# Transcriptomic and Phenomic Insights into the Epimutant Rogue Locus in *Pisum sativum*

James Lancaster
Bachelor of Science

A thesis submitted in total fulfilment of the requirements for the degree of Master of Science at
**La Trobe University, Victoria, Australia** in November 2021
School of Life Sciences
College of Science, Health and Engineering

## Abstract

The epigenetic drive termed "paramutation" was first discovered in pea (*Pisum sativum*) over a century ago, where the abnormally high penetrance of the rogue phenotype consisting of reduced and elongated stipules and leaflets was documented. Paramutagenic alleles induce heritable epigenetic changes to susceptible alleles, resulting in the paramutated epiallele being inherited at greater frequency than Mendelian inheritance would otherwise predict. The paramutation phenomenon and its mechanisms have been extensively studied in maize and among other eukaryotic organisms, though the rogue epimutation in *P. sativum* has yet to be identified and its mechanism of inheritance remains unknown. With the recent release of a complete genome and advances in genomic technologies, we can now re-investigate a century-old mystery. I was able to systematically test rogue phenotype inheritance in hybrid crosses and begin the development of a machine learning model with the ability to classify individuals into their correct epigenotype, based on leaf morphometrics and phenomic analysis. I then used a combination of transcriptomic techniques involving RNA-seq, small-RNA-seq and long-read transcriptome sequencing in an attempt to identify and characterise the rogue locus. Through long-read sequencing on the Oxford Nanopore platform, I was able to improve on the current transcriptome annotation and increase the sensitivity of my analysis. I discovered that epimutated rogue peas of multiple cultivars display differential expression in mRNAs and sRNAs relative to their wild type counterparts. From this, I was able to develop a candidate list of genes for the elusive rogue locus. Genes involved in key regulatory and developmental pathways were differentially expressed, such as ZINC INDUCED FACILITATOR-LIKE 1 (ZIFL1) for example, were down-regulated in rogues compared to wild types. In some cases, candidate genes were also found to have nearby clusters of 24-nt sRNAs, which are well known for their involvement in transcriptional regulation and silencing pathways, such as the RNA-directed DNA Methylation pathway (RdDM). Integrating transcriptomic evidence for the rogue locus in *P. sativum* has enabled construction of the first candidate list of genes a century after the rogue pea's discovery, enabling future research to further elucidate the complex epigenetic interaction known as paramutation.

**Statement of Authorship**

Except where reference is made in the text of the thesis, this thesis contains no material published elsewhere or extracted in whole or in part from a thesis accepted for the award of any other degree or diploma. No other person's work has been used without due acknowledgment in the main text of the thesis. This thesis has not been submitted for the award of any degree or diploma in any other tertiary institution.

Signature:

Print Name: James Lancaster

Date: 30/11/2021

**Acknowledgements**

Here I extend my gratitude to all the individuals who contributed to making my Masters an excellent experience and opportunity.

First and foremost, I would like to thank my supervisor Mathew Lewsey for his remarkable support and guidance since I first expressed interest in studying plant genomics with his lab - I couldn't have asked for a better supervisor. The research community you have developed at the Lewsey Lab is something to be very proud of.

I also would like to thank my secondary supervisor Quentin Gouil, for sharing his expertise in genomics, epigenomics and bioinformatics - the amount of effort and support you provided me throughout my studies was exemplary, and I can't thank you enough.

The members of the Lewsey Lab were always supportive, happy to provide helpful suggestions for my research, and provided a welcoming social circle during my Masters. I am grateful to have worked alongside you all.

Lastly, I would like to thank my partner, Hannah Petocz, for putting up with me during my studies.

This research was supported by an Australian Government Research Training Program (RTP) Scholarship.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Gregor Mendel (1822-1884) made his ground-breaking discoveries on inheritance of monogenic traits in *Pisum sativum* (Pea), where he conducted hybridisation experiments between varieties with obvious phenotypic differences [1, 2]. The key phenotypic traits Mendel analysed were seed colour (green or yellow) and coat (wrinkled or smooth) [2]. Through extensive hybridisation experiments, he discovered that these traits exhibited dominant and recessive relationships, and segregated in subsequent generations following a 3:1 ratio in F2 generations [2]. This study allowed Mendel to establish his famous laws of genetic inheritance: Independent Assortment, Dominance, and Segregation [1]. However, the complex nature of biology was later demonstrated when exceptions to Mendel's laws were discovered in other angiosperms such as incomplete dominance, sex-linked genes or polygenic traits [3–5].

Ironically, it is also in *P. sativum* that exceptions to Mendel's laws were revealed in the early 1900's. A novel pea phenotype termed the 'rogue' phenotype was found to be inherited in drastically increased proportions compared to what Mendelian laws would predict [6, 7]. The morphological characteristics of the phenotype consist of smaller, narrow and pointed leaflets and stipules (Figure 1.1), noted to resemble rabbit ears, whereas wild type peas generally have large and broad leaves with rounded leaf apices [6]. In some cases, a curved pod phenotype was also identified in rogue peas. The origin and nature of rogue peas was first studied by Bateson and Pellew (1915) [6], through extensive crossing studies which investigated the inheritance properties of the phenotype. Due to slight variations in the rogue phenotype observed during crossing studies, a basic classification system was used to measure the severity of 'rogueness' [6]. Class 1 described wild type peas displaying what was considered normal phenotypic properties for peas of the Sutton's Early Giant and Duke of Albany cultivars used in the study. Class 2 described plants as mostly wild type, though with slight curvature of pods. Class 3 described intermediate phenotypes, where initial nodes of individuals displayed wild type characters, and progressively became more rogue-like in leaf morphology and pod morphology as individuals developed to mature plants. Class 4 individuals were mostly rogue in phenotype, with slightly larger stipules and leaflets and curved pods. Finally, Class 5 described complete rogues.

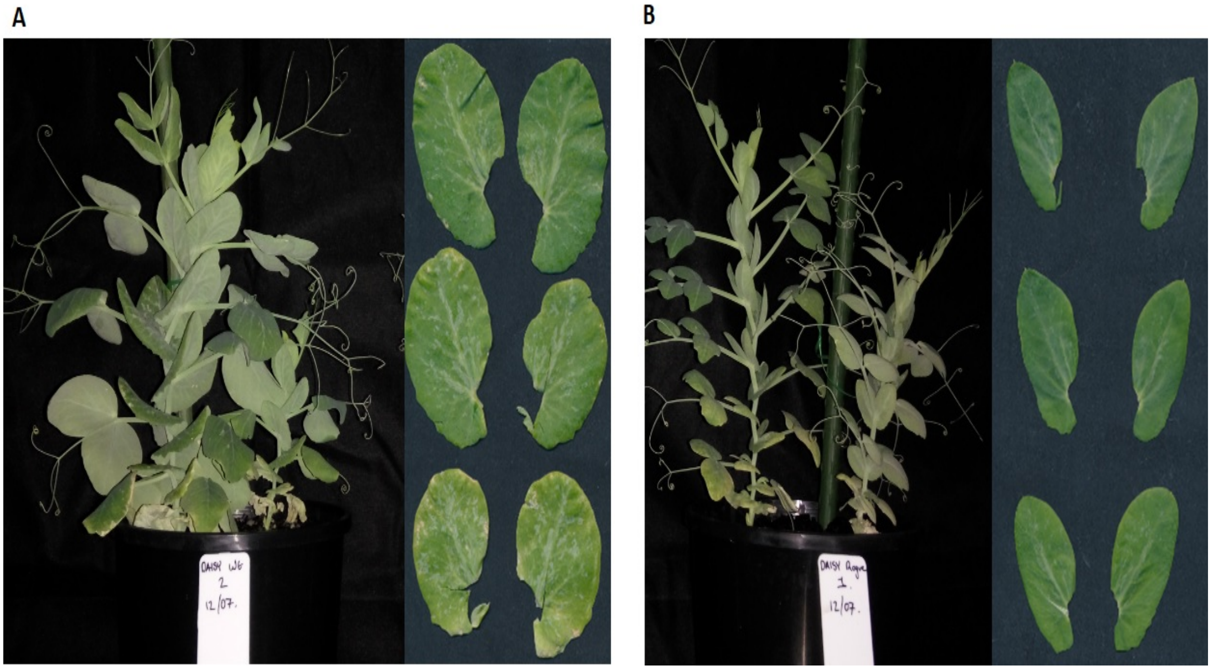


FIGURE 1.1: **The rogue phenotype.** Wild type individual from the Daisy cultivar (A) displays its typical leaf morphology, whereas the rogue variation of this cultivar (B) displays an altered phenotype where stipules and leaflets are reduced in size, and have pointed "rabbit-eared" tips.

