Central Washington University
ScholarWorks@CWU

All Master's Theses | Master's Theses

Summer 2019

DNA Methylation and Genetic Divergence Associated with an Inducible Defensive Response in Mimulus guttatus

David Farr
Central Washington University, farrda@cwu.edu

Follow this and additional works at: https://digitalcommons.cwu.edu/etd

 Part of the Bioinformatics Commons

Recommended Citation

Farr, David, "DNA Methylation and Genetic Divergence Associated with an Inducible Defensive Response in Mimulus guttatus" (2019). *All Master's Theses*. 1249.
https://digitalcommons.cwu.edu/etd/1249

This Thesis is brought to you for free and open access by the Master's Theses at ScholarWorks@CWU. It has been accepted for inclusion in All Master's Theses by an authorized administrator of ScholarWorks@CWU. For more information, please contact scholarworks@cwu.edu.

DNA METHYLATION AND GENETIC DIVERGENCE ASSOCIATED WITH AN INDUCIBLE DEFENSIVE RESPONSE IN *MIMULUS GUTTATUS*

A Thesis

Presented to

The Graduate Faculty

Central Washington University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

Biology

---

by

David Louis Farr

June 2019

CENTRAL WASHINGTON UNIVERSITY

Graduate Studies

We hereby approve the thesis of

David Louis Farr

Candidate for the degree of Master of Science

APPROVED FOR THE GRADUATE FACULTY

_____ _____
Dr. Alison Scoville, Committee Chair

_____ _____
Dr. Ian Quitadamo

_____ _____
Dr. Linda Raubeson

_____ _____
Dean of Graduate Studies

ABSTRACT

DNA Methylation and Genetic Divergence Associated with an Inducible Defensive Response in *Mimulus guttatus*

by

David Louis Farr

June 2019

Phenotypic plasticity allows many organisms to respond to their environment by changing their phenotype, but the mechanisms to do so are not well understood. Yellow Monkeyflower (formerly *Mimulus guttatus*; now *Erythranthe guttata*) is one such organism that can serve as a model to promote our understanding of these mechanisms due to its striking response to insect herbivory. Monkeyflower responds to leaf damage by increasing the number of hair-like glandular trichomes, a putative defensive trait that reduces the magnitude of damage by insects. This plastic response is transgenerationally inherited in a way that is sensitive to genome-wide demethylation when transmitted through the maternal but not the paternal germline. Investigation of this phenomenon has been hampered by a lack of computational tools to analyze pooled methylome and genome sequence data. In this study, two distinct software pipelines were developed and tested on data from Monkeyflower. The first pipeline detects regions that are differentially methylated and identifies adjacent candidate genes, using Nanopore data. This was tested on data from a Monkeyflower recombinant inbred line (RIL) subject to either parental damage or control conditions. The second pipeline uses pooled DNA sequence data to identify genomic regions that exhibit statistically significant divergence in allele frequencies. This was tested on genome sequence data from an experiment

involving artificial selection for increased trichome production. Results indicate that epigenetic inheritance of the damage response in a particular RIL is associated with 59 differentially methylated regions. Relevant functions, including anatomical structure development and response to abscisic acid, are significantly overrepresented in the set of genes that lie closest to these DMRs. Artificial selection for high trichome production produced one highly divergent region adjacent to a gene associated with seed coat mucilage development. These findings identify candidate epigenetic and genetic factors associated with glandular trichome development while providing an effective test case for the development of two new software pipelines.

ACKNOWLEDGMENTS

I would like to thank my graduate committee mentor, Dr. Alison Scoville, for her extraordinary support and attention to detail as both an educator, advisor, and research mentor whom I have worked with both as a graduate and undergraduate student. Dr. Scoville has shown me great patience as I embarked on a long journey and asked innumerable questions, and always found a way to encourage me to learn and press forward.

Dr. Ian Quitadamo has proven time and again his excellence as a thoughtful, powerful, and inspiring educator and continues to encourage his students to be empowered to think critically and approach the world with a passionate, scientific approach that sets a high standard for communication and compassion. Dr. Quitadamo encouraged me not only to develop a more comprehensive style of scientific communication but also took extra time to help me as I taught at community college and university levels as an instructor for nearly a year.

Dr. Linda Raubeson has proven a fierce advocate for attention to detail and high academic standards and welcomed me into expanding my understanding of genetics through nutrigenomics as a graduate teaching assistant. Dr. Raubeson took the time to see where I was and pushed me to grow as an academic.

I applied for and received several rounds of grant funding, enabling this research. Specifically, I would like to thank the CWU Office of Undergraduate Research, the CWU College of the Sciences – SURE, the CWU Biology Mycology, and Botanical Research Fund, the Washington State Distinguished Fellowship in Biology, and the CWU Graduate Studies department for their generous financial support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Phenotypic plasticity allows organisms to adapt to their environment by mounting a phenotypic response to challenges. This sometimes results in dramatic variation, such as the increase of glandular trichomes in plants--hair-like structures that can secrete chemicals to effectively deter insects from damaging leaves. The mechanisms used by organisms to achieve this striking variation are not well known, though the study of the specific genetic and epigenetic changes involved are often accomplished through the use of genome and epigenome sequencing. As sequencing technology has become less expensive and more widely available, computational methods to analyze the resultant datasets have lagged. In response, the development of new software and analysis pipelines has become imperative. This study involves the development of two novel analysis pipelines, as well as their application to existing datasets from experiments addressing the epigenetic and genetic control of glandular trichome production in Yellow Monkeyflower.

New Methods in Genome Sequencing

The development and use of DNA sequencing technology have progressed rapidly over the last few decades, resulting in a demand for high-throughput genomic analyses that has at times outpaced the availability of appropriate computational methods. This has been a major driver for researchers to not only develop new methods in computational analysis but to automate and streamline multiple steps in data processing. One of the core informatics strategies to address these issues is the development of software pipelines. In

addition to their ability to improve analytical methods, pipelines can increase accessibility for the broader community of biologists by bridging the fragmented series of programs that are often needed to produce straightforward results.

As the efficiency and availability of sequencing technology have increased, the cost associated with sequencing DNA has greatly decreased. Whole-genome sequencing (WGS) has become more prevalent, enabling more biologists to use industry-standard Illumina and emerging Nanopore technology (Besser et al. 2018) to answer compelling questions about genomic divergence and differential regulation associated with adaptation. Nanopore has emerged as one of the most affordable next-generation sequencing (NGS) methods, and its value can be extended by pooling the DNA of multiple individuals from a population before WGS. Pooled WGS allows for estimating allele frequencies within a group of individuals, without the need to separately sequence each individual. This type of experimental design works well for detecting genomic regions associated with variation between groups that differ in one key trait. Developing software that accommodates this design can provide a valuable platform for biologists investigating any species.

Phenotypic Plasticity in Yellow Monkeyflower

Mimulus guttatus, recently renamed *Erythranthe guttata* and commonly known as Yellow Monkeyflower, is an emerging model organism that displays striking phenotypic variation between geographic populations. This species serves as an excellent test case to develop software that can detect genetic and epigenetic divergence associated with defensive traits. Like many plants (Maes & Goosens 2010, Huchelmann et al. 2017, Scoville et al. 2011), Yellow Monkeyflower naturally produces two distinct types of

trichomes – hair-like appendages on leaves and other aerial tissues that can reduce the frequency and intensity of herbivory (Coliccio et al. 2013, Holeski 2013). Interestingly, populations from Point Reyes Natural Seashore (PR) and Iron Mountain (IM) in the Cascade mountains differ significantly in their trichome production. In Point Reyes, Monkeyflower grows in dense perennial clusters that experience more frequent insect interactions. Members of this population produce more glandular trichomes, which secrete chemicals that have been shown to passively deter insects (Holeski 2007, Holeski et al. 2013, Harborne 1993). In the Iron Mountain population, Monkeyflower grows as a smaller, sparser annual that experiences fewer insect interactions. Members of this population primarily produce structural trichomes that do not secrete chemicals but may actively restrict insect movement and inhibit egg deposition (Levin 1973).

