Spring 5-15-2022

THE FUNCTION AND PHYLOGENY OF A DUPLICATE SQUALENE-HOPENE CYCLASE GENE IN METHYLOBACTERIUM EXTORQUENS CM4

Rowan Behnke
Washington University in St. Louis, rowanbehnke@gmail.com

WASHINGTON UNIVERSITY

Department of Earth and Planetary Sciences

THE FUNCTION AND PHYLOGENY OF A DUPLICATE SQUALENE-HOPENE CYCLASE GENE IN *METHYLOBACTERIUM EXTORQUENS* CM4

by

Rowan Behnke

A thesis presented to the Department of Earth and
Planetary Science of Washington University in
partial fulfillment of the requirements for the degree
of Bachelor of Arts with Honors.

May 2022

Saint Louis, Missouri

Table of Contents

Table of Figures

Table of Tables

Abstract

Hopanoids, a group of isoprenoid lipids produced by certain bacteria, are common biomarkers that are informative about past life. However, several mysteries remain regarding their purpose and production *in vivo*. One of these mysteries is why certain strains of bacteria possess two copies of a gene, *sqhC*, whose product, the enzyme squalene-hopene cyclase, is instrumental in the production of hopanoids. It is unusual for bacteria to have two copies of a gene due to the added energy cost of replicating the second copy. Therefore, it is posited that there is some evolutionary advantage to this second copy of *sqhC*. Additionally, every *sqhC2*-containing strain examined also has a gene for a TetR-family transcriptional repressor immediately upstream of *sqhC2*, which may regulate *sqhC2* expression. Starting from Pearson et al.'s (2007) list of genera whose strains may have two copies of *sqhC*, phylogenetic trees were constructed to examine the evolutionary relatedness of *sqhC1*, *sqhC2*, and the *tetR* gene associated with *sqhC2*. These indicate that *sqhC2* and *tetR* may have co-evolved, supporting the idea that this TetR protein regulates *sqhC2*, and that different genera evolved *sqhC2* independently several times, forming several orthologous clades, supporting the idea that *sqhC2* and its associated *tetR* carry an evolutionary advantage. The *tetR* gene associated with *sqhC2* was unable to be knocked out in *Methylobacterium extorquens* CM4, which may indicate that regulation is necessary to the evolutionary advantage conferred by *sqhC2*, but this is pure speculation.

Introduction

Hopanoids are pentacyclic isoprenoid lipids produced by certain bacteria. A representative hopanoid structure is shown in Figure 1 below, adapted from Kulkarni et al. (2015). Hopanoids are typically well-preserved in the rock record, making them an extremely

common organic component in sediment and sedimentary rocks up to 1.64 Ga (Belin et al. 2018). As such, hopanoids are useful in deciphering geobiological history, particularly when reconstructing bacterial ecology. However, despite existing work using hopanoids to investigate paleobiology, much remains unknown about the role of hopanoids in living organisms. In order to be able to use hopanoids to their full potential in paleo- and geobiological research, it is imperative that their biological role be understood in great detail in modern bacteria. This project aims to expand existing knowledge about the biosynthetic pathway of hopanoids.

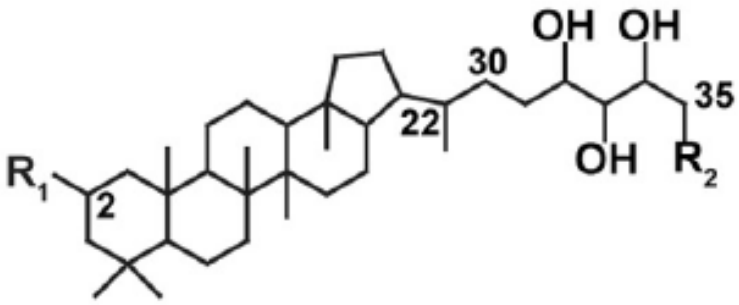

Figure 1: Representative hopanoid structure (adapted from Kulkarni et al. 2015)

While some debate remains (Saenz 2015), the most widely-accepted function of hopanoids in live bacteria is as a control on cell membrane fluidity, similar to the role of sterols (tetracyclic isoprenoid lipids) in eukaryotes (Belin et al. 2018). Despite remaining questions as to the molecules' function, the process of hopanoid ring biosynthesis is well-characterized. Hopanoids are synthesized by first cyclizing a squalene molecule and then modifying the side chain in a variety of ways to produce diverse hopanoids (Belin et al. 2018). The cyclization of squalene for hopanoid formation is performed in one concerted reaction, catalyzed by the enzyme squalene-hopene cyclase (SqhC), as shown in Figure 2 below, adapted from Pearson et al. (2007). SqhC cyclization of squalene produces either diploptene or diplopterol, pictured in Figure 3 below, adapted from Belin et al. (2018). Only ~10% of bacteria are believed to possess

the *sqhC* gene that encodes SqhC, which defines the species capable of hopanoid production (Pearson et al. 2007).

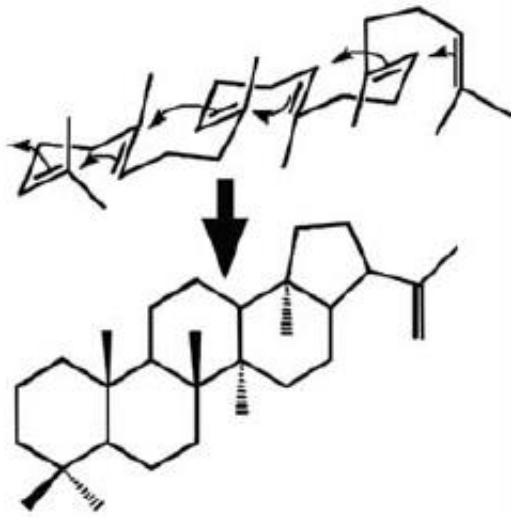

Figure 2: Electron-pushing diagram showing how squalene is cyclized to diploptene in one concerted reaction (adapted from Pearson et al. 2007)

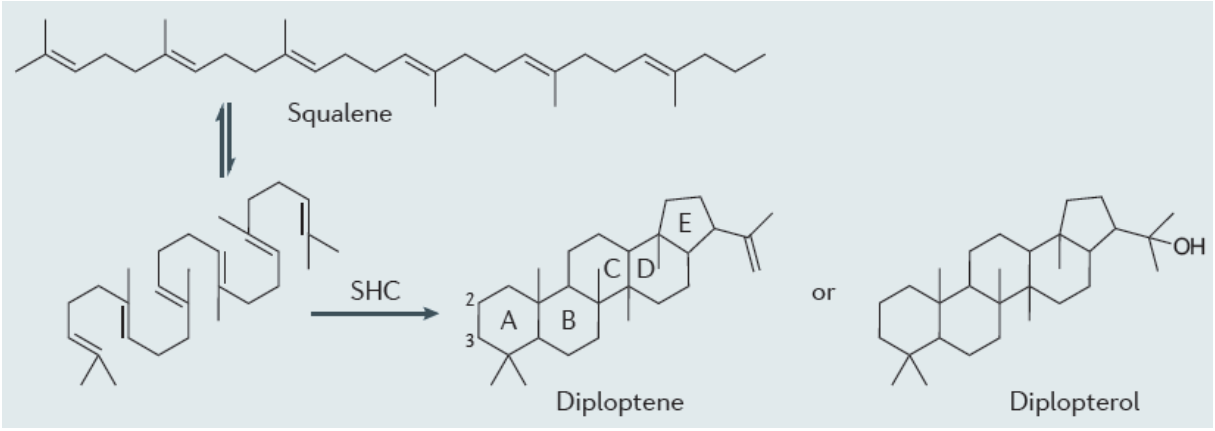

Figure 3: Squalene-hopene cyclase-mediated cyclization of squalene to form diploptene or diplopterol (adapted from Belin et al. 2018)

Despite how well-characterized hopanoid biosynthesis is, some aspects of their production remain unknown. One of these revolves around a set of bacteria that appear to have two copies of the *sqhC* gene. In preliminary work by Pearson et al. (2007), the duplicate copies appear to be orthologous, more closely related to one another than any paralogous pair possessed by one organism. Additionally, this duplicate copy, dubbed *sqhC2*, appears to be associated with

a *tetR*-family transcriptional repressor found directly upstream of *sqhC2*. Here, the phylogeny of these duplicate *sqhC* genes, as well as that of the *tetR* genes associated with *sqhC2*, are examined in more detail, and the function of each copy is examined in *Methylobacterium extorquens* CM4. It is hypothesized that the two copies of *sqhC* may be involved in the production of different hopanoids, or that *sqhC2* simply produces more of the same hopanoids under some as-yet-unknown conditions. Especially in the latter case, which could be achieved through simple regulation of one copy of *sqhC*, the evolutionary advantage of this system must be characterized. The replication of an additional copy of a gene carries an energy cost that reduces fitness (Adler 2014), so it is believed that *sqhC2* must increase fitness in some way to compensate for this penalty. Furthermore, the TetR repressor encoded upstream of *sqhC2* must be confirmed to act on *sqhC2* itself. Phylogenetic evidence presented here suggests that the *sqhC2-tetR* system evolved multiple times in different clades, lending credence to the idea that this system carries an evolutionary advantage, and that *sqhC2* and *tetR* coevolved, lending credence to the idea that this TetR regulates *sqhC2*. The lipid products of the enzymes encoded by *sqhC1* and *sqhC2* in *M. extorquens* CM4 were unable to be compared, but this, in itself, suggests the possibility that the regulation afforded by TetR is crucial to the increase in fitness afforded by the *sqhC2-tetR* system.

