CHARACTERIZATION OF A PROKARYOTIC K^+ CHANNEL FROM *ESCHERICHIA COLI*

A Dissertation

by

SARAH DANIELLE BEAGLE

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Steve W. Lockless
Committee Members,	James C. Hu
	Deborah A. Siegele
	James L. Smith
	Joseph A. Sorg
Head of Department,	Thomas D. McKnight

December 2019

Major Subject: Microbiology

ABSTRACT

K^+ channels are widely conserved among all species. Although bacterial K^+ channels are often used in structural and biophysical assays as models for their eukaryotic homologs, little is known about their physiological role in bacteria. *In vivo* characterization of bacterial channels has been difficult because an obvious phenotype is not always associated with a null mutant. The *E. coli* genome contains one K^+-selective channel, Kch, which remains poorly understood despite attempts at characterization. In order to elucidate the physiological function of Kch, we performed a large-scale computational protein co-evolution analysis to predict protein interaction partners of Kch. We hypothesized that determining which proteins the K^+ channel was predicted to interact with could reveal insight into its function. Linking the channel to proteins with known biological functions would allow for targeted experimental validation of the predicted interactions and further *in vivo* characterization. Our analysis revealed that Kch was predicted to co-evolve with proteins involved in oxidation-reduction processes, cell division, and metabolism.

We first asked if loss of the channel resulted in a growth defect in various media. In rich media, the Δ*kch* strain exhibited a slight growth defect upon entering mid-exponential phase, which could be rescued by the addition of a fermentable, but not an oxidizable, carbon source. Replacement of the Δ*kch* mutation with a functional *kch* gene failed to rescue the growth defect, and whole genome sequencing revealed additional mutations in the background, including a point mutation in *ubiH* (*ubiH*^{V223G}). *UbiH* is

required for biosynthesis of the electron carrier molecule, ubiquinone, which functions in the aerobic electron transport chain (ETC). Characterization revealed that the *ubiH*^{V223G} mutation acts to reduce the overall efficiency of the aerobic ETC, leading us to hypothesize that Kch was involved in modulation of the membrane potential ($\Delta\Psi$). CRISPR interference (CRISPRi)-mediated *kch* depletion results in growth defects and $\Delta\Psi$ fluctuations, indicating that the native function of Kch is rapid modulation of $\Delta\Psi$. Using a variety of novel approaches, we demonstrate that Kch is important for adaptation to conditions that promote rapid growth, expanding our limited understanding of the physiological functions of microbial K^+ channels.

DEDICATION

To my parents, who sacrificed a lot to make my dreams possible and supported me during every step in this process.

For my future researchers – Addison, Joshua, and Rachel – never stop asking questions.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Steve Lockless, for his patience and support throughout the course of this work. Thanks also to my committee members, Dr. Siegele, Dr. Sorg, Dr. Smith, and Dr. Hu, for their advice and feedback on my projects.

I am grateful to current and former members of the Lockless lab: Shian Liu, Jibran Khan, Katrina Hofstetter, and Ashley Hudson, who became my Texas family. Their friendship and support made this process much more enjoyable. Thanks to Alicia Burris for the constant friendship, for the moral support, and for correcting me all the times I tried to give carbon five bonds. Finally, thanks to my husband, Aaron. His encouragement, love, and belief in me have meant so much. I could not imagine doing this without him.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Dr. Steve Lockless, advisor, and Drs. Deborah Siegele, James Smith, and Joseph Sorg of the Department of Biology and Dr. James Hu of the Department of Biochemistry and Biophysics.

Membrane potential data using $DiOC_2(3)$ was collected and analyzed by M.A. Hudson. Membrane potential data using Thioflavin T was collected with the assistance of Dong-yeon D. Lee in the Süel Laboratory at the University of California San Diego. Assistance with Adobe InDesign and Illustrator for figure generation was provided by Aaron Beagle. Figures generated outside of Adobe were created with the BioRender platform or Prism GraphPad. All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by Texas A&M University startup funds to Steve Lockless. This work was also made possible in part under Grant Number A1742 from the Welch Foundation. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Office of Graduate and Professional Studies.

NOMENCLATURE

$\Delta\Psi$	Membrane Potential
ΔpH	Proton Gradient
PMF	Proton Motive Force
RCK	Regulator of K^+ Conductance
PPI	Protein-Protein Interaction(s)
SCA	Statistical Coupling Analysis
Q_8	Ubiquinone

TABLE OF CONTENTS

PAGE

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

The cytoplasm of cells is protected from the external environment by semi-permeable lipid bilayers. K^+ channels are transmembrane proteins that allow for the selective movement of K^+ ions across lipid membranes. These proteins are highly conserved and can be found in all domains of life[1,2]. In higher organisms, K^+ channels are well-characterized for their role in action potentials and electrical signaling in excitable cells, but they have additional critical roles in the physiology of both excitable and non-excitable cells. A few examples of known functions are the involvement of K^+ channels in K^+ recycling for electrolyte balance in renal epithelial cells[3], the hyperpolarization events required for mitogenesis and proliferation of T and B cells[4], the secretion of insulin from pancreatic beta cells[5], and the regulation of cell volume and shape in erythrocytes[6]. The extensive understanding of the physiological role of K^+ channels in various cell types led to an increased appreciation for how mutations or disruptions to K^+ channels contribute to human disease. Many renal, cardiac, and metabolic disorders are rooted in channel dysfunction, making K^+ channels attractive and effective pharmacological targets to alleviate pathologies[7–9]. Despite the ubiquitous distribution of K^+ channels, our understanding of the physiological role of K^+ channels has come primarily from studies of higher eukaryotic organisms and much remains to be learned about the role of K^+ channels single-celled organisms and viruses.

Discovery of Prokaryotic K^+ Channels

In contrast to the wealth of functional information available for eukaryotic K^+ channels, the physiological role for bacterial K^+ channels is poorly understood. Because of the early characterization of K^+ channel involvement in neuronal action potentials, it was thought that K^+ channels were present only in eukaryotic organisms[10]. In fact, the identification of bacterial K^+ channels did not occur as a result of an active search for them, but rather as a serendipitous discovery. Nearly twenty years ago, Kch (for K^+ channel) was discovered in *Escherichia coli* by Roger Milkman[11]. Interested in the *trp* operon in *E. coli*, he sequenced through the surrounding region for comparison, and in doing so identified an open reading frame with significant sequence identity to known voltage-gated K^+ channels[11]. The unexpected discovery of Kch was soon followed by reports of K^+ channels in other bacteria, archaea, and even viruses[2,12]. The explosion of whole genome sequencing technologies has revealed that K^+ channels are widely conserved in single-celled organisms (eukaryotic, prokaryotic, and archaeal) and viruses. In fact, the genome of the single-celled ciliate, *Paramecium tetraurelia*, contains ~3x the number of K^+ channels as the human genome[13,14]. Instead of being restricted to neurons or even eukaryotic systems, K^+ channels are ubiquitously found, highlighting their critical importance to cellular physiology.

Key Characteristics of K^+ Channels

Topology of Prokaryotic and Eukaryotic K^+ channels

The primary sequence of viral, prokaryotic, and eukaryotic K^+ channels reveal similarities in the topology of K^+ channels (Fig 1). The simplest arrangement is that of the Chlorella virus K^+ channel, Kcv, which contains an N-terminal transmembrane domain (S_1), a pore helix (P), a K^+

selectivity filter sequence, and a C-terminal transmembrane domain (S_2). Four of these subunits would come together to form a functional channel. This basic structure is found in many prokaryotic and eukaryotic K^+ channels with some variation (Fig 1).

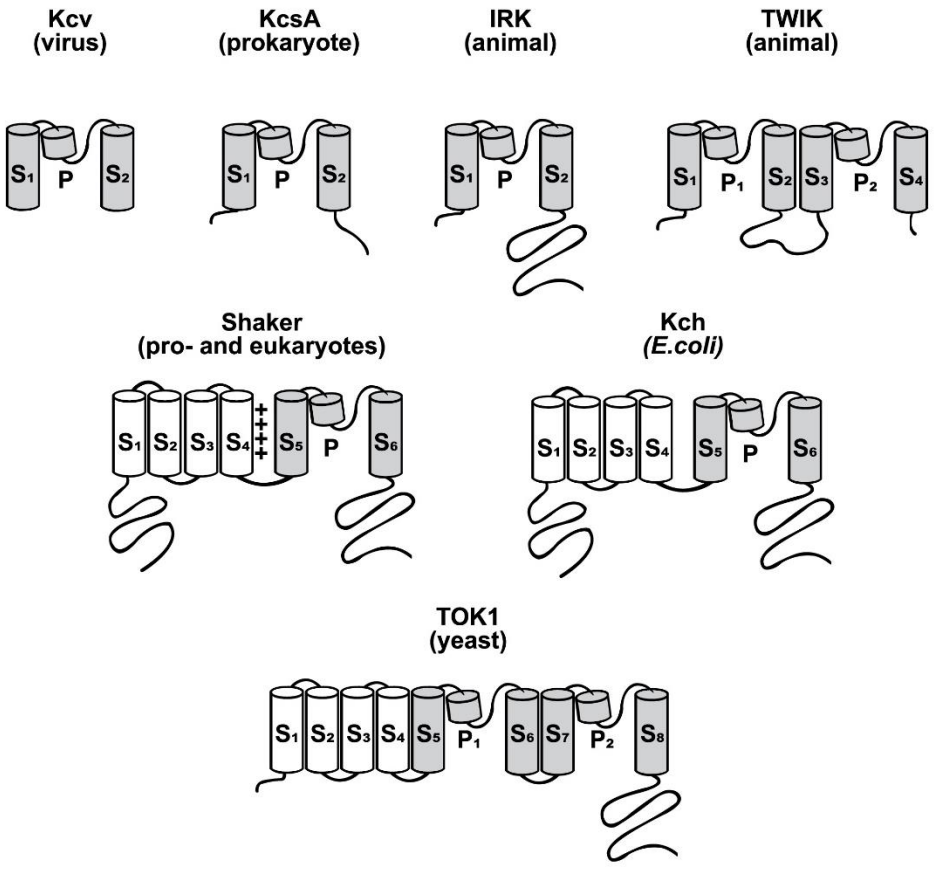

Figure 1. Topology of Prokaryotic and Eukaryotic K^+ channels
Examples of known Prokaryotic and Eukaryotic K^+ channel subunits. Functional channels would result after multimerization of subunits, typically as homotetramers. A notable exception is the two-pore-domain K^+ channels found in mammals (TWIK). Each subunit has two pore domains that dimerize to form a functional channel[15]. Adapted from Kubalski *et al*[10].

Selectivity

Eukaryotic and prokaryotic K^+ channels share many common structural features. In fact, the first structures of K^+ channels, which gave insight into their function, were of bacterial channels[16,17]. The structure of *Streptomyces lividans* KcsA revealed that the pore of the channel

was comprised of four subunits, with each subunit contributing two transmembrane α-helices and a pore helix to create a path for ion conduction. The pore helix spans nearly half the membrane and ends in a large, water-filled cavity. Selectivity for K^+ ions over Na^+ ions is mediated by a selectivity filter sequence located near the extracellular side of the pore[16,18]. This overall architecture is conserved in K^+ channels (Fig 2). The selectivity filter is highly conserved, and drastic variation from the consensus TVGYG sequence is rare[18]. The selectivity filter creates four binding sites for K^+, and the carbonyl oxygens mimic the eight water oxygens that surround a K^+ ion in aqueous solution. On average, there are two K^+ ions located within the selectivity filter at a given time occupying either sites 1 and 3 or sites 2 and 4 with a water molecule separating them. This configuration sets up an ion conduction cycle where a K^+ ion enters from one side of the filter and a different K^+ ion exits from the opposite side, with the direction of ion movement dependent on the electrochemical gradient for K^+[16,18].

The KcsA structure revealed how K^+ channels can be both highly selective for K^+ over other cations and highly conductive: the carbonyl oxygens in the selectivity filter mimic the normal hydration state of a K^+ ion, and having multiple ions in the filter causes repulsion between the ions that overcomes the affinity of the ion for the binding site[16,18].

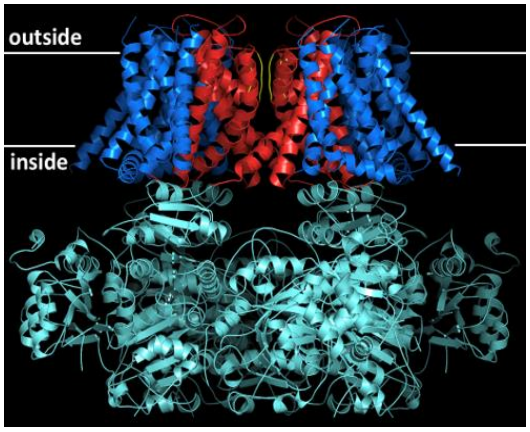

Figure 2. Structure of slo1 BK K^+ Channel
The pore domain is shown in red, the selectivity filter in yellow, voltage-sensing domains in blue, and RCK domains in cyan[19].

Gating

In Eukaryotes, sub-families within the broad K^+ channel family are based largely on the physiological stimuli that cause the K^+ channel to open. For example, some K^+ channels are voltage-gated, which means that the opening of the channel pore is coupled to the movements of a membrane-integrated voltage sensor that senses changes in voltage across the membrane[20]. There are also ligand-gated K^+ channels that depend on the binding of a nucleotide, an ion, or another protein to a regulatory domain for pore opening[18]. One type of regulatory domain is the Regulator of K^+ Conductance (RCK) domain, which is present in some eukaryotic, Ca^{2+}-activated K^+ channels and is commonly found in prokaryotic K^+ channels[21,22]. Electrophysiology data for Prokaryotic channels is limited, and given the difficulty in obtaining this information, a characterization of bacterial K^+ channels based the number of transmembrane helices (TM), the type of regulatory domains, and homology to known K^+ channels has been proposed. This type of organization has resulted in ~10 different categories for bacterial K^+ channels[1] (Fig 3).

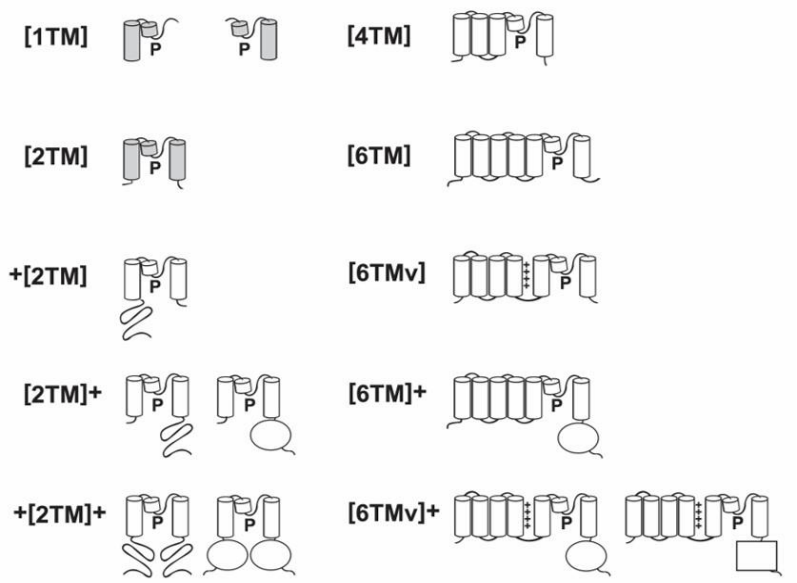

Figure 3. Topology of known Prokaryotic K^+ channel forming subunits
All topologies contain the K^+ selectivity filter in the pore region, but vary in the number of transmembrane domains, the presence of voltage-sensing domains, and the type (if any) of C-terminal regulatory domains. Circles represent Regulation of K^+ Channels (RCK) domains, squares represent Cyclic Nucleotide Binding Domains (CNBDs), and lines represent unknown domains. Adapted from Kubalski *et al.*[10]**.**

The Importance and Transport of K^+ in Prokaryotes

In all cells, K^+ represents the major intracellular monovalent cation. The intracellular K^+ concentration is highly regulated as K^+ is involved in many critical biological processes. It serves as an osmo-protectant and is rapidly accumulated in response to hyperosmotic stress[23–25]. K^+ is important for protein synthesis, as it is required for proper structure and function of the ribosome[26]. Additionally, the gradients established by the intracellular sequestering of K^+ and exclusion of Na^+ are important for transport of other molecules as well as maintenance of proper membrane potential[27]. Because of the importance of K^+, there are a number of widely-conserved K^+ transport systems in bacteria.

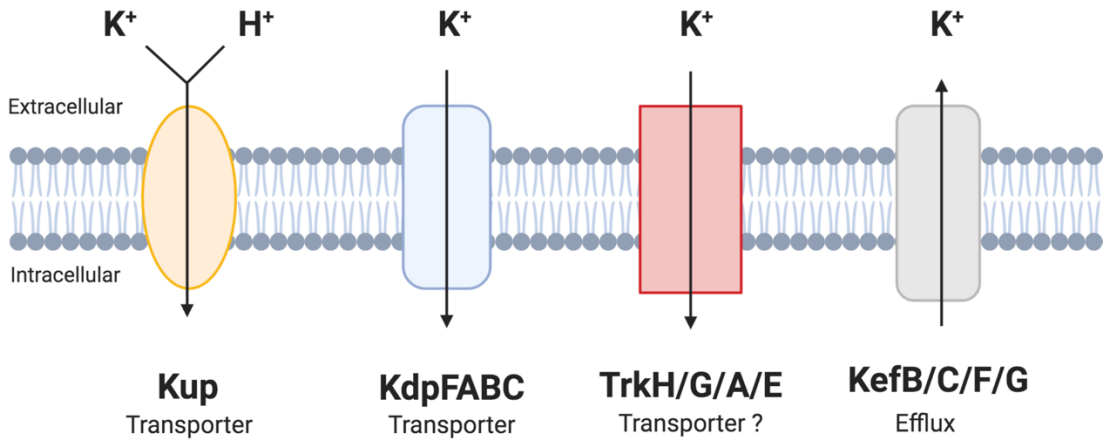

Figure 4. K^+ transport systems in *E. coli*
The K^+ transport systems in *E. coli* are illustrated (see text for description of each system). Arrows indicate the direction of K^+ transport.

In *E. coli*, K^+ transport and homeostasis are mediated via dedicated K^+ transport systems: Trk, Kdp, and Kup (Fig 4). The Trk system (Transport of K^+) is a constitutively expressed system that contains two transporters: TrkH and TrkG[28]. TrkG, which shares 41% amino acid sequence identity with TrkH, was likely introduced to the genome via a *rac* prophage as it has an unusual codon usage and low overall GC content[29] . Disruption of both *trkH* and *trkG* resulted in disrupted K^+ transport in a strain also lacking the other transport systems[30]. TrkH and TrkG are both dependent on a cytoplasmic, RCK domain-containing protein, TrkA, for function. Another cytoplasmic regulatory protein, TrkE, also has been shown to be required for TrkH function and to alter the kinetics of K^+ transport by TrkG[31].

While the Trk proteins were originally identified during screens for proteins involved in K^+ transport[30], the characterization of Trk as a K^+-selective transport system has been challenged by recent work. Structural and *in vitro* work on TrkH/A from *Vibrio parahaemolyticus* suggests that TrkH/A functions as a non-selective channel, conducting both K^+ and Na^+, with only a slight

preference for K^{+}[32,33]. The high sequence similarities of the TrkH selectivity filter sequence between *V. parahaemolyticus* and *E. coli* suggest that this may also be true in *E. coli*, but further work will be needed to determine this.

The Kdp system (K^{+} dependent protein) is comprised of a P-type ATPase encoded by *kdpFABC*, whose expression is regulated by the two-component system KdpDE[34,35]. This system is a high affinity transporter that is induced under conditions of limiting extracellular K^{+} or osmotic stress, although the exact stimulus is still debated. Initial studies suggested that the Kdp system responded to changes in turgor pressure, but extracellular K^{+} concentration, intracellular ionic strength, and ATP levels have also been suggested as possible stimuli[36–38]. *In vitro* work has demonstrated that under high salt concentrations the universal stress protein, UspC, is able to bind to the sensor kinase, KdpD, resulting in increased levels of phosphorylated KdpE. This provides a possible mechanism for increased KdpFABC expression under conditions where the intracellular K^{+} concentration is high, such as under osmotic stress[39]. In enterohemorrhagic *E. coli* (EHEC), KdpD/E have been implicated in pathogenesis, as they control expression of several virulence factors[40].

The Kup (K^{+} uptake) system is a constitutively expressed, low-affinity K^{+} transport system[30]. Under hyper-osmotic stress in low external pH (5.5), the uptake of K^{+} is eliminated in the presence of a protonophore, suggesting that Kup functions as a K^{+}:H^{+} symporter[41,42]. Kup is comprised of a single protein with two domains: an integral membrane domain with 12 transmembrane helices and a cytoplasmic C-terminal domain. Plasmid-encoded *kup* allows for growth of a mutant lacking functional Trk, Kdp, and Kup systems in low K^{+} medium (0.1 mM K^{+})[43]. Kup is thought to be the primary K^{+} uptake system under acidic hyper-osmotic conditions.

An additional system that is absent in *E. coli*, but commonly found in other bacteria is the Ktr (K^+ transport) system. In *Bacillus subtilis*, Ktr transporters are critical for K^+ uptake and response to osmotic shock[44]. Additionally, the Ktr transport system is important for virulence, as a *Staphylococcus aureus* strain with a disrupted Ktr system was more sensitive to antibiotics and was out-competed by a wild-type strain in a mouse bacteremia model[45].

In addition to K^+ uptake systems, bacteria also possess K^+ efflux systems that serve to export K^+. The Kef (K^+ efflux) systems are conserved in *E. coli* and other Gram-negative bacteria[46–48]. The Kef system is comprised of two independent K^+ efflux systems, KefB and KefC, which are regulated by cytoplasmic RCK domains. In *E. coli*, KefB and KefC function as homodimers with KefG and KefF respectively. With the Kef system, export of K^+ is coupled with H^+ import, which serves to protect the cells against electrophilic compounds by acidification of the cytoplasm[49,50].

K^+ transport systems are widely conserved in both Gram-negative and Gram-positive bacteria, highlighting the importance of K^+ to cellular physiology. The multiplicity of K^+ transporter systems has complicated the understanding of K^+ channels. It is difficult in many cases to distinguish if K^+ channels function primarily as another K^+ uptake system and this activity is masked by the functionally redundant transport systems in the cell, if channels have K^+ uptake activities under specific conditions, or if they have functions that are separate or in addition to K^+ homeostasis. The growing interest in bacterial ion channels has resulted in some clarification of these possibilities for K^+ channels in certain bacterial species.

Physiological Functions of Bacterial K^+ Channels

Prokaryotic K^+ channels have contributed greatly to our understanding of the biophysical properties of K^+ channels. Information regarding how K^+ channels are both highly selective and highly conductive came from the first crystal structures of bacterial K^+ channels[18,21]. However, this structural and biochemical information has not greatly contributed to our understanding of cation channel function in bacteria. While most bacterial genomes contain a readily identifiable K^+ channel, there remains little functional characterization of these proteins.

The limited reports of functional characterization for specific bacterial K^+ channels tend to be in organisms that lack well-characterized K^+ transport systems. For example, the genome of the Gram-positive bacterium, *Corynebacterium glutamicum*, does not contain any homologues of the Trk, Ktr, or Kdp systems, but it does possess a Kup-like transporter and a putative K^+ channel, CglK. Disruption of CglK resulted in severely diminished K^+ transport and an impaired response to acid stress[51]. Similarly, the Gram-negative pathogen, *Helicobacter pylori*, lacks any obvious Trk, Kdp, Kup, or Ktr homologues, but has a putative K^+ channel, HpKchA. Disruption of HpKchA altered K^+ uptake and drastically reduced host colonization in a murine stomach infection model[52]. These data suggest that in the absence of known K^+ transport systems, one function of K^+ channels in these organisms is to mediate K^+ uptake and maintain K^+ homeostasis.

However, in bacteria whose genomes contain dedicated K^+ transport systems as well as a K^+ channel, it is much less clear what physiological function the channel plays. *E. coli*, which possesses Trk, Kdp, and Kup K^+ transport systems, also possesses a poorly characterized K^+-selective channel, Kch. It is unlikely that the primary function of Kch is K^+ transport or K^+ homeostasis, as several screens for proteins involved in K^+ transport failed to identify *kch*[30], and

a null *kch* mutant failed to generate any observable K^+ transport phenotypes[13]. The lack of observable phenotypes for channel deletion strains has complicated the study of K^+ channels.

Signaling Roles for Bacterial K^+ Channels

In higher eukaryotic organisms, K^+ channels are involved in modulation of membrane potential, the difference in electrical charge between the intracellular and extracellular sides of the membrane. Changes in concentrations of charged species, such as K^+, across the membrane result in alterations to the membrane potential. The membrane potential ($\Delta\Psi$), along with the proton gradient across the membrane (ΔpH), compose the proton motive force (PMF)[53]. In bacteria, the PMF is required for nutrient transport, ATP production, flagellar rotation, and other critical biological processes[54–57]. The strong conservation of eukaryotic and prokaryotic K^+ channels in both sequence identity and structure suggests that prokaryotic channels would also have roles in signaling, but there are few clear examples of such behavior. However, there is emerging evidence for signaling roles for bacterial K^+ channels.

In the Cyanobacterium, *Synechocystis* sp. PCC 6803, deletion of a Ca^{2+}-dependent K^+ channel, SynCaK, resulted in depolarization of the cells, indicating that the channel alters the membrane potential[58]. Recent work in *B. subtilis* has revealed that bacterial K^+ channels are important for electrical signaling and long-range communication within biofilms. The Süel laboratory demonstrated that, upon reaching a threshold size, continued growth of a *B. subtilis* biofilm occurs in periodic cycles[59]. This oscillatory growth was due to metabolic conflict between interior cells and peripheral cells. The metabolically active peripheral cells rapidly consumed glutamate, starving the interior cells. In turn, the glutamate starvation of the interior cells reduced their production of ammonium, which was needed for growth of the peripheral

cells. Replenishment of glutamate to the interior cells restored ammonium production and growth of the peripheral cells. Intracellular maintenance of both glutamate and ammonium concentrations is dependent on membrane potential[60,61]; therefore, the role of electrical signaling in the communication of metabolic stress was evaluated.

Measurement of membrane potential using a voltage-sensitive dye revealed that there were periodic fluctuations across the biofilm, and that these oscillations were quenched by the addition of glutamate. Further investigation revealed that K^+ efflux was correlated with changes in membrane potential, suggesting that K^+ was involved in propagation of signals within a biofilm community. Consistent with previous observations that YugO was critical for biofilm formation[62], disruption of *yugO* reduced the propagation of electrical oscillations under glutamate limiting conditions. YugO-mediated K^+ release results in transient depolarization of nearby cells, altering their nutrient (glutamate) uptake and retention (ammonium); propagation of this signal throughout the biofilm can be used to communicate metabolic stress to distant sites[59,63].

Electrical signaling from *Bacillus* biofilms, mediated by YugO, attracted motile, planktonic cells of diverse species (*B. subtilis* and *Pseudomonas aeruginosa*) to the biofilm. The YugO-mediated electrical signaling resulted in modulation of the planktonic cell's membrane potential, which altered the PMF and subsequently motility, and resulted in an attraction to the biofilm[64]. Together, these data suggest that bacterial K^+ channels are able to participate in electrical signaling, reminiscent of eukaryotic K^+ channels in excitable cells.

Introduction to the *E. coli* K^+ Channel, Kch

Despite being the first identified bacterial K^+ channel, very little is understood about the *in vivo* role of Kch. Identified due to its sequence identity to known voltage-gated eukaryotic K^+ channels, Kch has a Shaker-like topology comprised of six transmembrane α-helices (S1-S6) with a pore region and K^+ selectivity filter between S5 and S6, and a cytoplasmic RCK domain linked to S6[14] (Fig 3). However, the traditional voltage sensing motif in S4 is degenerate so it is unlikely that Kch responds to voltage changes[11]. The traditional genetic approach to dissecting the function of bacterial genes is to generate a null mutation and then examine the changes in behavior associated with the mutation. However, this approach yielded little insight regarding the role of Kch, as most studies found no discernable phenotypes for Δ*kch* strains [13,14,65]. The lack of apparent phenotypes for *kch* null strains resulted in a shift towards *kch* over-expression for characterization of the protein.

Over-expression studies confirmed that Kch localized to the inner membrane as expected for an ion channel[66,67]. Ungar *et al.* reported that over-production of *kch* drastically altered the intracellular K^+ concentration via increased K^+ leakage, but this K^+ leakage could not be blocked by known K^+ channel inhibitors[13]. Munsey *et al.* reported that uninduced leaky expression of *kch* alone was lethal to *E. coli*. The authors observed that the lethality of *kch* over-expression could be overcome by the addition of 50 mM K^+ to the growth medium. Additionally, they demonstrated that the lethality could be suppressed by the addition of K^+, Rb^+, or NH_4^+ but not by the addition of Na^+, consistent with the selectivity of a K^+ channel[65]. However, Kuo *et al.* found a conflicting result: that over-expression of *kch* was unable to be rescued by excess K^+ in the medium (200 mM K^+). The detrimental effects of *kch* over-expression were exacerbated by ionic strength and were not compensated for by mutations that disrupt the K^+ selectivity filter,

arguing that the growth defects were due to over-expressing a membrane protein - not *kch* specifically[14]. Discrepancies among the overexpression studies make it difficult to interpret any conclusions drawn from them. Attempts at studying the electrophysiological properties of *kch* have also proven unsuccessful. Expression of *kch* in *Xenopus* oocytes failed to elicit any detectable currents[65]. A second attempt by creating a chimeric Shaker protein with a *kch* pore sequence also failed to generate a detectable current[65].

The best evidence that Kch forms a K^+ channel *in vivo* comes from a gain-of-function screen[14]. The Kch coding sequence was cloned into vectors under control of an IPTG-inducible promoter, randomly mutagenized, and then screened in a Δ*kch* background for increased sensitivity to 200 mM extracellular K^+. This screen was done in the absence of inducer to reduce any toxicity effects. Seven mutants that were able to grow on LB but unable to grow in the presence of 200 mM K^+ were explored further, as these are expected to represent channels that have a higher open probability (gain-of-function). These mutants were not sensitive to extracellular Na^+ or sorbitol, and the phenotype was retained when the mutagenized channels were placed under native promoters. Mapping of these mutations revealed that they were all near the C-terminal end of the protein, likely affecting the RCK regulatory domain and subsequently the gating of the channel (Fig 3). If the loss of viability in the presence of 200 mM K^+ was due to mis-regulated K^+ flux through these mutant Kch channels, then disruption of the K^+ selectivity filter should rescue this growth defect. The GYG portion of the selectivity filter sequence was targeted for mutagenesis. Forty-one suppressor mutants were isolated, and all contained mutations that disrupted the selectivity filter, while mutations that failed to rescue the phenotype were those that retained the structure of the filter (TVGFG). This result strongly suggests that Kch selectively transports K^+ *in vivo*[14,65].

In addition to the conflicting results surrounding the *kch* over-expression studies, contradictory results have been reported for the viability of *kch* null mutants. Initial studies of *kch* reported that there were no obvious growth defects or observable phenotypes for a Δ*kch* strain[13,14]. In a modified transposon screen, *kch* was identified as a gene required for growth at 37 °C in rich media but dispensable for growth in minimal media or at cold temperatures[68]. However, under similar growth conditions (37°C in rich media) a viable null *kch* mutant was generated during the creation of the Keio collection, marking the gene as non-essential in these conditions[69]. Despite a series of attempts at functional characterization, much remains to be learned about the functional role of Kch in *E. coli.*

Using Protein Co-Evolution to Identify Functional Interactions

Despite the attempts of several independent laboratories using a variety of traditional biochemical and genetic approaches, little progress has been made in understanding the role of Kch and other bacterial K^+ channels. An alternative and complementary method that could aid in understanding the function of Kch is protein co-evolution. The ability of co-evolution to predict functional interactions between amino acids within proteins has been well-established[70,71]. Here we expand this concept to detect interactions between proteins. As proteins rarely function in isolation, understanding which proteins within the cell are interacting yields valuable information regarding the organization and cooperativity of biological processes within the cell. However, detecting protein-protein interactions is not a trivial task, and many current methods have undesirable technical issues[72].

The protein co-evolution analysis, detailed in chapter 2, is based on the concept that proteins interacting directly or indirectly should experience a selective pressure at the residues

responsible for maintaining that interaction. Our sequence-based analysis is utilized to detect networks of co-evolving amino acid residues between proteins. The presence of these functionally connected amino acid networks between proteins allows us to determine proteins that are co-evolving and, therefore, interacting with each other in some capacity. This method will allow us to identify well-characterized interactions as well as predict novel interactions that can be further validated experimentally.