In addition to variation in constitutive production of glandular or structural trichomes, a significant increase in glandular trichomes can be induced by simulating insect damage to leaves (Holeski 2007). The effects of this induction are readily seen in subsequent generations produced by damaged parents, where the effects not only persist for three generations (Akkerman et al. 2016, Holeski 2007), but the combination of maternal and paternal damage produces a sex-dependent, additive increase in glandular trichome production (Akkerman et al. 2017). Despite strong evidence that trichome development in Monkeyflower is a phenotypically plastic trait (Akkerman et al. 2017, Holeski 2007), the genetic architecture and differential regulation involved in this rapid response are poorly understood.

The status of Yellow Monkeyflower as an emerging model for ecological genomics is especially useful to ongoing research in evolutionary biology. Scientists who

seek to develop universal analytical methods often use well-studied models. This allows for greater reproducibility and the use of rich existing data sets, such as an annotated reference genome. Monkeyflower is an attractive model for advancing and testing computational methodology due to existing genomic resources. Also, because of the role glandular trichomes play in plant defenses and production of medically or economically important compounds, our understanding of this trait has important implications for broader agricultural and ecological study.

Software Development for Genetic and Epigenetic Analysis

Genome-wide association studies (GWAS) are often conducted by comparing individual genotypes that vary in a specific phenotype. Using this basic premise, genetic variations associated with higher baseline trichome production can be detected based on comparisons between naturally divergent populations such as IM and PR. More specifically, genetic divergence detected at the level of SNPs in a pooled WGS study can expand our understanding of the genetic architecture for trichome production. One of the first test statistics that allow for a test of divergence between populations is B*, developed by Kelly (2013). The B* test identifies windows of SNPs that exhibit significant divergence in allele frequency. These windows can then be compared with the published, annotated *Mimulus* reference genome (Helsten et al. 2015) to generate a list of candidate genes that may be involved in producing related phenotypic variation.

Nanopore sequencing can extract the DNA sequence of long fragments. Importantly, it also outputs raw signal data that can be used by Nanopolish (Simpson et al. 2017) to detect specific patterns of methylation in DNA, which can contribute to epigenetic transcriptional regulation without any change in a DNA sequence. Applied to

WGS, Nanopolish can detect methylation on cytosine nucleotides, allowing for the construction of a methylome dataset.

These methods for separately analyzing pooled, whole-genome sequences for genetic and epigenetic data analysis are new (Kelly et al. 2013, Simpson et al. 2017) and therefore have significant gaps in translation of computational methodology as a result of having no public software release, or requiring a complex process for analysis that can present a barrier for novice researchers. Using the genomic data sourced from two separate experiments, I use a three-fold approach to further elucidate the molecular architecture of glandular trichome development in Yellow Monkeyflower:

1. Develop a software pipeline for genome analysis that identifies significant divergence in allele frequencies based on original C# development by Farr and translated to an R package by McKinnon (unpublished work, 2019).
2. Apply the proposed genome analysis pipeline to an existing variant call format file resulting from Neuffer (2015) to identify significant single nucleotide polymorphisms (SNPs) associated with evolutionary divergence due to artificial selection for increased glandular trichome production.
3. Use Nanopore whole-genome sequencing of pooled tissue samples from Akkerman et al.'s (2017) experiment on epigenetic inheritance to discover significant differentially methylated regions (DMRs) associated with increased glandular trichome production. As a principal component of this analysis, a pipeline will be proposed that handles the Nanopolish (Simpson et al. 2017) methylation frequency data processing and DMR-

Scan (Colicchio et al. 2018) analysis. This pipeline will extend the original functionality of DMR-Scan to include identifying gene, annotation, and whether or not each DMR is nearby a coding sequence or regulatory region.

This comprehensive approach will contribute to our understanding of specific methylation and evolutionary patterns in Yellow Monkeyflower as well as expand the toolset for analysis in future research on any organism.

CHAPTER II

LITERATURE REVIEW

Glandular Trichome Production in Response to Biotic Stress

Glandular trichomes are found in more than 30% of vascular plants (Huchelmann et al. 2017) and are found on leaves where they form from extensions of the plant epidermis. Trichomes exist as specialized unicellular or multicellular structures that contribute to the secretion of tannins, essential oils, mucilage, and resinous structures (Levin 1973). In Monkeyflower, glandular trichomes secrete phytochemicals, including phenylpropanoid glycosides (PPGs), that likely contribute to defense against insect herbivores (Holeski et al. 2013, Holeski 2007, Scoville et al. 2011). An increased number of such polyphenolic secondary metabolites has been associated with a decreased rate of herbivory (Coley et al. 1985; Holeski et al. 2013). Plants that produce and secrete these phytochemicals incur a continuous energetic cost to maintain even small concentrations in their trichome secretions. The half-lives of such phytochemicals can vary broadly from 10 hours to six days among some agricultural species, although some can be recycled during senescence (Coley et al. 1985). Holeski et al. (2013) observed an increase in PPG concentration as an induced response when Monkeyflower leaves were damaged, which suggests a plastic mechanism for the plant to increase its defensive capabilities in response to simulated leaf herbivory.

Epigenetic Mechanisms and Trait Plasticity

When environmental factors trigger a change in phenotype, this phenomenon is referred to as phenotypic plasticity. Such phenotypic changes often involve epigenetics,

defined as the sum of regulatory mechanisms that dictate whether a gene can be transcribed, without resulting in permanent modifications to DNA. Multiple forms of epigenetic regulation exist together in most organisms (Maunakea et al. 2010; Satyaki 2017). However, one of the most common forms of epigenetic modifications observed in plants is the presence of cytosine-phosphate-guanine (CpG) modifications. CpG commonly appears as large islands, or repetitive sequences of CpG nucleotides along the same strand where a methyl group has been added to the cytosine. As the organic methyl group is added to a CpG, the cytosine is converted into 5-methylcytosine (Jablonka & Raz 2009) 2009). The presence of CpG modifications that occur in a promoter region, typically preceding a coding gene, prevent transcription factor proteins from binding to the promoter to initiate transcription of the coding sequence (Grant-Downton & Dickenson 2005). The effects of this mechanism generally act as a way to silence a gene (Finnegan et al. 1998); however, if a protein product such as a repressor is silenced from transcription, an increase in transcriptional activity of another coding gene may be apparent in the phenotype.

Germline epigenetic regulation is compatible with the transgenerational inheritance of epigenetically regulated genes, similar to the process of gene imprinting (Satyaki & Gehring 2017). Much of this process was summarized by Satkayi (2017) as a general mechanism for the process of gene imprinting in plants. While gene imprinting is only one potential mechanism for epigenetic inheritance, it is an example of how parents contribute to methylation patterns in offspring. Methylation is frequently established from the germline contributions of the paternal line where the MET1 protein, a form of methyltransferase studied in the model organism *Arabidopsis thaliana*, can establish a

fully methylated 5-methylcytosine from a hemimethylated cytosine that acts in opposition to demethylase (DME) activity. In the central cell of the plant ovule, DME upregulation is coupled with low expression of MET1, whereas in sperm cells DME is absent and MET1 is expressed at high levels. This contributes to a classic presentation of a hypomethylated ovule and a methylated pollen grain where the hemimethylated CpG residues are targets for MET1 and are subsequently methylated.