The emergence of rogues among populations and their inheritance properties were identified through crossing rogue and wild type individuals. First, rogue peas emerge spontaneously in populations of self-fertilized individuals. Rogue individuals can then only produce rogues. Crosses between rogue and wild type peas produce F1 plants with intermediate phenotypes (Class 3 individuals), where the lower nodes display wild type leaf morphology and subsequent nodes develop rogue-like traits as the individual develops. This could suggest that the activity of the underlying genetic or epigenetic mechanism occurs during plant growth rather than at fertilisation. The F2 progeny from hybrid crosses then display a complete morphological shift to the rogue phenotype, with a penetrance range of 97-99% - vastly exceeding expected inheritance ratios in Mendelian genetics [6, 8]. Notably, no segregation is observed either when rogue-like F1s or F2s are back-crossed to a wild type [6, 7]. There are occurrences where progeny from back-crosses revert to intermediate phenotypes, though these individuals then switch back to rogue phenotypes as the plant develops [8]. From historical literature on the rogue phenotype, it is clear Mendelian laws do not apply to rogue peas, and that other genetic or epigenetic mechanisms are present.

The causal locus of the rogue phenotype, and its inheritance mechanism remain unidentified, despite a long history of *P. sativum* research as a model legume species [9]. *P. sativum* has been widely utilised as a genetic, biochemical and physiological model organism, owing to its short generation time, ease of growth and the ability to perform crosses. However, the advancement of genomic research in pea has been slower than in other model plants and this has limited progress in defining the mechanism underlying the rogue phenotype [9]. This is due to the large and complex pea genome (4.5Gb), consisting mainly of repetitive DNA and transposable elements which made modern genomics approaches challenging [10, 11]. However, with the first publicly available pea genome and transcriptome released in 2019, 'omics techniques can now be more readily used in *P. sativum* to investigate the genomic locus and the mechanism of the rogue phenotype [11].

The non-Mendelian inheritance of the rogue phenotype matches the epigenetic phenomena known as paramutation. Paramutation is an epigenetic drive that alters the way in which genes are inherited. The paramutated alleles of affected loci are inherited at much greater frequencies than expected in traditional models of Mendelian Inheritance [7, 12, 13]. Paramutation involves an epimutation-inducing allele, termed the "paramutagenic" allele, and alleles susceptible to paramutation, termed "paramutable". Alleles which remain unaffected by epimutation are "neutral". *P. sativum* was likely the first example of paramutation in any species, documented over a century ago, though the definition of the phenomena was not yet established [6].

Subsequently paramutation has been observed and characterised in other plants such as maize, petunia, and tomato, as well other eukaryotic organisms like *Caenorhabditis elegans*, *Drosophila melanogaster* and *Mus musculus* have been investigated at the molecular level with great advances in elucidating how paramutation operates [13–18]. Our greatest understanding of paramutation mechanisms is derived from extensive studies on maize, which has become a strong model for the epigenetic phenomena [19].

Mechanistic investigation of paramutation across multiple model eukaryotes is important, as it can improve molecular techniques and develop our understanding of genetic and epigenetic inheritance. Once we understand what is required for paramutation, it could be intentionally induced to fix valuable crop traits, aiding crop development and research and potentially increase global food security [9, 20–22]. Maize has served as an excellent organism to study paramutation, though there are still many gaps in our understanding of paramutagenic inheritance - particularly surrounding how this phenomena is first established within populations, and how it operates across multiple paramutant organisms. There are currently a limited number of paramutant examples, with an over-representation of affected loci related to pigmentation [23]. Because of this, our understanding of paramutation in terms of conserved mechanisms and the frequency with which epimutation occurs is lacking. Incorporating more examples of non-Mendelian inheritance like the rogue paramutation in *P. sativum* can increase our understanding of epigenetic inheritance, and identify potential conserved mechanisms in paramutant eukaryotes. Therefore, using what has been identified at well characterised paramutant loci such as those in maize set examples of molecular characters to look for in other paramutant angiosperms like *P. sativum*.

Our current understanding of the mechanisms of paramutation relies heavily on studies in maize. Multiple paramutagenic loci were identified through mapping studies: *r1* (*red1* or *coloured1*), *b1* (*booster1* or *coloured plant1*), *pl1* (*purple plant 1*), *p1* (*pericarp colour* 1) and *lpa1* (*low phytic acid 1*). All paramutated loci except *lpa1* are involved in flavonoid biosynthesis pathways affecting the pigmentation of various plant organs [23]. Thanks to forward genetic screens, several genes involved in the paramutation mechanism at these loci in maize were identified: the Mediator of paramutation (MOP) genes and Required to Maintain Repression (RMR) genes. Orthologues of MOP and RMR genes in *Arabidopsis thaliana* encode proteins associated with small-interfering RNA (siRNA) synthesis and the RNA-directed DNA Methylation pathway (RdDM) [24–27], leading to the model that the RdDM pathway is the key driver of paramutation. The orthologue of *MOP1* in *A. thaliana* encodes RNA-DEPENDENT RNA POLYMERASE 2 (RDR2), which is a key component of RdDM [24]. Targeted methylation of repetitive regions of the genome like transposons is carried out by a complex formed by RDR2 and the plant-specific DNA-dependent RNA polymerase IV [24, 28]. This RdDM complex produces short double-stranded RNAs which are then cleaved by Dicer-like endonucleases. The resulting 24-nt siRNAs direct Argonaute proteins to Pol V transcripts and recruit a cytosine methyltransferase to methylate the target genomic region, reducing or completely silencing gene expression [14, 29–32]. The claim that RdDM drives paramutation in maize is strongly supported, as loci affected by 5-methylcytosine patterns and transcriptional repression through orthologous RdDM machinery imitate the well described regulatory pathway in *A. thaliana* [19, 29, 33–36]. The extensive assays and research in maize serve the purpose of elucidating epimutation mechanisms, and better inform us of which transcriptomic constituents to investigate in other examples of paramutant species.

Similar mechanisms to those in maize also appear to be driving paramutant interactions in other species. Much like in maize, the *sulfurea* paramutation in tomato correlates with siRNA production and differential methylation at the promoter region for the *SLTAB2* locus. A chlorophyll-deficient phenotype in paramutant individuals is a result of transcriptional repression of *SLTAB2* [18, 37]. The commonality in paramutation mechanisms also extends to *C. elegans*, *D. melanogaster* and *M. musculus* where the involvement of heritable RNA-induced silencing and differential DNA methylation patterns has been observed [15, 17, 38, 39]. In *C. elegans*, nuclear RNA-interference (RNAi) and chromatin factors induce multi-generational inheritance of piwi-interacting RNAs (piRNAs) and environmental RNAi silencing of transgenes for up to 20 generations [38, 39]. Similar studies in *D. melanogaster* were able to induce a strong paramutagenic effect by insertion of transgenes in distant euchromatic regions from transposable P-element tandem repeats [17]. Maternally inherited sRNAs matching the tandem repeat region induced a paramutated state to transgenes, which then became paramutagenic themselves in subsequent generations [17]. The RdDM pathway, siRNAs, and repeat-associated mechanisms identified in model organisms prompted us to investigate whether a similar mechanism may be associated with the non-mendelian epigenetic inheritance of rogue phenotypes in *P. sativum*, and whether these hallmarks of paramutation may be used to identify the genomic locus responsible for the rogue phenotype.

The first study with the aim of investigating the pea rogue locus at the molecular level was published by Santo et al. (2017) [40]. A genome-wide methylation assay using methylation-sensitive restriction enzymes (HpaII & MspI) and reverse phase high-performance liquid chromatography (RP-HPLC) revealed no significant global differences in cytosine methylation between rogue and wild type epigenotypes of the Onward cultivar [40]. They then attempted to differentiate epigenotypes by methylome in leaf DNA, using methylation-sensitive amplified fragment length polymorphism (MS-AFLP) analysis to increase analytical sensitivity, and identified 22 polymorphic markers between both epigenomes out of 2,238 tested. For 12 of these markers, DNA methylation differences between rogue and wild type were also conserved in pollen DNA [40]. Of the polymorphic markers identified through MS-AFLP, 13 of 22 had similarities with expressed sequences in *Medicago truncatula* and *Cicer arietinum* - identified using the Basic Local Alignment Search Tool (BLAST, NCBI). Santo et al. (2017)[40] then assessed expression of all 22 sequences between epigenotypes by quantitative real-time RT-PCR, but was not able to detect 8. The remaining 14 sequences were successfully detected, though no differences in expression were observed regardless of polymorphic methylation patterns [40]. This study unfortunately suffered from severe limitations. First of all, the authors did not take into account that two lines of peas that have been grown separately for many generations will accumulate epigenetic differences that are unrelated to the rogue phenotype. Secondly, the candidate approach based on 2,338 markers would only cover about 0.01% of the pea genome, assuming an average size of 300 bp, making it a 1 in a 10,000 chance that the rogue locus is included this set. Thirdly, the differentially methylated sequence hypothetically responsible for the paramutation may be quite distant from the gene it regulates, as is the case for the maize *b1* gene whose differentially methylated regulatory sequence is located 100 kb upstream, so searching for expressed genes directly in DNA markers is unlikely to succeed. Overall, this study contributed little to the search for the rogue locus.