A potentially useful application of coevolutionary analysis is to proteins that have limited or no characterization information. Uncharacterized proteins can be challenging to study, as it is difficult to design experiments without a predicted function. Additionally, a null mutant may fail to show clear phenotypes because of functional redundancy or because the correct conditions were not used during screening. Having a list of predicted interactions for uncharacterized proteins allows for connections to be drawn to proteins and possibly protein networks with known functions. These predictions will provide a starting framework for targeted experimental design and in vivo characterization studies. Protein co-evolution could be a powerful tool for characterization of proteins such as Kch, where traditional experimental methods did not yield a clear understanding of the *in vivo* function. The results of this analysis will be useful in guiding experimental design for validation of novel interactions and networks, ultimately leading to a better understanding of how biological processes interact within the cell.

In Chapter II, I will discuss my work on extending the co-evolution approach to predict protein-protein interactions on a genome-wide level. The results of this analysis will be used to design assays and experiments that will aid in determining the functional role of Kch. Chapters III details the *in vivo* approaches used to characterize Kch, revealing a novel function in

membrane potential modulation. In Chapter IV, I describe a swimming pattern formation first identified in a Δ*kch* mutant.

CHAPTER II

USING EVOLUTIONARY INFORMATION TO IDENTIFY NOVEL PROTEIN-PROTEIN INTERACTIONS

Introduction

Protein interactions are responsible for mediating nearly all biological processes; therefore, identifying novel interactions on a genome-wide scale is essential to understanding how biological systems function. Understanding how a multitude of pathways and reactions are integrated together at the cellular level remains a major challenge, even in extensively studied model organisms. Identification of protein-protein interactions on a genome-wide scale will provide a more comprehensive view of the evolution and maintenance of signaling networks and other pathways critical to cellular communication. However, reliably detecting *in vivo* interactions is not a trivial task.

Current methods employed for the detection of protein-protein interactions (PPI) may be able to detect high affinity physical interactions, but it is unclear how sensitive they are to transient, low affinity, or functional interactions[73]. These types of interactions are wide-spread and highly valuable for regulating cellular processes. Any type of protein modification (*i.e.*, phosphorylation, glycosylation, etc.) is mediated via a necessarily transient protein-protein interaction[74–78]. Additionally, many proteins, such as enzymes within the same biosynthetic pathway, may have genetic or indirect interactions mediated through a biosynthetic intermediate, a small molecule, or another protein. Many such interactions may go undetected because most techniques require a strong physical interaction between proteins for the interaction to be detected. Protein co-evolution is an attractive method for predicting protein-protein interactions

as it is based on the evolutionary history of the proteins and is not dependent on a strong physical interaction between proteins; it could be used to identify interactions between proteins and systems that are inherently difficult to detect with current methods[79–81].

Statistical coupling analysis (SCA) is a method to predict functional interactions between amino acids within proteins[70]. SCA has been successfully applied to protein families and has revealed the presence of functionally coupled networks of co-evolving amino acids that serve to connect the active sites to spatially distant allosteric sites in the proteins[70,71]. These co-evolving amino acid networks are sparse and reflect the requirement for communication between spatially distant regions of the protein. Here, we show that the application of SCA can be expanded to identify functionally connected networks of co-evolving amino acids between proteins. The presence of amino acid co-evolution between proteins would predict that those proteins are interacting either physically or functionally.

Materials and Methods

Creation of a large protein dataset

The Orthologous Matrix (OMA) database was utilized to obtain high quality orthologous protein sequences (December 2012 release)[82]. For a given protein, a multiple sequence alignment was generated by obtaining all the protein sequences within the OMA group that the protein is found in, along with all the sequences in the close groups. To automate this process for every protein in the *E. coli* genome, the OMA SOAP::API was queried using an in-house Perl script. To avoid ambiguity, the OMA close groups for each protein were compared and any shared groups were eliminated. The MSA was generated using the t-coffee quickaln function. Any MSA

containing fewer than 150 sequences was eliminated from the dataset. The final dataset contained MSAs for 1,717 proteins out of a total of 4,263 MSAs generated.

Large-scale co-evolution analysis

An all-against-all alignment step was performed for all 1,717 proteins in the dataset, where an MSA was concatenated to every other MSA to create a joint MSA. The concatenation step was performed in a species-dependent manner, meaning that the *E. coli* sequence in the first MSA was joined to the *E. coli* sequence in the second MSA, etc. Any protein sequences with >80% sequence identity were removed from the joint alignment to reduce the possibility of false positives. The joint MSAs were again screened for size, and any joint alignments with fewer than 150 sequences were discarded. The remaining joint MSAs were imported into Matlab for processing and analysis. Additionally, joint MSAs were further removed from analysis if 1) positions/perturbations were only found in one of the MSAs and 2) if the number of positions/perturbations in one MSA was less than 10% of the total number of positions/perturbations in the MSA. The co-evolution analysis was performed with the following occupancy and conservation parameters. Occupancy: All positions within the joint alignment were evaluated and any positions with >20% gaps were eliminated from the alignment. Conservation: The conservation parameters were set such that an informative site has a specific amino acid present 40-80% of the time. The program will perturb every position where the occupancy and the conservation requirements are met. With each perturbation, changes in the amino acid frequencies at all other sites were evaluated and recorded. The results of the analysis were reported as a matrix of free energy values and visualized as a heat map.

To automate the evaluation of joint alignments for co-evolutionary signal, the clustered heatmap is divided into nine sections and the mean signal in each section is calculated. The section with the most signal is used to determine if the joint alignment is positive or negative for co-evolution by evaluating the distribution of amino acid positions from each protein that cluster together. If the two proteins are co-evolving, positions within the second protein will respond to perturbations made in the first protein and vice versa and therefore cluster together. However, if two proteins are not co-evolving, positions in first protein will not respond to perturbations made to positions in the second protein and these will not cluster together. The statistical significance of the mixing of positions from both proteins is determined by an unpaired t test, with p values closer to 1 indicative of co-evolution and a p values closer to 0 indicative no co-evolution. Pairwise comparisons with P values =/> 1.0×10^{-12} were considered positive for co-evolution.

Comparison of known protein-protein interactions

The PDB database was downloaded from https://www.rcsb.org/ (accessed October 2014). The database was filtered to obtain PDB ids that represent co-structures and complexes of proteins that were also in the final co-evolution dataset of 1,717 proteins, and from that a list of unique protein-protein interactions was generated. Protein interaction data was downloaded from the MPIDB database http://www.jcvi.org/mpidb/ (accessed April 2015). The unique protein interactions were filtered to include proteins that were present in the final co-evolution dataset. For both sets of interaction data, the number of known protein interactions predicted to co-evolve was calculated, along with the number of interactions we were unable to evaluate and the number of interactions that were negative for co-evolution.

The PDB co-structures were filtered against the co-evolution dataset of 1,717 proteins to filter unique interactions. The PDB interaction data was separated into two sets: one set of unique PPI that excludes the ribosome and one set that includes the ribosome. For the non-ribosomal set, there were 78 co-structures that contained proteins in the co-evolution dataset. Of those 78 interactions, only 38 of the interactions could be evaluated by the co-evolution analysis. Thirty of the thirty-eight interactions were predicted to co-evolve. With the PDB dataset including the ribosome, every protein in the structure was evaluated. A total of 1,019 unique PPI were obtained from the PDB data. Of these 1,019 PPI, 134 interactions were unable to be evaluated, resulting in 885 interactions. Of the 885 interactions, 861 interactions were predicted by co-evolution. Of the 2,102 unique interactions from the MPIDB dataset, 724 interactions involved proteins in the co-evolution dataset. Of the 724 PPI, 438 interactions were unable to be evaluated, resulting in 286 PPI that could be analyzed. The SCA analysis found that 235 PPI were predicted to co-evolve. These results demonstrate that proteins with known interactions can be recovered using SCA.

GO term enrichment analysis

Broad classification of the proteins in the final dataset was obtained by mapping the proteins to a custom GO slim subset using the Map2slim.pl script [https://metacpan.org/pod/distribution/go-perl/scripts/map2slim]. The enrichment analysis was performed using custom Matlab scripts. For the protein of interest, the total number of evaluated proteins was calculated and compared to a subset of proteins with strong co-evolution scores ($N_{coevolve}$, P value =/> 1.0^{-6}). A 1,000 randomization trials were performed where we randomly selected the number of proteins equal to $N_{coevolve}$ from the total proteins that were evaluated. For

each trial, the number of proteins that mapped to each specific GO slim term was noted. The averages and standard deviations were determined for each group based on the randomizations and were used to calculate a Z-score to determine significance. GO slim terms with a Z-score of 1 or greater were considered to be biological processes that were enriched.

Results

Detection of co-evolving amino acid networks

Functionally connected sites within a protein can be identified through mutagenesis approaches[83], but the drawback to this method is that it requires significant time and resources to complete and is not feasible for large-scale analysis. Instead of relying on mutational experiments, we utilize a sequence-based approach to detect functionally interacting residues. We exploit the fact that evolutionary pressures lead to constraints within protein sequences, and that these evolutionary signatures can be detected and used to determine which sites within a protein communicate with each other[70,71].

Likewise, if two (or more) proteins are interacting in some capacity, there will be a mutual evolutionary constraint at the amino acid positions required for maintaining that interaction. Networks of co-evolving amino acids between proteins should impose a constraint such that changes within the networks of one protein would result in compensatory changes in the other protein. This constraint is detectable in a sufficiently diverse multiple sequence analysis (MSA). In order to evaluate two proteins for evidence of co-evolution, we must first generate a large, diverse multiple sequence alignment for each protein that accurately represents its evolutionary history. Orthologous protein sequences are used to create each MSA, as protein orthologs arise out of a speciation event and tend to retain similar evolutionary pressures and the

same function in other organisms. This is in contrast to paralogous sequences that arise out of a gene duplication event and tend to experience weaker selective pressures and diverge and evolve new functions[84].

Next, the two individual MSAs are joined together in a species-dependent manner, meaning that the *E. coli* K-12 sequence in the first protein MSA is joined to the *E. coli* K-12 sequence in the second MSA and so on (Fig 5). The resulting joint alignment is then evaluated for informative sites. If an amino acid position is not contributing to the overall structure or function of a protein, the lack of evolutionary pressure at that position should result in a more random distribution of amino acids[70], indicating that there is no strong selection for a specific amino acid at that site. However, if an amino acid position is important for the structure or function of a protein, there is expected to be a strong selective pressure, resulting in a preference for a specific amino acid(s) at that site. Thus, positions within the joint MSA whose amino acid distributions vary from a mean distribution (*i.e.*, have a strong preference for a specific amino acid at that site) are considered informative sites if they meet these occupancy and conservation thresholds.

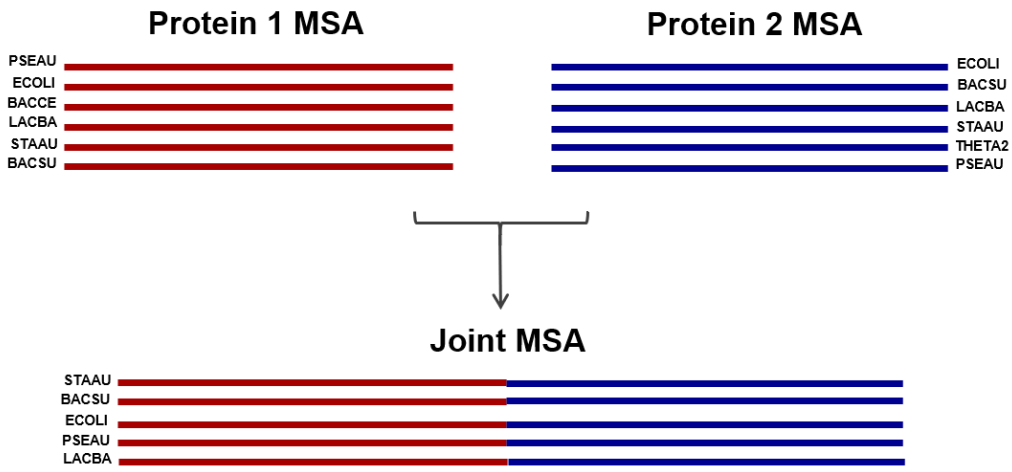

Figure 5. Generation of joint MSAs
The final dataset included MSAs for 1,717 individual proteins. These proteins were compared in a pair-wise manner. To create the joint alignment, the two individual MSAs are joined in a species-specific manner. The final joint MSA must have a minimum of 150 sequences to be further evaluated.

Once informative sites are identified, we next needed to determine which positions are co-evolving with each other (Fig 6). To accomplish this, a perturbation is made to an informative site by shifting the amino acid distribution. Any functionally interacting sites will experience a shift in their amino acid distributions in response to that initial perturbation. The response of the amino acid distributions at all other positions to a perturbation is recorded, and the next informative site is perturbed. Thus, making perturbations to all informative sites within the joint alignment and recording the responses to those perturbations, networks of co-evolving amino acids between proteins can be detected.

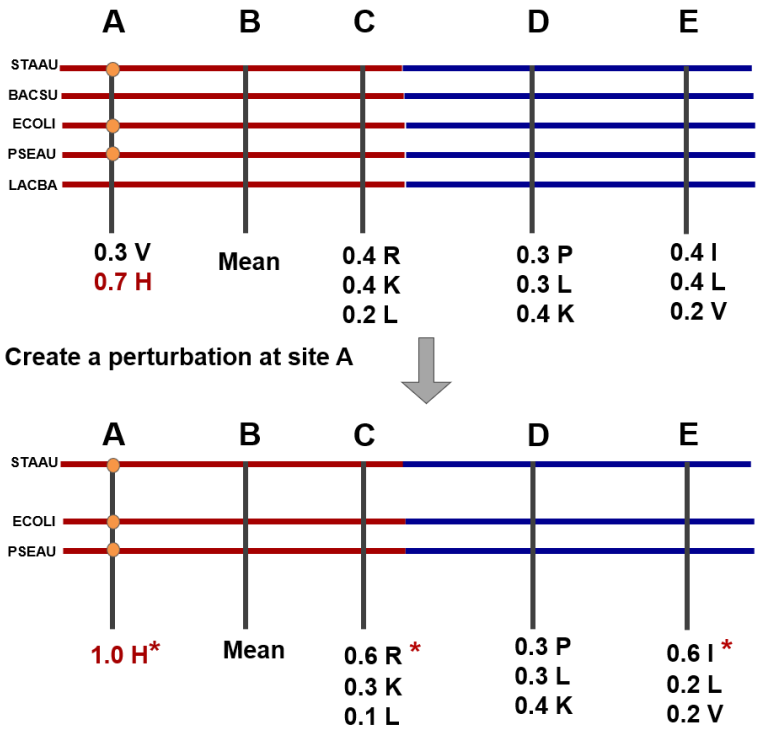

Figure 6. Detecting networks of co-evolving amino acids
For positions vital to the structure or function of a protein, it is expected that there would be evolutionary constraints at those sites, meaning amino acid frequencies would vary from the mean value. When site A is perturbed by selecting sequences with a H at that position, we observe a shift in the amino acid frequency at sites C and E. This indicates that those sites share a mutual constraint and are co-evolving with each other. Whereas site D is evolutionarily constrained but evolving independently of site A. Adapted from Süel *et al*[71].

Amino acid co-evolution can be detected between proteins with known interactions

We first asked if we could detect co-evolutionary signal between proteins by applying our method to two different sets of proteins known to physically and functionally interact: a subset of structural flagellar proteins and the Uvr endonuclease (Fig 7). These systems were selected for several reasons, the first being that both systems contain well-characterized functional and physical interactions. The biosynthesis and regulation of flagella is a complex biological process that includes more than 50 genes in *E. coli* [85]. Several of the Flg proteins were selected for analysis as they have critical structural roles in anchoring the flagella to the

membrane. The Uvr endonuclease is a multi-protein complex that mediates the response to UV-induced DNA damage[86,87]. The three Uvr proteins, UvrA, B, and C, interact in a pair-wise manner. First, UvrA-UvrB form a complex that scans DNA looking for DNA lesions. When a lesion is located, UvrA leaves, UvrC is recruited, and the UvrB-UvrC complex cleaves the DNA upstream of the damaged site[86,88]. These two control systems allow us to test the ability of SCA to detect both physical and functional interactions.

The second benefit of these systems is that they represent different gene arrangements. The Flg proteins are arranged within an operon and are subject to regulation by a common promoter. However, the Uvr proteins are not located within operons and are found in different regions of the chromosome. The genomic arrangement of the control proteins allows us to confirm that the presence of co-evolutionary signal is due to the presence of an interaction and not an artifact of being located within an operon. CorA, a magnesium transporter was selected for comparison with the control proteins, as it has no known or expected interactions with either of these systems[89].

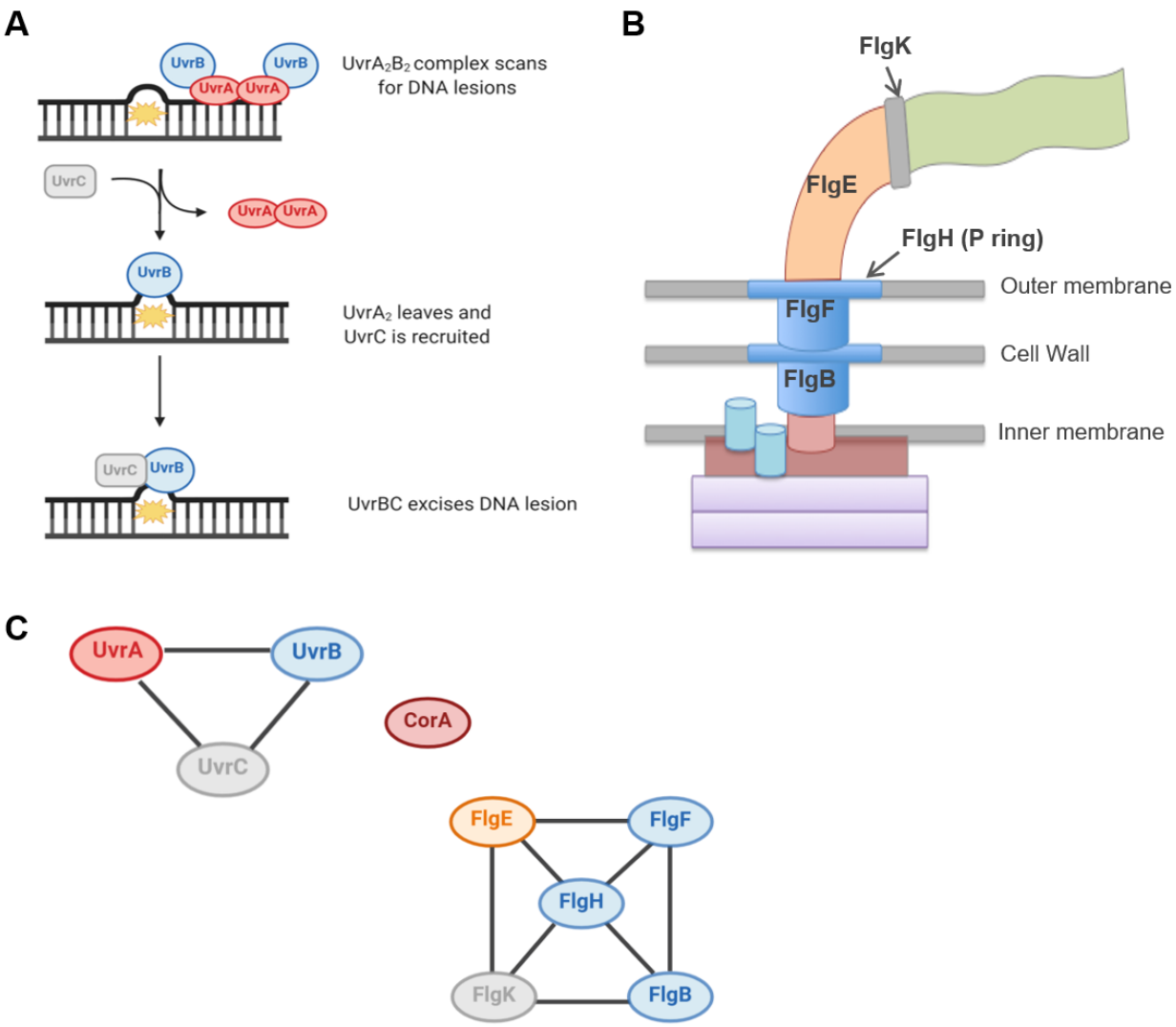

Figure 7. SCA can detect known protein interactions
To test the effectiveness of our method at predicting co-evolution between proteins, sets of proteins with known physical and functional interactions were selected as controls. (A) The UvrABC endonuclease is involved in the Nucleotide Excision Repair (NER) pathway, and (B) the Flg proteins are structural flagellar proteins. These systems also allow us to determine if we can predict protein interactions between proteins encoded at both nearby and distant regions of the chromosome, as genes encoding flg proteins are arranged within an operon, but genes encoding the Uvr endonuclease are in distant regions of the chromosome. (C) We find evidence of co-evolution between all Uvr proteins and between all flagellar proteins, and none between any proteins and CorA (negative control).

The proteins in each system were analyzed in a pairwise manner. We were able to detect evidence of co-evolution between all of the tested Flg proteins as well as between all three Uvr proteins (Fig 7 & 8), consistent with the known interactions between these proteins. No evidence of co-evolution was detected between CorA and any of the other proteins (Fig 9). These results indicate that we can detect co-evolutionary signal between proteins that have a physical interaction, as well as between proteins that are involved in the same complex but may not have direct physical connections (i.e., several of the Flg proteins and UvrA-UvrC). Importantly, this signal is not dependent on the genomic organization of the proteins. Next, we wanted to perform a large-scale analysis to predict PPIs on a genome-wide scale.

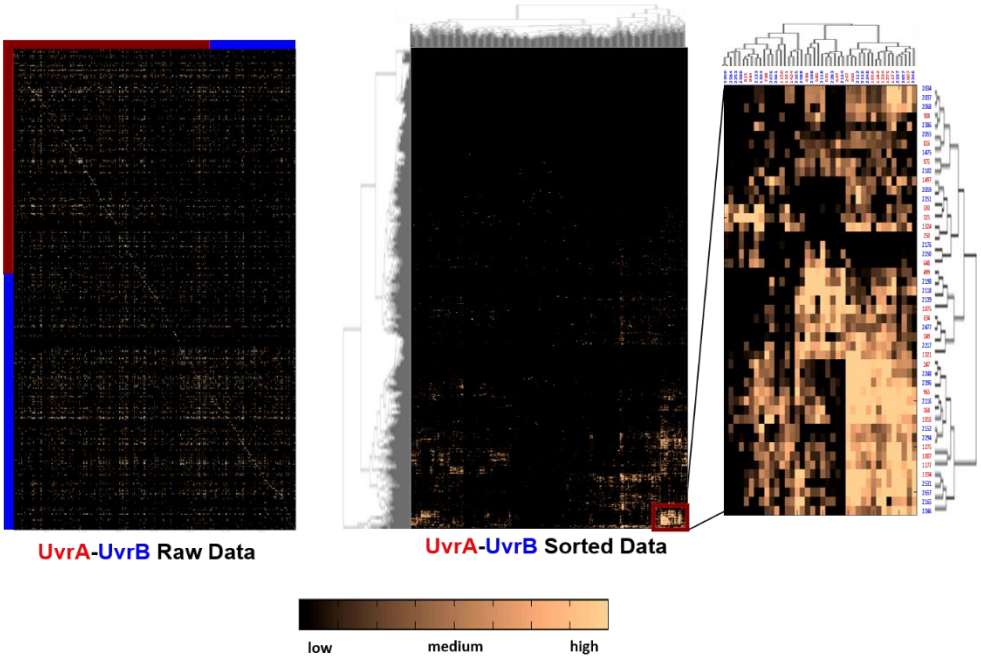

Figure 8. Co-evolution signal for proteins with known interaction- UvrA-UvrB
A heatmap is used to visualize protein co-evolution results. The left-most panel is the visualization of the signal from the linear sequence, where the X-axis represents the perturbations made in the joint alignment and the Y-axis represents all positions in the joint alignment. Perturbations and positions flanked by red are those found in UvrA, and perturbations and positions flanked by blue are found in UvrB. The middle panel is the raw data after a round of two-dimensional hierarchical clustering. If two positions are co-evolving with each other, they will have similar coupling patterns to all perturbations and will be clustered together. Likewise is true for two perturbations, if they have similar patterns of coupling to all positions they would be clustered together. We can visually capture the strength of the interaction based on the color scale (low=black high=copper). The right-most panel is a subset of the clustered dataset. The clustering together of positions and perturbations from both proteins (designated by the respective colors) indicates that these two proteins are co-evolving.

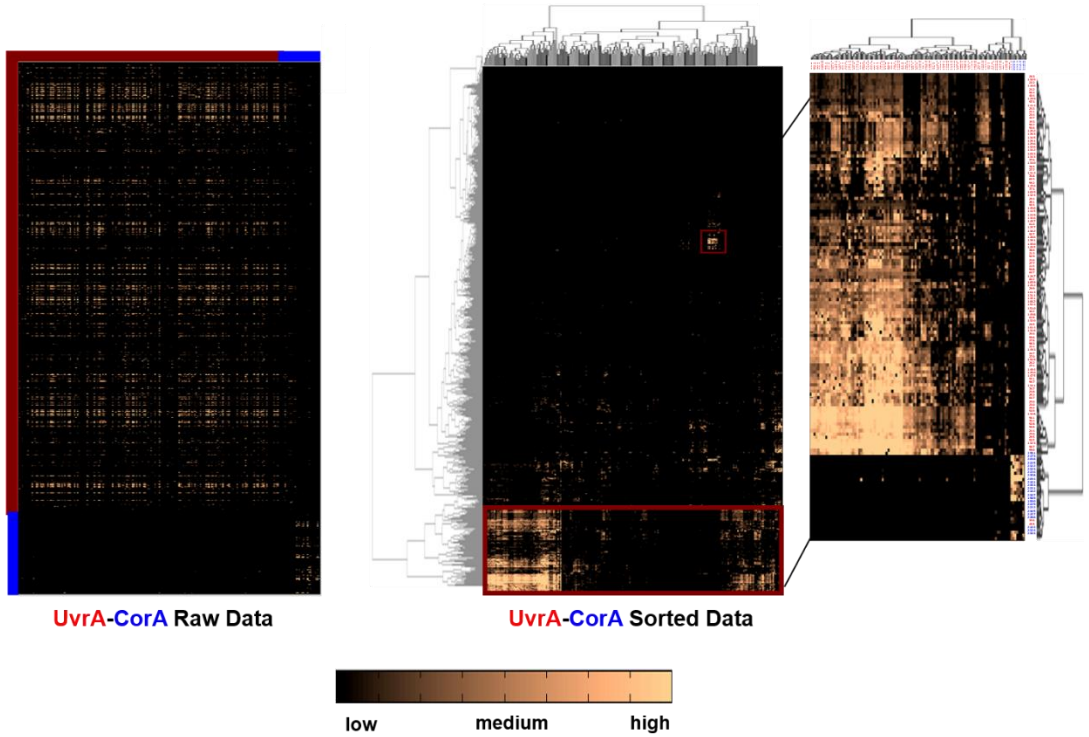

Figure 9. Co-evolution signal for proteins with no known interactions - UvrA-CorA
CorA was used as a negative control because there were no known or expected interactions between our control proteins and CorA. An example of the signal seen with with CorA is shown here in the UvrA-CorA joint alignment. The left-most panel is the visualization of the signal from the linear sequence, where the X-axis represents the perturbations made in the joint alignment and the Y-axis represents all positions in the joint alignment. Perturbations and positions flanked by red are those found in UvrA, and perturbations and positions flanked by blue are found in CorA. The middle panel is the raw data after a round of two-dimensional hierarchical clustering. The right-most panel is a subset of the clustered dataset. Clustering of positions and perturbations is observed in each protein individually, but the lack of clustering between UvrA and CorA indicates that these proteins evolve independently of each other as expected.

Generation of a dataset of bacterial proteins for SCA analysis

We selected the *E. coli* K-12 genome, which encodes ~4,300 proteins, to serve as the template genome for a large-scale protein co-evolution analysis. As a model organism, the *E. coli* genome benefits from a high-level of characterization and annotation, but a significant

number of proteins (~30%) remain uncharacterized or poorly annotated[90]. Poorly characterized proteins may have unknown roles in characterized biological processes or represent entirely novel cellular pathways; information on predicted interaction partners may help link uncharacterized proteins to known proteins or pathways. Thus, analysis of the *E. coli* genome will allow for validation of our method by identification of known protein interactions and complexes, as well as provide insights into novel interactions through the analysis of poorly characterized proteins.

To create a dataset for analysis, a multiple sequence alignment (MSA) must be generated for every protein in the *E. coli* genome (Fig 10 A). Orthologous protein sequences were used to create each MSA to ensure that the evolutionary history of the protein is represented. Because orthologous proteins arise out of a speciation event, they tend to retain the same function and similar evolutionary pressures in different organisms. In contrast, paralogous proteins arise from a gene duplication event, and they often diverge and evolve new functions due to weaker selective pressure[84]. Next, the MSAs for each protein are joined together in a species-dependent manner, meaning that the *E. coli* protein sequence from the first MSA is joined to the *E. coli* protein sequence in the second MSA. Another size threshold is imposed and joint alignments with <150 sequences are eliminated. This resulted in 1,717 x (1,717 -1) ÷ 2 = 1,473,186 pairwise comparisons. Of all possible comparisons, ~500,000 comparisons had enough sequences to be analyzed further (Fig 10B).

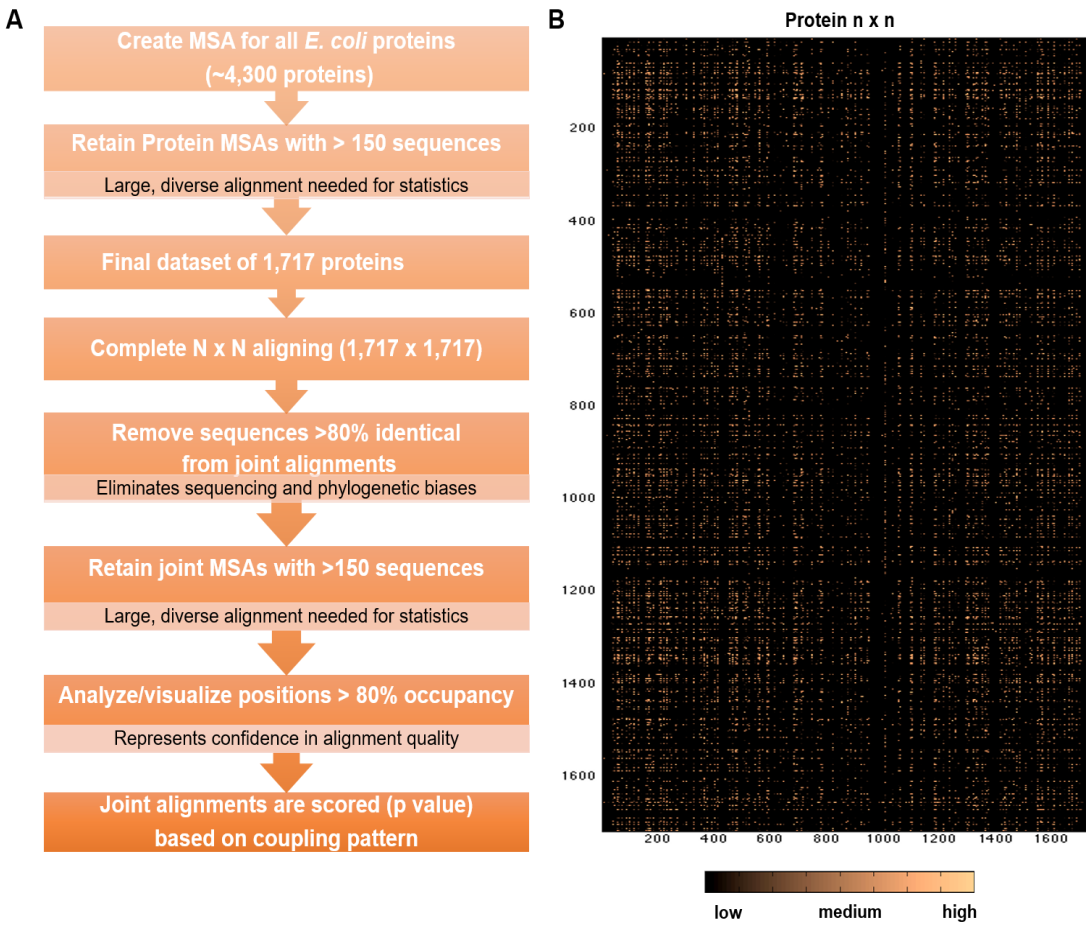

Figure 10. Generation of a large protein dataset

In order to perform a large-scale co-evolution analysis, we first had to generate a dataset of proteins that could be further evaluated. (A) Workflow to generate large protein dataset. The *E. coli* K-12 genome was used as a template, and MSAs were built for every K-12 protein using orthologous sequences obtained from the OMA database. Any proteins with an MSA containing fewer than <150 sequences were removed from the dataset. The final dataset contained 1,717 proteins. In an all-against-all joining step, all 1,717 proteins were aligned to all other proteins in the dataset. Joint alignments were further filtered to remove any sequences >80% identical, and any joint alignments with <150 sequences were not analyzed. Within each joint alignment only positions with >80% occupancy are evaluated. Joint alignments are evaluated for co-evolution and scored based on the coupling pattern. (B) Raw matrix for the n x n comparison. Pairwise comparisons that show no evidence of co-evolution have low values (black color) and those that are positive for co-evolution have a high value (copper color).