Methods

Phylogenetic Tree Building and Comparison

Sequence Generation using BLAST

These sequences and the phylogenetic trees generated in the next section were ultimately not used, but are included for completeness. The preliminary method of sequence collection used GenBank and the BLAST tool to find and compile the nucleotide sequences of *squalene-hopene cyclase* (*sqhC*) genes in species known to have two copies based on the work of Pearson et al.,

2007. The BLASTN program was used to gather these sequences using the nucleotide sequence of *M. extorquens* CM4 acquired earlier in the history of this project as the query sequence. This query sequence was used in a search against the genomes of each of the species identified by Pearson, et al., in turn. This was performed by entering a species name (e.g., *Geobacter sulfurreducens*) in the "Organism" field and selecting each strain in the GenBank library in turn and searching using the BLASTN program for slightly similar sequences and its default parameters.

Each of the high-coverage, high-fidelity alignments (typically the first two to three alignments sorted by E-value) was examined in GenBank by following the "GenBank" link next to the range of the alignments under the "Alignments" tab of the BLAST results. From this page, the "Graphics" link beneath the name and locus of the GenBank record was followed to the location of the alignment in the subject region. If this alignment block fell within an annotated gene labeled *shc*, this gene's nucleotide FASTA sequence was collected by following the "FASTA record" link that appears when hovering over the name of the gene above the track (not the track itself).

The region upstream of each of the annotated *sqhC* genes was then investigated for the presence of an annotated *tetR* family transcriptional regulator. This was done by hovering over the tracks for gene annotations immediately upstream of the *sqhC* genes and determining whether the "Name" field indicated a *tetR* regulator. Those that do have such names as "transcriptional regulator, TetR family." The nucleotide FASTA sequence of each *tetR* gene identified was obtained in the same way as the *sqhC* sequences.

Tree Generation using PhyML

The FASTA sequences for the *sqhC* genes acquired using the preceding method were aligned using the Clustal Omega tool at https://ebi.ac.uk/Tools/msa/clustalo/, with the output

format set to "ClustalW with character counts" and with default settings. The alignment was then converted back to FASTA format using the tool at https://sequenceconversion.bugaco.com/converter/biology/sequences/clustal_to_fasta.php, and the number of sequences and length of the alignment were added at the top (in this case, "15 2494"). This file was then used to build a phylogenetic tree at https://atgc-montpellier.fr/phyml/, using different combinations of parameters.

Sequence Generation using SEED

Subsequently, *sqhC* and *tetR* sequences were gathered from the SEED database found at https://pubseed.theseed.org. These were found by searching "squalene hopene cyclase [genus]" for the genera presented in Pearson et al., 2007, Figure 4 (e.g., *Geobacter*). Each unique *sqhC* sequence was obtained directly from the SEED viewer. Each was also examined for upstream *tetR* and *acrR* (*acrR* is also in the *tetR* family) genes by looking at the "Visual Region Information" tab. Where *tetR* or *acrR* genes were found, their tracks in the viewer were clicked on, which directs the user to the page for that gene, from which the sequence can be obtained. These sequences are available in Supplemental Material.

Tree Generation using RAxML

The FASTA sequences for the *sqhC* genes acquired using the preceding method were aligned using the Clustal Omega tool at https://ebi.ac.uk/Tools/msa/clustalo/, with the output format set to "ClustalW with character counts" and with default settings. The alignment was then converted back to FASTA format using the tool at https://sequenceconversion.bugaco.com/converter/biology/sequences/clustal_to_fasta.php. This file was then uploaded as the input alignment in RAxML-GUI and the program was run with default parameters. The resulting tree was visualized by uploading the best tree file produced by RAxML to https://itol.embl.de/upload.cgi. This process was repeated for a file containing *tetR*

and *acrR* sequences, as well as one containing only the *sqhC* sequences associated with these *tetR* and *acrR* genes—*sqhC2*.

Tree Comparison

The *sqhC2* and *tetR* trees were put into this website: https://eti.pg.edu.pl/TreeCmp/WEB, which returned a weighted Robinson-Foulds metric based on clusters. Two random trees with the same number of leaves as the *sqhC2* and *tetR* trees (22) were also generated for comparison at this site: http://www.trex.uqam.ca/index.php?action=randomtreegenerator&project=trex, and their weighted Robinson-Foulds metrics based on clusters calculated at https://eti.pg.edu.pl/TreeCmp/WEB.

tetR Knockout Generation

The procedure outlined below was ultimately unsuccessful, despite nine months of effort and troubleshooting. Possible explanations for this will be explored in the results section. Despite its failure to produce results, the procedure used in the attempt to knock out the *tetR* gene upstream of *sqhC2* in *M. extorquens* CM4 is described here. An "in-out" allelic exchange method was used for this purpose. In this method, the sequences immediately upstream and downstream of the gene to be knocked out are cloned onto a plasmid containing an antibiotic resistance cassette, as well as the *sacB* gene. *sacB* encodes levansucrase, which, when expressed in gram-negative bacteria, converts sucrose to a lethal compound. The knockout plasmid sequence is allowed to exchange with the wild-type gene, inserting the knockout plasmid into the genome. Organisms with the knockout plasmid inserted into the genome are selected for through exposure to antibiotics that the knockout plasmid carries resistance for. Exposure to sucrose is then used to select for those organisms which have undergone exchange again to eject the knockout plasmid, either reverting to wild-type or taking on the knockout allele from the knockout plasmid. This second exchange is further confirmed by simultaneous plating with and

without the antibiotics that the knockout plasmid confers resistance to. Those that only grow in the absence of these antibiotics can be either wild-type or knockouts, and gel electrophoresis and sequencing are used to identify knockouts. The knockout plasmid used in this case is pAB274, generated by Jeremy Pomerantz, a previous undergraduate in the Bradley Lab, by modifying pCM433, a vector generated by Christopher Marx for the purpose of "in-out" allelic exchange, containing *sacB* and a resistance cassette for the antibiotics ampicillin, chloramphenicol, and tetracycline. The full "in-out" allelic exchange procedure is represented graphically in Figure 4, adapted from Marx (2008). Two different strategies were used in different attempts to insert the knockout plasmid (pAB274) into *M. extorquens* CM4: triparental mating and electroporation.