5-azacytidine (5-aza) has been shown to decrease methylation in progeny and has been used previously as a human cancer treatment by inhibiting the ability for methyltransferase, specifically MET1, which is associated with high levels of activity during replication and in the plant embryo (Christman 2002; Satyaki 2017). In Akkerman et al. (2016), damage to parental Monkeyflower resulted in a significant increase in glandular trichome density in progeny. When both the maternal and paternal parents had been damaged, an additive effect on the increase of trichomes was noted. Treatment of the seeds with 5-aza resulted in the loss of the maternally transmitted high-density response. This suggests that 5-aza directly antagonizes the process by which the maternal response to damage is passed on to offspring. In contrast, the paternal effect of damage was not erased by 5-aza treatment, suggesting an unknown alternate mechanism involved in paternal epigenetic inheritance. While it is possible that paternally associated RNAs could re-establish the damage-induced paternal methylation patterns after treatment with 5-aza, this possibility remains unexplored and yet confirms a transgenerational, epigenetic basis for regulation of trichome production in response to simulated herbivory (Akkerman et al. 2016).

Genetic Adaptation for Plant Defense

The specific genetic architecture that may be responsible for glandular trichome generation in Monkeyflower is not well understood. The basis for phenotypic variation in many organisms, however, can originate from both genetic differences and epimutation (Cohen 1999, Colicchio 2017, Morishita et al. 2012). Transposable elements make up a large component of many plant genomes – as much as 90% of Maize – and are common, mobile vehicles for methylation (Underwood et al. 2017). Changes to epigenetic silencing of these vast arrays of transposons can provide a pathway to phenotypic variation. In addition, spontaneous conversion of a 5-methylcytosine to a thymine base can introduce further deleterious, favorable, or silent genetic variation (Morishita et al. 2012).

Due to the cost associated with a regular turnover of energetically expensive phytochemicals within trichomes (Colicchio 2017, Coley et al. 1985) constitutive expression of high numbers of glandular trichomes is unlikely to be a favorable trait. It has been observed that Monkeyflower trichome production is associated with a decreased magnitude of leaf damage by herbivory rather than a decrease in the frequency of herbivory events, which suggests a somewhat proportional response and therefore a tradeoff in trichome development (Colicchio 2017). These effects have been observed in a variety of isolated populations of Monkeyflower and vary drastically depending on the frequency of insect interactions as well as growth conditions (Colicchio 2017).

The existence of allelic and phenotypic diversity allows for a relatively straightforward study design where artificial selection can be used to capture the genomic signal observed in a partial sweep. In a population of RILs created from a single F1

individual, single nucleotide polymorphisms – the most fundamental unit of genetic variation – can be analyzed from the standpoint of reference versus alternate allele frequency (Kelly et al. 2013). When artificial selection is used to select for one specific phenotype, such as glandular trichome density, a resulting increase in allele frequency divergence can be observed (Kelly et al. 2013, Neuffer 2015). The linkage disequilibrium existing even after multiple generations of cross-breeding requires a conservative methodology; however, an additional benefit of analyzing SNP-level variation between sample populations is a decrease in confounding epistatic interactions (Kelly et al. 2013).

Review of Common Computational Analyses

One of the traditional approaches to the discovery of genetic variation associated with a phenotype of interest is genome-wide association studies (GWAS), which has played a major role in understanding the genetic basis of human disease and pathology (Bush & Moore 2012). In GWAS experimental design and analysis, the general goal is to identify SNP variation between populations; however, the methods traditionally used in GWAS (Bush & Moore 2012) can preclude analysis of pooled genomic data where contributing individuals of each sample cannot be determined. Furthermore, pooled genome sequencing applied to a GWAS-style construct cannot be analyzed for the effects of linkage disequilibrium (Kelly et al. 2013).

Traditional GWAS F_{st} analysis, as well as the B* test proposed by Kelly (2013) are used to identify genetic divergence. However, careful interpretation is required due to the potential for a non-causal variation to appear associated with the studied trait. The B* test is more conservative and, when paired with the GenWin (Beissinger et al. 2015) analysis, is less likely to miss or obscure windows of SNPs associated with causal

variation. Use of GenWin reduces the probability of choosing a window size, according to the methods proposed by Kelly (2013), that arbitrarily weights calculated B-value divergence too high or low. Without the discovery of a median window size for analysis, a window that is too small will result in very noisy data and will tend to inflate the number of regions that appear to be significantly different (Kelly 2013 and Beissinger 2016). Windows that are too large result in a higher probability of missing a truly significant window due to the frequency of repeats such as transposable elements that occur in large regions throughout the genome.

Python is one of the most commonly used computational languages in bioinformatics research, due in part to its agnostic treatment of various operating systems and its cost-effectiveness. Many of the foundational tools used for the processing of raw genomic data produced through sequencing methods are written in Python, such as SAMTools, minmap2, and Nanopolish for nanopore data published by Li (2009, 2018) and Simpson (2017) respectively. While these tools are frequently utilized, the computational experience required to correctly install their dependencies and carry out their pipelines is extensive. Fortunately, many of the developers of these and other essential applications released their work under MIT, GNU, or open-source licensing and this makes their use a reliable component of software pipelines that handle data throughout a sequence of various processing methods. This also allows for the introduction of original code designed to summarize, display, or connect standalone software in novel pipelines. The development of these pipelines can improve general accessibility by decreasing the technical experience required (Leipzig 2017), as well as by integrating existing software to achieve a novel output method.

"R" is another common statistical analysis programming language which, like Python, is free for common use and focuses on statistical and graphical analysis and generation rather than the potentially more diverse user base for Python in commercial software development. RStudio is a successful integrative development environment (IDE) for working with R, which contributes to its accessibility for researchers that lack extensive software development experience. Perhaps one of the most important features of R is that published R scripts and packages can more easily allow for the installation of required third party R libraries, reducing or eliminating the need for a user to manually install dependencies by simply including simple commands in the R script itself. This has resulted in a wide array of bioinformatics tools that are either in R or Python, appealing to a wide array of researchers interested in -omics research such as genomics, transcriptomics, and methylome applications.

CHAPTER III

METHODS

Tissue Collection and Library Preparation

Mimulus guttatus RIL 85 served as a single source line for a full factorial experiment measuring the effects of the damage and 5-azacytidine treatment as detailed in Akkerman et al. (2016). The goal of the original experiment was to examine whether induction of the mechanism to increase glandular trichome damage in response to simulated insect herbivory could be transgenerationally inherited and if it was sex-dependent. Preserved leaves from the progeny of samples that had both maternal and paternal damage, as well as progeny of parental samples that had no damage, were used in this study. Samples from the progeny of the maternal and paternal damage can also be referred to as double-damaged, and the effects of this damage resulted in evidence for additive effects in Akkerman et al. (2016). For each of these treatments (double damage or no damage), samples from the 6th leaf pair for each of the 6 individual progeny of the 6 independent parent pairs were stored in liquid nitrogen, resulting in a total of 36 individual plants in each treatment.

In preparation for DNA extraction, the leaf tissue samples for each of the two groups were thawed and ground as preparation for pooled sequencing. Pooled DNA from each group was then extracted using a standard Urea extraction (Appendix B) and the sequencing library was prepared according to the PCR-free Oxford Nanopore Technologies protocol.