Majority of pair-wise interactions are positive for co-evolution

One of the unexpected results of our large-scale analysis was a large amount of positive co-evolutionary signal. On average, a given protein was predicted to co-evolve with ~60-80% of the evaluated proteins. This was a much higher number than we anticipated, but several key steps were taken to minimize factors that could result in false positive signal. The first step was imposing a stringent sequence identity cutoff during the formation of the joint MSAs. The bacterial genomes for which whole-genome sequencing is available is heavily biased towards model laboratory organisms such as *E. coli* and *B. subtilis*[91,92]. When creating protein MSAs, this bias in the sequencing databases can result in an MSA that may have a large number of sequences but is significantly biased towards specific species. To avoid signal that would erroneously arise from a lack of sequence diversity, we imposed a sequence identity threshold and removed highly similar sequences (>80% identity). Secondly, we imposed a size threshold for the number of sequences in the joint alignment. To ensure that the joint MSA is sufficiently diverse, we required that the final joint MSA contain a minimum of 150 sequences after the removal of highly similar sequences. The use of stringent parameters likely reduces the total number of interactions that can be evaluated, but drastically increases our confidence that the signal obtained is indicative of an interaction. The high amount of positive signal likely reflects the highly conserved nature of the proteins included in our final dataset (see Discussion).

Large-scale co-evolution analysis recovers known PPI networks

Our initial data suggested that proteins that are coevolving are also interacting in some capacity (Fig 7). In order to determine how well co-evolution predicts interactions, we compared our results to two datasets of known interactions: the Protein Data Bank (PDB) and the Microbial

Protein Interaction Database (MPIDB). Data from the PDB database represents physical interactions, and the MPIDB provides curated interactions supported by at least one experimental evidence[93,94]. Interactions from these two datasets were filtered against the proteins in the coevolution dataset, and the number of protein-protein interactions predicted by co-evolution was calculated. Overall, the overlap in the number of known interactions and those that were able to be evaluated in our dataset was low (see methods), but in the subset of known interactions evaluated in our analysis, we were able to recover a significant number (235/286 interactions from MPIDB and 30/38 from PDB – see methods) (Fig 11). These included well-described interaction networks including flagellar and chemotaxis proteins and cytoskeletal proteins. Consistent with our preliminary results, this result demonstrates that our analysis can recover known interactions and interaction networks in a large-scale analysis.

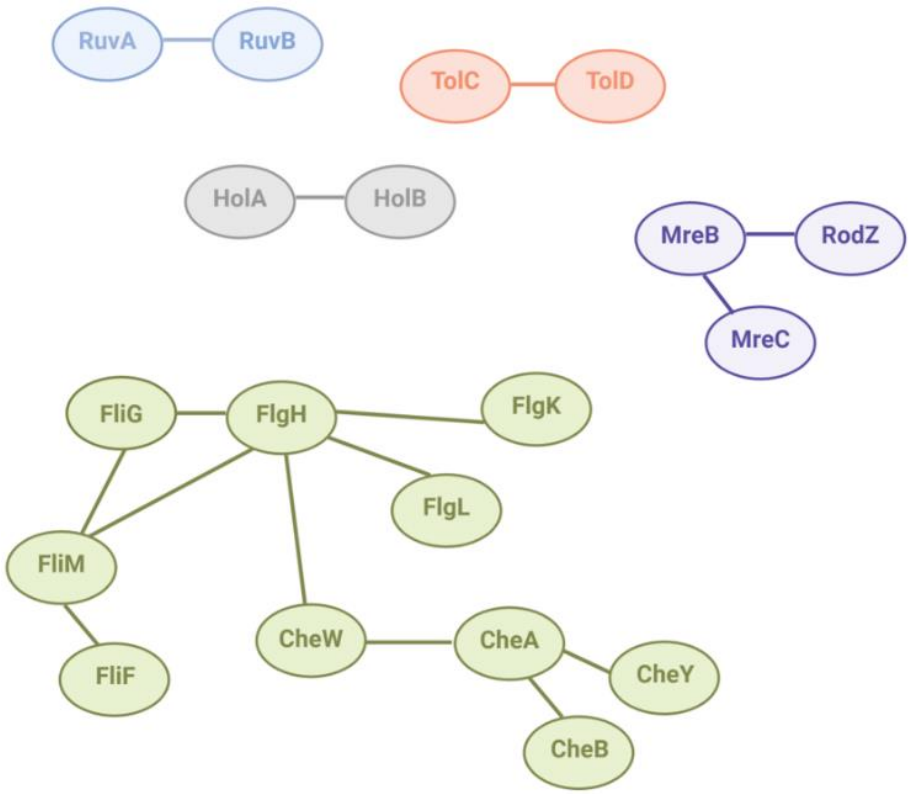

Figure 11. Large-scale co-evolution analysis recovers known PPI networks
Known protein-protein interactions and interaction networks were recovered in the co-evolution analysis. All experimentally reported PPI, as indicated by lines between proteins, were predicted to co-evolve.

GO term enrichment analysis

Given the large number of predicted interactions generated by this analysis, it can be difficult to decipher patterns from a list of protein names that are predicted to interact. In order to better understand the cellular role of individual proteins, we wanted to be able to determine what broad biological pathways a given protein is interacting with. To do this, we utilized Gene Ontology (GO terms) to broadly classify the proteins in our dataset into biological processes[95]. Next, for a protein of interest, we performed a GO term enrichment analysis by determining which biological processes are significantly over-represented within the subset of protein

partners predicted to co-evolve. This type of analysis allows us to visualize which biological systems a given protein is interacting with, yielding insight into its function. For example, analysis of UvrB, a protein involved in mediating repair of UV-induced DNA damage, is predicted to interact with proteins involved in DNA repair, DNA recombination, DNA replication, and Translation consistent with the known biological function of UvrB (Fig 12). Use of the gene ontology enables us to quickly determine which pathways the protein of interest may be involved in.

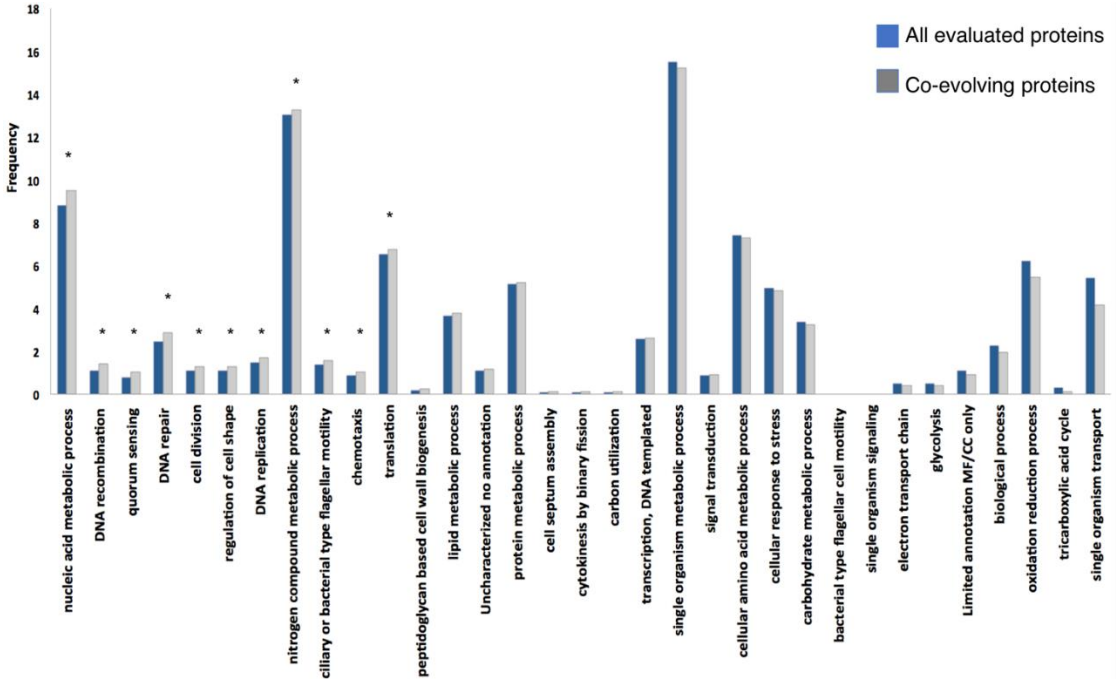

Figure 12. GO term enrichment analysis for UvrB
Predicted interaction partners are enriched for nucleic acid metabolic processes, DNA recombination, DNA repair, DNA replication, regulation of cell shape, Translation, and motility/chemotaxis processes. Biological processes with an asterisk have a Z score > 1 and are considered significantly enriched.

Application of co-evolutionary information to poorly characterized proteins

One of the most useful applications of the co-evolution analysis is to proteins that lack thorough experimental characterization. Of the 1,717 proteins in the final dataset, 394 (~23%) are proteins that carry y-names, indicating that they are poorly characterized. Although their cellular function may be unknown, the fact that they are highly conserved among bacterial species implies that they have significant physiological roles. Prediction of interaction partners for y-proteins, links them to biological pathways, allowing for the inference of function. Additionally, the use of co-evolutionary information to predict interactions has the potential to reveal not just novel interactions, but novel protein complexes and networks. For example, GO term enrichment analysis of a y-protein in our dataset, YrbD, revealed that the only biological process significantly enriched was Lipid Metabolic processes. YrbD was predicted to interact with several other y-proteins as well as proteins involved in lipid metabolism, indicating that this may represent a novel pathway (Fig 13). Recent experimental characterization of YrbD revealed that it formed a complex with YrbE, YrbF, YrbB, and together with YrbA and YrbC functions as a retrograde outer membrane phospholipid trafficking system. These y-proteins were renamed MlaA-F for maintenance of outer membrane lipid asymmetry[96,97]. This result demonstrates that co-evolutionary information can be successfully applied to poorly characterized proteins to reveal novel interactions that yield insight into their function.

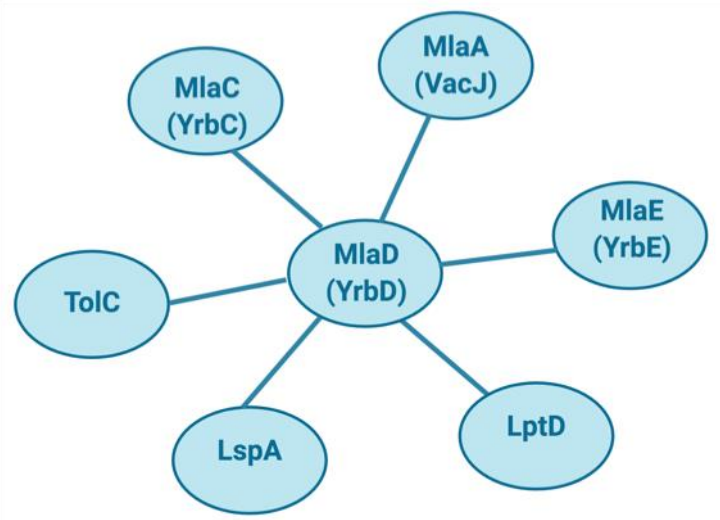

Figure 13. Application of co-evolutionary information to uncharacterized proteins
Characterization of YrbD revealed that the y-proteins were all components of an outer membrane phospholipid trafficking system. A small subset of predicted interaction partners are represented above. TolC is an outer membrane protein involved in efflux[98]. LptD is an outer membrane protein involved in LPS assembly at the outer membrane[99]. LspA is a prolipoprotein signal peptidase[100].

Kch is predicted to interact with proteins involved in redox, cell size regulation, and metabolic processes

Results from our analysis indicate that co-evolutionary information can recover known interactions as well as be used to predict novel interactions. We next wanted to apply this analysis to Kch, a relatively uncharacterized K^+ channel in the *E. coli* genome. Attempts at characterization of Kch have had limited success; therefore, many questions remain about the in vivo function of this protein. We were able to evaluate Kch with 161 proteins, of which, 94 were predicted to co-evolve. GO term enrichment analysis revealed that Kch was predicted to interact with proteins in oxidation-reduction processes, cell size regulation, and metabolic processes. Information regarding which processes in the cell Kch is predicted to interact with will allow us to design more targeted experiments to elucidate the *in vivo* function of the channel. Subsequent chapters of this dissertation present experiments designed to validate these predictions.

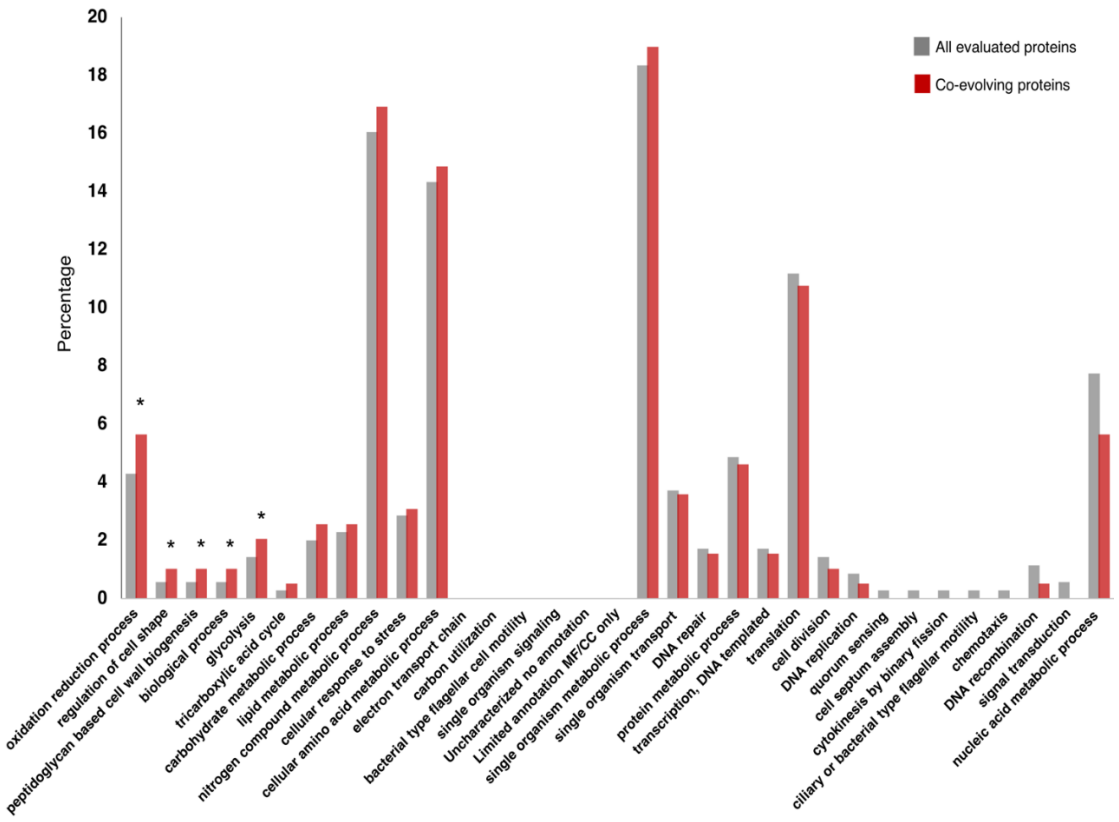

Figure 14. GO term enrichment analysis for Kch
Predicted interaction partners are enriched for oxidation-reduction processes, regulation of cell shape, cell wall biosynthesis and metabolic processes. Biological processes with an asterisk have a Z score > 1 and are considered significantly enriched.

Discussion

Protein interactions are vital to carrying out all cellular processes. Identification of interacting proteins is important for understanding how the various cellular pathways cooperate and how information is propagated within the cell. Detection of PPIs is difficult as some interactions may be low affinity, condition-dependent, or transient and are therefore hard to capture with traditional methods. Here we applied a sequence-based co-evolutionary analysis to

predict protein interactions on a genome-wide scale. Application of our analysis to a large dataset of conserved bacterial proteins (>1,700) both recovered well-established protein interactions and as well as predicted novel interactions.

One unexpected result from the analysis was the high number of pairwise protein comparisons that were positive for co-evolution. One factor that likely contributes to the high percentage of predicted interactions is the stringent criteria we imposed during the creation of the MSAs and subsequent evaluation of joint alignments (elimination of close groups, total number of sequences >150, sequence identity of 80 % or less). The final dataset included 1,717/4,263 proteins (~40%), and these proteins represent a highly conserved subset of the proteome. Critical cellular processes must be able to cooperate and communicate with each other in the cell; therefore, it would be expected to see a large number of predicted interactions. Additionally, because this method is not limited by experimental conditions, we have the ability to predict a larger number of interactions than traditional methods. For example, some protein-protein interactions may only occur under certain conditions such as specific growth stages, in the presence of a specific stressor, certain temperatures or media, or the interactions could require the presence of multiple proteins simultaneously. These types of interactions could be missed by more traditional PPI detection methods. This approach is also not limited by the strength of the interactions, meaning transient or functional interactions will be predicted in this analysis. Another factor that could also contribute to the high percentage of positive interactions is that not every interaction could be evaluated. This missing information would likely contain interactions that would be negative for co-evolution, lowering the number of average connections significantly.

Using the co-evolutionary analysis, we demonstrated that proteins with known interactions were also predicted to co-evolve; we were able to recover known PPI as well as PPI networks (Figs 11). Coupling the co-evolution results with a GO term enrichment analysis, allows us to predict which biological pathways a given protein is interacting with. Application of this analysis to uncharacterized y-proteins revealed novel interactions between YrbD (now MlaD) and proteins involved in lipid metabolic processes and other uncharacterized proteins (Fig 13). Characterization of YrbD (MlaD) by other laboratories revealed that these uncharacterized y-proteins represented a novel retrograde phospholipid trafficking complex in the outer membrane[96,97]. This result demonstrates the ability of co-evolutionary information to predict novel interactions and highlights the value in application of co-evolutionary information to predict the function of uncharacterized proteins.

Future Directions

We hypothesized that co-evolutionary information could be used to recover known protein-protein interactions as well as predict novel interactions. Our large-scale implementation of SCA does indeed recover well-characterized protein interactions (Figs 7 & 11). We also demonstrated that co-evolutionary information could be useful in predicting both novel interactions and potential biological functions for uncharacterized proteins (Fig 13). Just over 20% of the proteins in the final protein co-evolution dataset are y-proteins that are widely conserved in bacterial genomes but have limited-no-functional information. As was seen in the case of YrbD (MlaD), the exploration of uncharacterized proteins will reveal novel protein complexes and novel biological pathways, increasing our understanding of bacterial physiology.

Future work will focus on further analysis of y-proteins and other poorly characterized proteins. GO term enrichment analysis for y-proteins will be performed to determine which biological processes they may be involved in. Y-proteins could be grouped based on which biological processes are enriched to facilitate validation. For example, if a subset of y-proteins were enriched for protein partners in chemotaxis or other motility processes, deletion strains from the Keio collection could be screened for motility defects or altered chemotactic responses. If a subset were enriched for protein partners involved in DNA repair, these deletion strains could be screened for altered responses to DNA damaging agents. The use of co-evolutionary information provides a direction to focus experiments to begin validation of predictions. Further investigation of the predictions from this study will no doubt result in a clearer understanding of the role of uncharacterized proteins, providing insight into their conservation in bacteria.

CHAPTER III

FUNCTIONAL CHARACTERIZATION OF THE *ESCHERICHIA COLI* K^+ CHANNEL, KCH

Introduction

Potassium channels are highly conserved proteins found in all domains of life that allow for selective movement of K^+ ions across the membrane. The physiological roles of individual members of this diverse channel family are best understood in eukaryotic systems, where they have critical functions in a variety of cellular processes including signaling, muscle contractions, cell size regulation, electrolyte balance, and cardiac function[3,7,101]. The wealth of functional characterization of eukaryotic K^+ channels has resulted in a clearer understanding of how channel dysfunction contributes to disease states. Indeed, disruption of K^+ channel function underlies many cardiac, neuronal, renal, and metabolic disorders, and this understanding of how channel function affects physiology has led to the successful use of eukaryotic K^+ channels as pharmacological targets to alleviate pathologies[102,103].

Although many of the initial structural and biophysical studies that gave insight into eukaryotic K^+ channel function were performed on bacterial homologs, the contribution of K^+ channels to bacterial physiology remains poorly understood. Historically, bacterial K^+ channels were hypothesized to function primarily in K^+ homeostasis and osmoregulation due to their selective transport of K^+ and the ability of K^+ to serve as an osmoprotectant that can be accumulated to high intracellular concentrations[63]. However, the majority of bacterial K^+ channels have no demonstrated functional characterization, leaving their physiological role unclear. Recently, K^+ channels have been implicated in intercellular communication within

biofilm communities, suggesting that they may have a more substantial role in bacterial physiology than previously expected[11].

E. coli possesses a single K^+-selective channel, *kch*, which has the distinction of being the first described bacterial K^+ channel[11]. As a member of the voltage-gated K^+ channel superfamily, Kch retains several key structural features of eukaryotic K^+ channels, including monomers with six transmembrane domains that assemble to form a homotetramer, a conserved K^+-selectivity filter in the pore region, and the presence of a C-terminal RCK domain. However, it is unlikely that Kch responds to voltage changes, as the amino acid motif associated with voltage-sensing is degenerate[66,104]. Despite its serendipitous discovery more than twenty years ago, initial attempts at Kch characterization were plagued by conflicting results and much remains to be understood about how the channel functions in *E. coli* physiology.

Early studies demonstrated that the Kch protein localized to the inner membrane and displayed behaviors characteristic of K^+ channels[13]. Alteration of channel concentrations through overexpression or deletion yielded drastically different results between laboratories. For example, one study found that *kch* overexpression resulted in increased membrane permeability for K^+, but that known inhibitors of K^+ channels failed to block the K^+ leakage[65]. Another study found overexpression of *kch* to be lethal, but supplementation of the media with 50 mM extracellular K^+ suppressed the lethality of overexpression[14]. Yet another study found a conflicting result that the toxicity of *kch* overexpression could not be rescued by external K^+. In fact, both extracellular K^+ and Na^+, but not equi-osmolar sorbitol, were found to aggravate the toxicity phenotype. Additionally, the toxicity persisted even when over-expressing a mutant with a collapsed selectivity filter, suggesting an issue with ionic strength and overexpression of a

membrane protein and not from *kch*-specific overexpression[14]. The discrepancies between various studies make it difficult to draw reliable conclusions from these experiments.

The most direct evidence that Kch forms a K^+-selective channel *in vivo* comes from a gain-of-function screen[14]. Kch was cloned into a plasmid, randomly mutagenized, and then screened in a Δ*kch* background in the absence of inducer to minimize lethal effects from over-expression. Mutants were isolated that grew normally on LB, but were sensitive to 200 mM extracellular K^+, as these likely represent channels with a higher open probability (gain-of-function). Next, suppressors of this phenotype were selected, and the suppressor mutations were mapped and found to reside within the selectivity filter, suggesting that the channel is functional *in vivo* and capable of transporting K^+ ions[13,14].

Similarly, attempts at analyzing *kch* null mutants resulted in opposite phenotypes. Initial reports of a Δ*kch* strain reported no new K^+-selective transport phenotypes or any additional obvious phenotypes[68]. A random transposon-mutagenesis screen found that *kch* was dispensable for growth at low temperatures in rich media or in minimal media at 37 °C, but that it was required for growth of *E. coli* in rich media at 37 °C[69]. However, under similar growth conditions (LB, 37 °C), a viable *kch* null mutant was generated during the creation of the Keio collection, marking *kch* as a non-essential gene[105]. Similarly, a large-scale transposon screen for essential genes also reported *kch* as a non-essential gene[106].

Given the failure of classical techniques in revealing the *in vivo* role of *kch*, we used a novel computational assay to gain insight into this cryptic K^+ channel. We used statistical coupling analysis (SCA) to predict which proteins in the cell Kch was co-evolving with and interacting with in the cell. Being able to connect Kch to proteins of known functions could help

reveal the biological processes that the channel is involved with, as well as aid in designing targeted experiments for further validation.

Materials and Methods

Strains and Reagents

Strains used in this study are listed in Table 1. The BW25113 Δ*kch::kan,* Δ*ubiH::kan,* and Δ*trpC::kan* strains were obtained from the Coli Genetic Stock Center (CGSC). The Δ*kch::frt* strain (SDB6) was generated following transformation of SDB2 with the pCP20 plasmid[107]. The Δ*kch::kch*$^{+}$ strains (SDB222 & 223) were generated by transducing in a closely linked auxotrophic marker, Δ*trpC* from SDB9. The Δ*trpC::kan* deletion was rescued via P1*vir* transduction with a wild-type lysate from SDB1, and transductants were selected for growth on M9 Glucose media lacking tryptophan. The Δ*ubiH::ubiH*^{V223G} strain (SDB11) was generated by transducing in the *ubiH*^{V223G} mutation from SDB2 into the Δ*ubiH* strain. Transductants were selected for their ability to grow on minimal media with succinate as the sole carbon source. The *ubiH*^{V223G}*::ubiH*$^{+}$ rescue strain (SDB28) was generated by conjugation of SDB11 with the Hfr strain EA1005. Exconjugants were selected for on M9 glucose media with chloramphenicol. All genotypes were confirmed via both PCR and Sanger sequencing. Independent BW25113 Δ*kch::kan* isolates were made by using P1*vir* to transduce the Δ*kch::kan* cassette from SDB2 into the SDB1 strain. Each isolate was obtained on selective media from an independent transduction, and was PCR checked to confirm the Δ*kch::kan* deletion prior to WGS.

Table 1. Strains and Plasmids used in Chapter III

Strain or Plasmid Number	Description	Relevant Genotype	Source or Reference
Strains			
SDB1	BW25113	*F⁻ DE(araD-araB)567 lacZ4787(del)::rrnB-3 LAM⁻ rph-1 DE(rhaD-rhaB)568 hsdR514 wrbA^{G96V} NemA^{Y359N}*	Paul Straight
SDB2	BW25113 Δ*kch::kan*	Δ*kch::kan ubiH^{V223G} lrhA198::IS5*	CGSC #9121
SDB6	BW25113 Δ*kch::frt*	Δ*kch::frt ubiH^{V223G} lrhA198::IS5*	
SDB8	EA1005	KL228 *srl::Tn10*	Ry Young
SDB9	BW25113 Δ*trpC::kan*	Δ*trpC::kan*	CGSC #9131
SDB10	BW25113 Δ*ubiH::kan*	Δ*ubiH::kan*	CGSC #10230
SDB11	BW25113 *ubiH^{V223G}*	SDB10 *ubiH^{V223G} lrhA198::IS5*	This study
SDB28	BW25113 *ubiH^{V223G} ::ubiH⁺*	SDB11 *srl::Tn10*	This study
SDB222	BW25113 Δ*kch::kch⁺*	*kch⁺ ubiH^{V223G} lrhA198::IS5*	This study
SDB223	BW25113 Δ*kch::kch⁺*	*kch⁺ ubiH^{V223G} lrhA198::IS5*	This study
Plasmids			
pCP20	Used to excise selection marker	Vector carrying Flippase used to excise kanamycin marker	Paul Straight
pSB12	Addgene vector #44249	Vector expressing dCas9 from aTc-Inducible promoter	Addgene
pSB14	Addgene vector #44251	Vector expressing mRFP-targeting sgRNA from a constitutive promoter	Addgene
pSB24	44251 + *kch* seed region	Vector expressing *kch*-targeting sgRNA from a constitutive promoter	This study
pSB25	44251 + *araC* seed region	Vector expressing *araC*-targeting sgRNA from a constitutive promoter	This study

Strains were routinely cultured in Luria-Bertani broth (10 g/L tryptone, 10 g/L NaCl, and 5 g/L yeast extract) or Tryptone Broth (10 g/L tryptone and 8 g/L NaCl). M9 minimal media was prepared as described[107]. EZ rich defined media with glucose or succinate as the sole carbon source was obtained from Teknova Inc (https://www.teknova.com/mops-ez-rich-defined-medium-kit.html). 10X MOPS buffer was prepared using the following recipe: 83.7 g/L MOPS, 13.6 g/L Sodium Acetate, trihydrate, 3.7 g/L EDTA, disodium dihydrate. NaOH was used to adjust the pH to 7.0. Antibiotics were used at the following concentrations: kanamycin, 50 μg/mL, ampicillin, 100 μg/mL, and chloramphenicol 25 μg/mL. Membrane potential dyes, $DiOC_2(3)$ and Thioflavin T, were obtained from Thermofisher and Acros organics, CAS: 2390-

54-7 respectively. The dyes were used in final concentrations of 30 μM for $DiOC_2(3)$ and 10 μM for Thioflavin T.

Plasmids carrying the CRISPRi systems (Addgene numbers 44249 (dCas9) and 44251 (sgRNA)) were kind gifts of Stanley Qi. Seed regions designed to target either *kch* or *araC* genes were designed as described[108]. These regions were cloned into the sgRNA plasmid (Addgene # 44251) using the IVA cloning method[109]. Primers were obtained from Integrated DNA Technologies (IDT).

Growth curves

Overnight cultures were diluted 1:100 into 50 mL of fresh, pre-warmed Tryptone broth (TB) in 250 mL flasks and were grown at 37 °C with aeration (250 RPM). OD_{600} readings to monitor growth were taken every 30 minutes. Carbon sources were supplemented to a final concentration of 0.4%[110].

H_2O_2 challenge assays

Freshly streaked isolates were used to start 3 mL overnight cultures in LB in 15 mL round-bottom tubes and incubated at 37 °C. Each overnight culture was diluted 1:100 into two independent 125 mL flasks containing 30 mL of fresh LB. Growth was monitored and once they reached an OD_{600} ~= 0.35, cells were challenged with LB + H_2O_2 (final concentration of 2.5 mM) or an equivalent volume of LB only. Growth was monitored every twenty minutes using OD_{600}.

Growth profiles during CRISPRi knockdowns

E. coli BW25113 (SDB1) strain was transformed with either a dCas9 plasmid or a dCas9 + sgRNA plasmids. Freshly streaked isolates were used to start 3 mL overnight cultures in LB in 15 mL round-bottom Falcon tubes and were grown for 18 hrs at either 30 °C or 37 °C with 250 RPM. Overnight cultures were diluted 1:100 into fresh media with either 2 μM anhydrous tetracycline (aTc) or EtOH only and aliquoted in 200 μL volumes into a 96-well plate. The 96-well plate was incubated in the dark at either 30 °C or 37 °C with 250 RPM aeration to protect the inducer from light. OD_{600} readings to monitor growth were taken on a BMG PolarStar Omega plate reader (BMG Labtech Inc, USA).

Oxygen consumption measurements

Oxygen consumption measurements were performed on an Oroboros oxygraph-2k (http://www.oroboros.at, Oroboros Instruments, Innsbruck, Austria) as described[111]. A volume of 2 mL of EZ rich defined succinate media (Teknova Inc) was used to fill the chambers prior to the addition of cells. Overnight cultures grown in the same media were diluted 1:100 into pre-warmed fresh media and grown at 37 °C with aeration until early exponential phase (OD_{600} = ~0.2). Five milliliters of cells were pelleted by centrifugation at 4,000 RPM for 3 minutes, resuspended in 1 mL of 1X MOPS buffer, adjusted to an OD_{600} = 0.5, and a 100 μL volume was injected into the chambers. O_2 consumption was monitored for at least 15 minutes. The oxygen consumption rate (OCR) was calculated for each strain and normalized to the SDB1 strain.

Membrane potential measurements with $DiOC_2(3)$

Measurements were performed as described previously[112]. Overnight cultures were diluted 1:1000 into pre-warmed fresh LB and grown at 37 °C with aeration until mid-exponential phase (OD_{600} = 0.5-0.6). Cells were harvested by centrifugation at 2,400 x g for 10 mins at room temperature and resuspended in 1X PBS. EDTA (10 mM) was added and cells were incubated for 5 mins at room temperature to permeabilize the outer membrane. Cells were pelleted by centrifugation at 2,400 x g for 10 mins at room temperature and resuspended in resuspension buffer composed of 130 mM NaCl, 60 mM Na_2HPO_4, 60 mM NaH_2PO_4, 10 mM glucose, 5 mM KCl, and 0.5 mM $MgCl_2$. The pH was adjusted 7.0 using NaOH, and the resuspension buffer was sterilized using a 0.22 µM filter. Resuspended cells were dispensed into opaque 96-well plates along with $DiOC_2(3)$ to a final concentration of 30 µM, and membrane potential was monitored using a Varioscan Lux plate reader with the following parameters: excitation at 450 nm and emission at 500 nm and 670 nm.

Membrane potential measurements with Thioflavin T during Kch depletion

Membrane potential measurements with Thioflavin T (ThT) were performed as previously described[113–116]. Briefly, freshly streaked strains were used to inoculate 3 mL of EZ rich defined glucose media and grown at 37 °C with aeration until early exponential phase. Cultures were loaded into a commercial microfluidics plate (EMD Millipore) and was connected to a CellAsic Onix2 device to control media flow. Base media (EZ rich defined glucose media with appropriate antibiotics and 10 µM ThT) was flown into each chamber at a rate of 1.5 psi throughout the experiment and the chamber temperature was maintained at 37 °C. After a period of flowing base media only, 2 µM aTc was added to induce expression of the CRISPRi system.