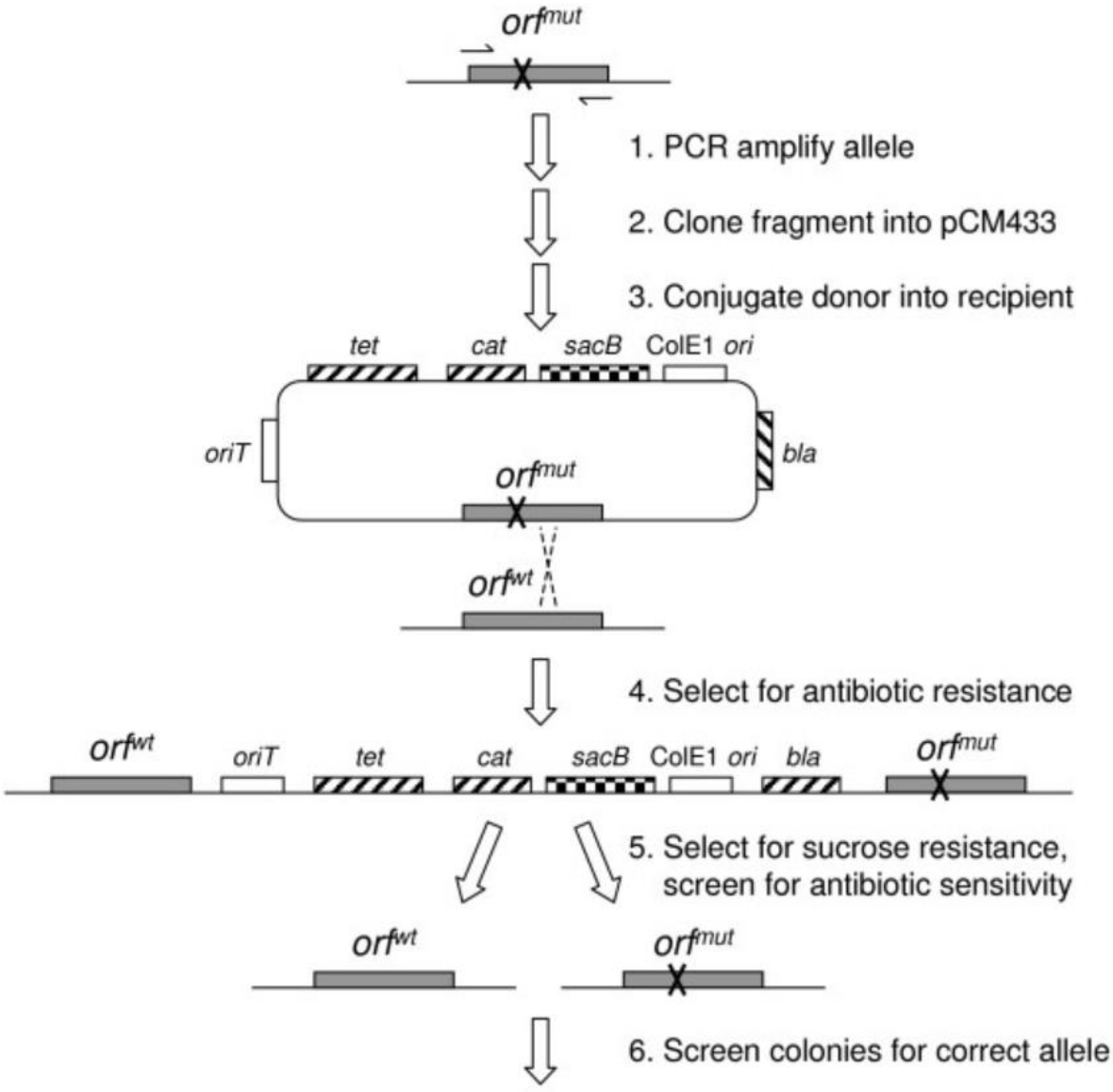

Figure 4: Graphical representation of "in-out" allelic exchange using pCM433, as is used here to generate *tetR* knockouts to derepress *sqhC* transcription (adapted from Marx 2008)

Triparental Mating

Triparental mating transfers a plasmid from one bacterium into another using a "helper strain" capable of conjugation and DNA transfer. The helper strain first accepts the plasmid from the donor strain and then transfers it into the target strain. In this case, the donor strain was *E. coli* containing pAB274, the helper strain was *E. coli* containing pRK2073, and the target strain was *M. extorquens* CM4. Each were first grown overnight in shaking incubators, the bacteria

containing pAB274 in 10 ml Luria broth (LB) and 10 µl tetracycline at 37ºC, the bacteria containing pRK2073 in 10 ml LB and 10 µl streptomycin at 37ºC, and CM4 in 10 ml hypho and 200 µl succinate at 30ºC. These were then centrifuged, their supernatants were poured off, and roughly equal amounts of bacteria were resuspended, combined, and dispensed together, without spreading, on nutrient agar. This plate incubated on the benchtop overnight before all of its contents were suspended in molecular grade water and used in selective plating, as explored below.

Electroporation

Electroporation forces bacteria to take up a plasmid by electrically shocking the bacteria, disrupting the cellular envelope so that the plasmid can enter. This requires treating the bacteria to make them "electrocompetent" so that they are not killed by the electricity. For this experiment, electrocompetent *M. extorquens* CM4 was previously prepared by Jeremy Pomerantz. However, the knockout plasmid, pAB274, needed to be extracted from *E. coli*. To this end, *E. coli* containing pAB274 was grown in 10 ml LB and 10 µl each chloramphenicol and tetracycline in a shaking incubator set to 30ºC for three days. It was then miniprepped using Promega's PureYield Plasmid Miniprep System, resulting in 117 ng/µl pAB274 DNA. On ice, 1 µl pAB274 was pipette-mixed with 40 µl electrocompetent *M. extorquens* CM4, which had been thawed on ice. The mixture was then transferred to a pre-chilled electroporation cuvette with a 0.1 cm electrode gap. The cuvette was wiped clean and inserted into the electroporation device, then pulsed at 1.8 kV for 3.70 ms (the Ec1 setting). The cuvette was removed and 1 ml SOC medium was immediately added. Then, the mixture was transferred to a 1.5 ml tube and incubated, shaking, at 30ºC for 1 hour. This was then plated on 25 ml agar from a batch of 300 ml hypho + 1410 µl 20% succinate + 6 g agar + 300 µl tetracycline (HSTet plates) and incubated at 30ºC until colonies grew for selective plating.

Selective Plating

There were many iterations of this process from many attempts at triparental mating and electroporation. What follows is just one example of timing and number of colonies picked, but the general procedure remained the same. Four days after electroporation, 16 colonies were picked from the electroporation plates and restreaked on new HSTet plates. 19 days later, once enough colonies had grown large enough, 14 colonies from these plates were picked and restreaked onto new HSTet plates once again. After 13 more days, 2 colonies from these plates were picked and grown in flasks containing 10 ml hypho and 200 µl 20% succinate for 4.5 hours. The flasks' contents were centrifuged at 5000 rpm for 10 minutes, the supernatant was poured off, and the cell pellets were resuspended in 200 µl molecular grade water. These were plated on hypho agar with 5% by volume sucrose and 0.1% by volume succinate. After three days, 24 colonies were picked from these plates and restreaked on both HS and HSTet plates.

Knockout Confirmation

Three days after plating on HS and HSTet, 22 colonies that grew on HS alone were picked with autoclaved toothpicks, swirled into 50 µl molecular grade water, and processed at 98ºC for 10 minutes in order to extract the DNA. Polymerase chain reaction (PCR) was run on these samples as in Table 1 below, amplifying the region around the *tetR* gene that was intended to be knocked out using primers previously generated by Alexander S. Bradley. PCR products were subjected to gel electrophoresis for 15 minutes at 90 V on a 1.5% agarose gel (0.45 g agarose in 30 µl TAE buffer and 3 µl SYBR safe stain) in TAE buffer. 25 µl of PCR product were run with 4 µl BioLabs loading dye purple (6X) for all samples, and those which could have been knockouts were run the same way again using the remaining 25 µl of PCR product, alongside a wild-type control. The second resulting gel is pictured in Figure 5 below. The wild-type PCR product is expected to be ~800 bp long, and the knockout ~200 bp. Lane I appeared to

be a knockout. Its bands were cut out of both gels and stored in the same 1.5 ml tube. These gel bands were dissolved and cleaned using the Invitrogen PureLink PCR Purification kit, resulting in 0.5 ng/µl DNA. PCR was run on this DNA using the same procedure and primers as on the DNA extracted directly from the bacteria, resulting in 10.3 ng/µl DNA. This DNA was sequenced by Eurofins and found not to be the knockout. However, sequencing results were not obtained until after Strain I's lipids were extracted for comparison to those of wild-type *M. extorquens* CM4, as though it were the *tetR* knockout.

	Master Mix	H2O (uL)	GC buffer (uL)	DMSO (uL)	Betaine (uL)	dNTPs (uL)	Total volume
	Per Reaction	8.75	5	2.5	2	0.5	20
Reaction #	16	140	80	40	32	8	300

	PCR Reaction	f primer (uL)	r primer (uL)	Template DNA (uL)	Master Mix (uL)	Phusion
	Per Tube	0.25	0.25	0.5	18.75	0.25

Thermocycler settings

Step	T deg C	Time (min:sec)	Note
Denaturation	98	0:30	
Denaturation	98	0:10	
Annealing	58	0:20	
Extension	72	0:30	
Final Extension	72	8:00	
Hold	4	---	

Template DNA	boil preps
Pos Ctrl DNA	lambda

Using Rowan program pcr4

Primer Sets	**Forward**	**Reverse**	**PCR Product Name**
Biotinylated:	AB-orf85_F2_I	orf85_R2	AB_orf85_biotin
	AB-orf86_F2_I	orf86_R2	AB_orf86_biotin
	AB-orf88_F2_I	orf88_R2	AB_orf88_biotin
Non-biotinylated:	orf85_F	orf85_R	AB_orf85
	orf86_F	orf86_R	AB_orf86
	orf88_F	orf88_R	AB_orf88
Expression:	orf85_F3_expr	orf85_R3_express	AB_orf85_express
	orf86_F3_expr	orf86_R3_express	AB_orf85_express
TetR k/o check:	TetR_check_F	TetR_check_R	AB_TetR_check
TOTAL:			