Methylome Sequencing

Methylome sequencing was performed using a Minion nanopore sequencer from Oxford Nanopore Technologies, according to the manufacturer's protocol for 1D Genomic DNA by ligation (SQK-LSK108, Version GDE 9002 v108 revU 18Oct2016), using three R9.4.1 flow cells for pooled DNA from undamaged parents and four R9.4.1 flow cells for pooled DNA from damaged parents. Raw sequence data were collected from each flow cell for 48 hours, using ONT's MinKNOW program. MinKNOW's Albacore real-time basecaller was used and the default read quality control sorted the raw read fragments according to the Mimulus v.2.0 reference genome. Once the base-called FASTQ data was obtained, reads that passed quality control were analyzed using the recommended pipeline for Nanopolish (Simpson et al. 2017) for detection of 5-methylcytosine in conjunction with aligning, sorting, and mapping the basecalled data to the *Mimulus guttatus* v2.0 reference genome (Helsten et al. 2015) obtained from Phytozome (Neupane *et. al* 2011). To obtain the 5-mC calls, the call-methylation function of Nanopolish was used. This produces individual log-likelihood ratio probabilities of methylation for every resulting methylated fragment, which can be summarized by using the methylation frequency function to generate a concise table of methylation frequencies. This process was replicated to generate frequency data for both undamaged and double (parental) damaged groups.

After sequencing and mapping, Qualimap 2.2.1 (Okonechnikov et al. 2016) was used to generate a summary report on the damaged and undamaged genomes to aggregate statistics and information about the respective BAM files using the default settings.

Finally, Samtools "stats" (Li et al. 2009, 2011) was used to evaluate the average read/fragment length for each sample.

Genomic Analysis from Artificial Selection

After the artificial selection experiment conducted by Neuffer (2015), a variable-call format (VCF) file was generated based on pooled-sequencing of multiple populations: the source population, two replicate control populations, and two replicate treatment populations. The control and treatment populations were successively bred for four generations. In each generation, 30 plants were selected for breeding based on a random number generator (control populations) or their status as the highest trichome producers in the populations (treatment populations). The pooled genomic data that contributed to the VCF was sequenced using Illumina technology at the University of Kansas. McKinnon helped generate an R package (unpublished work, 2019) that incorporated a C# application developed by Farr, based on an original unpublished Python script to run the analysis proposed in Kelly (2013) after identifying the median window for analysis of SNPs from the VCF file based on GenWin (Beissinger et al. 2015). These methods were applied to the existing VCF file sourced by Neuffer (2015) where the final output of the pipeline, developed by Farr and ported to R by McKinnon (2019, unpublished work), results in a file that provides a B* test statistic, a BH-adjusted P-value (Benjamini & Hochburg 1995), and the genomic position of the mid-point of each sliding window.

The B* test (Kelly et al. 2013) involves calculating a B value for every window of SNPs defined by the median value calculated from GenWin (Beissinger et al. 2015) in the dataset based on allele frequency differences. The median window was calculated to

be 7. The B and B*, which is a test statistic tractable to a Chi-square distribution, is calculated and reported for every window. This genomic position was then compared to the *Mimulus guttatus v.2.0* (Hellsten et al. 2013) annotated reference genome published online on Phytosome (Neupane et al. 2011) to identify the closest gene, using BEDTools "closest" (Quinlan et al. 2010).

Methylome Analysis

Methylome analysis was accomplished by modifying the DMR-Scan R-Script from Colicchio (2018) to accept output from Nanopolish's methylation frequency function. The script was further modified to allow for analysis of our experimental design which only contains CpG methylation and the two pooled samples. Further settings, such as the use of Changepoint (Yokoyama et al. 2015) PELT manual penalty of 1.4 was used based on implementation of previous research using similar methylation data analysis (Colicchio et al. 2018). Only methylation frequencies that differed by a minimum of 4% were included in the analysis. Consistency of differential methylation in the resulting putative DMR regions was analyzed after scaling and re-centering the segments defined by Changepoint. A modified generalized linear model was used to predict methylation frequency as a function of parental treatment, with a logit link function and binomial distribution of error terms, using the lme4 package (Bates *et. al* 2015). This model provides p-values for individual DMRs. The integrated "p.adjust" function was used to perform a Benjamini-Hochberg (1995) correction for multiple comparisons and false discovery rate analysis on the results, where multiple comparison adjusted p-values of

less than 0.05 were used to indicate significantly differentially methylated regions and were retained for further analysis.

Once the modified DMR-Scan script provided the list of significant DMRs, the BEDTools program (Quinlan & Hall 2010) was used to discover proximity to the closest gene using the "closest" function, and then whether the DMR intersected ("intersect" function) with a coding domain sequence (CDS) or regulatory 5'-UTR region according to the *Mimulus guttatus* v2.0 repeat-masked assembly and annotation.

Gene Ontology Enrichment & Analysis

Once the closest genes to each DMR were identified, BLAST annotations were obtained by using BEDTools "intersect" (Quinlan et al. 2010) to map these genes to function. The list of genes was subject to gene ontology (GO) enrichment analysis through PlantRegMap (Jin et al. 2017), which utilized a Fisher's Exact Test with a manually defined $\alpha = 0.05$ to discover DMR-associated GO terms that were significantly over-represented. ReviGO (Supek et al. 2011) was used to summarize and visualize the GO enrichment results, using the default settings. Due to a small number of terms resulting from the genomic analysis, no GO enrichment was performed on the genomic dataset; however, the GO terms associated with each significant result were identified using Dicot Plaza 4.0, an online tool for searching data from many dicot organisms.

CHAPTER IV

RESULTS

Differentially Methylated Regions

Qualimap (Okonechnikov et al. 2016) reports indicated that for the double damaged (maternal and paternal damage) sequence data, the mean genomic coverage for scaffolds 1-14 was 2.77 with an overall mapping quality of 22.65 (Phred score) and a general error rate of 18.75%. The error rate is calculated as the ratio of total collected edit distance vs. the number of alignment mismatches reported by SAMTools (Okonechnikov et al. 2016). These numbers included scaffolds beyond the total number of chromosomes in Monkeyflower, so the actual general error rate may somewhat lower than reported. One of the contributing factors to this error rate could be a large number of sequence scaffolds that are unlocalized between the reference and sample genome sequences. For the undamaged sequence data, the mean genomic coverage was 1.40x with a mean mapping quality of 21.28 and a general error rate of 20.47%. Again, these reports were calculated including scaffolds above the total number of chromosomes, so these numbers may be inflated. The average read/fragment length reported by SAMTools (Li et al. 2009) for the double damaged sample was 1,589 base pairs, and for undamaged the length was 2,114 base pairs.

The resulting data indicated 59 unique DMRs, identified from the pipeline developed for the methylation analysis (available at https://www.github.com/davidfarr). Of these, 17 DMRs intersected genomic features such as 5'-UTR regulatory regions or CDS coding sequences. One DMR was long enough to intersect both of these key features. These 17 DMRs are summarized in Table 1.

Table 1 Results of Nanopolish and DMR-Scan analysis pipeline. Each row represents an individual DMR based on differences to mean methylation frequency. Only DMRs that were intersecting or within 2 kilobases of a genomic feature are listed. Distance in base pairs and DMRs that are not near a genomic feature may be found in the appendices. The difference in mean methylation is calculated based on the methylation frequencies for methylated segments of the parental damaged and undamaged individuals. P-values are FDR adjusted for $\alpha = 0.05$.
Significance: All are P <0.05; *P < 0.01; **P < 0.001; ***P < 0.0001.

