2017

IN SEARCH OF A FUNCTION FOR AN UNCHARACTERIZED CONSERVED PROTEIN IN Streptococcus sanguinis SK36

Ayana Scott-Elliston
Virginia Commonwealth University

IN SEARCH OF A FUNCTION FOR AN UNCHARACTERIZED CONSERVED PROTEIN IN *Streptococcus sanguinis* SK36

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

By

Ayana Scott-Elliston

Biochemistry B.S., University of North Carolina at Greensboro, 2011

Principal Investigator: Dr. Peter Uetz

Associate Professor in the Center for Biological Complexity

Virginia Commonwealth University

Richmond, VA

May 2017

Acknowledgements

I have the honor to give thanks to all of the individuals who aided in the completion of this thesis. I owe a debt of gratitude to my advisor, Peter, who let me enter his lab, trusted me with a project, and supported me through my journey. I would also like to show my gratitude to all my colleagues in the lab and our visiting colleagues: Harry, Neha, Jitender, Shari, Chris, and Monica. From each of my lab members, I learned a great deal about various experimental methods and science in general. I especially wish to thank Harry for the additional guidance, moral support, and aid he gave me in handling Bioinformatics, which I had no background before entering Peter's lab. I would also like to thank Dr. Kitten and Dr. Ping, for allowing me to use their equipment and mutants, and for always being available to answer my questions about *Streptococcus sanguinis* and the oral microbiome.

I would also like to give thanks to my family and my boyfriend without whom I would have been unable to complete the degree. They encourage me through the rough times, allowed me to stay at our family home while I pursued this degree. They were my biggest cheerleaders when I myself felt as if I could not continue, and I am grateful and blessed to have them.

Finally, I would like to thank Virginia Commonwealth University, and the wonderful staff and faculty whom I have met such as Sonya Washington, Cynthia Cornelissen, Dave Brohawn, Katy Beth and Debbie Roberts that have helped me to navigate this degree and present me with all options.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

%	percentage
~	approximately
AA	Amino acids
BHI	Brain Heart Infusion
B. subtilis	*Bacillus subtilis*
COG	Clusters of orthologous groups
dH_2O	deionized water
E. coli	*Escherichia coli*
ENOG	Evolutionary non-supervised orthologous groups
g	gram
H_2O	water
hr	hour
L	liter
mg	milligram
min	minute
mL	milliliter
ng	nanogram
NGS	Next Generation Sequencing
OD	optical density
PCR	polymerase chain reaction
recA	recombination protein A
recF	recombination protein F

ss	single stranded
S. mutans	*Streptococcus mutans*
S. sanguinis	*Streptococcus sanguinis*
sec	second
TE Buffer	Tris-EDTA buffer
µg	microgram
µL	microliter
WT	wildtype
wt/vol	weight/volume

ABSTRACT

PHENOTYPIC SCREENING AND CHARACTERIZATION OF PROTEINS WITH UNKNOWN FUNCTION IN *Streptococcus sanguinis* SK36

By Ayana Scott-Elliston, B.S.

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

Virginia Commonwealth University

Richmond, VA

May 2017

Principal Investigator: Dr. Peter Uetz

Associate Professor in the Center for Biological Complexity

With the number of fully sequenced bacterial genomes increasing in the past 7 years, it has been discovered that a large percentage of the putative protein coding genes have no known function. This lack of knowledge leaves scientists with an incomplete understanding of bacteria. In this study, conserved hypothetical protein mutants from *Streptococcus sanguinis* SK36 were screened on solid media with various environmental conditions. From these screens, the

candidate protein, SSA_2372, displayed a sensitivity to acidic conditions. Its homolog in *Bacillus subtilis* 168, BSU00030, also displayed a sensitivity to pH conditions at its acid tolerance extremes unlike its other homolog in *Escherichia coli*, YbcJ. When the growth rate and cell yield was acquired, the sensitivity was shown to be significant for both SSA_2372 and BSU00030 mutants. Through data mining, it was determined that Firmicutes in this homolog family COG2501 may function as a regulator for recombination protein F.

Chapter 1

Introduction

I. Homology

The definition of homology is a heavily debated subject [1]. Homology has evolved with the increase in genomic technology and computer advancements. Its origins involved scientists such as Darwin and Owen, and these scientists defined homology based upon physical similarity between anatomical appendages and organs of different species. Currently, its study determines associations between organs and molecular machines based on genetics. A general working definition of homology that I will be using is the similarity of structure and function between species based on common descent.

Today, homology is largely interwoven with phylogeny, which is the study of relationships in evolution. With this intersection of themes, new terminology was necessary; some of these terms are orthologs, paralogs, and xenologs. Orthologs are homologs that are a result of speciation and maintain the same function [2]. They may not have the same nucleic acid and amino acid sequence, but the similarity exists in its final structure and general function. Paralogs and xenologs are interrelated. When homology is the result of gene duplication and the two genes evolved together within an organism, this homolog is termed a paralog [2]. A xenolog is also a duplication of a gene; yet, the duplicate gene has been proven to be a gene transferred by horizontal or lateral gene transfer (HGT) from a different species or from a subpopulation of the same species [2].

The nineteenth-century morphologist Richard Owens defined homologs as "the same organ in different animals under every variety of form and function" [2]. Genetic homology utilizes sequence, structure, and function data from previous studies to predict potential

homologs between species. Owen's concept of homology can be applied to bacteria, and it will be the utilized in this study. Specifically, I will apply the overall theme of homology based on Owen, but I will use the techniques of genetic homology to find connections between proteins [1].

II. Uncharacterized Proteins

The number of fully sequenced bacterial genomes has increased exponentially in the past 7 years, but many of the putative proteins in these bacterial species have no known function [3]. Since Next Generation Sequencing (NGS) was developed, the number of fully sequenced genomes has increased exponentially, and we have reaped benefit from this knowledge. For characterized proteins, we know where in the genome they are located, which is advantageous for drug targeting studies of pathogenic bacteria and for the study of human-microbe interactions in health. For example, knowing the full sequence of prokaryotic genomes has allowed for techniques such as CRISPR to be found and developed, which has allowed for targeted mutagenesis and targeted disease treatment development.

Unfortunately, 25% to 60% of these proteomes are composed of proteins with unknown function, also known as hypothetical or uncharacterized proteins [3]. Hypothetical proteins are proteins that are predicted to be expressed from an open reading frame, but for which there is no experimental evidence of translation from mRNA [4]. Based on computational analysis, these hypothetical proteins often fit into important functional classes, such as hydrolases, transferases, transporters, DNA binding proteins, and RNA binding proteins [3]. Having such a large proportion of proteins with unknown function creates a gap in our understanding of these organisms as biological systems alone and within a host system. Therefore, acquiring this

knowledge could help to create more efficient treatments and to prevent unnecessary secondary reactions that we observe in current treatments by understanding the protein pathways fully [5].

Among hypothetical proteins, there are proteins conserved among multiple organisms from several phyla [4]. These proteins are called conserved hypothetical proteins and have no functional validation. As of October 2014, GenBank labeled 48,591,211 hypothetical sequences in NCBI, of which 34,064,553 were in bacteria [4]. Elucidating the function of one of these proteins will not only annotate its own protein/gene dataset and supplement our knowledge of the specific bacterium, but it will also provide functional information about its homologs.

III. DNA

A. DNA Structure and DNA Damage

Deoxyribonucleic acid (DNA) serves as the genetic blueprint for all living creatures from large mammals such as *homo sapiens* to microscopic bacteria. DNA is composed of two antiparallel strands of covalently linked nucleotides [6]. Nucleotides have a nucleobase (adenine, thymine, guanine, or cystosine), a monosaccharide, and a phosphate group, which forms the covalent bond that establishes the sugar-phosphate backbone of DNA. The antiparallel strands of DNA associate with each other by hydrogen bonds between the nucleobases. The phosphodiester bond is one of the more kinetically stable bonds found in biological functional groups [7]. DNA itself is stable at pH 7, and it can maintain its association in slightly alkaline conditions (pH 8 to 9). The presence of sodium cations helps to increase the stability of DNA by preventing the large negative charges of nucleophilic phosphate groups from destabilizing the DNA.

Unfortunately, DNA has a susceptibility to drastic pH change, UV damage, and chemical damage. UV and chemical damage cause double stranded breaks [8]. At low pH, the

phosphodiester bonds of the DNA's sugar phosphate backbone hydrolyze, and an increase in pH above 9 causes the hydrolysis of the hydrogen bonds between the two antiparallel DNA strands [9]. Fortunately, bacteria have developed repair mechanisms to combat the effects of these environmental stressors.

B. **DNA Repair**

To combat environmental factors that damage DNA, such as UV damage and pH change, the bacterial and eukaryotic cell has developed pathways to correct double stranded DNA damage. Double stranded breaks are targets for exonucleases, which can lead to gene deletions, to incorrect DNA ends connecting, and to chromosomal rearrangements that are commonly present in tumors [8]. Bacteria, as well as eukaryotes, have developed two general pathways for repair of double stranded DNA damage, which are non homologous end joining and homologous recombination. Non homologous end joining (NHEJ) is the reconnection of the broken DNA ends by the utilization of DNA ligase and of Ku [8]. Ku is a conserved protein that binds the ends of the broken DNA strands to protect it from exonuclease degradation, and Ku recruits DNA ligase to the break site to reform the bonds [8]. The second general pathway for remedying double stranded DNA damage, homologous recombination, is performed for more difficult breaks and has multiple stages that use multiple proteins in each stage.

Homologous recombination is composed of three stages: presynapsis, synapsis, and postsynapsis, and the protein groups may vary depending on the bacterial phyla [10]. In presynapsis, 3' single stranded DNA ends are generated to bind with the homologous sequence on the sister chromosome [8]. The proteins used in *Escherichia coli* for presynapsis is the enzyme complex RecBCD (recombination protein BCD), which is a multifunction protein complex with multiple domains [10]. Upon exposure to the Chi sequence signal

(5'GCTGGTGG) on dsDNA, the recombination function of RecBCD is initiated, which causes the enzyme to bind to DNA ends and to generate overhangs [8,10]. Synapsis presents the 3' DNA overhang formed by RecBCD to the homologous duplex of the sister chromosome via the enzyme Recombination Protein A (RecA) [8]. RecA requires ATP to bind to each ssDNA overhang, to create a right handed helix for inserting into the homologous duplex DNA, and to arbitrate annealing [8]. RecA has to be loaded onto the ssDNA due to Single Strand Binding proteins having a higher affinity to ssDNA than RecA [8]. The loading of RecA onto the ssDNA depends on the bacterial species. However, the RecFOR/FLOR pathway is conserved in many species for that function. The RecFOR consists of proteins recombination protein F, recombination protein O, and recombination protein R. For *E.coli,* the process is handled by one of two epistatic groups, RecBCD or RecFOR[8]. *Bacillus subtilis* has five epistatic groups, α (*recFLOR* and *recN*), β (*addAB*), γ (*recP* and *recH*) , ε (*ruvA*, *recD*, and *recU*), and ζ (*recS*). During the final stage of homologous recombination, postsynapsis, RecG and RuvA helicases generate a four stranded exchange, also known as Holliday junctions [8].

IV. Acid Tolerance in Bacteria

Since DNA is essential to the function and propagation of all cells, the cell's ability to maintain the stability of its DNA is paramount. In the gut microbiome the pH can range drastically from 6.3 to 1.5, which is a million fold increase in the hydrogen proton concentration. Subsequently, the bacterial cell has developed methods to maintain its intracellular pH despite the drastic changes in its biome's pH. Two examples of the bacterial cell's ability to maintain internal pH is *E.coli* and *B. subtilis,* neutrophiles. *E.coli* can maintain an internal pH range of 7.2 to 7.8 in an extracellular pH range of 5.5 to 9, and *B. subtilis,* a neutrophile, can maintain a

constant pH of 7.4 at the pH range of 6 to 8 [11]. The protective mechanisms that allow these bacteria the capability to stabilize are distinguished into 2 categories XAR (extreme acid resistance) and ATR (acid tolerance response). ATR mechanisms are adaptive responses at nonlethal mildly acidic pH, such as pH 5.5 for *E.coli* K12, that enhances the bacteria's ability to handle severe acidic challenge as low as pH 3 [11]. Most bacteria possess one ATR based mechanism. XAR mechanisms allow cells that have not had the opportunity to acclimate to their acidic conditions survive at pH conditions less than 2.5 [11]. An environment with such an extreme high concentration of protons produces a biome that does not permit the growth of neutrophilic bacteria.

VI. Streptococcus sanguinis

The oral ecosystem is a complex biome. The temperature is maintained around 35°C - 36 °C, except in inflammation induced conditions where it increases to 39 °C [12]. The majority of the bacterial species in the mouth are facultative anaerobes or obligate anaerobes, yet there is a small population of aerobes in the oral cavity [12]. The difference between the two types of anaerobes is that facultative anaerobes can exist with or without oxygen present, where obligate anaerobes cannot exist with oxygen. These oral microflora proliferate in distinct regions of the mouth depending on their capability of handling various percentages of oxygen [12]. The most anterior portion of the mouth is exposed to ~20% O_2, and the most posterior buccal folds is ~0.3%. Additionally, the oral cavity ranges distinctly in pH over time and in different regions, which its inhabitants need to handle. A healthy mouth's pH lies between 6.5 to 6.9 at rest, and as a mouth is stimulated while eating its pH increases to between 7.0 to 7.5 [12]. The maintenance of this pH is handled on regions of the mouth by saliva [12]. An individual with poor oral health and dental caries tends to have a more acidic pH in their mouth. Normally, after the consumption

of foods high in carbohydrates, the catabolism of sugars by the bacteria will decrease the pH. However, the pH will return gradually to resting conditions [12]. The bacteria associated with good oral health can handle brief instances of high concentrations of hydrogen protons, but the bacteria associated with dental caries thrive in acidic conditions for a prolonged period of time.

One of these healthy oral microbes is *Streptococcus sanguinis*. It is a facultative anaerobe, and its pH growth range in the mouth under normal conditions is from 5.0 to 7.8. *S. sanguinis* inhibits the growth of bacteria involved in dental caries, e.g. *Streptococcus mutans* by the release of hydrogen peroxide - an end product of its metabolism [12].

VII. Research Objective

The Aims of this Study are:

1. To phenotypically screen conserved uncharacterized proteins of *Streptococcus sanguinis* SK36
2. To find a single protein without paralog that displays a phenotype to one of the environmental conditions
3. To study the protein using bioinformatics, systems biology, and molecular biology to potentially elucidate its function.

The overall objective of this study is to investigate and to characterize a conserved uncharacterized protein family to help in the effort of closing our gap in knowledge of bacterial proteins.

Chapter 2

Materials and Methods

Bacterial Culture Conditions. The *Streptococcus sanguinis* SK36 wild type and its mutants used in this study were created and provided by the Ping Xu Lab [13]. They are listed in Table 1. *S. sanguinis* SK36 and its mutants were revived from 80% glycerol freezer samples in Brain Heart Infusion Broth (BD Difco) broth under microaerophilic conditions (7 – 10% CO_2 and 6% O_2) at 37 °C in an AnaeroPack 2.5 L Rectangular Jar (Thermo Scientific). For all experiments, *S. sanguinis* SK36 and its mutants were grown in microaerophilic conditions (7 – 10% CO_2 and 6% O_2) at 37 °C.