Gradient

Lane	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**10**	**#**	**#**
T deg C										
Reaction id										
Reaction id				I - X (minus T and W)						
Reaction id										
Reaction id										
Reaction id										
Reaction id										
Reaction id										
Reaction id										

Table 1: Master Mix and PCR program used to amplify the region that would contain *tetR* in order to identify knockouts

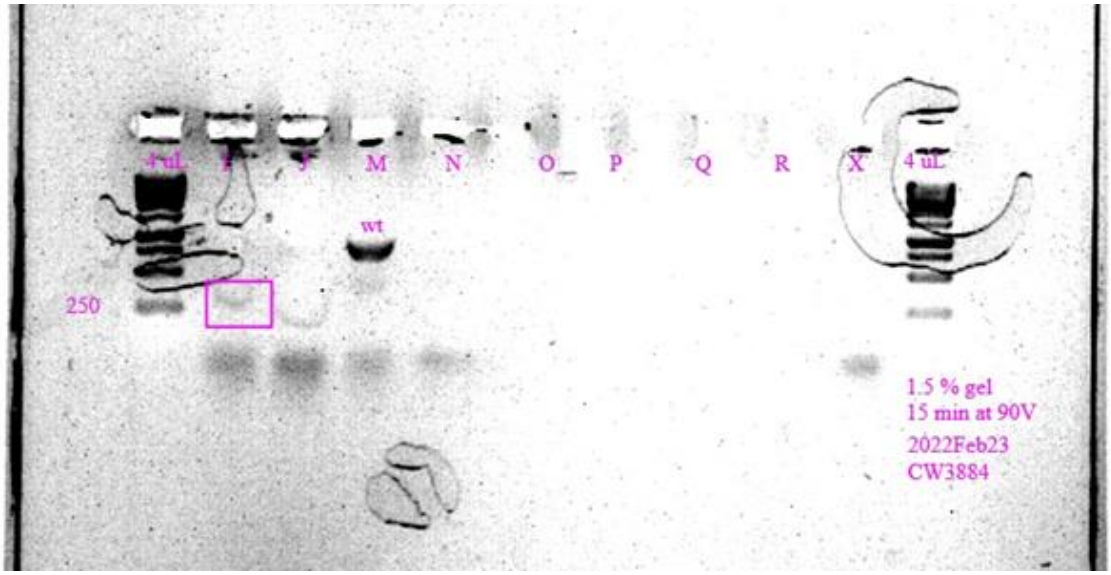

Figure 5: Gel electrophoresis results for select potential knockouts. The boxed band in Lane I is the right length for a *tetR* knockout, so it was cut out and the associated bacteria were treated as a successful knockout and renamed RB01. The band labeled wt in Lane M is the right length for a wild-type CM4 which still contains *tetR*.

Lipid Extraction and Comparison

This procedure was intended to be performed on the *tetR* knockout, but this was never done successfully. However, it was performed on Strain I under the impression that this was the knockout. Strain I was grown in separate flasks containing 10 ml hypho and either 100 µl 20% succinate, 100 µl 2.96 M methylamine, or 100 µl 2.96 M methanol in an incubating shaker set to 30°C for four days, then 25 µl from each were transferred to flasks containing 10 ml hypho and either 200 µl 20% succinate, 100 µl 2.96 M methylamine, or 100 µl 2.96 M methanol, respectively. Wild-type CM4 from freezer stock was grown in flasks containing the same. Both grew for three days in a 30ºC incubating shaker.

125 µl from each flask were reserved for growth curve generation, discussed below. The remainder of each flask was centrifuged 5 min at 5000 rpm and the supernatant was poured off. The cell pellet was resuspended in the remaining supernatant and Bligh-Dyer lipid extraction was performed as follows. In 4 ml glass vials, 100 µl chromasolv water, 250 µl chloroform, and 500 µl methanol were added to the cell pellets. The vials were sonicated for 15 min to break up

cellular material. 250 µl each water and chloroform were added to each vial, mixed by swirling, and allowed to separate overnight at -20ºC. The bottom layer of chloroform and extracted lipids was extracted into fresh 4 ml vials. Another 250 µl chloroform was added to the remaining material and the chloroform layer again extracted to the new vial. This was repeated one more time. Since some organic material remained, each chloroform sample was run through a filtering 6 in Pasteur pipet containing glass wool plugs and combusted sand, rinsed with pure chloroform. Each sample was then dried under nitrogen gas and resuspended in 500 µl dichloromethane (DCM).

50 µl of the samples in DCM were transferred to gas chromatography vial inserts. 15 µl pyridine and 20 µl N,O-Bis(trimethylsilyl)trifluoroacetamide (BSTFA) were added to each sample, all were dried under nitrogen gas, and all were run on the Agilent gas chromatograph-mass spectrometer (GC-MS) using the ABPlantwax5_50M program previously established by the Bradley Lab for use on bacterial lipids.

The remainder of each sample in DCM was subjected to bacteriohopanepolyol cleavage in order to examine the side chains of hopanoids alone. The samples were dried under nitrogen gas. 1 ml of periodic acid solution (100 mg periodic acid per 1 ml 8:1 tetrahydrofuran:water) was added to each sample, and the solutions were left to sit for an hour. 1 ml dichloromethane (DCM)-extracted deionized (DI) water was then added to each sample. 1 ml 4:1 diethyl ether:hexane was added to each sample, then the top (organic) layer was extracted into a 15-ml vial. This was repeated three more times. Anhydrous Na_2SO_4 was added to the organics from each sample to remove water. 3 ml sodium borohydride solution (100 mg $NaBH_4$ in 3 ml methanol) were added to each sample and the solution was allowed to sit for 4 hours at room temperature. 1.5 ml potassium phosphate solution (100 mM in DCM-extracted DI water) were

added to each sample. 1 ml 4:1 hexane:diethyl ether was added to each sample, and the top (organic) layer was extracted into 12 ml vials. This step was repeated two more times. The samples were dried under nitrogen gas, then transferred to GC-MS microvials twice using 100 µl hexane. This was then evaporated under nitrogen gas and 15 µl pyridine and 20 µl acetic anhydride were added to each, and all were run on the Agilent gas chromatograph-mass spectrometer (GC-MS) using the ABPlantwax5_50M program.

Growth Curve Generation

The 125 µl reserved from each flask of CM4 and Strain I on each substrate were reinoculated into flasks containing 10 ml hypho and 25 µl of their respective substrates (20% succinate, 2.96 M methylamine, or 2.96 M methanol). Blanks of 10 ml pure hypho and 10 ml hypho with 25 µl substrate were also prepared. 640 µl from each flask were added to various wells in a 48-well plate as in Table 2 below, which was generated randomly. "RB01" refers to strain I, which was believed to be the *tetR* knockout when growth curves were generated. The plate was loaded into the Biotech Epoch 2 plate reader and run, agitating, at 30ºC for 50 hours, and the optical density at a wavelength of 600 nm (OD_{600}) of each well was read every 15 minutes to construct growth curves.

	1	2	3	4	5	6	7	8
A	Hypho	Hypho	Hypho	Hypho	Hypho	Hypho	Hypho	Hypho
B	Hypho	RB01_M	CM4_S	RB01_Y	CM4_M	CM4_S	RB01_S	Hypho
C	Hypho	CM4_S	RB01_M	RB01_Y	CM4_Y	Hypho	RB01_M	Hypho
D	Hypho	CM4_Y	CM4_M	CM4_Y	CM4_Y	RB01_S	RB01_S	Hypho
E	Hypho	RB01_Y	RB01_Y	RB01_S	CM4_M	RB01_M	CM4_M	Hypho
F	HyphoS	HyphoY	HyphoM	Hypho	Hypho	Hypho	Hypho	Hypho

Table 2: The array of samples and controls incubated on the plate reader for growth curve generation. M refers to cultures grown on methanol, S to cultures grown on succinate, and Y to cultures grown on methylamine.

Heterologous Expression of *sqhC1* and *sqhC2* from CM4 in CM3945

In order to compare the lipid products of CM4's SqhC1 and SqhC2, their genes will be expressed individually in CM3945, an altered form of the strain *M. extorquens* PA1 (which is

closely related to CM4) that has had its native *sqhC* gene removed. To do this, the *sqhC1* and *sqhC2* genes of CM4 need to be made into plasmids, these plasmids replicated in *E. coli*, the genes (from plasmid form) ligated into inducible expression plasmids, the new expression plasmids grown in *E. coli* again, and the expression plasmids transferred into CM3945 using electroporation or tri-parental mating. At the time of writing, *sqhC1* and *sqhC2* may be in expression plasmids in *E. coli*, but the organisms are growing less robustly than expected.