Three different locations in each chamber were imaged every 15 minutes under a 40X objective lens using an Olympus inverted microscope with a motorized stage.

Image analysis and ThT quantification

Image analysis was performed using Fiji (Image J)[117]. Cell boundaries were automatically detected by using the threshold function on the phase contrast image, with manual adjustment of cell boundaries performed as needed. These values obtained were saved to the ROI and then applied to the corresponding fluorescent image to obtain cell area and mean ThT intensity.

Statistical analysis

All experiments were performed with a minimum of three biological replicates. Error bars represent the standard error of the mean unless otherwise indicated. Statistical differences were determined using T-tests in GraphPad Prism.

Whole genome sequencing and analysis

Genomic DNA was prepared as previously described [112]. DNA was submitted to AgriLife Genomics at Texas A&M University for library preparation and Illumina Novaseq paired-end sequencing. Alignment of the fastq paired-end reads to the NCBI BW25113 reference genome (NZ_CP009273.1) was performed using Bowtie2 and BWA aligners[118,119]. Variant calling was performed using BCFtools and FreeBayes[120]. Novel insertion elements were detected using ISmapper[121,122]. All sequence variants were confirmed with Sanger sequencing.

Results

Deletion of kch results in growth defects on non-fermentable carbon sources

In a previous co-evolution analysis, Kch was predicted to interact with proteins involved in redox reactions, cell size regulation, and metabolism. To begin validating the results of this analysis, we performed targeted experiments to evaluate the involvement of *kch* in these cellular processes. Because of the predicted connection to metabolism, we first asked if loss of the channel resulted in any growth defects in various media. We found that when grown in Tryptone Broth (TB), the Δ*kch* strain (SDB2) exhibited a decreased growth rate when the culture reached an OD_{600} = ~0.5 (Fig 15A). If grown in the presence of a fermentable carbon source, glucose, no decrease in growth rate was observed (Fig 15 B). However, if the cells were grown in the presence of a non-fermentable carbon source, succinate, a decreased growth rate was again observed when the cultures reached an OD_{600} = ~0.5 (Fig 15C), suggesting that the loss of the channel affected aerobic respiration.

To test whether loss of the channel was responsible for the observed defects, we created a genotypic rescue using P1*vir* transduction, replacing the deletion cassette with a wild-type copy of *kch* under its native promoter and regulation. However, we found that the Δ*kch::kch*$^{+}$ (SDB222) strain retained the growth defects seen in SDB2, suggesting that there were additional mutations present in the background. We performed whole genome sequencing (WGS) to identify what other mutations were present in the background.

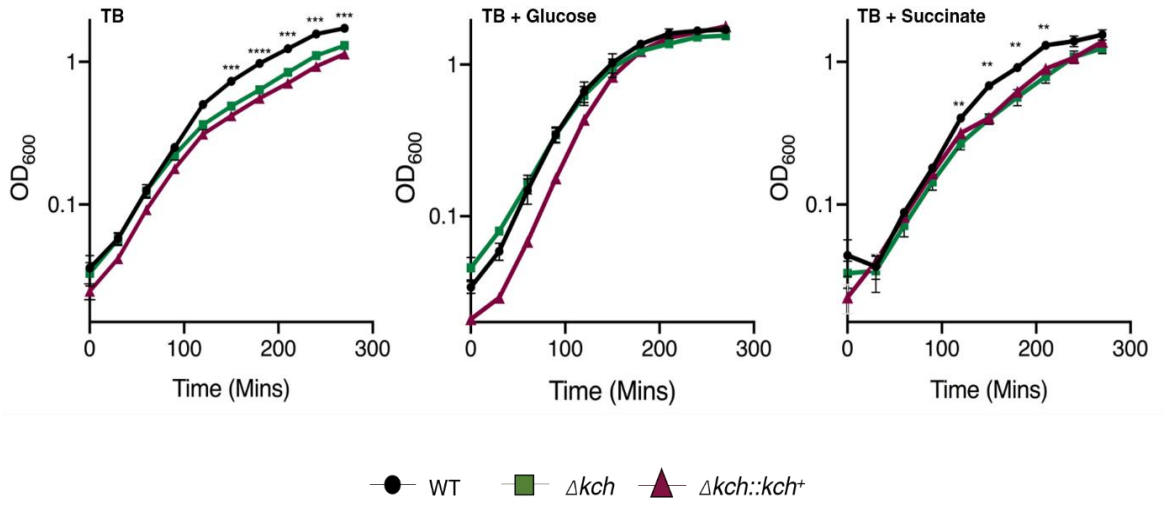

Figure 15. BW25113 Δ*kch* (SDB2) shows a slight growth defect
Strains SDB1, SDB2, and SDB222 were grown as described in methods in (A) Tryptone broth (TB), (B) TB + 0.4% glucose, and (C) TB + 0.4% succinate. N=3-6 for each strain, error bars = SEM. Significance determined by an unpaired t test, p values **** = < 0.0001, *** = < 0.0005, ** = <0.005.

Whole genome sequencing of Δkch Keio mutant reveals non-isogenic background

In addition to the Δ*kch::kan* deletion, WGS of the SDB2 strain revealed two point mutations, one in an intergenic region (genomic pos # 1317393) and one in *ubiH*, and a novel IS5 insertion disrupting *lrhA*. LrhA is a transcription factor that regulates genes involved in Type I fimbriae production, and indirectly regulates genes involved in motility and chemotaxis, via regulation of the master motility regulator *flhDC*[123]. UbiH is an enzyme required for biosynthesis of ubiquinone, a lipid-soluble electron carrier molecule that is critical in the aerobic electron transport chain (ETC)[124]. In both eukaryotes and prokaryotes, the ETC plays a critical role in cellular energy production by generating a proton gradient that is used to drive the formation of ATP. Ubiquinone, together with its reduced form, ubiquinol, forms a redox cycle that is essential for proper functioning of ETC[125]. Given the important role that ubiquinone plays in aerobic respiration, the *ubiH* point mutation, which resulted in an amino acid change at

position 223 (*ubiH*^{V223G}), is the likely cause for the metabolic phenotypes observed. Because this metabolic SNP arose in the background of the Δ*kch* strain it could represent a suppressor mutation and understanding how the mutation affects growth may yield insight into Kch function.

The ubiH^{V223G} *mutation likely decreases intracellular ubiquinone levels*

To understand how the mutation affected ubiquinone levels in the cells, we assayed how strains carrying the *ubiH*^{V223G} mutation grew on succinate as a sole carbon source. Aerobic growth on non-fermentable carbon sources such as succinate requires ubiquinone, as succinate dehydrogenase (complex II) has a strict requirement for ubiquinone[126]. A Δ*ubiH* strain (SDB10) fails to grow on succinate as the sole carbon source, but SDB1, which carries a wild-type *ubiH,* is able to grow (Fig 16A). The SDB2 strain, which contains the *ubiH*^{V223G}, is also able to grow aerobically on succinate. Additionally, transduction of the *ubiH*^{V223G} mutation into Δ*ubiH* strain (SDB10) restores the ability of the strain to grow on succinate (SDB11), indicating that the mutation does not result in a full loss of function.

To further determine if the *ubiH*^{V223G} was a gain- or reduction-of-function mutation, we determined how strains carrying *ubiH*^{V223G} responded to challenge with exogenous reactive oxygen species (ROS) stress (Fig 16B). In *E. coli,* defects in ubiquinone levels are associated with increased sensitivity to exogenous H_2O_2[127,128]. Upon addition of exogenous H_2O_2 to exponential cultures, the parental strain experiences a delay of ~ 70 minutes, but eventually resumes growth (red solid circles). However, strains carrying the *ubiH*^{V223G} mutation (SDB2, SDB11) fail to resume growth following the H_2O_2 challenge, a phenotype that is consistent with a reduction in ubiquinone levels. Replacement of the *ubiH*^{V223G} with a wild-type *ubiH* allele

(SDB28) returns a wild-type response to exogenous H_2O_2 (red, open squares). Based on these results, we conclude that the *ubiH*^{V223G} point mutation is a reduction-of-function mutation that likely reduces the ubiquinone levels in the cell.

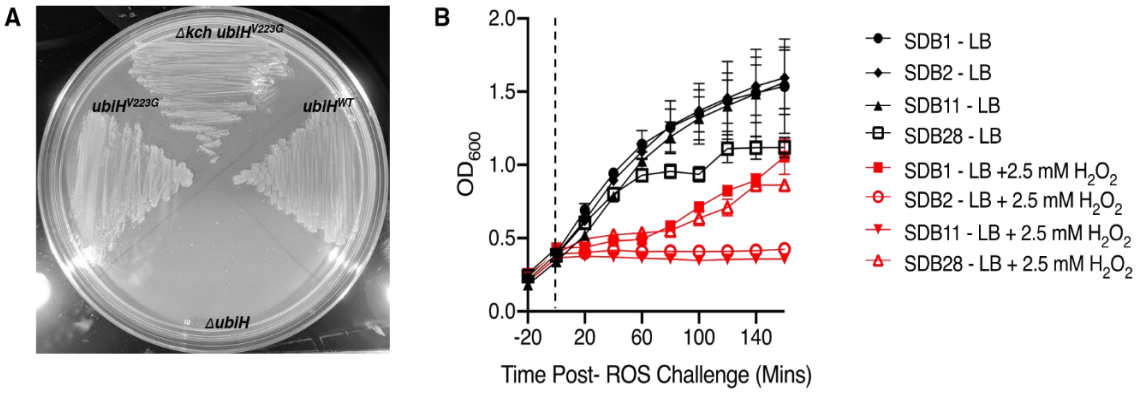

Figure 16. The *ubiH*^{V223G} mutation results in decreased intracellular ubiquinone levels (A) The ability of the WT *ubiH* (SDB1), Δ*kch ubiH*^{V223G} (SDB2), *ubiH*^{V223G} (SDB11) and the Δ*ubiH* (SDB10) to grow on succinate as the sole carbon source was determined. (B) The response of strains carrying either WT *ubiH* or *ubiH*^{V223G} to exogenous ROS stress. At t=0 (dashed line), either LB only (black lines) or LB + 2.5 mM H_2O_2 (red lines) was added. Rescue of the *ubiH*^{V223G} mutation with a WT *ubiH* (SDB28) returns a WT response to exogenous ROS stress (red open triangles). N = 3-5 for each strain in each condition. Error bars represent SEM.

The ubiH^{V223G} *mutation reduces ETC efficiency and depolarizes the cell*

Based on our result that the *ubiH*^{V223G} mutation likely reduces ubiquinone levels in the cell, we explored the effect this would have on ETC function. As ubiquinone (Q_8) functions as an electron carrier in the aerobic ETC, a reduction in Q_8 levels should slow the transfer of electrons to O_2, the terminal electron acceptor. To further test our hypothesis that the *ubiH*^{V223G} mutation leads to reduced Q_8 levels, we compared the efficiency of ETC in a wild-type strain to a strain carrying the *ubiH*^{V223G} mutation.

During aerobic respiration, O_2 is the terminal electron acceptor; therefore, O_2 consumption can be measured to assess the overall efficiency of ETC. We measured the rate of

O_2 consumption of wild-type and *ubiH*^{V223G} strains during growth on succinate as a sole carbon source and observed that O_2 consumption was reduced ~20% in a *ubiH*^{V223G} strain as compared to wild-type (Fig 17A).

Given that subsequent O_2 consumption is lowered in the *ubiH*^{V223G} strain, we hypothesized that the membrane potential would also be altered in these strains. If Q_8 levels and O_2 consumption are lowered, then we predicted that H^+ translocation across the membrane would also be reduced, resulting in a decreased $\Delta\Psi$ (more internally positive) as compared to the wild-type strain. To monitor membrane potential in these backgrounds, we utilized $DiOC_2(3)$, which is a positively-charged, fluorescent dye that acts as a Nernstian membrane potential indicator[127]. The more negatively charged (hyperpolarized) the inside of the cell is, the more dye will be taken up and sequestered. At dilute concentrations, $DiOC_2(3)$, has a green fluorescence but at high intracellular concentrations the fluorescence emission shifts to red; thus, an increase in red fluorescence is indicative of hyperpolarization (inside of the cell becoming more negative)[129–131]. Using this approach, membrane potential measurements of wild-type cells and *ubiH*^{V223G} strains during early exponential phase revealed that indeed the

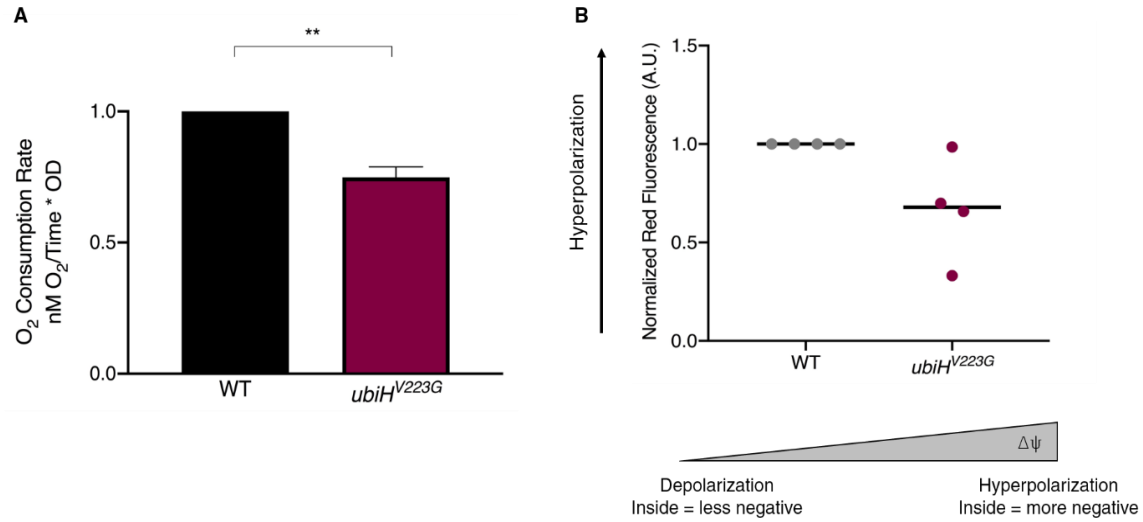

Figure 17. *ubiH*^{V223G} reduces ETC efficiency
(A) O_2 consumption by WT and *ubiH*^{V223G}strains. (B) ΔΨ measurements of WT and *ubiH*^{V223G}strains using $DiOC_2(3)$. For panel A, n = 3-4, and for panel B, n = 4. Error bars represent SEM. Significance determined by an unpaired t test, p values ** = <0.005. Data in panel B was collected and analyzed by M.A. Hudson.

ubiH^{V223G} strains were depolarized as compared to the wild-type (Fig 17B). Together this data suggests that the *ubiH*^{V223G} mutation acts to reduce the levels of ubiquinone in the cell, which subsequently slows electron flux through ETC reducing the overall efficiency and depolarizing the cell.

Model: Kch is critical for ΔΨ modulation during rapid growth

Based on our WGS data and characterization of the *ubiH*^{V223G} mutation, we propose the following model for the *in vivo* function of Kch. During rapid growth, electron flux through ETC is high, and the accompanying efflux of H^+ from the cell results in a large charge difference across the membrane with the inside of the cell being more negative (hyperpolarized). We hypothesize that Kch functions to rapidly modulate large variations of the membrane potential,

serving as a buffer against extreme hyperpolarization that would occur when strong gradients are created during rapid growth. In the absence of the channel and its buffering capabilities, one way for the cell to survive an inability to regulate extreme membrane potential changes would be to alter the strength of the gradients themselves. Mutations, such as *ubiH*^{V223G}, that reduce ETC efficiency would result in slowed electron movement and subsequent H^+ pumping, reducing the likelihood of encountering extreme changes.

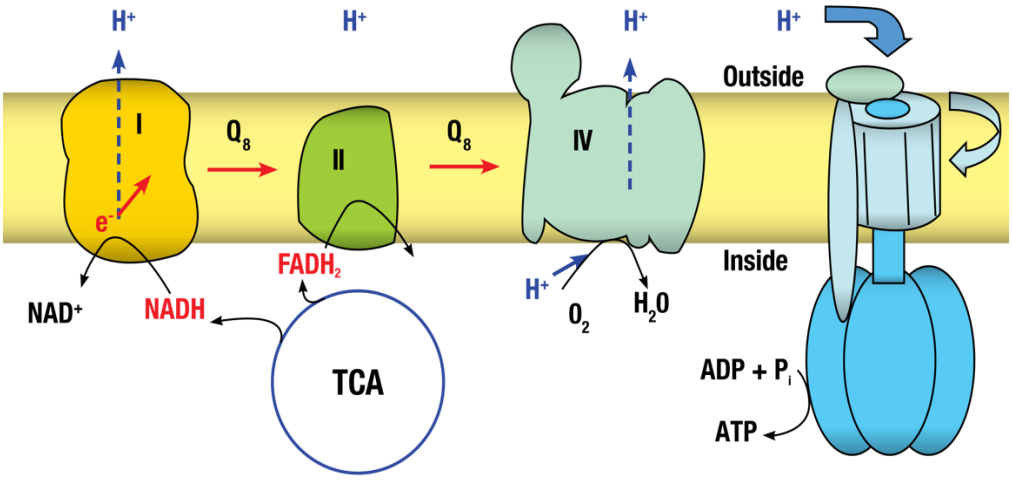

Figure 18. The electron transport chain couples electron flow to H^+ pumping
NADH and $FADH_2$ donate electrons that are shuttled between respiratory complexes by ubiquinone (Q_8). The flow of electrons through ETC is associated with H^+ translocation across the membrane. Movement of H^+ across the membrane contributes to the PMF, which is used to drive a number of biological processes including ATP synthesis. During rapid growth, the electron flux through ETC is high and generates strong gradients. One response to the loss of the channel is to reduce the efficiency of ETC (*ubiH*^{V223G}), preventing the cell from encountering strong gradients that it cannot recover from**.**

Depletion of Kch in rich media at 37 °C results in growth delay

As a test of our model, we wanted to demonstrate that Kch directly alters membrane potential *in vivo*. CRISPR interference (CRISPRi), a recently developed technology, has been used successfully to identify and to characterize the role of essential genes in a number of bacterial species[129]. This technology relies on a single guide RNA (sgRNA) and a catalytically

dead Cas9 (dCas9) to target a gene of interest for transcriptional silencing. Targeting a gene using CRISPRi causes rapid, selective transcriptional inhibition of that gene, resulting in a depletion of the gene product as a function of cell division[130,132]. Because this method results in a reduction of the targeted gene product below optimal levels and not a full loss of the gene product, CRISPRi has been used to probe how the cells respond to suboptimal levels of a variety of essential genes, including those involved in central metabolism and cell wall synthesis[69]. We utilized this approach to determine how *E. coli* responded to a targeted knockdown of *kch* under various growth conditions.

We designed a sgRNA to target *kch*, and first tested how knockdown of the channel altered the growth of the SDB1 strain in a variety of conditions. Our model would predict that depletion of Kch in rich media at 37 °C would be detrimental to growth, as we expect that the channel is critical to modulating $\Delta\Psi$ fluctuations under these conditions. However, it would also predict that lowering the growth rate either by reducing the temperature or by growth in a minimal media would produce less extreme $\Delta\Psi$ fluctuations, making Kch function less critical here. Therefore, Kch depletion is expected to be less detrimental under these conditions.

In rich media at 37 °C with an induced CRISPRi system, we found that depletion of *kch* resulted in a severe growth defect that was not seen in control cells carrying the dCas9 only (Fig 19, D). The basal expression alone of dCas9 when coupled with the *kch* sgRNA was also resulted in an ~ 1 hour delay in reaching exponential growth. To further confirm that the growth defects seen at 37 °C in LB were a result of the loss of channel function, we designed a sgRNA to target *araC*, a gene known to be non-essential for growth in these conditions[59,63] . Knockdown of *araC* in LB at 37 °C did not result in a severe growth defect (Fig 19, C). Consistent with our model, knockdown of the channel under conditions that resulted in a slower growth rate (minimal

glucose media at 37 °C or LB at 30 °C) showed no difference in growth from the dCas9-only or from the *araC* sgRNA control strains (Figs 20 and 21).

The CRISPRi results suggest that Kch is required for growth of *E. coli* under optimal conditions that allow for rapid growth (rich media at 37 °C), but that it is dispensable for growth in minimal media at 37 °C or rich media at lower temperatures. This result supports our model of Kch functioning to rapidly respond to $\Delta\Psi$ fluctuations and is also consistent with the transposon mutagenesis screen that identified Kch as important for growth at 37 °C in rich media.

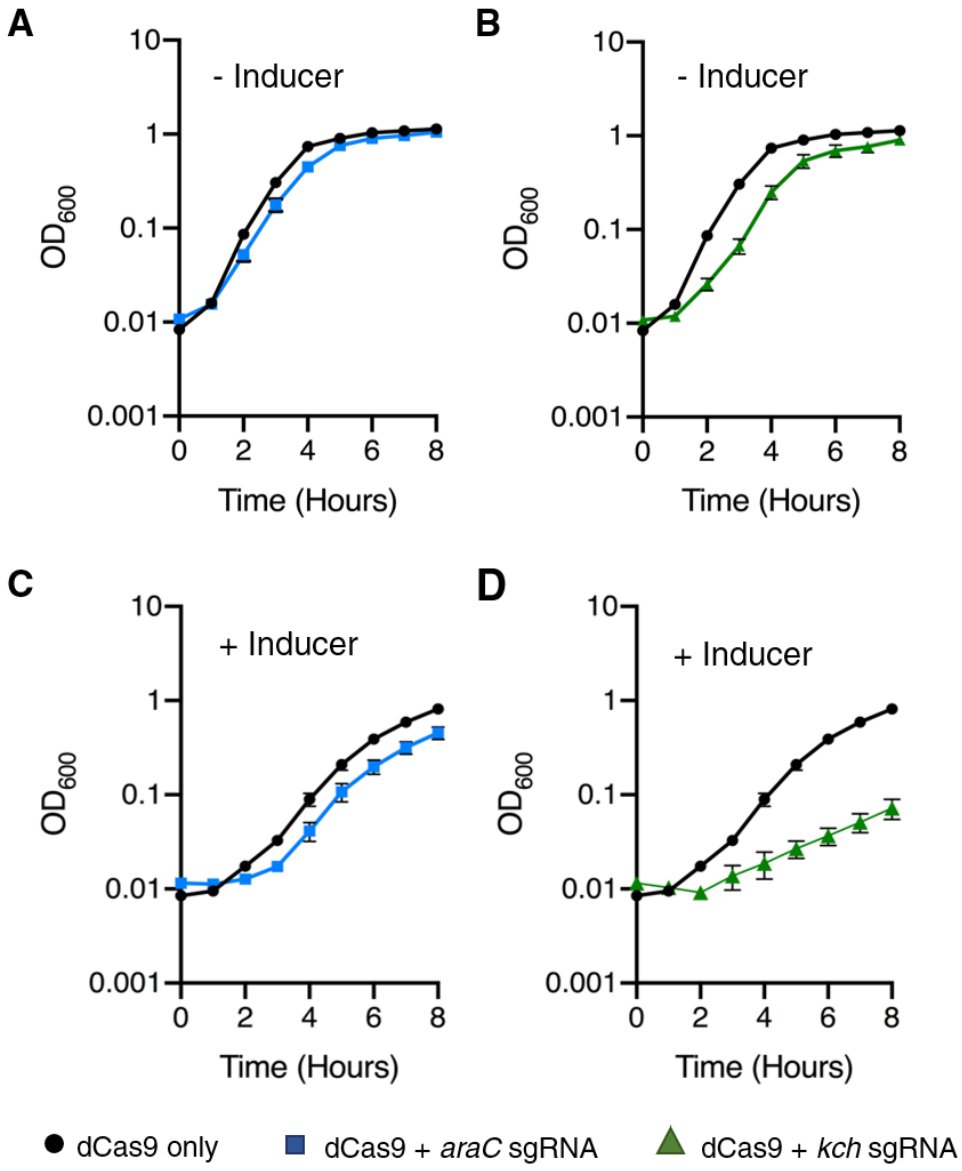

Figure 19. CRISPRi-mediated depletion of Kch results in growth defect in rich media at 37 °C

Growth during CRISPRi-mediated depletion of either *kch* (green triangles) or *araC* (blue squares) in the absence (top) or presence (bottom) of inducer. Growth curves were performed as described in methods. Strains were grown at 37 °C in LB with appropriate antibiotics and 2 µM aTc inducer as indicated. N=7-8 for all strains, error bars = SEM.

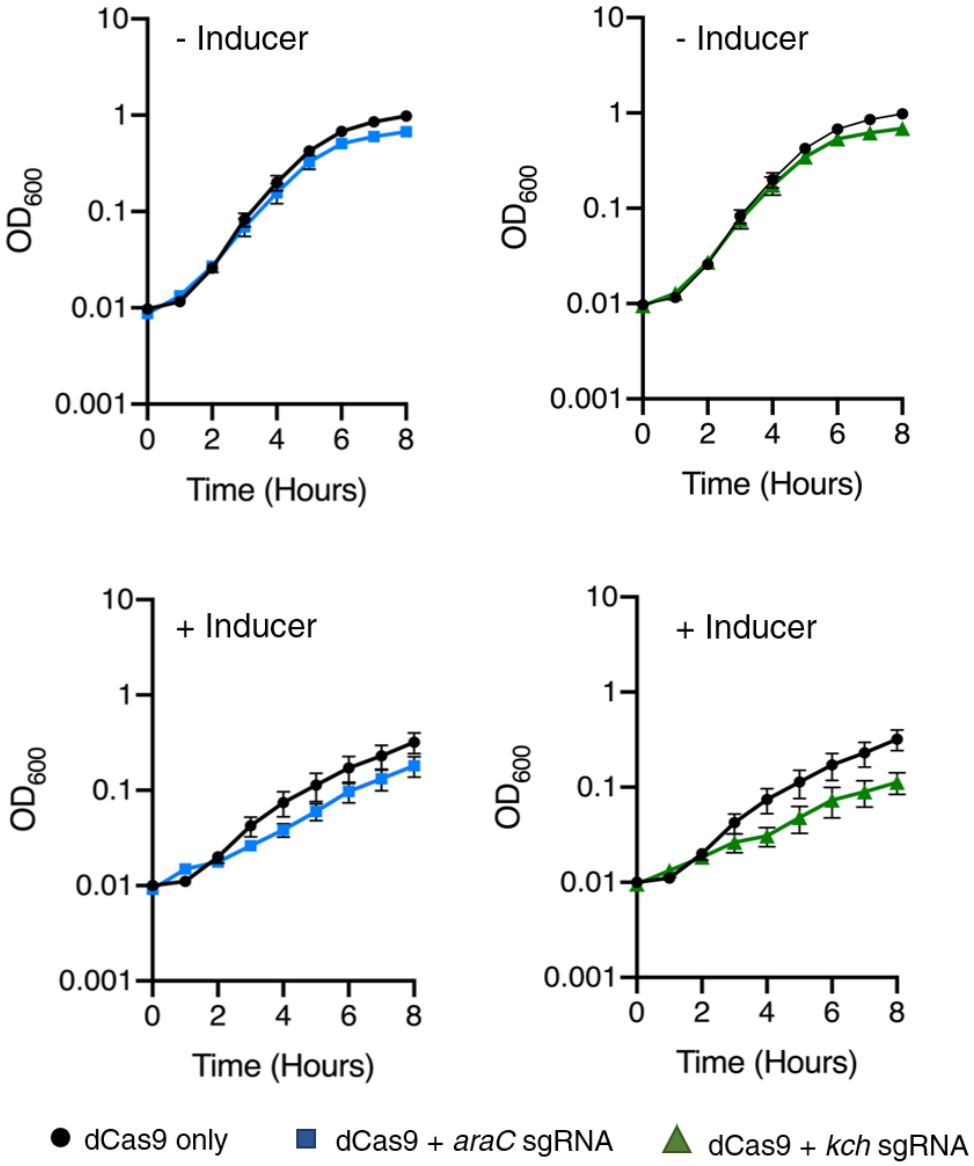

Figure 20. CRISPRi-mediated depletion of Kch in LB at 30 °C does not result in growth defect

Growth during CRISPRi-mediated depletion of either *kch* (green triangles) or *araC* (blue squares) in the absence (top) or presence (bottom) of inducer. Growth curves were performed as described in methods. Strains were grown at 30 °C in LB with appropriate antibiotics and 2 µM aTc inducer as indicated. N=4 for all strains, error bars = SEM.

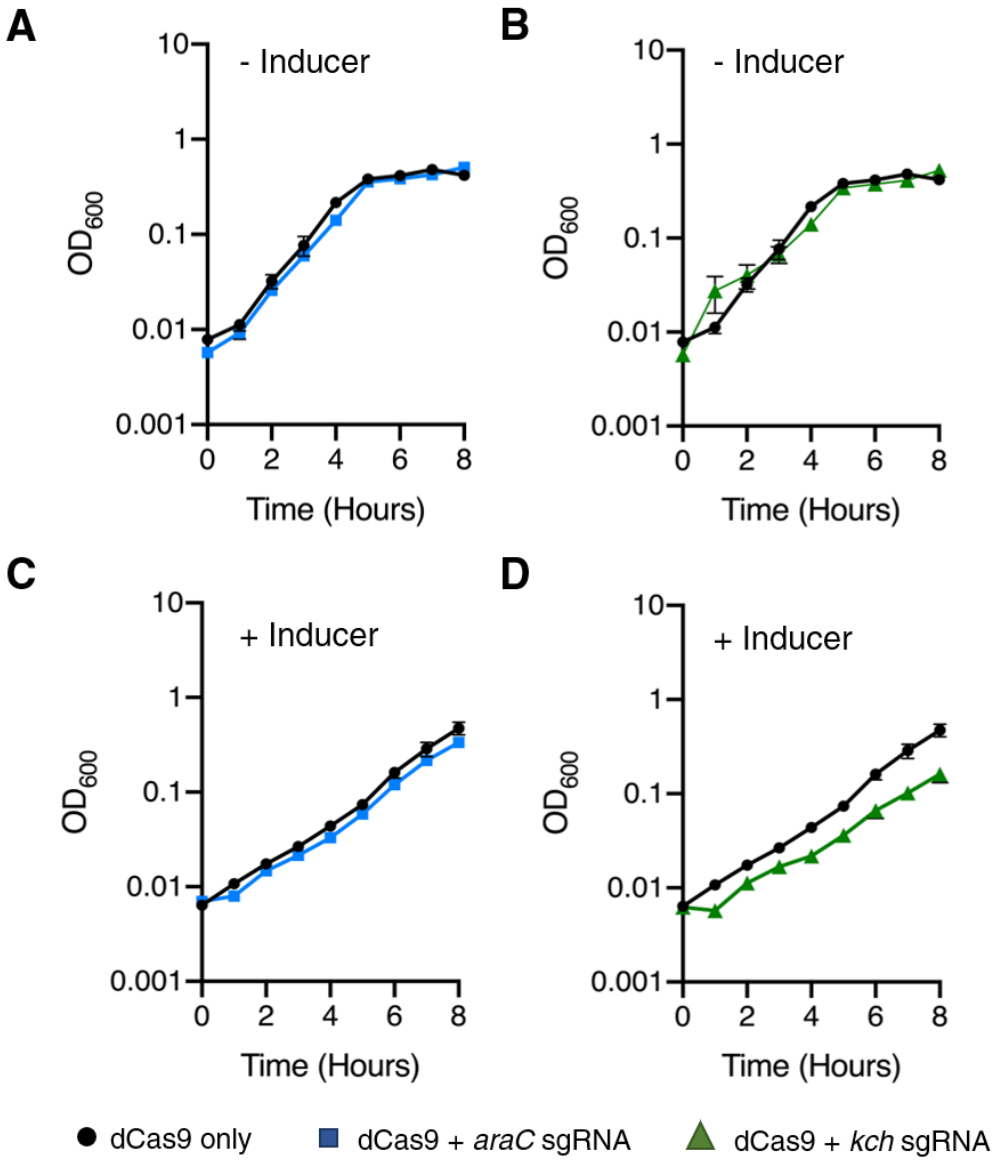

Figure 21. CRISPRi-mediated depletion of Kch in M9 glucose media at 37 °C does not impact growth

Growth during CRISPRi-mediated depletion of either *kch* (green triangles) or *araC* (blue squares) in the absence (top) or presence (bottom) of inducer. Growth curves were performed as described in methods. Strains were grown at 37 °C in M9 glucose media with appropriate antibiotics and 2 µM aTc inducer as indicated. N=4 for all strains, error bars = SEM.