B. subtilis 168 and *E.coli* K-12 wild type and mutants used in this study are listed in Table 2. *B. subtilis* 00030 mutant was provided by the Bacillus Genetic Stock Center, and *E.coli ybcJ* mutant and *oxaC* control strain was provided by the Osterman lab at Sanford Burnham Prebys Medical Discovery Institute. The *Bacillus subtilis* 168 WT, *B. subtilis* 168 00030 mutant, *Eschericihia coli* K-12 WT, and *E. coli* K-12 *ybcJ* mutant were revived in LB broth in aerobic conditions with 1 ug/mL erythromycin (*B. subtilis)* and 1 ug/mL kanamycin (*E.coli*).

Broth Preparation: Brain Heart Infusion. 37 grams of BHI powder was dissolved in 1000mL deionized H_2O using a stir bar and heating at 100 °C on a hot plate. It was then autoclaved for 15 minutes at 121°C. The sealed autoclaved broth was stored at 4 °C when not in use.

Broth Preparation: pH specific BHI media. 37g of BHI powder was dissolved in 1000mL deionized H_2O using a stir bar and heating at 100 °C. The dissolved BHI broth was adjusted to the following pH values depending on the experiment: 4, 5, 5.5, 6, 6.5, 7, 8, & 10. The pH adjustments to create the acidic pH values were achieved by pipetting 12.1 M hydrochloric acid until the broth reached the appropriate value. The alkaline pH values were reached by pipetting volumes of 0.0316M NaOH (pH 12.5) into the BHI broth and mixing after each addition. When the BHI was at the appropriate value, the Erlenmeyer flasks were autoclaved for 15 minutes at 121°C.

Agar Preparation: pH specific BHI plates. 52g of BHI agar powder was dissolved in 1000mL deionized H_2O using a stir bar and heating at 100 °C. The dissolved BHI broth was adjusted to the following pH values depending on the experiment: 4, 5, 5.5, 6, 6.5, 7, 8, & 10. The pH adjustments to create the acidic pH values was achieved by pipetting 12.1 M hydrochloric acid until the broth reached the appropriate value. The alkaline pH values were reached by pipetting volumes of 0.0316M NaOH (pH 12.5) into the BHI broth and mixing after each addition. When the BHI was at the appropriate value, the Erlenmeyer flasks were autoclaved for 15 minutes at 121°C. When the Erlenmeyer flasks reached 70 °C, the molten agar was poured into omnitrays. Once the plate solidified, it was stored in sterile containers in a cold room until used.

Agar Preparation: Chemically Defined Media with Various Sugars. All of the values of the Chemically Defined Media used in Ping Xu's study had to be doubled because it would be diluted in half when the agar was introduced [13]. For every 1000 mL deionized H2O, 20g of selected sugar, 4.0 g of L-glutamic acid, 0.4 g of L-cysteine, 1.8 g of L-leucine, 2.0 g of NH_4Cl,

5.0 g of K_2HPO_4, 5.0 g of KH_2PO_4, 8.0 g of $NaHCO_3$, 2.4 g of $MgSO_4.7H_2O$, 0.04 g of $MnCl_2.4H_2O$, 0.04 g of $FeSO_4.7H_2O$, 1.2 g of sodium pyruvate, 2.0 mg of riboflavin, 1.0 mg of thiamine HCl, 0.2 mg of D-biotin, 2.0 mg of nicotinic acid, 0.2 mg of p-aminobenzoic acid, 1.0 mg of Ca-pantothenate, 2.0 mg of pyridoxal HCl, 0.2 mg of folic acid, and adjusted to pH 7.0 with H3PO4. In order to add the sugar, riboflavin, thiamine HCl, D-biotin, nicotinic acid, p-aminobenzoic acid, Ca-pantothenate, pyridoxal HCl, and folic acid, a stock solution of each was created with dH_2O. Then, the appropriate volume from each stock solution was added to the 1000 mL solution. Once all of chemicals were added, the solution was filter sterilized with Stericup filter bottle Units (Millipore). The selected sugars were glucose, mannose, galactose, and lactose. The agar was made by adding 15g of agar to 1000mL dH_2O, which was autoclaved at 121 °C for 15 minutes. The agar was cooled in a hot water bath to 55 °C, which is lower than the 60 °C degradation temp for the vitamins and the minimum temperature to maintain the liquid phase of agar without bubbles.

Phenotype Screen: Solid Phase. Following the revival of the strains in a 96 well microtiter plate, these mutants were pinned to the respective agar plate using a pin tool attached to the Biomek FX robot (Beckman Coulter). The motion of the pin tool followed the respective program I designed, and between pinning a new plate, it always included a four-step wash process – a dH_2O rinse, a 10 second ethanol soak, a 30 second 10% bleach soak, and a final dH_2O rinse. These plates were incubated for 48 hours in an Anoxomat jar (Spiral Biotech) in which the microaerophilic conditions were generated by the Anoxomat system (Mart).

Quantification of Phenotype: Liquid Phase. 200μL of the BHI broth of a specific pH value was added to the appropriate column in a black 96 well plate. For the *S. sanguinis* SK36 mutant phenotype quantification, the wells were inoculated from an overnight growth by the pintool of the Biomek FX robot. For the *E.coli* and *B. subtilis* strains phenotype quantification, the wells were inoculated with 1 μL from an overnight growth. For both quantification studies, the source plate mutants had an OD of 0.9 to 1.0 at 600 nm. The inoculum was grown overnight at 37 °C with atmospheric conditions of 6% O_2 and 6% CO_2 in the BioTek Synergy H1 Hybrid Multi-Mode Monochromator Fluorescence Microplate Reader (BioTek). Readings were taken every ten minutes at 600nm.

Bioinformatics. The sequence analysis and homolog search was performed using NCBI Protein Blast non redundant protein sequences database; homolog determination was also performed via Eggnog v 4.5.1 and UNIPROT [14,27]. The comparative multiple sequence alignment of the COG2501 homologs were performed with ClustalOmega version 1.2.1 [15]. The visualization of the phylogenetic tree data was created using the interactive Tree of Life (iTOL) version 3.5, and the code utilized to separate the species from EGGNOG into families with stars was created by Harry Caufield [16]. Protein abundance data was extracted on COG2501 from PaxDB version 4.0 [17]. The source of the data was either presented in the file extracted from the website, or it was acquired by contacting the PaxDB curators [17]. Hypothetical COG2501 structures and secondary structure information was calculated by I-Tasser version 5.1 and RaptorX [18,19, 20, 21]. The NCBI taxonomy database was utilized for information on phyla [22], and the Microbial Genome Context Viewer (MGcV) was used to visualize the genome context for all proteins in this study [29].

Statistics, Growth Rate, & Doubling Time Calculations.

All statistical tests were performed utilizing R and R Studio [23,24]. The necessary packages for these statistical tests were stats version 3.3.2, methods version 3.3.2, datasets version 3.3.2, and utils version 3.3.2. The R code used is in the Supplemental (Supplemental Table 2). Exponential growth rate is a first order reaction, and the equation used to calculate the growth rates and doubling times are below. In the growth rate equation, t represents time, and c represents the optical density obtained at 600nm.

$$K = \text{growth rate} = \frac{\ln (C_f/C_i)}{(t_f - t_i)}$$

$$t_{2/1} = \text{doubling time} = \frac{\ln (2)}{k}$$

Table 1: *S. sanguinis* strain and its mutants without paralogs used in this study

Conserved hypothetical protein mutants	Mutation	Species Count
SK36 wildtype		
SSX_0064	SSA_0064 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	157
SSX_0066	SSA_0066 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	217
SSX_0073	SSA_0073 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	280
SSX_0101	SSA_0101 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	530
SSX_0152	SSA_0152 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	7
SSX_0161	SSA_0161 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	15
SSX_0183	SSA_0183 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	67
SSX_0188	SSA_0188 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	582
SSX_0224	SSA_0224 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	23
SSX_0238	SSA_0238 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	66
SSX_0252	SSA_0252 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	26
SSX_0273	SSA_0273 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	9
SSX_0274	SSA_0274 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	120
SSX_0281	SSA_0281 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	13
SSX_0288	SSA_0288 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	2
SSX_0293	SSA_0293 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	1073
SSX_0299	SSA_0299 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	376
SSX_0305	SSA_0305 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	77
SSX_0311	SSA_0311 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	186
SSX_0316	SSA_0316 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	917
SSX_0339	SSA_0339 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	284
SSX_0341	SSA_0341 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	507
SSX_0353	SSA_0353 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	74

SSX_0354	SSA_0354 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	62
SSX_0390	SSA_0390 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	142
SSX_0431	SSA_0431 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	256
SSX_0446	SSA_0446 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	117
SSX_0505	SSA_0505 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	70
SSX_0591	SSA_0591 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	312
SSX_0595	SSA_0595 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	540
SSX_0624	SSA_0624 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	1499
SSX_0627	SSA_0627 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	182
SSX_0629	SSA_0629 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	239
SSX_0646	SSA_0646 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	96
SSX_0647	SSA_0647 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	157
SSX_0659	SSA_0659 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	438
SSX_0664	SSA_0664 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	79
SSX_0668	SSA_0668 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	205
SSX_0670	SSA_0670 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	89
SSX_0672	SSA_0672 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	1599
SSX_0711	SSA_0711 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	411
SSX_0752	SSA_0752 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	692
SSX_0755	SSA_0755 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	121
SSX_0802	SSA_0802 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	749
SSX_0803	SSA_0803 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	130
SSX_0815	SSA_0815 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	323
SSX_0843	SSA_0843 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	664
SSX_0844	SSA_0844 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	93
SSX_0864	SSA_0864 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	464

SSX_0868	SSA_0868 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	1303
SSX_0874	SSA_0874 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	736
SSX_0885	SSA_0885 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	108
SSX_0915	SSA_0915 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	58
SSX_0978	SSA_0978 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	169
SSX_0983	SSA_0983 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	55
SSX_0985	SSA_0985 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	108
SSX_0990	SSA_0990 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	114
SSX_0992	SSA_0992 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	77
SSX_1067	SSA_1067 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	64
SSX_1076	SSA_1076 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	321
SSX_1078	SSA_1078 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	121
SSX_1083	SSA_1083 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	75
SSX_1154	SSA_1154 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	996
SSX_1156	SSA_1156 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	80
SSX_1161	SSA_1161 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	173
SSX_1180	SSA_1180 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	139
SSX_1201	SSA_1201 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	1506
SSX_1203	SSA_1203 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	200
SSX_1211	SSA_1211 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	136
SSX_1214	SSA_1214 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	109
SSX_1238	SSA_1238 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	95
SSX_1247	SSA_1247 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	80
SSX_1248	SSA_1248 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	74
SSX_1249	SSA_1249 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	95
SSX_1280	SSA_1280 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	341

SSX_1313	SSA_1313 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	191
SSX_1348	SSA_1348 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	78
SSX_1353	SSA_1353 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	303
SSX_1384	SSA_1384 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	78
SSX_1404	SSA_1404 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	59
SSX_1418	SSA_1418 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	495
SSX_1439	SSA_1439 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	932
SSX_1454	SSA_1454 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	60
SSX_1455	SSA_1455 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	215
SSX_1471	SSA_1471 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	729
SSX_1480	SSA_1480 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	256
SSX_1526	SSA_1526 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	360
SSX_1535	SSA_1535 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	66
SSX_1537	SSA_1537 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	177
SSX_1540	SSA_1540 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	67
SSX_1545	SSA_1545 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	112
SSX_1549	SSA_1549 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	302
SSX_1554	SSA_1554 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	83
SSX_1562	SSA_1562 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	552
SSX_1585	SSA_1585 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	343
SSX_1586	SSA_1586 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	327
SSX_1587	SSA_1587 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	338
SSX_1610	SSA_1610 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	149
SSX_1618	SSA_1618 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	594
SSX_1627	SSA_1627 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	302
SSX_1646	SSA_1646 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	226

SSX_1651	SSA_1651 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	69
SSX_1686	SSA_1686 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	129
SSX_1704	SSA_1704 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	137
SSX_1707	SSA_1707 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	75
SSX_1746	SSA_1746 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	51
SSX_1765	SSA_1765 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	94
SSX_1770	SSA_1770 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	593
SSX_1772	SSA_1772 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	169
SSX_1786	SSA_1786 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	207
SSX_1808	SSA_1808 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	83
SSX_1817	SSA_1817 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	98
SSX_1833	SSA_1833 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	47
SSX_1837	SSA_1837 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	1126
SSX_1844	SSA_1844 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	98
SSX_1855	SSA_1855 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	998
SSX_1899	SSA_1899 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	404
SSX_1910	SSA_1910 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	346
SSX_1911	SSA_1911 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	1076
SSX_1921	SSA_1921 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	53
SSX_1941	SSA_1941 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	234
SSX_1960	SSA_1960 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	92
SSX_1981	SSA_1981 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	75
SSX_1987	SSA_1987 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	176
SSX_2000	SSA_2000 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	85
SSX_2013	SSA_2013 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	118
SSX_2021	SSA_2021 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	69

SSX_2030	SSA_2030 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	1272
SSX_2036	SSA_2036 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	241
SSX_2038	SSA_2038 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	150
SSX_2076	SSA_2076 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	139
SSX_2126	SSA_2126 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	1653
SSX_2170	SSA_2170 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	278
SSX_2188	SSA_2188 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	361
SSX_2189	SSA_2189 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	269
SSX_2198	SSA_2198 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	52
SSX_2252	SSA_2252 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	218
SSX_2337	SSA_2337 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	78
SSX_2340	SSA_2340 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	130
SSX_2354	SSA_2354 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	283
SSX_2369	SSA_2369 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	277
SSX_2372	SSA_2372 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	184
SSX_2386	SSA_2386 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	52
SSX_2390	SSA_2390 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	377
SSX_2390	SSA_2390 replaced with *aphA*- 3 gene encoding kanamycin resistance using recombinant PCR method	1653

Table 2: *B.subtilis* & *E.coli* strains and COG2501 homologs used in this study.

Bacterial Strain	**Strain Mutant**
B. Subtilis 168	*yaaA*, BKE00030
E. coli K-12	*ybcJ*, b0528, JW5070
E. coli K-12	*uxaC*

Chapter 3

Environmental Based Screening of 149 *S. sanguinis* SK36 *mutants* & the Candidate Proteins that were Found

Identification of Candidate Protein

The qualitative rate of growth of 147 *S. sanguinis* SK36 mutants in different pH conditions (pH 5.8, pH 6, pH 7, pH 8, & pH 10) was analyzed based on the diameter of the inoculation pin on BHI agar. All of the selected mutants grew with diameters the size of the pins at pH 7, which was our pH control for growth (Figure S1). At slightly acidic pH of 6, the mutant in row D column 9, SSX_2372, had diminished growth (Figure 1). When you zoom into row D, you can see the mutants in columns 1 to 8 have full pins in which you unable to distinguish individual colonies from the pin (Figure 2). SSX_2372 was not affected by a change in carbon source, and it grew the same with glucose, mannose, lactose and galactose (Figure 3, 4, 5, 6). All colonies grew to the same size (Figure 3, 4, 5, 6).

To get a clearer visual comparison of the two candidate proteins SSA_2372 & SSA_1480, the two candidate proteins were pinned to agar alongside the wildtype and two pH control mutants, SSX_0204 and SSX_1507 (Figure 7). These pH control mutants were identified in previous studies by the Xu lab. SSX_0204 has an acid resistant phenotype, and SSX_1507 is acid sensitive. At pH 7, SSX_1507 has a diminished diameter of its inoculation pin, and its individual colonies are visible (Figure 7). SSX_1507 has no growth at pH 7 (Figure 7). These mutants were discovered to have these phenotypes in Dr. Ping Xu's lab [13]. When compared to *S. sanguinis* SK36 pH control mutants, SSX_2372 has pronounced pH phenotype (Figure 7). SSX_2372 at pH 6 looks like SSX_1507 does at pH 7 (Figure 7). You can see individual colonies, and the diameter of the inoculation pin is smaller (Figure 7).