The recent publication of a reference genome for *P. sativum* enables us to revisit the problem of identifying the rogue locus by searching genome-wide. While whole-genome DNA methylation sequencing would be informative, it remains very costly. The search space and cost may be reduced by searching the mRNA transcriptome and small-RNAome for common markers of paramutation-silenced alleles and differential sRNA abundances, then using targeted methylation assays on the candidate loci. Study design must also consider that genetic and epigenetic drift contribute to variations between rogue and wild type lines, as they have been kept separate for over 100 years. To counter this, multiple rogue pea cultivars should be studied and commonly implicated loci identified. However when using multiple rogue cultivars, one must keep in mind the causal locus of the rogue phenotype may be cultivar-specific. Many rogue *P. sativum* cultivars are available through the John Innes

Center (UK), providing the necessary material for this approach. Recent progress in long-read transcriptome sequencing enables better annotation and quantification of transcripts, particularly for complex genomes as in *P. satiuvm* [41–43]. However, the throughput of long-read sequencing is limited, and therefore it must be paired with short-reads to achieve a comprehensive and quantitative coverage of the pea transcriptome.

Detailed parameterization of the phenotype is another gap in characterising the rogue epimutation of *P. sativum*. Few studies to date have investigated rogue phenotypic properties using modern phenomics approaches. Brotherton (1923) [8] manually assessed the length-to-width ratios of wild types, rogues and hybrid crosses of the Gradus cultivar, identifying shifts in F1 phenotype at the 5th-7th leaflet node. The leaf area and cell size of rogue and wild type epigenotypes has also been assessed in 3 cultivars (Thomas-laxton, Greengolt and Sutton's Early Giant) [44]. Results confirmed a mean leaf area reduction of 35%, and stipule area reduction of 44%. There were no differences observed in cell size between epigenotypes - only less cells present in rogues. Current phenomics software packages and machine learning techniques could be utilised to increase analysis throughput, and to develop phenotypic standards for paramutant cultivars of *P. sativum*. A reference for the phenotypic classification of epigenotypes would aid the future study of the rogue phenotype, and provide a quantitative, high-throughput and unbiased scoring system for rogue penetrance.

It is clear further molecular investigation of the rogue locus is required to identify the epimutated region, the mechanism driving non-mendelian rogue inheritance, and affected developmental pathways inducing shifts in leaf morphology. The availability of the new pea reference genome and transcriptome means transcriptomic analysis will increase in accuracy, enabling greater confidence and accuracy in identifying differentially expressed genes which may be involved with the rogue phenotype and paramutation. Today's sequencing technologies provide ample cost-effective sequencing, read length and depth to identify the causal locus. Phenomic analysis packages such as Momocs increase analysis throughput when utilised in conjunction with machine learning platforms [45]. A combination of the two can develop classification standards for all paramutated cultivars.

The aims of my study were to:

1. Develop a list of candidate genes for the rogue locus generated through a combination of long- and short-read mRNA sequencing, as well as small-RNA sequencing, in 4 *P.*

*sativum* cultivars (Thomas Laxton, Black-Eyed Susan, Daisy and Sutton's Hundredfold),

2. Create an improved reference transcriptome annotation through long-read sequencing to increase the sensitivity of differential gene expression analysis,
3. Confirm the inheritance pattern of the rogue phenotype in selected cultivars is true to that described in prior literature,
4. Develop an automated phenomics classification method to quantify the rogue phenotype for selected cultivars enabling systematic testing for rogue trait inheritance.

Based on similarities in the phenotypic shift from wild type to rogue in multiple *P. sativum* cultivars, I hypothesise that a single locus may be responsible for the rogue paramutation across cultivars. Based on the RdDM model of paramutation, I also hypothesise that the affected gene is down-regulated in rogue epigenotypes and that the down-regulation correlates with small RNA production at regulatory sequences. I expect the causal locus is involved in developmental pathways which can affect leaf morphology. I also expect that F2 generation hybrids from rogue and wild type crosses will display complete transitions to the rogue phenotype, as described in early literature.

In this study I identified candidate genes for the rogue locus, further developed the reference transcriptome annotation, and developed a framework for the quantitative characterisation of the rogue phenotype. I describe my research in the following sections.

# Chapter 2

# Methods

## 2.1 Plant growth conditions

Individuals from *P. sativum* cultivars 'Thomas Laxton', 'Daisy', 'Black-Eyed Susan' and 'Sutton's Hundredfold' (supplied by the Australian Grains Genebank and the John Innes Centre), along with previously crossed F1 and F2 generation hybrids were grown within a controlled glasshouse environment at an average ambient temperature of 22°C and subjected to long-day light cycles (8 h dark, 16 h light). The soil medium macro/micro nutrients and other constituents consisted of 0.5 L Vermiculite, 0.5 L Perlite, 35 g Macracote Coloniser Plus 4-month slow release fertiliser (15N : 3P : 9K), 30 g Nitrogen slow-release fertiliser (40N : 0P : 0K), 25 g Water holding granules, 15 g Trace elements (6Mg : 6.5FE :5.4S : 1.5Mn : 0.4Zn : 0.14B : 0.07Mo), and 5 g Garden lime per 30 L of soil.

## 2.2 RNA extraction

Whole stipules were harvested from the third leaflet node of individual seedlings 2 weeks post-emergence and snap-frozen in liquid nitrogen. Whole stipule tissue samples were then stored at -80°C until required for extraction. Each sample was crushed into a fine powder by pestle and mortar within a laminar flow cabinet and prevented from thawing using liquid nitrogen. Samples were then weighed on a microgram scale to 1 mg of whole stipule tissue per extraction. Total RNA was then isolated using Invitrogen TRIzol$^{\mathrm{TM}}$ reagent (ThermoFisher Scientific) using the manufacturer's protocol (TRIzol reagent user guide, ThermoFisher

Scientific). Minor modifications to the protocol included an extended centrifugation for 15 minutes, as some RNA precipitate was still suspended in solution for some samples. RNA precipitate pellet was washed twice rather than once, as it was found this removed further impurities and improved the overall quality of the RNA. Centrifugation between washes was also extended to 10 minutes, to ensure the RNA pellet was anchored to the bottom of the Eppendorf tube during wash steps. Total RNA pellets for each sample were then re-suspended in RNase-free water, and subsequently quality-checked and quantified utilising the NanoDrop$^{\text{TM}}$ spectrophotometer.

## 2.3 mRNA sequencing libraries

Library preparation was carried out utilising the Illumina TruSeq stranded mRNA kit following the manufacturer's protocol. Libraries were quantified utilising the Qubit 4 Fluorometer, and quality-checked on the Agilent TapeStation, to ensure the expected final product fragment size peaked at 260 bp. Libraries were then normalised to 10 nM per sample and pooled for sequencing on the Illumina Nextseq 550 platform. Pooled libraries were sequenced using an Illumina NextSeq 550 instrument (v2.5 75 cycle kit, single end) at a total concentration of 1.8 pM, along with a PhiX 76 SE + 6 bases index spike-in control.

## 2.4 Long-read sequencing libraries

Long read libraries for Thomas Laxton and Daisy cultivars were prepared with the Oxford Nanopore PCR-cDNA Barcoding kit (SQK-PCB109) following the manufacturer's guidelines. Fifty-nanograms total RNA input was used for each of the 12 samples (3 wild types, 3 rogues for each cultivar). Samples were reverse-transcribed into cDNA and PCR amplified and barcoded for 15 cycles in a thermocycler. Two microlitres of each amplified library was pooled, for a total of 24 µl as per the protocol. The final sample volume was made to 150 µl with the addition of 75 µl sequencing buffer and 51 µl of loading beads required for nanopore sequencing. One micro-litre of the final pool was then carefully loaded onto a Flongle (Oxford Nanopore Technologies Limited), and quality checked by sequencing the pool for 300 k reads on a MinION (MK1c, Oxford Nanopore Technologies Limited). It was important to ensure all barcode identifiers were present and pore occupancy was high, as this is indicative of a high-yield of long reads. Once satisfied with quality, the total pool was then loaded onto a PromethION flow cell and sequenced on the PromethION P24 (Oxford Nanopore Technologies Limited) over 72 hours.

## 2.5 Small-RNA sequencing libraries

The same total RNA used for long-read sequencing from pea cultivars Thomas Laxton and Daisy was also utilised to produce small-RNA libraries, with the addition of the Sutton's Hundredfold cultivar. Triplicates of each epigenotype made a total of 18 samples. Ten micrograms of total RNA per sample was pre-size selected for small-RNAs using a 15% denaturing polyacrylamide gel (4.2 g Urea, 0.5 ml 10X MOPS, 3.75 ml (w/v) 19:1 acrylamide bis-acrylamide, 2.5 ml RNase-free $H_2O$, 70 µl 10% (w/v) ammonium persulphate, 3.5 µl TEMED) along with 5ul denatured NEB miRNA marker as reference. Prior to loading, total RNA was first vacuum dried and re-suspended in 15 µl of loading solution for the denaturing gel (2.75 µl Formaldehyde (40%), 7.5 µl de-ionised Formamide, 0.75 µl 10X MOPS, 2 µl 10X dyes (cylene cyanol FF, bromophenol blue). The gel was pre-run at 100V for 30 mins with 0.5X MOPS as buffer. Samples were then loaded, and the gel was run at 50V until both dyes had entered the gel, then turned up to 100 V until the bromophenol blue band had reached the end of the gel (approx. 2.5 hours). The gel was then stained for visualisation on a UV transilluminator with 2.5 µl SYBR gold, and bands were cut between the 20–30 bp region (A.3). Excised gel bands were then shredded in a shredder tube and rotated overnight in 400 µl 0.3M NaCl to extract the small-RNAs from the gel. Extracted small RNAs were then re-precipitated and re-suspended in 6 µl RNase-free H2O. Three micro litres of RNA per sample was utilised for library preparations using NEBNext Multiplex Small RNA kit for the Illumina platform – the manufacturer's guidelines were followed for all preparations. Library quality was assessed after PCR amplification using the Agilent TapeStation HS-D1000 screentape to ascertain average fragment size, concentration and to check for any primer dimers or other potential quality issues. Libraries were then cleaned-up using a 2.2X concentration of magnetic beads synthesised in-house [46]. Clean libraries were then pooled to make up 20 µl at 5 nM concentration, and again quality assessed using the HS-D1000 screentape. A second size selection gel and extraction were conducted on the pool, to remove remaining primer dimers and produce a final clean pool ready for sequencing. As a final quality check before a full sequencing run, 3 libraries were spiked into a MiSeq sequencing run to ensure libraries sequenced as expected, with correct barcoding, read length and read quality. Once the quality of the spiked-in libraries was assessed, the Illumina NextSeq 550 system was used to sequence the small-RNA pool using the 75SE High output run kit.