Kch modulates membrane potential in vivo

We next asked if we could detect changes to ΔΨ while Kch is being depleted. In rich media, we would expect to see hyperpolarization events upon loss of the channel. To measure the ΔΨ in real time, we utilized a different fluorescent dye, Thioflavin T (ThT). ThT is a fluorescent, positively-charged dye that will accumulate intracellularly in response to membrane potential, so an increase in intracellular ThT signal correlates to a hyperpolarization event [133]. Unlike $DiOC_2(3)$, ThT does not require EDTA pre-treatment to be taken up by *E. coli* (Fig 22). Cells were only fluorescent in the presence of ThT, and they were efficiently labeled without EDTA treatment (Fig 22, D).

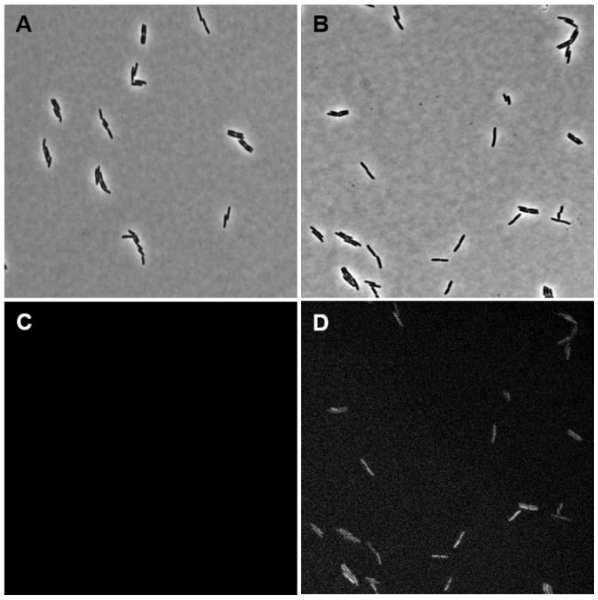

Figure 22. ThT fluorescently labels *E. coli*
SDB1 was grown in rich defined glucose media at 37 °C, and 1.5 µL of culture at an OD_{600} = 0.2 were spotted onto agarose pads with no ThT (A & C) or with 10 uM ThT (B & D).

The CRISPRi strains were grown in colorless, rich defined media in a microfluidics device that allowed imaging of individual cells and monitoring of ΔΨ over time (Fig 23). Within 30 minutes

of the addition of inducer to activate the CRISPRi-mediated Kch depletion, an increase in the intracellular ThT signal was seen (Fig 23, bottom panels). This increase in ThT signal persisted over 240 mins of exposure to the ThT. The response was specific to the *kch* sgRNA**,** as an increase in intracellular ThT signal was not seen in a strain containing dCas9 only or dCas9 and an *araC*-targeting sgRNA (Fig 24). The significant increase in ThT seen during Kch depletion suggests that the native function of Kch is to prevent severe hyperpolarization events during rapid growth.

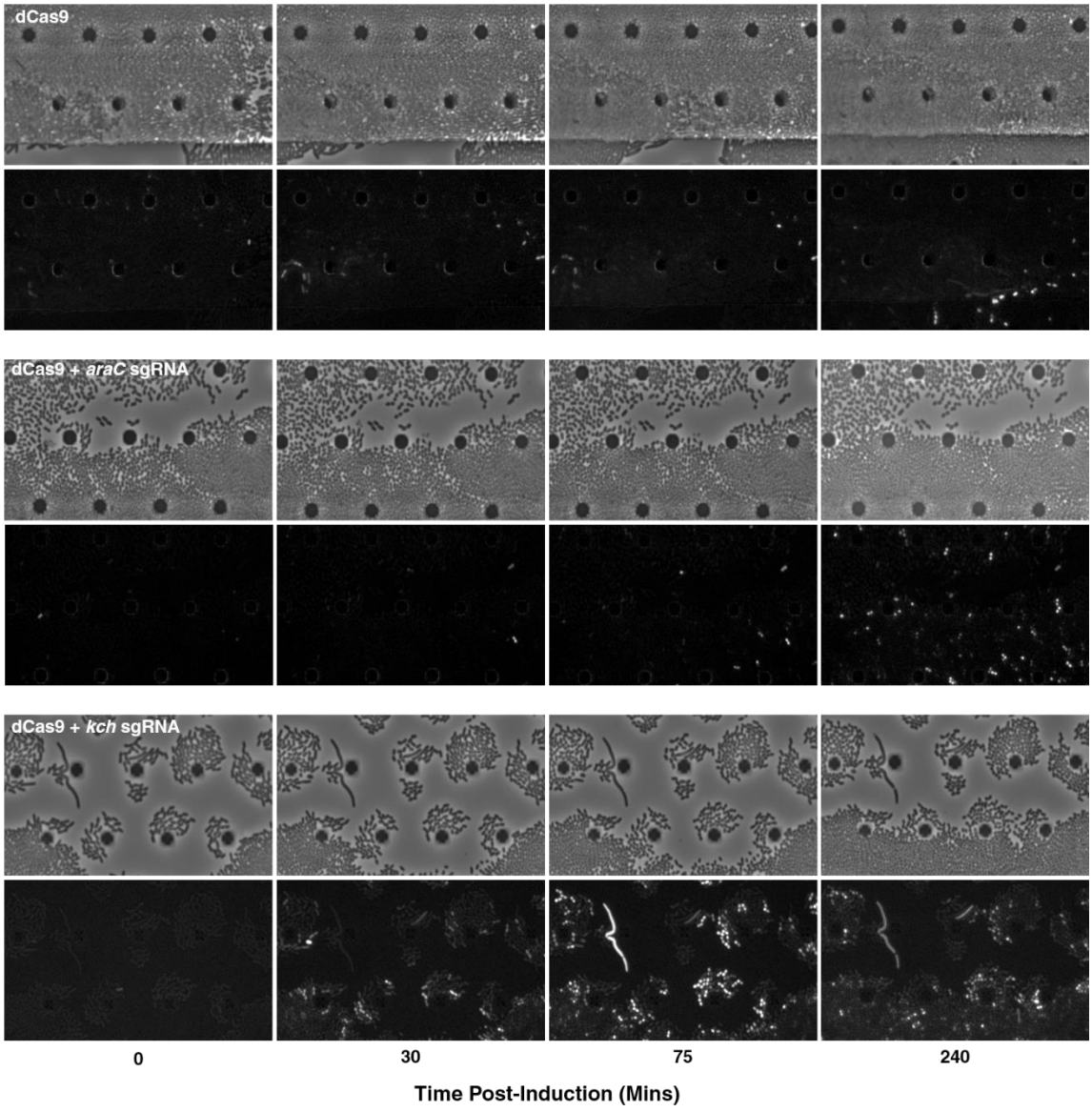

Figure 23. Depletion of Kch results in hyperpolarization events
SDB1 carrying dCas9 only or dCas9 (top panels) and sgRNAs targeting *araC* (middle panels) or *kch* (bottom panels) were loaded into a microfluidics chamber and imaged before and after inducer was added into the chambers. Base media was EZ rich, defined glucose + appropriate antibiotics + 10 μM ThT. Inducer (aTc) was added to a final concentration of 2 μM.

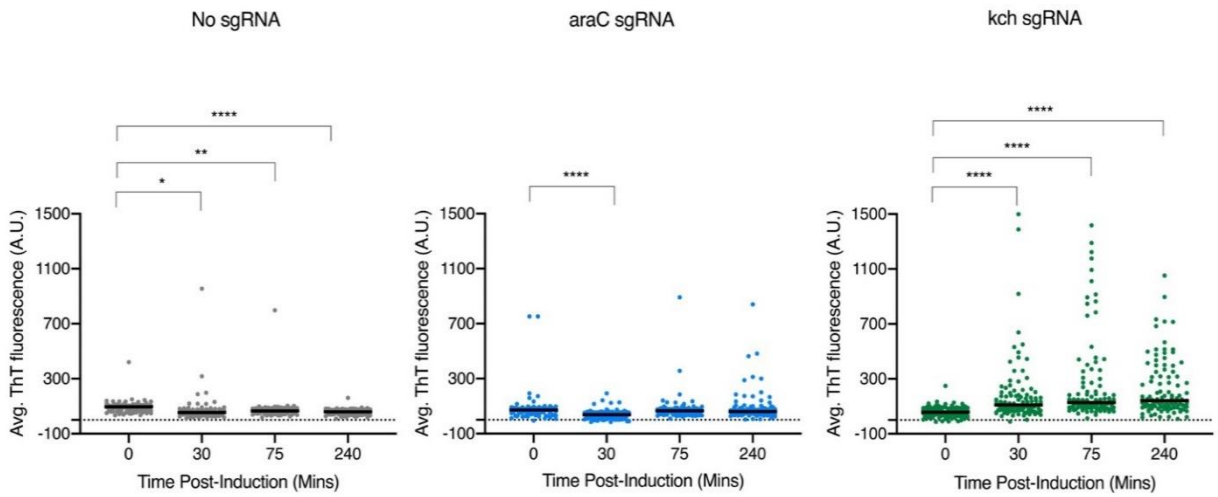

Figure 24. Intracellular ThT concentration increases in Kch-depleted cells
The average intracellular ThT concentration in single cells from Fig 23 was quantified before and after induction of CRISPRi. Data from 103 cells were analyzed for each strain at each time point. Significance determined by an unpaired t test, p values **** = < 0.0001, *** = < 0.0005, ** = <0.005.

Determining if Kch is essential

The unexpected finding of the *ubiH*^{V223G} mutation in the background of the Δ*kch* strain from the Keio collection (SDB2) led us to ask if this mutation or other mutations that reduced the efficiency of the ETC were always associated with loss of the channel? Using P1*vir* transduction, we generated seven independent *kch* null isolates by transducing the Δ*kch::kan* cassette from SDB2 into SDB1. Transductants were selected for growth on kanamycin at 37 °C and deletions were confirmed by PCR. Confirmed independent Δ*kch::kan* isolates were submitted along with our parental strain for WGS analysis. The WGS revealed that the only other mutation identified in the seven independent Δ*kch::kan* isolates was a mutation in a gene called *nemA* (*nemA*^{Y359N}). *NemA* is a N-ethylmaleimide (NEM) reductase that is also involved in the maintenance of intracellular redox by reduction of a wide variety of electrophiles, including quinones[134,135].

However, it is unclear if the *nemA*^{Y359N} mutation represents a suppressor mutation, as this mutation was also found in the background of our parental strain. Interestingly, our parental strain had several mutations compared to the NCBI BW25113 reference genome. One of those mutations was the *nemA*^{Y359N} SNP, and the other SNP was in a gene called *wrbA* (*wrbA*^{G96V}). *WrbA* is an FMN-dependent NAD(P)H:oxidoreductase that is also thought to be involved in the maintenance of intracellular redox by quinone detoxification[136]. The genomic locations of *kch* and *wrbA* are not sufficiently close to have a wild-type *wrbA* allele co-transduced in with the Δ*kch::kan* cassette, so it is unclear why only the *nemA*^{Y359N} mutation is present in the background. The lack of additional mutations in the independent Δ*kch*::*kan* transductants was unexpected, given the data that suggest that *ubiH*^{V223G} is a suppressor mutation in the Δ*kch* background. There are several possible explanations for that observation.

First, the *nemA*^{Y359N} mutation alone is enough to compensate for the loss of the channel. Neither the *wrbA* nor the *nemA* point mutations are present in the Keio Δ*kch::kan* strain, but in the absence of those mutations the *ubiH*^{V223G} mutation was acquired. It is possible that the presence of mutations in these redox genes creates an environment where additional suppressor mutations are not required to allow for viability in the absence of the channel. The second possibility is that the *ubiH*^{V223G} mutation does not represent a suppressor of Δ*kch* lethality, and instead represents a mutation that arose through strain handling. The modulation of ΔΨ by Kch is important to the cell, but its loss can be compensated for through mechanisms other than suppressor mutations. The third possibility is that during the genome analysis a structural variant (*i.e.*, duplication, inversion, or translocation) within these independent knockouts was missed. Further work will be needed to determine which of these possible scenarios is likely correct.

Discussion

Despite many attempts by several different laboratories at characterizing the *E. coli* K^+ channel, Kch, it has remained difficult to study and its function poorly understood. In Chapter 2 we applied a co-evolutionary analysis and found that Kch was predicted to interact with proteins involved in redox maintenance, regulation of cell size, and several metabolic processes. That led us to perform growth experiments in various media, where we identified a growth defect for the Δ*kch* strain when grown on non-fermentable carbon sources. This defect was unable to be rescued by a functional channel indicating additional mutations were present in the background, and WGS revealed the presence of several unexpected mutations in the background.

Results from this study highlight the importance of WGS of strains. Single gene knockout collection strains are presumed to be isogenic, lacking only the gene of interest, but our sequencing result demonstrates that this is not the case. While we report the presence of additional mutations in a single strain from a large collection, similar results from other strain collections suggest that this is a wide-spread complication[137–139]. Our sequencing data reiterates the extreme caution that should be used in the interpretation of data from high-throughput phenotypic screens done on the single-gene knockout strain collections. While working with single-gene knockout strains is ideal to parameterize experiments, it is apparent that deletion of single genes, even those assumed non-essential, can have large impacts on the genome that may confound experimental interpretation.

One of the mutations identified in the background of the Δ*kch* strain (SDB2) was the $ubiH^{V223G}$ point mutation, which we demonstrated acts to reduce the overall efficiency of aerobic ETC and to reduce the membrane potential. Based on the effects of the $ubiH^{V223G}$ mutation, we hypothesized that Kch was involved in modulation of $\Delta\Psi$ to prevent severe hyperpolarization.

CRISPRi-mediated depletion of Kch leads to a decreased growth rate and cell filamentation. In conditions where growth is slower, either due to lower temperature or minimal media, the function of *kch* appears to be more dispensable and there is no discernable growth delay from channel depletion. Using microfluidics devices and fluorescent $\Delta\Psi$ dyes, hyperpolarization of the cell was reported after induction of CRISPRi-mediated *kch* depletion, supporting the hypothesis that the native function of Kch is to prevent extreme hyperpolarization during rapid growth.

Together our data suggest that Kch is important for adaptation to conditions that promote rapid growth. During rapid growth, electron flux through ETC is high and the accompanying H^+ translocation generates strong gradients, where the inside of the cell would be highly negative. In the absence of the channel, the reduction in ETC efficiency serves to prevent extreme membrane potential fluctuations. This functional coupling of the K^+ channel to cellular respiration allows the cell to respond to channel disruptions by reducing ETC efficiency, thereby preventing the cell from ever encountering situations of extreme hyperpolarization of the membrane that it cannot recover from.

The functional role of bacterial K^+ channels has been poorly understood. Recent work in Bacillus subtilis revealed that the YugO K^+ channel is involved in both communicating metabolic stress with the biofilm as well as attracting motile cells to the biofilm via modulation of the target cell's membrane potential[140]. Here we also demonstrate that the K^+ channel in *E. coli* is important for growth in rich media at 37 °C.

Given the high conservation of K^+ channels in bacterial genomes, it is clear that K^+ channels play an important role in bacterial physiology. It is possible that one functional role for K^+ channels is rapid modulation of membrane potential, allowing for adaptation to different

growth environments. Our results also suggest bacterial K^+ channels could be exploited as effective drug targets. Inhibition of bacterial K^+ channels in pathogens could lead to impaired energy generation, making the cells less capable of competing with the host's natural microbiome or the host's immune system.

CHAPTER IV

IDENTIFICATION OF A PERIODIC BANDING PATTERN FORMED BY *ESCHERICHIA COLI* MUTANTS

Introduction

Despite being unicellular organisms, prokaryotes engage in complex social behaviors that resemble those of their eukaryotic counterparts. Well-characterized cooperative microbial behaviors include quorum sensing, biofilm formation, and swarming motility, but perhaps the most visually striking example of cooperative behavior is the vast array of complex growth patterns that have been described in bacteria. Examples include branching dendritic patterns formed by gliding *Bacillus subtilis* cells[115], cellular aggregation in soft agar in response to a chemoattractant in *Escherichia coli* and *Salmonella typhimurium*[114,141], and the striking, concentric "bulls-eye" pattern created by swarming *Proteus mirabilis* cells[113].

Pattern formation in single-celled organisms requires coordinated gene expression and a variety of intercellular interactions; it represents a distinct transition from autonomous growth and function to multicellular-like behaviors. The advantages of such a transition include access to nutrients, protection from predation, and increased resistance to antibiotics and other stresses[142,143]. The growing number of bacteria for which a complex growth stage or pattern formation has been observed suggests that cooperative growth is more prevalent in the microbial world than previously thought. Here we identify and characterize a striking banding pattern formed by *Escherichia coli* mutants during swimming motility.

Materials and Methods

Strains, media, and growth conditions

The single deletion Keio strains were obtained from the Yale Coli Genetics Stock Center. To generate double mutant strains, the kanamycin cassette was excised from the BW25113 Δ*kch::kan* strain (SDB2), resulting in BW25113 Δ*kch::frt* (SDB6). P1*vir* transduction was used to transduce the Δ*tsr,* Δ*lsrB,* Δ*luxS*, and Δ*qseC* mutations individually into MG1655 and then into the BW25113 Δ*kch::frt* (SDB6) strain. Double mutants were selected for on LB + kanamycin at 37 °C, and the genotypes were confirmed with PCR and by the loss of the ability to grow on arabinose as a sole carbon source. Replacement of the Δ*kch::frt* with a functional kch gene was done by first transducing in the closely linked Δ*trpC::kan* auxotrophic marker from SDB9 into SDB6, and then transducing in a wild-type *trpC* gene. The final strain was screened for its ability to grow on minimal media lacking tryptophan, and all genotypes were confirmed with PCR and Sanger sequencing.

Table 2. Strains used in Chapter IV

Strain Number	Description	Relevant Genotype	Source or Reference
Strains			
SDB1	BW25113	*F⁻ DE(araD-araB)567 lacZ4787(del)::rrnB-3 LAM⁻ rph-1 DE(rhaD-rhaB)568 hsdR514 wrbA^{G96V} NemA^{Y359N}*	Paul Straight
SDB2	BW25113 Δ*kch::kan*	Δ*kch::kan*	CGSC #9121
SDB6	BW25113 Δ*kch::frt*	Δ*kch:frt ubiH^{V223G} lrhA198::IS5*	This study
SDB7	BW25113 Δ*tsr::kan*	Δ*tsr::kan*	CGSC #11812
SDB9	BW25113 Δ*trpC::kan*	Δ*trpC::kan*	CGSC #9131
SDB12	MG1655 Δ*qseC::kan*	Δ*qseC::kan*	This study
SDB13	MG1655 Δ*tsr::kan*	Δ*tsr::kan*	This study
SDB14	MG1655 Δ*luxS::kan*	Δ*luxS::kan*	This study
SDB18	BW25113 Δ*kch* Δ*luxS::kan*	Δ*kch:frt* Δ*luxS::Kan ubiH^{V223G} lrhA198::IS5*	This study
SDB20	BW25113 Δ*kch* Δ*qseC::kan*	Δ*kch:frt* Δ*qseC::Kan ubiH^{V223G} lrhA198::IS5*	This study
SDB22	BW25113 Δ*kch* Δ*tsr::kan*	Δ*kch:frt* Δ*tsr::Kan ubiH^{V223G} lrhA198::IS5*	This study
SDB222	BW25113 Δ*kch::kch⁺*	*kch⁺ ubiH^{V223G} lrhA198::IS5*	This study
SDB223	BW25113 Δ*kch::kch⁺*	*kch⁺ ubiH^{V223G} lrhA198::IS5*	This study
SDB267	BW25113 Δ*qseC::kan*	Δ*qseC::kan*	CGSC #10295
SDB268	BW25113 Δ*luxS::kan*	Δ*luxS::kan*	CGSC #10100

Routine culturing of strains was done in Luria-Bertania (LB) broth – 10 g/L Tryptone, 10 g/L NaCl, and 5 g/L Yeast Extract. Kanamycin was used at a 50 ug/mL concentration as needed. Tryptone Broth motility media was made with 10 g/L Tryptone, 8 g/L NaCl, and 0.25% w/v agar. Standard Motility Media was made with 10 g/L Tryptone, 5 g/L Yeast extract, and 0.25% w/v agar with either 0.17M NaCl (SMM-Na) or 0.17M KCl (SMM-K). Media was dispensed in 25 mL volumes into sterile petri dishes. Sterile toothpicks were used to inoculate a single colony into the center of the plate, and plates were incubated at 37 °C for 16 hours unless otherwise indicated. Amino acids were obtained from Sigma-Aldrich and were supplemented to SMM-K to a final concentration of 10 mM.

Whole genome sequencing

Genomic DNA was prepared as previously described [144,145]. DNA was submitted to AgriLife Genomics at Texas A&M University for Illumina paired-end sequencing. Alignment of the fastq paired-end reads to the NCBI BW25113 reference genome was performed using Bowtie2 and BWA aligners[118,146]. Variant calling was performed using BCFtools and Freebayes[120]. Novel insertion elements were detected using ISmapper[143]. All sequence variants were confirmed with Sanger sequencing.

Results

K^+ channel mutant exhibits both motility and pattern formation in LB at 37 °C

E. coli encodes a single K^+-selective ion channel called *kch*. While exploring the effect of Δ*kch* on swimming motility, we serendipitously discovered that a Δ*kch* strain exhibited a striking pattern formation when swimming through Luria-Bertani (LB) semi-soft agar at 37 °C (Fig 25, B). Under these conditions, a wild-type strain is non-motile, but the Δ*kch* strain (SDB2) is both motile and pattern-forming. Once the cells start swimming out of the inoculation site, they form concentric bands approximately every two hours. The front of cells moving out from the inoculation site is followed by an increase in cell division, which sets up an alternating pattern of low density (dark) and high cell density (bright) bands. The bands differ from chemotaxis rings in that the bands are static and do not move once formed, whereas chemotaxis rings move and expand with the cell front.

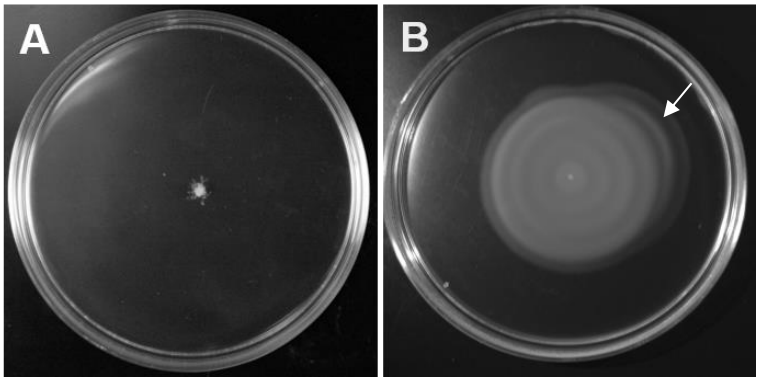

Figure 25. BW25113 Δ*kch* mutant forms periodic banding pattern
Motility agar (LB $^{+}$ 0.25% w/v agar) was inoculated with either wild-type BW25113 (A) or BW25113 Δ*kch* (B) and incubated at 37 °C for 16 hrs. The white arrow indicates offshoots that form and swim without pattern formation**.**

We asked how media composition and temperature affected the formation and appearance of the bands. While growth in LB at 37 °C are common laboratory growth conditions for *E. coli*, they are non-standard conditions for assessment of motility. Therefore, we explored the motility behaviors in Tryptone Broth (TB) and in LB at a range of temperatures from 25-37 °C. At 37 °C in TB, the Δ*kch* strain was motile, but failed to form the banding pattern (Fig 26). We tested a range of temperatures between 25-37 °C, and we found that decreases in temperature were associated with a loss of the banding pattern and a reappearance of chemotaxis rings when grown in LB (Fig 27). The conditions in which the patterning arises, which are non-standard for motility assays explains why this pattern formation in *E. coli* has not been described previously, given that *E. coli* is a model organism for the study of chemotaxis and motility.

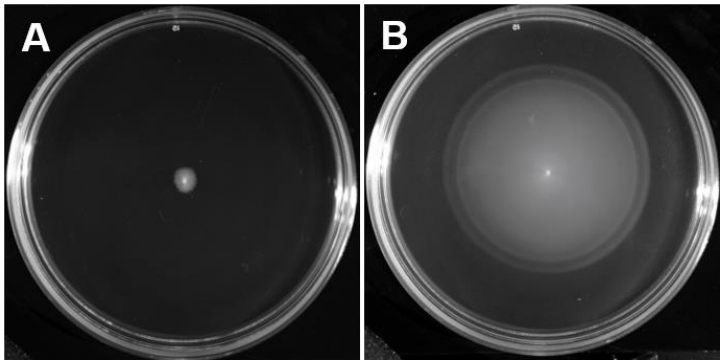

Figure 26. Periodic banding pattern is specific to LB media
TB motility agar (TB + 0.25% w/v agar) was inoculated with either wild-type BW25113 (A) or BW25113 Δ*kch* (B) and incubated at 37 °C for 16 hrs.

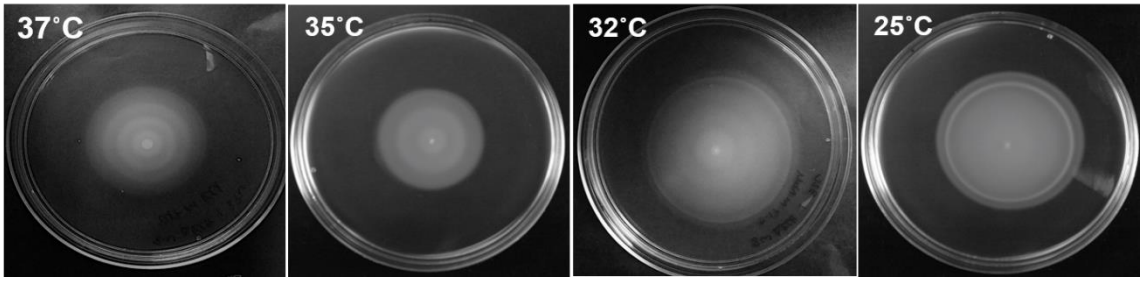

Figure 27. Lower temperatures suppress pattern formation
The effect of slower cell growth on the banding pattern was determined by lowering the incubation temperature. Strains were grown on SMM-K media for 15 hrs. (37 °C and 35 °C) or 20 hrs (32 °C and 25 °C) prior to imaging.

Addition of K^+ fails to rescue the banding pattern

To give insight into what processes might be involved in the pattern formation, we asked if we could identify conditions that suppress the formation of this banding pattern. Because this patterning was identified in a null K^+ channel mutant, we first asked if the addition of K^+ to the media could mask the bands. The salt concentration in Miller LB is ~170 mM, so we replaced the NaCl with an equimolar concentration of KCl and assayed the swimming behavior. The Δ*kch* strain still formed concentric bands in the presence of 170 mM KCl (Fig 28). Replacing NaCl in

the LB with equimolar concentrations of KCl, results in subtle changes to the patterning, with the bands being thicker in the presence of KCl. The Δ*kch* strain is motile and banding at high molar concentrations, but eventually becomes non-motile at salt concentrations above 0.5 M. The appearance of banding behaves differently in the two different salt species; it persists even in 0.5M KCl, but the cells are non-motile at the same concentration of NaCl. The addition of KCl results in more obvious banding and is used in all future experiments at a concentration of 170 mM.

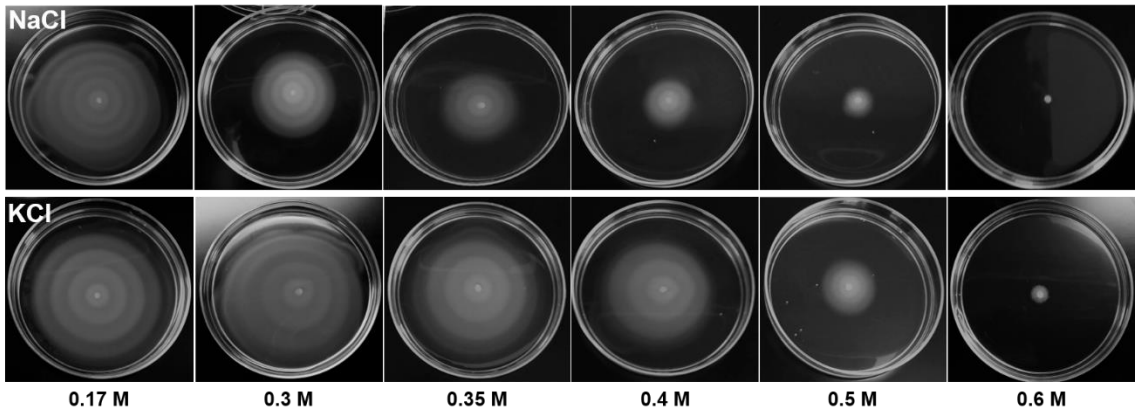

Figure 28. Supplementation of K^+ fails to suppress pattern formation
The effect of increasing NaCl or KCl on the pattern formation. Increasing concentrations of either KCl or NaCl was added to motility media (1% Tryptone, 0.5% Yeast Extract, 0.2 % w/v agar). Plates were incubated at 37 °C for 16 hours.

Hypermotility suppresses pattern formation

Cells isolated from either the high or low cell density bands reliably reproduce the same motility pattern (data not shown); however, we noticed that towards the end of the incubation period there were offshoots that would form and swim without the pattern formation (Fig 25 B, white arrow). Cells from these offshoots were isolated and re-tested to determine their swimming behavior at 37˚C in LB soft agar. Compared to cells isolated from banding regions, cells isolated

from the offshoots swam faster and without pattern formation (Fig 29, A). The offshoot isolates were renamed Δ*kch-HM* for their hyper-motile phenotype.

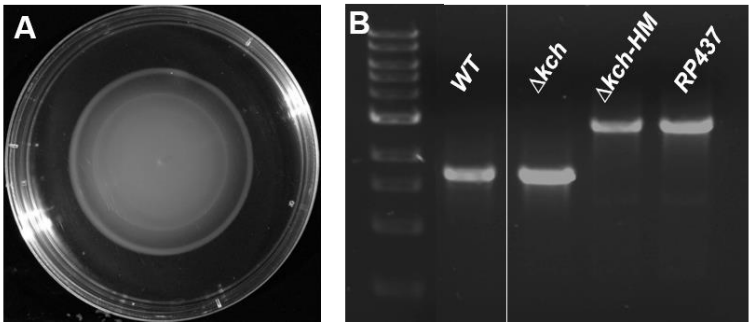

Figure 29. BW25113 Δkch hypermotile isolates swim without pattern formation

(A) The isolated hypermotile cells faster and without the pattern formation. SMM-K motility plates were incubated for ~7 hours at 37 °C. Photo taken by K. Hofstetter. (B) PCR amplification of the *flhDC* regulatory region in the WT, Δ*kch,* and Δ*kch-HM* strains. The size of the band in the Δ*kch-HM* is consistent with that of RP437 which carries an IS5 disruption in that region. Extra lanes between the WT and the Δ*kch* lanes were removed for clarity.

In *E. coli*, it has been reported that mutations that increase motility are often due to novel IS elements incorporating upstream of the master flagellar regulator, *flhDC*[143]. The disruption of negative regulatory sites leads to an overall increase in flagellar gene expression and higher flagellar numbers[147,148]. We compared the *flhDC* regulatory region of the wild-type strain, the Δ*kch* strain, and the Δ*kch-HM* isolate to that of RP437, a commonly used chemotaxis strain that is hyper-motile due to an IS5 insertion upstream of the master regulator[149]. The Δ*kch-HM* strain had a band size consistent with RP437, indicating that the hyper-motility was due to a novel IS5 insertion upstream of *flhDC* (Fig 28B). The loss of pattern formation is associated with the acquisition of a novel IS element upstream of *flhDC*, which results in a hypermotile phenotype.

Addition of serine suppresses band formation

We asked whether the addition of specific nutrients could also mask the banding pattern. To assess how the supplementation of amino acids and specific carbon sources would alter the banding, we supplemented each of the 19 essential amino acids (Tyrosine was insoluble at all concentrations tested) at a 10 mM concentration to the SMM-K motility media (Fig 30).