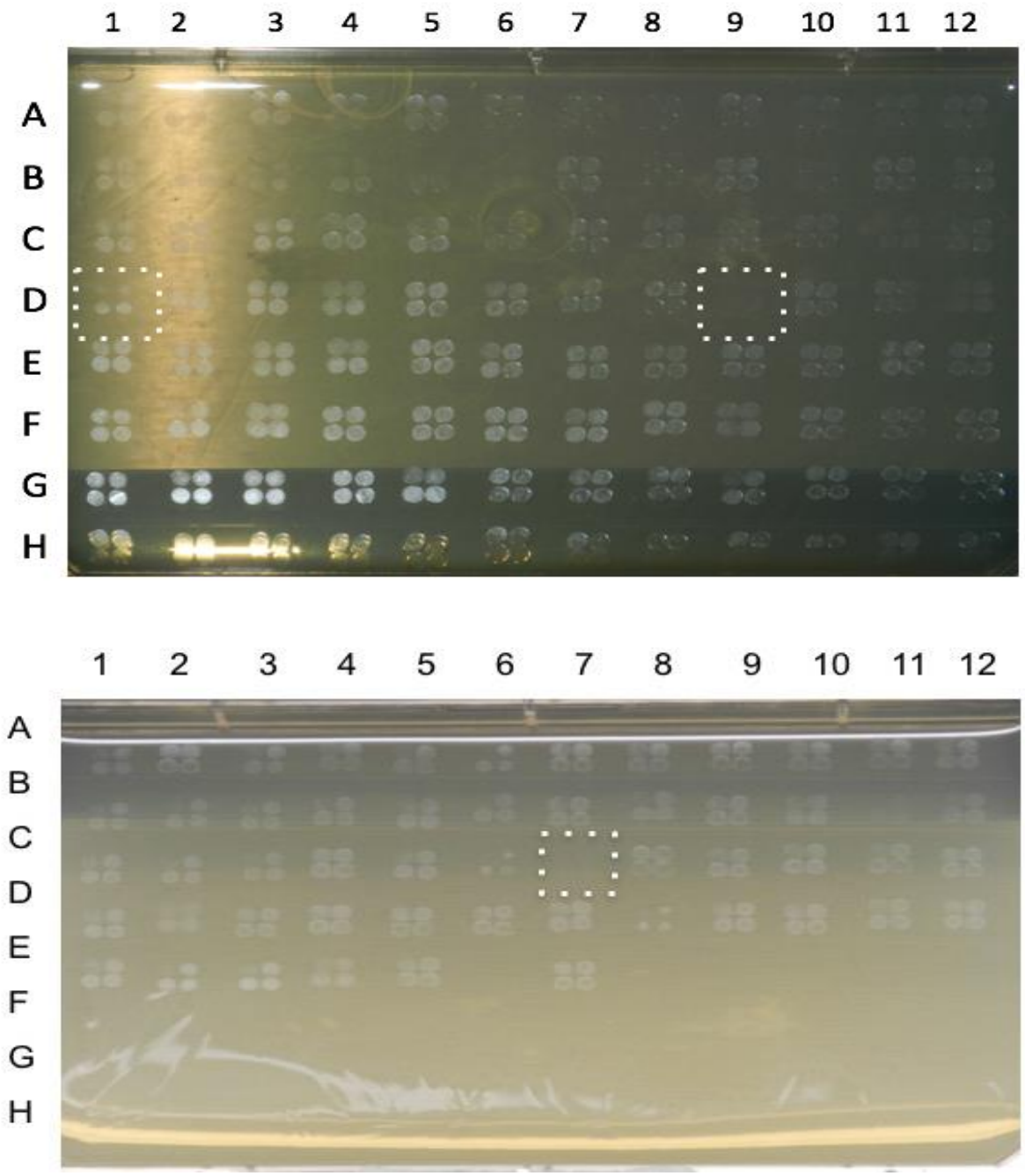


Figure 1. BHI Agar Plate (pH6) Pinned with *Streptococcus sanguinis* SK36 mutants. On these two plates, 95 mutants were pinned in quadruplicate, and the plate was incubated at 37 °C under microaerophilic conditions for 48 hours. The dotted box indicates the wells with extreme diminished growth.

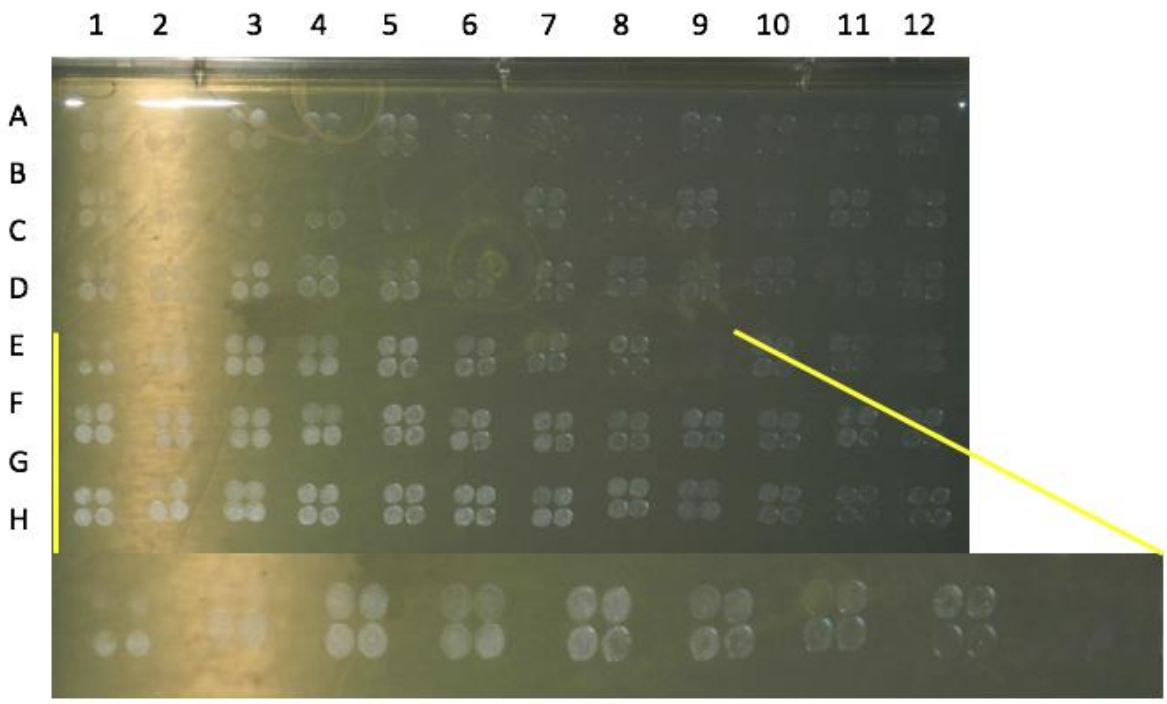


Figure 2. BHI Agar Plate (pH 6) Pinned with *Streptococcus sanguinis* SK36 mutants Zoom Into Candidate Protein, SSA_2372. On these two plates, 95 mutants were pinned in quadruplicate, and the plate was incubated at 37 °C under microaerophilic conditions for 48 hours. Row D was enlarged to make the well, D9, that displayed the acid sensitivity clear.

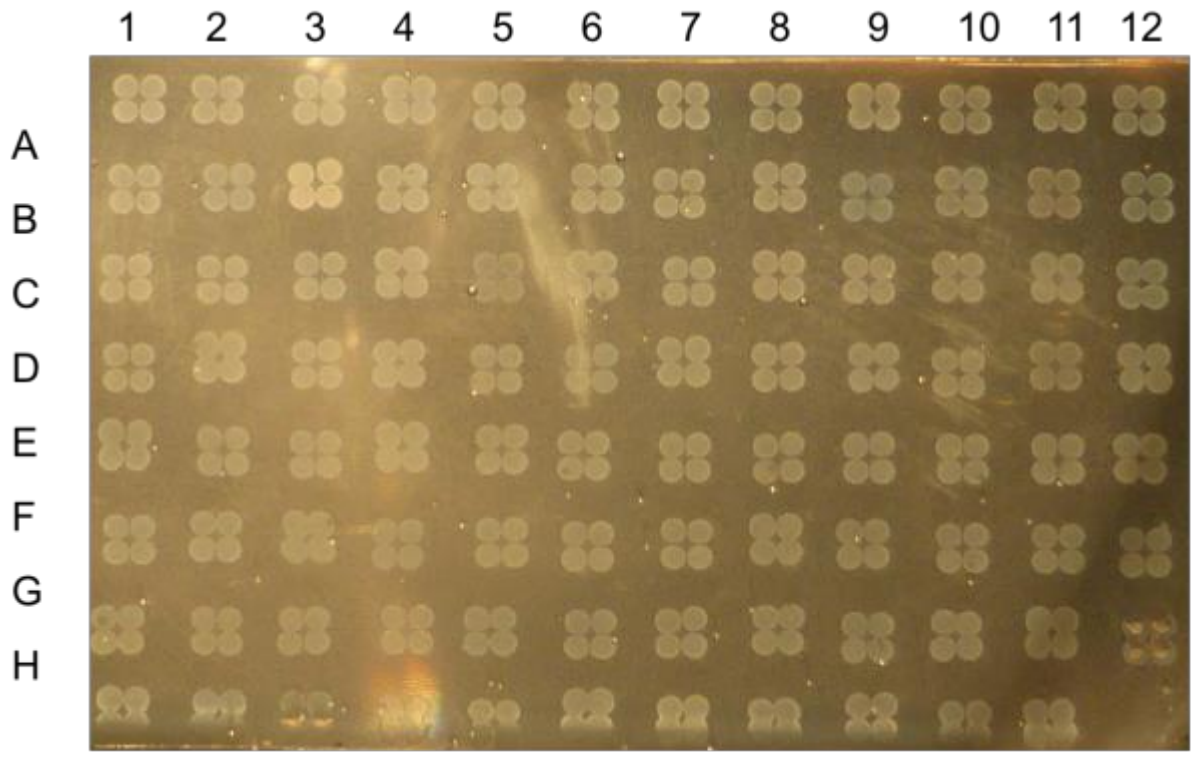


Figure 3. Chemically Defined Agar with Glucose. This plate contains the minimal amount of nutrients necessary for a *S. sanguinis* SK36 bacterium to survive. The selected carbon source for this plate is glucose. It was grown at 37 °C for 48 hours under microaerophilic conditions.

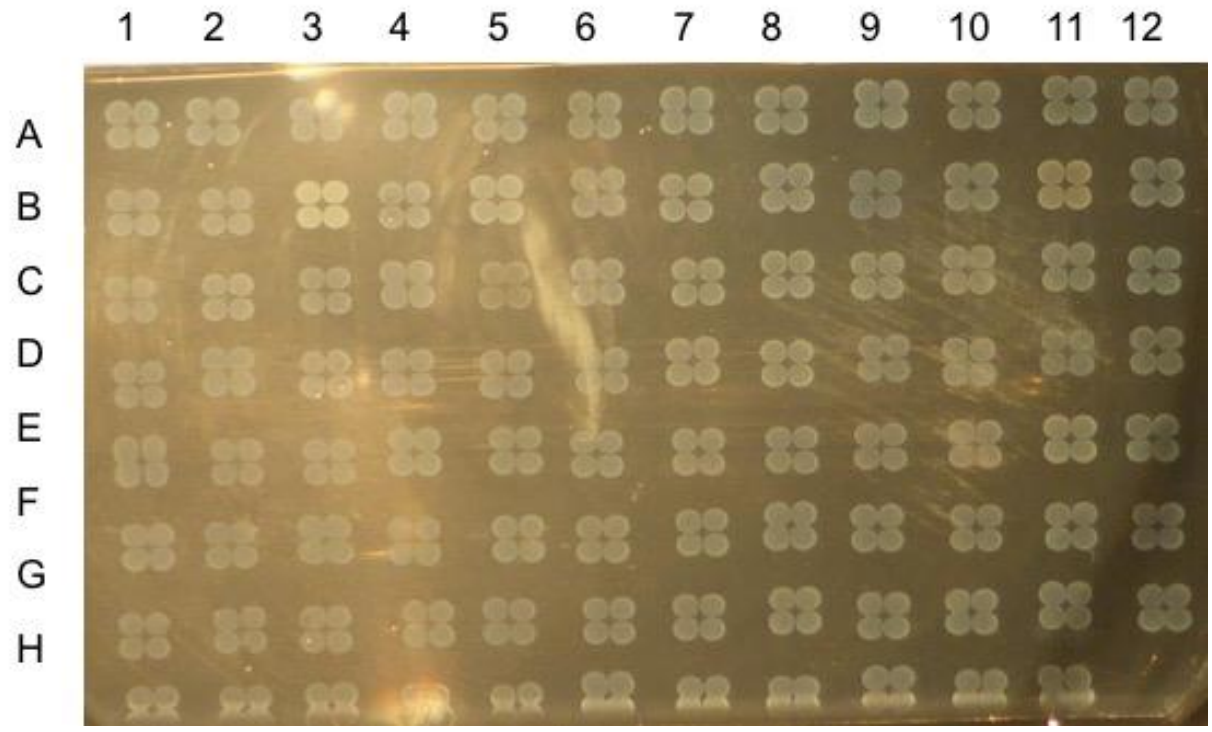


Figure 4. Chemically Defined Agar with Lactose. This plate contains the minimal amount of nutrients necessary for a *S. sanguinis* SK36 bacterium to survive. The selected carbon source for this plate is lactose. It was grown at 37 °C for 48 hours under microaerophilic conditions.

