Mapping Conditional Loss of Growth Mutations in *Physcomitrella patens*

By

Leah Pervere

Submitted to the Bioinformatics & Computational Biology and Biochemistry Programs in partial fulfillment of the requirements for the degree of

Bachelor of Science

at

WORCESTER POLYTECHNIC INSTITUTE

April 2015

Certified by ……………………………………………………………………….

Patrick Flaherty

Assistant Professor, Biomedical Engineering Department

Certified by ……………………………………………………………………….

Luis Vidali

Associate Professor, Biology & Biotechnology Department

Mapping Conditional Loss of Growth Mutations in *Physcomitrella patens*

by

Leah Pervere

Submitted to the Bioinformatics & Computational Biology and Biochemistry Programs on April 30, 2015, in partial fulfillment of the requirements for the degree of

Bachelor of Science

ABSTRACT

To identify genes important for growth, we generated conditional mutants that experience loss of growth at high temperature (CLoG mutants). We outcrossed one of these mutants, CLoG4, to a polymorphic strain and sequenced the whole genome of a pool of segregants that maintained temperature sensitivity to map the mutation causing the phenotype. I also performed a rescue experiment to confirm the identity of the phenotype-causing mutation of a second strain, CLoG1, which had already been mapped.

MQP Primary Advisor: Luis Vidali
Title: Associate Professor
Department: Biology & Biotechnology Department

MQP Secondary Advisor: Patrick Flaherty
Title: Assistant Professor
Department: Biomedical Engineering Department

ACKNOWLEDGMENTS

I would like to thank Professors Vidali and Flaherty for their continued guidance and support throughout this project, as well as Xinxin Ding ’14 for her helpful correspondence as I continued her work. I would also like to thank all members of the Vidali laboratory for their support and insightful comments.

TABLE OF CONTENTS

INTRODUCTION

The Importance of Understanding Plant Growth

Some of the biggest problems that face our society today include hunger and sustainable fuel. These two issues, though very different, both show promise of resolution by one seemingly-simple thing: an increase in crop production. Plants such as corn and sugar cane have become a viable source of biofuel (Byrt et al., 2011) and many crops can be genetically engineered to resist insects and herbicides (Christou, 2013), to produce more nutrients (Christou, 2013), to resist drought (Lawlor, 2013), or to be biochemically inclined to produce higher yields (Whitney et al., 2011). As the world's population continues to grow, this type of engineering may be crucial for maintaining an adequate food supply and lessening hunger around the world. Higher yields of biofuel-compatible crops could also aid in solving the energy crisis our world faces today without putting additional strain on the global food supply. However, the genetic engineering necessary to make this a reality is not possible without a deep understanding of plant growth. Only once plant growth is well understood can methods of increasing growth, efficiency, and overall production be developed. Although many of the basic mechanisms of plant growth are already known, there is still much work to be done on identifying the specific genes that are most responsible for the growth and development of plants. Once these genes are identified, researchers can begin to modify them to suit the needs of the world today.

Physcomitrella patens as a Model Organism

A model organism that is very useful for studying plant growth is the moss *Physcomitrella patens*. *P. patens* allows for the study of many growth characteristics of larger vascular plants, starting as early as the gametophyte stage (Prigge and Bezanilla, 2010). As described by Prigge, there are three aspects of *P. patens* that make it amenable to reverse genetics studies, or the study of phenotypes resulting from genetic changes (Alonso and Ecker, 2006): 1) its ability to integrate DNA via homologous recombination, allowing for studies to be performed on targeted genes; 2) the ability to generate protoplasts (moss cells without their cell wall) and the ease with which plants can be propagated at any stage of life; and 3) the relative ease of performing transformations, or the insertion of non-wild-type DNA into moss protoplasts. In addition to these reverse genetics approaches, forward genetic studies are also possible which study the genetic causes of phenotypes observed in the organism (Alonso and Ecker, 2006) by inducing mutations in *P. patens* (Engel, 1968), which is made possible by the fact that *P. patens* is haploid and therefore only has one copy of each chromosome (Cove and Knight, 1993).

Additionally, the *P. patens* genome – derived from the wild-type “Gransden” strain – has been sequenced and is publicly available (JGI, n.d.), allowing for genotyping studies to be conducted much more easily. The moss is also growing in popularity, as is evident from a quick search of the word *Physcomitrella* on Google Scholar; for date range 1995-2005 there are ~2,180 results while there are ~12,000 hits for 2005-2015. This growing popularity means more and more information is being discovered and published about its genes and growth mechanisms. All of these factors combine to make *P. patens* a prime candidate for experiments studying plant growth and development.

The Utility of Conditional Mutants

A primary setback of studying genes vital for growth is the fact that mutations in these genes are oftentimes lethal to the plant. When they are not lethal, they can cause the plant line to be unsustainable (due to conditions such as inhibited growth or impaired fertility), making it difficult to study the line's genotype. One method of avoiding this problem is the production of conditional mutants. When studying plant growth, conditional mutations are ones that cause the plant to grow normally under one set of conditions but abnormally under another. This project revolved around the use of temperature-sensitive (TS) mutants. These mutants are characterized by a temperature threshold that separates the plants' "restrictive" temperature from their "permissive" temperature. At the permissive temperature a normal phenotype is seen, but at the restrictive temperature the phenotype is mutated (Bajaj et al., 2008). Here, the mutant phenotype is seen in plants – lines of *P. patens* whose mutations were generated randomly through ultraviolet (UV) mutagenesis (Ikehata and Ono, 2011). These plants grow normally at the permissive temperature of 25°C but exhibit highly stunted growth at the restrictive temperature of 32°C. By maintaining copies of the mutant cell lines at both their restrictive and permissive temperatures, genotypes relevant to cell growth can be studied without causing death of the line.

Characteristics and Frequencies of Temperature-Sensitive Mutations

Temperature-sensitive mutants have been used in numerous previous studies for a variety of organisms (Suzuki et al., 1967; Esposito and Esposito, 1969; Bajaj et al., 2008; Lockwood et al., 2011). To identify the causal mutation of a TS phenotype, mapping procedures of different types can be used to narrow down the number of possible gene candidates, but rarely can they automatically narrow it down to one. Because of this, oftentimes the remaining candidates must be examined for characteristics common to temperature-sensitive mutations. Previous studies are extremely valuable for identifying what some of these characteristics may be, allowing for the identification of the probable causal mutation. These studies can also provide very useful information about the frequency with which TS mutations occur.

No information is available about the frequency of TS mutations in *P. patens*. However, studies in other organisms have provided statistics that could be useful in estimating this frequency. For example, a study of TS mutations in *Drosophila melanogaster* found that 6.3% of lethal or “semilethal” mutations induced using ethyl methanesulfonate (EMS) were temperature-sensitive (Suzuki et al., 1967). A study in the yeast *Saccharomyces cerevisiae* found that of 896 survivors of UV mutagenesis, 75 showed a temperature-sensitive phenotype for sporulation (Esposito and Esposito, 1969). Finally, a study of an *Escherichia coli* toxin aimed to produce every possible single-site mutation of the 101-residue-long protein (Bajaj et al., 2008). Seventy five percent of these mutants were obtained and of those, 16% were temperature-sensitive. This study also found that 52% of those mutants were the result of mutations at buried sites (Bajaj et al., 2008). This agrees with Lockwood et al.’s finding that “TS mutations tend to occur at buried and rigid residues, and are located at conserved protein domains” (Lockwood et al., 2011). Therefore knowledge of possible mutations, and their effect on protein structure when applicable, can be very useful in identifying the causal mutation of TS phenotypes.

Some studies have used their experimental data to make claims about the number of genes capable of producing temperature-sensitive mutations. A study of bacteriophage T4D in *E. coli* determined that the 382 TS mutations isolated as part of the study could “with reasonable assurance” be assigned to a total of 37 genes (Edgar and Lielausis, 1964). Another study, by Brown et al., isolated nine temperature-sensitive mutants of herpes simplex virus (HSV) which were found to be located on eight separate genes. The authors used this information and a Poisson distribution to estimate that the HSV genome contains more than 30 genes which can result in temperature-sensitive mutations (Brown et al., 1973). If this same type of distribution could be developed for *P. patens*, it would allow the total number of possible temperature-sensitive mutations to be estimated.

Next-Generation Sequencing

The identification of genetic causes for specific phenotypes has become more readily achievable with the advent of cheaper and more efficient sequencing technologies. For decades the sequencing technology of choice was Sanger sequencing, a method developed by Frederick Sanger in 1977 that quickly grew in popularity due its relative efficiency and adaptations that lacked radioactive chemicals (Sanger et al., 1977). Over time, the cost of Sanger sequencing decreased and its efficiency increased thanks to the use of capillary electrophoresis and automated sequencers (Liu et al., 2012). However, Sanger sequencing is limited in the relatively low number of reads that can be processed at one time. New sequencing technologies have come into popularity, known as "next-generation" sequencing, that allow millions of reads to be processed at once with one or two runs of the sequencer (Mardis, 2008). There are four main next-generation sequencing technologies: Illumina/Solexa, ABI/SOLiD, 454/Roche, and Helicos (Morozova and Marra, 2008).

This study utilized Illumina sequencing, which is performed using a method of sequencing-by-synthesis. Single-stranded DNA fragments are attached to a flow cell and nucleotides are added one base at a time. Fluorescent markers are used to analyze which nucleotide is added at each step (Morozova and Marra, 2008), allowing the fragment's sequence to be determined. Compared to other available sequencing technologies, Illumina sequencing has the largest output and the lowest cost (Liu et al., 2012). However, these advantages come at the cost of read length. Illumina only produces reads of 35-175 base pairs, compared to the 454/Roche method's approximate length of 400 base pairs per read and the reads of 900-plus base pairs achieved through traditional Sanger sequencing (Reis-Filho, 2009). With a genome sequence comprised of large numbers of short reads, it very quickly becomes necessary to develop computational techniques capable of making sense of the sequencing data and to identify regions of interest, such as the identification of single nucleotide polymorphisms (SNPs).

Mapping Strategies

With the development of these sequencing technologies and the ever-increasing presence of "big data," mapping strategies have been developed to make sense of whole-genome sequencing reads. These strategies all share a common basic procedure of aligning reads together to form a genome-long sequence, followed by the identification of SNPs and mutations of interest. However, this procedure presents some limitations which researchers have worked to overcome. This has resulted in the development of various mapping strategies, each of which has its pros and cons. All of these strategies can be used to "map" the causal mutation (to identify the mutation causing the phenotype in an organisms' population), even if some may work better for one organism than others.

One of the main limitations of whole-genome sequencing that researchers have aimed to overcome is a lack of accuracy. This is primarily caused by the short read lengths characteristic of next-generation sequencing when compared to traditional Sanger technologies. One strategy, described by Hillier et al., aimed to improve the accuracy of read placement given the fact that shorter reads are more likely to have multiple possible placements in the genome than longer reads. They did this by developing a computational technique to identify reads of the *C. elegans* genome that may be susceptible to having multiple possible placements (Hillier et al., 2008).

Accuracy of read placement can also suffer for fragments of DNA where indels (insertions or deletions) are present. When developing a mapping strategy for human genome sequences, DePristo et al. aimed to improve this accuracy and also to improve the accuracy of distinguishing between actual variations in the sequence and sequencing errors. They did this through the development of a new computational framework, which they demonstrated on human DNA sequences (DePristo et al., 2011).