CHR	Start BP	Diff. Mean Methylation	DMR Size (BP)	Nearest Gene	Associated Genomic Feature
1	28582	8.00E-02	329	Migut.A00002	5'-UTR
5	34877	-4.97E-02	623	Migut.E00005	5'-UTR
6	31237	-1.03E-01	115	Migut.F00003	5'-UTR
7	31750	9.64E-01	13	Migut.G00001	5'-UTR
10	29964	-2.12E-01	65	Migut.J00005	5'-UTR
10	48388	-3.49E-01	22	Migut.J00007	5'-UTR
11	10821	-4.20E-02	545	Migut.K00001	5'-UTR
11	12382	-4.97E-02 **	1126	Migut.K00001	5'-UTR
11	65607	-2.57E-01 ***	62	Migut.K00005	CDS
11	74531	-2.99E-01 *	58	Migut.K00008	5'-UTR
11	79413	6.43E-01	13	Migut.K00009	5'-UTR
13	2796	-8.51E-02 *	218	Migut.M00001	5'-UTR
13	30805	5.99E-01 *	16	Migut.M00003	5'-UTR
14	17052	-2.64E-01	54	Migut.N00004	5'-UTR, CDS
14	18012	-4.41E-02 ***	524	Migut.N00004	5'-UTR
14	18012	-4.41E-02 ***	524	Migut.N00005	5'-UTR
14	27582	8.75E-01	15	Migut.N00008	5'-UTR

The complete list of DMRs was converted to a format compatible for processing through the GO Enrichment analysis described in the methods above. The GO enrichment process uses Fishers Exact Test and FDR to produce a list of GO terms categorized by biological process, cellular component, and molecular function. The complete list is provided in appendix A and has been summarized according to biological process in Table 2. The full GO term results were analyzed and visualized with Revigo in Figure 1.

DMR GO Enrichment Analysis

Table 2 Summary of the GO enrichment analysis results related to biological process. Fisher's Exact Test was used to identify ontological terms that are significantly overrepresented based on the list of all genes identified from the DMR analysis. An FDR of 0.05 was used to account for multiple comparisons and in all cases, the q-value for the terms below was 1. Significance: All are $P < 0.05$; *$P < 0.01$

GO Term	Annotated	Count	Expected	p-value
response to abscisic acid	286	3	0.39	0.0065 *
response to alcohol	330	3	0.45	0.0097 *
response to lipid	370	3	0.51	0.0132
single-organism carbohydrate metabolic process	382	3	0.52	0.0144
single organism reproductive process	758	4	1.04	0.0175
developmental process involved in reproduction	830	4	1.14	0.0237
shoot system development	486	3	0.67	0.0271
response to hormone	878	4	1.2	0.0285
reproductive process	937	4	1.28	0.0351
response to endogenous stimulus	939	4	1.29	0.0353
carbohydrate metabolic process	971	4	1.33	0.0393
reproduction	239	2	0.33	0.0416
anatomical structure development	1459	5	2	0.0417
response to acid chemical	596	3	0.82	0.0455

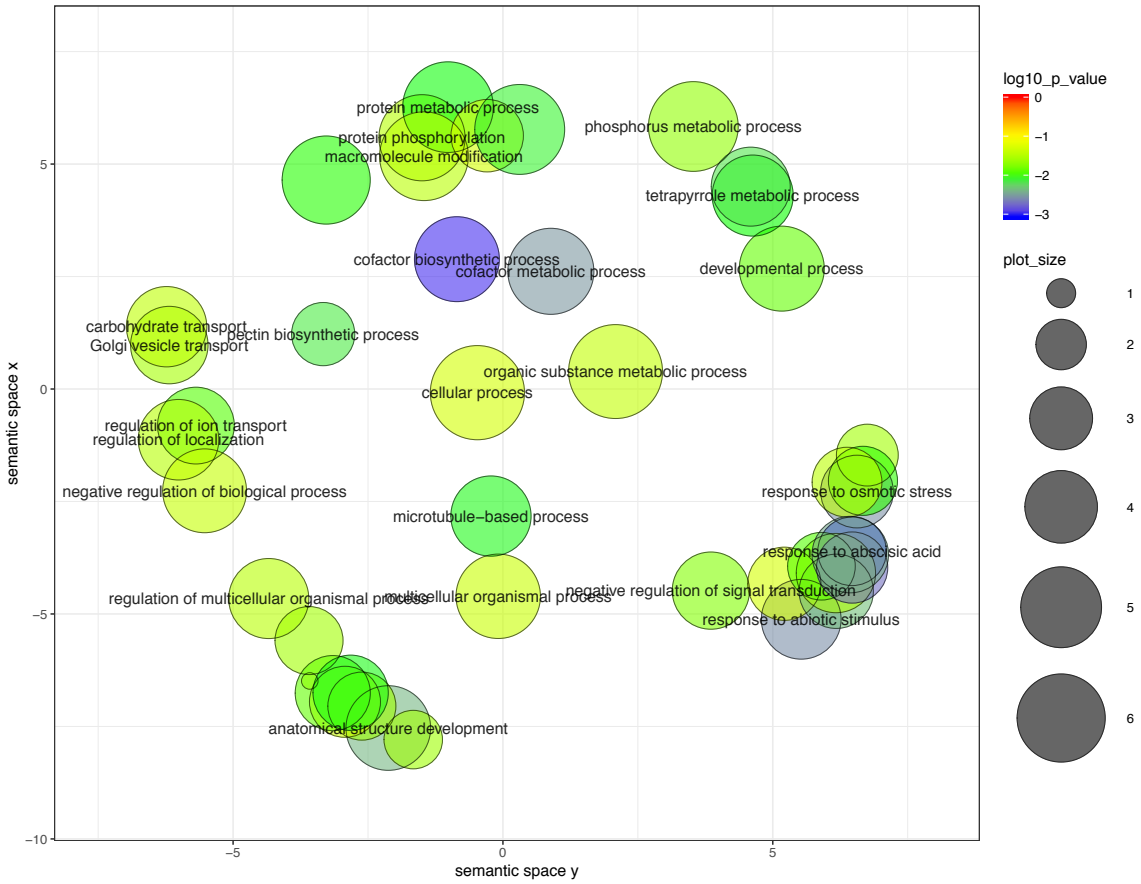

Figure 1 Visualization of GO enrichment analysis results provided by Revigo, which uses ontological terms to group genes by similar function in a semantic space. Circle size refers to the number of DMR-associated genes that fit the individual term and color indicates the significance of each term. Only terms that meet a dispensabilility threshold of 0.35 are included.

Genetic Analysis Results

The BAM files that contributed to the VCF included variants from a single pooled DNA sample from each of the following populations: the source population, control replicate 1, control replicate 2, treatment replicate 1, and treatment replicate 2. Genome sequence from IM 767 was also included to polarize the VCF format so that results would be expressed in terms of the frequency of the allele from the founding IM 767 parent. This analysis compared allele frequencies in the source and control populations to those in the replicate treatment populations to identify genomic regions exhibiting significant divergence. The BAM file was generated at the conclusion of research by Neuffer (2015) but had not yet been analyzed in detail.

Of the 3,756,767 SNPs in the initial B* calculation, this results in 1,073,360 reported windows with a B* and p-value reported. To adjust for multiple corrections and false discoveries, the "p.adjust" function was used using the FDR a.k.a "BH" method. After FDR analysis, adjusted P-values at an $\alpha = 0.05$ are reported in Table 3 along with nearest genes, genomic features, and distance to feature.

The results in Table 3 were compared to the list of differentially methylated regions in the event that there were any common genetic or epigenetic changes between the two separate experiments that would show support for a common mechanism. However, none of the significant SNPs based on the artificial selection experiment matched any DMRs from the epigenetic experiment.

Table 3 The list of candidate genes after B* analysis. BP values correlate to the middle or end of the window of 7 SNPs as identified through GenWin (Beissinger et al. 2016) and the B* value is the test statistic reported for that window. Distances were reported to the nearest 5'-UTR or CDS when it occurred within two kilobases, with a positive distance indicating a region of genomic divergence that is upstream of the feature. Adjusted P-values were rounded to six significant figures and their similarity is likely associated with a large number of results with a P-value of 0.5, which can affect the adjusted p-value. P-values before correction for multiple comparison testing will be available on https://github.com/davidfarr.