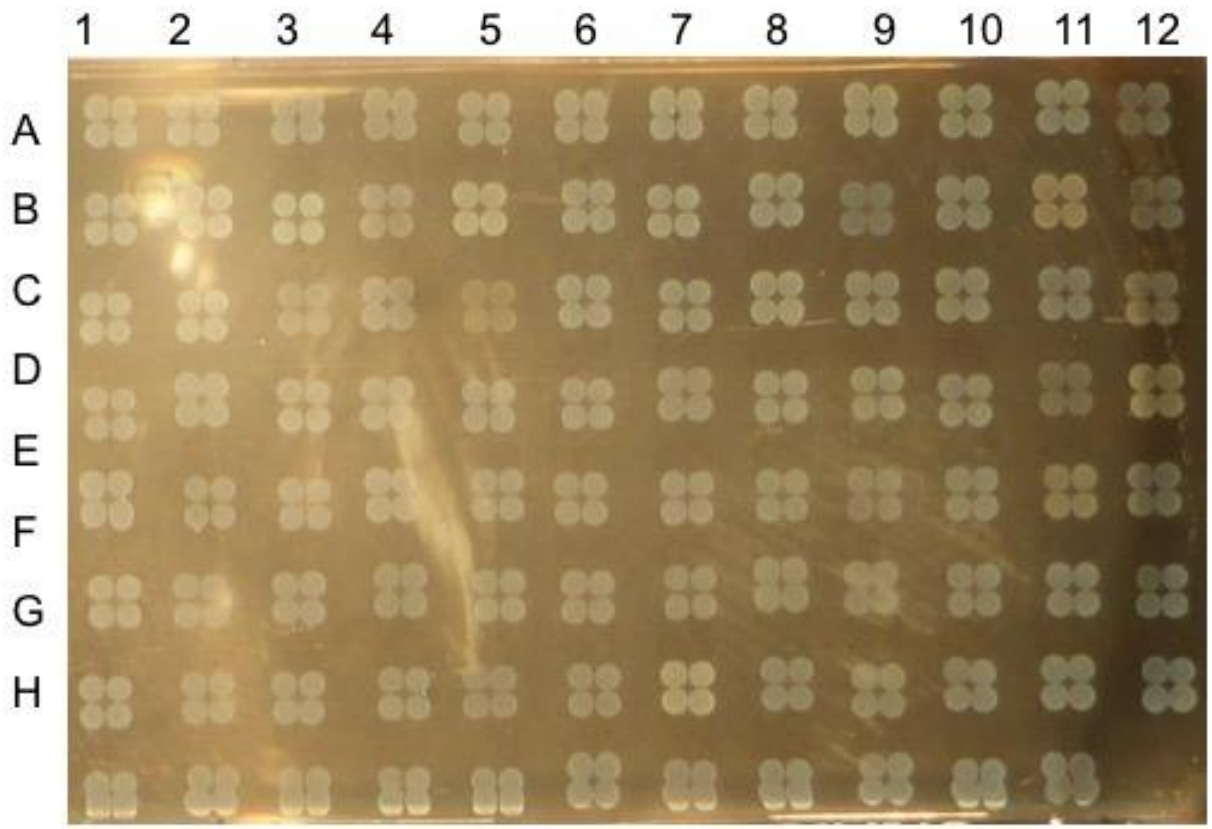


Figure 5. Chemically Defined Agar with Mannose. This plate contains the minimal amount of nutrients necessary for a *S. sanguinis* SK36 bacterium to survive. The selected carbon source for this plate is mannose. It was grown at 37 °C for 48 hours under microaerophilic conditions.

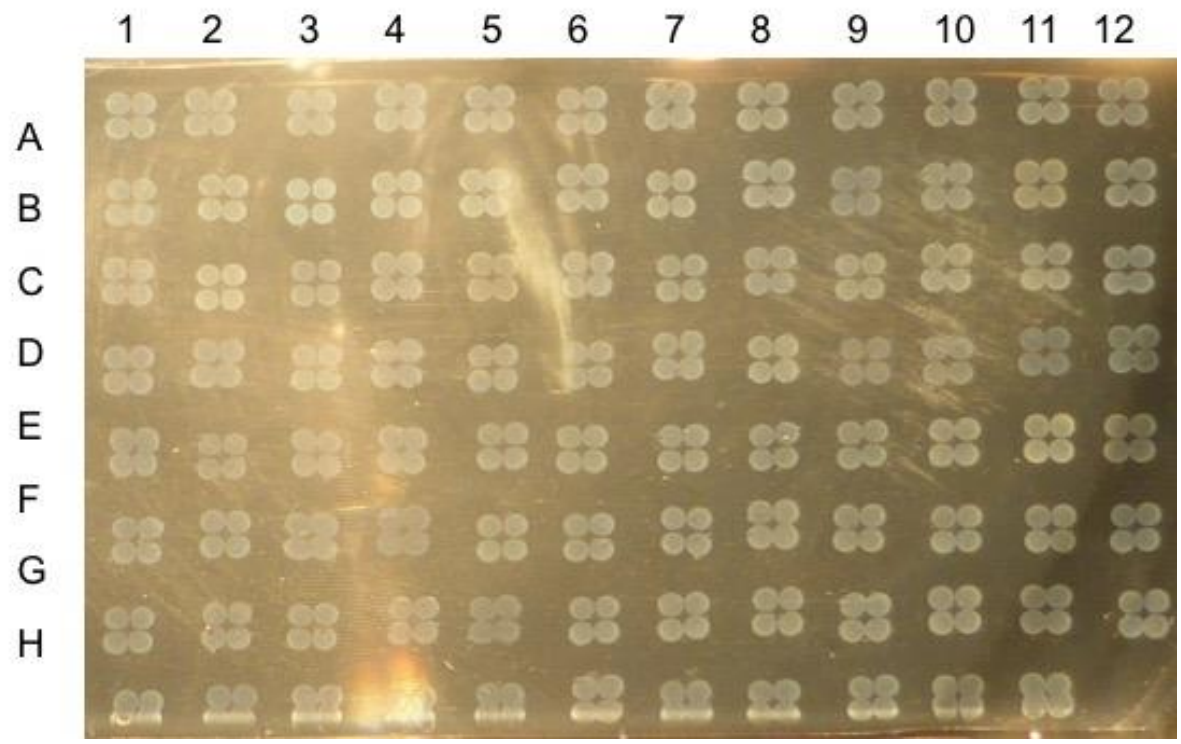


Figure 6. Chemically Defined Agar with Galactose. This plate contains the minimal amount of nutrients necessary for a *S. sanguinis* SK36 bacterium to survive. The selected carbon source for this plate is galactose. It was grown at 37 °C for 48 hours under microaerophilic conditions.

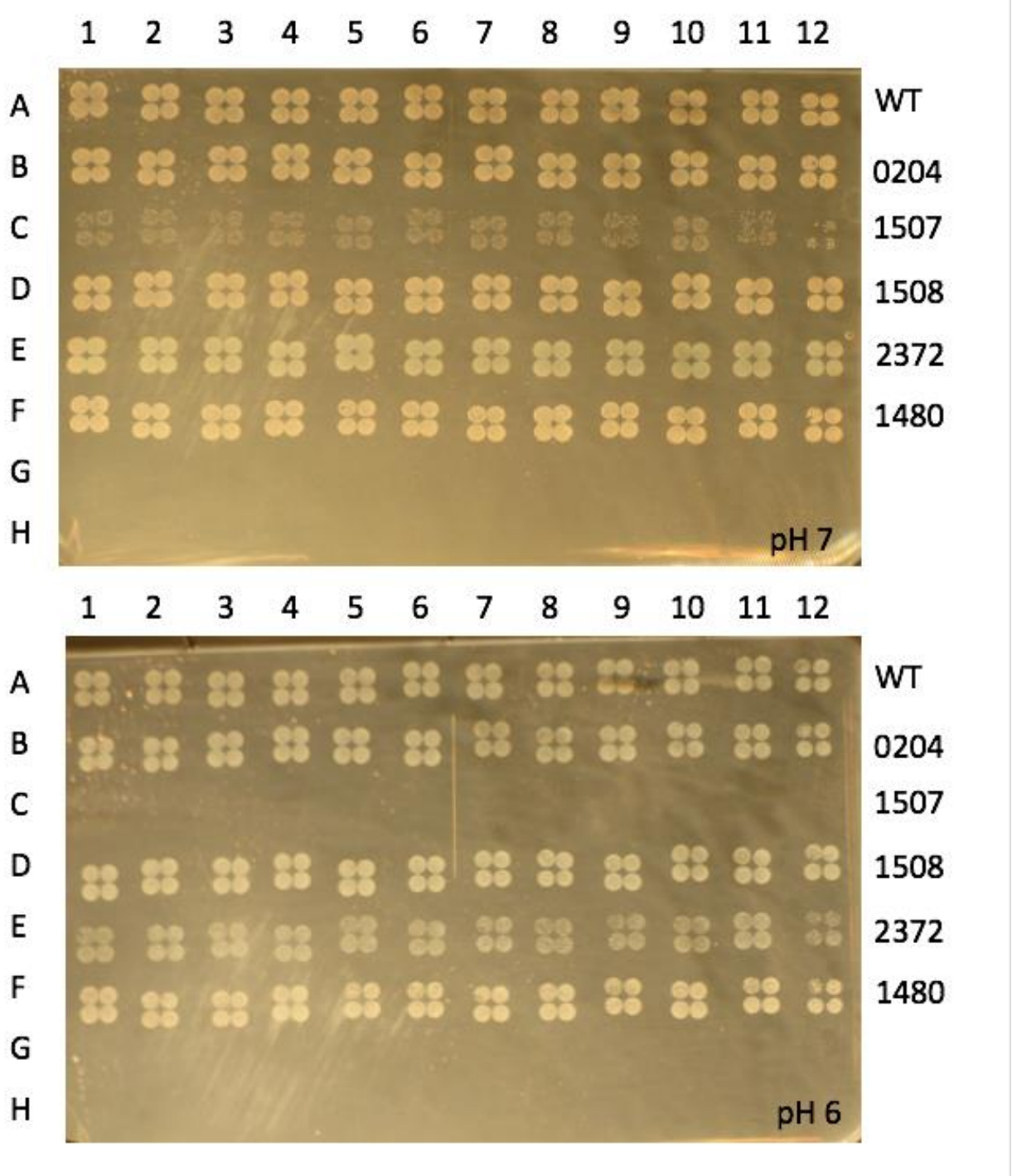


Figure 7. Qualitative Comparison at pH 6 and pH 7 of Streptococcus sanguinis SK 36 candidate mutants and control mutants on BHI Agar. The plates were pinned using a Biomek FX with a pin plate in quadruplicate for each mutant in BHI agar at pH 7 and pH 6. These plates were grown at 37 °C under microaerophilic conditions.

Quantification of Phenotype

The candidate proteins' cell yield was quantified and compared using optical density in a 96 well plate at pH 4, pH 5.8, pH 6, pH 7, pH8, and pH 10 in BHI broth (Figure 8). The doubling time of *S. sanguinis* SK36 wildtype is approximately 1 hr 11 minutes under microaerophilic conditions at pH 7, and it has a growth rate of 0.58 doublings per hour. At pH 6, *S. sanguinis* SK36 wildtype has a doubling time of 3 hours and 10 minutes. The acid sensitive control mutant, SSX_1507, grew the least at pH 7 as expected based on its growth on the agar, and it had negligible growth at pH 6 (Figure 8). The doubling time of SSX_1507 at pH 7 is approximately 1 hour 41 minutes, and its growth rate is 0.4 doublings per hour. The doubling time of SSX_1507 at pH 6 was 0.076 doublings per hour, and it had a doubling time of approximately 9 hours. SSX_1507's cell yield and growth rate are significantly different at a p value of 0.01 and 0.05 at both pH 6 and pH 7 than the wildtype of *S. sanguinis* SK36, which makes it a good acid sensitive control. The candidate mutant, SSX_2372, has a cell yield that is significantly smaller compared to the wildtype at confidence level of 95% and 99% at pH 7 and pH 6. SSX_2372 has a growth rate of 0.16 doublings per hour and a doubling time of approximately 4 hours and 18 minutes.

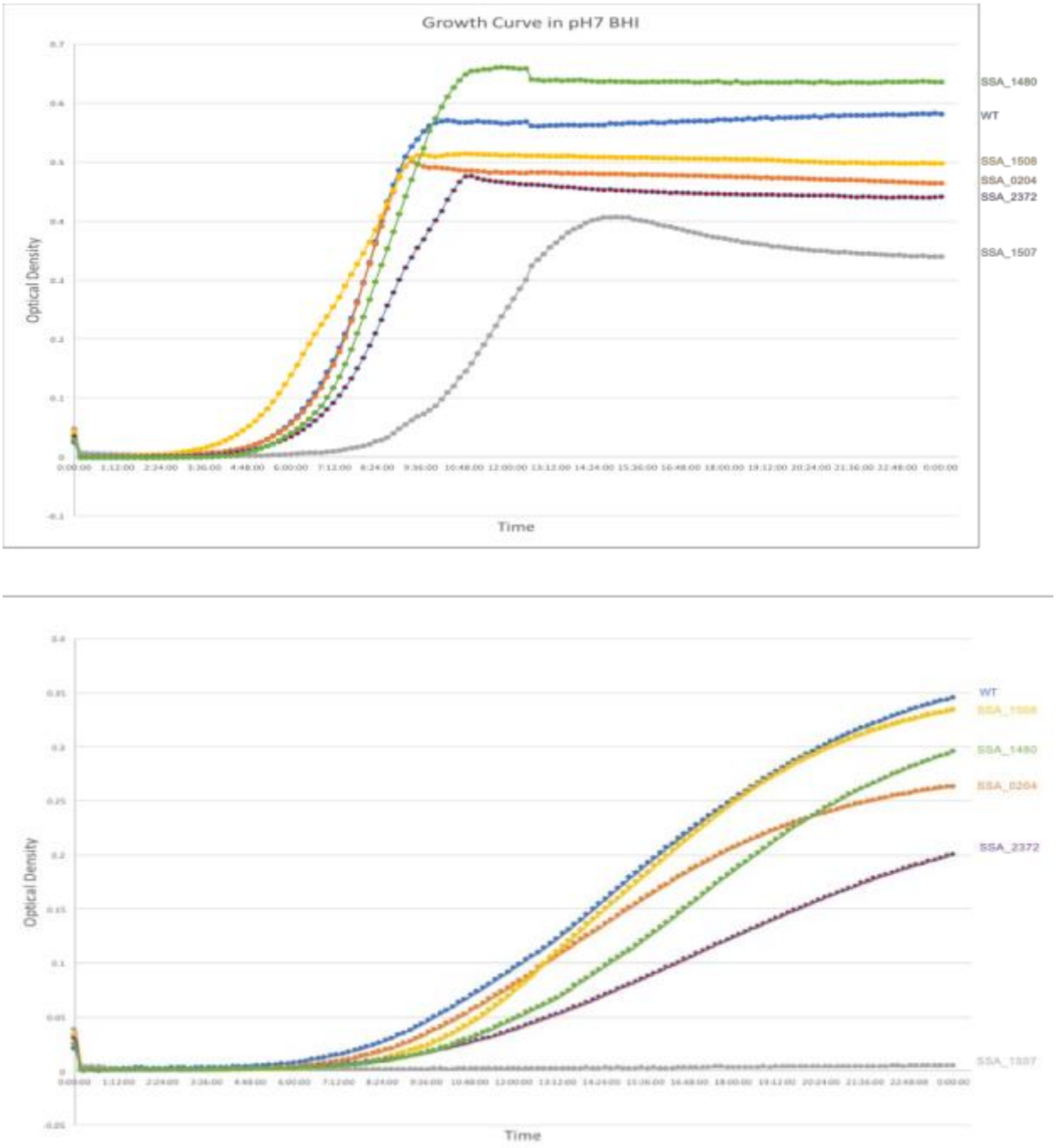


Figure 8. Growth Curves of the S. sanguinis SK36 candidate mutants and control mutants in BHI broth at pH 7 and pH 6. 96 well plates were inoculated in pH specific media, and it was grown overnight under microaerophilic conditions. Optical density readings at 600nm were recorded every 10 minutes.

Chapter 4

Bioinformatic Characterization of *S. sanguinis* SSA_2372 Protein & the Analysis of its Conserved Orthologous Group, COG2501

Conservation of COG2501

The conservation of this candidate protein was determined utilizing bioinformatics databases. COG2501 can be found in 776 species of bacteria (16). Homologs of SSA_2372 can be found in 5 major phyla and 2 superphyla (Figure 9). The phyla include Proteobacteria, Firmicutes, Actinobacteria, Cyanobacteria, and Fusobacteria, and the superphyla include PVC group and FCB Group (Figure 9). Homologs of COG2501 in the PVC superphyla can be found in Ilyobacter (order), unclassified Verrucomicrobiales (order), and Planctomycetaceae (family), and in the FCB group they are found in Flavobacteriales (order) and Sphingobacteriales (order) (Figure 9). Homologs are not found in the phyla of Chlamydiae, Thermodesulfobacteria, Acidobacteria, Thermotogae, and Spirochaetes, and Chloroflexi (Figure 9). In the phyla, homologs are found in 33 families of Firmicutes, 85 families of Proteobacteria, 26 families of Actinobacteria, and 2 families and 3 orders of Cyanobacteria.