A second limitation of these sequencing-based studies is the size of the population of crossed organisms (the "mapping population") that one must use to reliably sequence the phenotype's causal mutation. As many previous studies had used mapping populations of 1000 organisms or more, Austin et al. developed a new mapping strategy to decrease this to a more manageable population of 50 *Arabidopsis thaliana* plants or less (the authors claim as few as 10 segregants can be used, but recommend using at least 50) (Austin et al., 2011). Other studies have aimed to optimize the mapping process overall. One study by Doitsidou et al. took previously existing strategies and combined them to decrease the overall time of mutation mapping, from the experimental isolation of mutant organisms to the identification of their underlying genetic cause (Doitsidou et al., 2010). Another study aimed to develop suggestions for the mapping population and amount of sequencing coverage that should be used in these experiments. Simulations of these mapping experiments in outcrossed *Arabidopsis thaliana* plants revealed an optimal sequencing coverage of 25x for a mapping population of 100 or more to narrow the number of candidate causal mutations down to five (James et al., 2013). This study also determined paired-end sequencing to be more informative than single-ended sequencing. This information and the simulation strategy used in their study both aided in the development of the sequencing strategy used for this project, as vascular plants share many common characteristics with *P. patens*.

A strategy somewhat similar to ours is described by Zuryn et al., in which organisms were treated with EMS to produce linked mutations. The mutants were then crossed with un-mutagenized organisms, and only ones demonstrating the phenotype of interest were retained so as to eliminate mutations that were not responsible for the phenotype. Thus the location of the causal mutation could be identified by looking for clusters of EMS-induced variations in the crossed organisms' pooled DNA sequence. After this, EMS-induced mutations and those found in multiple mutants were subtracted so as to leave only a handful of variations to examine in a relatively small portion of the genome. From there, the authors were able to identify the mutation causing the phenotype of interest (Zuryn et al., 2010).

However, the strategy most relevant to this study is that described by Schneeberger et al., in which they crossed *A. thaliana* plants with EMS-induced mutations to a polymorphic strain: one for which a SNP marker is known approximately once every kilobase. Five hundred segregants were pooled and sequenced at 22x coverage. From this sequencing data, the software SHOREmap was used to develop interval plots much like those produced in this study, which allowed a narrow region in which the causal mutation may be located to be identified. From there, one mutation was identified and the researchers confirmed it to be the causal mutation. (Schneeberger et al., 2009).

As these mapping strategies have become more and more common for genetic studies, researchers have worked to optimize their computational strategies and also to make mapping resources more widely available. The result of this has been the development of publicly-available tools such as CloudMap, which allows users to map mutations using a cloud-based analytical system (Minevich et al., 2012). As resources such as these are developed, these types of mapping studies will become quicker and more readily achievable. This, combined with the decreasing cost of next-generation sequencing technologies, will make mapping studies such as the one performed in this project much more cost-effective in coming years. This will hopefully lead to a large increase in the number of genes identified that are vital for plant growth, bringing us to a better understanding of how plants must be engineered to meet the needs of today's society.

Previous Work

This project was a continuation of the work done by Xinxin Ding, WPI Class of 2014, as part of her Major Qualifying Project (Ding, 2014). Her project involved analyzing mutants of the moss *P.patens* produced through UV mutagenesis from the wild-type Gransden line, in addition to producing additional mutatns. As mentioned previously, the Gransden line is the one for which the full genome sequence is publicly available. Ding isolated two new temperature-sensitive plants, and the remainder of her work focused on six previously-isolated "Conditional Loss of Growth" cell lines, named CLoG1 through CLoG6. She crossed each of these lines with the Villersexel strain of *P. patens*. The Villersexel strain is polymorphic, meaning it has many known single nucleotide polymorphisms (SNPs) when compared to the Gransden line. These SNPs allow Villersexel DNA to be distinguished from Gransden DNA after homologous recombination occurs during crossing. The TS lines used in this study were derived from the Gransden line, meaning that after outcrossing to Villersexel, portions of the progeny's DNA can be determined to be of either Gransden or Villersexel origin. Ding expanded and screened more than 100 plants outcrossed from CLOG1, twenty-four of which maintained a TS phenotype. DNA was extracted from each plant, which was then pooled and sent for whole-genome sequencing. Ding worked to map the mutation causing this temperature sensitivity using Linux-based tools including Burrows-Wheeler Aligner (Li and Durbin, 2009), SAMTools (Li et al., 2009), and original MATLAB code.

This mapping was performed by taking advantage of the nature of meiotic homologous recombination. Homologous recombination is the process by which portions of two sequences of very similar DNA are exchanged. This process is generally used to repair double-stranded breaks in DNA, by using a similar sequence to replace the damaged portion of the DNA. It is also responsible for much of the genetic diversity developed during meiosis in the formation of gametes, as it allows portions of maternal chromosomes to be exchanged with portions of paternal chromosomes. This gives the offspring a mixture of both the maternal and paternal DNA; as there are only two sources of genetic information, it is expected that for any given nucleotide in a progeny's genome, there is a 50% chance that nucleotide came from the paternal DNA and a 50% chance that nucleotide came from the maternal DNA. In this study, the Gransden-derived TS CLoG1 line and the polymorphic Villersexel line served as the parents. Therefore in their progeny – the outcrossed plants that were screened for temperature sensitivity – for each nucleotide, there should be a 50% chance that the nucleotide came from the CLoG1 line and a 50% chance that it came from the Villersexel line.

The mapping strategy used in this study took advantage of the fact that this 50/50 ratio would be maintained for all but one small region of the *P. patens* genome: the location of the mutation causing the plants' temperature sensitivity. This is because the plants that were sequenced had all been selected specifically for their temperature sensitivity. Therefore at one location on their genome, every plant should have the same nucleotide: the TS' phenotype causal mutation that came from the CLoG1 line. When analyzing the reads that comprised the whole-genome sequence of these pooled TS segregants, this can be seen by plotting the ratio of alleles coming from the CLoG1 line to the total number of alleles seen at each position analyzed. This analysis could theoretically be performed for every nucleotide in the genome, but a more practical analysis is to calculate this ratio for every location of a Villersexel marker. This way, only portions of the genome where the Villersexel genome diverged from the Gransden genome in the first place are analyzed.

The result of this plotting is a trendline at approximately 0.5 (the result of the 50/50 distribution described previously) for the entire genome except near the region containing the TS phenotype's causal mutation. Here, a peak around 1.0 should be seen. This signifies that every read in that region came from CLoG1 DNA and is therefore very likely to contain the mutation that was selected for. Around this area, the ratio of CLoG1 alleles to the total stays above 0.5 for the length of the recombination fragment on which the mutation is located. This process narrows down the mutation's location to approximately a 3 Mbp region. From there, individual mutations in the region must be examined and evaluated for the likelihood of their being the causal mutation of the TS phenotype, based on the characteristics of TS mutations described previously.

Project Objectives

This project aimed to achieve the following objectives through the continuation of Ding's work: 1) expansion of plants produced from crosses between the CLoG4 and Villersexel strains; 2) extraction, pooling, and sequencing of their DNA; 3) mapping of the causal mutation for CLoG4 using the same procedure that Ding used for CLoG1; 4) documentation of the mapping procedure so as to make it more user-friendly and reproducible; 5) execution of a rescue experiment to confirm the mapped location of CLoG1's temperature-sensitive mutation.

Secondary objectives included comparison of this mapping strategy to that of CloudMap's and development of a method for estimating the frequency of TS mutations in the *P. patens* genome. CLoG4 was chosen for this project because its genome had been sequenced as part of Ding's work. This would allow the sequencing reads of CLoG4's TS crossed plants to be mapped to CLoG4's wild-type sequence instead of just to the Gransden sequence. It would also allow for the identification of mutations in CLoG1's sequence that were not the result of UV mutagenesis but rather were from naturally-accumulating mutations in the laboratory's Gransden line. By mapping the causal mutation for another line's temperature sensitivity, support would be given to this method of isolation, screening, and mapping as an effective way to identify genes important for proper cell growth. Documenting this procedure would allow it to be more commonly used in the future, and comparing it to the use of CloudMap would determine if there are quicker ways of performing this analysis. Confirming the location of CLoG1's causal mutation would also be invaluable in verifying the experiment's accuracy at identifying the proper gene.

MATERIALS AND METHODS

TS Mutant Expansion

Outcrossed plants produced by Ding (Ding, 2014) were expanded by grinding each plant in 1.2 mL dH_2O using a Fisher Scientific PowerGen 125 Homogenizer. This was divided between two small 60x15 mm Petri dishes of 10 mL $PpNH_4$ medium (see recipe, Appendix 2) overlaid with cellophane. Four-hundred microliters were added to each plate, and the remainder was placed on a plate of LB medium to monitor for contamination. These plants were left to grow at 25°C under a cycle of 16 hours light and 8 hours dark for at least two days. After this initial growth period, one plate of each plant was left to grow at 25°C for another seven days while the other plate was transferred to 32°C for the seven-day growth period.

TS Mutant Screening

After growing at the permissive and restrictive temperatures of 25°C and 32°C respectively for seven days, plants were viewed under a Zeiss SteREO Discovery V12 microscope to screen for a temperature-sensitive phenotype. Plants that appeared to be temperature-sensitive were passed again following the same procedure described in *TS Mutant Expansion* above and re-screened under the same microscope settings. Any plants that still demonstrated a temperature-sensitive phenotype after this second round of screening was then used for DNA isolation.

TS Mutant DNA Isolation, Pooling, and Sequencing

Water was blotted out of the tissue for TS plants before being weighed and combined with other mutants to obtain sufficient plant tissue for the DNA isolation (10-50 mg). Approximately equal weights of tissue were used for each plant in this pool. Then, DNA was isolated from the pooled plant tissue using the Mo Bio PowerPlant Pro DNA Isolation Kit and the resulting DNA's concentration was measured using a NanoDrop 2000c spectrophotometer. The DNA samples isolated in *TS Mutant DNA Isolation* above were pooled so that each temperature-sensitive plant's DNA was present at the same concentration. This pooled sample was then sent to BGI Americas Corporation for whole-genome sequencing.

Causal Mutation Mapping

The mutation causing the plants' temperature sensitivity was mapped using the workflow developed by Ding (Ding, 2014). This involved isolating the locations of single nucleotide polymorphisms (SNPs) from the publicly available sequence of the polymorphic Villersexel strain of *P. patens* – SRR072296.sra, downloaded from the National Center for Biotechnology Information (NCBI)'s Sequence Read Archive (SRA) (NCBI SRA, 2010) – to develop a list of markers. Whole-genome sequencing reads of the temperature-sensitive strain CLoG1 were then aligned to the *P. patens* v3.0 genome (JGI, n.d.) and from this a list of variants, a depth matrix (Flaherty et al., 2012), and information about each chromosome's size was extracted. The tools used for this mapping procedure were: fastq-dump.2.4.1 (from NCBI's SRA toolkit), bwa (Li and Durbin, 2009, 2010), samtools (Li et al., 2009), bcftools, and vcfutils.pl (Li et al., 2009), along with MATLAB code written by Ding (Ding, 2014) and basic Linux commands such as grep and cut.