CM4 was grown from freezer stock in flasks of 10 ml hypho and 47 µl 20% succinate for two days, then its whole genome was isolated using the Promega Wizard Genomic DNA Purification kit. PCR was performed as in Table 3 below using primers generated by Alexander Bradley to isolate *sqhC1* and *sqhC2*. The amplified DNA was run on 1% agarose gel in TAE buffer for 50 minutes at 90 V, resulting in Figure 6 below. 85A-C represent *sqhC1* and should be 2004 bp, and 86A-C represent *sqhC2* and should be 1938 bp. All are the expected length, so the remaining PCR product was cleaned using the Promega Wizard SV Gel and PCR Clean-Up System, quantified, and those which were most abundant (85C at 99.7 ng/µl and 86C at 135 ng/µl) were sequenced by Eurofins. The 85C and 86C sequences, available in Supplemental Material, align to *sqhC1* and *sqhC2*, respectively, in CM4, as desired. Therefore, they were used moving forward.

	Master Mix	H2O (uL)	GC buffer (uL)	DMSO (uL)	Betaine (uL)	dNTPs (uL)	Total volume
	Per Reaction	3.3	2	1	2	0.2	8.5
Reaction #	13	42.9	26	13	26	2.6	110.5

PCR Reaction	f primer (uL)	r primer (uL)	Template DNA (uL)	Master Mix (uL)	Phusion
Per Tube	0.5	0.5	0.3	8.5	0.2

Thermocycler settings

Step	T deg C	Time (min:sec)	Note
Denaturation	98	0:30	
Denaturation	98	0:10	
Annealing	72	0:20	Gradient details below
Extension	72	0:45	
Final Extension	72	8:00	
Hold	4	---	

Template DNA	CM4
Pos Ctrl DNA	Lambda

Primer Sets	Forward	Reverse	PCR Product Name
Biotinylated:	AB-orf85_F2_biotin	orf85_R2	AB_orf85_biotin
	AB-orf86_F2_biotin	orf86_R2	AB_orf86_biotin
	AB-orf88_F2_biotin	orf88_R2	AB_orf88_biotin
Non-biotinylated:	orf85_F	orf85_R	AB_orf85
	orf86_F	orf86_R	AB_orf86
	orf88_F	orf88_R	AB_orf88
Expression:	orf85_F3_express	orf85_R3_express	AB_orf85_express
	orf86_F3_express	orf86_R3_express	AB_orf85_express
TOTAL:			

Gradient

Lane	1	2	3	4	5	6	7	8	9	10	11	12
T deg C	65	65.2	65.6	66.3	67.2	68.1	68.9	69.8	70.7	71.4	71.8	72
Reaction id												
Reaction id												
Reaction id												
Reaction id								AB_orf85_A		AB_orf85_B		AB_orf85_C
Reaction id								AB_orf86_A		AB_orf86_B		AB_orf86_C
Reaction id								AB_orf88_A		AB_orf88_B		AB_orf88_C
Reaction id	Pos Ctrl											
Reaction id	Neg Ctrl											

Table 3: Master Mix and PCR program used to amplify the regions containing *sqhC1* and *sqhC2*.

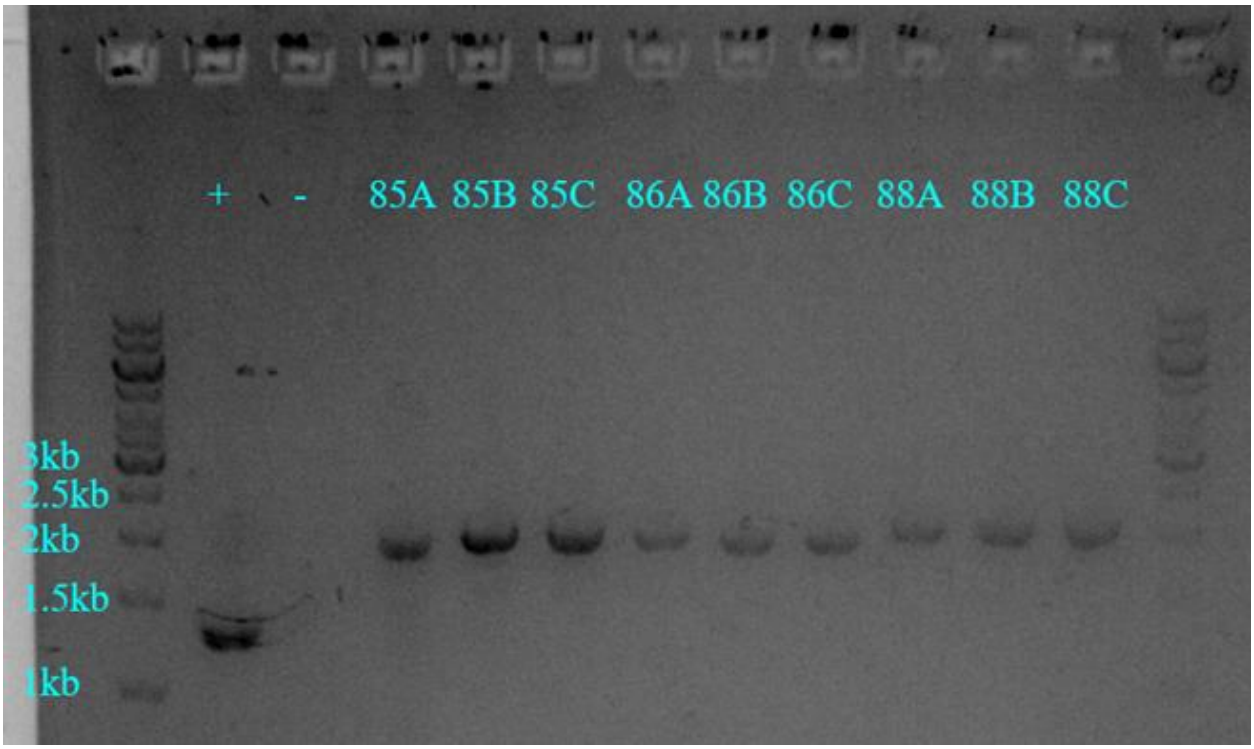

Figure 6: Gel electrophoresis results showing 85A-C (*sqhC1*) and 86A-C (*sqhc2*) both at roughly 2kb, as they should be, confirming that the correct regions have been amplified.

4 µl of 85C and 86C PCR products were used with the Invitrogen Zero Blunt TOPO PCR Cloning Kit to make pAB269 and pAB270, plasmid forms of *sqhC1* and *sqhC2*, respectively. These plasmids were each transformed into Top 10 competent *E. coli* by mixing 5 µl of rehydrated plasmids with tubes of Top 10, letting the mixtures sit on ice for 30 minutes, heat-shocking them in a 42ºC water bath for 30 seconds, adding 250 µl SOC medium to each on an ice bath, shaking and incubating the mixtures for 60 minutes at 30ºC, and plating 25 µl, 50 µl, 100 µl, and the balance of the mixtures on separate plates of LB agar with kanamycin. These plates were allowed to grow for three days, then three colonies each of *E. coli* containing pAB269 and *E. coli* containing pAB270 were picked and added to separate flasks of 10 ml LB and 10 µl kanamycin. These were incubated, shaking, at 37ºC for one day. 200 µl from these flasks were reserved for freezer stocks, and plasmids were isolated from the remainder of each flask using the Promega PureYield Plasmid Miniprep System. The pAB269 and pAB270

plasmids were sequenced using primers for *sqhC1* and *sqhC2*, respectively, and confirmed to contain their respective genes.

E. coli containing pLC290, a plasmid used for inducible expression of genes, was grown from freezer stock overnight in a flask of 10 ml LB and 10 µl kanamycin. pLC290 was isolated from this flask using the Promega PureYield Plasmid Miniprep System. pAB269 and some of pLC290 were digested with KpnI-HF and BglII restriction enzymes according to New England Biolabs' guidelines, and pAB70 and some more of pLC290 were digested with KpnI-HF and EcoRI-HF restriction enzymes according to New England Biolabs' guidelines. These were run on 1.2% agarose gel in TAE buffer for 50 minutes at 90 V, resulting in Figure 7 and Figure 8 below. Digested pAB269 and pAB270 were expected to be ~2 kb, and both pLC290 digests were expected to be ~8 kb. These are visible for all but pAB269, but all of these regions were cut out of the band anyway.