## 2.6 Short-read RNA-seq analysis

In this study I analysed two short-read transcriptome data sets, denoted mRNA1 and mRNA2. mRNA1 was the data generated in the lab experiments described above. mRNA2 was an additional data set generated prior to this study using the same workflow as mRNA1, with some slight differences; rogue and wild types were sourced from different seed banks (John Innes Centre and Australian Grains Genebank respectively), 2 replicates were used per epigenotype, sequencing was paired-end, and input tissue was whole leaf tissue rather than stipules alone (Table 2.1).

The mRNA-seq analysis pipeline used is presented in Figure 2.1. Sequencing data for each library from data sets mRNA1 and mRNA2 was quality-checked using FastQC [47], v0.11.9. Once initial quality was confirmed, libraries were then trimmed of their standard Illumina mRNA adapters using Trim Galore! v0.6.5, to ensure adapter removal for accurate downstream analysis [48]. The data was re-quality checked using FastQC to ensure adapters were successfully removed, and to further confirm the quality of the RNA-seq libraries. The *P. sativum* genome [11] v1a release was indexed using Hisat2 [49], v2.1.0, allowing reads to then be mapped and aligned to the indexed reference genome. The pre-processed data files were converted from SAM to BAM, to then allow for counting of assigned genomic features utilising the featureCounts function in the Rsubread package [50], v2.2.2. The resulting genomic feature counts were then analysed using the edgeR [51], v3.30 DE analysis package in R, following the software manual, with alterations made to suit the study design.

Lowly expressed genes were removed from the analysis if their expression failed to exceed 0.5 counts per million in at least 2 samples, as statistical power is lost on rare or absent transcripts. Gene counts were also normalised with the TMM method [52]. I used the general linear model (GLM) approach in edgeR to estimate common dispersion, trended dispersions and tagwise dispersions in my data, and then fitted the negative binomial generalised linear model. I performed the likelihood ratios test (LRT) to test for DE genes within my modelled count data. DE genes between rogue and wild type samples, as well as the F1 lower and

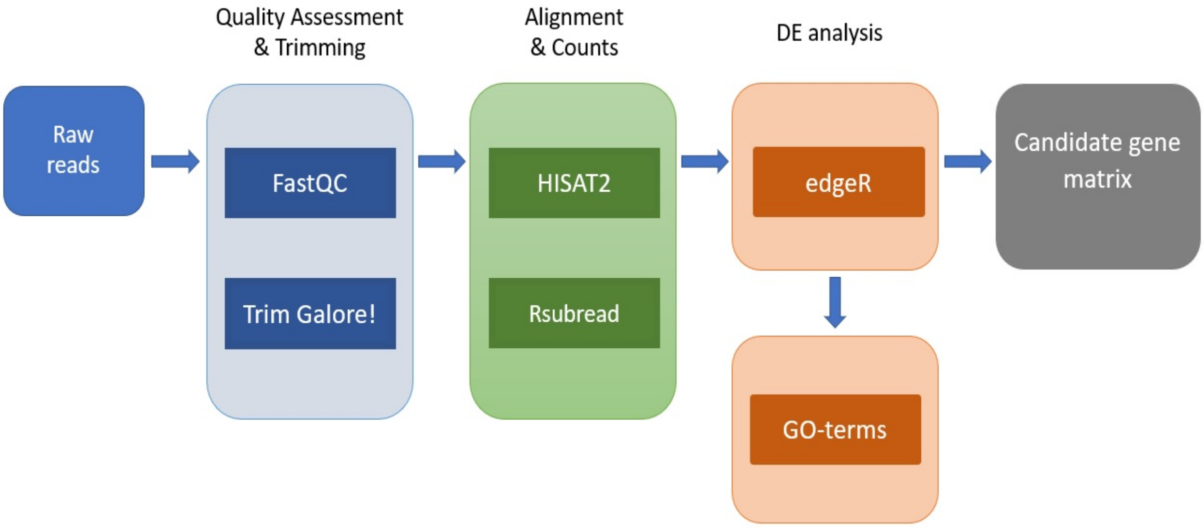


FIGURE 2.1: **DE analysis pipeline.** Raw read data was first processed in FastQC to determine initial read quality. Adapter sequences and poor quality bases were then removed using Trim Galore!, and the remaining good quality reads were again assessed by FastQC. Trimmed reads were then aligned to the *P. sativum* reference transcriptome, and subsequently processed by the featureCounts function in Rsubread to produce read counts mapping to reference features (exons). Counts were then analysed to in edgeR to identify DE genes between rogue and wild type samples. Top DE genes in each cultivar were then combined into a candidate gene matrix.

upper node, were called at a false discovery rate cut-off of p $\leq$ 0.1, and a log2-fold change of 1.5. The same analysis pipeline was also used to analyse the cultivars separately, which then allowed for a comparison of DE genes across cultivars, to identify any commonality in down or up-regulated genes.

## 2.7 Long-read RNA-seq analysis

Raw long read data were basecalled with Guppy v4.4.1 (Oxford Nanopore Technologies Limited) using the PromethION-specific high-accuracy DNA basecalling model. Read quality was checked using LongQC - a read quality assessment package tailored for long-read sequencing data [53], v.1.20b. Sequencing adapters were trimmed with Guppy. Trimmed long reads were then aligned to the reference transcriptome using Minimap2, with recommended options for long-read cDNA sequencing data [54], v2.17. BAM alignments from each sample were pooled using SAMtools [55], v1.9 for transcript discovery and quantification with the *bambu* R package (0.18129/B9.bioc.bambu, pre-publication release v1.02).

## 2.8 Small RNA-seq analysis

Raw sRNA reads were first quality assessed using FastQC to ensure expected read lengths were present, and that read quality criteria such as duplication, GC content, and adapter content were at acceptable levels. Raw reads were then trimmed using TrimGalore! to remove adapter sequences and poor quality reads, with settings set specific to the Illumina platform which ensured correct sRNA adapter sequences were removed. Following a second quality check with FastQC, reads were filtered using the 'awk' unix command to only retain reads within the size parameter of 21–25 nt in length. Filtered and trimmed sRNA reads were then aligned and clustered onto the reference genome using the ShortStack package [56], v3.6. ShortStack first indexed the reference, then utilised Bowtie [57], v1.2.1.1 as a dependency to align sRNA reads to the reference. Aligned reads were sorted into BAM files using SAMtools then de-novo clustered across the reference genome to identify sRNA cluster accumulation. ShortStack then proceeded with quantification for each individual read group (sample) to generate counts of aligned sRNA reads. Major sRNAs at each identified cluster mapping to a locus were called, based on the most abundant sRNA. ShortStack then made 'DicerCalls' for each identified sRNA cluster, which called the most dominant sRNA lengths at each given aligned cluster. RNAs not related to the RNAi process were given a DicerCall of 'N', which are often fragmented products of abundant RNAs.

Read counts were analysed using edgeR to detect DE sRNA clusters between rogue and wild types, using normalisation and statistical tests previously described for DE analysis of mRNA data. I further analysed sRNA cluster data using the plyranges and the GenomicRanges R packages [58], v1.42 to identify loci with sRNA clusters within 10 kb.

## 2.9 Whole-genome DNA methylation analysis

I used whole-genome CG methylation data previously generated by one of my supervisors, Quentin Gouil. Genomic DNA from leaf tissue of one Sutton's Hundredfold rogue and one

Caméor wild type was sequenced on one PromethION flow cell per sample. Reads were basecalled with Guppy v4.0.11 and methylation in the CG context was called with Nanopolish v0.13.2 [59].