CHR	BP	B*	Adjusted P-value	Gene	Feature	Distance from Feature (BP)
10	18469191	25.8936	0.03998	Migut.J01741	5'-UTR	642
10	18547232	28.9589	0.03998	Migut.J01753	NA	NA
10	18547259	31.2980	0.03998	Migut.J01753	NA	NA
10	18547277	27.9977	0.03998	Migut.J01753	NA	NA

The list of genes in Table 3 was not large enough to conduct a GO enrichment analysis. Instead, the gene ontology was accessed using Dicots Plaza 4.0 based on Plaza Integrative Orthology and InterPro. These results are summarized in Table 4.

Table 4 Summary of the gene ontology for the genes reported in Table 3. An AT source refers to a known ontology from *Arabidopsis thaliana*.

Gene	Biological Process	Molecular Function	Cellular Component	Provider	Source
Migut.J01741	mucilage biosynthetic process involved in seed coat development	1,4-beta-D-xylan synthase activity	Golgi apparatus	PLAZA Integrative Orthology	AT3G10320
Migut.J01753	NA	binding	NA	InterPro	NA

CHAPTER V

DISCUSSION

Epigenomic Analysis Pipeline

One of the key goals of this project was to expand on and generate a pipeline that can locate differentially methylated regions based on Nanopore data. The first stage of the pipeline can be considered to be the actual sequencing itself. Nanopore sequencing technology allows smaller institutions and groups to perform genetic sequencing with lower startup costs than Illumina (Besser et al. 2018). One of the greatest remaining challenges using the Nanopore device for whole-genome work – whether the goal is epigenetic or genetic analysis – is managing historically lower base mapping accuracy and read quality compared to other next-gen methods such as classic bisulfite sequencing (Simpson et al. 2017). However, Nanopore technology is subject to frequent revision and improvement by Oxford Nanopore Technologies, incrementally increasing its read quality and accuracy while maintaining the small form factor of the nanopore device itself (Oxford Nanopore Technology).

Popular methods and software for working with "-omic" work is built on an assumption that input data will come from Illumina, which results in fragmentation of the data processing pipeline for researchers utilizing Nanopore technology. For example, Nanopolish (Simpson et al. 2017) provides comprehensive documentation for analyzing CpG methylation in Nanopore-based sequence data based on log-likelihood probability. However, Nanopolish output is not directly compatible with other scripts, such as DMR-Scan (Colicchio et al. 2018), that were initially designed to accept Illumina data for downstream analysis.

To generate a pipeline for analysis of Nanopore data, an R script was developed that included Bash programming to expedite the process of generating a methylation frequency file that can be enumerated into the DMR-Scan R script. The current version of this new pipeline requires users to have the necessary python components installed prior to use, such as Nanopolish (Simpson et al. 2017), minmap2 (Li 2018), BCFTools (Li 2011), SAMTools (Li et al. 2009), BEDTools (Quinlan & Hall 2010), and their dependencies. All R libraries are installed and called running the pipeline R script and simply require R to run. The input format for the actual DMR analysis can be difficult to visualize into figures, so a report-generating tool such as Qualimap (Okonechnikov et al. 2016) is useful to summarize and visualize basic data and runs as a standalone third-party application.

The public release of the nanopore methylation pipeline will appear on https://github.com/davidfarr.

Genomic Analysis Pipeline

The statistical methods originally proposed by Kelly (2013) did not come with a public release of a script or software to complete the analysis. Kelly generated a series of unpublished python scripts that were hard-coded to support Neuffer (2015). In Farr (2019) an initial C# language translation of the python scripts was used to present preliminary data and software methods for discovery of SNPs and nearby genes that are associated with increased constitutive production of trichomes in response to artificial selection. The C# program was expanded to take advantage of a more object-oriented programming methodology and cater to a more diverse set of needs.

With the help of McKinnon (unpublished work, 2019), the software pipeline that was generated to run the B* analysis was converted to an R package, which is expected to expand its user base to more researchers – especially those who are using UNIX based systems such as Linux or macOS without dependence on third-party frameworks. The final version of the R package will allow for a range of user options so that the B* analysis can be used to locate SNPs displaying evolutionary divergence from any VCF input where the contributing BAM files are representative of pooled genomic DNA.

Differentially Methylated Regions

Using the pipeline for discovering differentially methylated regions, a total of 59 DMRs were located along with their closest genes, features, and gene ontology enrichment (Table 2, Figure 1). The full list of DMRs, which includes DMRs that were not within 2 kilobases of a CDS or 5'-UTR, are located in the appendices. All of the DMRs were returned as significant as a result of the analysis from DMR-Scan (Colicchio 2018). Many of the GO terms reported during enrichment were biologically interesting. Genes associated with terms for anatomical structure development and growth, as well as response to endogenous stimuli and abscisic acid, were particularly interesting as they support the biosynthetic and structural growth of trichomes as well as known plant signaling pathways that respond to damage (Colicchio et al. 2015). The exact mechanism by which Monkeyflower upregulates glandular trichome production is not well understood, so the results do not confirm a specific mechanism, but instead support a series of methylation changes that may be associated with the epigenetic response to simulated insect damage to leaves.

Candidate Genes for Glandular Trichome Production

One of the most important changes to the software pipeline developed by Farr (2019) that is included in the current R package developed with help from McKinnon (unpublished work, 2019) is the inclusion of the p.adjust method to supplement false discovery rate analysis with an additional correction for multiple comparisons, resulting in the generation of adjusted p-values. Rather than only limiting false discoveries, the test provides a shorter list of SNPs with significant adjusted p-values that are sourced from highly significant original B* results.

The most relevant set of ontological terms displayed in Table 4 for the gene identifier, Migut.J01741, is associated with mucilage development in the seed coat. Broadly, mucilage can be defined as a secretory product associated with glandular trichomes as well as the formation of the seed coat (Li 2009, Tsai 2017). Mucilage and transcription factors promoting genes for mucilage development are expressed at higher levels in *Arabidopsis thaliana* trichomes and are an essential component of seed coat development in reproduction (Li 2009, Tsai 2017). Mucilage can also be used in medications, and therefore has agricultural and economic importance as well (Malviya 2011, Prajapati 2013).

CHAPTER VI

CONCLUSION

Within the scope of this study, we observed that there was significant variation in glandular trichome production as both an epigenetic response and a genetic response to selection. The phenotypic variance between populations and individuals is likely due to complex interactions between genetic factors, epigenetic factors, and expression level changes that are difficult to untangle in a single experiment. Rather, our data suggest candidate genes of interest and differentially methylated regions associated with increased trichome production that require future investigation through gene knockout or knockdown studies, as well as a thorough exploration of the role of transcription factors and transposable elements.

Akkerman (2016) proposed that glandular trichome production as a defensive trait was transgenerationally inherited in a sex-dependent manner and that maternal transmission was specifically susceptible to interference by 5-azacytidine, which impedes the replication of CpG epigenetic modifications. The results summarized in Table 1 support the assertion that parental damage results in significantly differentially methylated regions and suggest a potential ontological basis for regulation of associated genes.

The results of the genomic analysis show an increase in the frequency of Point Reyes alleles located at one end of chromosome 10 (Table 3) and associated with increased constitutive trichome production. This region exists closest to gene ID Migut.J01741, which is associated with mucilage development and represents an attractive candidate gene for the inheritance of increased trichome production. Future

knockout or knockdown of this gene could be used to assess involvement with trichome production.