Rescue Experiment

PCR

To isolate the wild-type DNA necessary for this experiment, DNA was isolated from the wild-type line of *P. patens* using the Mo Bio PowerPlant Pro DNA Isolation Kit. This DNA was amplified via polymerase chain reaction (PCR) by running 20 50-µL reactions according to New England BioLabs Inc.'s "PCR Protocol for Phusion® High-Fidelity DNA Polymerase (M0530)," using 80.5 ng of DNA per reaction and primers CLoG1-mut(F) and CLoG1-mut(R) (see *Primer Design* below). The PCR was run using the following settings, as recommended by New England BioLabs Inc.'s PCR protocol and Tm calculator: 98°C for 30 seconds; 30 cycles of: 98°C for 10 seconds, 66°C for 30 seconds, and 72°C for 60 seconds; 72°C for 10 minutes; hold at 4°C. The resulting 1 mL of PCR product was purified using the Machery-Nagel NucleoSpin® Extract II kit according to their "Protocol for PCR clean-up." Measurement with the Nanodrop 2000c spectrophotometer determined the total yield to be 3.1 µg of DNA. To increase this yield, a second PCR was performed on this product. Using the same primers and temperature cycle as before, 20 50-µL reactions were run with 10 ng PCR product per reaction according to New England BioLabs Inc.'s "PCR reaction using Phusion Hot Start II DNA Polymerase (F-549)" protocol. One µL of each primer – CLoG1-outF and CLoG1-outR – was used per reaction, at a concentration of 10µM. This yielded at total of 28.73 µg of wild-type DNA.

Transformation

Transformations were performed as described by Liu (Liu and Vidali, 2011). As counting with a hemocytometer determined there to be a total of 5,875,000 protoplasts, 3.7 mL of MMg was used for incubation. Two transformations were performed, each using 600µL of protoplasts: one with 30µg of a 3mEGFP hygromycin-resistant construct and another with this construct plus 14.68µg of wild-type DNA.

Transformant Expansion, DNA Extraction, and Sequencing

Twenty plants resulting from the transformation were expanded by grinding the plant in 1.5 mL of dH_2O using a homogenizer and the resulting mixture on two small $PpNH_4$ plates – 500 µL each – and the remainder on an LB plate to check for contamination. One plant developed contamination, so DNA was extracted from the remaining nineteen plates using the Mo Bio PowerPlant Pro DNA Isolation Kit. A 20-µL PCR was performed for each sample extracted from these plants. Following New England Biolab Inc.'s "PCR reaction using Phusion Hot Start II DNA Polymerase (F-549)" protocol with 1µL of 10µM CLoG1-outF, 1µL of 10µM primer CLoG1-outR, and 1µL of DNA per reaction. After running the reaction for 30 seconds at 98°C; 30 cycles of 10 seconds at 98°C, 30 seconds at 67°C , 60 seconds at 72°C; and 10 minutes at 72°C, the PCR product was loaded into a medium-sized 0.8% Seakem GTG agarose gel, stained with ethidium bromide and run at 220V for 19 minutes. The bands were removed from the gel, and DNA was purified from the samples using the Machery-Nagel NucleoSpin® Extract II kit according to their "Protocol for DNA extraction from agarose gels." The purified DNA was then sent to Eton Bioscience Inc. with primers CLoG1-inF and CLoG1-inR for sequencing.

Primer Design

The first set of primers designed for this experiment, CLoG1-mut(R) and CLoG1-mut(F) were used in the PCRs that produced the DNA necessary for the transformation. They were designed to amplify approximately 1kb on either side of the mutation's location at nucleotide 4,325,703 of *P. patens*'s v3.0 genome. CLoG1-mut(R) has a sequence of GCG TTG AAG TCA TCT GTA ATG AGG and CLoG1-mut(F) has a sequence of ATT CCT CCT CCT CTA CTT CCA ACT. The second set of primers, CLoG1-outF and CLoG1-outR, were designed to amplify approximately 150bp outside of the region amplified by CLoG1-mut(R) and CLoG1-mut(F). This was done so that only the targeted area would be amplified, to avoid any contamination from other sites of the genome where the PCR fragment might be unintentionally incorporated. CLoG1-outF has a sequence of TAC CTG AAA AAC TCT TCA CCA CCA and CLoG1-outR has a sequence of GTT CAC TAT TGG AGG TAA GCT GGA. A third set of primers, CLoG1-inF and CLoG1-inR, were designed to amplify approximately 150bp on either side of the mutation, for sequencing purposes. CLoG1-inF has a sequence of TTT GCT TTC AAT AAA CCG CAC AAC and CLoG1-inR has a sequence of TGT TGT GTT GTT CTT TGA GCA GTT. These primers' locations relative to the mutation are shown in figure 1 below.

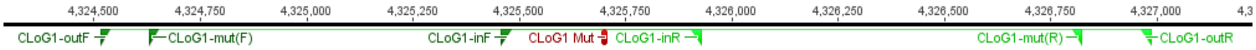

Figure 1: Primer Locations This figure shows the presence of the suspected causal mutation at position 4,325,703 on chromosome 24 (shown in red), and the locations of the forward primers (in dark green) and reverse primers (in light green) for the original selection primers, the external confirmation primers, and the internal sequencing primers. Screenshot taken of Geneious software.

Three more pairs of primers were designed to amplify mutations found elsewhere in chromosome 24 that were present in CLoG1's sequence but not in the Villersexel or wild-type sequences. These mutations were identified using MATLAB code, included in supplement 1. All mutations were downstream of the casual mutation, and the primers were designed to bind 250-300 base pairs away from the mutation on either side. The first mutation, located at base pair 7,533,237, was amplified using primers mutCLoG1-1F and mutCLoG1-1R with sequences of GGA AAG TTA AAG AGT GGC CAC G and TCC AAA TCA CTT TCA ATC TCT AGC respectively. The second mutation, located at base pair 10,809,758, was amplified using primers mutCLoG1-2F and mutCLoG1-2R with sequences of GTT CTT CCA AGC CTG CAC GAC C and TGC TTT AAG AAC TCC ATT TCC AAG respectively. Finally, the third mutation, located at base pair 12,848,580, was amplified using primers mutCLoG1-3F and mutCLoG1-3R with sequences of GGT TCA GTC ACC CCG TTA CAA and CCT CAG TCT TCC AAA AAC CCT AT respectively. The locations of these three mutations and their primers is shown in figure 2 below.

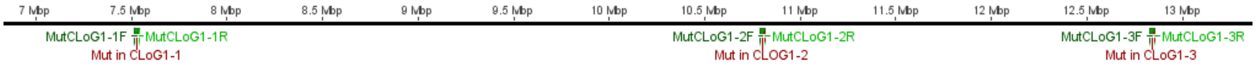

Figure 2: Confirmation Mutations – Primer Locations This figure shows the location of the three mutations used for confirming that the "rescued" plant was in fact rescued and was not the result of contamination. The mutations are located at base pairs 7,533,237, 10,809,758, and 12,848,580 on chromosome 24. The causal mutation is located further upstream, at position 4,325,703. Screenshot taken of Geneious software.

RESULTS

35 TS Segregants Isolated

Of 75 plants resulting from a cross between the temperature-sensitive line CLoG4 and the polymorphic Villersexel strain, 35 remained free of bacterial contamination and showed clear temperature-sensitive (TS) phenotypes after two rounds of screening. Representative samples of both wild-type and TS CLoG4 plants are shown in figure 3 below.

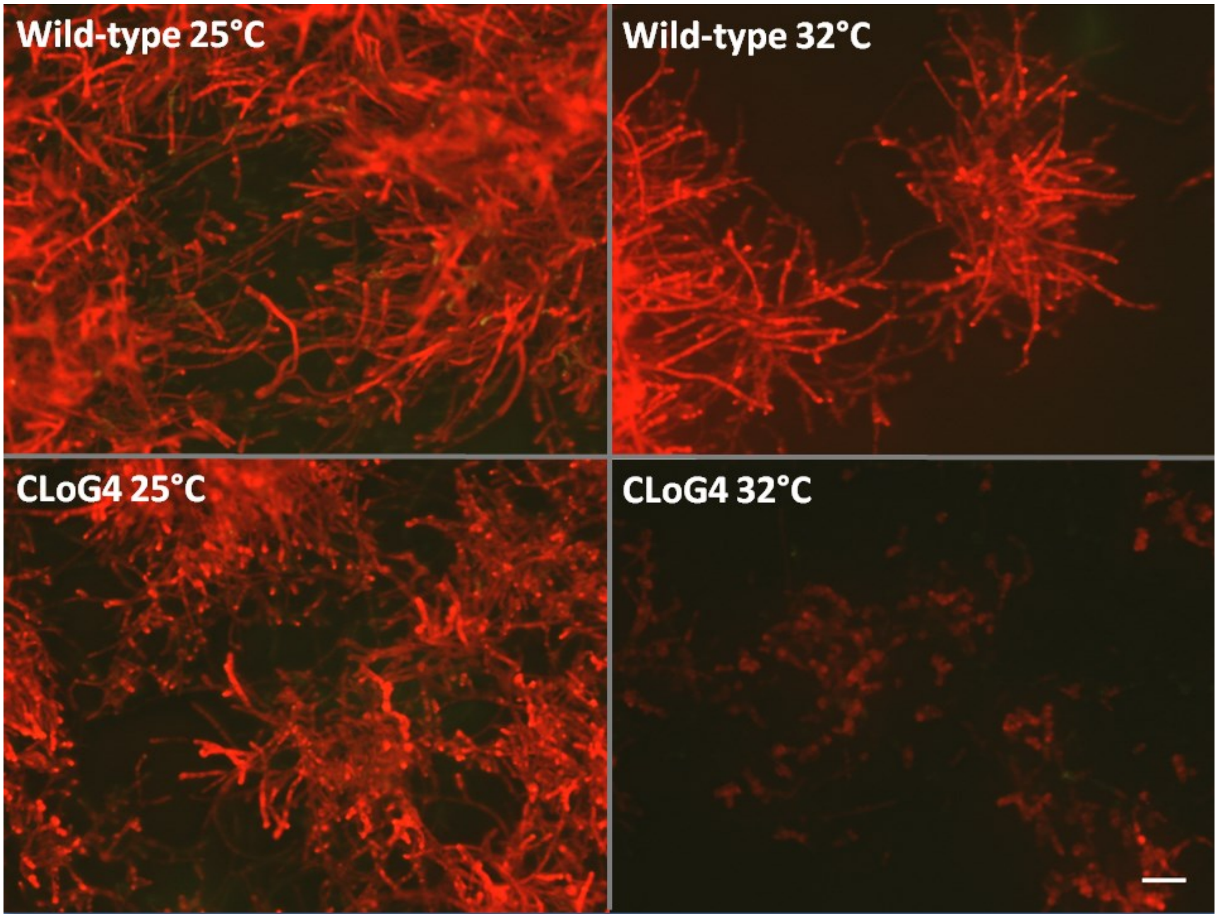

Figure 3: Images of Wild-type and TS Plants Pictures of wild-type and TS CLoG4 plants, each after one week of growth at 25°C or 32°C. Although they grow much like the wild-type plants at 25°C, TS plants show a clear phenotype of stunted growth when grown at 32°C. Scale bar ≈ 100 μm.

A 12.04 µg sample of pooled DNA was produced from these 35 mutants, at a concentration of 0.344 µg per mutant. This sample was then sent to BGI-Americas for whole-genome sequencing at 10X coverage.

Mapping Procedures Performed

CLoG1 Mapping Reproduced and Documented

One aim of this project was to map CLoG1's mutation again, for the sake of confirming the mapping procedure's reproducibility while also documenting it to make it more user-friendly. Ding (Ding, 2014) had mapped the location of the causal mutation for the TS line CLoG1 to position 4,325,703 on chromosome 24 of the *Physcomitrella patens* v3.0 genome (JGI, n.d.). At this location, there is a guanine present in the wild-type DNA whereas there is an adenine present in CLoG1. The goal was to obtain this same result after attempting to reproduce her mapping strategy from scratch. Doing this resulted in an alignment that showed the same mutation that she identified. This was seen using the alignment's visualization in The Broad Institute's Integrative Genomics Viewer, or IGV (Robinson et al., 2011), as shown in figure 4 below.