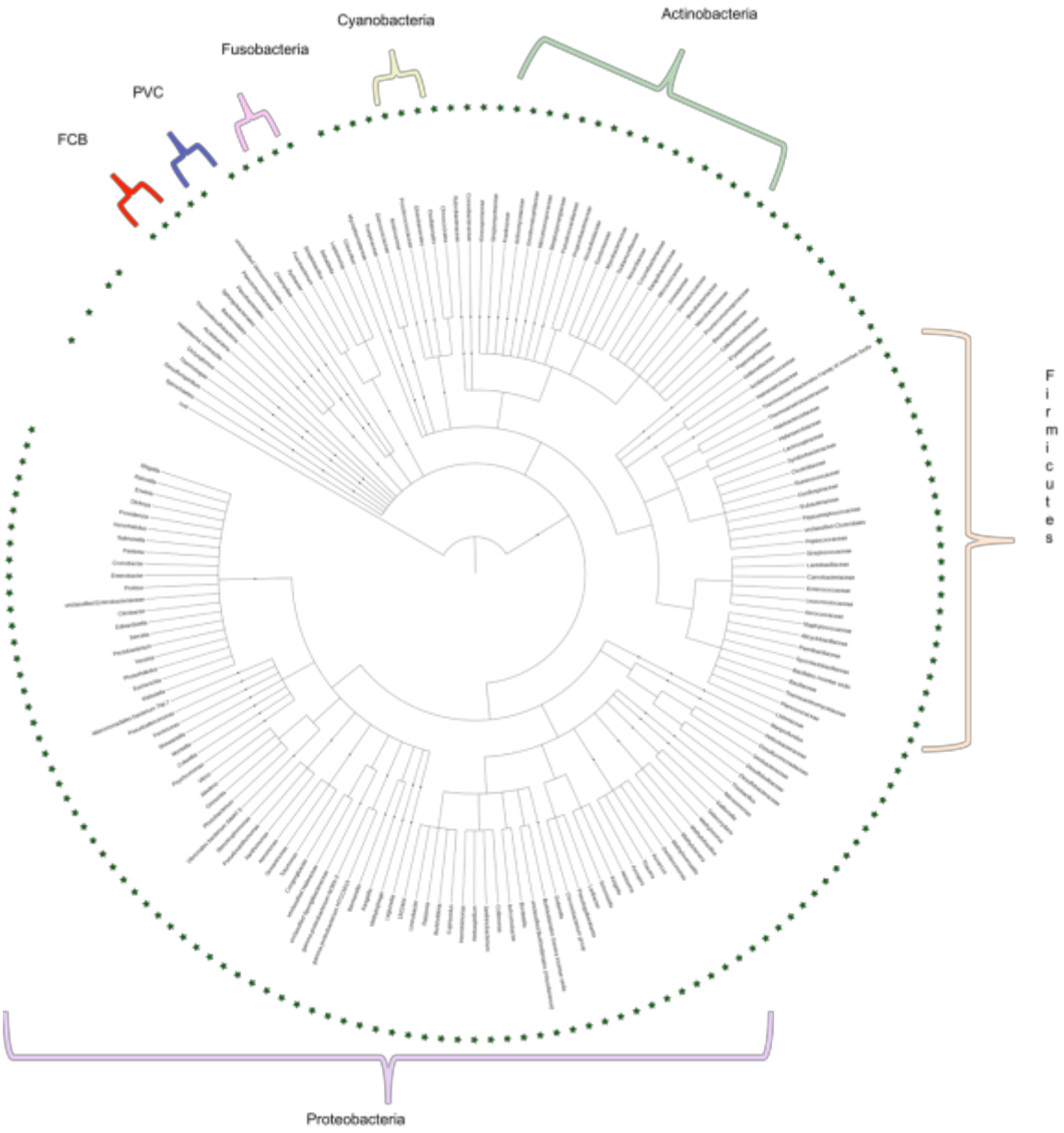


Figure 9. Phylogenetic Tree displaying the conservation of COG2501 at the phyla and superphyla level. Homolog list with NCBI identifiers was extracted from the EGGNOG database, and the dataset was inputted into iTOL v3.5 for phylogenetic tree visualization.

Functional Annotation, Expression, and Structural Prediction of SSA_2372

In depth data mining and utilization of structural genome prediction software was performed to gather functional and structural information on this gene/protein family. Based on NCBI protein blast comparisons and other various protein comparison software, SSA_2372 has been established as having an S4 domain (Table 3) [25]. It has two domains: a conserved domain, S4, and a coil-coil domain. The S4 domain is approximately 64 AA in length, and it is predicted to be comprised of 3 alpha helices and 5 beta sheets (Table 3). The S4 domain is also a part of a well conserved domain family [25]. The coil-coil domain is a non-conserved domain family, and it is predicted to be composed of alpha helices in a scissor shape. The RaptorX structure prediction, which is a software that predicts structure on sequence alone, has 3 alpha helices and 5 beta sheets (Figure 10) [20]. I-Tasser, which utilizes both sequence and crystal structures from homologs, also has 3 also has 3 alpha helices and 5 beta sheets (Figure 12). In RaptorX, the coil-coil domain, curls towards the other domain forming a pocket (Figure 11). However, in I-Tasser, the coil-coil domain extends away from the S4 domain (Figure 12).

The protein expression level of COG2501 depends on the species. The *Streptococcus pyogenes* homolog (Firmicute), SPY_2203, expresses under normal BHI conditions at 5.11 ppm and 9.69 ppm (Table 4). When the growth media is supplemented with 0.3% yeast, it expresses at a slightly higher value of 31.5 ppm (Table 6). When compared to the expression of all proteins in *S. pyogenes*, it is in the bottom 10% to 25% (Table 4). The *E.coli* homolog (Proteobacteria) YbcJ, expresses under various growth conditions in the top 25%. Some of the abundance values are 805 ppm, 617 ppm, and 352 ppm (Table 4).

Genomic Context and Sequence Comparison of COG2501

COG2501 varies both in sequence similarity and in size from a single domain protein approximately 68 AA to 80 AA in length to a two-domain protein approximately 130 to 140 AA in length (Figure 10, 13, 14). In all species of the genus *Streptococcus*, COG2501 has two domains and is approximately 130 AA in length. In all homologs in the genus Lactococcus, COG2501 also has two domains and is ~130 AA in length (Figure 10). In Bacillus homologs such as *B. subtilis*, the protein only has a single domain and is 70 AA in length, which is the S4 domain (Figure 10). *B. subtilis* has a sequence similarity of ~60% (Figure 13). For all other homologs such *E. coli* and *V. cholera*, the protein is ~70 AA in length with the single predicted S4 domain (Figure 10). *E.coli* has a sequence similarity of ~ 20% (Figure 14). The Streptococcus, Lactococcus, and Bacillus genus is in the Firmicutes phyla, and *E.coli* is in Proteobacteria [22].

Interactions of COG2501

The interactions of COG2501 family of proteins depends on the species. In *E.coli,* YbcJ interacts with the subunits of the ribosome, 30S ribosomal proteins and 50S ribosomal proteins, but not the polysome (26). This data was determined by pull down assays and immunoblot analysis [25, 28]. These ribosomal subunits include: 30S ribosomal protein S6, 30S ribosomal protein S2, 30S ribosomal protein S3, 50S ribosomal protein L1, 50S ribosomal protein L7/L12, 50S ribosomal protein L9, 50S ribosomal protein L11, 50S ribosomal protein L19, 50S ribosomal protein L20, and 50S ribosomal protein L33 (Table 4). ybcJ also interacts with metabolic enzymes and degradation proteins. It interacts with Pnp-polyribonucleotide nucleotidyltransferase, aceF- dihydrolipoyllsine-residue acetytyltransferase component of pyruvate dehydrogenase complex, dnaJ- chaperone protein DnaJ, and Eno – alpha-enolase (Table 4). Pnp is involved in mRNA degradation and tRNA processing; it catalyzes the

phosphoroylsis of ss polyribonulceotides in the 3' to 5' direction and works with RNase II (Table 4). AceF catalyzes the conversion of pyruvate to acetyl-CoA and oxygen, and DnaJ works with 2 other proteins to disaggregate stress-denatured and unfolded proteins [27]. The *S. pneumoniae* homolog, SP_2226, was found to interact by Y2H in the Uetz lab with SP_0064 - PTS system IIA component – and by microfluidics with SP_2227 – recombination protein F.

Table 3. Annotation of SSA_2372 from NCBI & UNIPROT.

SSA_2372 Annotation		
Sequence	MEYKLFDDYITLQALLKETGIIQSGGAIKTFLSEYPVLFNGEPENRRGKKLR VNDRISLPEQGIEINLTAPSQEEILQHQKEIAEKKRVAELVKAMNKDLKKSQ TSKKEKRKNTGIPKKQKTTKSPVRFPGI	
Domains	S4 Domain	Coil-coil
	Approximately 64 AA in length	Approximately 30 AA in length
	3 alpha helices and 5 beta sheets	Alpha helices in scissor shape
	Conserved domain	Non-conserved domain
EGGNOG	COG2501	
NCBI	YP_001036279.1	
UNIPROT	A3CRC4	
No identical proteins outside of *Streptococcus sanguinis*		
Found in all sequenced strains of *Streptococcus sanguinis*		

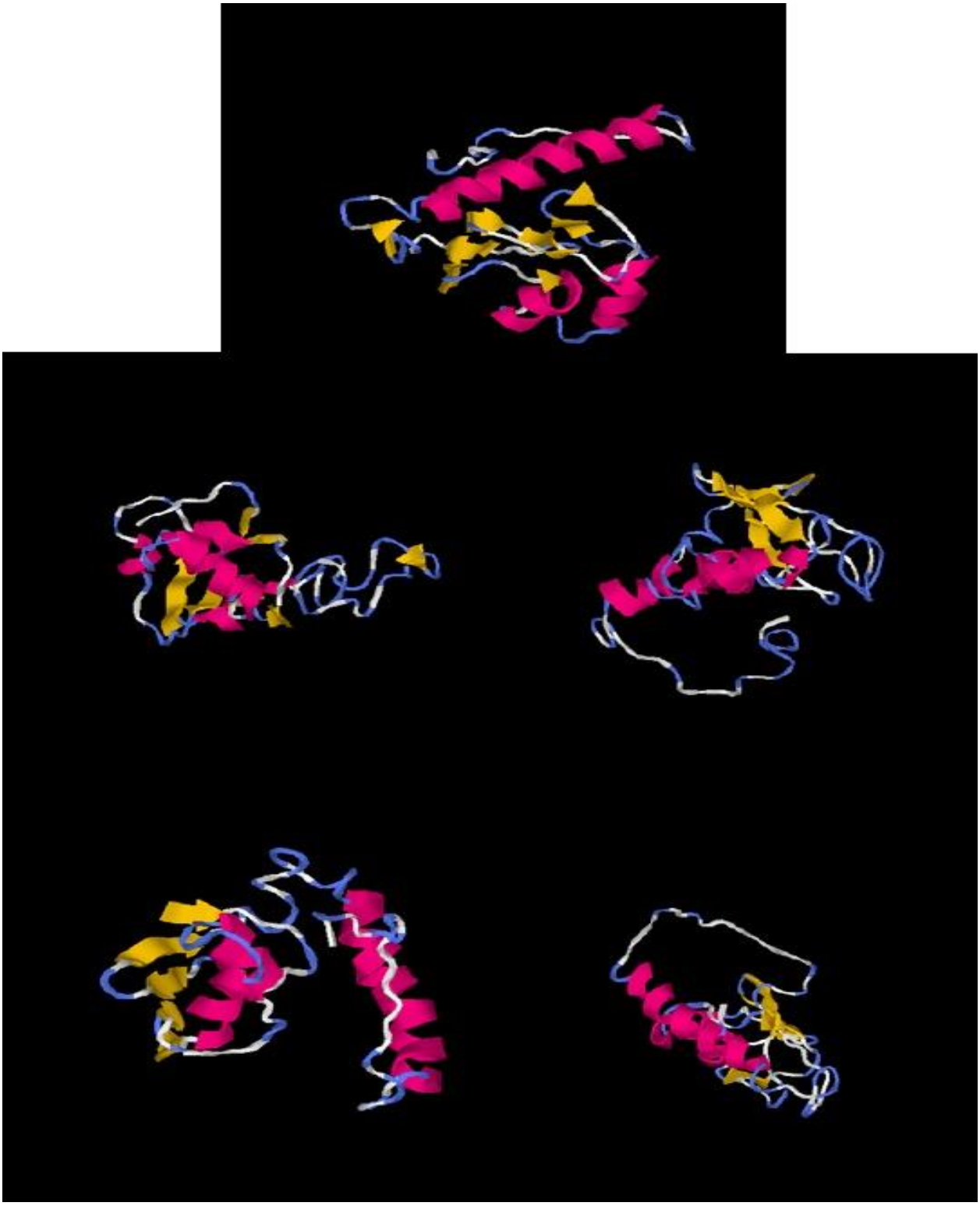


Figure 10. Structural rendering of SSA_2372 using RaptorX. RaptorX developed this 3-D rendering based on the sequence alone using its in house developed software.

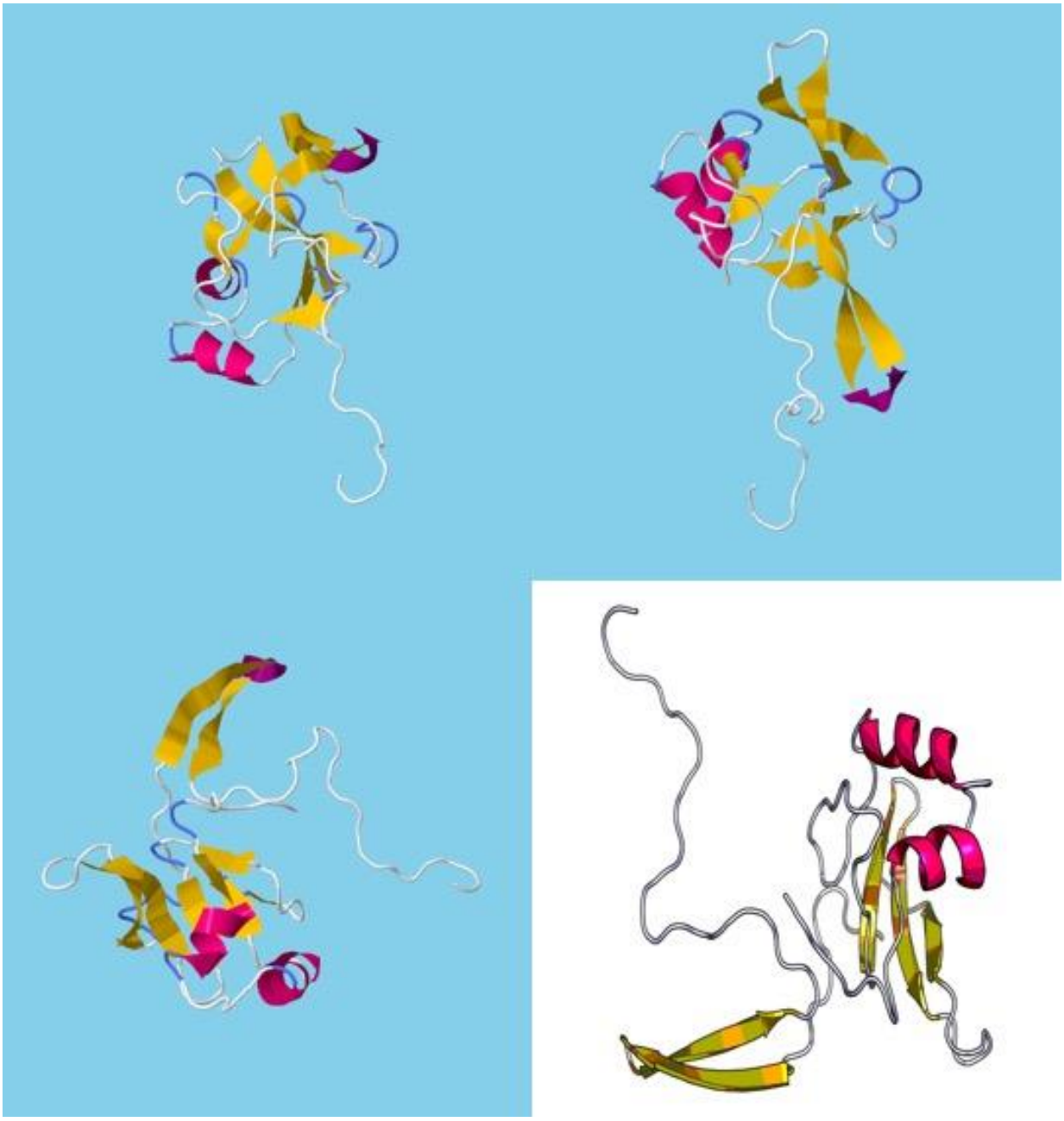


Figure 11. Structural rendering of SSA_2372 using iTasser. iTasser developed this 3-D rendering for SSA_2372 using it sequence and homolog modeling. In the S4 protien family, there currently are only 3 crystal structures in PDB: ybcJ, S4, and tyrosyl-tRNA synthetase.