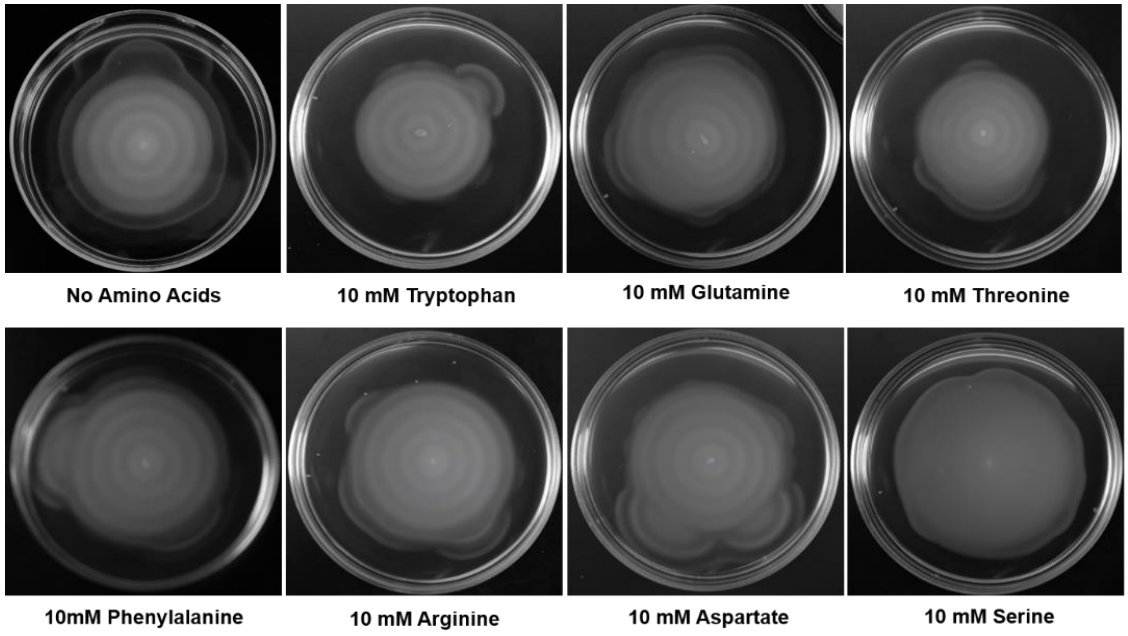

Figure 30. Addition of serine suppresses band formation
The effect of amino acid supplementation on pattern formation was tested. Only All nineteen amino acids were tested for their ability to suppress the pattern formation, but only serine was able to suppress the pattern formation. No significant alterations to the pattern were seen with any of the other tested amino acids. SMM-K plates were incubated at 37 °C for 16 hours**.**

The only amino acid that significantly altered the appearance of the banding was serine, which resulted in a confluent swimming pattern. A range of serine concentrations were tested, and significant alterations were seen at 5 mM serine concentrations (Fig 31).

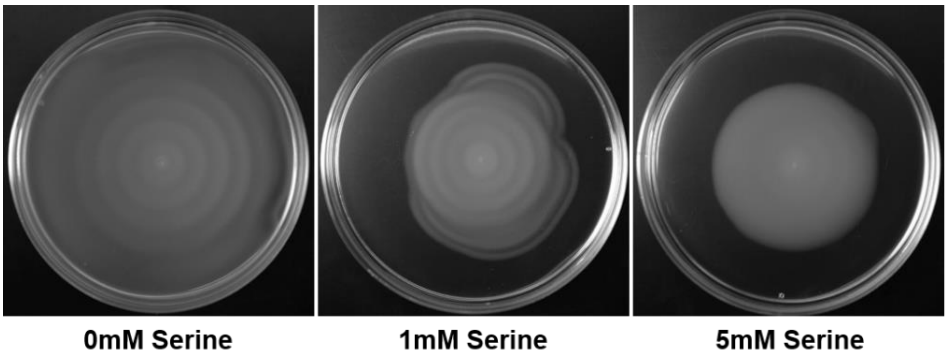

Figure 31.The effect of a range of serine concentrations on pattern formation
A range of serine concentrations was tested its effect on pattern formation. At 1 mM, serine does not have a distinct effect on the patterning, but at 5mM serine the pattern formation is significantly altered. Base media is SMM-K, and plates were incubated at 37 °C for 16 hours**.**

Pattern formation requires intact Tsr

Because serine was the only amino acid tested that was able to suppress the banding pattern, we investigated whether Tsr was required for proper band formation. Tsr is one of four methyl-accepting chemoreceptors in *E. coli*, and it mediates chemotaxis towards serine and away from a number of repellents[150]. We generated a double mutant, Δ*kch* Δ*tsr,* and assayed its ability to form bands in SMM-K at 37 °C. The double mutant remained motile and swam out from the inoculation site, but failed to generate a pattern, indicating that Tsr is required for band formation (Fig 32E). In addition to serine chemotaxis, Tsr also mediates the chemotactic response towards and the auto-aggregation behavior in response to the quorum sensing molecule, autoinducer-2. We next asked if quorum sensing played any role in pattern formation.

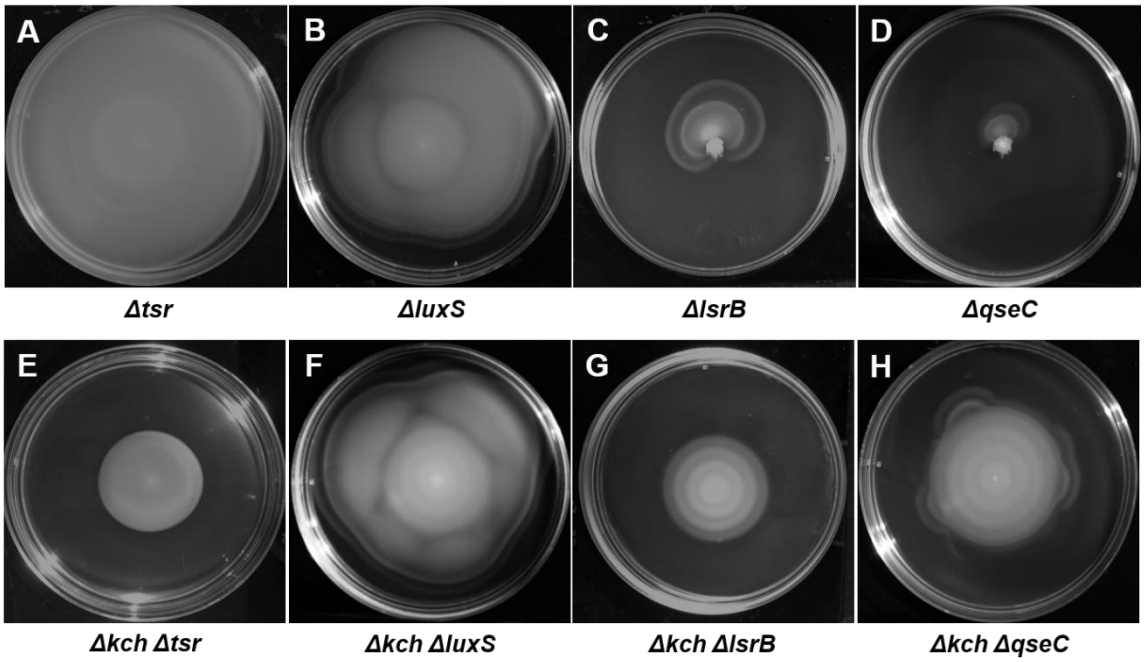

Figure 32. Tsr is required for pattern formation, but the quorum sensing systems are individually dispensable
Disruption of *tsr* results in a loss of pattern formation (E). Pattern formation is not dependent on either AI-2 production or sensing (F & G) or *qseC* (H). SMM-K plates were incubated for 12 or 16 hours at 37 °C.

Individual quorum sensing systems are dispensable for pattern formation

Organizing social behaviors, such as pattern formation, requires cell-to-cell communication and coordinated gene expression. One mechanism for intercellular communication in bacteria is quorum sensing. Quorum sensing allows bacterial populations to utilize chemical cues to both sense and to respond to increasing cell density through coordinated gene expression. Quorum-controlled biological processes are then relegated to being active only when the cell density has reached a critical threshold and the behavior is most advantageous[150].

There are two quorum-sensing systems in *E. coli*, a *luxS*-dependent system and a *qseBC*-dependent system. AI-2 is generated from S-adenosylmethionine (SAM), a methyl donor that contributes to many cellular processes and is involved in one-carbon metabolism. Conversion of

SAM yields S-adenosylhomocysteine (SAH), which is processed to S-ribosylhomocysteine (SRH) and adenine. LuxS converts SRH to homocysteine and 4,5 dihydroxy, -2,3- pentanedione (DPD), and spontaneously cyclized DPD is further processed to AI-2. As AI-2 is generated from SAM, the autoinducer can serve as both a reporter of cell density and an indicator of the metabolic state of the population[151,152].

The *luxS*-mediated biosynthetic pathway for AI-2 production is highly conserved in bacteria, and AI-2 may represent an interspecies quorum sensing molecule. In *E. coli*, AI-2 differs from traditional autoinducers in that it the peak extracellular AI-2 concentration occurs in mid-to-late exponential phase and drops sharply upon entrance into stationary phase. This sharp drop is due to the internalization and subsequent processing of AI-2 via the *lsr* operon. The reason for production, export, and re-uptake of AI-2 remains unclear[153].

The *qseBC* quorum sensing system has been best characterized in enterohemorrhagic *E. coli* (EHEC), where it has been shown to be a unique system that allows for interkingdom signaling. In EHEC, *qseBC* represents a two-component system, where *qseC* is a sensor kinase that responds to the bacterial quorum sensing molecule, autoinducer-3, as well as host produced signals, epinephrine (epi) and/or norepinephrine (NE), and *qseB* is the response regulator. Together *qseBC* have been shown to positively regulate the expression of flagellar and motility genes[153]. In *E. coli* K-12, expression of the *qseBC* operon has also been shown to be activated by AI-2, but a direct effect on the phosphorylation state of *qseC* by AI-2 remains to be demonstrated[121].

We asked if either of these two quorum sensing systems were required for the banding pattern in *E. coli*. For the AI-2 system, we tested whether the production of AI-2 or the sensing of AI-2 was important for the pattern formation. We generated double mutants that were either

unable to synthesize AI-2 (Δ*luxS* Δ*kch*) or impaired in AI-2 internalization (Δ*lsrB* Δ*kch*) and assayed how their motility behavior was altered. However, the behavior of the double mutants was unchanged from that of the Δ*kch* strain; they remained motile and capable of banding at 37˚C, indicating that neither AI-2 synthesis or detection were required for pattern formation (Fig 32 F & G).

Next, we asked if an intact *qseBC* two-component system was required for formation of the banding pattern by deleting the histidine kinase, *qseC*. A Δ*qseC* Δ*kch* double mutant also remained motile and banding at 37 °C (Fig 32, H). Together, these data show that the individual quorum sensing systems are dispensable for formation of the banding pattern. It is possible that there is some compensation occurring between the two quorum sensing systems, and a triple mutant would be required to rule out that possibility.

Pattern formation in Δkch strain is not due the channel deletion

To test whether the pattern phenotype was solely attributable to the loss of the channel, we created a genotypic rescue of the Δ*kch*. Using P1*vir* transduction, we generated a Δ*kch::kch*$^+$ strain (strains SDB222 & SDB223), replacing the deletion cassette in strain SDB2 with a wild type copy of the channel under its native promoter and regulation. However, even with a functional K^+ channel, the Kch rescue strains still formed a banded motility pattern in SMM-K media at 37˚C (Fig 33). This unexpected result led us to perform whole genome sequencing (WGS) on the Δ*kch* strain to identify what additional mutation(s) in the background of this strain are responsible for the banding formation. WGS revealed the presence of two SNPs and a novel IS5 element insertion event in the background, in addition to the Δ*kch::kan* deletion (table1&2).

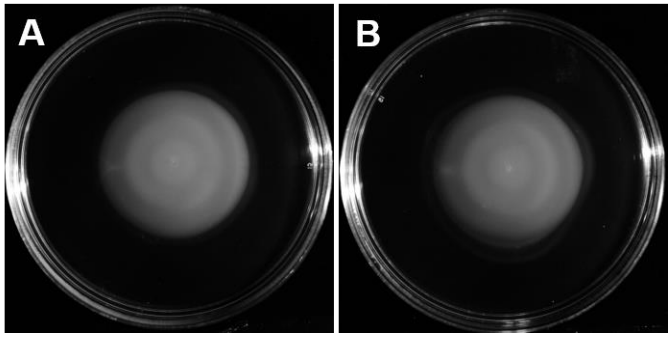

Figure 33.Genotypic rescue of the Δkch mutation with a functional kch gene fails to rescue pattern formation
A functional *kch* was transduced into the Δ*kch::kan* background, but this fails to return a wild-type motility behavior. Two independently obtained isolates, SDB222 (A) and SDB223 (B), remain both motile and pattern forming, indicating that neither phenotype is attributable to the channel deletion. SMM-K plates were inoculated in the center with a single colony and incubated at 37 °C for 14 hrs.

A likely candidate for the causative mutation for banding is the disruption of *lrhA* by an IS5 element. LrhA is a transcription factor that is known to negatively regulate motility in *E. coli* through transcriptional repression of *flhDC*[154,155]. Loss of the negative regulator, *lrhA*, results in depression of *flhDC*, allowing for the motility under non-permissive temperatures that we observe in this strain background. Further work will be required to determine if the banding is due to the *lrhA* disruption. Screening of other Keio knockout strains revealed additional strains that were also capable of pattern formation (Fig 34). It will be interesting to determine if the patterning in these backgrounds is due to the primary mutation or if any of the mutations identified in the Δ*kch* strain are present in these strains also.

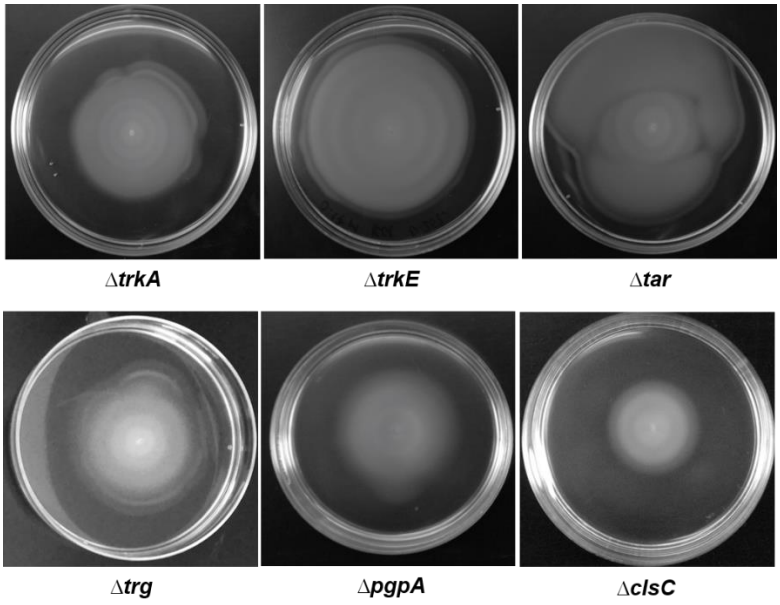

Figure 34.Additional pattern forming mutants
Several additional Keio mutants were identified that also form motility patterns in SMM-K. Plates were incubated at 37 °C for 16 hrs.

Discussion

Pattern formation is a ubiquitous phenomenon in nature. We identified a novel concentric banding pattern formed by *E. coli* mutants when swimming in LB soft agar at 37˚C. While it is unclear if bacterial pattern formation has a biological advantage or relevance, these patterns provide a clear framework for investigation into mechanisms of cell-to-cell communication, signal transduction, and physical interactions. The relative ease with which bacterial pattern formation can be manipulated and studied has made it a tractable model for understanding morphogenesis and scaling in higher organisms[154]. Additionally, bacterial patterns can be engineered as a result of synthetic gene circuits. For example, Liu *et al.* created synthetic genetic circuits that placed *E. coli* motility under the control of cell density. At low cell densities, the engineered cells are motile, and at high cell density, motility is repressed. This synthetic circuit

results in concentric bands of low and high cell density under conditions similar to those used for our motility assays (10 mL LB soft agar, 37˚C)[156].

Several *E. coli* mutants were identified that exhibited a banded swimming motility behavior. The formation of this banding pattern was media dependent and only formed in LB soft agar. Temperature also had a large impact on the band formation, with formation being more prominent at 37 °C. Decreases in temperature reduces the number of bands and widens their appearance, until at lower temperatures (25-30 °C) where the bands are not visible and chemotaxis rings have reappeared. We demonstrated that the pattern formation does not require either the *luxS*-mediated or the *qseBC*-mediated quorum sensing systems individually; however, a functional Tsr is required for the banding to occur. Tsr mediates chemotaxis towards serine, suggesting that detection of serine is required for banding pattern formation.

Based on our data, we propose the following model: the disruption of *lrhA* results in cells being motile at 37 °C, allowing cells to swim out from the inoculating site. The cell front is consistently moving outwards towards higher concentrations of serine, a strong chemoattractant[114]. The cell front rapidly consumes serine and secretes an unidentified molecule. The concentration of this unidentified molecule increases above a critical threshold at a point behind the cell front. Above the critical threshold, the serine-depleted cells behind the front begin to respond to the secreted molecule, which results in lowered motility and an increase in cellular aggregation. This type of behavior in response to a secreted molecule has been reported previously for *E. coli* mutants[121]. These cells continue to divide and aggregate, resulting in the formation of a band. The cell front, unaffected by the cells behind it, continues moving outward, and the pattern repeats. At lower temperatures the growth rate is slower, and therefore the rate of serine depletion and subsequent secretion of the molecule by the cell front is reduced, preventing

cell aggregation and band formation. Additionally, hypermotility results in loss of pattern formation because the increase in swimming speed reduces cell aggregation. The ability of high concentrations of serine to suppress the pattern formation suggests that serine is available in such excess it never drops below a critical threshold.

Future Directions

In this study, we identified one chemoreceptor, Tsr, that was required for pattern formation. Additionally, screening of other Keio mutants revealed that disruption of *tar* and *trg* also resulted in pattern formation, but it is unclear at this point if these mutations are causative for pattern formation. Further work is needed to determine what role other chemoreceptors have in the pattern formation. Assaying the effects of single chemoreceptor deletions on pattern formation in this background will determine if additional receptors are required for pattern formation.

Because of the coordination required for pattern formation, we investigated the role of quorum sensing systems in the pattern formation. We found that neither the *luxS*-mediated quorum sensing system or the *qseBC*-mediated quorum sensing system was individually required for the banding pattern to form; however, there is a possibility of compensation occurring between the two systems. To rule out this possibility, a triple mutant *(*Δ*luxS* Δ*qseB* Δ*kch*) will be generated to confirm the results that quorum sensing is dispensable for pattern formation.

Several additional *E. coli* mutants were identified that were capable of pattern formation, but in at least one strain (Δ*kch*), we demonstrated that the motility at 37 °C and the pattern formation was not due to the primary deletion. WGS revealed additional mutations in the background of the Δ*kch* strain, one of which was a novel IS insertion event that disrupted a

transcription factor, *lrhA*. LrhA is a global regulator that negatively regulates motility[157], and we believe that this disruption is what allows motility observed at 37 °C. However, it is unclear if the patterning is directly due to the *lrhA* disruption, a combination of one or both of the SNPs and the *lrhA* disruption, or if the pattern represents some underlying process that is typically not observed due to the variation in temperature and media. Transduction of the identified mutations in the Δ*kch* strain into clean strains will be done to determine if any of these mutations individually are sufficient to cause pattern formation. All pattern-forming mutants identified will be screened to determine if they share any of these background mutations to determine if a common mutation(s) is present.

CHAPTER V

CONCLUSIONS

Summary

The goal of this work was to characterize the functional role of a bacterial K^+ channel, Kch, from *Escherichia coli.* Although Kch was the first identified prokaryotic K^+ channel, it remains a poorly understood protein. Initially, deletion of *kch* was reported to have no obvious phenotypes[13,14,65], and subsequent studies focused on over-expression methods to confirm its role as a K^+ channel and to gain insight into its function[65]. However, these studies were plagued by conflicting results, making it difficult to interpret results. Attempts at demonstrating that *kch* was a functional K^+ channel through electrophysiology experiments were unsuccessful with both kch and a chimeric Shaker:Kch channel[14]. A gain-of-function screen provided the best evidence that *kch* functions *in vivo* as a K^+-selective channel. Kuo *et al.* mutagenized *kch* and screened for K^+ channels with gain-of-function mutations, resulting in channels with a higher open probability. These channel mutants all had mutations that likely affected the RCK regulatory domain and were able to grow on normal LB but unable to grow in the presence of high K^+ (200 mM). This growth defect was rescued only by suppressor mutations that likely resulted in collapse of the K^+ selectivity filter, arguing that Kch does transport K^+ *in vivo*[68].

While results indicated that *kch* forms a functional K^+ channel *in vivo*, its importance to *E. coli* physiology was also debated. A modified transposon screen

identified *kch* as a gene required for growth in rich media at 37 °C but dispensable for growth in rich media at cold temperatures or in minimal media at 37 °C[69]. However, a *kch* null mutant was obtained at 37 °C in rich media during the generation of the Keio collection, marking *kch* as non-essential under these conditions[78]. Despite numerous attempts and various methodologies, *kch* function remained cryptic and poorly understood. The purpose of this dissertation was to expand on the limited understanding of bacterial K^+ channels by using novel approaches to determine the physiological function of *kch* in *E. coli*.

The first novel approach was the use of protein co-evolution information to predict protein interaction partners of Kch. Given that more traditional methods for functional characterization had failed to yield significant insight in Kch function, the use of our computational approach could predict which proteins in the cell Kch interacted with. By compiling a list of predicted interaction partners, we were able to draw connections between the relatively uncharacterized protein, Kch, and proteins with known functions. These connections revealed biological processes that Kch was predicted to interact with and allowed us to design targeted experiments to better understand the *in vivo* function of Kch. One of the benefits of our approach was that it allowed us to evaluate a large number of proteins with Kch, as many commonly used methods for PPI detection are difficult to scale significantly[130]. Using a sequence-based analysis allows for predictions of interactions that are transient or even conditionally-specific. For example, some protein-protein interactions may be heavily influenced by environment factors and only occur in certain growth stages, media, temperatures, pH,

etc. These types of interactions may be missed by more conventional methods of PPI detection, as it is not typically feasible to perform large-scale PPI screens in numerous media under a variety of conditions. Additionally, our computational approach could detect other types of interactions, such as transient, low-affinity, or functional interactions, that would also tend to be missed by other methods that require interactions be high affinity and stable in order to be detected.

The large-scale co-evolution analysis predicted that Kch interacted with proteins involved in oxidation-reduction reactions, cell wall synthesis, and metabolism. With this information, we were able to design specific experiments to observe phenotypes with a *kch* null strain. Given the predicted interactions with metabolic proteins, we first asked if loss of *kch* resulted in alterations to growth in a variety of media. We found that Δ*kch* strains exhibited a slight growth defect upon entering mid-exponential phase. This defect could be rescued by the addition of a fermentable carbon source, glucose, but not by the addition of an oxidizable carbon source, succinate, indicating that the loss of the channel was altering aerobic respiration. However, after replacement of the Δ*kch* with a functional channel, the defect persisted, leading to us perform whole genome sequencing to identify the additional mutations present in the background. We identified a point mutation in a gene called *ubiH*, a gene involved in biosynthesis of ubiquinone, the primary electron carrier in aerobic respiration. Kch was predicted to interact with a number of redox proteins, including proteins involved in ubiquinone biosynthesis, which have important roles in respiration and metabolism (Fig 9). In the Keio Δ*kch* mutant (SDB2), the presence of the $ubiH^{V223G}$ mutation provides another line of evidence for a

functional link between the channel and aerobic respiration. The ability to connect Kch to proteins of known function through the co-evolution analysis allowed us to ask more focused questions to gain insight into Kch function.

Another novel approach was the use of CRISPR interference (CRISPRi) coupled with a fluorescent ΔΨ-monitoring dye and high-resolution time-lapse microscopy to monitor ΔΨ changes during depletion of *kch*. CRISPRi-targeting of a gene results in transcriptional repression, leading to depletion of the protein over time. This method has been useful in probing how cells respond to suboptimal levels of essential proteins[124]. Depletion of *kch* resulted in an increased intracellular ThT signal within 30 minutes of CRISPRi induction and persisted throughout the length of the experiment. The goal of this work was to better characterize the *in vivo* function of Kch, and this approach allowed us to demonstrate that at least one function of Kch is to respond to and to modulate ΔΨ fluctuations during rapid growth. Collectively the approaches taken during this work have successfully identified a novel function for Kch in ΔΨ potential modulation and demonstrated that Kch function is critical for adaptation to conditions that lead to rapid growth.

Our result that Kch modulates membrane potential in *E. coli* adds to a growing body of evidence that this may be a conserved function for viral and bacterial K^+ channels. For example, in Chlorella viruses, the Kcv K^+ channel was demonstrated to be important for depolarization of the host cell, which aids in the delivery of viral DNA into the host[158]. Additionally, deletion of the $Ca2^+$-dependent K^+ channel, SynCaK, in the Cyanobacterium *Synechocystic* sp. PCC 6803 resulted in a reduction of membrane

potential (depolarization), indicating that the K^+ channel is involved in modulation of membrane potential. Recent work in *B. subtilis* biofilms revealed that the YugO K^+ channel coordinates long-range signaling of metabolic stress within a bacterial community[140]. YugO-mediated release of K^+ also results in attraction and incorporation of motile cells of diverse species into the biofilm. This attraction appears to be via modulation of the membrane potential in the motile cells, which orients swimming towards the biofilm [13,14]. Current progress in the functional characterization of microbial K^+ channels has revealed novel roles in modulation of membrane potential for these proteins in a diverse number of organisms, indicating that this may be a conserved function. The studies performed in *B. subtilis* biofilms also demonstrated that bacterial K^+ channels are involved in both intercellular and interspecies communication. These results highlight how functional characterization of bacterial ion channels reveals novel insights into microbial physiology and communication.

Discussion

The results of this work revealed a functional connection between the K^+ channel, Kch, and aerobic respiration. Whole genome sequencing of the Keio Δ*kch* strain revealed a potential suppressor mutation in $ubiH^{V223G}$. As a critical component of the ubiquinone biosynthesis process, *ubiH* has an important role in aerobic ETC, a vital component of energy production in Prokaryotes. We demonstrated that this mutation reduces the overall efficiency of the aerobic ETC and propose the following model based on our results:

In *E. coli*, under aerobic conditions, the electron donors NADH and $FADH_2$ are oxidized by NADH dehydrogenases (*nuo, ndh*) and Succinate dehydrogenase respectively. The electrons are then transferred to ubiquinone, the aerobic electron carrier molecule present in the membrane. Ubiquinol, the reduced form of ubiquinone, then transfers electrons to terminal oxidases (*cyo, cyd*), which in turn donate electrons to the terminal electron acceptor, O_2, reducing it to H_2O. The flux of electrons from donors to O_2 is accompanied by proton translocation across the membrane by several respiratory complexes (*nuo, cyo, cyd*)[134,135]. This movement of H^+ across the membrane contributes to both the pH and $\Delta\Psi$ gradients that comprise the total protonmotive force (PMF) of the cell. In bacteria, the PMF is used to power many critical cellular functions such as ATP production, flagellar rotation, protein translocation, and nutrient uptake.

During rapid growth, electron flux through ETC is high and the accompanying H^+ translocation generates strong gradients, where the inside of the cell would be highly negative. Our data indicate that one physiological role for Kch is to rapidly modulate $\Delta\Psi$ to prevent extreme hyperpolarization. In the background of the Keio Δ*kch* strain (SDB2), the $ubiH^{V223G}$ mutation resulted in lowered concentrations of available ubiquinone and reducing electron flux and subsequent H^+ transfer, resulting in a depolarized state. While reducing energy production may seem counter-intuitive, the reduction in efficiency prevents extreme membrane potential fluctuations that the cell is unable to respond to in the absence of the channel.

This model is supported by CRISPRi results where knockdown of *kch* in optimal growth conditions (rich media at 37 °C) resulted in growth defects, consistent with our

hypothesis that channel function is critical under these conditions. However, under conditions where we anticipated that Kch function was less critical to the cell, such as rich media at a lower growth temperature or minimal media at 37 °C, we see that *kch* depletion does not have significantly altered growth. Based on these results, we propose the Kch function is important to allow for adaptation to conditions that promote rapid growth. Kch modulation of ΔΨ during rapid growth prevents the cell from encountering situations of prolonged extreme hyperpolarization that would be detrimental.

It remains unclear why a highly negative charge would be lethal. The electrical and pH gradients generated by the ETC contributes to the overall PMF; therefore, it is possible that extreme changes in the PMF could disrupt the many cellular processes that depend on PMF. Additionally, extremes in membrane charge could lead to a break down in membrane integrity, causing irreversible damage. Further work will be needed to determine why severe hyperpolarization is detrimental.

Future Directions

Based on our characterization of Δ*kch* strain from the Keio collection (SDB2), we hypothesized that the $ubiH^{V223G}$ mutation was a suppressor of the channel deletion. To determine if suppressors always arise in response to loss of the channel, we generated independent Δ*kch* strains and submitted them for whole genome sequencing. Every independent channel knockout also carried a single mutation in a gene called *nemA* ($nemA^{Y359N}$). *NemA* is N-ethylmaleimide (NEM) reductase that is involved in intracellular redox maintenance by reduction of a variety of electrophiles, including

quinones[40,45]. However, sequencing of our parental strain revealed that it also carried this *nemA*^{Y359N} mutation along with another SNP in a gene called *wrbA* (*wrbA*^{G96V}). *WrbA* is an FMN-dependent NAD(P)H:oxidoreductase that is thought to help maintain proper intracellular redox by quinone detoxification[134,135] The presence of these mutations in our parental strain complicates the interpretation of the independent Δ*kch* strains. We anticipated finding suppressor mutations in the background of the independent Δ*kch* strains, but it is unclear if the *nemA*^{Y359N} SNP represents a suppressor since it is also present in the parental background. There are several possible explanations for these results.

First, the presence of *nemA*^{Y359N} is enough to compensate for the loss of the channel. It is possible that the presence of SNPs in the redox genes, *wrbA* and *nemA*, creates an environment where additional mutations are not required to allow for viability of Δ*kch*. To test this possibility, I will obtain a BW25113 strain that contains wild-type *wrbA* and *nemA* genes and to compare to my parental strain, SDB1, using the $DiOC_{2(3)}$ mV assay. If the *nemA*^{Y359N} and the *wrbA*^{G96V} mutations are affecting the cell in a manner similar to the *ubiH*^{V223G}, then I would expect that the SDB1 strain would be depolarized compared to a BW25113 strain with wild-type *wrbA* and *nemA* genes. Additionally, I will generate independent Δ*kch* isolates in the background of BW25113 with wild-type *wrbA* and *nemA* genes and submit them for whole genome sequencing. If the SDB1 strain is not depolarized compared to other BW25113 strains and new independent Δ*kch* isolates have no other mutations in the background, then these results would indicate that

while modulation of ΔΨ by Kch is important to the cell, loss of Kch can be compensated for in other ways.

Monitoring of ΔΨ during *kch* depletion during growth in rich, defined glucose media revealed an increase in intracellular ThT signal after the addition of inducer to the chamber. In some cells, there are morphology changes (Fig 23, bottom panel) associated with the *kch* knockdown, where the cell elongates without dividing. Preliminary experiments with *kch* depletion in rich, defined succinate media show a similar response, but is also associated with cell elongation (Fig 35).

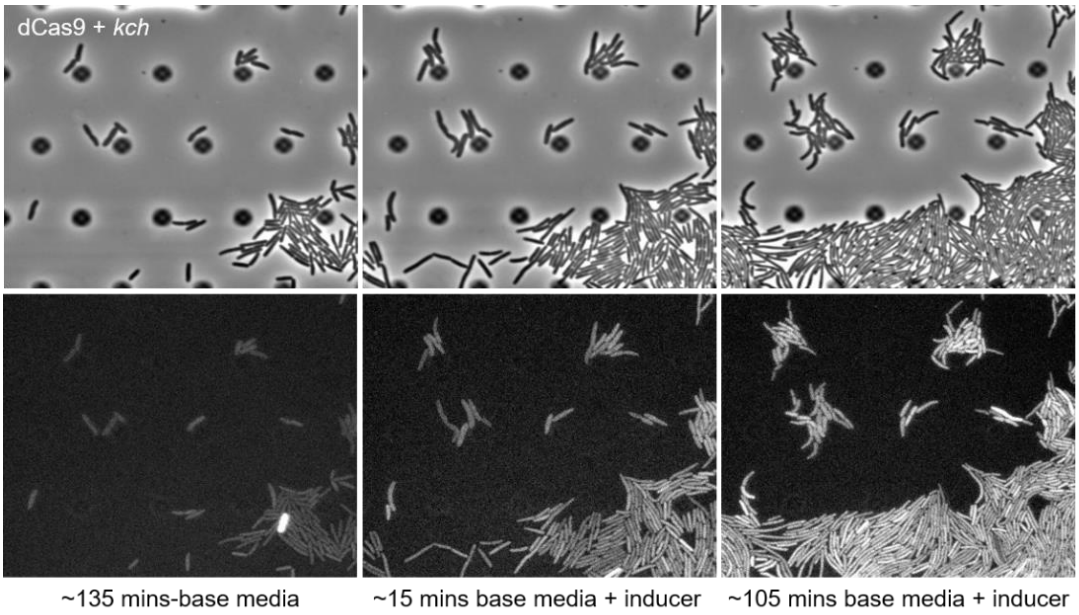

Figure 35.Depletion of kch in succinate media
Kch knockdown in EZ rich, defined succinate media results in hyperpolarization
Base media was EZ rich, defined glucose + appropriate antibiotics + 10 μM ThT.
Inducer (aTc) was added to a final concentration of 2 μM.

Growth on succinate requires aerobic respiration, potentially exacerbating the effects of the *kch* depletion. I will repeat these experiments with the control CRISPRi

strains (dCas9 only and dCas9 + *araC* sgRNA) to confirm their behavior under these conditions. I will track single cells to determine if there is a correlation between the intracellular ThT signal and cell length. One possible explanation for a cell elongation phenotype during loss of Kch is that the severe hyperpolarized state of the cell interferes with assembly of cell wall division machinery. An approach to determine this would be to perform the CRISPRi-mediated *kch* knockdown in the presence of fluorescently tagged cell division proteins. I would expect to see delocalized signal from the cell division proteins that would correlate with severe hyperpolarization events following the addition of inducer.