Sound data reporting from Nanopore sequencing typically requires high accuracy, quality, and genomic coverage (Kurdyukov et al. 2016, Ziller 2015), despite the potential value of extended read lengths not afforded cost-effectively from Illumina sequencing (Besser et al. 2018). These are areas which must be improved upon to present the most meaningful results of the methylation data. At present, due especially to a low coverage and quality, these factors diminish the statistical power of the epigenetic portion of this experiment, thus while it is possible that given greater coverage and quality mapping, some or all of the DMRs identified would remain significant, it is difficult to assess the probability of error and should be noted as such. The pipeline that was generated to carry the raw Nanopore data into DMR analysis will be useful for future studies and should be expected to produce actionable results when more quality DNA can be sequenced from the remaining samples from the Akkerman (2016) experimental population. Future challenges aside, the ability to apply differential methylation analysis to Nanopore data has not benefited from great documentation or software development, and the release and test case for the analysis is exciting as future studies into Monkeyflower unfold.

REFERENCES

Akkerman, K. C., Sattarin, A., Kelly, J. K., & Scoville, A. G. (2016). Transgenerational plasticity is sex-dependent and persistent in yellow monkeyflower (Mimulus guttatus). *Environmental Epigenetics*, *2*(2), dvw003.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software.*, *67*(51).

Beissinger, T. M., Rosa, G. J., Kaeppler, S. M., Gianola, D., & De Leon, N. (2015). Defining window-boundaries for genomic analyses using smoothing spline techniques. *Genetics Selection Evolution*, *47*(1), 1–9.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *51*(1), 289–300.

Besser, J., Carleton, H. A., Gerner-Smidt, P., Lindsey, R. L., & Trees, E. (2018). Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clinical Microbiology and Infection*, *24*(4), 335–341. [

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, *8*(12), e1002822.

Christman, J. K. (2002). 5-Azacytidine and 5-aza-2′-deoxycytidine as inhibitors of DNA methylation: Mechanistic studies and their implications for cancer therapy. *Oncogene*, *21*(35 REV. ISS. 3), 5483–5495.

Coley, P. D., Bryant, J. P., & Chapin, S. F. (1985). Resource availability and plant antiherbivore defense. *American Association for the Advancement of Science*, *230*(4728), 895–899.

Colicchio, J. (2017). Transgenerational effects alter plant defense and resistance in nature. *Journal of Evolutionary Biology*, *30*(4), 664–680.

Colicchio, J. M., Kelly, J. K., & Hileman, L. C. (2018). Parental experience modifies the Mimulus methylome. *BMC Genomics*, *19*(1), 746.

Colicchio, J. M., Monnahan, P. J., Kelly, J. K., & Hileman, L. C. (2015). Gene expression plasticity resulting from parental leaf damage in Mimulus guttatus. *New Phytologist*, *205*(2), 894–906.

Finnegan, E. J., Genger, R. K., Peacock, W. J., & Dennis, E. S. (1998). DNA Methylation in plants. *Annual Review of Plant Physiology Plant Molecular. Biology*, *49*, 223–247.

Grant-Downton, R. T., & Dickinson, H. G. (2005). Epigenetics and its implications for plant biology. 1. The epigenetic network in plants. *Annals of Botany*, Vol. 96, pp. 1143–1164.

Harborne, L. B. (1993). *Introduction to Ecological Biochemistry* (4th ed.). Academic Press.

Hellsten, U., Wright, K., Jenkins, J., Shu, S., Yuan, Y., Wessler, S. R., … Rokhsar, D. S. (2013). Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(48), 19478–19482.

Holeski, L. M. (2007). Within and between generation phenotypic plasticity in trichome density of Mimulus guttatus. *Journal of Evolutionary Biology*, *20*(6), 2092–2100. [

Holeski, L. M., Keefover-Ring, K., Bowers, M. D., Harnenz, Z. T., & Lindroth, R. L. (2013). Patterns of phytochemical variation in Mimulus guttatus (Yellow Monkeyflower). *Journal of Chemical Ecology*, *39*(4), 525–536.

Huchelmann, A., Boutry, M., & Hachez, C. (2017). Plant glandular trichomes: Natural cell factories of high biotechnological interest. *Plant Physiology*, *175*(1), 6–22.

Jablonka, E., & Raz, G. (2009). Transgenerational epigenetic inheritance: Prevalence, mechanisms, and implications for the study of heredity and evolution. *The Quarterly Review of Biology*, *84*(2), 131–176.

Jin, J., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J., & Gao, G. (2017). PlantTFDB 4.0: Toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research*, *45*(D1), D1040–D1045. [

Kelly, J. K., Koseva, B., & Mojica, J. P. (2013). The genomic signal of partial sweeps in Mimulus guttatus. *Genome Biology and Evolution*, *5*(8), 1457–1469. [

Kurdyukov, S., & Bullock, M. (2016). DNA methylation analysis: Choosing the right method. *Biology*, *5*(1), 3.

Leipzig, J. (2017). A review of bioinformatic pipeline frameworks. *Briefings in Bioinformatics*, *18*(3), 530–536.

Levin, D. (1973). The role of trichomes in plant defense. *The Quarterly Review of Biology*, *47*(1), 131–159.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993.

Li, H., Handsaker, B., Wysoker, A., Fennel, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, *25*, 2078–2079.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100.

Li, S. F., Milliken, O. N., Pham, H., Seyit, R., Napoli, R., Preston, J., … Parish, R. W. (2009). The Arabidopsis MYB5 transcription factor regulates mucilage synthesis, seed coat development, and trichome morphogenesis. *The Plant Cell Online*, *21*(1), 72–89.

Maes, L., & Goossens, A. (2010). Hormone-mediated promotion of trichome initiation in plants is conserved but utilizes species- and trichome-specific regulatory mechanisms. *Plant Signaling and Behavior*, *5*(2), 205–207.

Malviya, R. (2011). Extraction characterization and evaluation of selected mucilage as pharmaceutical excipient. *Polimery w Medycynie*, *41*(3):39-44.

Maunakea, A. K., Nagarajan, R. P., Bilenky, M., Ballinger, T. J., Dsouza, C., Fouse, S. D., … Costello, J. F. (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, *466*(7303), 253–257.

Morishita, S., Qu, W., Hashimoto, S.-I., Shimada, A., Nakatani, Y., Saito, T. L., … Takeda, H. (2012). Genome-wide genetic variations are highly correlated with proximal DNA methylation patterns. *Genome Research*, *3*, 1419–1425.

Neuffer, S. J. (2015). The genetic architecture of trichome production in Mimulus guttatus (Yellow Monkeyflower) *Unpublished*.

Neupane, R., Rokhsar, D. S., Mitros, T., Goodstein, D. M., Hayes, R. D., Dirks, W., … Hellsten, U. (2011). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, *40*(D1), D1178–D1186.

Okonechnikov, K., Conesa, A., & Garcia-Alcalde, F. (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, *32*(2), 292–294.

Oxford Nanopore Technology. How it works. [Internet] [Cited 3 July 2019]. Available from: https://nanoporetech.com/how-it-works

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842.

Satyaki, P. R. V., & Gehring, M. (2017). DNA methylation and imprinting in plants: machinery and mechanisms. *Critical Reviews in Biochemistry and Molecular Biology*, *52*(2), 163–175.

Scoville, A. G., Barnett, L. L., Bodbyl-Roels, S., Kelly, J. K., & Hileman, L. C. (2011). Differential regulation of a MYB transcription factor is correlated with transgenerational epigenetic inheritance of trichome density in Mimulus guttatus. *New Phytologist*, *191*(1), 251–263.

Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., & Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, *14*(4), 407–410.