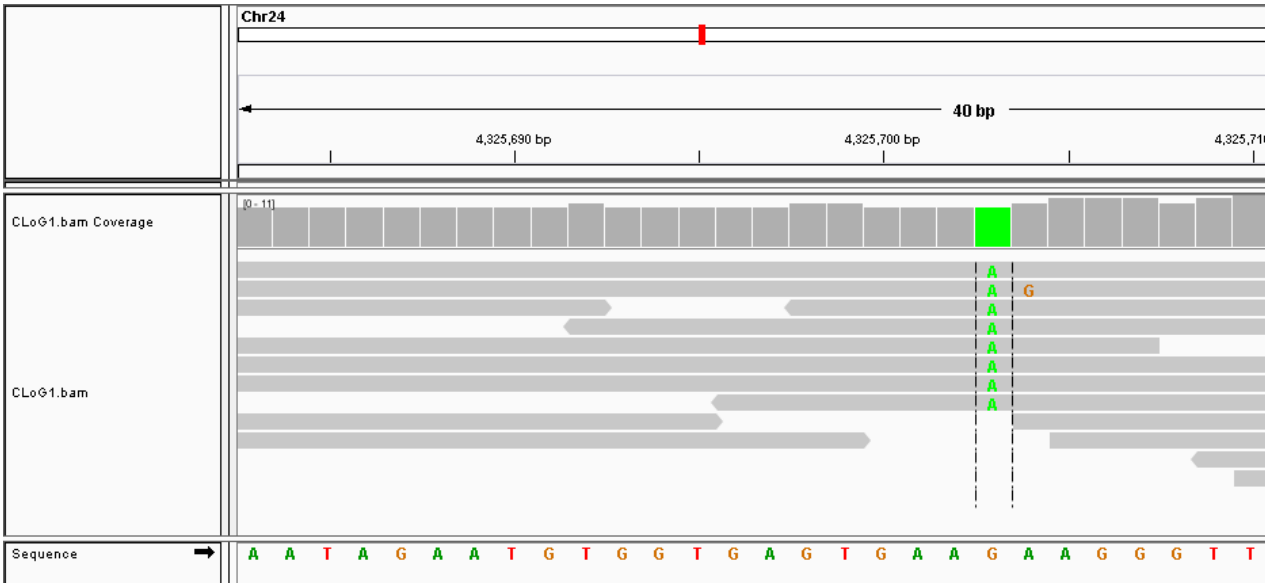

Figure 4: CLoG1 Mutation Location This alignment resulted from reproducing the mapping procedure for CLoG1 developed by Ding. The mutation that she identified – a G>A mutation seen in every read covering position 4,325,703 on chromosome 24 – is confirmed to be present in the reproduced alignment as expected.

A broad overview of the mapping procedure is described in the chart shown in figure 5. To document the whole procedure, a flowchart was developed to show the different steps and files involved throughout the entire mapping process. This flowchart, with the proper input files and access to these Linux-based software packages, should allow the mapping process to be reproduced for any CLoG mutant with relative ease. The flowchart is shown in figure 6 below.

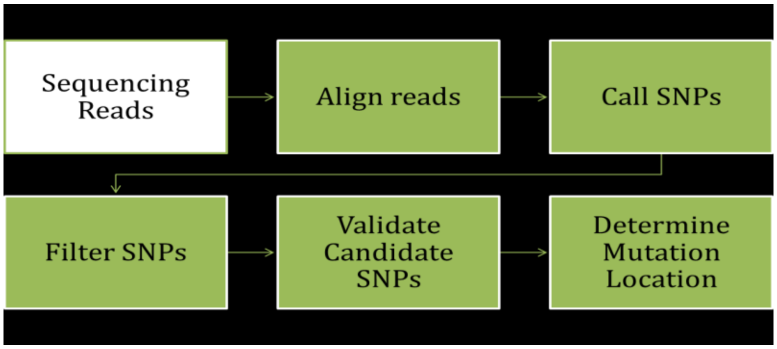

Figure 5: Mapping Overview This flowchart gives a very broad overview of the major steps in the mapping procedure. A detailed flowchart of the procedure can be found in figure 6.

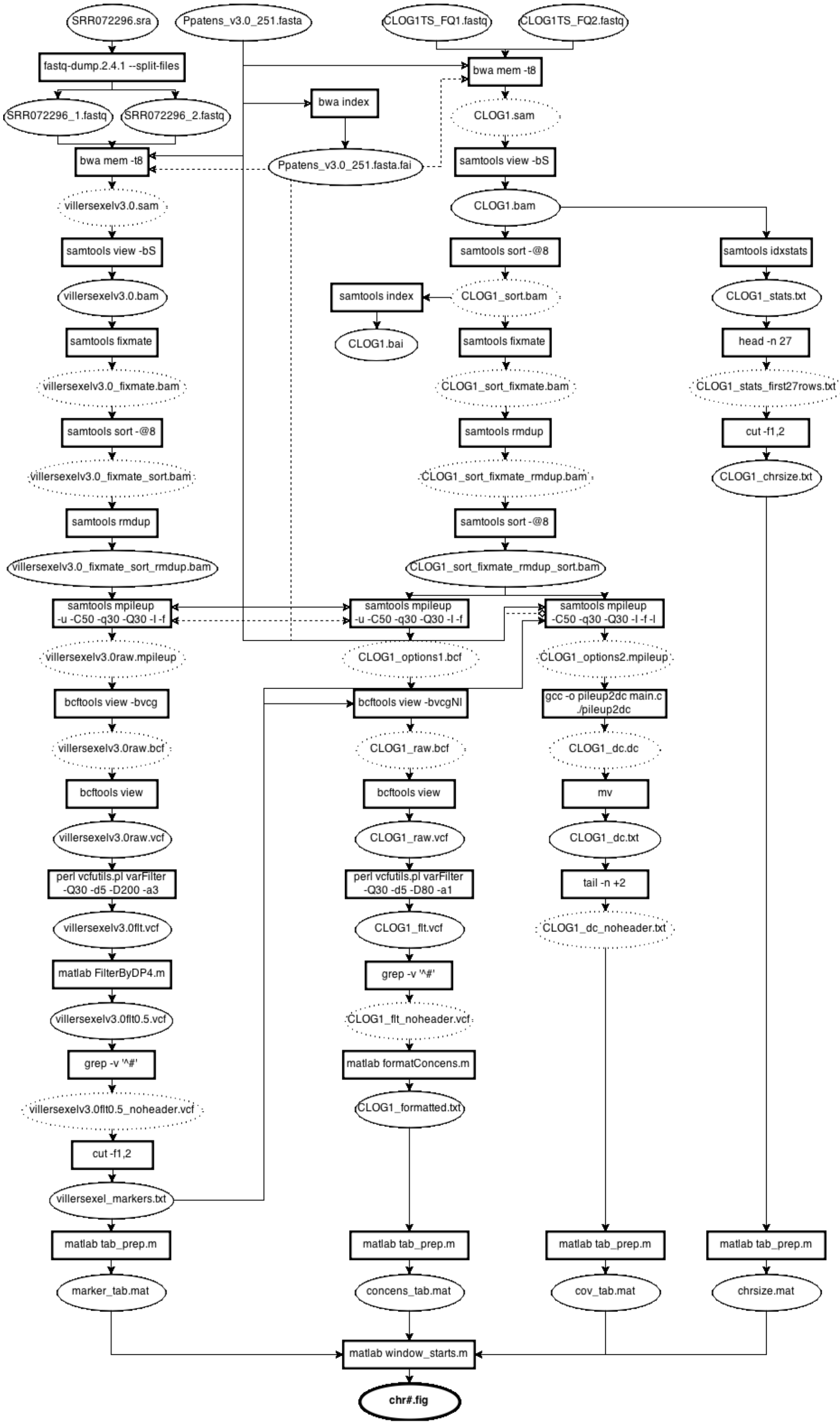

Figure 6: Mapping Procedure Flowchart This figure shows a flowchart of the entire mapping process from start to end. Boxes represent Linux commands while circles represent files – dotted lines for files that are discarded after mapping and solid lines for ones that are retained.

Attempting to reproduce the mapping process resulted in plots similar to Ding's unpublished results, but these plots were not exactly the same due to discrepancies seen at the beginning of almost every chromosome. It did, however, produce the same peak on chromosome 24 that Ding's did. Both versions of the plot for chromosome 24 are shown in figure 7.

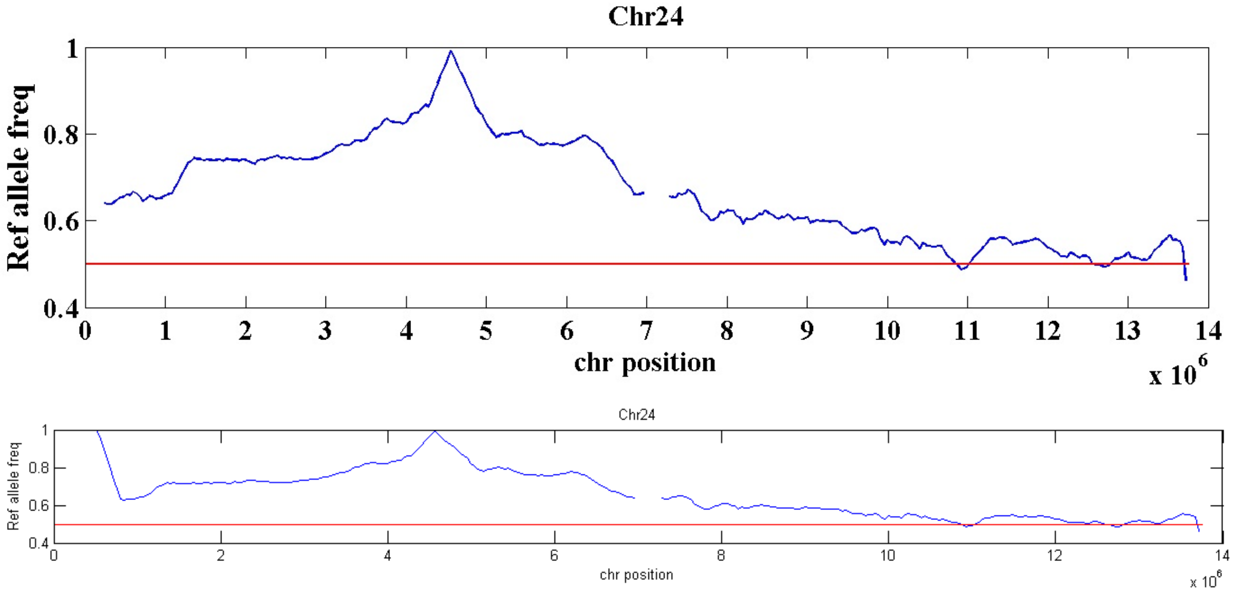

Figure 7: Mapping Result Plots The mapping results achieved by Ding (top) and my reproduction (bottom). The plot represents the ratio of CLoG1 DNA to Villersexel DNA at each position along the chromosome. It is still unclear why there is an off-the-plot peak at the beginning of my plot and not Ding's, so further work is needed to identify the cause of this discrepancy. However, larger trends further down the chromosome remain the same for both plots, such as the peak seen around position 4,300,000.