The gel bands were all cleaned using the Thermo PureLink PCR Purification Kit, and all were confirmed to contain DNA using the NanoDrop (73.9 ng/µl pAB269, 84.6 ng/µl pAB270, 90.4 ng/µl pLC290 cut for pAB269, and 97.1 ng/µl pLC290 cut for pAB270). The pairs cut with the same restriction enzymes (pAB269 and pLC290, pAB270 and pLC290) were ligated using the Thermo Rapid Ligation Kit and transformed into 5α competent *E. coli* using the same procedure outlined previously for the transformation of pAB269 and pAB270 into Top 10 competent *E. coli*. Several colonies have been picked from these plates and grown in 10 ml LB and 10 µl kanamycin, both at 37ºC and 30ºC, but none have grown sufficiently, even after several days, which is extremely unusual for *E. coli*.

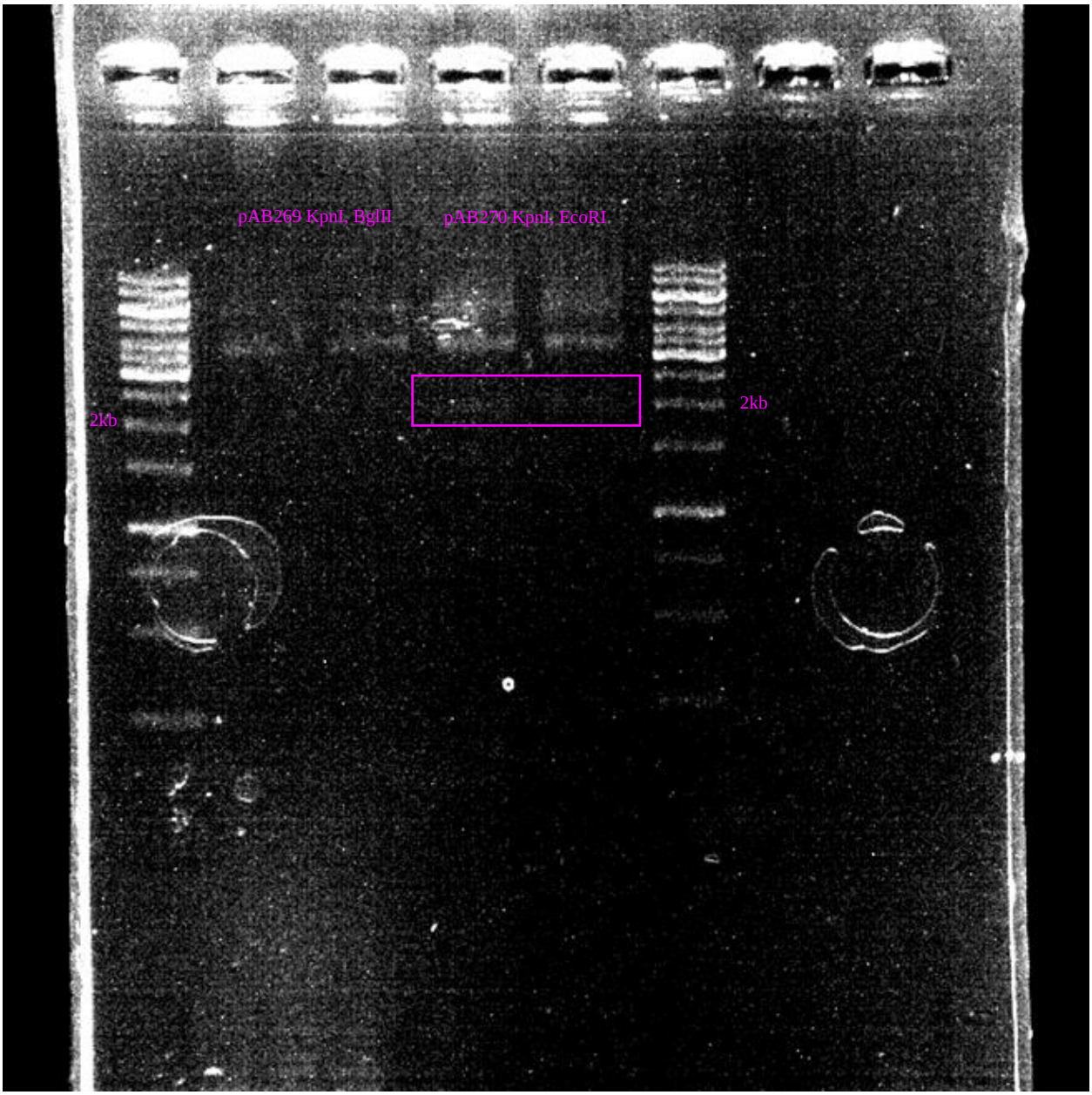

Figure 7: Gel electrophoresis results for pAB269 digested with KpnI and BglII and pAB270 digested with KpnI and EcoRI. Bands are visible at 2 kb for pAB270 (boxed), but not for pAB269. Both regions were excised anyway.

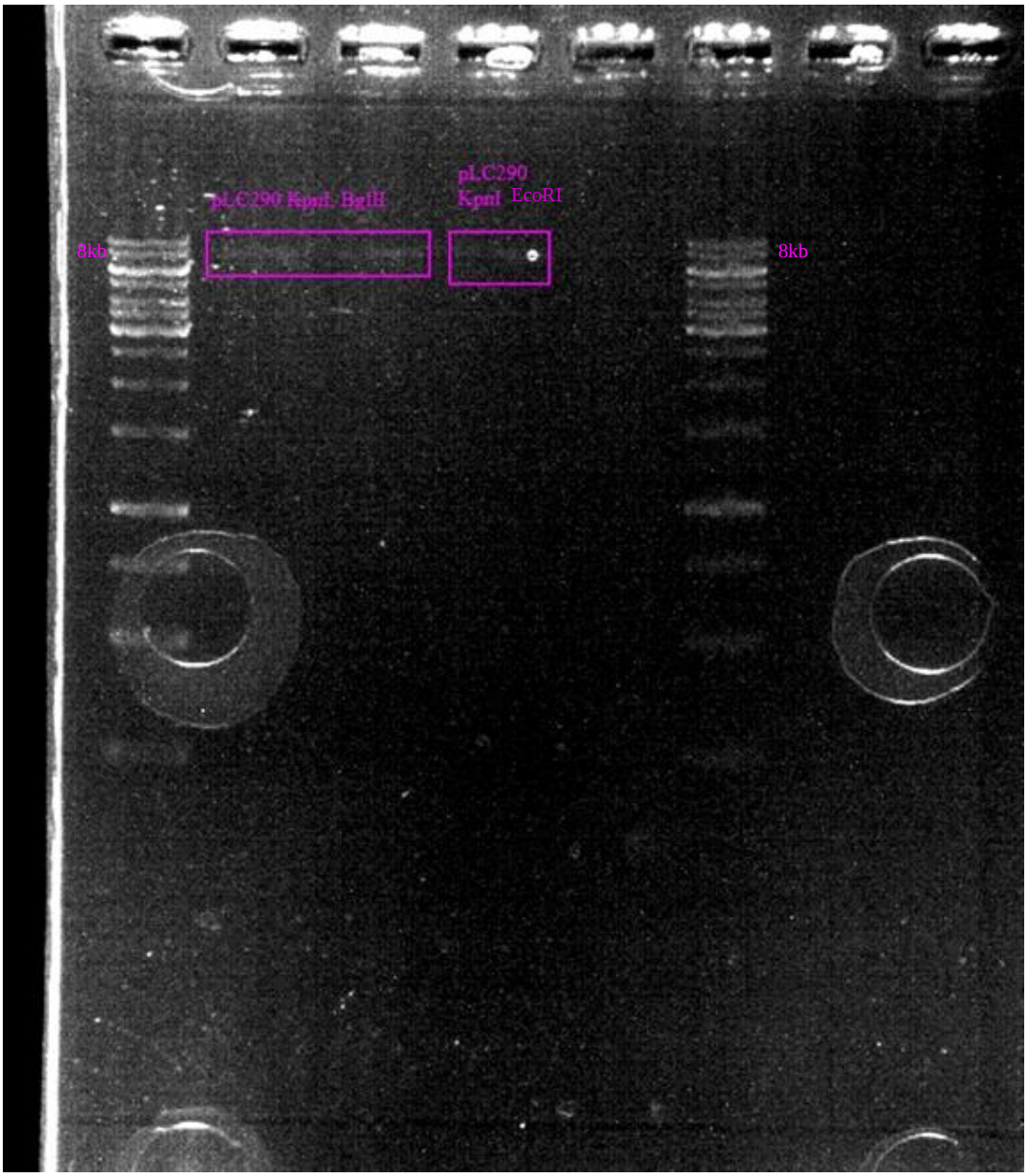

Figure 8: Gel electrophoresis results for pLC290 digested with KpnI and BglII and pLC290 digested with KpnI and EcoRI. Bands are visible at 8 kb for both (boxed) and both were excised.

