile strains (**Figure 14**). Similar results were obtained with a competition between a spontaneous *flhA* mutant and a wild type strain (**Figure 15**), emphasizing the role for *E. acetylicum* motility in stable colonization of its vertebrate host.

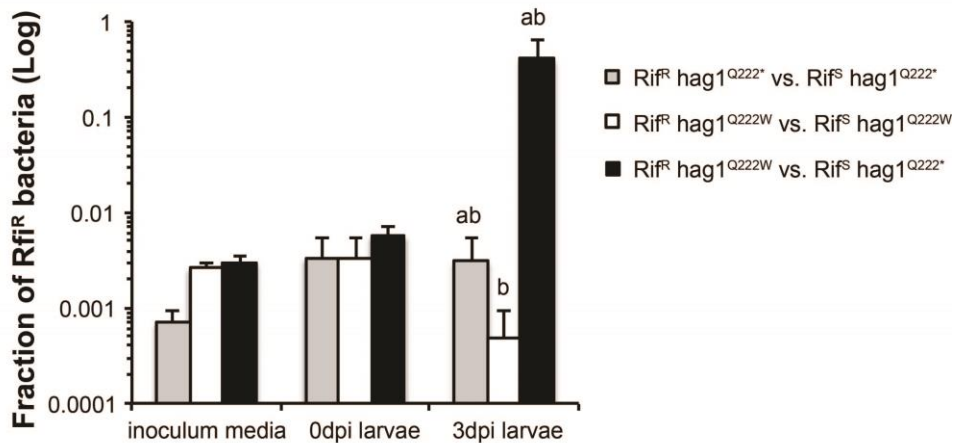


Figure 14: Motility enhances *E. acetylicum* colonization of germ free zebrafish.

Rifampin (Rif) resistant and sensitive versions of an *E. acetylicum* strain with a nonsense mutation in Hag1 (Hag1^{Q222*}) and its motile suppressor isogenic derivative (Hag1^{Q222W}) were placed in direct competition for colonization of 6 days post-fertilization germ-free zebrafish embryos. Inoculum media consisted of Rif resistant strains mixed with sensitive strains at a 1:500 ratio. The relative frequency of each strain in the inoculum media at 6 days post-fertilization, in association with animals immediately after colonization at 0 days post-inoculation (0dpi) and after 3 days of association (3dpi) were determined by assessing the percentage of Rif resistant bacteria. Error bars represent standard deviation. Letters indicate $P < 0.05$ compared to respective inoculum media (a) or 0dpi larvae (b) using Kruskal-Wallis test.

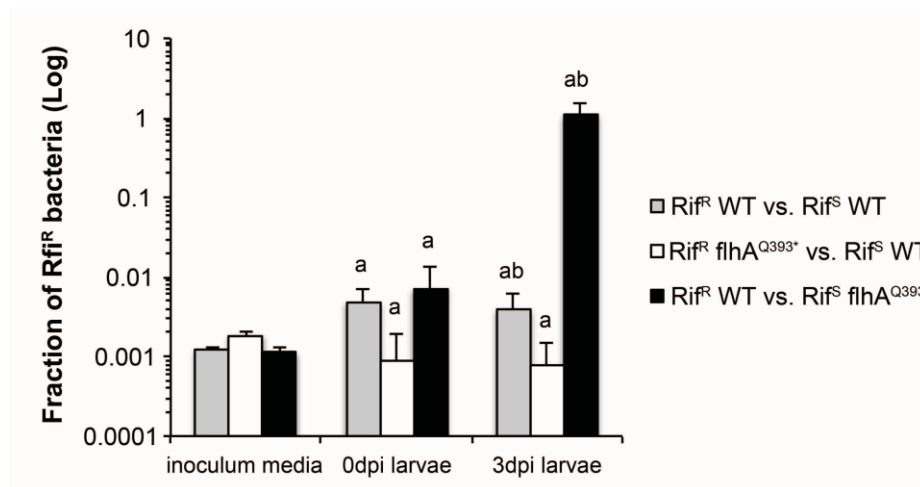


Figure 15: Motility contributes to the efficient colonization of zebrafish

Rifampin (Rif) resistant and sensitive versions of an *E. acetylicum* strain with a nonsense mutation in flagellar biosynthesis gene, *flhA*, (*Flh^{AQ393*}*) and wild type strain were placed in direct competition for colonization of 6 days post-fertilization germ free zebrafish embryos. Inoculum media consisted of Rif resistant strains mixed with sensitive strains at a 1:500 ratio. The relative frequency of each strain in the inoculum media at 6 days post-fertilization, in association with animals immediately after colonization at 0 days post-inoculation (0dpi) and after 3 days of association (3dpi) were determined by assessing the percentage of Rif resistant bacteria. Error bars represent standard deviation. Letters indicate $P < 0.05$ compared to respective inoculum media (a) or 0dpi larvae (b) using Kruskal-Wallis test.

We extended genetic suppression analysis to define the role of genes annotated with ambiguous or unknown functions that we identified as motility genes by MEAPS. We isolated suppressors of nonsense alleles of uncharacterized genes in RII: *ea2862*, encoding a protein with domains with predicted diguanylate cyclase (GGDEF) and phosphodiesterase activities (EAL), respectively (87); *ea2619*, encoding a hypothetical protein; and *ftsX*, encoding a protein associated with septation and sporulation (88) (Figure 16).

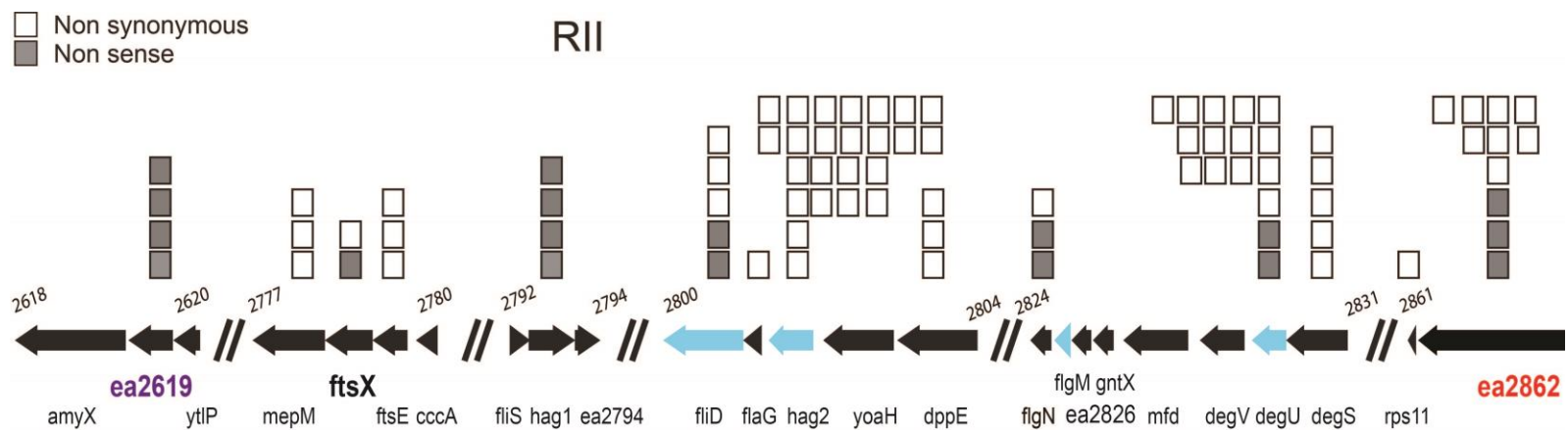


Figure 16: Motility Region II of *E. acetylicum* encodes for new motility genes.

Non-synonymous mutations in uncharacterized genes and close homologues of predicted motility genes (blue arrows) in the RII motility region are overrepresented among non-motile *E. acetylicum* mutants. The number of independent non-sense and non-synonymous mutations identified is represented by gray and white squares, respectively.

GGDEF/EAL domain proteins regulate the formation of c-di-GMP, a signaling molecule that controls multiple cellular behaviors including motility and biofilm formation (87). The *E. acetylicum* genome encodes ten proteins with tandem GGDEF/EAL domains, but only two genes encoding GGDEF/EAL proteins displayed a significantly higher mutational load among non-motile mutants (**Figure 17**).

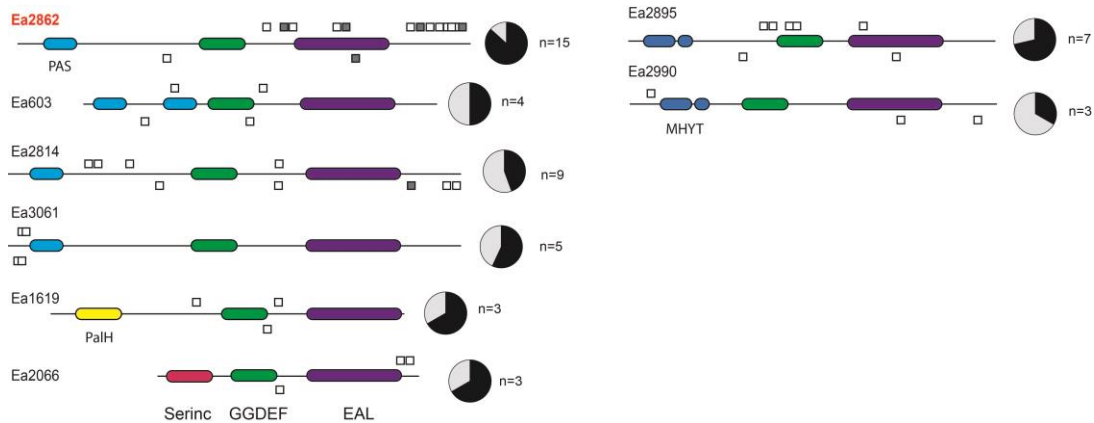


Figure 17: GGDEF/EAL domain proteins in *E. acetylicum*. Only genes with more than 3 total non-synonymous mutations are shown.

Pie charts indicate the proportion of total mutations identified among non-motile (black) and unselected (grey) mutant pools. Squares on top and bottom of each gene represent the location of mutations identified in the selected (non-motile) and unselected group, respectively. GGDEF, diacylglycerol kinase domain; EAL, phosphodiesterase domain; PAS, sensor for signal transduction; MYHT, bacterial signaling (Pfam03707); Serinc, serine incorporator (pfam03348); PalC, proteolytic processing (pfam08733).

For the *ftsX* nonsense mutant, as with structural flagellar genes, the suppressor mutations isolated were intragenic and consisted of reversions of the original mutation or nucleotide changes that restored translation (**Figure 18**). In *B. subtilis*, the FtsXE complex is required for the secretion of peptidoglycan hydrolases CwlO and LytE and proper septum assembly during sporulation and elongation (89), and peptidoglycan remodeling is important for flagellar biosynthesis and function in Gram positive bacteria (90). Interestingly, although FtsX in *B. subtilis* and *E. acetylicum* are 49% identical and share similar chromosomal location adjacent to motility genes, FtsX has not been

associated with motility in *B. subtilis*. Flagella were readily apparent in *ftsX* mutants indicating that CwlO, LytE, and related peptidoglycan remodeling enzymes are not required for flagellar assembly (**Figure 18**).

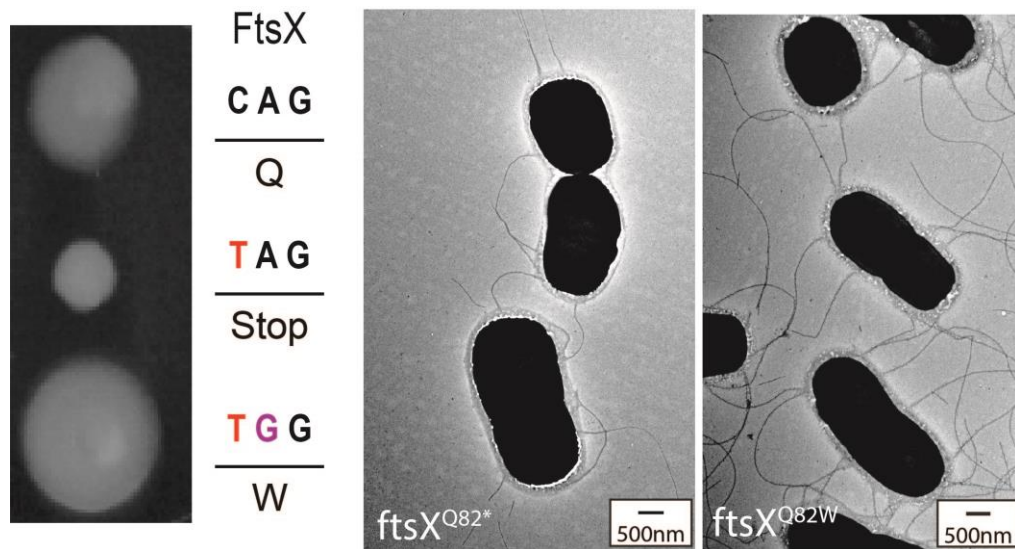


Figure 18: Motility behavior and flagellar assembly of a FtsXQ82* nonsense *E. acetylicum* mutant and a spontaneous variant that regained motility (FtsXQ82W).

TEM analysis of mutants and their suppressor indicate that FtsX is not required for flagellar assembly.

We propose that the negative impact of *ftsX* mutations on motility may be the result of improper assembly of additional factors required for flagellar function, as has been suggested for gliding motility in Flavobacteria (90, 91). Consistent with this premise, mutations in the putative ATPase FtsE and multiple peptidoglycan lyticases (Lyt) were also overrepresented among non-motile *E. acetylicum* mutants (**Appendix B**).

2.3.3 Discovery of *E. acetylicum* specific motility genes

For nonsense mutations in *ea2862* and *ea2619* we obtained multiple extragenic suppressor mutations that restored motility (**Appendix C**). Common suppressor mutations of at least three independent nonsense alleles of *ea2862* and of *ea2619* included point mutations in the flagellar switch proteins FliM and FliN, regulators of chemotaxis (CheY and PtsI) (92, 93), and the kinase DegS (94) (**Figure 19**).

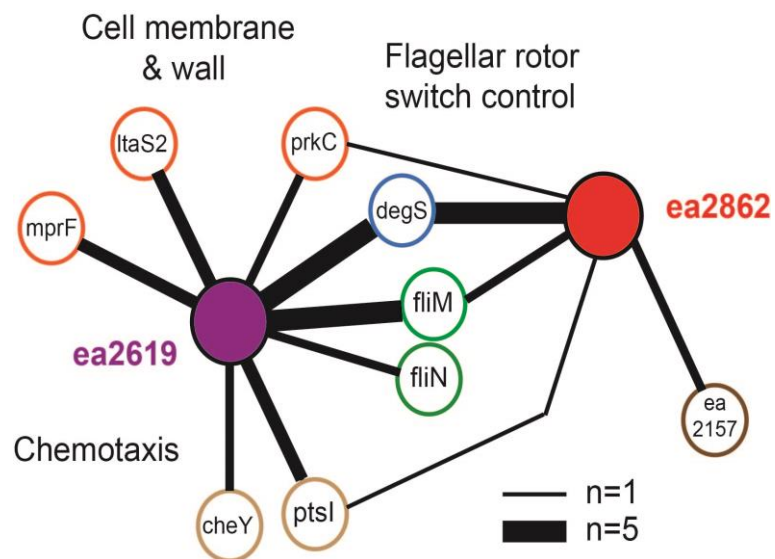


Figure 19: Extragenic suppression analysis of non-motile *E. acetylicum* mutants identifies a role for cell wall modifications and c-di-GMP sensing in commensal Firmicutes motility.

Genetic suppressors of loss of function alleles in two novel motility genes reveal indicates that loss of motility in mutants defective for *ea2619* and *ea2862* can be bypassed by changes in flagellar rotor switch control and chemotaxis. Motile variants of mutants bearing independent nonsense alleles of *ea2619* or *ea2862* were isolated. Common suppressor mutations (open circles) mapped to chemotaxis genes (brown), rotor control genes (green), regulators of flagellar gene transcription (blue), and cell membrane homeostasis (orange). Thickness of lines connecting nodes is proportional to the number of independent suppressor mutations identified (**Appendix C**).

Mutations in these genes were also common in spontaneous *E. acetylicum* mutant strains that had been selected for hypermotility on soft agar (**Appendix D**). Overall, these results indicate that mutations in *ea2862* and *ea2619* likely regulate the frequency and direction of motility as opposed to flagellar assembly or function. Consistent with this observation, the suppressor mutations we identified in FliM cluster in regions predicted to regulate interactions with FliG and other motility regulators (95, 96) and potentially the ability for FliM to self-assemble (97). Similarly, mutations in the homologue of the *B. subtilis* Ser/Thr kinase PrkC, which is linked to cell wall metabolism (98) also emerged as common suppressors. Additional suppressors of *ea2619* included mutations in a lipoteichoic acid biosynthetic protein (LtaS2) and MprF, a phosphatidyl lysil transferase, which are involved in maintenance of the cell envelope (99, 100). We speculate that cell wall modifications in these strains alter the rotation capacities of the flagellum (**Figure 20**).

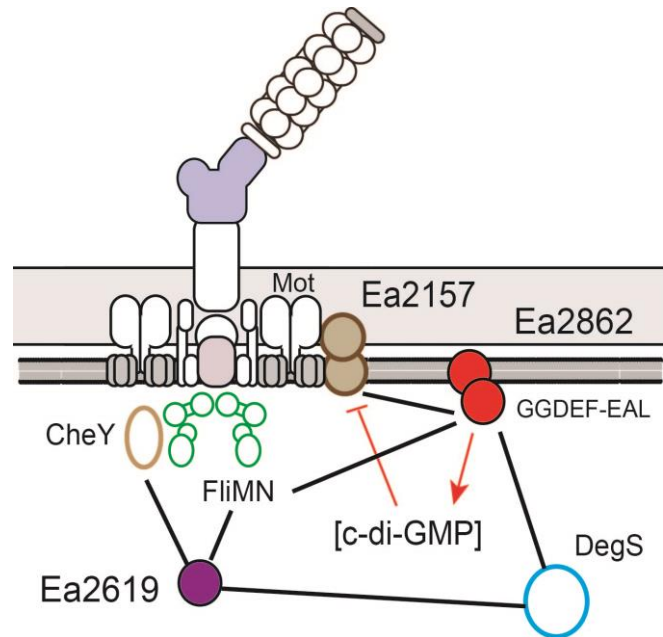


Figure 20: Schematic of suppressor mutations linking Ea2619 and Ea2862 to the regulation of swimming speed and direction.

Ea2157 indicate a direct link for the levels of c-di-GMP regulated by Ea2862 as central to the control of flagellar motility in *E. acetylicum* (red arrows). *ea2157* was independently identified by MEAPS as a putative motility gene.

Some suppressor mutations were gene specific. For instance, two independent amino acid changes in Ea2157, a PilZ domain protein predicted to bind c-di-GMP, specifically suppressed to independent nonsense mutation in *ea2862*. The N-terminus of Ea2157 has homology to the YcgR family of proteins, which in enteric bacteria control chemotaxis and swimming velocity by interacting directly with stator proteins (95, 101). These findings confirm the important role of c-di-GMP in the regulation of flagellar motility in Firmicutes (102). In addition to GGDEF and EAL domains, Ea2862 also has a PAS domain that is predicted to sense small molecules such as metabolites and gases (103).

Indeed, the automated PROKKA annotation program ascribed Ea2862 as a homologue of the oxygen sensing protein DosP (104). In *B. subtilis*, loss of the EAL protein PdeH causes elevated levels of c-di-GMP that inhibits motility through the PilZ domain protein YpfA and the flagellar motor protein MotA (101, 105, 106). Aside from the EAL domains in Ea2862, this protein does not share any homology with its *B. subtilis* counterpart. The role of Ea2862, or Ea2157, in motility could have not been predicted solely on bioinformatic analysis and highlights the utility of MEAPS for the functional annotation of a microbial genome (**Fig. 19**). We hypothesize that the role of Ea2862 in integrating multiple signals, through its PAS domain, to regulate c-di-GMP levels and motility is more complex than that performed by PdeH in *B. subtilis*. The observation that many *E. acetylicum* GGDEF/EAL proteins also have PAS domains is intriguing (**Fig. 16**) given that this domain can bind diffusible signal factors, such as unsaturated fatty acids (107), to potentially regulate c-di-GMP signaling and intra and interspecies cell communication in the intestine.