Table 2.1: **Summary of transcriptomic and methylation data sets**

| **Target** | **Dataset** | **Data** | **Wild type replicates** | **Rogue replicates** | **Cultivars** | **Purpose** |
|---|---|---|---|---|---|---|
| mRNA | mRNA1 | mRNA 75SE sequencing | 3 | 3 | Thomas-Laxton, Black-Eyed Susan, Daisy | Discover DE genes between epigenotypes |
| mRNA | mRNA2 | mRNA 75PE sequencing | 2 | 2 | Thomas-Laxton, Black-Eyed Susan, Sutton's Hundredfold | Discover DE genes between epigenotypes |
| mRNA | Long-read | Long-read PCR cDNA sequencing | 3 | 3 | Thomas-Laxton, Daisy | Improve transcriptome annotation |
| sRNA | sRNA | small-RNA sequencing | 3 | 3 | Thomas-Laxton, Daisy, Sutton's Hundredfold | Identify sRNA clusters neighbouring genes |
| gDNA | CpG Methylation | Nanopore direct DNA sequencing | 1 | 1 | Caméor, Sutton's Hundredfold | Identify differentially methylated regions or loci |

## 2.10 Candidate gene selection criteria

Candidate gene selection for the rogue locus from transcriptomic analysis was mainly determined by the commonality of DE genes across cultivars, as I hypothesised the rogue locus is conserved across paramutant pea cultivars. If a given candidate gene was repressed in some cultivars but not others, additional criteria based on assumptions of the RdDM model were used: presence and differential accumulation of small RNAs, presence and differential abundance of DNA methylation (Figure 2.2). Other selection criteria included consistency in the mRNA2 data set, and potential candidate gene function and implicated gene pathway in other organisms . Gene functions were investigated using the Basic Local Alignment Search Tool (BLASTn/BLASTx) from the National Center for Biotechnology Information (NCBI), using the standard nucleotide database (highly similar sequences) and the non-redundant protein sequences database (default search options).

Candidate genes which translated to functional proteins were then further investigated using the UniProt protein database to obtain information about the protein function and associated gene families (Uniprot.org, accessed 02/08/2021). Using my selection criteria, I then created a candidate ranking matrix to identify the strongest candidates, and to reduce any bias in candidate gene selection for the rogue locus. Each gene receives a score based on how many selection criteria conditions the gene has met; e.g. if gene1 is significantly down-regulated in 3 rogue cultivars, and has sRNA clusters within 10 kb, it receives a score of 4.

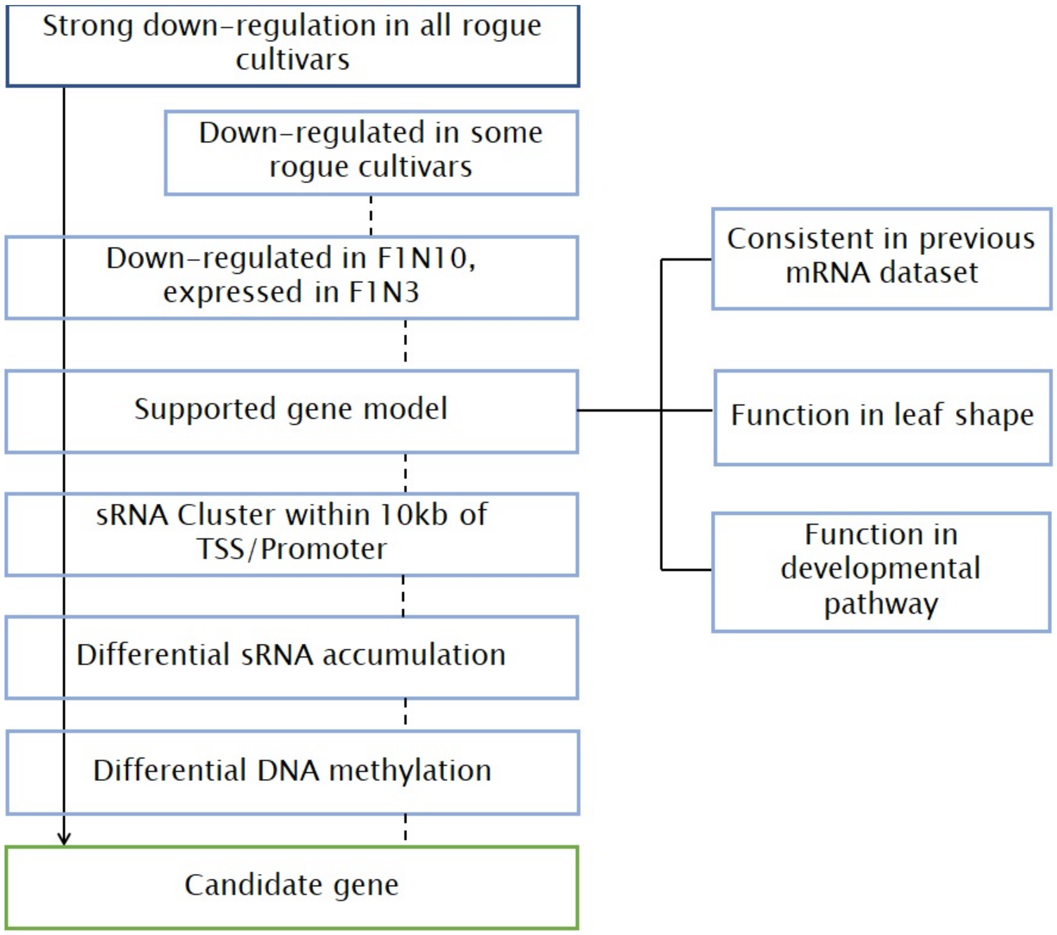

Strong down–regulation in all rogue cultivars
Down–regulated in some rogue cultivars
Down–regulated in F1N10, expressed in F1N3
Supported gene model
sRNA Cluster within 10kb of TSS/Promoter
Differential sRNA accumulation
Differential DNA methylation
Candidate gene
Consistent in previous mRNA dataset
Function in leaf shape
Function in developmental pathway

FIGURE 2.2: **Candidate gene selection.** Selection criteria for rogue candidates largely focused on repressed genes identified through DE analysis which are common across all assessed cultivars, coinciding with my hypothesis that the rogue locus is conserved across cultivars of *P. sativum*. Genes which may not be repressed in all cultivars, but fit other selection criteria such as presence of sRNA abundances and gene models involved in plant developmental pathways are also considered as good candidates. F1N10 = 10th leaflet node of F1 hybrid, F1N3 = 3rd leaflet node of F1 hybrid. TSS = transcription start site. Solid arrow indicates direct gene candidature, dotted lines indicate further evidence to be considered if candidate isn't down-regulated in all cultivars.

## 2.11 Crossing & phenomics analysis

Mature wild type and rogue peas for each cultivar were crossed by selecting inflorescences which had dehisced their pollen and applying the pollen to appropriately developed recipient flowers. Rogue peas were used as donors and wild types as recipients, with 10 replicate crosses per cultivar. F1 generation seeds were harvested once pods had dried and were re-sown under the same growth conditions and soil medium as the previous parent generation. Controls from parent lines were also re-sown to allow later phenomic comparison of the F1s, to capture any potential shift in phenotype, which may be indicative of a successful cross. Once individuals had developed 10 leaf nodes, whole stipules from the first 8 nodes were carefully excised, and image-scanned straight away for later analysis. Raw image files of whole stipules for each individual were first processed in GIMP (GNU Image Manipulation Program) in order to manually segment each stipule into a single image file. The segmented stipule images were then converted to binary images and filled to ensure no gaps were present in the image after conversion, using a custom ImageJ Fiji script. The stipule images originating from the left side of the plant were flipped, so the image orientation matched the stipules from the right side. Manipulated images were then imported into R and analysed using the Momocs morphometrics package ([45]). I first used the Elliptical Fourier Transformation (EFT) to decompose the leaf shapes into harmonics, which I then used to determine and compare the average shape for each epigenotype. I continued by using Momocs 'coo' functions to calculate the length (coo_length), width (coo_width) and area (coo_area) of stipule samples, which I then used to calculate length-to-width ratios for differentiating between rogue and wild types samples, as well as F1 and F2 hybrids for each cultivar.

## 2.12 Automated classification model

Stipule area and length-to-width ratio data generated from the Momocs analysis pipeline was exported and utilised in the KNIME Analytics Platform as phenotypic descriptor variables to differentiate between rogues and wild types. The Decision Tree Learner model was used to characterise stipule samples by randomly selecting 80% of the data set as training data for the model, and the remaining 20% as the test set. Due to the reduced sample size for the Black-eyed Susan cultivar, 65% of the data was used as the training set, and 35% as the test set in this case, though results were uninformative due to the small sample size. The Gini Index splitting method was used to calculate the most powerful variable which differentiated between rogues and wild types as the root node, and to subsequently determine node splits. Reduced error pruning was used to reduce tree complexity and avoid over-fitting data. Through systematic testing of modeling variables I found the Gini Index and reduced error pruning methods best suited my data and produced the most accurate results. The model ran iterations of the analysis until all stipule samples had been categorised as rogue or wild type. Classifications were then summarised in a confusion matrix, detailing correct classifications, mis-classifications, model accuracy, error rate and the Cohen's kappa coefficient (k).