CLoG4 Alignment Performed

The reads resulting from the sequencing of CLoG4's TS segregants were aligned to the *P. patens* v3.0 genome (JGI, n.d.) using Burrows-Wheeler Aligner (Li and Durbin, 2009) to produce an alignment file. This alignment file was visualized using IGV (Robinson et al., 2011) as shown in figure 8 below.

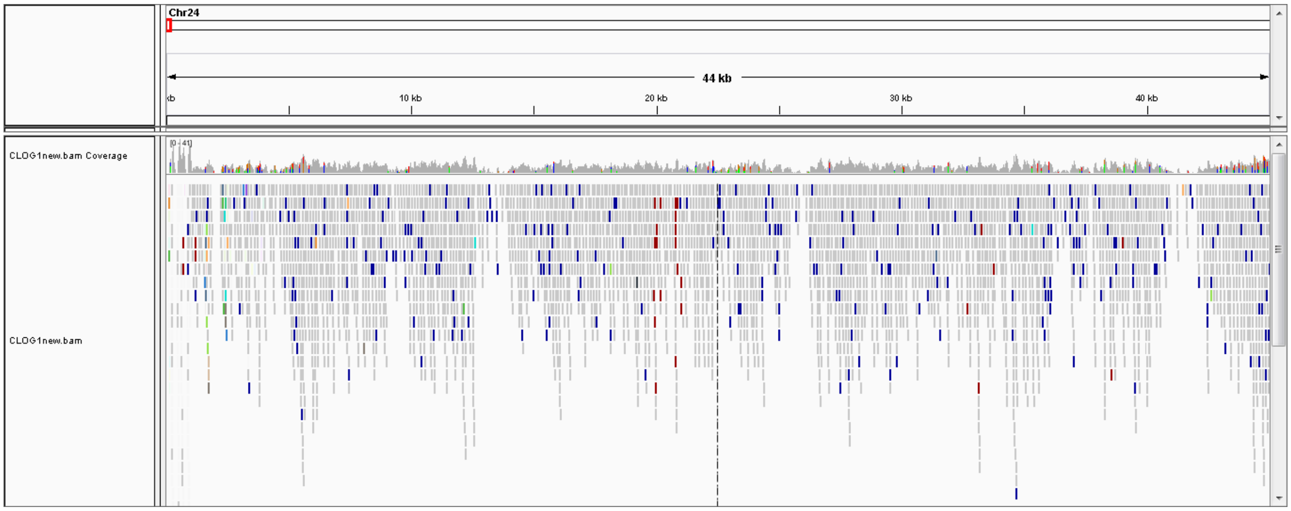

Figure 8: CLoG4 Alignment Visualization. This screenshot shows what the alignment of CLoG4's TS segregant sequencing reads to the *P. patens* v 3.0 genome looks like when visualized with IGV.

Comparison to CloudMap

The mapping procedure used in this study was compared to Galaxy's CloudMap (Minevich et al., 2012) to evaluate its relative efficiency. In doing so, CLoG1 was mapped to show the same peak as our procedure, as shown in figure 9 below. CLoG4 was also mapped, but no clear peak emerged that allowed definitive conclusions to be drawn.

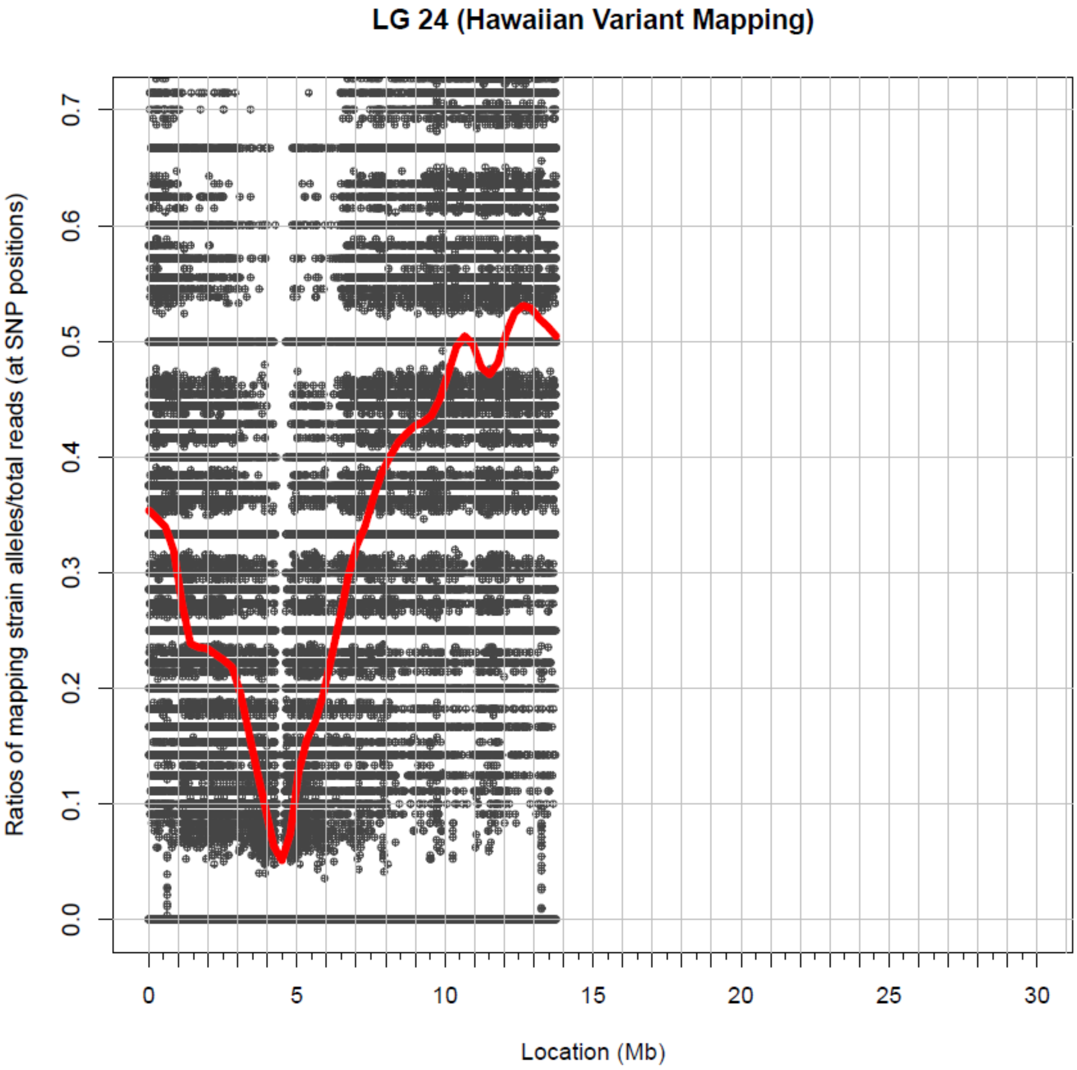

Figure 9: CloudMap results for CLoG1 This figure shows the map produced by CloudMap for CLoG1 at chromosome 24. Note that here the ratio is inverted compared to our procedure. Therefore "valleys" down to zero are indicative of the mutation's location rather than the peaks

to one seen previously. However, the location identified for CLoG1's mutation remains consistent between our mapping procedure and CloudMap's.

Rescue Experiment Performed

To confirm that this mutation was responsible for the line's temperature-sensitive phenotype, CLoG1 cells were transformed with wild-type DNA. This DNA was the product of a PCR on wild-type DNA amplified using primers CLoG1-mut(R) and CLoG1-mut(F). Nineteen of the resulting plants' DNA was amplified via PCR using primers CLoG1-outF and CLoG1-outR, and the resulting amplified DNA for nine of those plants is shown in figure 10 below.

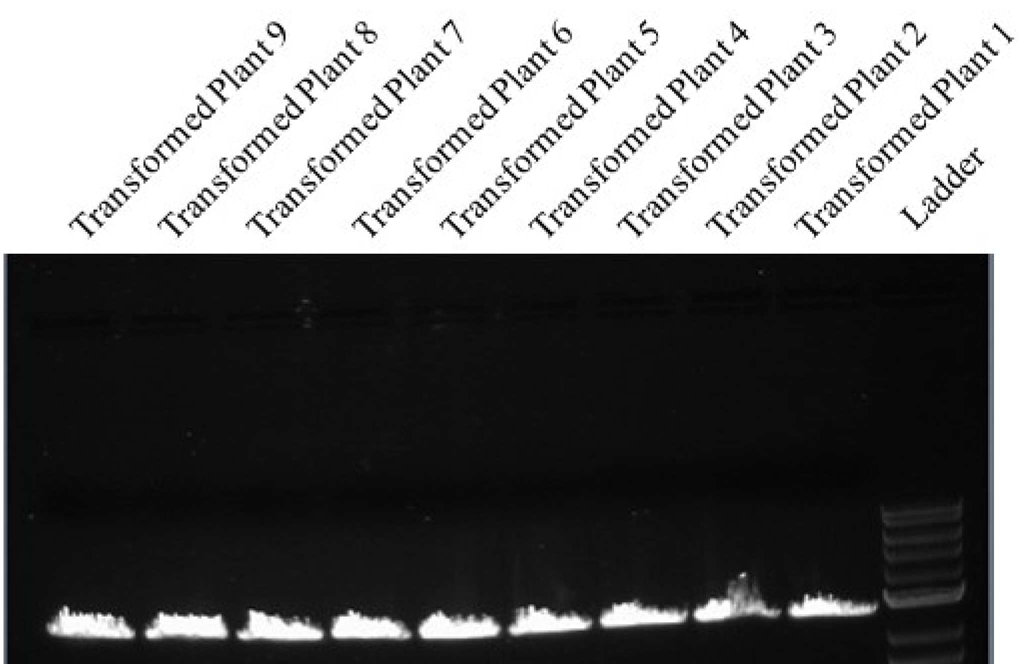

Figure 10: Transformant DNA Amplification This figure shows a gel with nine of the transformed plants after amplification via PCR.

After isolation from the gel, these fragments were sequenced using primers CLoG1-inF and CLoG1-inR to identify plants that had been rescued, meaning ones that took up the wild-type

DNA via homologous recombination. Sequencing results for these plants – shown in figure 11 below – appeared to show that one plant had been rescued, based on the presence of guanine (characteristic of wild-type DNA) at position 4,325,703. The remaining eighteen plants contained the guanine-to-adenine mutation characteristic of CLoG1's DNA at this position and therefore had not been rescued.