2.4 Discussion

The paucity of experimental tools to manipulate bacterial genomes has emerged as a major roadblock to understand how microbial communities assemble and influence human and environmental health. Here, we describe how coupling of chemical mutagenesis, phenotypic selection, suppression analysis, and genomic sequencing-based mutational mapping, can be applied to rapidly derive strong phenotype-genotype

correlations in a microbe with no pre-established molecular genetic tools leading to a functional annotation of previously uncharacterized genes. A similar conceptual framework has been proposed to explore gene function in the obligate intracellular pathogen *Chlamydia trachomatis* (47), conservation in protein function in bacteriophages (49), and to identify genes required for magnetosome formation (108) and microbial drug resistance (109, 110). By extending the application of genomic sequencing to the large-scale analysis of suppressor mutations, we genetically confirmed the contribution of specific genes to a complex trait and further revealed functions for new proteins whose activity could not be inferred solely from sequence homologies. We anticipate that this approach will significantly improve our ability to define the molecular mechanisms by which any culturable bacterial species interacts with their environments, aid with genome annotations, and provide biological tools to probe the function of specific genes in complex microbial communities.

3. Genomic-sequencing mutational analysis to identify essential genes in a genetically intractable microbe

3.1 Introduction

Essential genes can be defined as genetic material indispensable for the growth and survival of an organism. Essential genes are responsible for biological functions required for replication and survival and can vary between different microbes and in response to the specific environment in which the organism resides. In the strictest sense, essential genes are required for life in the most “optimal” growth conditions in which no selective pressures are placed. For example, essential genes include those required for DNA replication, transcription and translation. Accurate identification of essential genes may allow for the discovery of novel and specific drug targets for antimicrobial development.

A set of essential genes have been defined in model bacterial organisms, including *Bacillus subtilis*, a relative of *E. acetylicum*. *B. subtilis* has 271 essential genes which were identified by individually knocking out target genes by inserting a nonreplicating plasmid (111). A gene was defined as essential if the target genes was resistant to genetic disruption. This approach is powerful, but is not one that can be applied to microorganisms that lack molecular genetic tools (“genetically intractable”). We hypothesized that we can expand on our previous work, the identification of motility genes in *E. acetylicum*, to identify essential genes in the same microbe.

Our approach to identify essential genes in *E. acetylicum* consist of i) computational modeling to estimate the number of mutations we would need to generate to distinguish between essential and non essential genes and ii) generation of complex libraries of chemically mutagenized bacteria and iii) developing methods to inexpensively map transition mutations by WGS. Because nonsynonymous mutations can cause a disruption of protein function, our definition of “essential” for a gene is that the gene will have a relatively low tolerance for the accumulation of nonsynonymous mutations as compared to other genes in the genome. Our goal is to generate and sequence enough mutant strains such that we can achieve statistical confidence that based on the rate of nonsynonymous mutations we identify all non-essential genes. To test feasibility of this approach, we first introduced *in silico* mutations in the *E. acetylicum* genome and tracked the rate of nonsynonymous mutation between essential and non-essential genes.

3.2 Material and Methods

3.3 Results

3.3.1 Computational models to estimate the number of mutations needed to identify essential genes in chemically mutagenized *E. acetylicum*.

We developed a computational model to estimate number of mutations that are needed to identify essential genes by calculating the fraction of non-synonymous

mutation over total mutations. Our model introduced *in silico* transition mutations (C:G base to T:A base), which mimics those generated by EMS mutagenesis. Next, we chose random C:G bases in the *E. acetylicum* genome with the assumption that 10 to 15 mutations are generated per mutant strain. These mutations were evenly distributed across the genome; reflecting the overall distribution of GC bases. The *in silico* mutagenesis led to synonymous, non-synonymous amino acid changes as well as mutations in non-coding regions with a probability of 0.6, 0.3, and 0.1 respectively. These rates are consistent with the rate of mutation we have observed experimentally. Anywhere from 5 to 20% of all genes have been defined as essential in other model organism including *B. subtilis* (~7%) (111), *Mycobacterium tuberculosis* (~10%) and *Staphylococcus aureus* (23%) (112). For our simulations, we arbitrarily assigned different percentages of essential genes (10, 20 and 30% of the *E. acetylicum* genome) to mirror potential experimental outcomes. We expect that the smaller the percentage of the genome that is devoted to essential function, the greater the number of mutations we will need to map to identify essential genes.

For the computational models we made the following assumptions: i) essential genes for growth and survival in any given condition have 80% lower tolerance for non-synonymous mutations as compared to non-essential genes. To mimic this, all G or C bases have equal chance to have transition mutation and then 80% of total non-synonymous introduced in randomly selected essential genes were removed. ii) only

nonsynonymous mutations disrupt protein function (i.e. we ignore mutations in regulatory regions). Next, we calculated the ratio of substitution rate at nonsynonymous and synonymous sites per gene.

$$dN = \text{Rate of nonsynonymous substitution} = \frac{\text{Observed nonsynonymous mutations}}{\text{Number of nonsynonymous mutation site}}$$

$$dS = \text{Rate of synonymous substitution} = \frac{\text{Observed synonymous mutations}}{\text{Number of synonymous mutation site}}$$

In general, a $dN/dS = 1$ suggests that amino-acid substitutions are mostly neutral and a $dN/dS < 1$ implies that amino-acid substitution might be caused from negative selection. Genes were ranked by dN/dS ratio in the simulation and an essential gene discovery rate was established by examining the fraction of essential genes that represent in bottom 10, 20 and 30% dN/dS ratio of the genes respectively. With a low number of mutations ($n = 50,000$), the dN/dS ratio can display fluctuations between genes because of few mutations can skew the dN/dS ratio which does not have enough statistical power to make a prediction as to whether a gene is essential or not (**Figure 21**) whereas dN/dS ratio of overall genes are a lot more stable with 200,000 mutations.

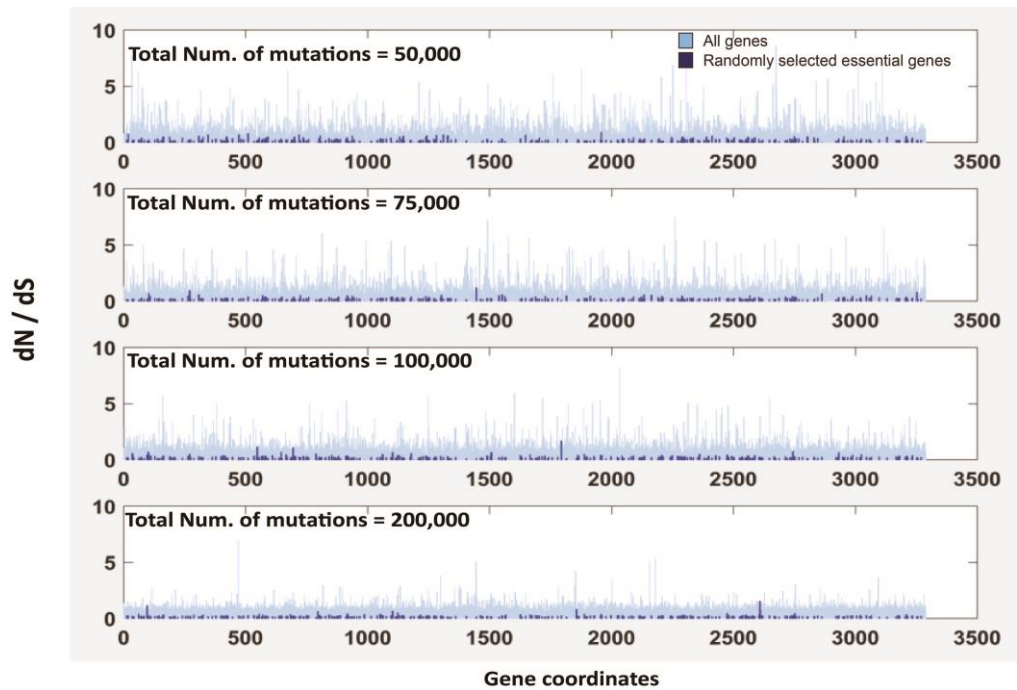


Figure 21: Comparative analysis of the distribution of dN/dS ratio by different mutation load

The final average essential gene discovery rate is the result from 1,000 simulations and our module randomly assigned essential genes each round of simulation. Simulations were performed with MATLAB R2015b (Mathworks). The simulations indicated that 100,000 mutations are needed to capture 80% of essential genes if we use dN/dS information alone. This will require 10,000 mutants to be sequenced under assumption of average 10 mutations/genome (**Figure 22**).

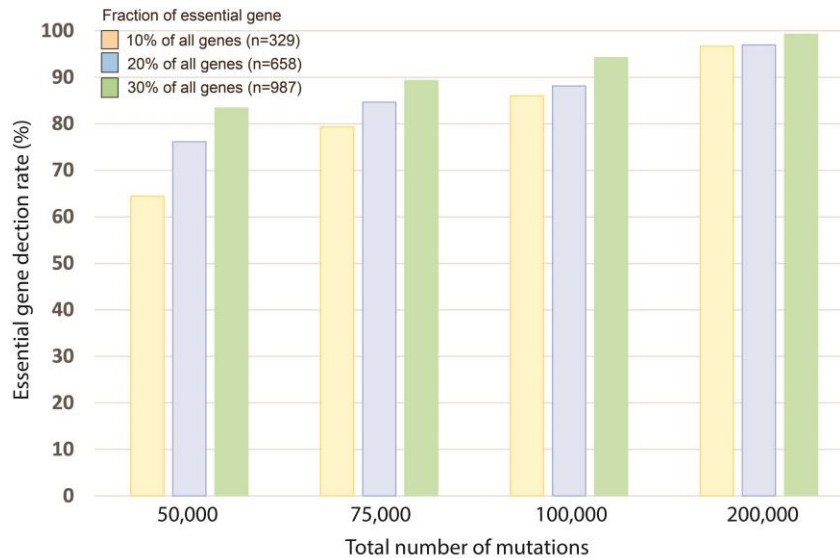


Figure 22: Simulation of the rate identification of essential genes based on the sequencing of mutants with different mutational loads

Further increases in the total number of mutations over 100,000 only showed marginal increases in the rate of detection of essential genes. We also tracked the fraction of genes which have a least one nonsense mutation. Nonsense mutations are very likely to disrupt the function of protein by early termination of protein translation, and thus represents a unique class of mutations as they provided more information than an average non synonymous mutation. If assume that there are no essential genes, the simulation suggest that 100,000 random mutations will lead to nonsense mutation(s) in 66% of all genes. Two fold increase in total number of mutations, (n=200,000), lead to nonsense mutation(s) in 85% of all genes. Because the generation of nonsense mutations rely on the codon compositions in each gene -only six codons (CAG, CAA, TCA, TAC,

CGA and TGG) are convertible to stop codon by EMS mutagen and these codon only compose 8% of total codons in *E. acetylicum* genome – it is highly unlikely that we can generate by EMS nonsense mutation in all non-essential genes. Nonetheless, analysis of nonsense mutations, and the genes in which they reside provide valuable information as to their relative requirement for bacterial viability.

3.3.2 Sequencing strategy to maximize the number of mutations identified in complex mutant pools.

In previous work, we pooled 20 individual *E. acetylicum* mutants to reduce the cost of WGS (\$12/genome). With this pooling strategy, the estimated cost of sequencing 10,000 *E. acetylicum* mutants is 120,000 dollars, which is prohibitive to perform a routine analysis of essential genes in multiple microbe or under multiple environmental conditions. To reduce the high cost of identifying essential genes, we tested the detectable range of single nucleotide polymorphism (SNP) frequency that can be obtained by WGS in an Illumina sequencing platform (HiSeq4000) by increasing the number of mutants that we simultaneously sequence. The overall goal is to minimize the cost per mutation identified.

We examined two potential experimental shortcomings that can lead to the misidentification of SNPs when sequencing of complex pools: i) large variance in sequence library size, ii) duplicate reads in sequence data. First, the size of the library

fragments directly affect to clustering amplification on the flowcell of an Illumina sequencer. A fragment size between 300-500bp excluding barcode is the optimal size for cluster amplification. Larger fragments sizes, up to 1000bp, might not amplify in dense clusters on the flowcell surface and short DNA fragments and primer dimers are overly amplified. Uneven size of library fragments or large size of fragment can also lead to inefficient clustering amplification on the flowcell surface and low sequencing yields. To avoid these issues, we selected a DNA fragmentation size around 450bp and amplified the selected fragmentations by PCR.

The second major issue is PCR duplicates that invariably arise during sequencing library preparation. PCR amplification steps can yield multiple fragments that are not representative of the abundance of the starting material and that can constitute from 4% to 30% of the sequencing library. Ideally during the sequencing library creation, each molecule generated from fragmented genomic DNA should be copied evenly so that all the original genomic DNA molecules are represented on the flowcell. Large variance in the size of DNA fragments highly influences the frequency of duplicate reads, because short fragments are preferentially copied during PCR amplification step. Technical duplication can be influenced by both genomic DNA input and fragment size. The detection rate of SNPs in pooled samples is especially sensitive to degree of technical duplicate reads, as SNP frequency is calculated as the number of mapped sequence reads in a given position over the total number of reads. So, duplicate reads could lead

to erroneous variant detection in DNA-seq data (113). To minimize the effect of duplicate reads, we size selected DNA fragments before the PCR amplification steps.

For microbes with small genomes, identical sequencing reads can also arise by chance because of the limited number of ways a genome can be fragmented and the length of the sequence reads. These sequence duplicates are indistinguishable from PCR duplicates. We average 325,000,000 sequence reads from one lane of a HiSeq4000 sequencing run. Assuming there are no PCR duplicate reads in the data, we still expect to observe 100 identical reads per read that comes from independent DNA molecule fragments ($325,000,000 \text{ reads} / 3,200,000 \text{ bp genome} = \sim 100 \text{ copies}$). A compression step increases the rate of detection to true mutation-induced variants by reducing the “noise” of the vast excess of reads that map to mutagenized sequences in the reference genome and thus offer no information. After compression, 80~90% of total reads have more than one exact match. Only the unique reads were used for SNV detection.

Application of the duplicate reads removal steps in pool of 20 mutant sequence data, we observed 5-7 fold increase in SNVs frequency as 10-17% per variant. It shows that increase in pooling samples would maintain detectable range of SNVs frequency and reduce cost of sequencing. To test the effect of removing duplicate reads in the accuracy of SNV detection, we have sequence pools of 100, 200 and 400 individual mutants. As a control, we included in the pooled DNA, samples of strains that we previously sequenced to determine how effective our methods were at allowing the

identification of these pre-identified SNVs. Without compression of duplicate sequence reads step, we were unable to identify any of the SNVs present in the pre-sequenced strains by variant calling tools, GATK and SNVer (76, 114) . The expected SNP frequency of a mutant from pool of 400 strains is 0.025% which is four times smaller than the sequencing error rate, so true mutations would be discarded and considered as a sequencing errors without the compression step. After compression of unique reads and applying calling SNVs with unique sequence reads, we successfully detected 80~90% of the mutations present, which averaged 2-3% SNV frequency, in the pre-sequenced strain. The pre-sequenced strains consisted of 11 different strains including 302 mutations. This method allowed us to increase by 20 fold the number of mutants that could be sequenced as pools.

3.3.3 A strategy to reduce the identification of false mutations.

For our purposes, the compression of sequencing reads was a critical step to identify SNV in a complex sequencing pool that were below the average error rate of the instrument. However, a drawback of this approach is the increased likelihood of false calls in our assignment of mutations. Sequence read quality of each bases would not be informative because the compression steps only accounts for sequence identity, not base read quality. Because low base quality calls are a major predictor of false SNPs, the loss of discriminating information leads to potential introduction of false positive SNVs by

treating all reads as equally informative. Fortunately, Illumina sequencing chemistry has specific sequencing error patterns; including miss assignments in CCG bases and inverted repeated (115, 116), which we can account for in our SNV assignments; although discriminating sequencing errors from true SNVs can be challenging if errors occur within the expected C:G to T:A conversion patterns. To further reduce the rates of false calls, we opted to use sequence reads that have no or more than one base mismatch from the reference genome. From our extensive single mutant genome sequencing data, we know that individual mutants generated under our EMS exposure parameters have on average 10 to 15 mutations and that these mutations are evenly distributed across the genome. It is highly unlikely that two mutations would map within the 100 base pairs present within a sequence read.

3.3.4 Approaches to identify putative *E. acetylicum* essential genes and conservation of these genes among *Exiguobacterium* sp.

We isolated and sequenced 8,500 EMS or ENU generated mutants. Overall, we identified 87,060 SNVs; 52,054 nonsynonymous SNVs, 25,667 synonymous SNVs and 11,338 non-coding region SNVs (**Figure 23**).

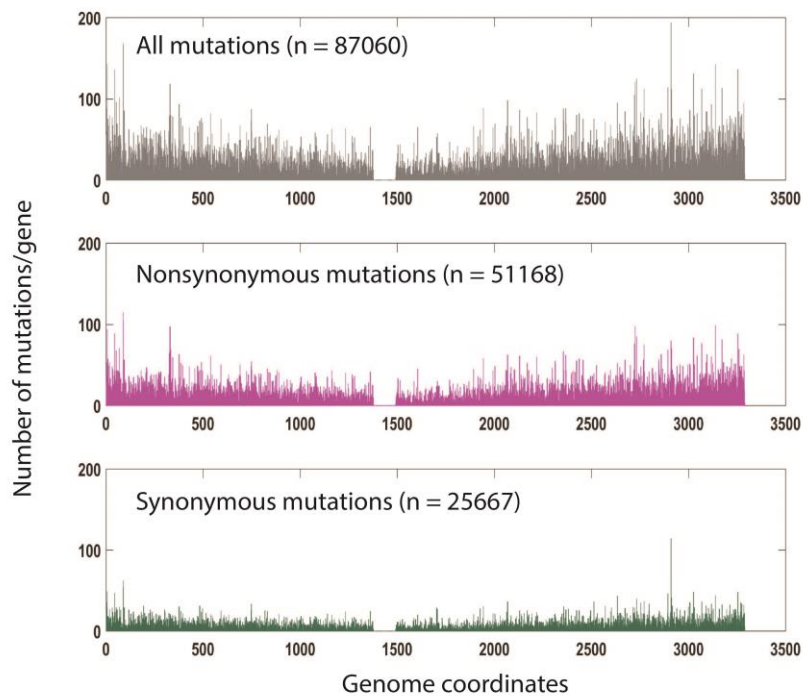


Figure 23: Comparative analysis of the distribution of synonymous and nonsynonymous mutation across *E. acetylicum* genomes of 8500 mutants

Among these, 2173 SNVs led to nonsense alleles in 1382 genes, with 582 genes having two or more nonsense mutations. Excluded from analysis are SNVs that map to repetitive regions and the genome coordinates from 1,400,000 to 1,500,000; which overall includes 114 genes. To identify genes whose mutational load was lower than expected by chance, we applied multiple filtering criteria; first, we sought to remove genes that contained a larger number of mutations. The average number of nonsynonymous mutation per gene is 16.5 (the median, $n = 13$) and average number of synonymous

mutation per gene is 8 (the median. $n = 6$). We considered genes with more than 10 nonsynonymous mutations ($n=2012$) as more likely to be dispensable which accounts for two-thirds of the average number of nonsynonymous mutations. Second, we removed all genes with nonsense mutations ($n=1382$) because the high probability of disabled protein functions with no effect on viability suggests that these genes are not essential. Third, genes which have no nonsynonymous and synonymous mutation were not considered, because there is not enough information for statistical analysis – this group constitutes mostly very small genes. Finally, genes having a dN/dS ratio > 1 are also less likely to be essential gene as multiple non conservative amino-acid substitutions do not appear to be detrimental for viability. After this filtering process 499 genes were categorized as potential essential genes of *E. acetylicum*.