A single model to encompass all cultivars was unsuitable due to cultivar-specific phenotypic properties which impeded the pipeline's ability to accurately classify individuals into their correct epigenotype. I tested the classification model using combined phenomics data from individuals used in both mRNA1 and mRNA2 data sets, though this caused higher mis-classification rates were present due to variances between each data set. This may be the result of plant material for each data set originating from different accessions and seed batches. Therefore, only phenomics data from individuals in the mRNA1 dataset were used for epigenotype classification of individual cultivars.

I then later used all mRNA1 rogue and wild type parent data for each cultivar to re-train the classification model, and assessed the inheritance pattern in F1 and F2 hybrids of each respective cultivar. As most hybrids transitioned from wild type to rogue from the 4th node on-wards, I used a 'majority rules' method to classify hybrids as rogue or wild type using nodes 4-8. This allowed me to classify whole individuals as rogue or wild type-like, rather than just at the stipule level used in parent classification. As an example, if plant X's hybrid nodes 4-7 are classified rogue, I deem the entire individual as rogue.

$Gini = 1 - \sum_{i=1}^{n}(pi)^2$

where $pi$ is the probability of an object being classified to a particular class.

$k = \frac{2\times(TP\times TN - FN\times FP)}{(TP+FP)\times(FP+TN)+(TP+FN)\times(FN+TN)}$

where TP are the true positives, FP are the false positives, TN are the true negatives, and FN are the false negatives.

# Chapter 3

# Development of phenomics methods to enable accurate rogue characterisation and classification

## 3.1 Summary and Objectives

The first morphometric properties of rogue peas described in the early 1900's were the reduction in leaflet and stipule size and their change in shape [6]. These studies were conducted on the Gradus cultivar, where the unusually high penetrance of the rogue phenotype was noted through analysing the length-to-width ratio of leaflets and stipules of rogue plants, wild type plants, and the successive generations of their progeny. Though this key phenotypic property was identified, investigation of the epimutant phenotype was limited due to the use of a single mutant cultivar. To fully characterise and describe the rogue phenotype, the analysis of multiple cultivars and their associated hybrid crosses is required because it will more accurately represent the altered rogue morphometric traits.

In my experiments I aimed to establish a quantitative description of the rogue phenotype across multiple cultivars, and use it to systematically test the inheritance pattern of rogue leaf morphology in rogue and wild type hybrid crosses in the F2 generation. If crosses were successful, I would expect to see a shift in the leaf phenotype towards the rogue phenotype and a deviation from the wild type phenotype, particularly in leaf area and length-to-width ratio as the F1 hybrid develops. Prior studies on the rogue phenotype describe the inheritance as highly penetrant in the F2 generation, and in subsequent generations [6, 8]. I would expect the same result for my selected cultivars if they are indeed 'true' rogues. My crossing study was therefore used to investigate inheritance of the rogue phenotype and to characterise rogue stipule morphometric properties.

My crossing studies also served to establish phenomic parameters that can be used as a proxy for genotyping, or epigenotyping, my plant material. No markers are available to conduct a standard genotype screening due to the DNA sequence similarity between epigenotypes within cultivars and the fact that the causal rogue locus has not been identified. This prevents high-throughput genotypic characterisation of offspring in crossing studies. I aimed to utilise the phenomic descriptor variables of length-to-width ratio and leaf area to begin developing a high throughput analysis technique for the rogue pea utilising and automated machine learning method. This approach should enable efficient characterisation of leaf morphometrics which will then allow for samples to be categorised to their correct epigenotype.

## 3.2 Quantifying leaf morphology in rogues and hybrids

Using phenomic analysis, I investigated if the rogue phenotype inheritance properties described in the literature are true in my selected cultivars. From the literature, I expected rogue peas to display reduced stipule area in comparison to wild types and a larger length-to-width ratio (l/w), which quantifies the pointed ‘rabbit-eared’ rogue phenotype [6, 8]. I expected the rogue phenotype would be inherited at high rates in rogue/wild type hybrid crosses, and an intermediate phenotype would be observed in F1s, while a more drastic shift to a complete rogue phenotype would be observed in F2s. To test this, I grew rogue and wild type parent populations of the Thomas Laxton cultivar, along with populations of rogue and wild type F1 and F2 hybrid crosses, and extracted morphometric data from image scans of their leaves.

Thomas Laxton rogue and wild type epigenotypes were clearly differentiated by both l/w ratio and stipule area (Figure 3.1 A, B). Wild type l/w ratios peaked at approximately 1.7 with a mean of 1.71, whereas rogues peaked at 2.35 with a mean of 2.3 (Table A.4). This differentiation between parent epigenotypes is also reflected by their stipule area, measured in pixels (k = thousand, px = pixels). Wild types show large variation in their size spanning between 50 kpx to 150 kpx with a mean of 111.9 kpx (Table A.4) while rogue samples have a distinct peak at 50 kpx (Figure 3.1 B) with a mean of 39.5 kpx (Table A.4). Some of the variation observed in stipule area is derived from the age of individual leaves, as multiple leaf nodes are plotted together - this is most obvious in the wild type samples (3.1).

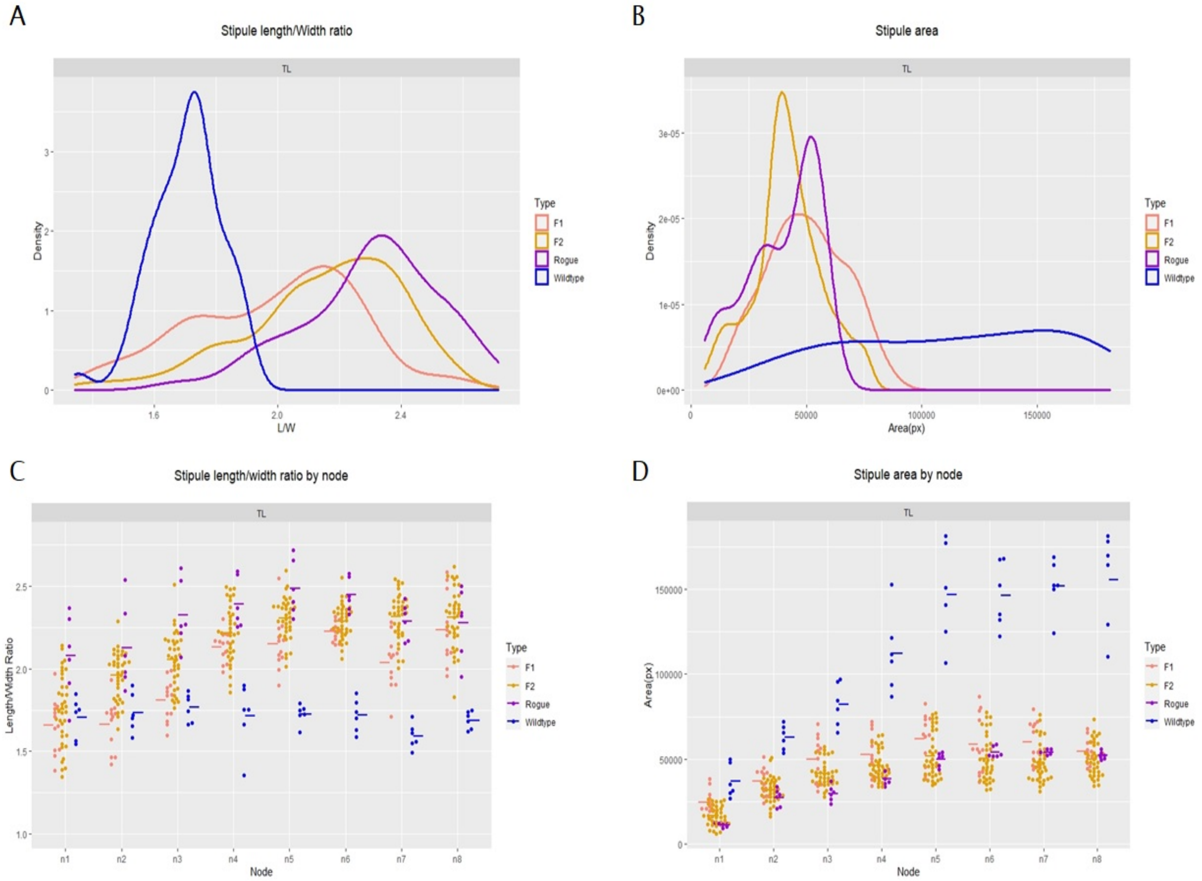


Figure 3.1: **Thomas Laxton rogue inheritance fits classical definitions.** Distribution of stipule length-to-width ratios (A) and areas (px) (B) of wild type and rogue epigenotypes, as well as F1 and F2 generation hybrid crosses. Individual data points for each stipule node for both l/w ratio (C) and area (D) are shown. (Stipules total = 480: Rogue = 48 (3 individuals), wild type = 48 (3 individuals), F1 = 96 (6 individuals), F2 = 288 (18 individuals)).