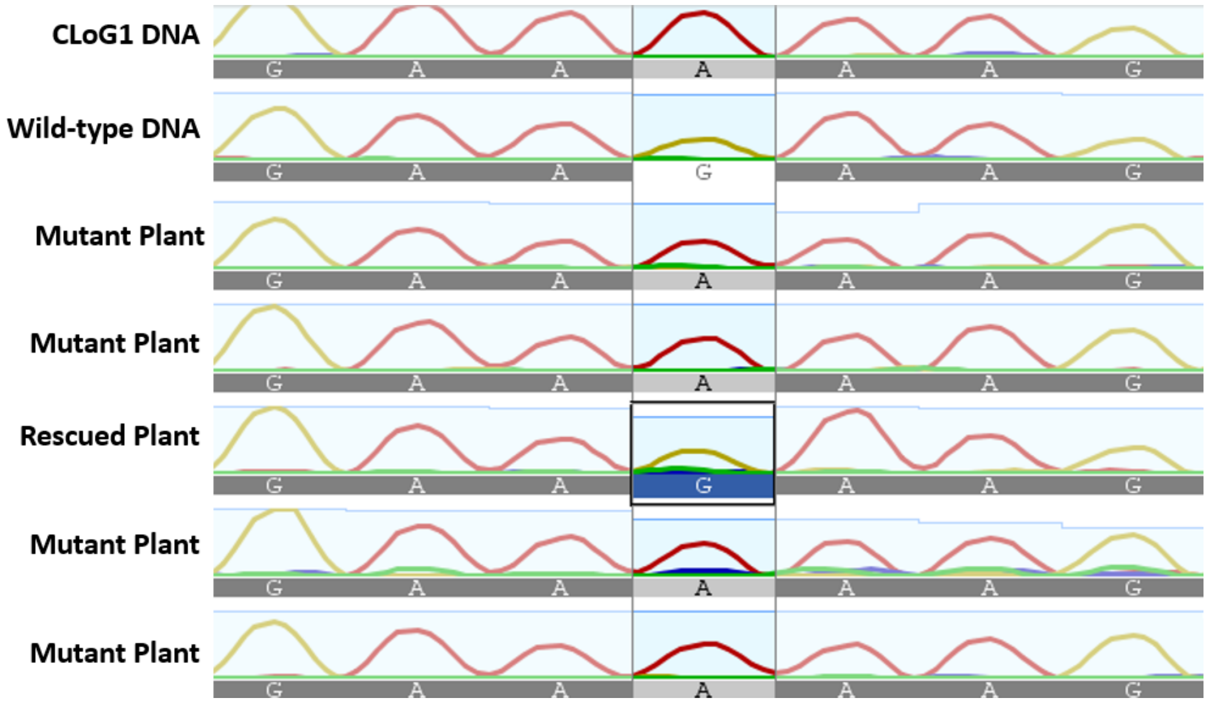

Figure 11: Rescue Experiment Sequencing Results. Sequences of base pairs 4,325,700-4,325,706 of chromosome 24 are shown for CLoG1, wild-type, and five transformed plants. Four of the plants retained the CLoG1 DNA with an adenine at base pair 4,325,703, while one plant (shown third from bottom) obtained the wild-type DNA with a guanine at this location, suggesting it may have been rescued. Screenshot taken of Geneious software.

To rule out the possibility of any contamination from wild type plants or DNA, two other mutations were identified in the CLoG1 genome that were not present in the Villersexel or wild-type lines. Primers were designed approximately 250-300 bp on either side of the mutation, to give a PCR product 500-600 bp in length. The PCR products for these locations in the wild-type, CLoG1, and rescued plant's DNA were all sent off for sequencing. Three mutations and sets of primers were intended for use, but only two were initially sent off for sequencing due to a lack of amplification as shown in figures 12 and 13 below.

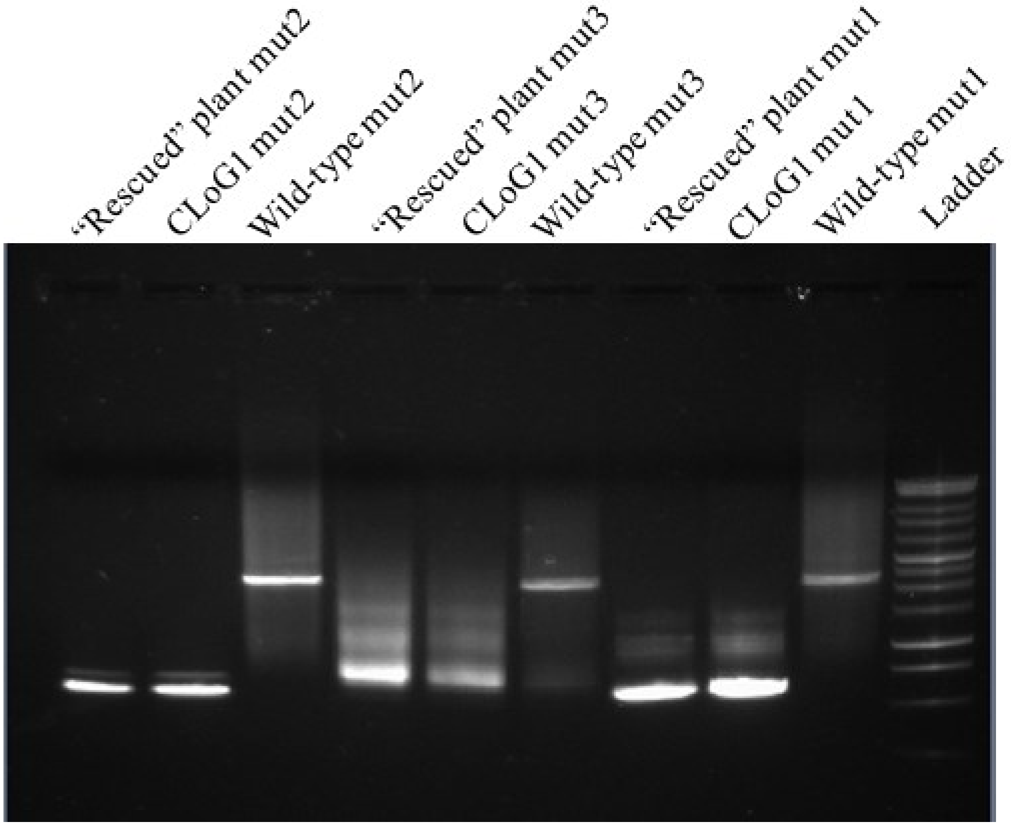

Figure 12: Confirmation Mutations Amplification An electrophoresis gel showing amplification of DNA fragments after PCR for each of the three chosen mutation locations. Each lane is labeled with the type of DNA amplified (Wild-type, CLoG1, or the rescued plant) and the primers used (mut1 = mutCLoG1-1F and mutCLoG1-1R, mut2 = mutCLoG1-2F and mutCLoG1-2R, and mut3 = mutCLoG1-3F and mutCLoG1-3R). Ladder has markers at 10000, 8000, 6000, 5000, 4000, 3000, 2500, 2000, 1500, 1000, 750, 500, and 250 base pairs. As can be seen on the gel, the rescued plant and CLoG1 DNA amplified a ~500 bp region as expected. The wild-type fragments were longer than the others – ~2500 bp – due to a mistake made in choosing which DNA sample to amplify from. Therefore the wild-type DNA was redone, as shown in figure 13.

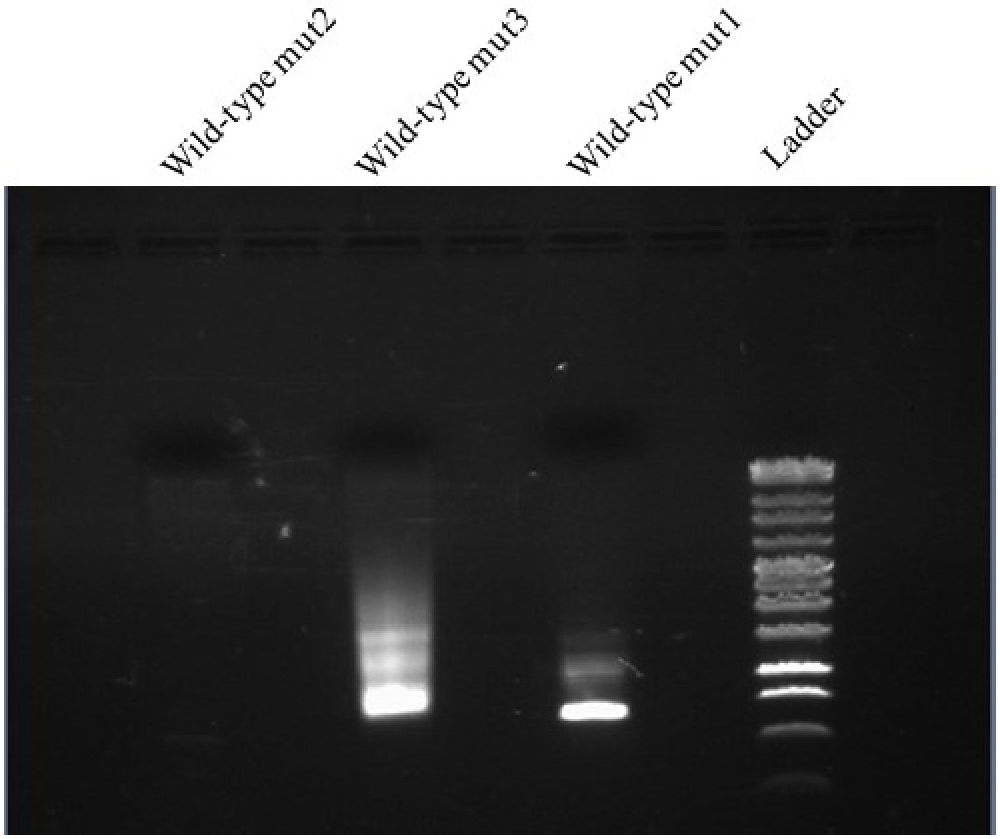

Figure 13: Wild-type DNA Amplification As the gel shown in figure 12 was not run with the proper wild-type sample, the PCR and gel electrophoresis were repeated with the proper DNA. As can be seen, the mut2 primers did not amplify anything. Therefore only the mut1 and mut3 samples were sent for sequencing. Again, the ladder used had markers at 10000, 8000, 6000, 5000, 4000, 3000, 2500, 2000, 1500, 1000, 750, 500, and 250 base pairs. Although the bands are not as distinct as those seen in figure 12, it is clear that these bands are in the expected 500-600 bp range unlike the 2500-bp wild-type DNA from before.

However, this sequencing proved to be inconclusive. Primers mutCLoG1-3F and mutCLoG1-3R produced very poor sequencing quality, while mutCLoG1-1F and mutCLoG1-1R showed the mutation as being present in CLoG1 despite our efforts to only use mutations that would not be present in the TS line. Therefore further work is necessary before the plant's rescue can be fully confirmed.

Probabilistic Model Developed

As mentioned previously, a secondary aim of this project was to develop a method for estimating the frequency of TS mutations in *P. patens*. It would be useful to have a model which, when given a certain configuration of unique genes that yield TS mutations, would estimate the total number of genes in the genome which could yield a TS phenotype when mutated. For the purpose of this section, these genes will be referred to as "TS genes" and a configuration will be defined as the count and locations for an observed set of TS genes (e.g., the six genes that correspond to the six CLoG lines). This would allow a researcher to determine the probability that they have reached saturation with their experimentally-obtained mutants, or the predicted remaining number of TS genes if saturation had not been achieved.

Let us define the true number of TS genes as the random variable G. Given an observed configuration of TS genes *x*; our object of inference is $p(G|x)$. Bayes' formula states that $p(B|A) = \frac{p(B \cap A)}{p(A)}$; since $p(B \cap A) = p(B) * p(A|B)$, Bayes' formula can be written as follows: $p(B|A) = \frac{p(B) * p(A|B)}{p(A)}$ (Durrett, 2009). Therefore, $p(G = g|x) = \frac{p(G=g) * p(x|G=g)}{p(x)}$. In this formula, each term has a name: $p(G = g|x)$ is the posterior, $p(G = g)$ is the prior, $p(x|G = g)$ is the likelihood, and $p(x)$ is the normalization constant. $p(x|G = g)$ can be calculated using a multinomial distribution function, which can be written as $\frac{(x_1 + \cdots + x_n)!}{x_1! * \ldots * x!} (\frac{1}{g})^{x_1} * \ldots * (\frac{1}{g})^{x_n}$ (Evans et al., 2000). For example, if *x* for CLoGs 1-6 is [2,1,1,1,1] with *n* = 5 (two of the mutations are mapped to the same gene while the rest are located on separate, individual genes; a total of five TS genes have been identified), the likelihood is $p([2, 1, 1, 1, 1]|G = g) = \frac{6!}{2!1!1!1!1!} \left(\frac{1}{g}\right)^2 (\frac{1}{g})^1 (\frac{1}{g})^1 (\frac{1}{g})^1 (\frac{1}{g})^1$, which can be calculated for as many values of *g* as desired.