An analysis of these 499 genes indicated that 10% encode for proteins homologous to ribosomal proteins, regulatory proteins, translation, transcription factors and cell walls component– many of which are predicted to be important for growth. Approximately half ($n=222$) of the filtered genes encode for hypothetical proteins of unknown function. Fifty five genes were removed from the essential candidate genes because they had no synonymous mutations, so the dN/dS ratio could not be calculated.

In parallel, we defined a list of 1132 genes that are conserved among 32 *Exiguobacterium* species and that have significant homology to bacterial genes found in databases of essential genes (www.essentialgene.org). Although high genetic

conservation does not necessarily mean that these are essential genes, we expect the majority of *E. acetylicum*'s essential genes to be among the most conserved *Exiguobacterium* genes. The *E. acetylicum* genome shows high similarity to the other 32 *Exiguobacterium* strains (**Figure 24**) with 80%-90% protein sequence similarity to about two thirds (n=19) of *Exiguobacterium* strains.

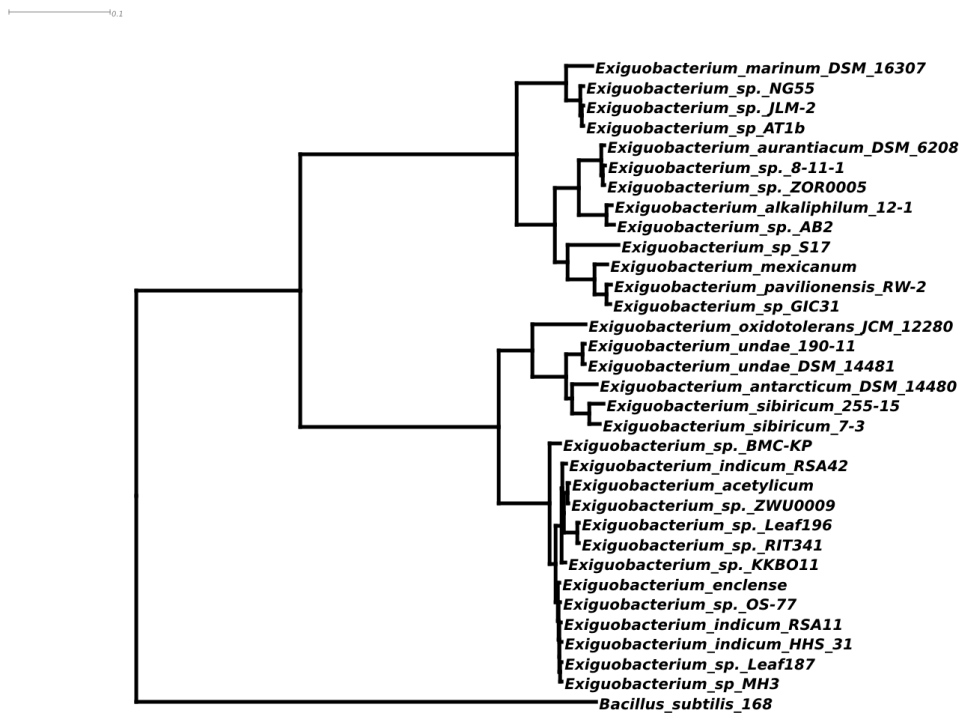


Figure 24: The phylogenetic reconstruction of *Exiguobacterium* based on a concatenated matrix of 32 genes and *B. subtilis* genes

Only 23% of candidate essential genes we generated by sequencing mutant pools (117 out of 499 genes) were found to be among the most conserved *Exiguobacterium* genes, including 11 hypothetical protein. Therefore, the comparison between conserved *Exiguobacterium* genes and candidate essential genes was not sufficient to help further refine gene essentiality. We next compared our list of candidate essential genes in *E. acetylicum* to those that have been experimentally defined in the related Firmicutes, *B. subtilis* (111). Out of 271 essential genes, homologues for 216 *B. subtilis* essential genes could be found in the *E. acetylicum* genome. However, of these 216 *B. subtilis* essential genes only 29 of their *E. acetylicum* counterparts were identified by our study as likely to be essential. These genes mostly involved in ribosomal assembly, cell shape and division, DNA and RNA metabolism.

Because only 6% of the candidate genes are homologous to *B. subtilis* essential genes, we re-examined our filtering criteria. We first compared GC bases between the 29 genes and the 421 genes respectively because our mutations depend on the number of GC bases per gene. The average number of GC bases (208 bp) of the 29 genes is two-fold smaller than the average number of GC bases (420bp) of *E. acetylicum*. This indicates that only small essential gene passed the filtering criteria and we biased against longer genes; most of longer essential genes were removed when we placed a cutoff of more than 10 nonsynonymous mutations per gene (**Figure 25**).

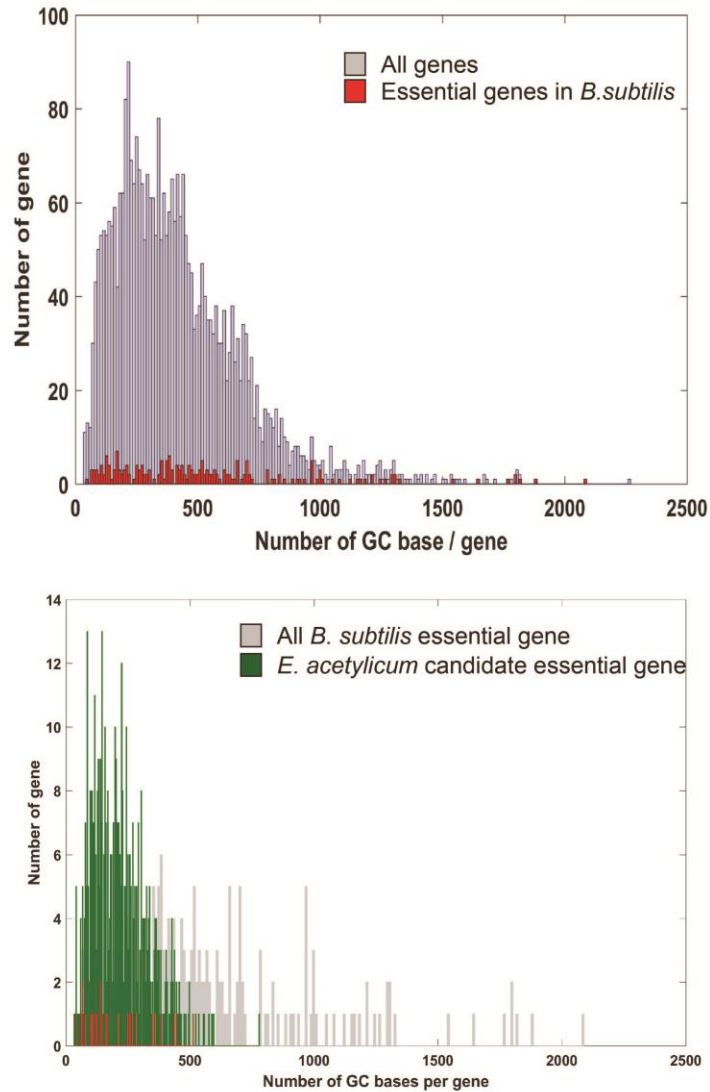


Figure 25: Comparison of number of GC bases per gene between all *E. acetylicum* genes and *B. subtilis* essential genes in *E. acetylicum* genome (upper panel). Comparison of number of GC bases per gene between candidate essential genes of *E. acetylicum* and *B. subtilis*.

For instance, the *rpoB* genes, encoding DNA-directed RNA polymerase subunit beta, is a well-characterized essential gene in most bacteria (111), that was not identified

as an essential gene in our data set because of high number of nonsynonymous mutations (n=115); The length of *rpoB* gene is 3557 bp including 1771 GC bases, which is four times higher than average gene length, with the dN / dS ratio of 0.99. In our experimental data, homologues for the 166 of *B. subtilis* essential genes have more than 10 accumulated nonsynonymous mutations genes and would have been filtered out. The cutoff point would have to be significantly relaxed to accommodate these longer genes, implying that the number of nonsynonymous mutations cannot be used as the filtering criteria.

3.3.5 Efforts at increasing the discovery rate of essential genes in the *E. acetylicum*.

We next used the essential genes of *B. subtilis*, to increase the discovery rate of true essential gene in *E. acetylicum*. Our previous analysis detected only 13% (29 out of 216 genes) of *B. subtilis* essential genes among the *E. acetylicum* candidate essential genes. Since the number of nonsynonymous mutation alone is not a reliable parameter to assign essentiality, we simplified our previous filtering steps: First, essential genes cannot tolerate nonsense mutation. Second, all essential genes should have a dN/dS ratio <1. Using this criteria 962 *E. acetylicum* genes were identified as potential essential genes and 37% (361 genes) of these genes are conserved among all *Exiguobacterium*. With this new criteria, a 100 genes (10.4%) overlapped with essential *B. subtilis* genes (**Figure 27**). Approximately half of these genes are responsible for protein synthesis. Some of

cell envelope, DNA metabolism and RNA metabolism genes were also identified (**Table 1**).

Table 1: Category Subcategory Number of gene

Category	Subcategory	Number of gene
DNA metabolism	Basic replication machinery	3
	Packing and segregation	4
	Basic replication machinery	5
RNA metabolism	Regulation	3
	Basic transcription machinery	2
	RNA modification	1
Protein synthesis	Ribosomal proteins	26
	Protein translocation	3
	Translation factors	6
	tRNA synthetases	13
Cell shape and division		8
Cell envelope	Cell wall	9
	Membrane lipids	7
Respiratory pathways	Menaquinone	4
	Isoprenoids	1
	Cytochrome biogenesis	1
	Thioredoxin	1
Glycolysis		1
Other		2

These modifications to our analysis increased by ~ 2-fold the number of putative essential *E. acetylicum* genes that are homologous to *B. subtilis* essential genes. The approach did no longer discriminate against the discovery of longer essential genes (**Figure 26**).

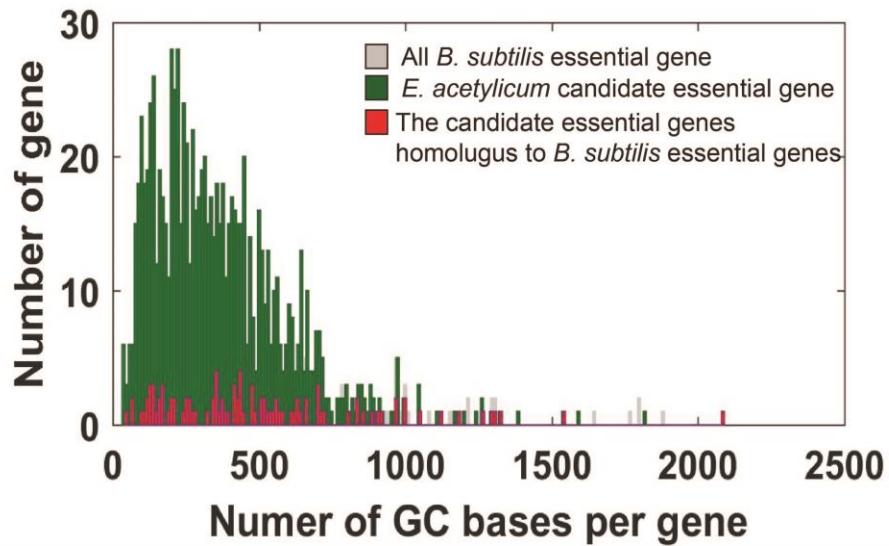


Figure 26: Comparison of number of GC bases per gene between candidate essential genes of *E. acetylicum*, *B. subtilis* essential genes in *E. acetylicum* genome and experimentally verified essential genes in *B. subtilis*

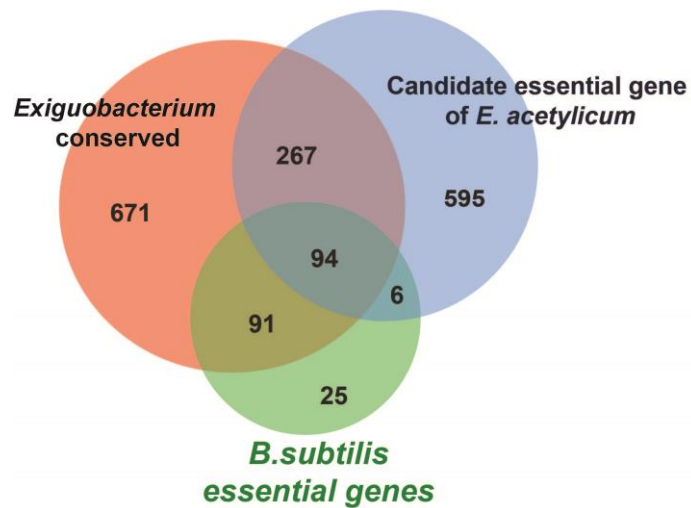


Figure 27: Overlapping set of putative *E. acetylicum* essential genes identified based on their homology to genes conserved among all *Exiguobacterium*, experimentally verified *B. subtilis* and candidate essential genes in *E. acetylicum* by analysis of mutational load

The high similarity between *B. subtilis* essential genes and 85% (185 out of 216 genes) of *Exiguobacterium* conserved genes (Figure) supports using the *B. subtilis* essential gene discovery rate as a convenient indicator of how well our essential gene discovery approaches are working. The size of the *B. subtilis* genome is 4.2Mb comprising of 4,100 coding sequence genes (111) and about 5% (n=271) of the *B. subtilis* genes have been verified as indispensable for growth in a nutritionally rich medium. *E. acetylicum* has a smaller genome, 3.2Mb, with fewer coding sequence genes, 3289 CDS, so the 962 candidate genes we have been identified using more relaxed filtering methods are possibly overestimating the number of candidate essential genes. The source of this overestimation is mostly likely due to the small number of mutations in a significant number of genes which can skew the dN/dS ratio. For instance, among our candidate essential genes, 60 genes have either only 1-2 mutations. These 60 genes were removed from further analysis.

Next, we focused on the 116 *B. subtilis* essential genes that were not present in our list of candidate essential genes. We found that 44 of the *E. acetylicum* homologues of essential *B. subtilis* genes had nonsense mutations, and that 16 of these genes had multiple homologues in the *E. acetylicum* genome, implying that the function of these genes is redundant in *Exiguobacterium*.

3.3.6 False positive essential genes removal and Effects of mutation on protein function

Overall, 84% (760/902 genes) of the candidate essential *E. acetylicum* genes and 67% (65/97 genes) of the *B. subtilis* essential genes with homologues in *Exiguobacterium*, have less than 600 GC bases. In addition, 75% (178 out of 238 genes) of hypothetical genes that are among our list of candidate essential *E. acetylicum* genes are low targets for mutagenesis (<300 GC bases per gene) (**Table2**). We observed 7 motility genes, which are expected to be non-essential genes, in the candidate essential genes. This implies that the candidate genes set includes false positives, which again is skewed towards short length genes as the dN/dS ratio fluctuate more easily by gaining few mutations. This false discovery rate can be improved by sequencing more mutants to increase statistical power.

Table 2: Distribution of the candidate essential gene production varied by number of GC bases/gene

Number of GC per gene (bp)	Number of the candidate gene	Num. of <i>B. subtilis</i> essential gene	Num. of hypothetical gene	Num. of motility gene
~ 300	391	28	178	4
301 – 600	369	37	55	3
601 – 900	110	17	5	0
901 – 1200	21	9	0	0
1201 – 1500	7	4	0	0
1501 ~	4	2	0	0

As not all nonsynonymous mutations have equal effect on protein function, a calculation of functional effect on each nonsynonymous mutations may improve the filtering process further. Non-essential genes are expected to tolerate non-neutral mutations without growth defect whereas non-neutral mutations on essential gene can influence on cell viability.

3.3.7 Curation of essential genes by using information of the mutational impact of nonsynonymous amino acid substitutions.

To test the functional impact of nonsynonymous mutations on the function of the gene product, we applied SNAP to identify non-acceptable protein polymorphisms (117). SNAP predicts functional effect of functional mutations by considering evolutionary information, secondary structure and solvent accessibility. The prediction is presented as a score range from -100 to +100, neutral to strong protein function alteration. We assume that essential genes would not tolerate non neutral mutations with high prediction score. To test this assumption, we calculated the SNAP score of one of candidate essential genes. A known *B. subtilis* essential gene *rpsQ*, encoding 30S ribosomal protein S17, also passed our essential gene filtering criteria; the dN/dS value of 0.8 and had mutations that led to 7 amino-acid substitutions. The prediction result of *rpsQ* mutations shows that only one out of 7 nonsynonymous mutations is natural mutation, no effect on native protein function. Rest of the 6 mutations were predicted as

non-neutral mutations which causes functional alteration of protein (**Table 3**). Opposed to our expectation, majority of nonsynonymous mutations of *rpsQ* gene were non-neutral, so *E. acetylicum rpsQ* gene might not be essential for growth or survival.

Table 3: *rpsQ* nonsynonymous mutations analysis and its predicted functional effect

Reference amino acid	Position	Variant amino acid	Predicted Effect	Score	Expected accuracy
M	1	I	Effect	74	85%
R	13	C	Effect	37	66%
V	15	I	Neutral	-11	57%
M	19	I	Effect	47	71%
T	26	I	Effect	3	53%
H	49	Y	Effect	46	71%
L	77	F	Effect	48	71%

A deletion of The *rpsQ* gene knockout of *E. coli* K-12 strain is failed to grow even in rich nutrient medium (118) whereas knockout *rpsQ* gene of *E. coli* is still viable with growth defect (119). Therefore, we could conclude that *rpsQ* gene is not dispensable for growth in *E. acetylicum*.

Due to the gene-target knockout method, list of targeted *B. subtilis* genes were relied on essential genes of other bacteria and essential genes of *B. subtilis* were not fully discovered or tested. It means that all the *E. acetylicum* candidate essential genes could not be found in the *B. subtilis* essential gene list. Essential gene identification is especially challenge when genes do not have nonsense mutations, because impact of single nonsynonymous mutation in a gene can be subtle overall cellular process or not lethal

enough for survival. However, above curation steps improve the assessment of gene essentiality.

3.4 Discussion

An important step towards advancing our understanding the biology of microbes, it is to determine the genetic requirements for growth and survival. Because essential genes are indispensable component of cellular survival, especially species specific essential could be a logical targets for developing antibiotics (120). Some genes are essential genes in some species, not essential for other species; comparative essential genes analysis of multi-species microbe shows that the universal essential genes could be less than 40 genes (121). To identify the essential genes of *E. acetylicum*, we have combined chemically induced mutants with collecting genetic variants by next generation sequencing. This approach was successfully applied for identification of gene set that are responsible for a certain phenotype (122). By development of a sequence analysis pipeline and an analysis framework, we identified 902 candidate essential genes. However, the candidate essential genes are also included genes with low GC content or low mutational loads. To better assess and curate our list of candidate essential genes, a development of an analytical method is necessary to derive statistical significance of single mutational impact on protein function, discussed in Chapter 5.

4. Implementation of a high-throughput sequence analysis framework to design synthetic genes

4.1 Introduction

It has been observed that there is apparent codon usage preference among proteins expressed at high versus low levels within the same organism and even between genes in the same operon (50, 55, 123, 124). However, the underlying mechanism of protein expression has not yet fully addressed because of the limitation of characterizing a high number of sequences encoding the same gene. A 30 amino-acid protein can be encoded by over three million different 90 base pair DNA sequences. Currently, it is impossible to test all of the sequences. To determine the relationship between gene variants, we designed and synthesized thousands of codon-variants of the AcGFP1 protein and categorized their expression levels. We quantified the effect of codon usage and mRNA structure throughout the entire gene, including N-terminal regions. Our observation shows the importance of N-terminal codon bias and overall mRNA structural effect on protein expression level.