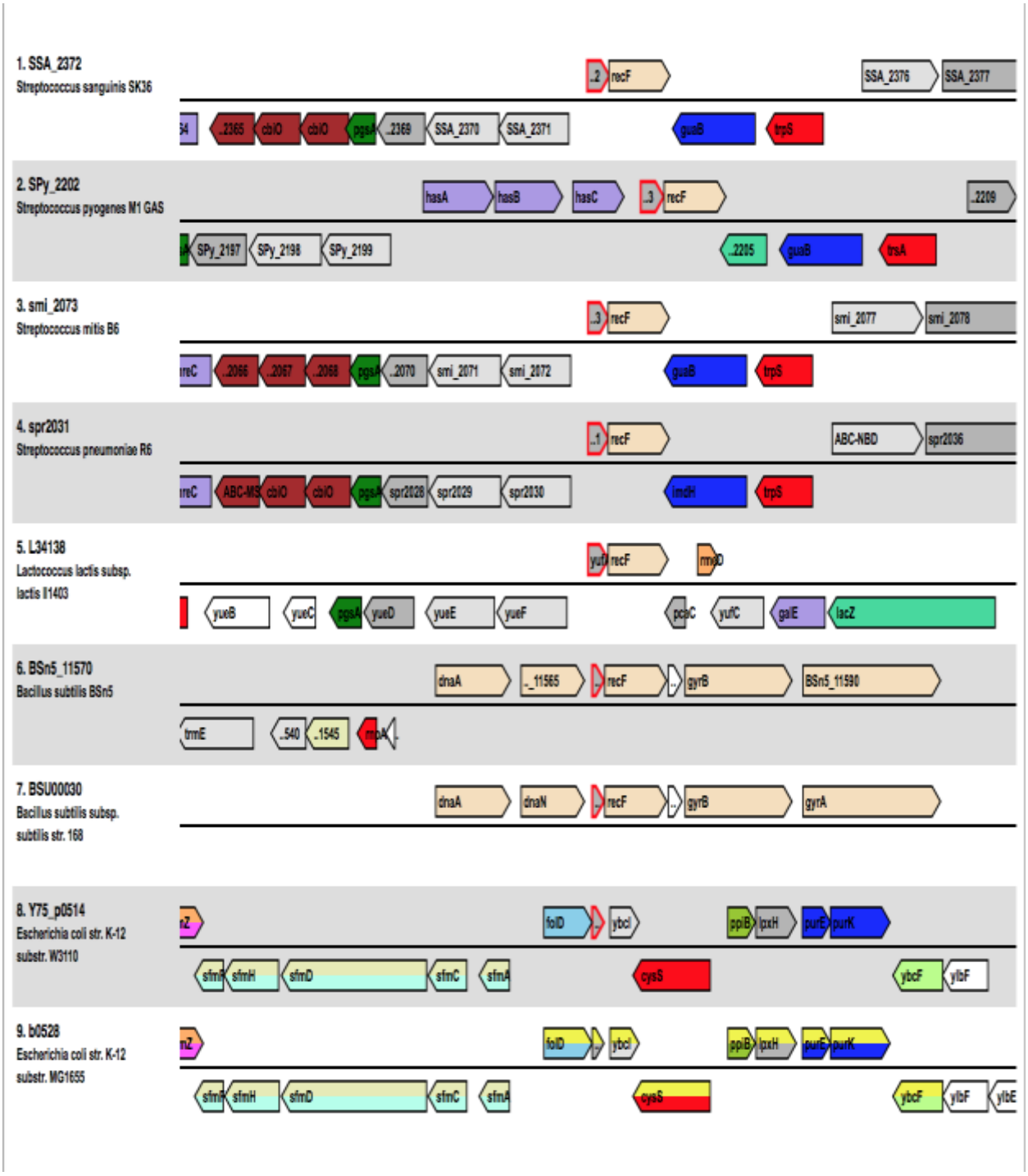


Figure 12. Genomic context of SSA_2372 and its homologs created using MgCV. NCBI locus tags were extracted from UNIPROT and EGGNOG, and these locus tags were used in MGcV to visualize SSA_2372 and COG2501 homologs in the context of their position in their respective genomes. SSA_2372 and the COG2501 homologs are framed in red.

Figure 13. Sequence alignment of SSA_2372 of *S. sanguinis* and yAAA of *B.subtilis*. The sequences were acquired from NCBI, and the sequence comparison was done using ClustalW.

Figure 14. Sequence alignment of SSA_2372 of *S. sanguinis* and ybcJ of *E. coli*. The sequences were acquired from NCBI, and the sequence comparison was done using ClustalW.

Table 4. Annotation of SSA_2372 from NCBI & UNIPROT.

COG2501	Molecule B	Interaction Detection Method	Name
ybcJ	rpsF	pull down	30s ribosomal protein S6
	pnp	pull down	Polyribonucleotide nucleotidyltransferase
	aceF	pull down	dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex
	dnaJ	pull down	chaperone protein DnaJ
	eno	pull down	alpha-enolase
	rplK	pull down	50S ribosomal protein L11
	rplL	pull down	50S ribosomal protein L7/L12
	rplS	pull down	50S ribosomal protein L19
	rplA	pull down	50S ribosomal protein L1
	rplT	pull down	50S ribosomal protein L20
	rpmG	pull down	50S ribosomal protein L33
	rplI	pull down	50S ribosomal protein L9
	rpsB	pull down	30S ribosomal protein S2
	rpsC	pull down	30S ribosomal protein S3

Table 5. Experimentally Confirmed Interactions of SSA_2372 Firmicute homologs: SP_2226, (*S. pneumoniae*), SP_2100 (*S. pneumoniae*), and BSU11620 (*B. subtilis*)

COG2501	Molecule B	Interaction Detection Method	Name
tyrosyl-tRNA synthetase (SP_2100)	SP_1851	miniaturized immunoprecipitation	Hypothetical protein
the RluA and RsuA family of c synthases (BSU11620)	BSU19400	two hybrid system	Superoxide dismutase-like protein YojM
SP_2226	SP_2227	Microfluidics	Recombination protein F
	SP_0064	Yeast two hybrid system (uetzlab)	PTS system; IIA component

Table 6. Protein Abundance of COG2501 homologs extracted from PaxDB.

Protein Abundance		
Gene Name	PPM	Percentage in the Entire Proteome of that extraction
ybcJ	805 ppm	Top 25%
	617 ppm	Top 25%
	352 ppm	Top 25%
	331 ppm	Top 25%
Spy_2203	5.11 ppm	Bottom 10%
	9.69 ppm	Bottom 10%
	31.5 ppm	Bottom 25%

Chapter 5

Phenotype Comparison of 2 homologs of SSA_2372: BSU00030 in *B. subtilis* and ybcJ in *E.coli*

To determine if the acid sensitivity phenotype was SSA_2372 specific, strains with mutations in two homologs, YbcJ and BSU00030, were grown in pH increments of 0.5 within the range of pH 5 – 8. At pH 7 and pH 6, BSU00030 had no significant difference in growth rate or cell yield to the *B. subtilis* WT (Figure 15). However at pH 5.5 and pH 5, BSU00030's cell yield was significantly different from *B.subtilis* WT (Figure 16). In the cases where the mutant was significantly different from the WT, it was the mutant that had significantly smaller cell yield. *E. coli* had different results. There was no significant difference at pH 7 between the control and the YbcJ mutant (Figure 17). At pH 6, there was a significant difference between control and YbcJ mutant, but in this case it was the control *E.coli* mutant, *oxaC* gene knockout, that had a significantly smaller cell yield compared to the COG2501 homolog, YbcJ. At pH 5, however, there was no significant difference between the YbcJ mutant and the control mutant in cell yield or growth rate (Figure 18).

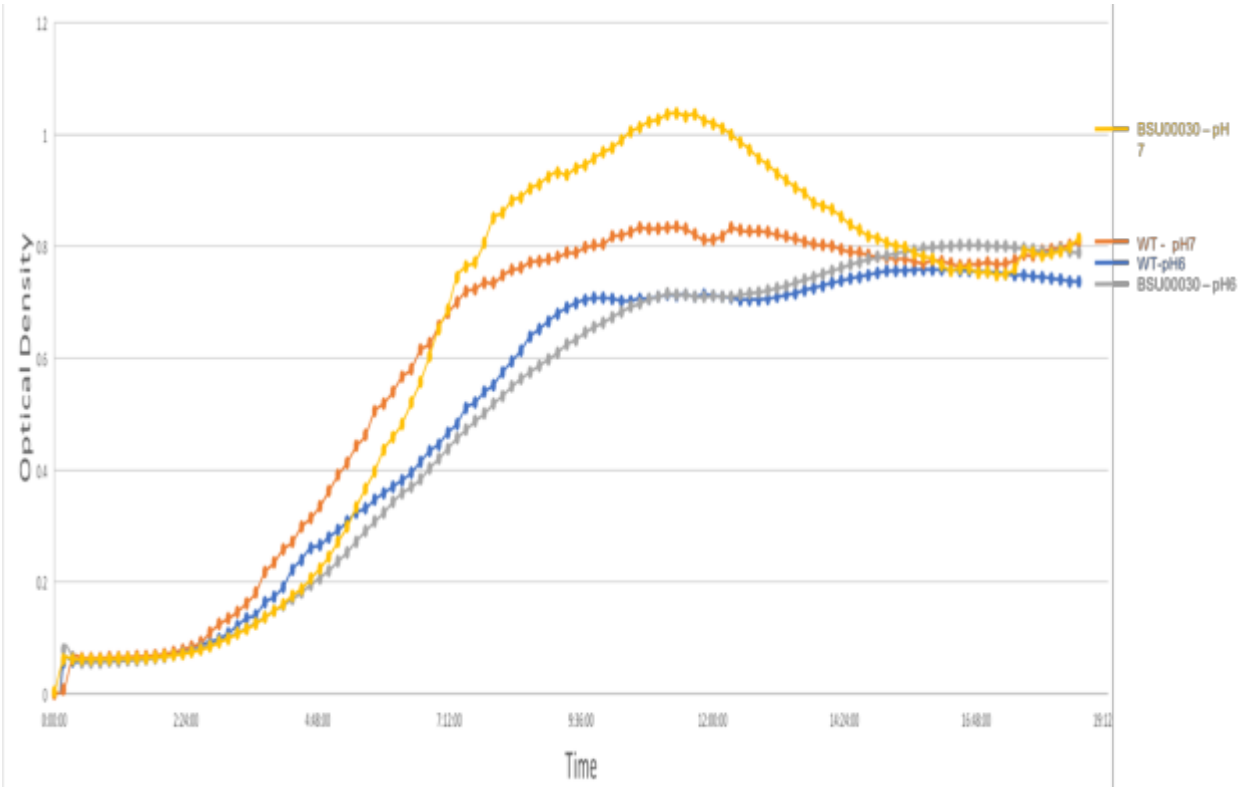


Figure 15. Quantification of Acid Tolerance Phenotype in *B. subtilis* WT and BSU00030 at pH 6 and 7. 96 well black walled plates had media containing BHI media of the following pH range: 5, 5.5., 6, 6.5, 7, and 8. In separate experiments, the OD value at 600nm of the WT and BSU00030 was recorded every 10 minutes for 20 hours. This graph shows the comparison of the growth values at pH 7 and pH 6. With a 95% confidence level, the p-value for a two sided t-test between the WT and the mutant at pH 6 is 0.899, which is not significant.

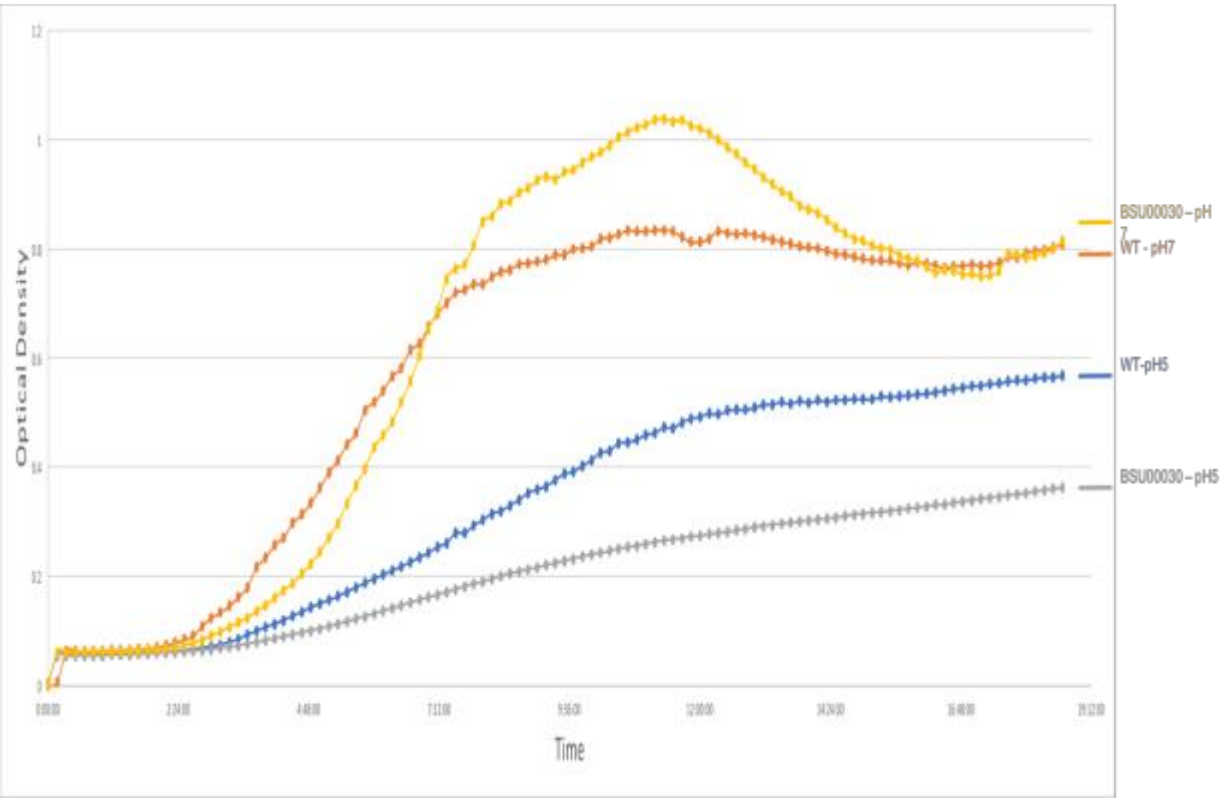


Figure 16. Quantification of Acid Tolerance Phenotype in *B. subtilis* WT and BSU00030 at pH 5 and 7. 96 well black walled plates had media containing BHI media of the following pH range: 5, 5.5., 6, 6.5, 7, and 8. In separate experiments, the OD value at 600nm of the WT and respective mutant was recorded every 10 minutes for 20 hours. This graph shows the comparison of the growth values at pH 7 and pH 5. With a 95% confidence level, the p-value for a two sided t-test of the WT and mutant at pH 5 is 5E-9, which is significant.