sqhC1 and *sqhC2* Knockout Plasmid Generation

Another attempt to compare the products of CM4's *sqhC1* and *sqhC2* was made by knocking out *sqhC1* and *sqhC2* individually in CM4. "In-out" allelic exchange, outlined above for the *tetR* knockout, is also used in this case. However, knockout plasmids must first be generated from scratch for this experiment. Synthetic plasmids containing the knockout

sequences for *sqhC1* and *sqhC2* (pAB275 and pAB276, respectively) were ordered from GenScript, but these sequences must be moved into pCM433 for "in-out" allelic exchange. To this end, pAB275 and pAB276 were transformed into 5α competent *E. coli* by mixing 5 µl of rehydrated plasmids with tubes of 5α, letting the mixtures sit on ice for 30 minutes, heat-shocking them by putting the tubes in a 42ºC water bath for 30 seconds, adding 250 µl SOC medium to each on an ice bath, shaking and incubating the mixtures for 60 minutes at 30ºC, and plating 25 µl, 50 µl, 100 µl, and the balance of the mixtures on separate plates of LB agar with kanamycin. These plates were allowed to grow for five days, then a single colony for each plasmid was inoculated into flasks containing 10 ml LB and 10 µl kanamycin, which were incubated, shaking, for one day at 30ºC. Freezer stocks were made of these by centrifuging the flasks' contents for 10 minutes at 5000 rpm, pouring off the supernatant, resuspending the pellets in 0.5 ml 8% DMSO in hypho, and freezing 0.5 ml of each at -80ºC.

Each of the freezer stocks of pAB275 and pAB276 in *E. coli* were grown overnight at 30ºC in a shaking incubator in flasks containing 10 ml LB and 10 µl kanamycin. pCM433, the plasmid generated by Christopher Marx for "in-out" allelic exchange, also in *E. coli*, was also grown overnight from freezer stock in a shaking incubator in a flask containing 10 ml LB and 10 µl each ampicillin, chloramphenicol, and tetracycline. All were miniprepped using Promega's PureYield Plasmid Miniprep System, resulting in 219.7 ng/µl pAB275, 254.2 ng/µl pAB276, and 528.7 ng/µl pCM433. It is intended to cut all of these plasmids with AatII and NotI, then ligate both pAB275 and pAB276 with pCM433 to form the knockout plasmids pAB277 and pAB278, respectively, but existing attempts have failed to generate sufficient amounts of cut plasmids. However, once pAB277 and pAB278 have been generated, they will be used to knock out *sqhC1*

and *sqhC2*, respectively, using the procedure previously explored for knocking out *tetR*, and to compare the lipid profiles of each to that of wild-type CM4.

Results

Phylogenetic Trees

The phylogenetic tree of both *sqhC1* and *sqhC2* genes is pictured in Figure 9 below. The full names of each strain present in this tree can be found in Supplemental Material. *sqhC1* genes are shown in orange, and *sqhC2* genes are shown in blue. Each *sqhC2* gene has a *tetR* gene immediately upstream. However, one strain, *Bradyrhizobium elkanii* USDA 76, has two copies of *sqhC*, shown in pink, neither of which have an upstream *tetR*. *sqhC* genes in strains that only have one copy of *sqhC* are shown in black. Most *sqhC2* genes form five clades, with the exception of that of *Syntrophobacter fumaroxidans* MPOB (which is the only *Syntrophobacter* strain examined that has an *sqhC2* gene). One clade contains the two *Methylobacterium* strains examined. Another contains only all of the *Frankia* strains examined, another only all of the *Zymomonas* strains, and another only all of the *Cupriavidus* strains. The remaining clade contains both all of the *Geobacter* and *Pelobacter* strains, but no others. The *Cupriavidus* and *Geobacter/Pelobacter* clades, seen on the bottom right in Figure 9, are separated by only one branch point, suggesting that they may be closely related to one another. *sqhC1* genes form clades within their genera as well, but these clades also contain the solo *sqhC* genes within the genera. Altogether, this suggests that several genera have independently evolved *sqhC2* genes accompanied by *tetR* which are orthologous to one another. This, in turn, suggests that this *sqhC2-tetR* system may hold an evolutionary advantage, in order for it to have evolved several times independently and been retained each time.

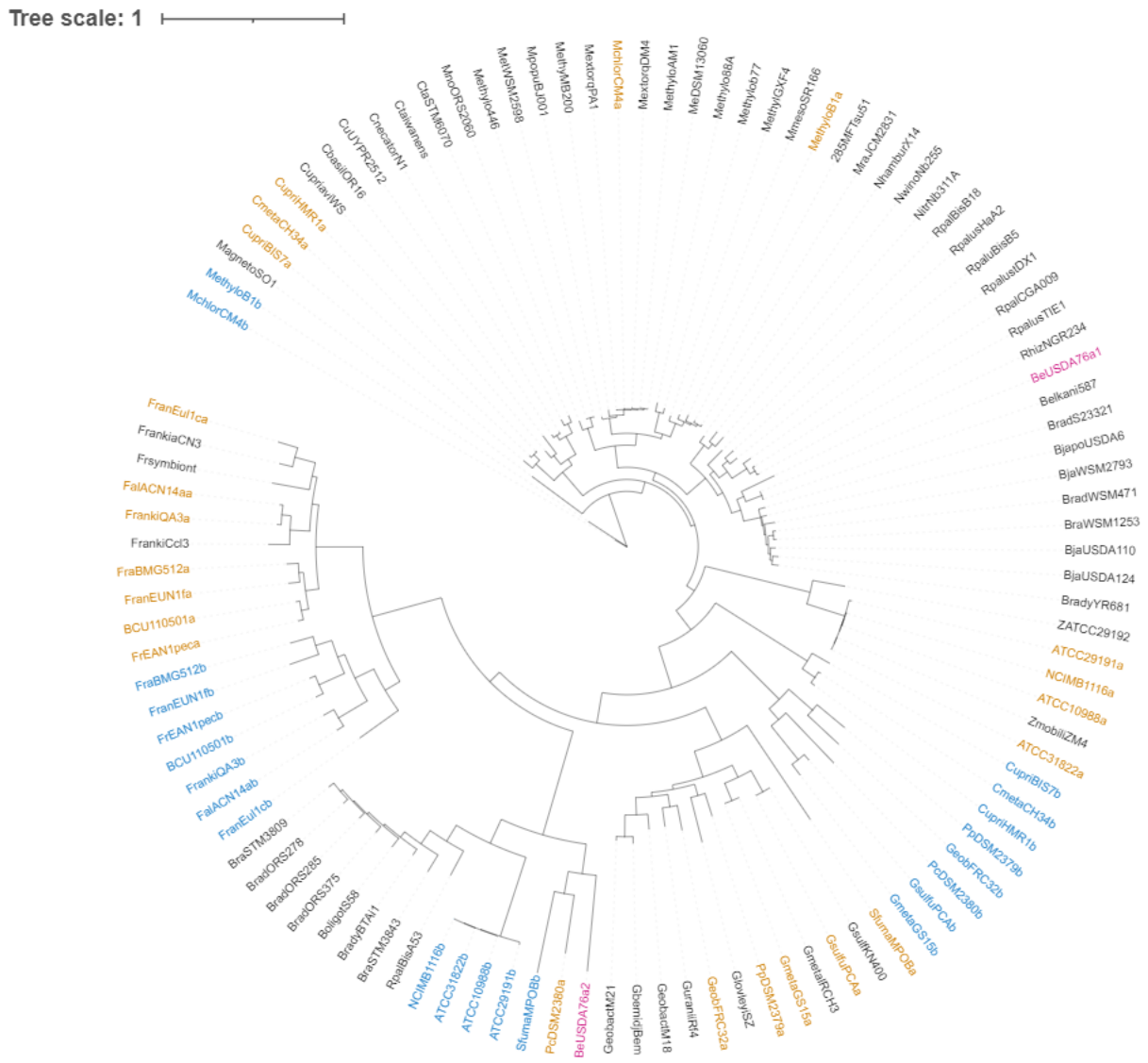

Figure 9: A phylogenetic tree for all *sqhC* genes gathered. *sqhC1* genes are shown in orange, and *sqhC2* genes are shown in blue. Each *sqhC2* gene has a *tetR* gene immediately upstream. *Bradyrhizobium elkanii* USDA 76, has two copies of *sqhC*, shown in pink, neither of which have an upstream *tetR*. *sqhC* genes in strains that only have one copy of sqhC are shown in black

Phylogenetic trees were also constructed of *sqhC2* and its upstream *tetR*, shown in Figure 10 and Figure 11 below, respectively. The same abbreviations found in Supplemental Material are used in these trees. On visual inspection, these trees appear similar, with genera again forming clades in both, but they are not identical. For quantitative comparison, a weighted Robinson-Foulds (RF) metric was calculated to compare the two trees. It is 3.7645. Given that the RF value for two randomly-generated 22-leaf trees is 21.9959, this was taken as evidence

that there is, indeed, significant similarity between the *sqhC2* and *tetR* trees. This suggests that *sqhC2* and its upstream *tetR* coevolved, supporting the proposition that this *tetR* gene is involved in the regulation of *sqhC2*.