4.2 Material and Methods

4.2.1 Oligonucleotides synthesis to introduce random variations of synonymous mutations

The AcGFP1 fluorescent protein multiplex library was created and tested for heterogeneous protein expression level. The degeneracy of the AcGFP1 genetic code was generated by our in-house software, which mapped to overlapping oligos. The full-

length AcGFP1 sequence was fragmented into 75mer oligos whose neighboring oligos share about 15 base pair regions, all with similar melting temperatures. Theoretically, the gene library has 3^{240} possible sequences that encode AcGFP1. All of the oligos were synthesized in house using Dr. Oligo DNA synthesizer (Azco Biotech). The synthesized oligos were cleaved with an ammonia solution and then dried using a vacuum concentrator (Eppendorf Vacufuge) overnight. The dried oligos were resuspended in ddH₂O and the concentrations were measured with the NanoDrop spectrophotometer (Thermo Scientific). Incompletely synthesized oligos, with lengths smaller than 17 base pairs, were removed by using the Qiaquick Nucleotide Removal Kit (QIAGEN).

4.2.2 Fragmented oligo assembly using the PCA method

To construct a combinatorial gene library, the polymerase cycling assembly (PCA) method was used to assemble single strand oligos into double stranded linear DNA constructs (125). Equal amounts of the purified oligos, totaling 200 ng, were pooled together and mixed with i) 0.5 μ l Phusion High-Fidelity DNA Polymerase (New England Biolabs), ii) 10 μ l Phusion HF Reaction Buffer, iii) 1 μ l dNTPs (40mM, New England Biolabs) mix and iv) ddH₂O. The following is the thermal cycling profile of PCA: an initial 30 s of denaturation at 98°C, 40 cycles which consisted of 10 s of denaturation at 98°C, slow ramping at 0.1°C/s from 70°C to 50°C before annealing at 50°C for 2 m, 15 seconds of extension process at 72°C, and then extra 5 minutes of extension time. The PCA product were examined by 1% agarose (Denville Scientific) gel

electrophoresis. The entire assembled AcGFP1 gene library has an overlapping region with the expression vector. The gene library was amplified with a primer set containing overlapping regions (pET_L: AGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACA, pET21_R: GGGCTTTGTTAGCAGCCGGATC). The thermal cycling profile follows: an initial denaturation at 98°C, 30 cycles which consisted of 10 seconds of denaturation, annealing at 55°C for 1 minute, and extension at 72°C for 15 seconds. The cycle was finished with an extended elongation step at 72°C for 5 minutes.

4.2.3 Plasmid library construction using the CPEC method

The commercial pET21a(+) vector (EMD Millipore) was modified in following ways: i) the sc101 origin of replication, with a copy number of less than five, replaced the pBR322 origin to minimize the effect of plasmid copy number on expression level, ii) all restriction sites are immediately downstream of the ribosome binding site, and iii) the T7-tag and His-tag site were removed. The pET21a(+) vector is under control of IPTG induction. In this system, the T7 RNA polymerase (T7 RNAP) is coded in the BL21(DE3) chromosome placed under the control of an IPTG-inducible lac promoter. The modified vector was linearized by PCR. The AcGFP1 gene library was cloned into the modified vector using the CPEC cloning method (126). 250ng of the linear modified pET21a(+) was mixed with the AcGFP1 gene library insert at a 1:2 molar ratio in the 25µl CPEC reaction using Phusion HF polymerase. The thermal cycling profile follows: an initial denaturation at 98°C, 20 cycles which consisted of 10 seconds of denaturation, annealing

at 55°C for 30 seconds, and extension at 72°C for 2 minutes and finished with an extended elongation step at 72°C for 5 minutes. The CPEC product was examined by 1% agarose gel electrophoresis.

4.2.4 High-throughput AcGFP protein screening using FACS

Three microliters of the successfully cloned product was directly transformed into BL21(DE3) competent cells. The entire transformed population was added to 50 mL of Lysogeny Broth (LB) medium containing 50 µg/mL of ampicillin antibiotic. These cells were incubated at 37°C shaking at 275rpm until log phase ($O.D_{600} = 0.6$); optical densities of cultures are measured with plate reader (Tecan GENios Pro Plate Reader). The AcGFP fluorescence was induced by adding 0.5mM IPTG followed by growth at 30°C on a rotor for 4 hours. GFP fluorescence was found to be higher at 30°C due to improved maturation and function (127). After AcGFP gene induction, 10^6 cells of the bacterial cultures were sorted into lowest and highest 20% of the library population with the FACS Aria II (BD Biosciences) based on relative green fluorescence levels. These populations are labeled “high” and “low”. The sorted populations were added to fresh LB and then incubated at 37°C in the presence of ampicillin until mid-log phase ($O.D_{600} = 0.6$). A final concentration of 0.5mM of IPTG was added into the sorted cells and incubated at 30°C for three hours. The fluorescence levels of 10^6 cells of the high and low sorted groups were analyzed by the FACSCanto flow cytometer (BD Biosciences).

The DNA plasmids of the high and low sorted populations were extracted using a plasmid miniprep kit (Qiagen). The AcGFP genes of the high and low gfp population were amplified with 6 cycles of 10 seconds of denaturation, annealing at 55°C for 30 seconds, and extension at 72°C for 15 seconds to increase the number of gene copies necessary for PacBioRS Next-gene sequencing library preparation. The PacBio RS library preparation and sequencing was done at GCB Duke University Sequencing Facility.

4.2.5 Quantitative real-time PCR (Q-PCR)

To test the transcriptional effect on high and low AcGFP expressed strains, Q-PCR was performed using the Power SYBR Green RNA-to-CT 1-Step Kit and the StepOne Real-Time PCR system according to the manufacturer's instructions (Applied Biosystems, Grand Island, NY, USA). Primers specific for high and low AcGFP are listed in Appendix 6. Primers against *E. coli* 16S rRNA were used to quantify the amount of bacteria in each sample. (**Appendix E**)

4.3 Results

4.3.1 Combinatorial gene library construction and the codon usage calculation matrix

Codon optimization is a method of using synonymous codons to eliminate features in the protein-coding sequence that may inhibit efficient protein expression. Many

publically available codon optimization software use multiple criteria to modulate protein expression level; the software use host favored codons (128), totally eliminate rare codons, and adjust GC content. They also eliminate significant lengths of repetitive sequences, unfavorable mRNA structures, and many others depending on the biological systems used and other specific concerns. Because of high number of permutations for synthetic genes, our in-house codon design software was used incorporate synonymous mutations using wobble positions. We synthesized and cloned the AcGFP gene library into the pET21a(+) vector using both the PCA and CPEC methods (**Figure28**).

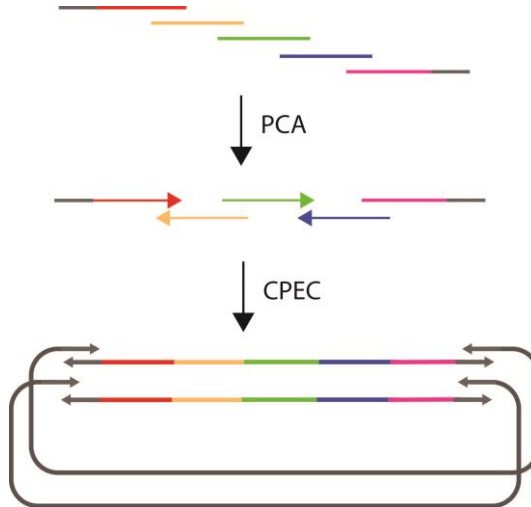


Figure 28: Combinatorial gene library cloning steps

To minimize the influence of other factors on protein expression level, we replaced the pBR322 replication origin, with a copy number of 15-20, with the pSC101

replication origin, with a copy number of 1-5. The AcGFP1 gene library was transformed in BL21(DE3) competent cells and expression was induced with 0.5mM IPTG. For high throughput screening, 10^6 cells expressing AcGFP1 were sorted by FACS into the lowest and highest 20% of green fluorescence levels. The screened high and low populations were sequenced with the PacBio RS next generation sequencer. PacBio RS has 99% sequencing accuracy with DNA molecules less than 1kb. Randomly introduced sequencing error can also be corrected by using the consensus of repeated base reads, known as Circular Consensus Sequencing (CCS). Thus, reads have higher accuracy rates with higher CCS reads.

4.3.2 High-throughput screening and NGS data analysis approach

We collected 36,334 high-GFP-expression and 33,031 low-GFP-expression sequences from PacBio RS sequencing reads. To harvest high accuracy reads, we extracted sequence reads that had more than 10 CCS reads and sequencing lengths greater than 95% of AcGFP1's length ($720\text{bp} \times 0.95 = 684\text{bp}$). After this filtering process, 23,397 high-expression and 21,209 low-expression reads remained. Even filtering for 10 consensus reads, we still observed insertion and deletion mutations in both groups of sequence data. We assumed that indel mutations were sequencing errors because the majority of sorted populations expressed fluorescence. Indel mutations would have caused frameshifts that would have prevented the expression of functional AcGFP.

Therefore, the indel mutations were replaced by WT AcGFP bases and unique sequences were consolidated for following calculations; 4419 high and 4498 low gfp expressed sequences were remained.

The following metrics are various representations of codon usage: codon adaptation index (CAI) (55), frequency of optimal codons (Fop) (61), codon bias index (CBI), and the effective number of codons (ENc)(129). Preferred codons in an organism also strongly correlates with the abundance of iso-accepting tRNA because of increased translational processing rates (130) and thereby this relation is quantified with the tRNA Adaptation Index, tAI (57). All of these metrics were calculated using the CodonW software (<http://codonw.sourceforge.net/>).

Synonymous mutations can also alter mRNA secondary structure. Certain secondary structures can affect the accessibility of mRNA sequences to ribosomes and regulatory molecules. Therefore, mRNA stability can lead to changes in protein level (131-134). The complexity of the mRNA structure can be represented as free folding energy (FFE) and can be calculated with the unified nucleic acid folding and hybridization package (UNAFold) (135).

We calculated those codon usage metrics and mRNA sequences of whole sequences in high and low expression sequences. Because we screened our gene library into high and low expression level, we calculated the average value of each metrics. But the average values were nearly the same between the high and low expressed sequences.

We then investigated the similarity of sequences within each group because a high frequency of similar sequences presented on each population could skew the calculated metric values. We found out that each sequences of high and low expression sequences were quite similar, so we clustered sequences that differ less than 25 bases (**Figure 29**). After the clustering and consolidating similar sequences, 658 sequences from the high expression group and 596 sequences from the low expression group remained.

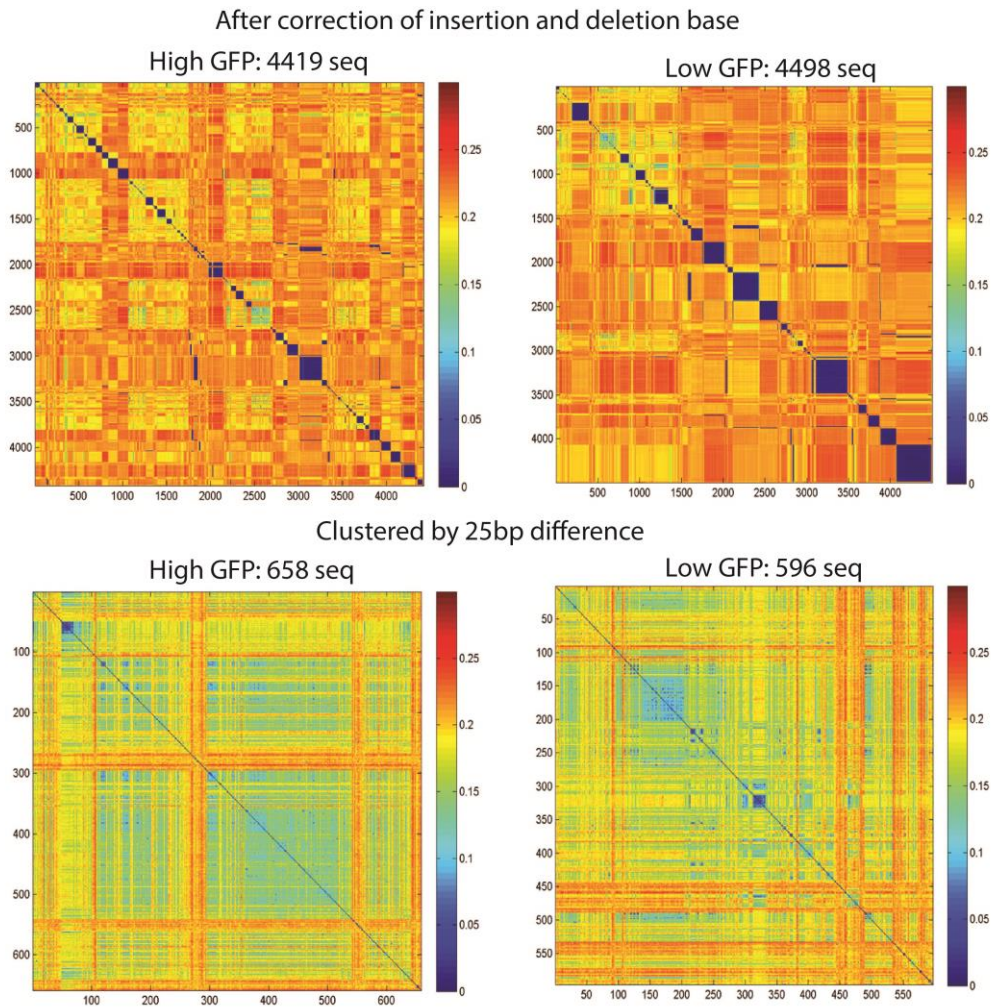


Figure 29: PacBioRS sequence data filtering for high and low expression groups

A) Sequence similarity between sequences in high and low expressed AcGFP sequence respectively after the correction of indels in the AcGFP gene library sequences. Red points represent low similarity and blue points represent high similarity between each sequence. B) Sequence similarity of high and low expression sequences after clustering and consolidating sequences that have less than 25 mismatch bases.

The clustered sequences were calculated by all the codon usage metrics and mRNA folding energy, but global level of codon usage or mRNA folding energy structures did not explain protein expression level. A previous studies claimed that mRNA structure

downstream of start codon is significantly correlated to protein expression level (52, 54) especially nt -4 to +37bp. We created sliding window length of 42bp of individual sequences to examine regional impact of codon usage and mRNA structure (**Figure 30**). We also observed more relatively weak mRNA secondary structure in high gfp expressed group compared to low gfp expressed group in N-terminal regions which are consistent with the previous studies (52, 54). Complex mRNA secondary structure may in N-terminal region could have delay on translation initiation to slow down overall protein level.

GC composition is also correlated to protein expression level and is even measured by local GC position at codon sites. Because many of wobbly codon locations are either second or third site of codons, codon variant sequences would have different GC2 and GC3 could be an indicator of synonymous codon usage (136). In our data, high expressed gfp sequences have low GC3 composition rate compared to low gfp expressed sequences. Considering weaker mRNA secondary structure of high gfp expressed sequences in N-terminal, the low GC3 composition in N-terminal region would contribute for weaker binding structure of mRNA.

The commonly followed recommendation is to replace codons that are rarely found in highly expressed host genes with more favorable codons throughout the target gene (137-140). This codon usage is indicated by the CAI, codon adaptation index, refers to geometric mean of relative codon adaptiveness through the sequence, range from

zero to one. The CAI value close to one means that favorable codons are used to encode a protein. However, on the CAI values of the high gfp expressed sequences are smaller than the CAI value of low gfp expressed sequences. The sliding window analysis indicates that high gfp expression sequences even encoded with not preferred codons compared to low groups. In addition, the CAI values between high and low sequence are overlapped each other in many regions. So, the codon bias usage did not contribute for expression level.

The favorably used codons in native host genes positively correlates with abundant tRNA availability. Similar trend was observed in the tAI value. At translation initiation, low gfp expressed sequences have more available tRNA source, but this environment did not result in high protein yields. As proved in the previous study, mRNA secondary structure near to the start codon site a key factor to optimize the expression level (52, 54). .

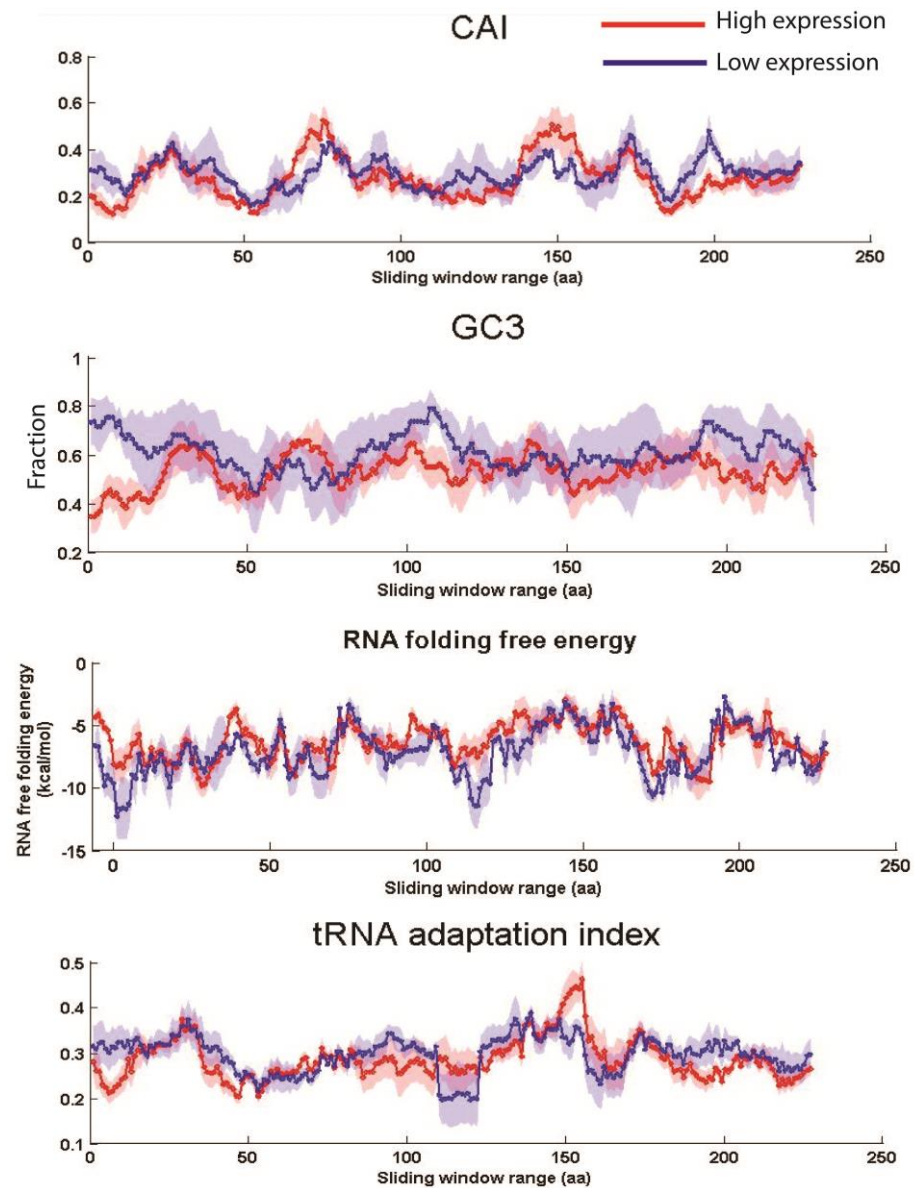


Figure 30: Sliding window analysis of CAI, GC3, mRNA folding energy and tAI of high and low AcGFP expressed population.

Red color represents high expressed and blue color represent low expressed AcGFP sequence. 2 standard deviation was displayed as light red and blue color. The sliding window length of 42nt.