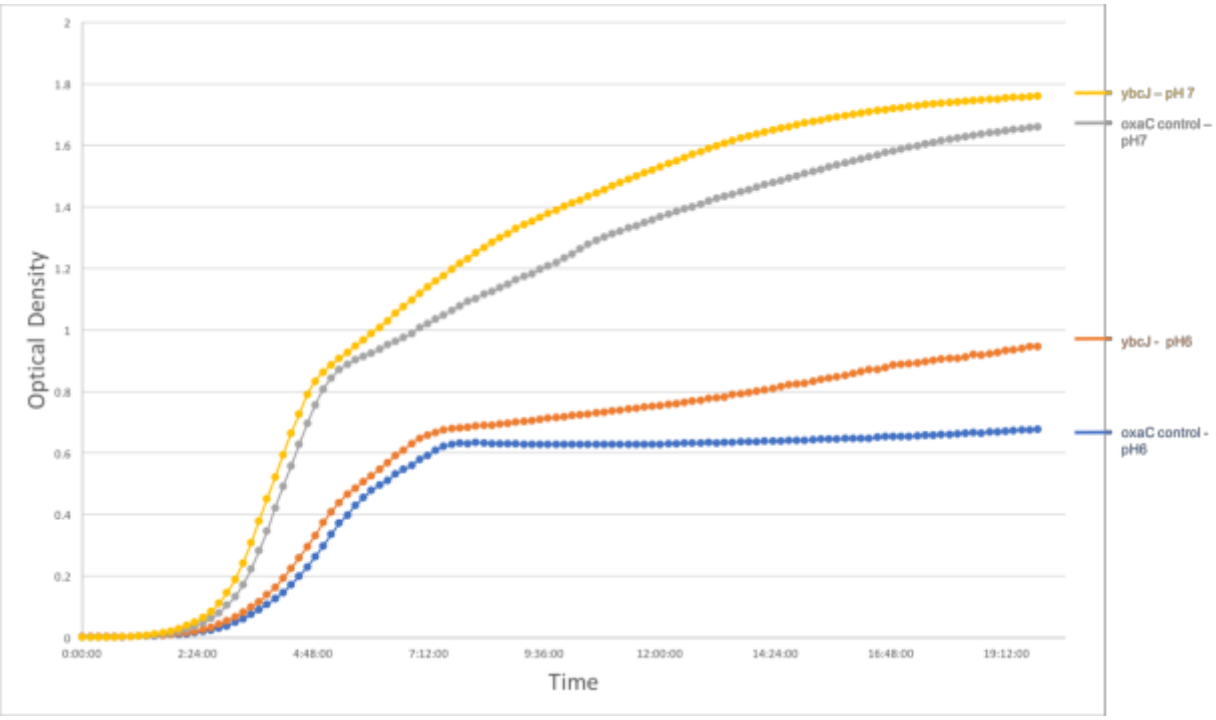


Figure 17. Quantification of Acid Tolerance Phenotype in *E.coli* oxaC kanamycin cassette control and ybcJ at pH 6 and 7. 96 well black walled plates had media containing BHI media of the following pH range: 5, 5.5., 6, 6.5, 7, and 8. In separate experiments, the OD value at 600nm of the WT and respective mutant was recorded every 10 minutes for 20 hours. This graph shows the comparison of the growth values at pH 7 and pH 6. With a 95% confidence level, the p-value for a two sided t-test between the control and the mutant at pH 6 is 0.002968, which is significant.

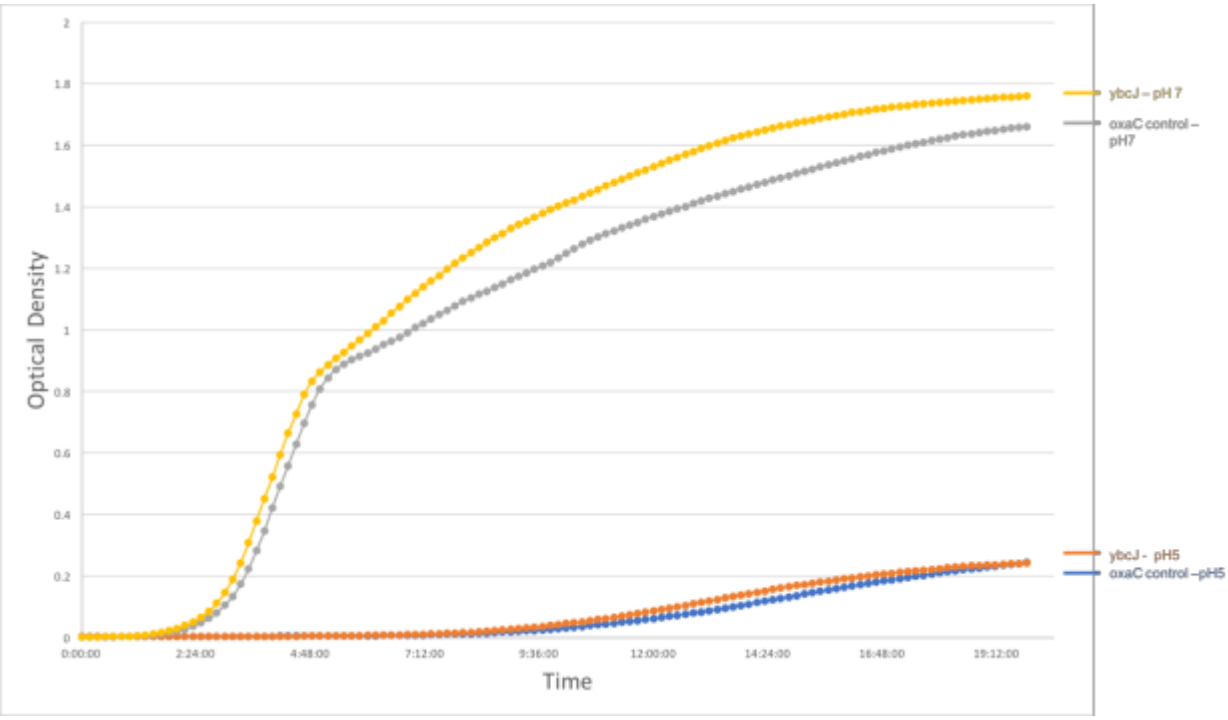


Figure 18. Quantification of Acid Tolerance Phenotype in *E.coli* oxaC kanamycin casette control and ybcJ at pH 5 and 7. 96 well black walled plates had media containing BHI media of the following pH range: 5, 5.5., 6, 6.5, 7, and 8. In separate experiments, the OD value at 600nm of the WT and respective mutant was recorded every 10 minutes for 20 hours. This graph shows the comparison of the growth values at pH 7 and pH 5. With a 95% confidence level, the p-value for a two sided t-test between the control and the mutant at pH 5 is 0.2896, which is not significant.

Chapter 6

Discussion

Screening mutants based on phenotypic features provides pertinent information that can be immediately annotated to its COG or ENOG; however, if you desire perspective on the bacteria's behavior in its natural environment, you must select your phenotype choice carefully. For our phenotypic screens, I selected different pH ranges and sugars as the natural biome of *S. sanguinis* SK36 is the mouth, which is exposed to different metabolites and pH ranges regularly. With these two types of screens, we could easily detect the impact to the mutants, and draw relevant conclusions. In the oral microbiome, pH decreases as a result of the metabolism of the native bacteria; it will also vary due to the pH of the items consumed, whether being acidic or alkaline. An individual's diet, which depending on your geographic region can be very specific, dictates the exposure to simple sugars that your oral microbiome experiences. For example, we used mannose as one of the screening sugars, which can be found in tomatoes, apples, oranges, and cranberries. If these foods aren't common in your environment, your bacteria may not have developed a pathway to process it, or it may exist as a non-essential pathway that was developed in early homo sapiens. In the case of the study, we found more than one mutant that showed susceptibility to low pH, such as SSA_1480, but none other with such a strong phenotype to low pH as SSA_2372. The decision to focus on SSA_2372, otherwise known as COG2501, was because the gene is conserved within multiple strains of *S. sanguinis*; it has a GC content similar to most of its genome, demonstrating that it was not horizontally transferred; and because it is a well conserved homolog in all phyla.

In deciding to perform phenotypic screening, one caution - that must be considered - is whether the phenotype observed is directly related to the gene's function or is indirectly related to the gene's function. The change in proton concentration intracellularly can impact various processes from DNA replication to protein stability. Therefore, when a mutant shows a phenotype of sensitivity to low pH, it cannot be assumed it is due directly to the protein's function, but must consider it could be related to the protein's interactions with other proteins that have a direct impact on intracellular pH maintenance or pH damage recovery systems. In the case of COG_2501 homologs in the Firmicutes phyla, the protein itself is annotated as an RNA binding protein based on similarity in sequence motif, but that should not alone have great impact on maintenance of intracellular proton concentration or DNA damage. However, all homologs in the Firmicutes phyla share an operon with recombination protein F, which is necessary to repair DNA damage. Without the homologous recombination pathway of DNA repair at a pH extreme to the particular bacteria, it will no longer be able to grow or survive. This is an example of an indirectly related phenotype where it is the protein's interaction with another protein that leads to the expression of a phenotype.

The method of screening, solid or liquid phase, can streamline your screening process. For the oral microbiome, the majority of the screening is performed in liquid phase, which is beneficial since it provides quantitative data immediately. However, performing the screening in solid phase provides qualitative and physical data, which is extremely useful. The problem arises when you try to do solid phase with chemically defined media. Certain vitamins degrade above a temperature of 60 C. Therefore, I had to design a method to create solid phase chemically defined agar plates. Agar remains in liquid phase until 55 C. Therefore, the flask has to be placed in a hot water bath to maintain it at this temperature after autoclaving. The temperature sensitive

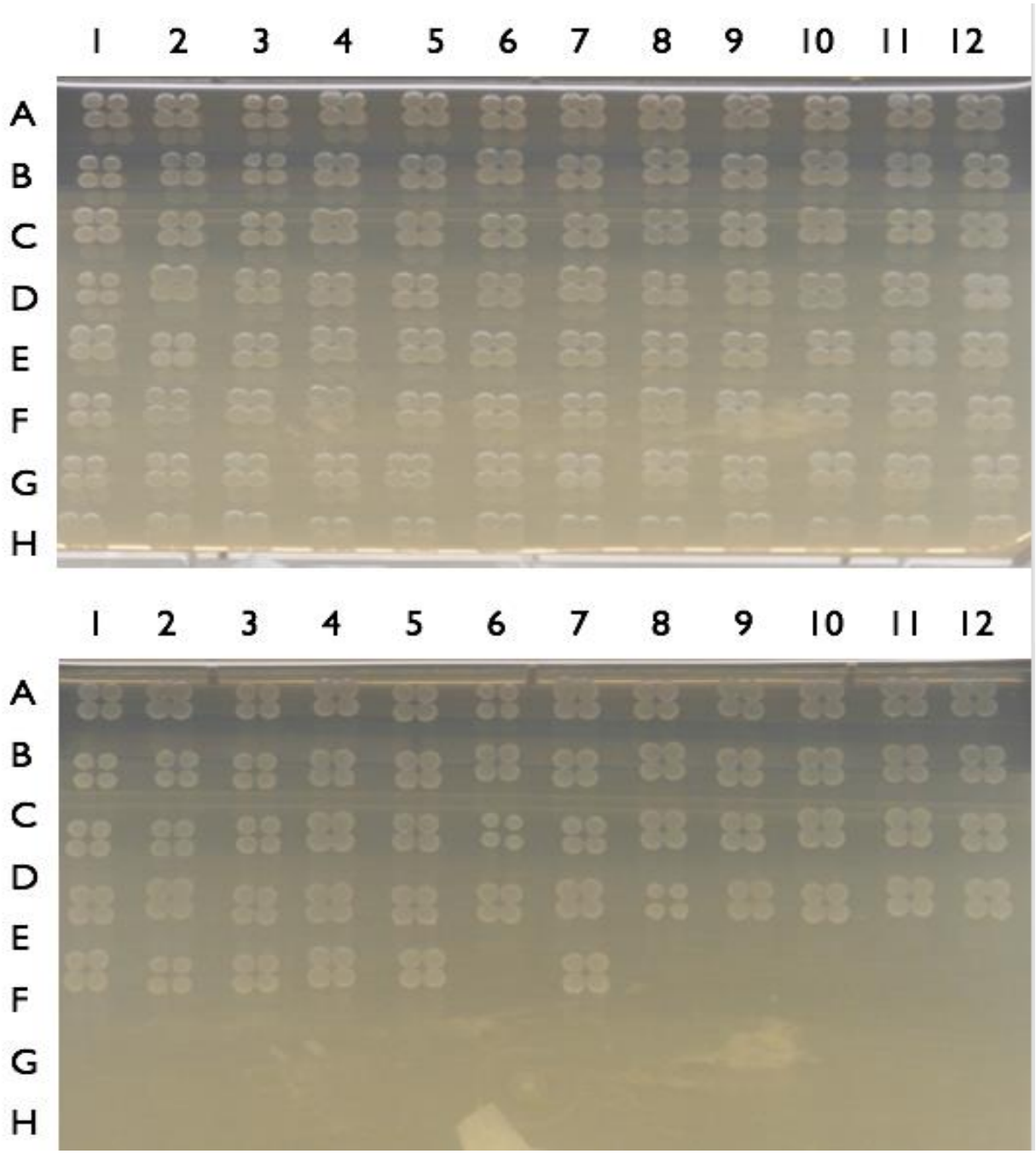
nutrients are added when the water bath temperature settles back to 55 C (Note: When you place a container of a higher temperature in water, it will increase the temperature of the water temporarily).