Based on results from this work, Kch is important for $\Delta\Psi$ modulation during rapid growth. Given the importance of this function and the strong conservation of K^+ channels in prokaryotes, it would be interesting to explore the role of bacterial K^+ channels in pathogens. There have been implications for the importance of K^+ transport systems in a variety of pathogens[159], but the role of K^+ channels has not been widely explored. A transposon screen in Enterohemorrhagic *E. coli* (*EHEC*), revealed that disruption of *kch* resulted in impaired adherence to intestinal epithelial cells *in vitro*, suggesting the channel may be important for virulence[159]. A clean Δ*kch* strain would be generated in an *E. coli* O157:H7 background and submitted for whole genome sequencing to confirm the genotype. The fitness costs associated with loss of the channel would be determined via $\Delta\Psi$ measurements and fitness competitions against the parental strain. Fitness competitions would be performed by labeling the parental and the Δ*kch* strains with different selectable markers. Equivalent starting cell numbers of each

labeled strain would be used to inoculate fresh LB broth. The co-culture would able allowed to grow for ~4.5 hours or until stationary phase is reached, at which point the culture would be serially diluted and plated onto selective media to enumerate the number of each strain in the co-culture. This would be repeated in a variety of media with supplementation with carbon sources such as glucose (fermentable) and succinate (non-fermentable) to see how this altered the ability of the Δ*kch* strain to compete with the parental strain.

Disruption of *kch* has been reported to interfere with adherence to intestinal epithelial cells. In order to determine if disruption of Kch alters the ability to adhere an efface epithelial cells *in vivo,* I propose utilizing a *C. elegans* infection model. *C. elegans* has been used successfully as a model system to assess bacterial colonization and infection[160,161]. Parental and Δ*kch* strains will be assessed for their ability to both colonize and cause infection in this model. Because *C. elegans* is transparent, the strains can be labeled with fluorescent proteins such as mcherry and GFP and bacterial colonization can be quantified microscopically. Additionally, the ability of the parental or Δ*kch* strain to adhere to and to colonize *C. elegans* after 48 hours of exposure will be quantified by washing worms with buffer to remove extracellular bacteria followed by a treatment with buffer and 1% Triton to disrupt the worms[160]. The worm lysates will be serially diluted and plated onto LB agar to determine CFUs. The ability of the parental and Δ*kch* strains to cause disease will be assessed by determining the viability of *C. elegans* over a period of 10 days post-infection. Viability of *C. elegans* fed a non-pathogenic *E. coli* strain (MG1655) will be assessed as a control.

REFERENCES

1. Kuo, M. M. C., Haynes, W. J., Loukin, S. H., Kung, C. & Saimi, Y. Prokaryotic K+ channels: From crystal structures to diversity. *FEMS Microbiology Reviews* **29**, 961–985 (2005).
2. Plugge, B. *et al.* A potassium channel protein encoded by chlorella virus PBCV-1. *Science* **287**, 1641–1644 (2000).
3. Hebert, S. C., Desir, G., Giebisch, G. & Wang, W. Molecular diversity and regulation of renal potassium channels. *Physiological reviews* **85**, 319–71 (2005).
4. Feske, S., Wulff, H. & Skolnik, E. Y. Ion Channels in Innate and Adaptive Immunity. *Annual Review of Immunology* **33**, 291–353 (2015).
5. Atwater, I., Rosario, L. & Rojas, E. Properties of the Ca-activated K+ channel in pancreatic β-cells. *Cell Calcium* **4**, 451–461 (1983).
6. Hoffmann, E. K., Lambert, I. H. & Pedersen, S. F. Physiology of Cell Volume Regulation in Vertebrates. *Physiological Reviews* **89**, 193–277 (2009).
7. Curran, M. E. *et al.* A molecular basis for cardiac arrhythmia: HERG mutations cause long QT syndrome. *Cell* **80**, 795–803 (1995).
8. Tomimitsu, H. *et al.* Mechanism of action of voltage-gated K+ channel antibodies in acquired neuromyotonia. *Annals of Neurology* **56**, 440–444 (2004).
9. Bauer, C. K. *et al.* Mutations in KCNK4 that Affect Gating Cause a Recognizable Neurodevelopmental Syndrome. *The American Journal of Human Genetics* **103**, 621–630 (2018).

10. Kubalski, A. & Martinac, B. *Bacterial Ion Channels and Their Eukaryotic Homologs*. (2005).

11. Milkman, R. An Escherichia coli homologue of eukaryotic potassium channel proteins. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 3510–4 (1994).

12. Schrempf, H. *et al.* A prokaryotic potassium ion channel with two predicted transmembrane segments from Streptomyces lividans. *The EMBO journal* **14**, 5170–5178 (1995).

13. Ungar, D. *et al.* Analysis of a putative voltage-gated prokaryotic potassium channel. *European Journal of Biochemistry* **268**, 5386–5396 (2001).

14. Kuo, M. M. C., Saimi, Y. & Kung, C. Gain-of-function mutations indicate that Escherichia coli Kch forms a functional K+ conduit in vivo. *EMBO Journal* **22**, 4049–4058 (2003).

15. Enyedi, P. & Czirják, G. Molecular Background of Leak K^+ Currents: Two-Pore Domain Potassium Channels. *Physiological Reviews* **90**, 559–605 (2010).

16. Doyle, D. A. *et al.* The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science (New York, N.Y.)* **280**, 69–77 (1998).

17. Jiang, Y. *et al.* Crystal structure and mechanism of a calcium-gated potassium channel. *Nature* **417**, 515–522 (2002).

18. MacKinnon, R. Potassium channels. *FEBS Letters* **555**, 62–65 (2003).

19. Tao, X., Hite, R. K. & MacKinnon, R. Cryo-EM structure of the open high-conductance Ca2+-activated K+ channel. *Nature* **541**, 46–51 (2017).

20. Seoh, S.-A., Sigg, D., Papazian, D. M. & Bezanilla, F. Voltage-Sensing Residues in the S2 and S4 Segments of the Shaker K+ Channel. *Neuron* **16**, 1159–1167 (1996).

21. Jiang, Y. *et al.* Crystal structure and mechanism of a calcium-gated potassium channel. 515–522 (2002).

22. Savalli, N., Pantazis, A., Yusifov, T., Sigg, D. & Olcese, R. The contribution of RCK domains to human BK channel allosteric activation. *The Journal of biological chemistry* **287**, 21741–50 (2012).

23. McLaggan, D., Naprstek, J., Buurman, E. T. & Epstein, W. Interdependence of K+ and glutamate accumulation during osmotic adaptation of Escherichia coli. *Journal of Biological Chemistry* **269**, 1911–1917 (1994).

24. Epstein, W. Osmoregulation by potassium transport in *Escherichia coli*. *FEMS Microbiology Letters* **39**, 73–78 (1986).

25. Epstein, W. The Roles and Regulation of Potassium in Bacteria. *Progress in Nucleic Acid Research and Molecular Biology* **75**, 293–320 (2003).

26. Rozov, A. *et al.* Importance of potassium ions for ribosome structure and function revealed by long-wavelength X-ray diffraction. *Nature Communications* **10**, 2519 (2019).

27. Skulachev, V. P. Membrane electricity as a convertible energy currency for the cell. *Canadian Journal of Biochemistry* **58**, 161–175 (1980).

28. Schlösser, A., Meldorf, M., Stumpe, S., Bakker, E. P. & Epstein, W. TrkH and its homolog, TrkG, determine the specificity and kinetics of cation transport by the Trk system of Escherichia coli. *Journal of bacteriology* **177**, 1908–10 (1995).

29. Schlösser, A., Kluttig, S., Hamann, A. & Bakker, E. P. Subcloning, nucleotide sequence, and expression of trkG, a gene that encodes an integral membrane protein involved in potassium uptake via the Trk system of Escherichia coli. *Journal of bacteriology* **173**, 3170–6 (1991).

30. Dosch, D. C., Helmer, G. L., Sutton, S. H., Salvacion, F. F. & Epstein, W. Genetic analysis of potassium transport loci in Escherichia coli: evidence for three constitutive systems mediating uptake potassium. *Journal of bacteriology* **173**, 687–96 (1991).

31. Schmid, R. *et al.* Identification of the ABC protein SapD as the subunit that confers ATP dependence to the K+-uptake systems TrkH and TrkG from Escherichia coli K-12. *Microbiology* **147**, 2991–3003 (2001).

32. Cao, Y. *et al.* Crystal structure of a potassium ion transporter, TrkH. *Nature* **471**, 336–340 (2011).

33. Cao, Y. *et al.* Gating of the TrkH ion channel by its associated RCK protein TrkA. *Nature* **496**, (2013).

34. Altendorf, K., Siebers, A. & Epstein, W. The KDP ATPase of Escherichia coli. *Annals of the New York Academy of Sciences* **671**, 228–243 (1992).

35. Polarek, J. W., Williams, G. & Epstein, W. The products of the kdpDE operon are required for expression of the Kdp ATPase of Escherichia coli. *Journal of bacteriology* **174**, 2145–51 (1992).

36. Sugiura, A., Hirokawa, K., Nakashima, K. & Mizurto, T. Signal-sensing mechanisms of the putative osmosensor KdpD in Escherichia coli. *Molecular Microbiology* **14**, 929–938 (1994).

37. Laermann, V., Ćudić, E., Kipschull, K., Zimmann, P. & Altendorf, K. The sensor kinase KdpD of Escherichia coli senses external K +. *Molecular Microbiology* **88**, 1194–1204 (2013).

38. Malli, R. & Epstein, W. Expression of the Kdp ATPase is consistent with regulation by turgor pressure. *Journal of bacteriology* **180**, 5102–8 (1998).

39. Heermann, R. *et al.* The Universal Stress Protein UspC Scaffolds the KdpD/KdpE Signaling Cascade of Escherichia coli under Salt Stress. *Journal of Molecular Biology* **386**, 134–148 (2009).

40. Njoroge, J. W., Gruber, C. & Sperandio, V. The interacting Cra and KdpE regulators are involved in the expression of multiple virulence factors in enterohemorrhagic Escherichia coli. *Journal of bacteriology* **195**, 2499–508 (2013).

41. Zakharyan, E. & Trchounian, A. K+ influx by Kup in Escherichia coli is accompanied by a decrease in H+ efflux. *FEMS Microbiology Letters* **204**, 61–64 (2001).

42. Trchounian, A. & Kobayashi, H. Kup is the major K+ uptake system in Escherichia coli upon hyper-osmotic stress at a low pH. *FEBS Letters* **447**, 144–148 (1999).

43. Schleyer, M. & Bakker, E. P. Nucleotide sequence and 3'-end deletion studies indicate that the K(+)-uptake protein kup from Escherichia coli is composed of a hydrophobic core linked to a large and partially essential hydrophilic C terminus. *Journal of bacteriology* **175**, 6925–31 (1993).

44. Holtmann, G., Bakker, E. P., Uozumi, N. & Bremer, E. KtrAB and KtrCD: two K+ uptake systems in Bacillus subtilis and their role in adaptation to hypertonicity. *Journal of bacteriology* **185**, 1289–98 (2003).

45. Gries, C. M., Bose, J. L., Nuxoll, A. S., Fey, P. D. & Bayles, K. W. The Ktr potassium transport system in Staphylococcus aureus and its role in cell physiology, antimicrobial resistance and pathogenesis. *Molecular microbiology* **89**, 760–73 (2013).

46. Bakker, E. P., Booth, I. R., Dinnbier, U., Epstein, W. & Gajewska, A. Evidence for multiple K+ export systems in Escherichia coli. *Journal of bacteriology* **169**, 3743–9 (1987).

47. Elmore, M. J. *et al.* Activation potassium efflux from Escherichia coli by glutathione metabolites. *Molecular Microbiology* **4**, 405–412 (1990).

48. Healy, J. *et al.* Understanding the Structural Requirements for Activators of the Kef Bacterial Potassium Efflux System. *Biochemistry* **53**, 1982–1992 (2014).

49. Booth, I. R., Epstein, W., Giffard, P. M. & Rowland, G. C. Roles of the trkB and trkC gene products of Escherichia coli in K+ transport. *Biochimie* **67**, 83–89 (1985).

50. Ferguson, G. P., Nikolaev, Y., McLaggan, D., Maclean, M. & Booth, I. R. Survival during exposure to the electrophilic reagent N-ethylmaleimide in Escherichia coli: role of KefB and KefC potassium channels. *Journal of bacteriology* **179**, 1007–12 (1997).

51. Follmann, M. *et al.* Potassium transport in corynebacterium glutamicum is facilitated by the putative channel protein CglK, which is essential for pH homeostasis and growth at acidic pH. *Journal of bacteriology* **191**, 2944–52 (2009).

52. Stingl, K. *et al.* Channel-mediated potassium uptake in Helicobacter pylori is essential for gastric colonization. *The EMBO Journal* **26**, 232–241 (2007).

53. Kashket, E. R. The Proton Motive Force in Bacteria: A Critical Assessment of Methods. *Annual Review of Microbiology* **39**, 219–242 (1985).

54. Manson, M. D., Tedesco, P., Berg, H. C., Harold, F. M. & van der Drift, C. A protonmotive force drives bacterial flagella. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 3060–4 (1977).

55. Taylor, B. L. Role of Proton Motive Force in Sensory Transduction in Bacteria. *Annual Review of Microbiology* **37**, 551–573 (1983).

56. Maloney, P. C., Kashket, E. R. & Wilson, T. H. A protonmotive force drives ATP synthesis in bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **71**, 3896–900 (1974).

57. Gunner, M. R., Amin, M., Zhu, X. & Lu, J. Molecular mechanisms for generating transmembrane proton gradients. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1827**, 892–913 (2013).

58. Checchetto, V. *et al.* Functional characterization and determination of the physiological role of a calcium-dependent potassium channel from cyanobacteria. *Plant physiology* **162**, 953–64 (2013).

59. Liu, J. *et al.* Metabolic co-dependence gives rise to collective oscillations within biofilms. *Nature* **523**, (2015).

60. Boogerd, F. C. *et al.* AmtB-mediated NH_3 transport in prokaryotes must be active and as a consequence regulation of transport by GlnK is mandatory to limit futile cycling of NH4+/NH3. *FEBS Letters* **585**, 23–28 (2011).

61. Tolner, B., Ubbink-Kok, T., Poolman, B. & Konings, W. N. Characterization of the proton/glutamate symport protein of Bacillus subtilis and its functional expression in Escherichia coli. *Journal of bacteriology* **177**, 2863–9 (1995).

62. Lundberg, M. E., Becker, E. C. & Choe, S. MstX and a Putative Potassium Channel Facilitate Biofilm Formation in Bacillus subtilis. *PLoS ONE* **8**, e60993 (2013).

63. Prindle, A. *et al.* Ion channels enable electrical communication in bacterial communities. *Nature* **527**, 59–63 (2015).

64. Humphries, J. *et al.* Species-Independent Attraction to Biofilms through Electrical Signaling. *Cell* **168**, 200-209.e12 (2017).

65. Munsey, T. S. Functional Properties of Kch, A prokaryotic homologue of eukaryotic K+ channels. *Decision Support Systems* **1005**, 10–16 (2002).

66. Johansson, M. & von Heijne, G. Membrane topology of Kch, a putative K+ channel from Escherichia coli. *Journal of Biological Chemistry* **271**, 25912–25915 (1996).

67. Voges, D. & Jap, B. K. Recombinant expression, purification and characterization of Kch, a putative Escherichia coli potassium channel protein. *FEBS letters* **429**, 104–8 (1998).

68. Serina, S. *et al.* Scanning the Escherichia coli chromosome by random transposon mutagenesis and multiple phenotypic screening. *Research in Microbiology* **155**, 692–701 (2004).

69. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular systems biology* **2**, 2006.0008 (2006).

70. Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295–299 (1999).

71. Süel, G. M., Lockless, S. W., Wall, M. A. & Ranganathan, R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology* **10**, 59–69 (2003).

72. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417**, 399–403 (2002).

73. Berggård, T., Linse, S. & James, P. Methods for the detection and analysis of protein–protein interactions. *PROTEOMICS* **7**, 2833–2842 (2007).

74. Benhar, M., Forrester, M. T. & Stamler, J. S. Protein denitrosylation: enzymatic mechanisms and cellular functions. *Nature Reviews Molecular Cell Biology* **10**, 721–732 (2009).

75. van Staalduinen, L. M. & Jia, Z. Post-translational hydroxylation by 2OG/Fe(II)-dependent oxygenases as a novel regulatory mechanism in bacteria. *Frontiers in microbiology* **5**, 798 (2014).

76. Loi, V. van, Rossius, M. & Antelmann, H. Redox regulation by reversible protein S-thiolation in bacteria. *Frontiers in microbiology* **6**, 187 (2015).

77. Kuhn, M. L. *et al.* Structural, kinetic and proteomic characterization of acetyl phosphate-dependent bacterial protein acetylation. *PloS one* **9**, e94816 (2014).

78. Phizicky, E. M. & Fields, S. Protein-protein interactions: methods for detection and analysis. *Microbiology and Molecular Biology Reviews* **59**, (1995).

79. Galtier, N. & Dutheil, J. Coevolution within and between genes. *Genome Dynamics* **3**, 1–12 (2007).

80. Cong, Q., Anishchenko, I., Ovchinnikov, S. & Baker, D. Protein interaction networks revealed by proteome coevolution. *Science* **365**, 185–189 (2019).

81. de Juan, D., Pazos, F. & Valencia, A. Emerging methods in protein co-evolution. *Nature Reviews Genetics* **14**, 249–261 (2013).

82. Altenhoff, A. M., Schneider, A., Gonnet, G. H. & Dessimoz, C. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Research* **39**, D289–D294 (2011).

83. Horovitz, A. Double-mutant cycles: a powerful tool for analyzing protein structure and function. *Folding and Design* **1**, R121–R126 (1996).

84. Koonin, E. v. Orthologs, Paralogs, and Evolutionary Genomics. *Annual Review of Genetics* **39**, 309–338 (2005).

85. Correction for Liu and Ochman, Stepwise formation of the bacterial flagellar system. *Proceedings of the National Academy of Sciences* **104**, 11507–11507 (2007).

86. Pakotiprapha, D., Samuels, M., Shen, K., Hu, J. H. & Jeruzalmi, D. Structure and mechanism of the UvrA-UvrB DNA damage sensor. *Nature Structural and Molecular Biology* **19**, 291–298 (2012).

87. Kisker, C., Kuper, J. & van Houten, B. Prokaryotic nucleotide excision repair. *Cold Spring Harbor Perspectives in Biology* **5**, (2013).

88. Kisker, C., Kuper, J. & van Houten, B. Prokaryotic nucleotide excision repair. *Cold Spring Harbor Perspectives in Biology* **5**, (2013).

89. Hmiel, S. P., Snavely, M. D., Miller, C. G. & Maguire, M. E. Magnesium transport in Salmonella typhimurium: Characterization of magnesium influx and cloning of a transport gene. *Journal of Bacteriology* **168**, 1444–1450 (1986).

90. Blattner, F. R. *et al.* The complete genome sequence of Escherichia coli K-12. *Science* **277**, 1453–1462 (1997).

91. Earl, A. M. *et al.* Whole-genome sequences of Bacillus subtilis and close relatives. *Journal of bacteriology* **194**, 2378–9 (2012).

92. Lukjancenko, O., Wassenaar, T. M. & Ussery, D. W. Comparison of 61 sequenced Escherichia coli genomes. *Microbial ecology* **60**, 708–20 (2010).

93. Rose, P. W. *et al.* The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Research* **43**, D345–D356 (2015).

94. Goll, J. *et al.* MPIDB: the microbial protein interaction database. *Bioinformatics* **24**, 1743–1744 (2008).

95. Gene Ontology Consortium: going forward. *Nucleic Acids Research* **43**, D1049–D1056 (2015).

96. Thong, S. *et al.* Defining key roles for auxiliary proteins in an ABC transporter that maintains bacterial outer membrane lipid asymmetry. *eLife* **5**, (2016).

97. Malinverni, J. C. & Silhavy, T. J. An ABC transport system that maintains lipid asymmetry in the gram-negative outer membrane. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 8009–14 (2009).

98. Koronakis, V., Eswaran, J. & Hughes, C. Structure and Function of TolC: The Bacterial Exit Duct for Proteins and Drugs. *Annual Review of Biochemistry* **73**, 467–489 (2004).

99. Wu, T. *et al.* Identification of a protein complex that assembles lipopolysaccharide in the outer membrane of Escherichia coli. *Proceedings of the*

National Academy of Sciences of the United States of America **103**, 11754–11759 (2006).

100. Innis, M. A. *et al.* Nucleotide sequence of the Escherichia coli prolipoprotein signal peptidase (lsp) gene. *Proceedings of the National Academy of Sciences of the United States of America* **81**, 3708–3712 (1984).

101. Shieh, C., Coghlan, M., Sullivan, J. P. & Gopalakrishnan, M. Potassium Channels : Molecular Defects , Diseases , and Therapeutic Opportunities. **52**, 557–593 (2000).

102. Epstein, W. The Roles and Regulation of Potassium in Bacteria. *Progress in Nucleic Acid Research and Molecular Biology* **75**, 293–320 (2003).

103. Gowrishankar, J. Identification of osmoresponsive genes in Escherichia coli: evidence for participation of potassium and proline transport systems in Identification of Osmoresponsive Genes in Escherichia coli : Evidence for Participation of Potassium and Proline Transport. *Journal of bacteriology* **164**, 434 (1985).

104. Voges, D. & Jap, B. K. Recombinant expression, purification and characterization of Kch, a putative Escherichia coli potassium channel protein. *FEBS letters* **429**, 104–8 (1998).

105. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 6640–5 (2000).

106. Goodall, E. C. A. *et al.* The essential genome of Escherichia coli K-12. *mBio* **9**, (2018).

107. M9 minimal medium (standard). *Cold Spring Harbor Protocols* **2010**, pdb.rec12295 (2010).

108. Larson, M. H. *et al.* CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nature Protocols* **8**, 2180–2196 (2013).

109. García-Nafría, J., Watson, J. F. & Greger, I. H. IVA cloning: A single-tube universal cloning system exploiting bacterial In Vivo Assembly. *Scientific Reports* **6**, 1–12 (2016).

110. Jeon, A. B. *et al.* 2-aminoimidazoles collapse mycobacterial proton motive force and block the electron transport chain. *Scientific Reports* **9**, 1513 (2019).

111. Lee, D. D. *et al.* Magnesium Flux Modulates Ribosomes to Increase Bacterial Survival. *Cell* 1–9 (2019). doi:10.1016/j.cell.2019.01.042

112. Bouillaut, L., McBride, S. M. & Sorg, J. A. Genetic manipulation of Clostridium difficile. *Current Protocols in Microbiology* (2011). doi:10.1002/9780471729259.mc09a02s20

113. Rauprich, O. *et al.* Periodic phenomena in Proteus mirabilis swarm colony development. *Journal of Bacteriology* **178**, 6525–6538 (1996).

114. Budrene, E. O., Berg & Berg, H. C. Dynamics of formation of symmetrical patterns by chemotactic bacteria. *Nature* **376**, 49–53 (1995).

115. Kinsinger, R. F., Shirk, M. C. & Fall, R. Rapid surface motility in Bacillus subtilis is dependent on extracellular surfactin and potassium ion. *Journal of bacteriology* **185**, 5627–31 (2003).

116. Eisenbach, M. Tar-dependent and -independent pattern formation by Salmonella typhimurium . These include : Tar-Dependent and -Independent Pattern Formation by Salmonella typhimurium. *J Bacteriol* **177**, 1683–1691 (1995).

117. Schindelin, J. *et al.* Fiji: An open-source platform for biological-image analysis. *Nature Methods* **9**, 676–682 (2012).

118. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. (2012).

119. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

120. Hawkey, J. *et al.* ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. *BMC Genomics* **16**, 667 (2015).

121. Lehnen, D. *et al.* LrhA as a new transcriptional key regulator of flagella, motility and chemotaxis genes in Escherichia coli. *Molecular Microbiology* **45**, 521–532 (2002).

122. Blumer, C. *et al.* Regulation of type 1 fimbriae synthesis and biofilm formation by the transcriptional regulator LrhA of Escherichia coli. *Microbiology* **151**, 3287–3298 (2005).

123. Nakahigashi, K., Miyamoto, K., Nishimura, K. & Inokuchi, H. Isolation and characterization of a light-sensitive mutant of Escherichia coli K-12 with a mutation in a gene that is required for the biosynthesis of ubiquinone. *Journal of Bacteriology* **174**, 7352–7359 (1992).

124. Dünnwald, P. & Unden, G. The Aerobic and Anaerobic Respiratory Chain of Escherichia coli and Salmonella enterica: Enzymes and Energetics. *EcoSal Plus* **1**, (2013).

125. Yang, X., Yu, L., He, D. & Yu, C. A. The quinone-binding site in succinate-ubiquinone reductase from Escherichia coli. Quinone-binding domain and amino acid residues involved in quinone binding. *The Journal of biological chemistry* **273**, 31916–23 (1998).

126. Søballe, B. & Poole, R. K. Ubiquinone limits oxidative stress in Escherichia coli. *Microbiology* **146**, 787–796 (2000).

127. Novo, D., Perlmutter, N. G., Hunt, R. H. & Shapiro, H. M. Accurate flow cytometric membrane potential measurement in bacteria using diethyloxacarbocyanine and a ratiometric technique. *Cytometry* **35**, 55–63 (1999).

128. Hudson, M. A. & Lockless, S. W. Use of an Optimized Fluorescence-Based Assay to Measure E. coli Membrane Potential Changes in Response to Small Molecule Inhibitors. *In Prep* (2019).

129. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).

130. Peters, J. M. *et al.* A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. *Cell* **165**, 1493–1506 (2016).

131. Liu, J. *et al.* Coupling between distant biofilms and emergence of nutrient time-sharing. *Science (New York, N.Y.)* **356**, 638–642 (2017).

132. Liu, X. *et al.* High-throughput CRISPRi phenotyping identifies new essential genes in *Streptococcus pneumoniae*. *Molecular Systems Biology* **13**, 931 (2017).

133. Lee, C., Shin, J. & Park, C. Novel regulatory system *nemRA* - *gloA* for electrophile reduction in *Escherichia coli* K-12. *Molecular Microbiology* **88**, 395–412 (2013).

134. Kishko, I. *et al.* Biphasic Kinetic Behavior of E. coli WrbA, an FMN-Dependent NAD(P)H:Quinone Oxidoreductase. *PLoS ONE* **7**, e43902 (2012).

135. Patridge, E. v & Ferry, J. G. WrbA from Escherichia coli and Archaeoglobus fulgidus is an NAD(P)H:quinone oxidoreductase. *Journal of bacteriology* **188**, 3498–506 (2006).

136. Liu, G. *et al.* Gene Essentiality Is a Quantitative Property Linked to Cellular Evolvability. *Cell* **163**, 1388–1399 (2015).

137. Claessen, D., Rozen, D. E., Kuipers, O. P., Sogaard-Andersen, L. & van Wezel, G. P. Bacterial solutions to multicellularity: a tale of biofilms, filaments and fruiting bodies. *Nat Rev Microbiol* **12**, (2014).

138. Shapiro, J. a. Thinking about bacterial populations as multicellular organisms. *Annual review of microbiology* **52**, 81–104 (1998).

139. Shapiro, J. a. Bacteria as Multicellular Organisms. *Scientific American* **258**, 82–89 (1988).

140. Humphries, J. *et al.* Species-Independent Attraction to Biofilms through Electrical Signaling. *Cell* **168**, 200-209.e12 (2017).

141. Woodward, D. E. *et al.* Spatio-temporal patterns generated by Salmonella typhimurium. *Biophysical journal* **68**, 2181–2189 (1995).

142. Fahrner, K. A. & Berg, H. C. Mutations that stimulate flhDC expression in Escherichia coli K-12. *Journal of Bacteriology* **197**, 3087–3096 (2015).

143. Barker, C. S., Prüß, B. M. & Matsumura, P. Increased motility of Escherichia coli by insertion sequence element integration into the regulatory region of the flhD operon. *Journal of Bacteriology* **186**, 7529–7537 (2004).

144. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754–60 (2009).

145. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357–9 (2012).

146. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987 (2011).

147. Mesibov, R. & Adler, J. Chemotaxis toward amino acids in Escherichia coli. *Journal of bacteriology* **112**, 315–26 (1972).

148. Hedblom, M. L. & Adler, J. Genetic and biochemical properties of Escherichia coli mutants with defects in serine chemotaxis. *Journal of bacteriology* **144**, 1048–60 (1980).

149. Surette, M. G. & Bassler, B. L. Quorum sensing in Escherichia coli and Salmonella typhimurium. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 7046–50 (1998).

150. Xavier, K. B. & Bassler, B. L. LuxS quorum sensing: more than just a numbers game. *Current Opinion in Microbiology* **6**, 191–197 (2003).

151. Sperandio, V., Torres, A. G. & Kaper, J. B. Quorum sensing Escherichia coli regulators B and C (QseBC): A novel two-component regulatory system involved in the regulation of flagella and motility by quorum sensing in E. coli. *Molecular Microbiology* **43**, 809–821 (2002).

152. Clarke, M. B. & Sperandio, V. Transcriptional regulation of flhDC by QseBC and σ28 (FliA) in enterohaemorrhagic Escherichia coli. *Molecular Microbiology* **57**, 1734–1749 (2005).

153. Sperandio, V., Torres, A. G. & Kaper, J. B. Quorum sensing *Escherichia coli* regulators B and C (QseBC): a novel two-component regulatory system involved in the regulation of flagella and motility by quorum sensing in *E. coli*. *Molecular Microbiology* **43**, 809–821 (2002).

154. Liu, C. *et al.* Sequential Establishment of Stripe Patterns in an Expanding Cell Population. *Science* **334**, 238–241 (2011).

155. Cao, Y. *et al.* Collective Space-Sensing Coordinates Pattern Scaling in Engineered Bacteria. *Cell* **165**, 620–30 (2016).

156. Mesibov, R. & Adler, J. Chemotaxis toward amino acids in Escherichia coli. *Journal of Bacteriology* **112**, 315–326 (1972).

157. Purves, D. *et al.* Ion Channels Underlying Action Potentials. (2001).

158. Frohns, F. *et al.* Potassium Ion Channels of Chlorella Viruses Cause Rapid Depolarization of Host Cells during Infection. *Journal of Virology* **80**, 2437–2444 (2006).

159. Tatsuno, I. *et al.* Isolation and characterization of mini-Tn5Km2 insertion mutants of enterohemorrhagic Escherichia coli O157:H7 deficient in adherence to Caco-2 cells. *Infection and Immunity* **68**, 5943–5952 (2000).

160. Schifano, E. *et al.* Virulence behavior of uropathogenic Escherichia coli strains in the host model Caenorhabditis elegans. *MicrobiologyOpen* **8**, (2019).