4.3.3 Regional effect of AcGFP protein expression level

Strong influence of mRNA secondary structure N-terminal region on protein level, we speculated how N-terminal regions are independent from the downstream sequences. To investigate regional impact on protein expression level, we swapped sequences near to the start codon sites between high, low, and WT AcGFP expression sequences. Top 5 abundant sequences in each high and low expressed sequences were pooled, synthesized and cloned into pET21+a vector. We also confirmed that high expressed AcGFP sequences expressed very bright level of gfp fluorescence that could visually determine by bacterial pellet (Figure 31).

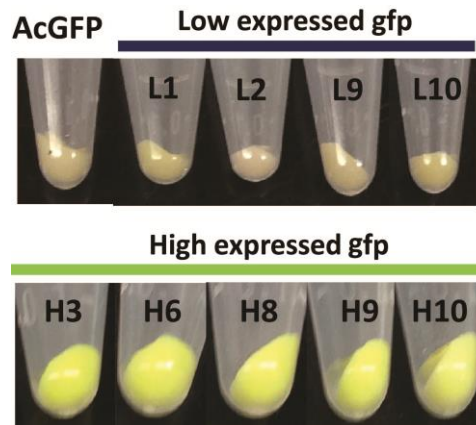


Figure 31: Quantification of translation levels of the top five most abundant sequence reads from the high, low and WT AcGFP expression groups.

AcGFP expression level displayed by bacteria pellet images of high, low and WT AcGFP sequences after 4 hours of incubation at 30°C following induction of 0.5mM of IPTG.

If mRNA folding energy near the start codon sites, nt -4 to +37bp (52, 54), solely determine the overall protein expression level, protein expression level would not

change depending on the downstream sequences. To test N-terminal sequence effect on bacterial protein expressions, we exchanged N-terminal regions between high, low and WT AcGFP expression sequences; swap regions are nt +1 through + 36bp and nt +1 through +75 bp (**Figure 1A**). Total 5 strains were used for the N-terminal region swap; 2 strains (labeled as H3, H9) from the 5 abundant high gfp expressed sequences, 2 strains (labeled as L2, L10) from the 5 abundant low gfp expressed sequences, and WT AcGFP sequences. We calculated mRNA free folding energy of the 5 strains using mfold software(141), to see mRNA secondary structure effect. In the -5 through +36nt region, the H3 and H9 strains have relatively weak mRNA structures, FFE = -8.63 and -9.87 kcal/mol, compared to L2, L10 and WT AcGFP strains, FFE = -11.3, -10.82 and -15.46 kcal/mol. Although high expressed sequence have higher folding energy, H9 and L10 sequences have only subtle difference in the folding energy. The subtle difference might lead a big change in expression level.

From start codon site, 36 and 75bp of the 5 sequences were swapped and the fluorescence were induced by IPTG (**Figure 1B**). Our experiment result shows that overall protein expression level could be predicted by the N-terminal sequences in many case. Sequences replaced N-terminal region with H3 sequenced generally enhanced gfp protein expression. However, the N-terminal sequence effect are not applicable for all mix-matched sequences. H9 sequence produce bring gfp protein level, but 36bp of N-terminal sequence of H9 fused with L10 sequence attenuated the protein expression

level compared to H9 sequence. If first 36bp of sequence determine the overall protein expression level, the downstream sequences should be independent. Our swap experiment results showed that optimization of protein level were also influenced by the sequences downstream of +36nt.

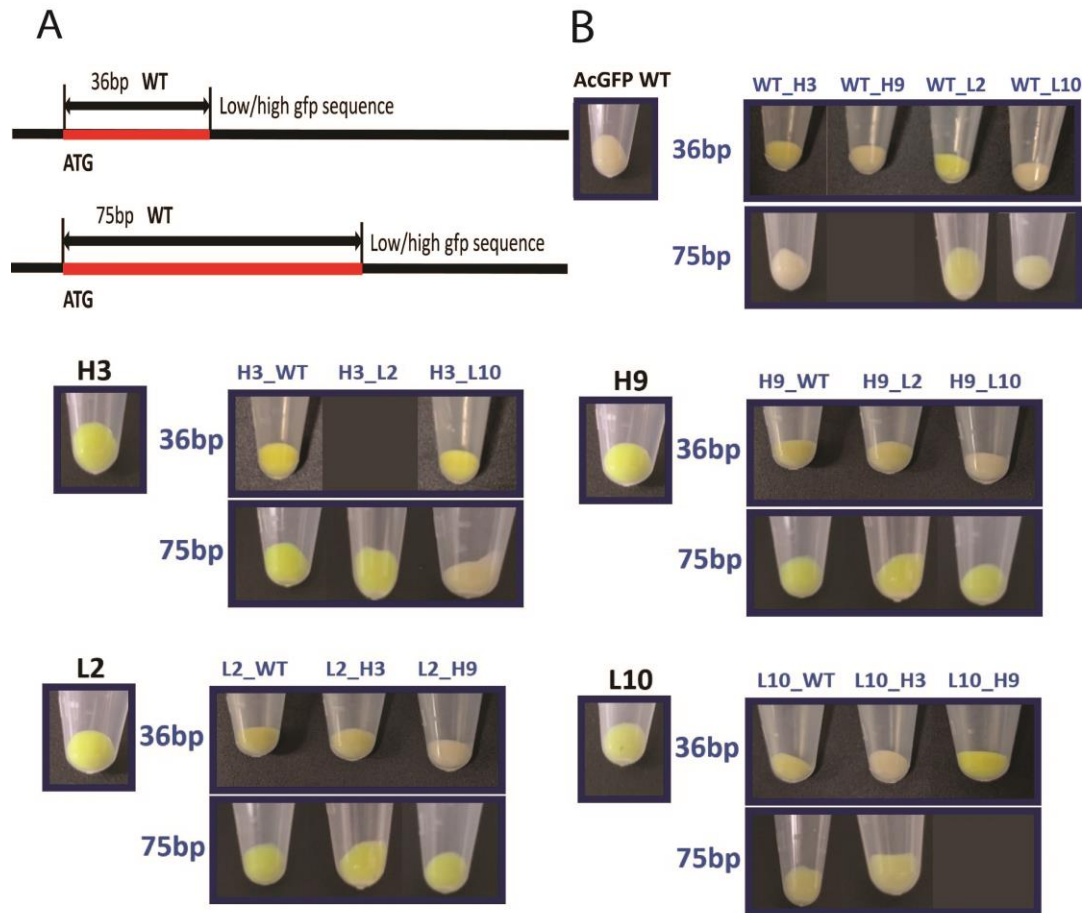


Figure 32: AcGFP protein expression levels of sequences with swapped regions between the low, high and WT AcGFP expression groups.

A) Schematic diagram of sequences with different lengths of swapped regions between high, low and WT AcGFP sequences. B) Comparison of AcGFP expression levels among displayed by bacteria pellet images of high, low and WT AcGFP sequences after 4 hours of incubation at 30°C following induction of 0.5mM of IPTG; beginning of 36bp and 72bp between high, low, and AcGFP were swapped and measured the corresponding expression levels. 36bp WT_H3 is

composed of nt +1 to +36bp of WT AcGFP1 and +37 to end of H3, high expressed AcGFP1 sequence.

To have mechanical understanding of synonymous mutational effect on protein expression, we designed, synthesized and construct AcGFP codon variant library. We applied high throughput screening and sequencing approach for population based codon usage difference between the high and low gfp expressed sequences. However, no obvious codon bias was not observed either the high or low groups. From sliding windows analysis, we confirmed crucial impact on protein level of mRNA structure near the start codon site and swapped the 5' coding sequences between high, low and WT AcGFP sequence to find other regions or factors modulate protein expression level. Also, codon usage and compositions did not have visible effect on protein level. Also population-based analysis was limited because of complex permutations of sequences pool in both the high and low gfp sequences. Individual quantification of protein level and corresponding codon-variant sequencing would be useful to develop a principle of synthetic gene design.

5. Future directions

5.1 Development of a statistical framework to identify essential genes based on chemical mutagenesis

To identify the essential gene set of a microbe when molecular genetic tools are not available, we developed a gene by gene analytical method. We sequenced over 8,000 EMS induced mutants and identified 902 candidate essential genes in *E. acetylicum*. About a quarter of the candidate genes were uncharacterized. Therefore, the essentiality of the hypothetical genes cannot be corroborated by comparative analysis to essential genes of closely related microbes or conserved genes among the same phylum.

A potential area for future work is to develop a statistical method which accounts for the functional impact of individual mutations on a gene. Essential genes would not be able to harbor high-impact, non-neutral mutations. Thus, identification of a minimum impact-score of non-neutral mutations would be helpful to derive the essentiality of gene. The statistical method would weigh the functional impact score of each mutation, the location of the gene, and the conservation of DNA sequences among Firmicutes species into the model.

Chemically induced mutants could harbor having high-impact non-neutral mutations on essential genes. Although essential genes have low tolerance to high-impact non-neutral mutations, but not all non-neutral mutations are lethal for cellular functions. Our mutants survived after chemical mutagen treatment, so non-neutral mutations on essential genes are still tolerable even though amino-substitutions heavily

influence the cellular process. Possibility of harboring high-impact non-neutral mutations on essential genes might hinder the identification of essential genes. We should develop a statistical model to assess the impact of individual mutations is required to identify essential genes.

5.2 Establish a sequence analysis method to identify mutations that are depleted after purifying selection

Traditional genetic manipulation tools create gene knockout strains by randomly introducing DNA inserts (e.g. transposons) into bacteria populations. What genes are required for growth in different mediums and conditions is determined by monitoring at a population level what genes do not bear DNA inserts. Although this method clearly determines essential gene *in vitro* and *in vivo* (43, 45, 142), but this method could not be applicable for genetically intractable microbe. We will use our mutant library to purify survival clones of zebrafish gut colonization. Then, the genetic variants of survival mutants would be identified by whole genome sequencing. During this selection scheme, mutations for critically defective host gut colonization would be depleted and would not be observed. Both enriched and depleted variants after the purifying selection would elucidate the genetic requirements for intestinal colonization.

Appendix A

Putative *E. acetylicum* motility genes identified by reciprocal BLAST analysis

Gene ID*	Gene Name	Predicted protein
121		'putative PIN and TRAM-domain containing protein precursor'
187	'ylxH_1'	'Flagellum site-determining protein YlxH'
296	'mcpB_1'	'Methyl-accepting chemotaxis protein McpB'
335	'hemAT_1'	'Heme-based aerotactic transducer HemAT'
357	'bdlA_1'	'Biofilm dispersion protein BdlA'
381	'mcp2'	'Methyl-accepting chemotaxis protein 2'
434	'cysL_1'	'HTH-type transcriptional regulator CysL'
446	'phoR_2'	'Alkaline phosphatase synthesis sensor protein PhoR'
509		'LemA family protein'
628	'smc_1'	'Chromosome partition protein Smc'
635	'corC_1'	'Magnesium and cobalt efflux protein CorC'
640	'mepM_1'	'Murein DD-endopeptidase MepM'
641	'mepM_2'	'Murein DD-endopeptidase MepM'
692	'ypdA_2'	'Sensor histidine kinase YpdA'
704	'liaS'	'Sensor histidine kinase LiaS'
747	'phoR_4'	'Alkaline phosphatase synthesis sensor protein PhoR'
829	'scrB'	'Sucrose-6-phosphate hydrolase'
830	'gmr_3'	'Cyclic di-GMP phosphodiesterase Gmr'
832	'yxlF_1'	'putative ABC transporter ATP-binding protein YxlF'
854	'yycG_1'	'Sensor histidine kinase YycG'
913	'mcpB_2'	'Methyl-accepting chemotaxis protein McpB'
959	'kinA'	'Sporulation kinase A'
1002	'xpsE'	'Type II secretion system protein E'
1003		'type IV pilin biogenesis protein'
1004	'xcpT'	'Type II secretion system protein G precursor'
1005		'hypothetical protein'
1006		'hypothetical protein'
1033	'recN'	'DNA repair protein RecN'
1131	'cheV'	'Chemotaxis protein CheV'
1138	'cysA'	'Sulfate/thiosulfate import ATP-binding protein CysA'
1164	'srrB'	'Sensor protein SrrB'
1221	'drrA_1'	'Daunorubicin/doxorubicin resistance ATP-binding protein DrrA'
1296		'Fluoroquinolones export ATP-binding protein/MT2762'
1400	'yxlF_2'	'putative ABC transporter ATP-binding protein YxlF'
1467	'mcpB_3'	'Methyl-accepting chemotaxis protein McpB'

1647		'Fluoroquinolones export ATP-binding protein/MT2762'
1695	'bdIA_2'	'Biofilm dispersion protein BdlA'
1717	'drrA_4'	'Daunorubicin/doxorubicin resistance ATP-binding protein DrrA'
1732	'bdIA_3'	'Biofilm dispersion protein BdlA'
1737	'mcpA'	'Methyl-accepting chemotaxis protein McpA'
1761	'hemAT_2'	'Heme-based aerotactic transducer HemAT'
1854	'bdIA_4'	'Biofilm dispersion protein BdlA'
1957	'gltR_2'	'HTH-type transcriptional regulator GltR'
2066	'cph2_6'	'Phytochrome-like protein cph2'
2140		'tetratricopeptide repeat protein'
2144	'cheR'	'Chemotaxis protein methyltransferase'
2157		'PilZ domain protein'
2174	'spoIIIE'	'DNA translocase SpoIIIE'
2201	'sigD'	'RNA polymerase sigma-D factor'
2202	'cheD'	'Chemoreceptor glutamine deamidase CheD'
2203	'cheC'	'CheY-P phosphatase CheC'
2204	'cheW'	'Chemotaxis protein CheW'
2205	'cheA'	'Chemotaxis protein CheA'
2206	'ylxH_2'	'Flagellum site-determining protein YlxH'
2207	'flhF'	'Flagellar biosynthesis protein FlhF'
2208	'flhA'	'Flagellar biosynthesis protein FlhA'
2209	'flhB_1'	'Flagellar biosynthetic protein FlhB'
2210	'fliR'	'Flagellar biosynthetic protein FliR'
2211	'fliQ'	'Flagellar biosynthetic protein FliQ'
2212	'fliP'	'Flagellar biosynthetic protein FliP precursor'
2214	'cheY'	'Chemotaxis protein CheY'
2215	'fliN'	'Flagellar motor switch protein FliN'
2216	'fliM'	'Flagellar motor switch protein FliM'
2218		'Flagellar protein (FlbD)'
2219	'flgG_1'	'Flagellar basal-body rod protein FlgG'
2220	'flgD'	'Basal-body rod modification protein FlgD'
2221		'Flagellar hook-length control protein FliK'
2222		'MgtE intracellular N domain protein'
2224	'yscN'	'putative ATP synthase YscN'
2225		'flagellar assembly protein H'
2226	'fliG'	'Flagellar motor switch protein FliG'
2227	'fliF'	'Flagellar M-ring protein'
2229	'flgC'	'Flagellar basal-body rod protein FlgC'
2230	'flgB'	'Flagellar basal body rod protein FlgB'
2370	'gltR_3'	'HTH-type transcriptional regulator GltR'

2399	'mecA'	'Adapter protein MecA 1'
2410	'dppE_3'	'Dipeptide-binding protein DppE precursor'
2456	'tlyC_3'	'Hemolysin C'
2463	'minD'	'Septum site-determining protein MinD'
2471		'Competence protein A'
2472	'comC'	'Type 4 prepilin-like proteins leader peptide-processing enzyme'
2473	'pulG'	'Type II secretion system protein G precursor'
2474	'epsF'	'Type II secretion system protein F'
2475	'pilT'	'Twitching mobility protein'
2476	'epsE'	'Type II secretion system protein E'
2479		'hypothetical protein'
2480		'hypothetical protein'
2481		'hypothetical protein'
2483		'Bacterial Ig-like domain (group 2)'
2493	'icaA'	'Poly-beta-1%2C6-N-acetyl-D-glucosamine synthase'
2500	'tig'	'Trigger factor'
2556	'phoR_9'	'Alkaline phosphatase synthesis sensor protein PhoR'
2599		'hypothetical protein'
2696	'metN_2'	'Methionine import ATP-binding protein MetN'
2739	'rbsB_2'	'D-ribose-binding periplasmic protein precursor'
2765		'tetratricopeptide repeat protein'
2775	'minJ'	'Cell division topological determinant MinJ'
2799	'fliS'	'Flagellar protein FliS'
2800	'fliD'	'Flagellar hook-associated protein 2'
2802	'hag_2'	'Flagellin'
2819		'hypothetical protein'
2820	'fliW'	'Flagellar assembly factor FliW'
2822	'flaB2'	'Flagellar filament 33 kDa core protein'
2823	'flgK'	'Flagellar hook-associated protein 1'
2825		'anti-sigma28 factor FlgM'
2830	'degU'	'Transcriptional regulatory protein DegU'
3026	'ywqD'	'Tyrosine-protein kinase YwqD'
3037	'gtfC'	'Glucosyltransferase-SI precursor'
3056		'pheromone autoinducer 2 transporter'
3059	'flgG_2'	'Flagellar basal-body rod protein FlgG'
3060	'flgF'	'Flagellar basal-body rod protein FlgF'
3070	'atpA'	'ATP synthase subunit alpha'
3083	'hssS'	'Heme sensor protein HssS'
3096		'Telomeric repeat-binding factor 2'
3101	'yohaH_2'	'Putative methyl-accepting chemotaxis protein YoaH'

3153	'phoR_11'	'Alkaline phosphatase synthesis sensor protein PhoR'
3173	'prtP'	'PII-type proteinase precursor'
3224	'aes'	'Acetyl esterase'
3255	'ydfJ'	'Membrane protein YdfJ'
3287	'pomA'	'Chemotaxis protein PomA'
3288	'motB'	'Motility protein B'
3292	'phoR_12'	'Alkaline phosphatase synthesis sensor protein PhoR'
3303	'swrC_2'	'Swarming motility protein SwrC'
3339	'rpfG_2'	'Cyclic di-GMP phosphodiesterase response regulator RpfG'
3356	'bspRIM'	'Modification methylase BspRI'
3383	'ytrB_4'	'ABC transporter ATP-binding protein YtrB'

*In NCBI or JGI, *E. acetylicum* loci is identified as "EZWU0009_" followed by five digit number. NCBI Gene IDs can be searched by adding prefix "EZWU0009_" and suffix "0" to the Gene ID number (e.g. Gene ID number 126 = EZWU0009_01260).