When using interaction network databases, the networks presented must be evaluated carefully. For SSA_2372, the string database presents multiple functional partners including DNA gyrase B, chromosomal initiation protein, DNA-directed RNA polymerase subunit delta, and DNA polymerase III subunit beta. This network would suggest that SSA_2372 is involved in DNA replication. The string database annotates it as being due to genomic context. However, when you look at the genomic context for SSA_2372, it is only adjacent to recombination protein F. When you continue data mining, it is discovered that the genes labeled as within the neighborhood and as appearing concurrently appears only in the model organism, *E. coli.* Most times the data presented in databases like String will be based on only one bacterium, but it will present the data as if it were based on that particular species or genus. There needs to be better contextual relationships presented for these networks, and actual interaction data should be utilized such as from Intact or by data mining.

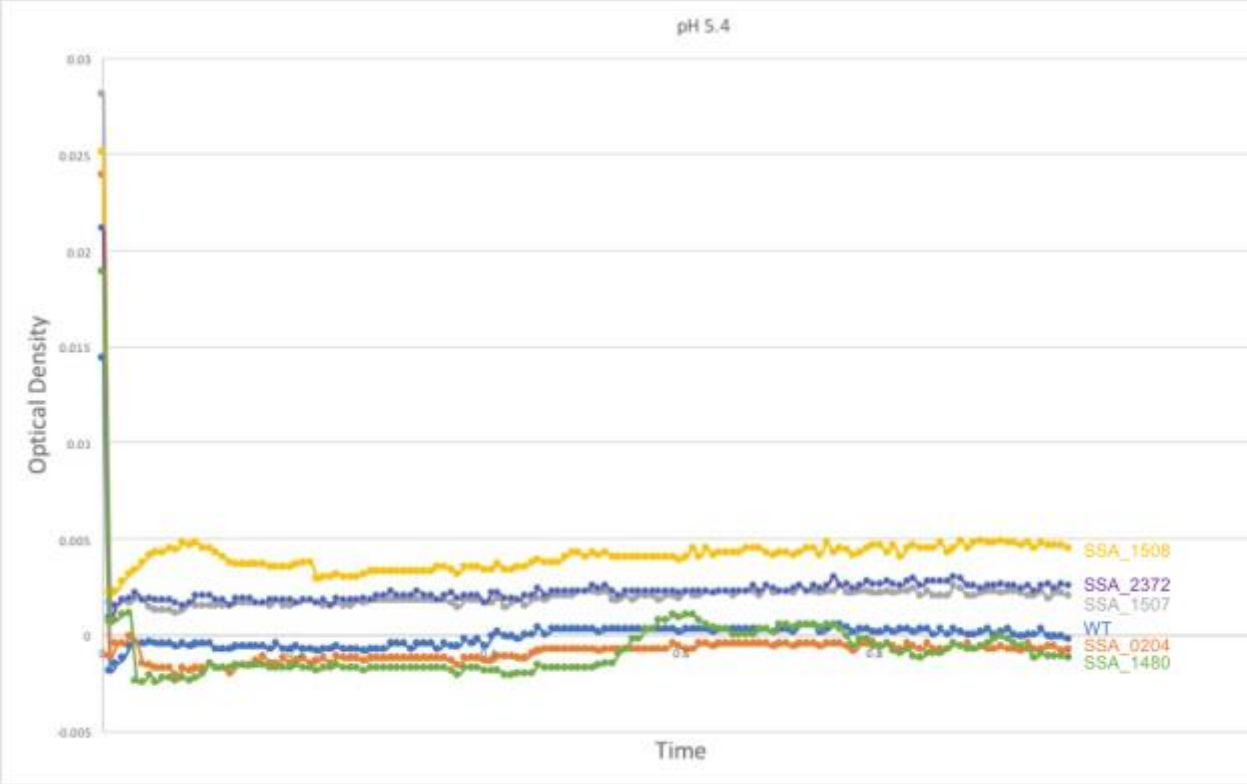
Based on the genomic context of *S. sanguinis* SK36 and a microfluidic experiment performed by a collaborator (Meier et al), SSA_2372 is involved in some degree with the protein recombination protein F. For all bacteria in the phyla Firmicutes, the homologs, COG2501, are adjacent (2 to 5 AA apart) to *recF*. In a microchip array study with cDNA spotted in individual wells, recF as a bait expressed in a well interacted with the SP_2226 (*Streptococcus pneumoniae* TIGR4) protein solution. Since the cDNA is transcribed to RNA and translated to protein in the well, there is a potential that the SP_2226 interacted with both the DNA/RNA and protein of SP_2227, RecF.

Based on our data and the data mined, there is a great likelihood that our protein is a transcription factor or regulator of recombination protein F. The reason that I believe it is a regulator is because of its protein expression level. Proteins in the cytosol that bind RNA or bind other proteins tend to be expressed in great abundance (30). However, transcription factors or other regulators tend to be expressed in low abundance. Our protein is expressed in the bottom 10 to 25% of total proteins expressed in a bacterium (Table 6). However, its homolog which is known to bind RNA is expressed in the top 25% repeatedly (Table 6). Based on the microfluidics data of *S. pneumoniae*, its only interactor is recombination protein F after being screened against all proteins is only recF. Now these wells will contain both protein, cDNA, and RNA. Therefore, I cannot be certain at which stage SSA_2372 regulates its expression. Nevertheless, when dsDNA damage occurs in response to a high proton concentration without the presence or functionality of SSA_2372, the growth rate and cell yield is diminish significantly. There is another piece of information that suggests it is involved in the homologous recombination pathway. In Y2H, it bound PTS system IIA component (Table 5). PTS system IIA component whether cytosolic or membrane bound catalyzes the pathway that leads to the transcription and later translation of catabolic proteins. RecF requires ATP in its recFOR pathway in order to load recA to the dsDNA breaks. Therefore, it would make sense potentially that a protein involved in regulating recF expression would also be involved in metabolism. Since SSA_2372 has two binding domains that also increases the feasibility of it performing 2 functions (Table 3).

Supplemental



Supplemental Figure 1. BHI Agar Plate (pH7) Pinned with *Streptococcus sanguinis* SK36 mutants. On these two plates, 95 mutants were pinned in quadruplicate, and the plate was incubated at 37 °C under microaerophilic conditions for 48 hours.



Supplemental Figure 2. Growth Curves of the S. sanguinis SK36 candidate mutants and control mutants in BHI broth at pH 5.4. 96 well plates were inoculated in pH specific media, and it was grown overnight under microaerophilic conditions. Optical density readings at 600nm were recorded every 10 minutes.

R Code. Statistical Tests performed to determine statistical significance were performed using R Studio and the following code.

Eleventh<-read.csv(file.choose(),header = TRUE)
Eleventh$X6y
Eleventh$X6oxac
sixy<-(Eleventh$X6y)
sixy
sixoxac<-(Eleventh$X6oxac)
Eleventh$X5point4y
Eleventh$X5point4oxac
fivefoury<-(Eleventh$X5point4y)
fivefoury

```
fivefouroxac<-(Eleventh$X5point4oxac)
msixy<-mean(Eleventh$X6y)
msixoxac<-mean(Eleventh$X6oxac)
t.test(sixy, sixoxac, alternative = c("two.sided"), mu = 0, paired = FALSE, conf.level = 0.95)
t.test(fivefoury, fivefouroxac, alternative = c("two.sided"), mu = 0, paired = FALSE,
conf.level = 0.95)
thirdtwenty<-read.csv(file.choose(),header = TRUE)
thirdtwenty$X5pHoxac
t.test(thirdtwenty$X5pHoxac, thirdtwenty$X5pHybcj, alternative = c("two.sided"), mu = 0,
paired = FALSE, conf.level = 0.95)
t.test(thirdtwenty$X5.5pHoxac, thirdtwenty$X5.5pHybcj, alternative = c("two.sided"), mu =
0, paired = FALSE, conf.level = 0.95)
t.test(thirdtwenty$X6pHoxac, thirdtwenty$X6pHybcj, alternative = c("two.sided"), mu = 0,
paired = FALSE, conf.level = 0.95)
msixy
thirdtwenty$X6pHybcj
sixy
Ecoli<-read.csv(file.choose(),header = TRUE)
t.test(Ecoli$X7pHoxac, Ecoli$X7pHybcj, alternative = c("two.sided"), mu = 0, paired =
FALSE, conf.level = 0.95)
t.test(Ecoli$X6pHoxac, Ecoli$X6pHybcj, alternative = c("two.sided"), mu = 0, paired =
FALSE, conf.level = 0.95)
t.test(Ecoli$X5pHoxac, Ecoli$X5pHybcj, alternative = c("two.sided"), mu = 0, paired =
FALSE, conf.level = 0.95)

June<-read.csv(file.choose(),header = TRUE)
May<-read.csv(file.choose(), header=TRUE)
June$sixsk36
May$sixsk36
fourteen<-(June$six1480 + May$six1480)
meanfour<-((fourteen)/(2))
meanfour
twentythree<-(June$six2372+May$six2372)
meanthree<-((twentythree)/(2))
meanthree
sk36<-(June$sixsk36 + May$sixsk36)
meansk36<-((sk36)/(2))
t.test(meansk36, meanfour, alternative = c("two.sided"), mu = 0, paired = FALSE, conf.level =
0.95)
t.test(meansk36, meanthree, alternative = c("two.sided"), mu = 0, paired = FALSE, conf.level
= 0.95)
fifteen<-(June$six1507 + May$six1507)
meanfif<-((fifteen)/(2))
meanfif
```

t.test(meansk36, meanfif, alternative = c("two.sided"), mu = 0, paired = FALSE, conf.level = 0.95)
t.test(meansk36, meanthree, alternative = c("two.sided"), mu = 0, paired = FALSE, conf.level = 0.99)
t.test(meansk36, meanfour, alternative = c("two.sided"), mu = 0, paired = FALSE, conf.level = 0.99)
meansk36
June$sevenSK36
May$sevenSK36
SK36neutral<-(June$sevenSK36+May$sevenSK36)
meansk36neu<-((SK36neutral)/(2))
meansk36neu
June$seven1507
May$seven1507
fifneutral<-(June$seven1507+May$seven1507)
meanfifneu<-((fifneutral)/(2))
meanfifneu
June$seven2372
May$seven2372
twothreeneutral<-(June$seven2372+May$seven2372)
meantwothreeneu<-((twothreeneutral)/(2))
meantwothreeneu
Bacillus<-read.csv(file.choose(),header = TRUE)
Bacillus
t.test(Bacillus$X5.5pHbs168, Bacillus$X5.5pHbsx, alternative = c("two.sided"), mu = 0, paired = FALSE, conf.level = 0.95)
t.test(Bacillus$X6pHbs168, Bacillus$X6pHbsx, alternative = c("two.sided"), mu = 0, paired = FALSE, conf.level = 0.95)
t.test(Bacillus$X5pHbs168, Bacillus$X5pHbsx, alternative = c("two.sided"), mu = 0, paired = FALSE, conf.level = 0.95)
t.test(Bacillus$X5pHbs168, Bacillus$X5pHbsx, alternative = c("two.sided"), mu = 0, paired = FALSE, conf.level = 0.99)
t.test(Bacillus$X7pHbs168, Bacillus$X7pHbsx, alternative = c("two.sided"), mu = 0, paired = FALSE, conf.level = 0.95)

References

1. Foundation N. *Homology.* Hoboken: Hoboken : Wiley; 2008.

2. Inkpen SA, Doolittle WF. Molecular Phylogenetics and the Perennial Problem of Homology. *J Mol Evol*. 2016.

3. Shahbaaz M, Bisetty K, Ahmad F, Hassan MI. Current Advances in the Identification and Characterization of Putative Drug and Vaccine Targets in the Bacterial Genomes. *Current Topics in Medicinal Chemistry*. 2016;16:1040-1069.

4. Ijaq J, Chandrasekharan M, Poddar R, Bethi N, Sundararajan VS. Annotation and curation of uncharacterized proteins- challenges. *Frontiers in genetics*. 2015;6:119.

5. Galperin MY, Koonin EV. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res*. 2004;32:5452.

6. Pommerville JC, author. *Fundamentals of Microbiology. Body Systems Edition.* Third edition, Body systems edition.. ed. Burlington, MA : Jones & Bartlett Learning; 2016.

7. Williams NH. DNA Hydrolysis: Mechanism and Reactivity. *Nucleic Acids and Molecular Biology*. 2004;13:3-17.

8. Kidane D, Sanchez H, Alonso JC, Graumann PL. Visualization of DNA double- strand break repair in live bacteria reveals dynamic recruitment of Bacillus subtilis RecF, RecO and RecN proteins to distinct sites on the nucleoids. *Mol Microbiol*. 2004;52:1627-1639.

9. Malacinski GM. *Essentials of Molecular Biology.* 4th ed.. ed. Boston: Boston : Jones and Bartlett; 2003.

10. Vaughan P. *DNA Repair Protocols Prokaryotic Systems.* Totowa: Totowa : Humana Press; 2000.

11. Lund P, Tramonti A, De Biase D. Coping with low p H : molecular strategies in neutralophilic bacteria. *FEMS Microbiol Rev.* 2014;38:

12. Marsh P. *Oral Microbiology.* Washington, D.C.: Washington, D.C. : American Society for Microbiology; 1980.

13. Xu P, Ge X, Chen L, et al. Genome-wide essential gene identification in Streptococcus sanguinis. *Scientific Reports*. 2011;1.

14. Powell S, Forslund K, Szklarczyk D, et al. eggNOG v4.0: nested orthology inference across 3686 organisms. Nucleic Acids Res. 2014;42:D231.

15. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high- quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*. 2011;7

16. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44:W242.

17. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, Mering C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*. 2015;15:3163-3168.

18. Bazzoli A, Tettamanzi AGB, Zhang Y. Computational Protein Design and Large-Scale Assessment by I- TASSER Structure Assembly Simulations. *J Mol Biol*. 2011;407:764-776.

19. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I- TASSER Suite: protein structure and function prediction. *Nature Methods*. 2014;12:7.

20. Yang J, Zhang Y. I- TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res*. 2015;43:W174.

21. Wang S, Li W, Liu S, Xu J. RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Res*. 2016;44:W430.

22. Federhen S. The NCBI Taxonomy database. *Nucleic Acids Res*. 2012;40:D136.

23. Gandrud C. *Reproducible Research with R and RStudio.* Second edition.. ed. Boca Raton : CRC Press, Taylor & Francis Group; 2015.

24. Schumacker R. *Understanding Statistics using R.* New York, NY : Springer New York : Imprint: Springer; 2013.

25. Aravind L, Koonin EV. Novel Predicted RNA-Binding Domains Associated with the Translation Machinery. *J Mol Evol*. 1999;48:291-302.

26. Jiang M, Sullivan SM, Walker AK, Strahler JR, Andrews PC, Maddock JR. Identification of Novel Escherichia coli Ribosome-Associated Proteins Using Isobaric Tags and Multidimensional Protein Identification Techniques. *The Journal of Bacteriology*. 2007;189:3434.

27. Martin E. UniProt. Oxford University Press; 2009

28. intact. Oxford University Press; 2008.

29. Overmars L, Kerkhoven R, Siezen RJ, Francke C. MGcV: the microbial genomic context viewer for comparative genome analysis. *BMC Genomics*. 2013;14:209-209.

30. Ishihama Y, Schmidt T, Rappsilber J, et al. Protein abundance profiling of the Escherichia coli cytosol. *BMC Genomics*. 2008;9:102-102.

VITA

Ayana Scott-Elliston was born on January 31, 1988 in Queens, NY. She graduated from Maggie L. Walker Governor's School for Government & International Studies in Richmond, VA in 2006. She received her Bachelor of Science in Biochemistry with a Minor in Classical Studies & Dance from the University of North Carolina at Greensboro in 2011.