This leaves two term to be calculated in Bayes' formula: $p(G = g)$ and $p(x)$. $p(G = g)$ is the prior over the number of true TS genes. We would have to derive the prior from previous literature about other organisms, as there is currently no estimate for how many genes exist in *P. patens* that could be mutated to obtain a TS phenotype. The normalization constant $p(x)$ can also be calculated, as it is equivalent to $p(x, G = 1) + \cdots + p(x, G = n)$, or in its expanded form: $p(x|G = 1) * p(G = 1) + \cdots + p(x|G = g) * p(G = g)$. Given the literature-based estimate for $p(k = j)$, this could be calculated using the formula for $p(x|G = g)$ derived above. Therefore once a distribution of values for $p(G = g)$ is determined and all six CLoG mutations have been mapped, we can estimate the posterior distribution over the number of TS genes.

DISCUSSION

Improvement upon Screening Procedure

Although CLoG4 proved to have a higher frequency of plants with a TS phenotype in its segregants after crossing with Villersexel than CLoG1 did (with ratios of TS plants:total being >1:3 for CLoG4 vs. <1:4 for CLoG1), the screening procedure is still a labor-intensive one. Seventy-five plants had to be expanded onto small Petri plates, grown in an incubator for a length of nine days, and then screened under a microscope. Then any plants that maintained a TS phenotype were passed and grown for another seven days for verifying the phenotype. This is a very time-consuming process, and anything to cut down on this time would greatly expedite the mapping process. One possible solution would be to expand the plants into 96-well plates that could be screened automatically using a Zeiss Axiovert 200M microscope. Currently a representative sample of 3-5 photos of each plant strain are taken manually. By allowing the microscope to take the pictures, one could greatly decrease the time and resources needed for the screening process. The size of the representative sample could also be increased if the pictures are being taken automatically.

The screening procedure could also benefit from an increase in accuracy. Currently the 3-5 pictures taken are screened visually and the research must decide which plants appear to be TS. This introduces the opportunity for human error more so than the statistical methods used for determining temperature sensitivity in the growth assays (Ding, 2014). The solidity and area of the plants are analyzed using an ImageJ routine, and currently the resulting data are used to develop boxplots that give an overview of the morphology for every plant in a sample. To make this method more feasible for use in the screening process, we could use prior data to develop thresholds of area and solidity that are likely to signify wild-type, TS, and questionable plants. Then a computer could easily output which plants appear to be TS and which are not, thus making the image analysis process more accurate as it will now rely on statistics instead of human judgment. An improvement in accuracy of this screening process would greatly improve the quality of the mutation mapping, as there would be less risk of background noise from plants that are not actually TS and would decrease the intensity of the peak present at the causal mutation's location.

Mapping Procedure

Discrepancies in CLoG1 Results

As seen in figure 7 above, discrepancies are seen between my maps produced for CLoG1 and those produced by Ding. This only appears to occur at the beginning of each chromosome, but it does appear to occur in nearly every one. This is not overly surprising as none of the files produced in this study exactly matched those produced by Ding. Most were extremely close, but small differences at the beginning of the mapping workflow may have very easily led to large discrepancies later on. The most probable cause of this would be changes in software. In the past year it is likely that software versions have been updated and modified slightly. As it is unknown which versions of each tool Ding used when mapping CLoG1, I have been unable to verify if this is the cause for the discrepancies between our maps. It is also possible that a step in the pathway was altered or not documented properly and that therefore a step in the workflow was performed incorrectly, leading to differences in the files.

My assumption, which remains to be tested, is that the Villersexel marker files produced by Ding and myself differed due to different software versions or a mistake in the filtering steps. It is also possible that the markers at the beginning of the file would be disproportionately affected by this difference or mistake, particularly if the sliding window analysis performed by MATLAB would be unable to run properly at the beginning of the chromosome. As the maps are based directly on the Villersexel marker positions, differing marker positions could easily lead to the discrepancies seen in figure 7. Why this would only occur at the beginning of each chromosome, however, is still up for exploration.

Comparison to CloudMap

As seen in figure 9, CloudMap's result for chromosome 24 of CLoG1 yields the same mutation location as our MATLAB-based mapping procedure. Therefore CloudMap (Minevich et al., 2012) is a viable and more automated alternative to our mapping process. However, CloudMap does have its drawbacks. As it is run on Galaxy's (Giardine et al., 2005; Blankenberg et al., 2010; Goecks et al., 2010) servers, wait times for experiments are variable depending on how many people are accessing the server at any given time. (Note that this is the case when using the free version of Galaxy as we were. Paid options are available.) Also, the procedure is run all at once. This gives the benefit of being more automated, as well as the drawback of not being able to examine files for any errors at each step. These factors mean that the user may start the procedure running in the morning, and if the servers are particularly busy the mapping procedure will not run until that night. Then the user may return the next morning just to find that an error occurred during one of the first steps, preventing the workflow from moving forward. Therefore this tradeoff must be taken into consideration.

Also, I have as of yet been unable to proceed past the mapping stage using CloudMap's workflow. After these maps are produced, a list of candidate mutations must be produced using either method. I ran into errors – which had unclear explanation messages – when trying to develop this. Therefore at the moment we are only capable of reproducing the maps using CloudMap but not fully identify the causal mutation of interest.

CLoG4 Mapping Preliminary Results

When CLoG4 was mapped using CloudMap, no clear peak was seen on any chromosome. Therefore it was decided that this mapping should be verified using our MATLAB-based approach. As documenting and troubleshooting the mapping procedure for re-creation of CLoG1's maps took much longer than expected, the mapping procedure has yet to be carried out for CLoG4. Therefore the only results we have for CLoG4 are those achieved using CloudMap. However, the deepest "valley" seen on these plots is one that reaches about 0.15 on chromosome 1. This means that 15% of reads at that area are still coming from the Villersexel strain, not the 0% expected for the causal mutation's location. Therefore no definitive conclusion can yet be drawn about CLoG4's mutation.

There are some possible causes for this. One is if the mutation is not actually a SNP as expected – for example, if it were caused by two mutations instead of one. However, if the phenotype was caused by multiple linked mutations, theoretically multiple peaks would arise. It would also be expected that a decreased number of TS segregants would be seen, not the increased number evident of the different frequencies of CLoG1 and CLoG4's screenings. The most likely explanation is that not all of the thirty-five segregants pooled for sequencing were truly TS. Despite two rounds of screening, it is possible that user error led to some segregants being included that were not truly TS. If the valley on chromosome 1 is the location of the causal mutation, its reaching to 0.15 would mean 15% of the mutants – or approximately 5 of the 35 used – were not TS. This is fairly high and it is doubtful that so many segregants were included mistakenly, however it is unclear what else would lead to such a result.

Comparison to Other Methods

Although many other mapping methods have been developed, most are fine-tuned for the use of one particular organism. As of yet, none have been developed specifically for *P. patens*. Most also call for a large mapping population. As discussed in the introduction, the strategy requiring the least number of segregants still recommends using fifty or more. As expanding, screening for, and selecting fifty to a hundred or more segregants is not feasible for the laboratory's resources, Ding developed a Monte Carlo simulation to determine the ideal number of segregants that should be used. This simulation found that although the more segregants the better, accuracy did not increase dramatically for sample sizes above twenty. Therefore our process can map a causal mutation with fairly high accuracy with as few as twenty to thirty segregants, a significant decrease from the fifty to five hundred that have been necessary for previous mapping strategies.

Rescue Experiment

Possible Causes of Inconclusive Sequencing

As mentioned previously, the sequencing of three extra mutations to confirm that CLoG1's mutation truly was rescued and was not just the result of contamination proved to be inconclusive. Sequencing was poor with a lot of background noise, and the reads could not be aligned properly to look at the mutation's location on each type of DNA. According to the Eton Biosciences troubleshooting page, some of the most common causes of this are not having enough DNA, contamination from salts or other substances that might inhibit the sequencing enzyme, excessive UV light exposure, or inefficient primer binding. As this inconclusive sequencing was only seen for one mutation and not the other, it is unlikely that they were contaminated with something inhibitory since both reactions were isolated at the same time using the same procedure.

DNA concentrations were also similar for both samples, so it is unlikely that these results were caused by not having enough DNA. As the bands containing the DNA were removed while being illuminated with UV light, prolonged UV exposure is a possibility but is unlikely considering the failed reaction's bands were removed from the gel before the more successful reaction's bands were. There were, however, very high 260/280 values (a measure of purity) for the samples. The mutCLoG1-3F and mutCLoG1-3R primer samples which failed to sequence properly had an average 260/280 value of 2.49, whereas the mutCLoG1-1F and mutCLoG1-1R samples which sequenced but with low quality had an average value of 2.12. As both of these values are above Eton Biosciences Inc.'s recommended value of 1.8-2.0 for sequencing, it is possible that this may be part of the problem. Theoretically this is a sign of high purity and should not be an issue, but it is possible that this is instead a sign of something being abnormal with our NanoDrop readings or our elution buffer. However, this should affect both samples equally. Therefore it is likely that this a contributing factor, but that the mutCLoG1-3F and mutCLoG1-3R primers should also be redesigned considering they were the only major difference between the sample that did sequence adequately and the one that did not. Hopefully re-designing these primers and re-doing the sequence after purifying the PCR product with a new elution buffer will lead to conclusive results that will allow us to confirm from the genotype if this plant was in fact rescued from its temperature sensitivity or not.

Rescue Efficiency

One of the main reasons we chose to carry out two types of confirmation experiments on the rescued plant – both the sequencing of other mutation sites and the growth assay – was the low rescue frequency of one plant out of nineteen. Based on this result, we cannot conclusively estimate the rescue efficiency of the transformation. Although it is possible that the efficiency is approximately the one plant out of nineteen observed in these experiments, it is also possible that the frequency is higher – closer to one out of 10 – or much lower – closer to one out of 100 – and a rescued plant was achieved by chance. However, previous studies have shown the efficiency of DNA integration to be much higher than this in *P. patens* (Schaefer and Zryd, 1997). Therefore the extra confirmation steps will allow us to ensure that the low frequency of the rescued phenotype is legitimate and that the one rescued plant did not result from a failed transformation that was contaminated with wild-type DNA.

Utility of Probabilistic Model

As discussed previously, part of this project was dedicated to developing a model for estimating the number of *P. patens* genes which can be mutated to obtain a TS phenotype. It is, however, important to note that much of the model will have to be based off literature. As the model uses Bayes' formula, a prior term is needed. As no information exists about the frequency of TS mutations in *P. patens*, developing this prior will involve using studies performed in other organisms. Therefore the researcher will have to make assumptions about these studies and how closely they resemble the mutation landscape involved in developing our *P. patens* CLoG mutants. These assumptions may vary in accuracy, so the accuracy of estimates resulting from the model will have to be considered with this in mind. However, this is simply the nature of Bayesian statistics; if an accurate prior already existed, the posterior would already be known and this model would not need to be developed. That being said, to increase the accuracy of the model a very diffuse prior will be used, allowing the posterior to be based on a large number of values and therefore not overly influenced by any one data value. This would allow the posterior to be based most heavily on the experimental data and less so on the prior. Despite this caveat on accuracy however, the model will provide us with valuable information about how large-scale a saturation study of CLoG mutations in *P. patens* would have to be.