Appendix B

Genes identified as potential motility genes by Mutational Enrichment Analysis after Phenotypic Selection (MEAPS)

Gene ID~	Gene Name	Protein product	NS el No ns yn	NSel Syn	Sel Nonsy n	Sel Syn	Corre cted Nons yn#	Nons ense Muta tions	Poten tial Nons ense	Rel. Nonsen se Freq	Nonse nse Freq score	Fisher exact test pvalue	Fisher Scor es	NS X F Score &	Reciproca l BLAST	Pfam
2224	'yycN'	'putative ATP synthase YycN'	1	0	28	0	27	4	30	0.13	5	#####	5	25	M	
2830	'degU'	'Transcriptional regulatory protein DegU'	0	0	16	0	16	2	12	0.17	5	#####	5	25	M	M
2227	'fliF'	'Flagellar M-ring protein'	1	0	11	3	10	10	42	0.24	5	0.007	5	25	M	M
2221	[]	'Flagellar hook-length control protein FliK'	0	0	7	0	7	7	42	0.17	5	0.015	5	25	M	M
2208	'flhA'	'Flagellar biosynthesis protein FlhA'	5	0	17	2	12	1	41	0.07	4	0.001	5	20	M	M
2219	'flgG_1'	'Flagellar basal-body rod protein FlgG'	1	0	10	0	9	1	16	0.06	4	0.009	5	20	M	M
2215	'fliN'	'Flagellar motor switch protein FliN'	2	0	10	1	8	3	30	0.1	4	0.011	5	20	M	M
2209	'flhB_1'	'Flagellar biosynthetic protein FlhB'	0	0	6	0	6	3	29	0.1	4	0.022	5	20	M	M
2793	'hag_1'	'Flagellin'	0	0	4	0	4	4	22	0.18	5	0.062	4	20		M
2226	'fliG'	'Flagellar motor switch protein FliG'	0	0	4	1	4	2	27	0.07	4	0.062	4	16	M	M
2862	'dosP'	'Oxygen sensor protein DosP'	2	3	13	2	11	3	97	0.04	3	0.004	5	15		M
2201	'sigD'	'RNA polymerase sigma-D factor'	2	0	9	0	7	1	20	0.05	3	0.02	5	15	M	M
2211	'fliQ'	'Flagellar biosynthetic protein FliQ'	0	0	3	0	3	3	14	0.21	5	0.124	3	15	M	M
2213	[]	'Flagellar biosynthesis protein%2C FliO'	0	1	3	0	3	3	14	0.21	5	0.125	3	15		M
3052	[]	'Sortase family protein'	0	0	3	0	3	1	8	0.13	5	0.125	3	15		
2800	'fliD'	'Flagellar hook-associated protein 2'	0	0	5	1	5	2	43	0.05	3	0.031	4	12	M	M
2225	[]	'flagellar assembly protein H'	2	0	5	2	3	4	35	0.11	4	0.125	3	12	M	M
2824	[]	'FlgN protein'	0	0	3	0	3	2	19	0.11	4	0.125	3	12		M
2212	'fliP'	'Flagellar biosynthetic protein FliP precursor'	1	0	4	0	3	1	8	0.13	5	0.216	2	10	M	M
2883	[]	'hypothetical protein'	0	0	4	0	4	1	29	0.03	2	0.062	4	8		

2216	'fliM'	'Flagellar motor switch protein FliM'	2	1	6	0	4	0	32	0.03	2	0.078	4	8	M	M
3038	[]	'Bacterial Ig-like domain (group 4)'	1	6	4	0	3	1	38	0.03	2	0.124	3	6		
1128	'rsbP'	'Phosphoserine phosphatase RsbP'	0	0	3	0	3	1	25	0.04	3	0.204	2	6		M
1442	'hmp_1'	'Flavo-hemoprotein'	1	0	4	0	3	1	19	0.05	3	0.248	2	6		
2802	'hag_2'	'Flagellin'	0	1	24	1	24	0	23	0	1	#####	5	5	M	M
2214	'cheY'	'Chemotaxis protein CheY'	1	1	9	0	8	0	4	0	1	0.012	5	5	M	M
2228	'fliE'	'Flagellar hook-basal body complex protein FliE'	0	0	2	0	2	2	9	0.22	5	0.499	1	5		M
338	'copA_1'	'Copper-exporting P-type ATPase A'	1	0	7	1	6	0	45	0	1	0.031	4	4		M
302	'ykoD_1'	'Putative HMP/thiamine import ATP-binding protein YkoD'	0	0	5	0	5	1	63	0.02	1	0.039	4	4		M
2831	'degS'	'Signal transduction histidine-protein kinase/phosphatase DegS'	1	0	6	1	5	0	29	0	1	0.062	4	4		M
3100	'alsT_2'	'Amino-acid carrier protein AlsT'	1	1	6	1	5	0	27	0	1	0.062	4	4		
1441	[]	'phosphoglycolate phosphatase'	0	0	4	1	4	0	15	0	1	0.062	4	4		
1418	[]	'peptidase T'	0	0	4	1	4	0	44	0	1	0.062	4	4		
3439	'yycG_5'	'Sensor histidine kinase YycG'	0	2	4	0	4	0	40	0	1	0.063	4	4		M
3426	[]	'hypothetical protein'	1	3	5	0	4	0	26	0	1	0.069	4	4		
2235	'trmFO'	'Methylenetetrahydrofolate--tRNA-(uracil-5-)-methyltransferase TrmFO'	0	1	4	2	4	0	23	0	1	0.077	4	4		
1432	[]	'hypothetical protein'	0	1	3	0	3	1	41	0.02	2	0.218	2	4		
985	'pstB3'	'Phosphate import ATP-binding protein PstB 3'	0	0	2	1	2	1	16	0.06	4	0.451	1	4		M
2778	'ftsX'	'Cell division protein FtsX'	2	2	4	0	2	1	14	0.07	4	0.499	1	4		M
2619	[]	'hypothetical protein'	3	1	5	1	2	4	41	0.1	4	0.5	1	4		
2323	[]	'hypothetical protein'	1	0	5	1	4	0	28	0	1	0.124	3	3		
527	'rbsC_1'	'Ribose transport system permease protein RbsC'	0	1	4	0	4	0	17	0	1	0.124	3	3		
2665	[]	'NADH dehydrogenase-like protein'	0	1	4	0	4	0	20	0	1	0.124	3	3		M
3348	[]	'hypothetical protein'	0	1	4	0	4	0	28	0	1	0.124	3	3		
2275	'fmt'	'Methionyl-tRNA formyltransferase'	0	0	4	0	4	0	25	0	1	0.124	3	3		

2410	'dppE_3'	'Dipeptide-binding protein DppE precursor'	0	0	4	0	4	0	35	0	1	0.124	3	3	M	M
2839	'lytC'	'N-acetylmuramoyl-L-alanine amidase LytC precursor'	0	2	4	1	4	0	23	0	1	0.124	3	3		
838	'glpD_2'	'Aerobic glycerol-3-phosphate dehydrogenase'	0	1	4	1	4	0	30	0	1	0.124	3	3		M
1823	[]	'ATP-dependent helicase HepA'	1	4	5	4	4	0	86	0	1	0.124	3	3		
6	'gyrA'	'DNA gyrase subunit A'	6	3	10	1	4	0	37	0	1	0.124	3	3		
2252	'smc_2'	'Chromosome partition protein Smc'	3	3	7	2	4	0	113	0	1	0.124	3	3		
2302	'divIVA'	'Septum site-determining protein DivIVA'	0	0	2	1	2	1	17	0.06	3	0.498	1	3		
2822	'flaB2'	'Flagellar filament 33 kDa core protein'	1	0	3	1	2	1	19	0.05	3	0.5	1	3	M	M
688	[]	'SNARE associated Golgi protein'	0	0	2	0	2	1	23	0.04	3	0.512	1	3		
928	'tagH_1'	'Teichoic acids export ATP-binding protein TagH'	0	0	2	0	2	1	19	0.05	3	0.53	1	3		M
2844	[]	'Acyltransferase family protein'	1	1	3	1	2	1	28	0.04	3	0.53	1	3		
2437	'aspS'	'Aspartate--tRNA ligase'	0	7	3	0	3	0	32	0	1	0.206	2	2		
1957	'glrR_2'	'HTH-type transcriptional regulator GlrR'	0	0	3	2	3	0	19	0	1	0.218	2	2	M	M
814	[]	'hypothetical protein'	0	0	3	2	3	0	42	0	1	0.218	2	2		
1885	'czcB'	'Cobalt-zinc-cadmium resistance protein CzcB'	0	0	3	1	3	0	21	0	1	0.218	2	2		M
524	'mtnA'	'Methylthioribose-1-phosphate isomerase'	0	0	3	1	3	0	30	0	1	0.247	2	2		
80	'tilS'	'tRNA(Ile)-lysine synthase'	0	0	3	1	3	0	47	0	1	0.248	2	2		
314	'yqik'	'Inner membrane protein Yqik'	0	0	3	1	3	0	28	0	1	0.248	2	2		M
216	'rebO'	'Flavin-dependent L-tryptophan oxidase RebO precursor'	0	2	3	0	3	0	38	0	1	0.248	2	2		
1737	'mcpA'	'Methyl-accepting chemotaxis protein McpA'	0	0	3	1	3	0	53	0	1	0.248	2	2	M	M
2895	'cph2_7'	'Phytochrome-like protein cph2'	2	0	5	2	3	0	43	0	1	0.249	2	2		M
2157	[]	'PilZ domain protein'	1	0	4	1	3	0	25	0	1	0.249	2	2	M	M
3043	'sdrI'	'Serine-aspartate repeat-containing protein I precursor'	1	0	4	1	3	0	23	0	1	0.249	2	2		
352	'ypdA_1'	'Sensor histidine kinase YpdA'	1	0	4	1	3	0	58	0	1	0.249	2	2		M
2958	'yfkN_3'	'Trifunctional nucleotide phosphoesterase protein YfkN precursor'	1	2	4	0	3	0	24	0	1	0.249	2	2		M

827	'pcrA_2'	'ATP-dependent DNA helicase PcrA'	1	2	4	0	3	0	56	0	1	0.249	2	2		
2415	'greA'	'Transcription elongation factor GreA'	0	0	3	0	3	0	5	0	1	0.249	2	2		
2808	[]	'hypothetical protein'	0	0	3	0	3	0	7	0	1	0.249	2	2		
616	'pspA_1'	'Phosphoserine phosphatase 1'	0	0	3	0	3	0	14	0	1	0.249	2	2		
2207	'flhF'	'Flagellar biosynthesis protein FlhF'	0	1	3	2	3	0	24	0	1	0.249	2	2	M	M
2779	'ftsE'	'Cell division ATP-binding protein FtsE'	0	0	3	0	3	0	13	0	1	0.249	2	2		M
1785	[]	'hypothetical protein'	0	1	3	2	3	0	42	0	1	0.249	2	2		
3021	'gmd'	'GDP-mannose 4%2C6-dehydratase'	0	0	3	0	3	0	24	0	1	0.249	2	2		
2183	'truB'	'tRNA pseudouridine synthase B'	0	1	3	0	3	0	15	0	1	0.249	2	2		
2171	'tmpC_2'	'Membrane lipoprotein TmpC precursor'	0	0	3	0	3	0	18	0	1	0.249	2	2		
93	'dusC'	'tRNA-dihydrouridine synthase C'	0	0	3	0	3	0	26	0	1	0.249	2	2		
556	'yerA'	'Putative adenine deaminase YerA'	0	1	3	2	3	0	51	0	1	0.249	2	2		
3284	'lgrD'	'Linear gramicidin synthase subunit D'	0	0	3	0	3	0	26	0	1	0.287	2	2		
383	'yedQ_1'	'putative diguanylate cyclase YedQ'	0	1	3	1	3	0	26	0	1	0.296	2	2		M
2062	'gltB'	'Glutamate synthase [NADPH] small chain'	1	1	4	1	3	1	40	0.03	2	0.373	1	2		M
2376	[]	'hypothetical protein'	0	0	2	0	2	1	44	0.02	2	0.548	1	2		M
2875	[]	'hypothetical protein'	0	1	2	0	2	1	37	0.03	2	0.624	1	2		
2914	'lytD'	'Beta-N-acetylglucosaminidase precursor'	4	0	6	1	2	3	86	0.03	2	0.687	1	2		M
2965	'uhpT'	'Hexose phosphate transport protein'	1	0	3	0	2	1	36	0.03	2	0.812	1	2		M
3140	'yvaA'	'putative oxidoreductase YvaA'	2	0	4	0	2	1	35	0.03	2	0.862	1	2		M
2044	'odhB'	'Dihydrolipoyllysine-residue succinyltransferase component of 2-oxoglutarate dehydrogenase complex'	0	0	3	0	3	0	18	0	1	0.324	1	1		M
632	'yfmT'	'Putative aldehyde dehydrogenase YfmT'	0	1	3	0	3	0	30	0	1	0.373	1	1		
265	'emrB'	'Multidrug export protein EmrB'	0	3	3	2	3	0	38	0	1	0.373	1	1		M
2210	'fliR'	'Flagellar biosynthetic protein FliR'	1	0	4	0	3	0	12	0	1	0.374	1	1	M	M
394	'rapA'	'RNA polymerase-associated protein RapA'	1	0	4	0	3	0	76	0	1	0.374	1	1		
3148	'mntB'	'Manganese transport system membrane protein MntB'	1	2	4	3	3	0	10	0	1	0.412	1	1		

2994	[]	'Bacterial dynamin-like protein'	1	1	4	2	3	0	51	0	1	0.413	1	1		M
2804	'dppE_5'	Dipeptide-binding protein DppE precursor'	0	2	3	1	3	0	28	0	1	0.43	1	1		M
2777	'mepM_7'	'Murein DD-endopeptidase MepM'	0	2	3	1	3	0	45	0	1	0.437	1	1		
28	'dnaX_1'	'DNA polymerase III subunit tau'	0	2	3	1	3	0	56	0	1	0.437	1	1		
2045	'odhA'	'2-oxoglutarate dehydrogenase E1 component'	0	2	3	1	3	0	69	0	1	0.437	1	1		
268	'pleD_1'	'Response regulator PleD'	1	1	4	1	3	0	24	0	1	0.437	1	1		M

* Mutations are binned by their impact on amino acid coding: Nonsynonymous (NonSyn) and synonymous (Syn) in selected (Sel) and non-motile selected (Sel) groups.

Color coding of genes reflect overall confidence in identified genes being required for motility. Lighter shades of red indicate decreased confidence & Arbitrary ranking based on combined tests for likelihood of a gene being required for motility (Fisher Exact test and Density of nonsense mutations)

% Two homology based tests, reciprocal BLAST (Appendix A) and Pfam domain, were used to identify potential motility (M) genes. Fifty-six of these genes (8.6%) were among the list of MEAPS-derived putative motility genes.

~ In NCBI or JGI, E. acetylicum loci is indentified as "EZWU0009_" followed by five digit number. NCBI Gene IDs can be searched by adding prefix "EZWU0009_" and suffix "0" to the Gene ID number (e.g. Gene ID number 126 = EZWU0009_01260).

Appendix C

Location of spontaneous suppressor mutations of non-motile *E. acetylicum* strains with nonsense mutations in *ea2862* and *ea2619*

Gene ID ~	Gene name	Protein product	Suppressor Mutation location	Reference ntd	Mutated ntd	Polymorphism type	Amino acid change	Parental non motile strain*	Homology based motility gene&	MEAPS identified motility gene!
2464	minC'	Septum site-determining protein MinC'	2263634	C	A	NON SENSE	E62*	Ea2862 Q817*_1		
2831	degS'	Signal transduction histidine-protein kinase/phosphatase DegS'	2620642	C	G	NON SYN	V289L	Ea2862 Q817*_2		M
2831	degS'	Signal transduction histidine-protein kinase/phosphatase DegS'	2620644	T	C	NON SYN	E288G	Ea2862 Q817*_2		M
2831	degS'	Signal transduction histidine-protein kinase/phosphatase DegS'	2620846	C	T	NON SYN	E221K	Ea2862 Q817*_1		M
2831	degS'	Signal transduction histidine-protein kinase/phosphatase DegS'	2620852	A	T	NON SYN	F219I	Ea2862 Q817*_1		M
2831	degS'	Signal transduction histidine-protein kinase/phosphatase DegS'	2620875	T	C	NON SYN	Q211R	Ea2862 Q817*_1		M
2157		PilZ domain protein'	1965833	G	A	NON SYN	A131V	Ea2862 Q558*_1	M	M
2216	fliM'	Flagellar motor switch protein FliM'	2029533	C	T	NON SYN	E147K	Ea2862 Q558*_1	M	
92	folK'	2-amino-4-hydroxy-6- hydroxymethylidihydropteridine pyrophosphokinase'	79078	A	G	NON SYN	E155G	Ea2862 Q558*_2		
262	tagB_2'	Putative CDP- glycerol:glycerophosphate glycerophosphotransferase'	213659	G	A	NON SYN	A168T	Ea2862 Q558*_2		
756		hypothetical protein'	709681	C	T	NON SYN	A55V	Ea2862 Q558*_2		
869	nplT_1'	Neopullulanase'	814339	A	G	NON SYN	M390T	Ea2862 Q558*_2		
959	kinA'	Sporulation kinase A'	907207	C	T	NON SYN	A453V	Ea2862 Q558*_2	M	
1102		hypothetical protein'	1036712	C	T	NON SENSE	Q199*	Ea2862 Q558*_2		
1358	arsB'	Arsenical pump membrane protein'	1281526	T	C	NON SYN	V129A	Ea2862 Q558*_2		

1394		hypothetical protein'	1319065	G	A	NON SYN	A136V	Ea2862 Q558*_2		
1470		hypothetical protein'	1381009	C	T	NON SYN	A30V	Ea2862 Q558*_2		
1745	ywnH_3'	Putative phosphinothricin acetyltransferase YwnH'	1609946	G	A	NON SENSE	W59*	Ea2862 Q558*_2		
1856		hypothetical protein'	1705498	G	A	NON SENSE	Q94*	Ea2862 Q558*_2		
1891		hypothetical protein'	1732739	G	A	NON SYN	V357I	Ea2862 Q558*_2		
2096		tRNA-specific adenosine deaminase'	1909885	C	T	NON SYN	R151Q	Ea2862 Q558*_2		
2157		PilZ domain protein'	1965869	A	G	NON SYN	F119S	Ea2862 Q558*_2	M	M
2272	prkC'	Serine/threonine-protein kinase PrkC'	2079501	G	A	NON SYN	A155V	Ea2862 Q558*_2		
804		hypothetical protein'	745379	A	G	NON SYN	Q92R	Ea2862 Q558*_3		
967	znuC_1'	High-affinity zinc uptake system ATP-binding protein ZnuC'	916098	A	C	NON SYN	Q82P	Ea2862 Q558*_3		
1427		hypothetical protein'	1342634	A	G	NON SYN	Q32R	Ea2862 Q558*_3		
2190	polC_1'	DNA polymerase III PolC-type'	2003415	T	C	NON SYN	Q960R	Ea2862 Q558*_3		
3020		hypothetical protein'	2833296	A	G	NON SYN	I351T	Ea2862 Q558*_3		
575	mprF'	Phosphatidylglycerol lysyltransferase'	529447	T	A	NON SENSE	L159*	Ea2619 W61*_1		
997	ltaS2'	Lipoteichoic acid synthase 2'	946218	A	G	NON SYN	Q220R	Ea2619 W61*_1		
2214	cheY'	Chemotaxis protein CheY'	2027779	A	C	NON SYN	M16R	Ea2619 W61*_1	M	M
2772		hypothetical protein'	2564376	A	C	NON SYN	V224G	Ea2619 W61*_1		
2114	ponA'	Penicillin-binding protein 1A/1B'	1929226	G	T	NON SYN	A142E	Ea2619 W61*_2		
2216	fliM'	Flagellar motor switch protein FliM'	2029533	C	T	NON SYN	E147K	Ea2619 W61*_2	M	
957		Putative GTP cyclohydrolase 1 type 2'	904575	C	G	NON SYN	R346G	Ea2619 W61*_3		
997	ltaS2'	Lipoteichoic acid synthase 2'	946947	C	T	NON SYN	A463V	Ea2619 W61*_3		
2216	fliM'	Flagellar motor switch protein FliM'	2029334	A	G	NON SYN	L213S	Ea2619 W61*_3	M	
202	sigW'	ECF RNA polymerase sigma factor SigW'	159807	C	A	NON SYN	P144Q	Ea2619 W61*_4		
2216	fliM'	Flagellar motor switch protein FliM'	2029532	T	C	NON SYN	E147G	Ea2619 W61*_4	M	
2659	ptsI'	Phosphoenolpyruvate-protein phosphotransferase'	2466047	C	T	NON SYN	T298M	Ea2619 W61*_4		
2831	degS'	Signal transduction histidine-protein	2620760	G	T	NON SYN	D249E	Ea2619 W61*_4		M

kinase/phosphatase DegS'