F1 and F2 hybrid offspring were more similar to rogues than to wild types in both l/w ratio and stipule area. F1 samples averaged a 1.99 l/w ratio, and F2 a 2.15 l/w ratio (Table A.4). Of note, most F1 and a subset of F2 stipules from lower leaflet nodes 1–4 more closely resembled the wild type stipules and were responsible for the increased number of observations with a l/w ratio below 2 (Figure 3.1 A, C). Similarities between rogue parent and hybrid samples were also observed in stipule area, where density peaks clustered together at approximately 50 kpx (Figure 3.1 B). Rogues and hybrids split from wild types by stipule area from as early as the 2nd node (Figure 3.1 D). Overlaying the mean leaf shapes for each epigenotype and their hybrid offspring confirmed that rogues displayed a much narrower and smaller stipule relative to its wild type counterpart (Figure 3.2 D). The F1 hybrid mean shape differed somewhat from the wild type (Figure 3.2 E), while the F2 mean shape shifted more strongly towards the rogue phenotype (Figure 3.2 C,F).

From these analyses I determined that the inheritance patterns observed in hybrid crosses for the Thomas Laxton cultivar are consistent with expected patterns of rogue inheritance and paramutation, and exhibit what we know as the rogue phenotype according to classical definition.

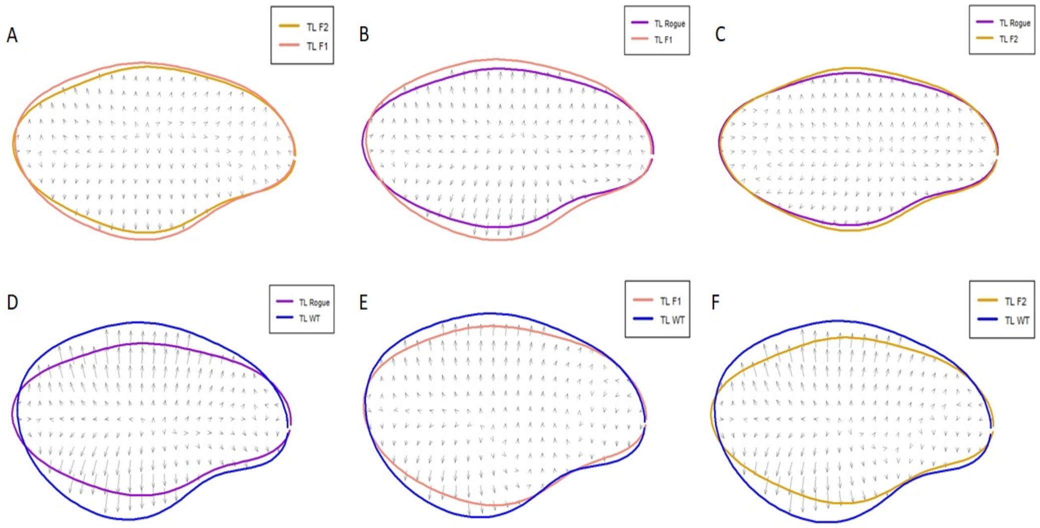


FIGURE 3.2: **Thomas Laxton mean stipule shape.** The average stipule shape for wild type, rogue, F1 & F2 samples generated using the Momocs morphometric and shape analysis tool in R. Stipules from all leaflet nodes were used to generate the mean shape. F1 and F2 mean stipule shapes closely resemble one another (A). F1s are slightly wider than rogues (B). F2s match very closely with rogues (C). Rogues and wild types show large variations from one another (D). The shift in phenotype can be observed comparing hybrids to wild type mean shapes (E,F). (Stipules total = 480: Rogue = 48(3 individuals), wild type = 48(3 individuals), F1 = 96(6 individuals), F2 = 288(18 individuals)).

I applied the same analytical approach to rogue inheritance patterns and leaf morphology traits in hybrid offspring of the Daisy cultivar. I found leaf morphology in the Daisy cultivar fulfilled rogue inheritance and phenotype expectations detailed in the literature. Wild type samples had a l/w ratio of 1.6, whereas rogue and hybrid cross samples peaked between

2–2.2 l/w ratio, reflecting a strong differentiation between wild type parents and the hybrid offspring which displayed rogue morphological traits (Figure 3.3 A). Rogue, F1 and F2 hybrid density peaks align closely, with similar mean ratios of 2.04, 1.87 and 1.95 respectively (Table A.3). Samples from nodes 1 and 2 for the F1 hybrid display similar l/w ratios and area before differentiating to rogue phenotypic properties in subsequent nodes. F2 hybrids primarily display rogue leaf morphology beyond node 1, though some samples took longer to display the rogue phenotype (Figure 3.3). This can be observed by the tails in the density curve for the F1 and F2 hybrids (Figure 3.3 A).

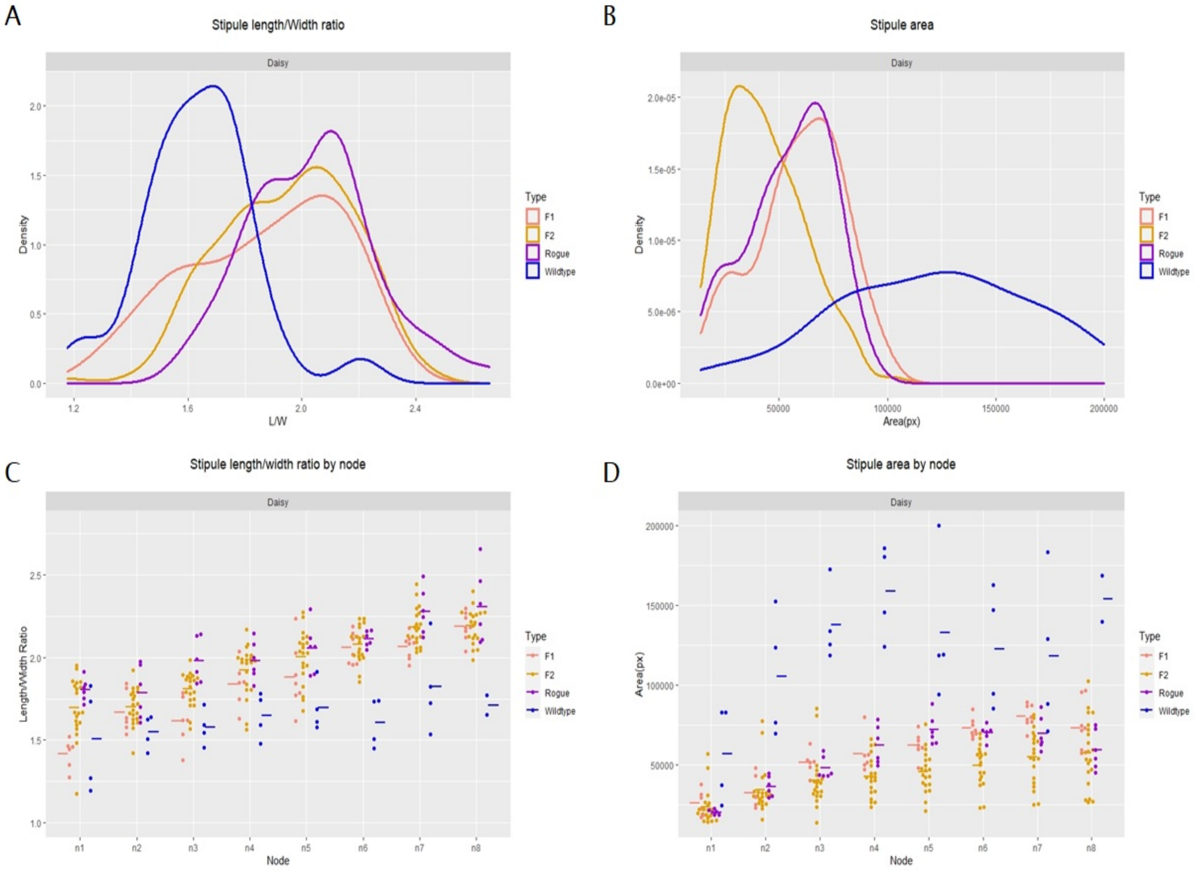


FIGURE 3.3: **Daisy hybrids have rogue phenotypes.** Density curve plots display the phenomic comparison of stipule length-to-width ratio (A) and area (px) of wild type and rogue epigenotypes (B), as well as F1 and F2 generation hybrid crosses. Individual data points for each stipule node for both l/w ratio (C) and area (D) are shown (Stipules total = 292: Rogue = 48 (3 individuals), wild type = 30 (2 individuals), F1 = 55 (4 individuals), F2 = 159 (10 individuals).