Future Work

Further Study of CLoG1

Ding's mapping identified CLoG1's causal mutation to be on a gene that has not yet been annotated. Therefore it is possible that we have identified a novel gene that is vital for proper plant growth. Once the rescue experiment has confirmed that this is the proper location, future work can focus on studying the "CLoG1 gene" to determine what its role in cell growth is. This could be done by inserting a GFP sequence in frame with the gene affected by the CLoG1 mutation. If this were done in a wild-type plant, the transcribed protein's localization within the cell could be visualized.

Conclusive Mapping of CLoG4

In addition to studying CLoG1, the plan is to also move forward and do the same things with CLoG4. As mentioned previously, preliminary results were achieved for CLoG4 but they were not conclusive. Therefore the next step is to resolve the issue(s) causing the discrepancies between Ding's plots and my own and to then map CLoG4's mutation using our MATLAB routine. If this gives the same inconclusive results as CloudMap, it would be worth exploring why this is happening before proceeding with other mapping experiments.

Mapping of Other CLoGs

There are a total of six isolated CLoG mutant strains thus far: CLoGs 1-6. Once CLoG1 and CLoG4 have been mapped conclusively and confirmed, work could begin on mapping the other CLoG mutations. Crosses have already been performed for CLoGs 3, 5, and 6 (CLoG2 showed a low – if not non-existent – outcrossing rate), meaning the screening and isolation procedure could begin on these three lines at any time. Therefore once the mapping workflow is adequately reproducible, identifying the mutations responsible for the TS phenotypes of these three lines should take much less time than it has for CLoG4.

WORKS CITED

Alonso, J.M., and Ecker, J.R. (2006). Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in Arabidopsis. Nat Rev Genet **7,** 524-536.

Austin, R.S., Vidaurre, D., Stamatiou, G., Breit, R., Provart, N.J., Bonetta, D., Zhang, J., Fung, P., Gong, Y., Wang, P.W., McCourt, P., and Guttman, D.S. (2011). Next-generation mapping of Arabidopsis genes. Plant J **67,** 715-725.

Bajaj, K., Dewan, P.C., Chakrabarti, P., Goswami, D., Barua, B., Baliga, C., and Varadarajan, R. (2008). Structural correlates of the temperature sensitive phenotype derived from saturation mutagenesis studies of CcdB. Biochemistry **47,** 12964-12973.

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol **Chapter 19,** Unit 19 10 11-21.

Brown, S.M., Ritchie, D.A., and Subak-Sharpe, J.H. (1973). Genetic studies with herpes simplex virus type 1. The isolation of temperature-sensitive mutants, their arrangement into complementation groups and recombination analysis leading to a linkage map. J Gen Virol **18,** 329-346.

Byrt, C.S., Grof, C.P., and Furbank, R.T. (2011). C4 plants as biofuel feedstocks: optimising biomass production and feedstock quality from a lignocellulosic perspective. J Integr Plant Biol **53,** 120-135.

Christou, P. (2013). Plant genetic engineering and agricultural biotechnology 1983-2013. Trends Biotechnol **31,** 125-127.

Cove, D.J., and Knight, C.D. (1993). The Moss Physcomitrella patens, a Model System with Potential for the Study of Plant Reproduction. Plant Cell **5,** 1483-1488.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., and Daly, M.J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet **43,** 491-498.

Ding, X. (2014). Isolation of temperature-sensitive mutants in the moss *Physcomitrella patens* and mapping of their causal mutation by genome sequencing of pooled segregants. BS. Senior Thesis (Major Qualifying Project), Worcester Polytechnic Institute. Advisors: L. Vidali and P. Flaherty

Doitsidou, M., Poole, R.J., Sarin, S., Bigelow, H., and Hobert, O. (2010). C. elegans mutant identification with a one-step whole-genome-sequencing and SNP mapping strategy. PLoS One **5,** e15435.

Durrett, R. (2009). Elementary probability for applications. (New York: Cambridge University Press).

Edgar, R.S., and Lielausis, I. (1964). Temperature-Sensitive Mutants of Bacteriophage T4d: Their Isolation and Genetic Characterization. Genetics **49,** 649-662.

Engel, P.P. (1968). The Induction of biochemical and morphological mutants in the moss *Physcomitrella patens*. American Journal of Botany **55,** 438.

Esposito, M.S., and Esposito, R.E. (1969). The genetic control of sporulation in Saccharomyces. I. The isolation of temperature-sensitive sporulation-deficient mutants. Genetics **61,** 79-89.

Evans, M., Hastings, N.A.J., and Peacock, J.B. (2000). Statistical distributions. (New York: Wiley).

Flaherty, P., Natsoulis, G., Muralidharan, O., Winters, M., Buenrostro, J., Bell, J., Brown, S., Holodniy, M., Zhang, N., and Ji, H.P. (2012). Ultrasensitive detection of rare mutations using next-generation targeted resequencing. Nucleic Acids Res **40,** e2.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W.J., and Nekrutenko, A. (2005). Galaxy: a platform for interactive large-scale genome analysis. Genome Res **15,** 1451-1455.

Goecks, J., Nekrutenko, A., Taylor, J., and Galaxy, T. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol **11,** R86.

Hillier, L.W., Marth, G.T., Quinlan, A.R., Dooling, D., Fewell, G., Barnett, D., Fox, P., Glasscock, J.I., Hickenbotham, M., Huang, W., Magrini, V.J., Richt, R.J., Sander, S.N., Stewart, D.A., Stromberg, M., Tsung, E.F., Wylie, T., Schedl, T., Wilson, R.K., and Mardis, E.R. (2008). Whole-genome sequencing and variant discovery in C. elegans. Nat Methods **5,** 183-188.

Ikehata, H., and Ono, T. (2011). The mechanisms of UV mutagenesis. J Radiat Res **52,** 115-125.

James, G.V., Patel, V., Nordstrom, K.J., Klasen, J.R., Salome, P.A., Weigel, D., and Schneeberger, K. (2013). User guide for mapping-by-sequencing in Arabidopsis. Genome Biol **14,** R61.

JGI. (n.d.). *Physcomitrella patens* V3.0 Early Release.

Lawlor, D.W. (2013). Genetic engineering to improve plant performance under drought: physiological evaluation of achievements, limitations, and possibilities. J Exp Bot **64,** 83-108.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25,** 1754-1760.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics **26,** 589-595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics **25,** 2078-2079.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. J Biomed Biotechnol **2012,** 251364.

Liu, Y.C., and Vidali, L. (2011). Efficient polyethylene glycol (PEG) mediated transformation of the moss *Physcomitrella patens*. J Vis Exp **50,** DOI 10.3791/2560.

Lockwood, S., Krishnamoorthy, B., and Ye, P. (2011). Neighborhood properties are important determinants of temperature sensitive mutations. PLoS One **6,** e28507.

Mardis, E.R. (2008). The impact of next-generation sequencing technology on genetics. Trends Genet **24,** 133-141.

Minevich, G., Park, D.S., Blankenberg, D., Poole, R.J., and Hobert, O. (2012). CloudMap: a cloud-based pipeline for analysis of mutant genome sequences. Genetics **192,** 1249-1269.

Morozova, O., and Marra, M.A. (2008). Applications of next-generation sequencing technologies in functional genomics. Genomics **92,** 255-264.

NCBI SRA. (2010). SRX030894: Physomitrella patens Villersexel library. Retrieved 04-28, 2015, from http://www.ncbi.nlm.nih.gov/sra/SRX030894%5baccn.

Prigge, M.J., and Bezanilla, M. (2010). Evolutionary crossroads in developmental biology: Physcomitrella patens. Development **137,** 3535-3543.

Reis-Filho, J.S. (2009). Next-generation sequencing. Breast Cancer Res **11 Suppl 3,** S12.

Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol **29,** 24-26.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A **74,** 5463-5467.

Schaefer, D.G., and Zryd, J.P. (1997). Efficient gene targeting in the *moss Physcomitrella patens*. Plant Journal **11,** 1195-1206.

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A.H., Nielsen, K.L., Jorgensen, J.E., Weigel, D., and Andersen, S.U. (2009). SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nat Methods **6,** 550-551.

Suzuki, D.T., Piternick, L.K., Hayashi, S., Tarasoff, M., Baillie, D., and Erasmus, U. (1967). Temperature-sensitive mutations in Drosophila melanogaster,I. Relative frequencies among gamma-ray and chemically induced sex-linked recessive lethals and semilethals. Proc Natl Acad Sci U S A **57,** 907-912.

Whitney, S.M., Houtz, R.L., and Alonso, H. (2011). Advancing our understanding and capacity to engineer nature's CO2-sequestering enzyme, Rubisco. Plant Physiol **155,** 27-35.

Zuryn, S., Le Gras, S., Jamet, K., and Jarriault, S. (2010). A strategy for direct mapping and identification of mutations by whole-genome sequencing. Genetics **186,** 427-430.

APPENDIX 1: RECIPES FOR MOSS GROWTH MEDIA

Medium	Recipe
Common elements (for media preparation)	1.03 mM $MgSO_4$, 1.86 mM KH_2PO_4, 45 µM $FeSO_4$ (add as powder), 9.93 µM H_3BO_3, 220 nM $CuSO_4$, 1.966 µM $MnCl_2$, 231 nM $CoCl_2$, 191 nM $ZnSO_4$, 169 nM KI, 103 nM Na_2MoO_4
Liquid $PpNH_4$	Common elements, 3.3 mM $Ca(NO_3)_2$, 2.72 mM Di-ammonium tartrate
Solid $PpNH_4$	Common elements, 3.3 mM $Ca(NO_3)_2$, 2.72 mM Di-ammonium tartrate, 7 g agar
Solid PRMB	Common elements, 3.3 mM $Ca(NO_3)_2$, 2.72 mM Di-ammonium tartrate, 10 g agar. Add 10 mM $CaCl_2$ after autoclaving.
Solid PRM-T	Same as PRMB solid medium, stored 50 mL per bottle.

APPENDIX 2: MATLAB CODE

```
%This script is designed to take in two CSVs of SNPs and compare them so that SNPs included in both files are removed.
%Each CSV must be imported into a table. This was done using the "Import Data" button on MATLAB's GUI.

%It's initial purpose is to compare the list of SNPs for CLoG1's segregant reads mapped to Ppatens_v3.0_251.fasta to the list of Villersexel SNPs
%when mapped to the v3.0 genome. It was also used to do the same thing for SNPs found when mapping CLoG4's direct sequencing reads to the v3.0 genome.
%This was done so that mutations could be identified in CLoG1's segregants which originated from CLoG1's genome, not the Villersexel line or
%naturally-occurring mutations in the laboratory's wild-type line from which the CLoG lines were developed.

SNPs1size=size(SNPtable1);
SNPs2size=size(SNPtable2);
SNPsOI = table;
rowToAdd = [];
i=1;
j=1;

while i<=SNPs1size(1) && j<=SNPs2size(1)
    if SNPtable1{i,2} < SNPtable2{j,2}
        rowToAdd = SNPtable1(i,:);
        SNPsOI = [SNPsOI; rowToAdd];
        i=i+1;
        %print statement optional, for user to monitor progress:
        fprintf('%s %d \n', 'i = ', i);
    else if SNPtable1{i,2} >= SNPtable2{j,2}
            j=j+1
            if SNPtable1{i,2} == SNPtable2{j,2}
                i=i+1;
            end
        end
    end
end

writetable(SNPsOI, 'SNPsOI.xlsx');
```