161. Ritchie, J. M. Animal Models of Enterohemorrhagic Escherichia coli Infection. in *Enterohemorrhagic Escherichia coli and Other Shiga Toxin-Producing E. coli* 175–195 (American Society of Microbiology, 2014). doi:10.1128/microbiolspec.ehec-0022-2013

APPENDIX A

CUSTOM MATLAB SCRIPTS FOR GO ANALYSIS

All custom functions below werre executable in Matlab version 2009b.

```
function run_goanalysis(prot_list,prot_info_file,mat_file, term_list, term_def_list)
%Example: run_goanalysis('high_partners.txt', 'updated_ml_prot_annot.txt',
'init_gSCA_1100_complete.mat', 'sb_goslim_list.txt', 'sb_goterm_def_list.txt')
%  Input: prot_list = list of OMA Ids for protein to be analyzed
%        annot_file = annotation file
%        mat_file = nxn matrix file
%        term_list = list of go slim terms
%        term_def_list = list of go slim term definitions
%
%Takes a list of proteins to be analyzed and runs them through the GO
%Analysis.
%
%Author: Sarah Beagle 7/2/14

names =load_goterms(prot_list);
[annot, rowheaders]=load_prot_annot(prot_info_file);

for i=1:length(names)
   val=names(i,:);
   char_val=char(val);
   new_array=sort_by_pval_mat_zscore(mat_file, prot_info_file, char_val);
   [ min_threshold_struct, max_threshold_struct, min_mapped_goterms_freq,
max_mapped_goterms_freq, all_mapped_goterms_freq, sorted_pvalues
]=thresholds_zscore(term_list, term_def_list, new_array, 1.0e-20, 1.0e-6, char_val);
end
end
```

```
function [ annot, rowheaders ] = load_prot_annot( prot_info_file )
%Usage: [annot, rowheaders]=load_prot_annot(prot_info_file)
%   Example: [annot, rowheaders]=load_prot_annot('updated_ml_prot_annot.txt')
%
% Reads in a tab-delimited file containing annotation information for each
% protein in the dataset, and stores info in a structure array (one
% field/column)
% Input : tab-delimited file containing annotation information
% Output : Structure array containing annotation info (one field/column in
%          the text file)
%          rowheaders : names of the structure fields
% author: Sarah Beagle -- 5/26/14

fid=fopen(prot_info_file);
c_array=textscan(fid, '%s %s %s %s %s %s %s %s %s', 'delimiter', '\t', 'HeaderLines', 1);
%Skips the first line of headers
fclose(fid);
%celldisp(c_array)

%Converting cell array to structure array %%
rowheaders={'OMA_Number', 'Protein_Name', 'UNIPROT_Identifier',
'UNIPROT_Accession_Number', 'OMA_Annotation', 'UNIPROT_Annotation',
'Protein_Family','GO_Slim_Term', 'PDB_ID'};
annot=cell2struct(c_array, rowheaders, 2);

end
```

```
function[mapped_goterms_total_counts,mapped_GOterm_freq,protcounts]=go_slim_
counts_zscore( sb_goslim_terms, mapped_GOterms, prot_name, title_name )
%
% example usage: [mapped_goterms_total_counts,
protcounts]=go_slim_counts('sb_goslim_list.txt', annot.GO_Slim_Term,'ECOLI_91')
% Determines the number of number of proteins mapped to each GO Slim term.
%
% Inputs: sb_goslim_terms = list of terms in the go slim file used for the mapping
%
%       mapped_terms = Ex. annot.GO_Slim_Term, field of a structure array that
contains the mapping info for each protein
%
%       prot_name = OMA identifier ('ECOLI_1', 'ECOLI_2', etc)
%
%       title_name = string that will allow for unique naming of graphs
%       and output files, (i.e. 'All', 'minthreshold', etc)
% Outputs: mapped_goterms_total_counts = matrix containing the counts of each term,
ordered by the sb_goslim_terms list
%
%        mapped_GOterm_freq=vector containing the frequency of each GO term
%
%       protcounts = matrix containing the number of GO terms associated with each
protein, ordered by annot file
%
%       - a tab-delimited text file containing the terms and their
%         associated counts and frequencies
%       - a bar graph of the goslim_counts
%       - a histogram of protein counts
%
% author: Sarah Beagle--4/28/14
% edited 5/7/14 to add graphs and counts for the #terms/protein
% edited 6/4/14 adjusted an error in the input--now handles multiple GO
% terms per protein (format GOterm1,GOterm2,GOterm3, etc). Added graph
% annotation.

%%reads list into cell array
terms=load_goterms(sb_goslim_terms);

%creates unique filename--adds date and time to name--processes to remove
%"- and :" to create a valid filename
%filename=sprintf('%s_sb_goslim_counts_zscore_%s_%s.txt', prot_name, title_name,
datestr(now));
%filename=(strrep(filename,'-','_'));
%filename=(strrep(filename,':','_'));
```

```
%filename=(strrep(filename,' ','_'));

%%counts the number of times a protein maps to a specific term

%Check dimensions of the mapped_GOterm array and reformat if needed (needs
%to be in a GO_Slim_Term: {25x1 cell}--added this because reassignment to a
%new structure can spontaneously switch the original dimensions, which
%breaks the function
if(size(mapped_GOterms,2)>1)
   mapped_GOterms=mapped_GOterms';
end

%Determine the number of times each unique GO term occurrs in the mapped
mapped_GOterms=regexprep(mapped_GOterms, '^"|"$', ''); % removes the "" that
bracket some of the strings in the cell array
mapped_GOterms=regexp(mapped_GOterms, ',', 'split');  % splits the string into
individual GO terms at the comma delimiter

for i=1:length(terms)
  for j=1:length(mapped_GOterms)
    val=terms{i};
    mapped_goterms_ind_count(j)=ismember(val, mapped_GOterms{j});
    mapped_goterms_total_counts(i,:)=sum(mapped_goterms_ind_count);
  end
end

%Determine frequencies
total_counts=sum(mapped_goterms_total_counts); %Determine total number of terms
mapped (can be higher than the # of proteins in dataset as 1 prot can map to mult. terms)
mapped_GOterm_freq=(mapped_goterms_total_counts./total_counts)*100;

%Print text file containing the GO term, actual counts, and frequencies
%fid=fopen(filename, 'W');
%fprintf(fid, 'GO_Term\tActual_Counts\tFrequencies\n');
%for i=1:length(terms)
% fprintf(fid,'%s\t%d\t%f\n', terms{i}, mapped_goterms_total_counts(i),
mapped_GOterm_freq(i));
%end
%fclose(fid);

%create bar graph of go slim counts
%Create unique filename to prevent overwriting
filename2=sprintf('%s_sb_goslim_counts_zscore_%s_%s.png',prot_name,title_name,dat
estr(now));
```

```
filename2=(strrep(filename2,'-','_'));
filename2=(strrep(filename2,':','_'));
filename2=(strrep(filename2,' ','_'));
%create a horizontal bar graph for easier reading of tick labels
 figure('visible','off');
 barh(mapped_goterms_total_counts);
% convert cell array containing GO terms to char array
char_terms=char(terms);
set(gca,'YTick', (1:33));
set(gca, 'YTickLabel', {char_terms});
xlabel('Percentage');
ylabel('GO terms');
name=sprintf('%s_%s_prot_parters_GO_terms',prot_name, title_name);
title(name,'Interpreter', 'none');
print('-dpng', filename2)

%counting the number of GO terms per protein and creating a histogram
%can't use regexp_counts(i)=numel(regexpi(mapped_GOterms{i}, string));
%because regexp outs an empty 1x1 cell if there is no match and numel
%counts the number of elements in a cell. So it will count non-matching
%cells as a 1 and skew the distribution.
string = 'GO:[0-9]+';
for i=1:size(mapped_GOterms,1)
   regexp_counts{i}=regexpi(mapped_GOterms{i}, string);
end

for i=1:length(regexp_counts)
  if (isempty(regexp_counts{i}{1}==1))
    protcounts(i)=0;
  else
    protcounts(i)=numel(regexp_counts{i});
  end
end

max_val=max(protcounts(:));
nbins=max_val+1;
hist(protcounts,nbins,'visible','off');
xlabel('Number of Terms');
ylabel('Number of Proteins');
name2=sprintf('%s_Number of GO terms per protein_%s', prot_name,title_name);
title(name2,'Interpreter', 'none');

filename3=sprintf('%s_prot_counts_%s_%s.png', prot_name,title_name, datestr(now));
filename3=(strrep(filename3,'-','_'));
```

```
filename3=(strrep(filename3,':','_'));
filename3=(strrep(filename3,' ','_'));

print('-dpng', filename3)
end
```

```
function [zscores, goterm_std, goterm_avg, no_of_draws,total_pool_size]  =
significance_zscores(sb_goslim_terms, max_threshold_struct,new_array,
max_threshold_goterms_counts,  no_of_cycles)
% Usage:
%   Example : [output]=significance(annot, new_array, 1000)
% Input: annot = structure array containing annotation info for all
%       proteins in the dataset
%
%       new_array = structure array containing annotation info for just
%       protein partners of the protein of interest (output from sort_by_pval_mat.m)
%
%       max_threshold_struct = new struct array that contains the annotation
%       info for all interaction partners greater than/equal to the max
%       threshold p value
%
%       max_threshold_goterm_counts = all_mapped_goterms_freq = frequencies for
each term for all
%       evaluated proteins ("background"--ordered by the sb_goslim_terms file)
%
%       number_of_cycles = number of times to select random datasets (ie
%       10, 100, 1000)
%
% Output: zscores = vector containing the significance values for each GO term
%       group, ordered by sb_goslim_terms
%
%       goterm_std = vector containing the std values calculcated from
%       the random trials for each GO term group, ordered by sb_goslim_terms
%
%       goterm_avg = vector containing the averages calculated from the
%       random trials for each GO term group, ordered by sb_goslim_terms
%
%       no_of_draws = number of proteins that are randomly sampled,
%       determined by the number of alignments that are equal/greater
%       than the max threshold cutoff
%
%       total_pool_size = the number of proteins that no_of_draws is
%       being sampled from--determined by the total number of evaluated
%       alignments
%
% Author: Sarah Beagle--6/9/14
% edited 8/7/14 --corrected issue with zscore calculation--was taking the
% actual counts value from the "all" dataset, but it should have been from
% the "max" dataset.
% Edited 9/2/14 -- to report no_of_draws/total_pool_size for printing in
```

```
% thresholds_zscore.m

%%reads list into cell array
terms = load_goterms(sb_goslim_terms);

%Getting frequencies

actual_counts=max_threshold_goterms_counts; %% actual counts for each term (all
protein partners that could be evaluated)
%F_post=max_threshold_goterms_freq; %% freqs for each term after an imposed pval
threshold

%Selecting random datasets
%tic
no_of_draws=size(max_threshold_struct.OMA_Number,2);
total_pool_size=size(new_array.OMA_Number,2);
rand_data=cell(no_of_draws, no_of_cycles);
rand_vals_ind=zeros(no_of_draws, no_of_cycles);
%rand_goterms = cell(no_of_draws, no_of_cycles);
rand_goterms_counts=zeros(length(terms), no_of_cycles);

for i=1:no_of_cycles
 rand_data(:,i)=randsample(new_array.OMA_Number, no_of_draws);
end
%finding the indices of randomly selected proteins and getting the GO annotation
%for s=1:no_of_cycles
for i=1:size(rand_data,2)
  for j =1:size(rand_data,1)
   val=rand_data(j,i);
   rand_vals_ind(j,i)=find(strcmp( val,new_array.OMA_Number));
   rand_goterms{j,i}=new_array.GO_Slim_Term(rand_vals_ind(j,i));
   rand_goterms_mat(j,i)=[rand_goterms{j,i}]';
   rand_goterms_mat(j,i)=regexp(rand_goterms_mat(j,i),',', 'split');
   rand_goterms_mat2{j,i}=regexprep(rand_goterms_mat{j,i}, '^"|"$', '');
  end
end

%Calculate GO distribution/round
for i=1:length(rand_goterms_mat2)
    temp_str=cat(2,rand_goterms_mat2{:,i}); %converts the row from a cell array of cells
to a cell array of strings
    for j=1:length(terms)
        val=terms{j};
        rand_goterm_counts(j,i)=numel(find(strcmp(val, temp_str)));
```

```
    end
end
% total_counts=sum(mapped_goterms_totcounts);

%toc
goterm_avg=mean(rand_goterm_counts, 2);
goterm_std=std(rand_goterm_counts,0,2);

%Calculating sig term
zscores=zeros(size(terms,1),1);
for i=1:length(terms)
   zscores(i,:)=(actual_counts(i,:)-goterm_avg(i,:))/goterm_std(i,:);
end

end
```

```
function new_array=sort_by_pval_mat( pval_matrix_matfile, annot_file, prot_name )
%Usage: sort_by_pval_mat(pval_matrix, annot, prot_name)
% Example : new_array=sort_by_pval_mat('init_gSCA.mat',
'updated_ml_prot_annot.txt', 'ECOLI_91');
%
%   Inputs: pval_matrix_matfile = nxn matrix containing the following:
%
%              1)p_value_mat = 2D matrix(1717x1717x2) that contains the
%                pos p values (1st) and the pert p values (2nd)
%
%              2)order_of_aln = vector containing the actual protein order
%                (protein names were randomized to hopefully prevent bias)
%
%           annot_file = tab-delimited text file containing protein annotation info
%           prot_name = OMA identifier of the protein of interest
%
%   Outputs: new_array=structure array containing only the annotation info
%           for the protein partners that were able to be evaluated
%
%           Text file containing a list of protein partners (just the
%           OMA number i.e. 973 instead of ECOLI_973)
%
%           text file with the annotation information for all interaction
%           partners plus an additional column for the p-values,sorted by
%           p value and filtered to remove alignments with errors
%
%  Author: Sarah Beagle 5/4/14 (May the fourth be with you!!!)
%
%%%%%%  Edited to take in init_gSCA.mat as input instead of the output of the
%  compilation.m script. %%%%
% EDITED 6/6/14--Includes corrected go_slim_counts.m function
% EDITED 7/9/14--Added the load_prot_annot function to reduce number of
% lines required to run the GO analysis. Also added the protein of interest
% (prot_name) to the new_array to facilitate reporting the GO slim terms
% that prot_name maps to in the final graph.
% Edited 9/3/14--Now handles proteins which don't have any evaluated
% alignments (all error codes)

%Load protein annotation information
[annot, rowheaders] = load_prot_annot(annot_file);

% Load pval_matrix.mat file
  load(pval_matrix_matfile);
```

```
% Find the TOTMI order of proteins in the matrix
   for i=1:size(annot.OMA_Number,1)
     temp=annot.OMA_Number{i,:};
     temp=temp(7:end);
     temp=str2num(temp);
     totmi_prot_ind(i)=find(order_of_aln == temp);
   end
%Find the protein of interest in annotation file and matrix
prot_name_annot_ind=find(strcmp(annot.OMA_Number, prot_name));
prot_name_matrix_ind=totmi_prot_ind(prot_name_annot_ind);


%creating max_p_value matrix. Compares the pos pval matrix to the pert
%value matrix, if the pos pval > pert pval, a 1 (true) is placed in that
%element in the gt_matrix. If pos pval < pert pval a 0 (false) is recorded.
%The gt_matrix is looped over--every position where a 0 is recorded, the
%value is retrieved from that location in the pert pval matrix, and every
%position with a 1 recorded the value is obtained from the pos pval matrix.
%Cases where the pos pval = pert pval (typically error codes), a 0 will be
%recorded and the max pval will be obtained from the pert pval matrix.
%This is faster than looping through each element, comparing them, and
%writing the larger value to a new matrix.
max_pval_matrix=zeros(1717,1717,1);
gt_matrix=(p_value_mat(:,:,1) > p_value_mat(:,:,2));

%tic
for i=1:size(gt_matrix,1)
 for j=1:size(gt_matrix,2)
     if(gt_matrix(i,j,1) == 0)
       max_pval_matrix(i,j,1) = p_value_mat(i,j,2);
     else
       max_pval_matrix(i,j,1) = p_value_mat(i,j,1);
     end;
  end;
end;
%toc

%Getting p values for all interaction partners of protein of interest
int_partners_maxpvals=max_pval_matrix(prot_name_matrix_ind,:);

%Sort interaction partners by max p value (descending order)
%Find and eliminate alignments that weren't able to be evaluated(NaN error
%code)
```

```
[nan_sorted_int_maxpval nan_sorted_int_maxpval_ind]=sort(int_partners_maxpvals,
'descend');
nan_temp_sorted_vals = isnan(nan_sorted_int_maxpval); %returns 1 for all elements
that equal NaN
nan_errors=sum(nan_temp_sorted_vals == 1); %returns the number of elements that ==
NaN
nan_filtered_int_maxpvals=nan_sorted_int_maxpval(nan_errors+1:end); %removes
NaN values from sorting
nan_filtered_int_maxpval_ind=nan_sorted_int_maxpval_ind(nan_errors+1:end);%remo
ves NaN value indices from sorting

%Removing names of alignments that had NaN errors
nan_sorted_order_of_aln=order_of_aln(nan_sorted_int_maxpval_ind);
nan_filtered_order_of_aln=nan_sorted_order_of_aln(nan_errors+1:end);

%Finding and removing all alignments associated with various error codes
[error_sorted_int_maxpval
error_sorted_int_maxpval_ind]=sort(nan_filtered_int_maxpvals,'ascend');
errors=sum(error_sorted_int_maxpval < 0); %all error codes are negative
final_filtered_int_maxpvals=error_sorted_int_maxpval(errors+1:end);
final_filtered_int_maxpval_ind=error_sorted_int_maxpval_ind(errors+1:end);

%removing names of alignments that had errors
error_sorted_order_of_aln=nan_filtered_order_of_aln(error_sorted_int_maxpval_ind);
final_filtered_order_of_aln=error_sorted_order_of_aln(errors+1:end);


%Finding the interaction partners that were analyzed in annotation file
%Checks to ensure that there alignments analyzed--if
%final_filtered_order_of_aln is empty then it prints a txt file with the
%prot_name, saying that that particular protein has no analyzable
%alignments. It then returns new_array without any fields (essentially
%an empty structure), and exits the function.

 TF=isempty(final_filtered_order_of_aln);
   if (TF == 1)
     filename_error=sprintf('%s_thresholds_zscore_%s.txt',prot_name, datestr(now));
     filename_error=(strrep(filename_error,'-','_'));
     filename_error=(strrep(filename_error,':','_'));
     filename_error=(strrep(filename_error,' ','_'));

     %figure('Position', [100 100 1250 1100], 'visible','off');

     fid=fopen(filename_error, 'w');
```

```
    fprintf(fid, 'Evaluated Protein Partners of %s\n', prot_name);
    fprintf(fid,'%s has no evaluated alignments!', prot_name);
    fclose(fid);

    new_array=struct;
    fprintf('%s has no evaluated alignments!', prot_name);
    return
  else
    header='ECOLI_';
    for i=1:length(final_filtered_order_of_aln)
    val=num2str(final_filtered_order_of_aln(i));
    str=strcat(header, val);
    int_annot_ind(i)=find(strcmp(annot.OMA_Number, str));
  end
end

%incorporating prot_name into new_array
int_annot_ind(end+1)=prot_name_annot_ind;
final_filtered_int_maxpvals(end+1)=1;

%Creating new array containing annotation info for evaluated protein
%partners (additional column for p values)

new_array=struct;
for i=1:length(int_annot_ind)
   val=int_annot_ind(i);
   new_array.OMA_Number{:,i}=annot.OMA_Number{val};
   new_array.Protein_Name{:,i}=annot.Protein_Name{val};
   new_array.UNIPROT_Identifier{:,i}=annot.UNIPROT_Identifier{val};

new_array.UNIPROT_Accession_Number{:,i}=annot.UNIPROT_Accession_Number{v
al};
   new_array.OMA_Annotation{:,i}=annot.OMA_Annotation{val};
   new_array.UNIPROT_Annotation{:,i}=annot.UNIPROT_Annotation{val};
   new_array.Protein_Family{:,i}=annot.Protein_Family{val};
   new_array.GO_Slim_Term{:,i}=annot.GO_Slim_Term{val};
   new_array.Max_Pval{:,i}=final_filtered_int_maxpvals(i);
end

%Determining GO counts;
[mapped_goterms_total_counts, protcounts]=go_slim_counts_zscore('sb_goslim_list.txt',
new_array.GO_Slim_Term, prot_name, 'all');
```

```
%Create unique filename for printing to text file-adds date and time to name to prevent
any potential overwriting--processes to remove
%"- and :" to create a valid filename for fopen to write to
filename=sprintf('%s_prot_partners_annot_info_%s.txt',prot_name, datestr(now));
filename=(strrep(filename,'-','_'));
filename=(strrep(filename,':','_'));
filename=(strrep(filename,' ','_'));

%Create and write annotation information to text file
fid=fopen(filename, 'w');
fprintf(fid,'OMA_Number\tProtein_Name\tUNIPROT_Identifier\tUNIPROT_Acesssion
_Number\tOMA_Annotation\tUNIPROT_Annotation\tProtein_Family\tGO_Slim_Term\
tMax_P_Value\n');
for i=1:length(int_annot_ind)
  val=int_annot_ind(i);
  fprintf(fid, '%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%d\n', annot.OMA_Number{val},
annot.Protein_Name{val}, annot.UNIPROT_Identifier{val},
annot.UNIPROT_Accession_Number{val}, annot.OMA_Annotation{val},
annot.UNIPROT_Annotation{val},
annot.Protein_Family{val},annot.GO_Slim_Term{val},final_filtered_int_maxpvals(i));
end
fclose(fid);

%Create and write full alignment information to text file, ordered in
%ascending p values and filtered to remove the alignments with errors
filename2=sprintf('%s_evaluated_prot_partners_%s.txt', prot_name, datestr(now));
filename2=(strrep(filename2,'-','_'));
filename2=(strrep(filename2,':','_'));
filename2=(strrep(filename2,' ','_'));

fid=fopen(filename2, 'w');
fprintf(fid, 'Evaluated Protein Partners of %s\n', prot_name);
%for row=1:size(final_filtered_order_of_aln,2)
 %  fprintf(fid, '%s\n', final_filtered_order_of_aln{row,:});
%end
dlmwrite(filename2,final_filtered_order_of_aln, '-append', 'delimiter', '\n');
fclose(fid);

end
```

```
function [ min_threshold_struct, max_threshold_struct, min_mapped_goterms_freq,
max_mapped_goterms_freq, all_mapped_goterms_freq, sorted_pvalues ] =
thresholds_zscore(sb_goslim_terms,sb_goslim_defs,
new_array,min_threshold,max_threshold,prot_name )
%usage = [threshold1_array, threshold2_array] =
thresholds(prot_info_list,sb_goslim_terms)
% Example: [output_args
]=thresholds('ECOLI_91_interaction_partners_08_May_2014_17_17_21.txt')
%
%   Sorts potential interaction partners (IPs) by p-value. Creates two subsets
%   of IPs based on the values given by min_threshold(less stringent) and
%   max_threshold(most stringent). Compares the two subsets to the whole
%   dataset, and calculates how significant the GO term distribution at the
%   most stringent threshold is compared to random.
%
% Input : sb_goslim_terms = text file containing GO slim terms
%
%       sb_goslim_defs = text file containing GO slim term definitions
%
%       new_array = structure array containing annot info and p values
%       for only protein partners that were able to be evaluated - output
%       from sort_by_pval_mat.m function
%
%       min_threshold = first cutoff value (less stringent), given in the following format :
1.0e-20
%
%       max_threshold = second cutoff value (more stringent), given in the following
format : 1.0e-01
%
%       prot_name = OMA identifier ('ECOLI_1', 'ECOLI_2') for use in the figure/file
names
%
% Output : min_threshold_struct = new struct array that contains the annotation
%        info for all interaction partners greater then/equal to the min
%        threshold p value
%
%       max_threshold_struct = new struct array that contains the annotation
%       info for all interaction partners greater than/equal to the max
%       threshold p value
%
%       all_mapped_goterms_counts = frequencies for each term for all
%       evaluated proteins ("background")
%
%       min_mapped_goterms_freq = frequencies for each term at the min threshold
```

```
%
%       max_mapped_goterms_freq = frequencies for each terms at the max
%       threshold
%
% Author: Sarah Beagle 6/6/14
% Edited: 7/10/14 -- added subplot to display GO Slim categories that the
% protein of interest mapped to for easier interpretation of graph
% Edited 9/2/14--added counts for the number of proteins drawn(size of the
% max_threshold structure--"strongest pvalue alignments)/the total
% number of proteins available to draw from (number of total evaluated
% alignments
% Edited 9/3/14--Now handles proteins which don't have any
% evaluated alignments (all error codes)

%%reads list into cell array
terms=load_goterms('sb_goslim_list.txt');
term_defs=load_goterms('sb_goterm_def_list.txt');

%Check new_array to see if it is empty--if array is empty, then the protein
%being analyzed didn't have any joint alignments that could be evaluated.
%The function will return the normal output as empty structures/arrays and
%exit. No figures will be generated.
TF=isempty(fieldnames(new_array));
if (TF==1)
   min_threshold_struct=struct;
   max_threshold_struct=struct;
   min_mapped_goterms_freq=zeros(1,1);
   max_mapped_goterms_freq=zeros(1,1);
   all_mapped_goterms_freq=zeros(1,1);
   sorted_pvalues=zeros(1,1);
 return
end

%Converts the P value field of the struct. array to a matrix (find function is not easily
used with structured arrays)
max_pval_mat = cell2mat(new_array.Max_Pval)';

%Determine the GO term distribution for all protein partners
[all_mapped_goterms_total_counts, all_mapped_goterms_freq,
all_protcounts]=go_slim_counts_zscore('sb_goslim_list.txt',new_array.GO_Slim_Term,
prot_name,'all_proteins');

%Select all interaction partners with p values equal to/greater than the min
%threshold and create a new struct array containing the prot info for those selected
```

```
min_threshold_ind=find(max_pval_mat >= min_threshold);


%tic
   min_threshold_struct=struct;
for i=1:length(min_threshold_ind)
   val=min_threshold_ind(i);
   min_threshold_struct.OMA_Number{i}=new_array.OMA_Number{val};
   min_threshold_struct.Protein_Number{i}=new_array.Protein_Name{val};
   min_threshold_struct.UNIPROT_Identifier{i}=new_array.UNIPROT_Identifier{val};

min_threshold_struct.UNIPROT_Accession_Number{i}=new_array.UNIPROT_Accessi
on_Number{val};
   min_threshold_struct.OMA_Annotation{i}=new_array.OMA_Annotation{val};

min_threshold_struct.UNIPROT_Annotation{i}=new_array.UNIPROT_Annotation{val
};
   min_threshold_struct.Protein_Family{i}=new_array.Protein_Family{val};
   min_threshold_struct.GO_Slim_Term{i}=new_array.GO_Slim_Term{val};
   min_threshold_struct.Max_Pval{i}=new_array.Max_Pval{val};
end
%toc
%Determine the GO term distribution for all protein partners with p values
%>= the min_threshold
[min_mapped_goterms_total_counts,min_mapped_goterms_freq,
min_protcounts]=go_slim_counts_zscore('sb_goslim_list.txt',
min_threshold_struct.GO_Slim_Term,prot_name,'min_threshold');

%Select all interaction partners with p values equal to/greater than the
%max threshold and create a new struct array containing the prot info for those selected
max_threshold_ind=find(max_pval_mat >= max_threshold);

max_threshold_struct=struct;
for i=1:length(max_threshold_ind)
   val=min_threshold_ind(i);
   max_threshold_struct.OMA_Number{i}=new_array.OMA_Number{val};
   max_threshold_struct.Protein_Number{i}=new_array.Protein_Name{val};

max_threshold_struct.UNIPROT_Identifier{i}=new_array.UNIPROT_Identifier{val};

max_threshold_struct.UNIPROT_Accession_Number{i}=new_array.UNIPROT_Access
ion_Number{val};
   max_threshold_struct.OMA_Annotation{i}=new_array.OMA_Annotation{val};
```

```
max_threshold_struct.UNIPROT_Annotation{i}=new_array.UNIPROT_Annotation{val
};
   max_threshold_struct.Protein_Family{i}=new_array.Protein_Family{val};
   max_threshold_struct.GO_Slim_Term{i}=new_array.GO_Slim_Term{val};
   max_threshold_struct.Max_Pval{i}=new_array.Max_Pval{val};
end

[max_mapped_goterms_total_counts,max_mapped_goterms_freq,
max_protcounts]=go_slim_counts_zscore('sb_goslim_list.txt',
max_threshold_struct.GO_Slim_Term,prot_name,'max_threshold');

%Determining significance for each group

[zscores, goterm_std, goterm_avg, no_of_draws,
total_pool_size]=significance_zscores('sb_goslim_list.txt', max_threshold_struct,
new_array, max_mapped_goterms_total_counts, 1000);

%Grouping the three datasets together
%All counts are ordered by the input GO slim list so indices will be the
%same. Sort the counts/freqs for the whole dataset by the significance values
%and then find the indices of the sorted data. The indices were used to pull the
%corresponding counts/freqs from the appropriate min/max threshold arrays.
%This will allow for plotting a graph containing all three datasets.
%order of the columns in matrices: 1) Whole dataset 2) min thresh 3) max
%thresh

%set all NaN values to zero
zscores(isnan(zscores))=0;

%Sort counts and frequencies by significance values
[sorted_zscores, sorted_zscores_ind]=sort(zscores, 'descend');
%sorted_goterm_std = goterm_std(sorted_sigs_ind);
sorted_gocounts_mat=zeros(length(terms),3);
sorted_gocount_freq_mat=zeros(length(terms),3);
sorted_terms=terms(sorted_zscores_ind);
sorted_term_defs=term_defs(sorted_zscores_ind);
sorted_goterm_std=goterm_std(sorted_zscores_ind);
sorted_goterm_avg=goterm_avg(sorted_zscores_ind);

 for j=1:length(sorted_zscores_ind)
   val2=sorted_zscores_ind(j);
   sorted_gocounts_mat(j,1)=all_mapped_goterms_total_counts(val2);
   sorted_gocount_freq_mat(j,1)=all_mapped_goterms_freq(val2);
```

```
  sorted_gocounts_mat(j,2)=min_mapped_goterms_total_counts(val2);
  sorted_gocount_freq_mat(j,2)=min_mapped_goterms_freq(val2);
  sorted_gocounts_mat(j,3)=max_mapped_goterms_total_counts(val2);
  sorted_gocount_freq_mat(j,3)=max_mapped_goterms_freq(val2);
 end

 %Converting zscores to one-tailed pvalues

 sorted_pvalues=zeros(length(sorted_zscores),1);
 for i =1:length(sorted_pvalues)
    sorted_pvalues(i,:)=normcdf(-abs(sorted_zscores(i)),sorted_goterm_avg(i),
sorted_goterm_std(i));
 end

%Creating a unique filename
filename=sprintf('%s_thresholds_zscore_%s_%s.png',prot_name, datestr(now));
filename=(strrep(filename,'-','_'));
filename=(strrep(filename,':','_'));
filename=(strrep(filename,' ','_'));

%Determining which GO Slim terms the protein of interest maps to
prot_name_ind=find(strcmp(new_array.OMA_Number, prot_name));
prot_name_goterms=new_array.GO_Slim_Term(prot_name_ind);
prot_name_goterms=regexprep(prot_name_goterms, '^"|"$', '');
prot_name_goterms=regexp(prot_name_goterms, ',', 'split');  % splits the string into
individual GO terms at the comma delimiter

%Checks to make sure the last cell doesn't just contain " marks--removes last cell if it
does
quotes= '';
YN=strcmp(prot_name_goterms{:}(end), quotes);
if (YN==1)
   prot_name_goterms{:}(end) = [];
end

for i=1:length(prot_name_goterms{:})
   is_a_match= ~cellfun(@isempty, regexp(terms, prot_name_goterms{:}(i), 'match'));
   vector_of_indices(i) = find(is_a_match);
end

prot_name_termdefs=term_defs(vector_of_indices);

%Graphing frequency data
fighandle=figure('Position', [100 100 1250 1100], 'visible','off');
```

```
subplot('Position', [0.075 0.45 0.75 0.525]);
width=.99;
bar(sorted_gocount_freq_mat, width, 'histc')
hold on;
x=1:33;
y=sorted_gocount_freq_mat(:,3);
%errorbar(x+0.54,y,sorted_goterm_std,'k.');
hold off;
char_sorted_goterm_def=char(sorted_term_defs);
set(gca, 'XTick', (1:33));
%set(gca, 'YTick', 1:2:25);
set(gca, 'XTickLabel', {char_sorted_goterm_def});
rotateXLabels(gca(), 90);
xlabel('GO terms');
ylabel('Frequency');
titlename=sprintf('%s_GO term distribution for various thresholds (%d/%d)', prot_name,
no_of_draws, total_pool_size);
title(titlename, 'Interpreter', 'none');
figlegend=legend('All','min thresh','max thresh');
%figlegend2=legend(ah,'%d/%d',no_of_draws, total_pool_size, 'location', 'NorthWest');

%Graphing significance values
subplot('Position', [0.075 0.05 0.75 0.15]);
bar(sorted_zscores);
set(gca, 'XTick', (1:33));
ylabel('zscore value');
xlabel('GO Terms');
title('Zscore Values for GO Terms');
set(gcf, 'PaperPositionMode', 'auto');

%Inserting textbox with the GO Slim categories that protein of interest
%mapped to for comparision with bar graph
subplot('Position', [0.84 0.85 0.15 0.12]);
hold on
titlename2=sprintf('%s mapped GOSlim categories' , prot_name);
title(titlename2, 'Interpreter', 'none', 'FontSize', 10 );
title(titlename2, 'Interpreter', 'none', 'FontSize', 10 );
annotation('textbox',[0.84 0.85 0.15 0.12],...
'String', prot_name_termdefs,'FontSize', 12);
set(gca, 'xtick', [])
set(gca, 'ytick', [])
hold off
print('-dpng', filename);
```

```
%creates unique filename--adds date and time to name--processes to remove
%"- and :" to create a valid filename
filename2=sprintf('%s_zscore_gotermstats_%s.txt', prot_name, datestr(now));
filename2=(strrep(filename2,'-','_'));
filename2=(strrep(filename2,':','_'));
filename2=(strrep(filename2,' ','_'));

%Print text file containing the GO term, actual counts, and frequencies
fid=fopen(filename2, 'W');
fprintf(fid,
'GO_Term\tGO_Term_Def\tAll_Actual_Counts\tAll_Frequencies\tMax_Actual_Counts\
tMax_All_Frequencies\tAverages\tStD\tZscore\tP_values\n');
for i=1:length(sorted_terms)
 fprintf(fid,'%s\t%s\t%d\t%f\t%d\t%f\t%d\t%f\t%f\t%f\n', sorted_terms{i},
sorted_term_defs{i}, sorted_gocounts_mat(i,1), sorted_gocount_freq_mat(i,1),
sorted_gocounts_mat(i,3),sorted_gocount_freq_mat(i,3),sorted_goterm_avg(i),sorted_go
term_std(i),sorted_zscores(i), sorted_pvalues(i));
end
fclose(fid);

end
```