2216	fliM'	Flagellar motor switch protein FliM'	2029332	T	C	NON SYN	N214D	Ea2862 W657*_3	M
2216	fliM'	Flagellar motor switch protein FliM'	2029895	T	C	NON SYN	E26G	Ea2862 W657*_2	M
2476	epsE'	Type II secretion system protein E'	2274422	G	T	NON SYN	T410K	Ea2862 W657*_3	M
2659	ptsI'	Phosphoenolpyruvate-protein phosphotransferase'	2465197	G	C	NON SYN	A15P	Ea2862 W657*_1	
704	liaS'	Sensor histidine kinase LiaS'	654669	C	T	NON SYN	E141K	Ea2619 W85*_1	M
2659	ptsI'	Phosphoenolpyruvate-protein phosphotransferase'	2465530	G	A	NON SYN	E126K	Ea2619 W85*_1	
2831	degS'	Signal transduction histidine-protein kinase/phosphatase DegS'	2620930	C	T	NON SYN	A193T	Ea2619 W85*_1	M
3408	ahpF'	NADH dehydrogenase'	3197840	T	G	NON SYN	Y259D	Ea2619 W85*_1	
2659	ptsI'	Phosphoenolpyruvate-protein phosphotransferase'	2465156	T	C	NON SYN	MIT	Ea2619 W85*_2	
2831	degS'	Signal transduction histidine-protein kinase/phosphatase DegS'	2620740	G	T	NON SYN	T256K	Ea2619 W85*_2	M
2914	lytD'	Beta-N-acetylglucosaminidase precursor'	2721746	G	A	NON SYN	C1041Y	Ea2619 W85*_2	
997	ltaS2'	Lipoteichoic acid synthase 2'	946737	C	T	NON SYN	T393M	Ea2619 W85*_3	
1389	yciC'	Putative metal chaperone YciC'	1312714	G	T	NON SYN	D264Y	Ea2619 W85*_3	
1389	yciC'	Putative metal chaperone YciC'	1312715	A	T	NON SYN	D264V	Ea2619 W85*_3	
1417		Tetratricopeptide repeat protein'	1336714	C	A	NON SYN	D170Y	Ea2619 W85*_3	
210	est_1'	Carboxylesterase'	167755	C	T	NON SYN	R231C	Ea2619 W161*_1	
344	malL_1'	Oligo-1%2C6-glucosidase'	294377	C	T	NON SYN	T486M	Ea2619 W161*_1	
554	purH'	Bifunctional purine biosynthesis protein PurH'	507076	A	G	NON SYN	T461A	Ea2619 W161*_1	
787	trmL'	tRNA (cytidine(34)-2''-O)-methyltransferase'	726723	A	G	NON SYN	H61R	Ea2619 W161*_1	
1187	mutS'	DNA mismatch repair protein MutS'	1120178	C	T	NON SYN	A214V	Ea2619 W161*_1	
1267		metal-dependent hydrolase'	1197646	G	A	NON SYN	S352N	Ea2619 W161*_1	
1896	ybhR'	Inner membrane transport permease YbhR'	1736309	G	A	NON SYN	G38R	Ea2619 W161*_1	

2203	cheC'	CheY-P phosphatase CheC'	2017483	T	C	NON SYN	H96R	Ea2619 W161*_1	M
2215	fliN'	Flagellar motor switch protein FliN'	2028518	A	G	NON SYN	F157L	Ea2619 W161*_1	M
2272	prkC'	Serine/threonine-protein kinase PrkC'	2079897	G	A	NON SYN	A23V	Ea2619 W161*_1	
2300	ileS_2'	Isoleucine--tRNA ligase'	2110335	T	C	NON SYN	H312R	Ea2619 W161*_1	
2423	alaS_2'	Alanine--tRNA ligase'	2223509	T	C	NON SYN	T333A	Ea2619 W161*_1	
2591	acsA'	Acetyl-coenzyme A synthetase'	2400188	G	A	NON SYN	H11Y	Ea2619 W161*_1	
3146	mgtE'	Magnesium transporter MgtE'	2958326	T	C	NON SYN	Q49R	Ea2619 W161*_1	
575	mprF'	Phosphatidylglycerol lysyltransferase'	529513	C	A	NON SENSE	S181*	Ea2619 W161*_2	
1834		hypothetical protein'	1688853	G	T	NON SYN	G328C	Ea2619 W161*_2	
2216	fliM'	Flagellar motor switch protein FliM'	2029370	G	A	NON SYN	S201L	Ea2619 W161*_2	M
2272	prkC'	Serine/threonine-protein kinase PrkC'	2079556	G	T	NON SYN	Q137K	Ea2619 W161*_2	
2215	fliN'	Flagellar motor switch protein FliN'	2028533	A	C	NON SYN	S152A	Ea2619 W161*_3	M
575	mprF'	Phosphatidylglycerol lysyltransferase'	529336	G	A	NON SYN	G122D	Ea2619 W161*_4	
2216	fliM'	Flagellar motor switch protein FliM'	2029769	G	A	NON SYN	T68M	Ea2619 W161*_4	M
997	ltaS2'	Lipoteichoic acid synthase 2'	947117	C	T	NON SYN	P520S	Ea2619 W43*_1	
2499	clpX'	ATP-dependent Clp protease ATP-binding subunit ClpX'	2305338	G	A	NON SYN	T307M	Ea2619 W43*_1	
2831	degS'	Signal transduction histidine-protein kinase/phosphatase DegS'	2620731	C	A	NON SYN	R259L	Ea2619 W43*_2	M
1128	rsbP'	Phosphoserine phosphatase RsbP'	1063614	T	C	NON SYN	M40T	Ea2619 W43*_3	M
2831	degS'	Signal transduction histidine-protein kinase/phosphatase DegS'	2620743	G	A	NON SYN	P255L	Ea2619 W43*_3	M

Extragenic mutations

			2039870	G	A			Ea2862 Q817*_1	
			2958574	C	T			Ea2862 Q817*_1	
			112341	G	T			Ea2862 Q558*_3	
			629399	A	T			Ea2862 Q558*_2	
			2040006	C	T			Ea2862 Q558*_3	

1909745	T	C	Ea2619 W61* _3
1740501	G	A	Ea2862 W657* _3
1383956	C	T	Ea2619 W161* _1
2039986	C	A	Ea2619 W161* _1

* Multiple suppressor strains were independently isolated from each parental non motile strain. The numbers indicates the lineage of the strains used.

& Putative motility genes identified by reciprocal BLAST analysis (Appendix A)

! Motility genes identified by MEAPS (Appendix B)

~ In NCBI or JGI, *E. acetylicum* loci is identified as "EZWU0009_" followed by five digit number. NCBI Gene IDs can be searched by adding prefix "EZWU0009_" and suffix "0" to the Gene ID number (e.g. Gene ID number 126 = EZWU0009_01260).

Appendix D

Location of spontaneous mutations of three *E. acetylicum* hypermotility mutant strains that had been selected on soft agar

Gene ID*	Gene name	Protein product	Suppressor Mutation location	Reference ntd	Mutated ntd	Polymorphism type	Amino acid change	Strain name
369	'copC'	'Copper resistance protein C precursor'	318738	G	A	NON SYN	V40I	H1
2216	'fliM'	'Flagellar motor switch protein FliM'	2029892	A	G	NON SYN	I27T	H1
147	'tuf'	'Elongation factor Tu'	129058	C	A	NON SYN	L168I	H2
997	'ltaS2'	'Lipoteichoic acid synthase 2'	946988	T	C	NON SYN	W477R	H2
2216	'fliM'	'Flagellar motor switch protein FliM'	2029390	A	C	NON SYN	N194K	H2
2831	'degS'	'Signal transduction histidine-protein kinase/phosphatase DegS'	2620774	G	A	NON SYN	P245S	H4

*In NCBI or JGI, *E. acetylicum* loci is identified as "EZWU0009_" followed by five digit number. NCBI Gene IDs can be searched by adding prefix "EZWU0009_" and suffix "0" to the Gene ID number (e.g. Gene ID number 126 = EZWU0009_01260).

Appendix E

Primer sequences used for RT-PCR

Primer Name	Sequence (5' – 3')
AcGFP_Foreward	AGG AGC GCA CCA TCT TCT TC
AcGFP_Reverse	TCC TCC TTG AAA TCG GTG CC
High1_Foreward	TAC GGC AAG CTG ACA CTG AA
High1_Reverse	GGT ACC GAG AGA AGC ACT GG
High3_Foreward	ATT ACA ACG CGC ACA ACG TC
High3_Reverse	CCA GTT GCA CAC TCC CAT CT
High8_Foreward	TTC AGC TGG CGG ACC ATT AC
High8_Reverse	CTT GCT CAG CGC ACT TTG TG
High9_Foreward	GAC CGA CAA AGC CAA GAA CG
High9_Reverse	CGA TGG GCG TAT TTT GCT GG
Low2_Foreward	GAA CGG GCA CAA GTT TAG CG
Low2_Reverse	GTA GGC CAA GGG ACT GGA AG
Low10_Foreward	TCA GCA GAA CAC GCC TAT CG
Low10_Reverse	TGG TCA CGC TTT TCA TTC GG
16sRNA_Foreward	AGA AGC TTG CTC TTT GCT GA
16sRNA_Reverse	CTT TGG TCT TGC GAC GTT AT

Appendix F

High, Low and AcGFP1 sequences of N-terminal swap region

>AcGFP

```
ATGGTGAGCAAGGGCGCCGAGCTGTTACCCGGCATCGTGCCCATCCTGATCGAGCTG
AATGGCGATGTGAATGGCCACAAGTTCAGCGTGAGCGGGCGAGGGCGATGCCACCTA
CGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGCAAGCTGCCTGTGCCCTGGCCCACCT
GGTGACCACCCTGAGCTACGGCGTGCAGTGCTTCTCACGCTACCCCGATCACATGAAGCAGCA
CGACTTCTTCAAGAGCGCCATGCCTGAGGGCTACATCCAGGAGCGCACCATCTTCTTCGAGGA
TGACGGCAACTACAAGTCGCGCGCCGAGGTGAAGTTCGAGGGCGATAACCCTGGTGAATCGCA
TCGAGCTGACCGGCACCGATTTCAAGGAGGATGGCAACATCCTGGGCAATAAGATGGAGTAC
AACTACAACGCCACAATGTGTACATCATGACCGACAAGGCCAAGAATGGCATCAAGGTGAA
CTTCAAGATCCGCCACAACATCGAGGATGGCAGCGTGCAGCTGGCCGACCACTACCAGCAGA
ATACCCCATCGGCGATGGCCCTGTGCTGCTGCCGATAACCACTACCTGTCCACCCAGAGCG
CCCTGTCCAAGGACCCCAACGAGAAGCGCGATCACATGATCTACTTCGGCTTCGTGACCGCC
CCGCCATCACCCACGGCATGGATGAGCTGTACAAGTGA
```

>H3

```
ATGGTAAGTAAAGGGCGCCGAGCTATTCACAGGGATAGTTCCTATACTCATCGAACTTA
ACGGTGACGTAAATGGACACAAATTCAGCGTTAGTGGTGAGGGCGAGGGCGACGCAACGTAC
GGCAAGCTTACTCTCAAATTCATATGCACCCTGGTAAACTTCCAGTTCCTGGCCCACGCTC
GTCACTACTCTCAGCTACGGAGTGCAGTGTTTTAGTCGGTATCCGGACCACATGAAGCAACAT
GATTTCTTCAAGAGCGCCATGCCAGAGGGTTACATCCAGGAGCGGACGATTTTTTTTCGAGGAC
GACGGGAATTACAAGAGCCGGGCCGAGGTCAAATTTGAAGGGGACACTCTGGTAAACCGGAT
CGAACTAACAGGCACTGACTTCAAGGAGGATGGTAACATACTCGGTAACAAGATGGAGTATA
ATTACAACGCGCACAACGTCTATATAATGACTGACAAAGCCAAGAATGGCATTAAAGTGAAT
TTCAAGATCCGCCATAATATTGAAGATGGGAGTGTGCAACTGGCCGACCACTACCAGCAAAA
TACCCCGATTGGGGACGGCCAGTCCTACTGCCGACAATCATTACCTGAGCACCCAAAGCGC
CCTAAGCAAGGATCCTAATGAAAAGCGTGACCATATGATTTATTTTGGGTTTGTACGGCGGC
TGCTATAACACACGGCATGGATGAGCTTTACAAATAA
```

>H9

```
ATGGTCAGCAAAGGTGCCGAGCTATTCACGGGAATAGTACCAATACTCATTGAGCTC
AATGGAGATGTTAATGGGCACAAATTCAGTGTAAGTGGCGAGGGCGAAGGCGACGCCACCTA
TGCCAAACTCACACTCAAGTTCATATGCACCCTGGGAACTGCCAGTGCCTTGGCCCACACT
GGTCACGACCCTAAGTTATGGTGTCCAATGCTTCAGCCGCTACCCCGATCACATGAAACAGCA
TGACTTCTTTAAAAGCGCAATGCCAGAAGGATATATCCAGGAACGCACTATTTTCTTCGAAGA
CGACGGAAACTATAAATCCCGGGCCGAGGTCAAGTTCGAAGGAGATACTCTCGTCAATCGTA
TCGAGCTAACCGGGACAGACTTCAAGGAGGATGGGAATATTTTGGGTAACAAGATGGAGTAC
AACTACAACGCCATAACGTGTACATCATGACCGACAAGCCAAGAACGGCATCAAAGTCAA
ATTCAAGATCCGACATAACATCGAGGATGGGTCCGTACAGTTGGCAGATCACTACCAGCAAA
ATACGCCCATCGGGGACGGCCCTGTCTCCTACCCGATAATCATTATCTAAGTACCCAAAGTG
```

CCCTGAGTAAAGATCCTAACGAGAAAAGAGACCATATGATTTATTTCTGGGTTTGTACACGGCAG
CTGCGATCACCCATGGCATGGACGAGCTATAACAAGTAA

>L2

ATGGTCAGCAAGGGTGCCGAGCTCTTTACGGCATTGTACCGATACTGATCGAGCTCA
ACGGGGACGTGAACGGGCACAAGTTTAGCGTGAGTGCCGAGGGCGAGGGGGACGCGACCTA
CGGGAAGCTCACACTCAAGTTTATTTGCACGACTGGCAAGCTTCCAGTCCCTTGGCCTACCCT
CGTAACAACACTAAGTTACGGGGTCCAGTGTTTTAGTCGGTACCCAGACCATATGAAACAACA
CGACTTCTTCAAGAGTGCGATGCCGGAGGGTTACATACAGGAGCGGACCATCTTTTTCGAGGA
TGACGGGAATTATAAAAAGTCGCGCCGAGGTCAAGTTTGAGGGTGACACACTCGTGAATCGAA
TAGAACTTACTGGAACCGATTTCAAAGAGGACGGAAATATACTTGGTAACAAGATGGAGTAT
AATTACAACGCGCACAATGTGTACATAATGACCGATAAAGCAAAGAACGGTATAAAAAGTTAA
CTTCAAGATTTCGGCATAATATTGAAGATGGAAGTGTACAACACTAGCTGATCATTATCAGCAGAA
TACGCCAATTGGTGATGGACCCGTAACCTACCTGACAACCACTACCTTAGTACCCAGAGCGC
GCTGAGTAAAGATCCGAACGAGAAGCGGGACCATATGATCTACTTTGGGTTCTGTGACGGCGG
CAGCAATTACTCATGGTATGGACGAACCTTACAAATAA

>L10

ATGGTGAGCAAGGGAGCTGAGCTCTTTACGGGCATAGTGCCCATCCTCATCGAGCTTA
ATGGTGATGTTAATGGTCATAAATTCAGTGTTAGCGGTGAAGGGGAAGGGGACGCAACATAC
GGAAAGCTAACATTAAGTTTCATCTGTACCACCGGTAAACTACCAGTGCCCTGGCCAACACTA
GTGACCACACTAAGTTACGGAGTCCAATGCTTTAGTCGGTACCCTGATCATATGAAACAACAC
GATTTTTTTAAATCTGCCATGCCC GAAGGGTATATTCAGGAACGTACCATTTTTTTTTGAGGATG
ATGGGAATTACAAGAGTCGGGCAGAGGTCAAGTTTCGAGGGAGACACGCTGGTCAACCGAATA
GAGTTGACCGGTACAGATTTTTAAGGAGGACGGCAACATTCTAGGAAATAAGATGGAGTATAA
TTACAACGCACACAACGTATATATAATGACCGATAAAGCAAAGAACGGTATAAAGGTCAATT
TTAAGATTTCGACACAACATAGAGGACGGCAGCGTGCAGCTGGCGGACCACTATCAGCAGAAC
ACGCCTATCGGGGATGGTCCCCTGCTCCTTCCCGACAATCACTACCTGAGTACACAGAGCGCA
CTTAGTAAGGATCCGAATGAAAAGCGTGACCATATGATATATTTCTGGGTTTGTGACGGCTGCC
GCCATAACTCACGGTATGGATGAGTTATATAAATAA

References

1. C. A. Petti, Detection and identification of microorganisms by gene amplification and sequencing. *Clin Infect Dis* **44**, 1108-1114 (2007).
2. D. A. Wheeler *et al.*, The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-876 (2008).
3. P. J. Turnbaugh *et al.*, Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc Natl Acad Sci U S A* **107**, 7503-7508 (2010).
4. W. B. Whitman, D. C. Coleman, W. J. Wiebe, Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* **95**, 6578-6583 (1998).
5. M. Rajilic-Stojanovic, W. M. de Vos, The first 1000 cultured species of the human gastrointestinal microbiota. *FEMS Microbiol Rev* **38**, 996-1047 (2014).
6. J. Qin *et al.*, A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65 (2010).
7. F. Backhed, R. E. Ley, J. L. Sonnenburg, D. A. Peterson, J. I. Gordon, Host-bacterial mutualism in the human intestine. *Science* **307**, 1915-1920 (2005).
8. R. E. Ley, D. A. Peterson, J. I. Gordon, Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124**, 837-848 (2006).
9. L. V. Hooper *et al.*, Molecular analysis of commensal host-microbial relationships in the intestine. *Science* **291**, 881-884 (2001).
10. P. J. Turnbaugh, F. Backhed, L. Fulton, J. I. Gordon, Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host Microbe* **3**, 213-223 (2008).
11. A. J. Macpherson, N. L. Harris, Interactions between commensal intestinal bacteria and the immune system. *Nat Rev Immunol* **4**, 478-485 (2004).

12. M. J. Claesson *et al.*, Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488**, 178-184 (2012).
13. B. W. Parks *et al.*, Genetic control of obesity and gut microbiota composition in response to high-fat, high-sucrose diet in mice. *Cell Metab* **17**, 141-152 (2013).
14. M. Rajilic-Stojanovic, H. G. Heilig, S. Tims, E. G. Zoetendal, W. M. de Vos, Long-term monitoring of the human intestinal microbiota composition. *Environ Microbiol*, (2012).
15. J. F. Rawls, M. A. Mahowald, R. E. Ley, J. I. Gordon, Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. *Cell* **127**, 423-433 (2006).
16. J. L. Round *et al.*, The Toll-like receptor 2 pathway establishes colonization by a commensal of the human microbiota. *Science* **332**, 974-977 (2011).
17. J. L. Round, S. K. Mazmanian, The gut microbiota shapes intestinal immune responses during health and disease. *Nat Rev Immunol* **9**, 313-323 (2009).
18. I. C. Arnold *et al.*, Helicobacter pylori infection prevents allergic asthma in mouse models through the induction of regulatory T cells. *J Clin Invest* **121**, 3088-3093 (2011).
19. E. Elinav *et al.*, NLRP6 inflammasome regulates colonic microbial ecology and risk for colitis. *Cell* **145**, 745-757 (2011).
20. S. K. Lathrop *et al.*, Peripheral education of the immune system by colonic commensal microbiota. *Nature* **478**, 250-254 (2011).
21. C. J. Robinson, B. J. Bohannan, V. B. Young, From structure to function: the ecology of host-associated microbial communities. *Microbiol Mol Biol Rev* **74**, 453-476 (2010).
22. R. E. Ley *et al.*, Obesity alters gut microbial ecology. *Proc Natl Acad Sci U S A* **102**, 11070-11075 (2005).
23. K. J. Pflughoeft, J. Versalovic, Human microbiome in health and disease. *Annu Rev Pathol* **7**, 99-122 (2012).