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of Gene Ontology terms. *PLoS One*.

Underwood, C. J., Henderson, I. R., & Martienssen, R. A. (2017). Genetic and epigenetic variation of transposable elements in Arabidopsis. *Current Opinion in Plant Biology*, *36*(Figure 1), 135–141.

Yokoyama, T., Miura, F., Araki, H., Okamura, K., & Ito, T. (2015). Changepoint detection in base-resolution methylome data reveals a robust signature of methylated domain landscape. *BMC Genomics*, *16*(1), 1–10.

APPENDIXES

APPENDIX A

Full List of Differentially Methylated Regions

Appendix A Table of the complete 59 differentially methylated regions resulting from the DMR-Scan analysis. Genomic features within 2 kilobases are described where available. A positive distance from genomic feature implies that the DMR was upstream of the feature where applicable. Adjusted P-values are calculated based on α=0.05. Significance: All are P <0.05; *P < 0.01; **P < 0.001; ***P < 0.0001.

CHR	Start BP	Diff. Mean Methylation	DMR Size (BP)	Nearest Gene	Genomic Feature	Distance from Genomic Feature (BP)
1	4127	9.23E-01	12	Migut.A00001	NA	NA
1	5455	3.39E-01	28	Migut.A00001	NA	NA
1	10412	-2.24E-01 **	85	Migut.A00001	NA	NA
1	18171	-3.58E-01 **	37	Migut.A00001	NA	NA
1	25599	-5.39E-02 ***	649	Migut.A00002	NA	NA
1	28582	8.00E-02	329	Migut.A00002	5'-UTR	955
2	23067	-8.20E-01 ***	17	Migut.B00003	NA	NA
2	28443	-1.33E-01 *	167	Migut.B00004	NA	NA
2	48986	-3.30E-01	39	Migut.B00006	NA	NA
3	2498	-3.44E-01	23	Migut.C00001	NA	NA
3	26607	7.77E-01	9	Migut.C00001	NA	NA
3	38308	2.81E-01	47	Migut.C00001	NA	NA
4	6582	7.86E-01	13	Migut.D00002	NA	NA
5	25254	1.78E-01 *	94	Migut.E00003	NA	NA
5	34877	-4.97E-02	623	Migut.E00005	5'-UTR	534
6	31237	-1.03E-01	115	Migut.F00003	5'-UTR	-1991
6	32407	4.17E-01 *	26	Migut.F00003	NA	NA
7	9403	1.32E-01 **	119	Migut.G00001	NA	NA
7	9775	-4.97E-02	565	Migut.G00001	NA	NA
7	16843	2.07E-01 ***	60	Migut.G00001	NA	NA
7	18957	-3.34E-01 *	35	Migut.G00001	NA	NA
7	25625	-6.81E-01 *	15	Migut.G00001	NA	NA
7	31750	9.64E-01	13	Migut.G00001	5'-UTR	0
7	36764	4.31E-01	29	Migut.G00002	NA	NA
8	8472	-1.77E-01 *	83	Migut.H00001	NA	NA
8	25101	2.92E-01 ***	77	Migut.H00002	NA	NA
8	29930	-6.11E-01	17	Migut.H00002	NA	NA
8	55375	7.07E-02 *	278	Migut.H00007	NA	NA

9	336	-2.26E-01	62	Migut.I00001	NA	NA
10	26693	1.77E-01	79	Migut.J00004	NA	NA
10	29964	-2.12E-01	65	Migut.J00005	5’-UTR	-64
10	40149	6.22E-01	14	Migut.J00006	NA	NA
10	42634	4.33E-01	17	Migut.J00007	NA	NA
10	48388	-3.49E-01	22	Migut.J00007	5’-UTR	1579
11	2350	-5.94E-02 *	406	Migut.K00001	NA	NA
11	10821	-4.20E-02	545	Migut.K00001	5’-UTR	-315
11	12382	-4.97E-02 **	1126	Migut.K00001	5’-UTR	-1876
11	14000	-1.24E-01 ***	276	Migut.K00001	NA	NA
11	29573	-8.48E-02 **	315	Migut.K00001	NA	NA
11	65607	-2.57E-01 ***	62	Migut.K00005	CDS	NA
11	74531	-2.99E-01 *	58	Migut.K00008	5’-UTR	-557
11	79413	6.43E-01	13	Migut.K00009	5’-UTR	1411
12	14273	-1.29E-01	124	Migut.L00003	NA	NA
12	19183	-8.33E-02	212	Migut.L00004	NA	NA
12	25208	1.02E-01	251	Migut.L00004	NA	NA
12	35035	5.91E-01	21	Migut.L00004	NA	NA
13	2796	-8.51E-02 *	218	Migut.M00001	5’-UTR	0
13	8699	2.57E-01 *	77	Migut.M00001	NA	NA
13	21383	-3.75E-01 *	47	Migut.M00002	NA	NA
13	21808	2.15E-01 ***	56	Migut.M00002	NA	NA
13	23480	1.59E-01 *	136	Migut.M00002	NA	NA
13	25721	-4.74E-02 ***	682	Migut.M00002	NA	NA
13	30805	5.99E-01 *	16	Migut.M00003	5’-UTR	-877
13	36044	2.56E-01 *	47	Migut.M00003	NA	NA
14	17052	-2.64E-01	54	Migut.N00004	5’-UTR and CDS	819
14	18012	-4.41E-02 ***	524	Migut.N00004	5’-UTR	-25
14	18012	-4.41E-02 ***	524	Migut.N00005	5’-UTR	-25
14	27582	8.75E-01	15	Migut.N00008	5’-UTR	0

APPENDIX B

Scoville Lab Urea Extraction Protocol

Day before grinding

1. Place mortar and pestle(s) in freezer
2. Place 70% ethanol in freezer
3. Can make 5M NaCl
4. Can make TE buffer

Day of extraction

5. Remove phenol:chloroform:isoamyl alcohol from fridge and place in hood. Protect from light. When equilibrated to 15 to 30 degrees, swirl thoroughly to form a single, clear, homogenous phase. (If necessary, it may be okay to use the lower, clear, organic layer at 2 to 8 degrees). Pipette out desired amount to retain protective argon layer in the bottle.
6. Need 5mL per extraction.
7. Make up lysis buffer
8. During step 7, make up RNaseA (0.5 mL per sample; 10mg/mL). For 8 extractions, need 40 mg in 4mL (or 42 mg in 4.2 mL)
9. During step 8, make up Chloroform:isoamyl alcohol 24:1

Locate/prep

10. 1 50 mL falcon tube (lysis buffer)
11. 5M NaCl (3.5 mL per 10 extractions)
12. spatula, weigh boats for tissue (1g) and RNAase A (42 mg)
13. platform shaker in hood
14. rocking platform
15. centrifuge for 15 mL falcon tubes; 3000 – 4000 rpm
16. 15 mL falcon tubes: 4 per sample
17. water bath or incubator at 37 degrees
18. ice

5M NaCl

1. Mix 14.61 g NaCl with 45 mL of ddH_2O by stirring.
2. Add ddH_2O until final volume is 50 ml.
3. Store at room temperature.

70% ethanol

1. Measure 73.68 mL of 190 proof (95% ABV) ethanol
2. Add ddH_2O to a final volume of 100 mL

1x TE buffer

1. 500 uL of 1M Tris-HCl (pH 8.0)
2. 100 uL of EDTA (0.5 M)
3. Add ddH_2O to a final volume of 50 mL

Chloroform:isoamyl alcohol 24:1

1. 48 mL chloroform: 2 mL isoamyl alcohol for 50 mL
2. 43.2 mL chloroform : 1.8 mL isoamyl alcohol for 45 mL