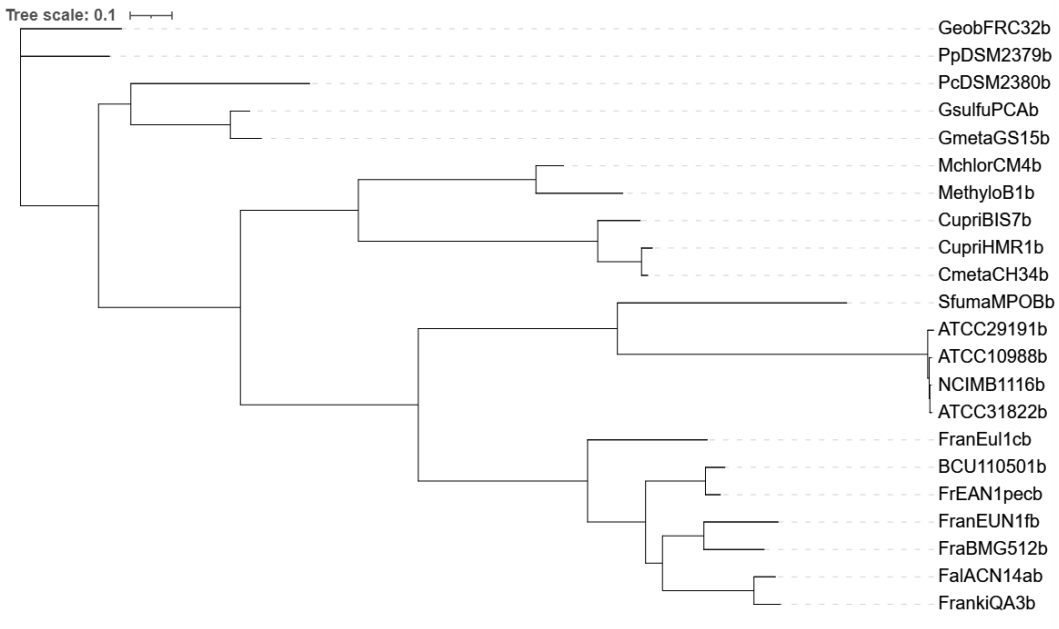

Figure 10: A phylogenetic tree of *sqhC2* genes. Similar genera-based clusters are seen here as in the phylogenetic tree of all of the *sqhC* genes.

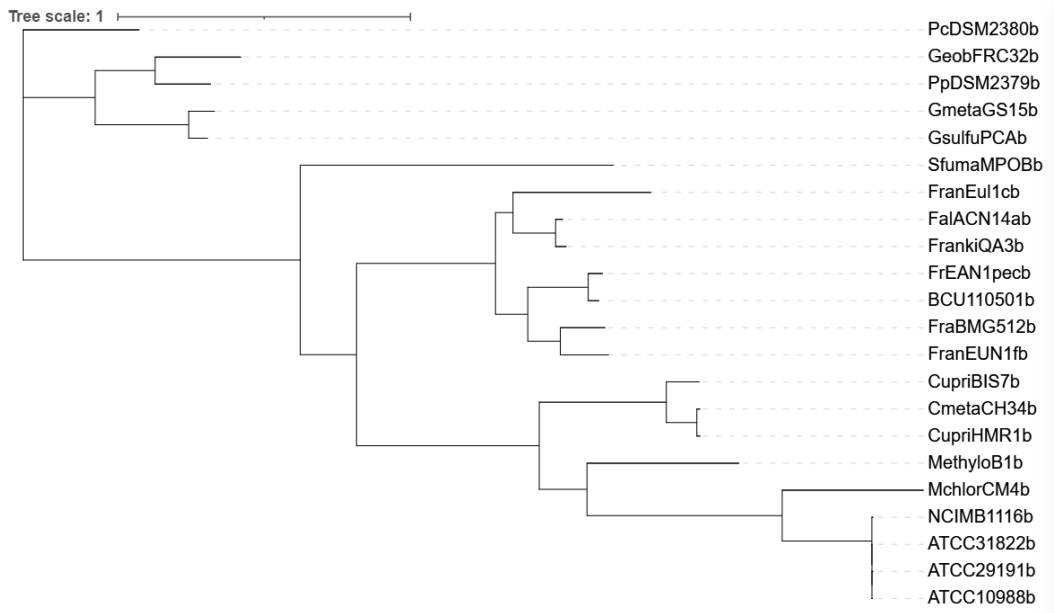

Figure 11: A phylogenetic tree of *tetR* genes associated with *sqhC2*. Similar genera-based clusters are seen here as in the *sqhC* phylogenetic trees, but this tree is not identical to the *sqhC2* tree.

tetR Knockout Failure

Despite nine months of effort, all attempts to knock out the *tetR* gene in *M. extorquens* CM4 were met with failure. There are several possible reasons for this. One possibility is that the regulation TetR provides to *sqhC2* is essential. This could be true, especially since *sqhC2* seems to always be accompanied by an upstream *tetR*. However, there are other possibilities. "In-out" allelic exchange, even when successful, ends with both colonies of the mutant and colonies which have reverted to wild-type. Simple bad luck in picking colonies could have resulted in missing the knockout. However, I do not believe this to be the case, based on the number of trials and picked colonies. On a slightly different tack, it is possible that knockout generation failed, not for any grand biological reason, but because these procedures are simply imperfect, and can never be perfect. While the first possibility would be interesting, the last is more likely, and, for this reason, attempts to knock out *tetR* in CM4 will continue.

Discussion

The work presented here is only a fraction of this project, which has been through several cycles of activity since 2014. Concrete results are admittedly sparse here, but the project will continue, and this work will be of use to it. *In silico* phylogenetic analysis has provided evidence to support an evolutionary advantage to the duplicate copy of the *sqhC* gene. It has additionally provided evidence of the coevolution of *sqhC2* and its upstream *tetR*, indicating that this TetR likely does regulate *sqhC2*. The failure to knock out *tetR* in CM4 could indicate that this regulation is itself evolutionary advantageous, or even necessary, but there is not enough evidence to state this conclusively. Existing work presented here can be used in continued attempts to knock out *tetR*. The same is true for attempts at heterologous expression of *sqhC1* and *sqhC2* in CM3945, as well as attempts to knock out these genes in CM4.

Works Cited

Adler, Marlen, Mehreen Anjum, Otto G. Berg, Dan I. Andersson, and Linus Sandegreen. "High Fitness Costs and Instability of Gene Duplications Reduce Rates of Evolution of New Genes by Duplication-Divergence Mechanisms." *Molecular Biology and Evolution* 31, no. 6 (2014): 1526-1535.

Belin, Brittany J., Nicolas Busset, Eric Giraud, Antonio Molinaro, Alba Silipo, and Dianne K. Newman. "Hopanoid Lipids: from Membranes to Plant-Bacteria Interactions." *Nature Reviews Microbiology* 16 (2018): 304-315.

Kulkarni, Gargi, Nicolas Busset, Antonio Molinaro, Daniel Gargani, Clemence Chantreuil, Alba Silipo, Eric Giraud, and Dianne K. Newman. "Specific Hopanoid Classes Differentially Affect Free-Living and Symbiotic States of *Bradyrhizobium diazoefficiens*." *mBio* 6, no. 5 (2015): e01251-15.

Marx, Christopher. "Development of a Broad-Host-Range *sacB*-Based Vector for Unmarked Allelic Exchange." *BMC Research Notes* 1, no. 1 (2008).

Pearson, Ann, Sarah R. Flood Page, Tyler L. Jorgenson, Woodward W. Fischer, and Meytal B. Higgins. "Novel Hopanoid Cyclases from the Environment." *Environmental Microbiology* 9, no. 9 (2007): 2175-2188.

Saenz, James P., Daniel Grosser, Alexander S. Bradley, and Kai Simmons. "Hopanoids as Functional Analogues of Cholesterol in Bacterial Membranes." *PNAS* 112, no. 38 (2015): 11971-11976.