24. P. J. Turnbaugh *et al.*, The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci Transl Med* **1**, 6ra14 (2009).
25. F. Armougom, M. Henry, B. Vialettes, D. Raccah, D. Raoult, Monitoring bacterial community of human gut microbiota reveals an increase in *Lactobacillus* in obese patients and Methanogens in anorexic patients. *PLoS One* **4**, e7125 (2009).
26. R. Jumpertz *et al.*, Energy-balance studies reveal associations between gut microbes, caloric load, and nutrient absorption in humans. *Am J Clin Nutr* **94**, 58-65 (2011).
27. J. Dicksved *et al.*, Molecular analysis of the gut microbiota of identical twins with Crohn's disease. *ISME J* **2**, 716-727 (2008).
28. L. M. Brown, *Helicobacter pylori*: epidemiology and routes of transmission. *Epidemiol Rev* **22**, 283-297 (2000).
29. J. S. Bakken, Fecal bacteriotherapy for recurrent *Clostridium difficile* infection. *Anaerobe* **15**, 285-289 (2009).
30. P. J. Turnbaugh *et al.*, A core gut microbiome in obese and lean twins. *Nature* **457**, 480-484 (2009).
31. M. Vijay-Kumar *et al.*, Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. *Science* **328**, 228-231 (2010).
32. L. Wen *et al.*, Innate immunity and intestinal microbiota in the development of Type 1 diabetes. *Nature* **455**, 1109-1113 (2008).
33. D. Mariat *et al.*, The Firmicutes/Bacteroidetes ratio of the human microbiota changes with age. *BMC Microbiol* **9**, 123 (2009).
34. A. Spor, O. Koren, R. Ley, Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat Rev Microbiol* **9**, 279-290 (2011).
35. R. E. Ley, P. J. Turnbaugh, S. Klein, J. I. Gordon, Microbial ecology: human gut microbes associated with obesity. *Nature* **444**, 1022-1023 (2006).

36. I. Semova *et al.*, Microbiota regulate intestinal absorption and metabolism of fatty acids in the zebrafish. *Cell Host Microbe* **12**, 277-288 (2012).
37. W. Z. Stephens *et al.*, Identification of Population Bottlenecks and Colonization Factors during Assembly of Bacterial Communities within the Zebrafish Intestine. *MBio* **6**, e01163-01115 (2015).
38. A. R. Pacheco *et al.*, Fucose sensing regulates bacterial intestinal colonization. *Nature* **492**, 113-117 (2012).
39. S. M. Lee *et al.*, Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature* **501**, 426-429 (2013).
40. V. K. Sharma, S. M. Bearson, Evaluation of the impact of quorum sensing transcriptional regulator SdiA on long-term persistence and fecal shedding of *Escherichia coli* O157:H7 in weaned calves. *Microb Pathog* **57**, 21-26 (2013).
41. B. R. Boles, L. L. McCarter, Insertional inactivation of genes encoding components of the sodium-type flagellar motor and switch of *Vibrio parahaemolyticus*. *J Bacteriol* **182**, 1035-1045 (2000).
42. L. Thomason *et al.*, Recombineering: genetic engineering in bacteria using homologous recombination. *Curr Protoc Mol Biol* **Chapter 1**, Unit 1 16 (2007).
43. T. van Opijnen, K. L. Bodi, A. Camilli, Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Methods* **6**, 767-772 (2009).
44. S. M. Wong, J. D. Gawronski, D. Lapointe, B. J. Akerley, High-throughput insertion tracking by deep sequencing for the analysis of bacterial pathogens. *Methods Mol Biol* **733**, 209-222 (2011).
45. A. L. Goodman, M. Wu, J. I. Gordon, Identifying microbial fitness determinants by insertion sequencing using genome-wide transposon mutant libraries. *Nat Protoc* **6**, 1969-1980 (2011).
46. T. van Opijnen, A. Camilli, Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat Rev Microbiol* **11**, 435-442 (2013).

47. B. D. Nguyen, R. H. Valdivia, Virulence determinants in the obligate intracellular pathogen *Chlamydia trachomatis* revealed by forward genetic approaches. *Proc Natl Acad Sci U S A* **109**, 1263-1268 (2012).
48. M. A. Harper *et al.*, Phenotype sequencing: identifying the genes that cause a phenotype directly from pooled sequencing of independent mutants. *PLoS One* **6**, e16517 (2011).
49. W. P. Robins, S. M. Faruque, J. J. Mekalanos, Coupling mutagenesis and parallel deep sequencing to probe essential residues in a genome or gene. *Proc Natl Acad Sci U S A* **110**, E848-857 (2013).
50. P. M. Sharp, W. H. Li, An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol* **24**, 28-38 (1986).
51. E. Angov, C. J. Hillier, R. L. Kincaid, J. A. Lyon, Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. *PLoS One* **3**, e2189 (2008).
52. G. Kudla, A. W. Murray, D. Tollervey, J. B. Plotkin, Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* **324**, 255-258 (2009).
53. M. Welch *et al.*, Design parameters to control synthetic gene expression in *Escherichia coli*. *PLoS One* **4**, e7002 (2009).
54. D. B. Goodman, G. M. Church, S. Kosuri, Causes and effects of N-terminal codon bias in bacterial genes. *Science* **342**, 475-479 (2013).
55. P. M. Sharp, W. H. Li, The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**, 1281-1295 (1987).
56. Z. E. Sauna, C. Kimchi-Sarfaty, Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet* **12**, 683-691 (2011).
57. M. dos Reis, R. Savva, L. Wernisch, Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* **32**, 5036-5044 (2004).

58. E. P. Rocha, Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Res* **14**, 2279-2286 (2004).
59. N. Zhi *et al.*, Codon optimization of human parvovirus B19 capsid genes greatly increases their expression in nonpermissive cells. *J Virol* **84**, 13059-13062 (2010).
60. Z. Zhou, P. Schnake, L. Xiao, A. A. Lal, Enhanced expression of a recombinant malaria candidate vaccine in *Escherichia coli* by codon optimization. *Protein Expr Purif* **34**, 87-94 (2004).
61. I. Mani, V. Singh, D. K. Chaudhary, P. Somvanshi, M. P. Negi, Codon optimization of the major antigen encoding genes of diverse strains of influenza a virus. *Interdiscip Sci* **3**, 36-42 (2011).
62. G. Marais, L. Duret, Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. *J Mol Evol* **52**, 275-280 (2001).
63. R. Hershberg, D. A. Petrov, Selection on codon bias. *Annu Rev Genet* **42**, 287-299 (2008).
64. E. Angov, Codon usage: nature's roadmap to expression and folding of proteins. *Biotechnol J* **6**, 650-659 (2011).
65. F. Supek, T. Smuc, On relevance of codon usage to expression of synthetic and natural genes in *Escherichia coli*. *Genetics* **185**, 1129-1134 (2010).
66. M. McFall-Ngai *et al.*, Animals in a bacterial world, a new imperative for the life sciences. *Proc Natl Acad Sci U S A* **110**, 3229-3236 (2013).
67. M. Nieuwdorp, P. W. Gilijamse, N. Pai, L. M. Kaplan, Role of the microbiome in energy regulation and metabolism. *Gastroenterology* **146**, 1525-1533 (2014).
68. A. L. Goodman *et al.*, Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe* **6**, 279-289 (2009).
69. M. Wu *et al.*, Genetic determinants of in vivo fitness and diet responsiveness in multiple human gut Bacteroides. *Science* **350**, aac5992 (2015).

70. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014).
71. T. A. Vishnivetskaya *et al.*, Draft genome sequences of 10 strains of the genus *exiguobacterium*. *Genome Announc* **2**, (2014).
72. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461 (2010).
73. H. S. Girgis, Y. Liu, W. S. Ryu, S. Tavazoie, A comprehensive genetic characterization of bacterial motility. *PLoS Genet* **3**, 1644-1660 (2007).
74. V. R. Malapaka, A. A. Barrese, B. C. Tripp, B. C. Tripp, High-throughput screening for antimicrobial compounds using a 96-well format bacterial motility absorbance assay. *J Biomol Screen* **12**, 849-854 (2007).
75. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).
76. Z. Wei, W. Wang, P. Hu, G. J. Lyon, H. Hakonarson, SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res* **39**, e132 (2011).
77. S. Mukherjee, D. B. Kearns, The structure and regulation of flagella in *Bacillus subtilis*. *Annu Rev Genet* **48**, 319-340 (2014).
78. L. N. Pham, M. Kanther, I. Semova, J. F. Rawls, Methods for generating and colonizing gnotobiotic zebrafish. *Nat Protoc* **3**, 1862-1875 (2008).
79. M. Erhardt, Strategies to Block Bacterial Pathogenesis by Interference with Motility and Chemotaxis. *Curr Top Microbiol Immunol*, (2016).
80. J. F. Rawls, M. A. Mahowald, A. L. Goodman, C. M. Trent, J. I. Gordon, In vivo imaging and genetic analysis link bacterial motility and symbiosis in the zebrafish gut. *Proc Natl Acad Sci U S A* **104**, 7622-7627 (2007).
81. T. C. Cullender *et al.*, Innate and adaptive immunity interact to quench microbiome flagellar motility in the gut. *Cell Host Microbe* **14**, 571-581 (2013).

82. C. Schwab *et al.*, Longitudinal study of murine microbiota activity and interactions with the host during acute inflammation and recovery. *ISME J* **8**, 1101-1114 (2014).
83. P. B. Carpenter, G. W. Ordal, Bacillus subtilis FlhA: a flagellar protein related to a new family of signal-transducing receptors. *Mol Microbiol* **7**, 735-743 (1993).
84. A. R. Zuberi, C. Ying, D. S. Bischoff, G. W. Ordal, Gene-protein relationships in the flagellar hook-basal body complex of Bacillus subtilis: sequences of the flgB, flgC, flgG, fliE and fliF genes. *Gene* **101**, 23-31 (1991).
85. A. R. Zuberi, D. S. Bischoff, G. W. Ordal, Nucleotide sequence and characterization of a Bacillus subtilis gene encoding a flagellar switch protein. *J Bacteriol* **173**, 710-719 (1991).
86. L. S. Cairns *et al.*, FlgN is required for flagellum-based motility by Bacillus subtilis. *J Bacteriol* **196**, 2216-2226 (2014).
87. T. Schirmer, U. Jenal, Structural and mechanistic determinants of c-di-GMP signalling. *Nat Rev Microbiol* **7**, 724-735 (2009).
88. S. Garti-Levi, R. Hazan, J. Kain, M. Fujita, S. Ben-Yehuda, The FtsEX ABC transporter directs cellular differentiation in Bacillus subtilis. *Mol Microbiol* **69**, 1018-1028 (2008).
89. J. Meisner *et al.*, FtsEX is required for CwlO peptidoglycan hydrolase activity during cell wall elongation in Bacillus subtilis. *Mol Microbiol* **89**, 1069-1083 (2013).
90. R. Chen, S. B. Guttenplan, K. M. Blair, D. B. Kearns, Role of the sigmaD-dependent autolysins in Bacillus subtilis population heterogeneity. *J Bacteriol* **191**, 5775-5784 (2009).
91. M. J. Kempf, M. J. McBride, Transposon insertions in the Flavobacterium johnsoniae ftsX gene disrupt gliding motility and cell division. *J Bacteriol* **182**, 1671-1679 (2000).
92. H. Szurmant, T. J. Muff, G. W. Ordal, Bacillus subtilis CheC and FliY are members of a novel class of CheY-P-hydrolyzing proteins in the chemotactic signal transduction cascade. *J Biol Chem* **279**, 21787-21792 (2004).

93. S. Neumann, K. Grosse, V. Sourjik, Chemotactic signaling via carbohydrate phosphotransferase systems in *Escherichia coli*. *Proc Natl Acad Sci U S A* **109**, 12159-12164 (2012).
94. G. Amati, P. Bisicchia, A. Galizzi, DegU-P represses expression of the motility *fla-che* operon in *Bacillus subtilis*. *J Bacteriol* **186**, 6003-6014 (2004).
95. K. Paul, V. Nieto, W. C. Carlquist, D. F. Blair, R. M. Harshey, The c-di-GMP binding protein YcgR controls flagellar motor direction and speed to affect chemotaxis by a "backstop brake" mechanism. *Mol Cell* **38**, 128-139 (2010).
96. S. Y. Park, B. Lowder, A. M. Bilwes, D. F. Blair, B. R. Crane, Structure of FliM provides insight into assembly of the switch complex in the bacterial flagella motor. *Proc Natl Acad Sci U S A* **103**, 11886-11891 (2006).
97. A. Pandini, J. Kleinjung, S. Rasool, S. Khan, Coevolved Mutations Reveal Distinct Architectures for Two Core Proteins in the Bacterial Flagellar Motor. *PLoS One* **10**, e0142407 (2015).
98. E. A. Libby, L. A. Goss, J. Dworkin, The Eukaryotic-Like Ser/Thr Kinase PrkC Regulates the Essential WalRK Two-Component System in *Bacillus subtilis*. *PLoS Genet* **11**, e1005275 (2015).
99. M. Levefaudes *et al.*, Diaminopimelic Acid Amidation in Corynebacteriales: NEW INSIGHTS INTO THE ROLE OF LtsA IN PEPTIDOGLYCAN MODIFICATION. *J Biol Chem* **290**, 13079-13094 (2015).
100. S. Samant, F. F. Hsu, A. A. Neyfakh, H. Lee, The *Bacillus anthracis* protein MprF is required for synthesis of lysylphosphatidylglycerols and for resistance to cationic antimicrobial peptides. *J Bacteriol* **191**, 1311-1319 (2009).
101. A. Boehm *et al.*, Second messenger-mediated adjustment of bacterial swimming velocity. *Cell* **141**, 107-116 (2010).
102. S. B. Guttenplan, D. B. Kearns, Regulation of flagellar motility during biofilm formation. *FEMS Microbiol Rev* **37**, 849-871 (2013).
103. J. T. Henry, S. Crosson, Ligand-binding PAS domains in a genomic, cellular, and structural context. *Annu Rev Microbiol* **65**, 261-286 (2011).

104. J. R. Tuckerman *et al.*, An oxygen-sensing diguanylate cyclase and phosphodiesterase couple for c-di-GMP control. *Biochemistry* **48**, 9764-9774 (2009).
105. X. Gao *et al.*, Functional characterization of core components of the *Bacillus subtilis* cyclic-di-GMP signaling pathway. *J Bacteriol* **195**, 4782-4792 (2013).
106. Y. Chen, Y. Chai, J. H. Guo, R. Losick, Evidence for cyclic Di-GMP-mediated signaling in *Bacillus subtilis*. *J Bacteriol* **194**, 5080-5090 (2012).
107. R. P. Ryan, S. Q. An, J. H. Allan, Y. McCarthy, J. M. Dow, The DSF Family of Cell-Cell Signals: An Expanding Class of Bacterial Virulence Regulators. *PLoS Pathog* **11**, e1004986 (2015).
108. L. Rahn-Lee *et al.*, A genetic strategy for probing the functional diversity of magnetosome formation. *PLoS Genet* **11**, e1004811 (2015).
109. D. A. Wride *et al.*, Confirmation of the cellular targets of benomyl and rapamycin using next-generation sequencing of resistant mutants in *S. cerevisiae*. *Mol Biosyst* **10**, 3179-3187 (2014).
110. M. Harper, L. Gronenberg, J. Liao, C. Lee, Comprehensive detection of genes causing a phenotype using phenotype sequencing and pathway analysis. *PLoS One* **9**, e88072 (2014).
111. K. Kobayashi *et al.*, Essential *Bacillus subtilis* genes. *Proc Natl Acad Sci U S A* **100**, 4678-4683 (2003).
112. R. A. Forsyth *et al.*, A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Mol Microbiol* **43**, 1387-1400 (2002).
113. M. A. DePristo *et al.*, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498 (2011).
114. A. McKenna *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
115. K. Nakamura *et al.*, Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* **39**, e90 (2011).

116. M. Schirmer, R. D'Amore, U. Z. Ijaz, N. Hall, C. Quince, Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* **17**, 125 (2016).
117. Y. Bromberg, G. Yachdav, B. Rost, SNAP predicts effect of mutations on protein function. *Bioinformatics* **24**, 2397-2398 (2008).
118. T. Baba *et al.*, Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**, 2006 0008 (2006).
119. M. Bubunencko, T. Baker, D. L. Court, Essentiality of ribosomal and transcription antitermination proteins analyzed by systematic gene replacement in Escherichia coli. *J Bacteriol* **189**, 2844-2853 (2007).
120. A. E. Clatworthy, E. Pierson, D. T. Hung, Targeting virulence: a new paradigm for antimicrobial therapy. *Nat Chem Biol* **3**, 541-548 (2007).
121. R. L. Charlebois, W. F. Doolittle, Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Res* **14**, 2469-2477 (2004).
122. S. Bae, O. Mueller, S. Wong, J. F. Rawls, R. H. Valdivia, Genomic sequencing-based mutational enrichment analysis identifies motility genes in a genetically intractable gut microbe. *Proc Natl Acad Sci U S A* **113**, 14127-14132 (2016).
123. M. Gouy, C. Gautier, Codon usage in bacteria: correlation with gene expressivity. *Nucleic acids research* **10**, 7055-7074 (1982).
124. P. M. Sharp, T. M. Tuohy, K. R. Mosurski, Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic acids research* **14**, 5125-5143 (1986).
125. J. R. TerMaat, E. Pienaar, S. E. Whitney, T. G. Mamedov, A. Subramanian, Gene synthesis by integrated polymerase chain assembly and PCR amplification using a high-speed thermocycler. *J Microbiol Methods* **79**, 295-300 (2009).
126. J. Quan, J. Tian, Circular polymerase extension cloning for high-throughput cloning of complex and combinatorial DNA libraries. *Nat Protoc* **6**, 242-251 (2011).

127. H. Ogawa, S. Inouye, F. I. Tsuji, K. Yasuda, K. Umesono, Localization, trafficking, and temperature-dependence of the Aequorea green fluorescent protein in cultured vertebrate cells. *Proc Natl Acad Sci U S A* **92**, 11899-11903 (1995).
128. Y. Nakamura, T. Gojobori, T. Ikemura, Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic acids research* **28**, 292 (2000).
129. F. Wright, The 'effective number of codons' used in a gene. *Gene* **87**, 23-29 (1990).
130. H. Akashi, Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* **136**, 927-935 (1994).
131. J. V. Chamary, L. D. Hurst, Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol* **6**, R75 (2005).
132. R. M. Goetz, A. Fuglsang, Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from *Escherichia coli*. *Biochem Biophys Res Commun* **327**, 4-7 (2005).
133. J. Duan *et al.*, Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Human molecular genetics* **12**, 205-216 (2003).
134. F. Capon *et al.*, A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups. *Hum Mol Genet* **13**, 2361-2368 (2004).
135. N. R. Markham, M. Zuker, UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol* **453**, 3-31 (2008).
136. X. F. Wan, D. Xu, A. Kleinhofs, J. Zhou, Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. *BMC Evol Biol* **4**, 19 (2004).
137. S. Wang *et al.*, Hemagglutinin (HA) proteins from H1 and H3 serotypes of influenza A viruses require different antigen designs for the induction of optimal protective antibody responses as studied by codon-optimized HA DNA vaccines. *J Virol* **80**, 11628-11637 (2006).

138. H. Foster *et al.*, Codon and mRNA sequence optimization of microdystrophin transgenes improves expression and physiological outcome in dystrophic mdx mice following AAV2/8 gene transfer. *Mol Ther* **16**, 1825-1832 (2008).
139. N. A. Burgess-Brown *et al.*, Codon optimization can improve expression of human genes in *Escherichia coli*: A multi-gene study. *Protein Expr Purif* **59**, 94-102 (2008).
140. N. Zhi *et al.*, Codon optimization of human parvovirus B19 capsid genes greatly increases their expression in non-permissive cells. *J Virol*, (2010).
141. M. Zuker, Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406-3415 (2003).
142. J. D. Gawronski, S. M. Wong, G. Giannoukos, D. V. Ward, B. J. Akerley, Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. *Proc Natl Acad Sci U S A* **106**, 16422-16427 (2009).

Biography

Sena Bae was born and grew up in South Korea. She received her Bachelor of Science degree in Computer Science, at California State University, Sacramento in 2009.