Similar patterns of leaf morphology differentiation were observed in stipule area (Figure 3.3 B) where wild type samples were larger in size in comparison to rogues and hybrid

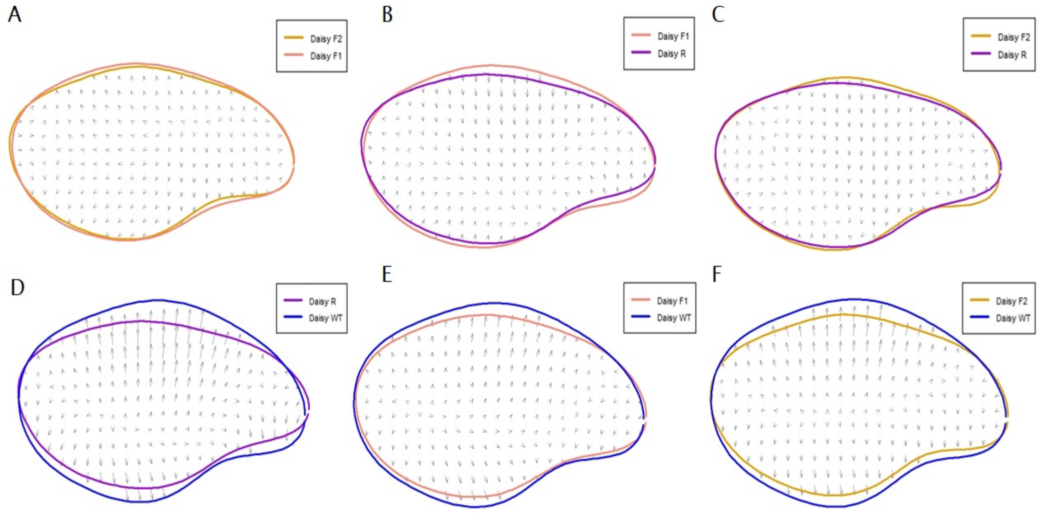


FIGURE 3.4: **Daisy mean stipule shape.**The average stipule shape for wild type, rogue, F1 & F2 samples generated using the Momocs morphometric and shape analysis tool in R. Stipules for all leaflet nodes were used to generate the mean shape. Rogues and wild types show large differentiation in leaf shape (D). Stipule shapes for both F1 and F2 hybrids closely resemble the mean shape of rogue parents (B, C). F1s and F2s also closely resemble each other (A). Both hybrid mean shapes are reduced in size relative to wild type parents (E, F) (Stipules total = 292: Rogue = 48 (3 individuals), wild type = 30 (2 individuals), F1 = 55 (4 individuals), F2 = 159 (10 individuals).

individuals - visualised by the broad density curve for wild types, and detailed by a mean stipule area of 121 kpx (Table A.3). As seen in Thomas Laxton, Daisy wild types also show a broad area distribution owing to analysing all leaflet nodes together, with variations occurring due to leaf age. Rogue and F1 hybrid sample peaks aligned almost perfectly with mean areas of 54.8 kpx and 59.1 kpx respectively, whereas the F2 offspring displayed greater reduction in stipule area with a mean stipule area of 43.5 kpx (Table A.3). The mean leaf shape comparison for Daisy rogue and wild types showed that rogue stipules were more elongated and reduced in size (Figure 3.4 D). Both F1 and F2 hybrid offspring shared leaf morphological traits consistent with rogue parent samples, and differed from the wild type parents (Figure 3.4 B, C, E, F). The Daisy F1 hybrids met expectations of an intermediate phenotype, as is detailed in the literature where lower leaf nodes first appear wild type-like and transition to rogues as they develop. The F2 generation closely align with rogue parents in l/w ratio and area, indicating a near-complete transition to the rogue phenotype in just 2 generations for this cultivar.

## 3.3 Automated epigenotype classification

As l/w ratio and area variables differentiate rogues and wild types generally well in both cultivars, they can be utilised as phenomic descriptor variables for the development of a machine learning classification method. My aims were to utilise the machine learning model to develop a non-arbitrary method of classifying individuals into their hypothesised epigenotype, based on leaf morphometric properties. An alternative approach based on leaf morphometric properties associated with each epigenotype was required because there were no known genetic markers between rogues and wild types within cultivars to conventionally genotype individuals. I developed a classification model pipeline for each cultivar based upon a decision tree learner using the KNIME analytics environment. With this method, I first to trained the learner to distinguish between rogues and wild type stipules in parent lines using phenomic properties of individual stipules. I then later re-trained the learner using all parent stipule data, in order to classify hybrid crosses into their hypothesised epigenotype at a whole-plant level using a majority vote approach for stipules from nodes 4 and above.

The l/w ratio and area variables of whole stipules for both Thomas Laxton and Daisy cultivars were sufficient for the decision tree learner to distinguish between rogues and wild types. In order to train the tree learner, 80% of parent data was subsetted for model training, and the remaining 20% used for model testing. The subsetted data for model training was processed in 2 branches which split by threshold values determined for each phenomics variable by the decision tree learner under the supervised learning approach (Figure 3.5). Thomas Laxton splits first by area, with a threshold value $> 59.742$ kpx which separated most wild type samples (Figure 3.5 A). The remainder of samples $< 59.742$ kpx area then split by l/w ratio, where samples $> 1.85$ were all rogues, and $< 1.85$ split the remainder of the wild type samples. Results of the Daisy decision tree splits were similar to Thomas Laxton, just with values specific to the cultivar (Figure 3.5 B). Once training was completed and classification thresholds were established, the remaining 20% of parent data was used to test the decision tree model.

I first tested classification for the Daisy cultivar, with the aim of achieving a classification rate of 90% to prove the viability of the decision tree learner method to non-conventionally assign hypothesised epigenotypes to individual cultivars, and later hybrid progeny. The model for the Daisy cultivar utilising area and l/w ratio descriptors achieved a classification

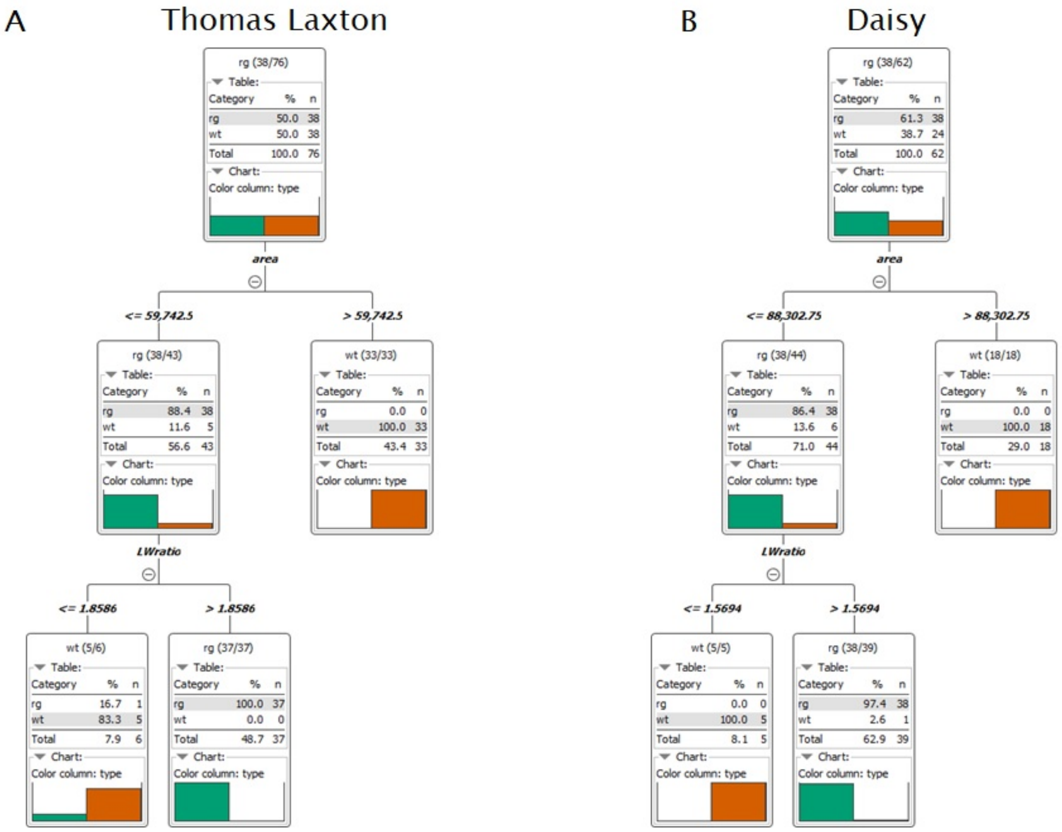


FIGURE 3.5: **Phenomic variables distinguish rogues and wild types** The training subset of phenomics data (80%) for parent lines of Thomas Laxton (A) and Daisy (B) cultivars set thresholds for l/w ratio and area phenomic variables. These thresholds were then used to classify the remaining stipule samples used as the test set (20%). (Daisy stipules total = 78: Rogue = 48 (3 individuals), wild type = 30 (2 individuals); Thomas Laxton stipules total = 96, Rogue = 48 (3 individuals), wild type = 48 (3 individuals))

accuracy of 93.75% (Table 3.1). Out of 16 stipule samples used as the test set, 15 were correctly classified under the supervised learning approach (Figure 3.6 & Table 3.1, k = 0.86). The Daisy test set results can be observed in figure 3.6 A + B for each pheonomic variable. The misclassified wild type sample can be observed amongst correctly classified rogue samples in both area and l/w ratio plots (Figure 3.6 A, B).

The ability of the classification model for the Daisy cultivar to correctly assign the majority of test samples showed the decision tree learning method is robust enough to detect differences between epigenotypes, and is suitable for non-conventional genotyping in this case. Success of the decision tree learner was similar for the Thomas Laxton cultivar. The model achieved a classification accuracy of 100% with 20/20 stipule samples correctly classified under the supervised learning model (Figure 3.6 & Table 3.1, cohen's kappa (k) = 1).

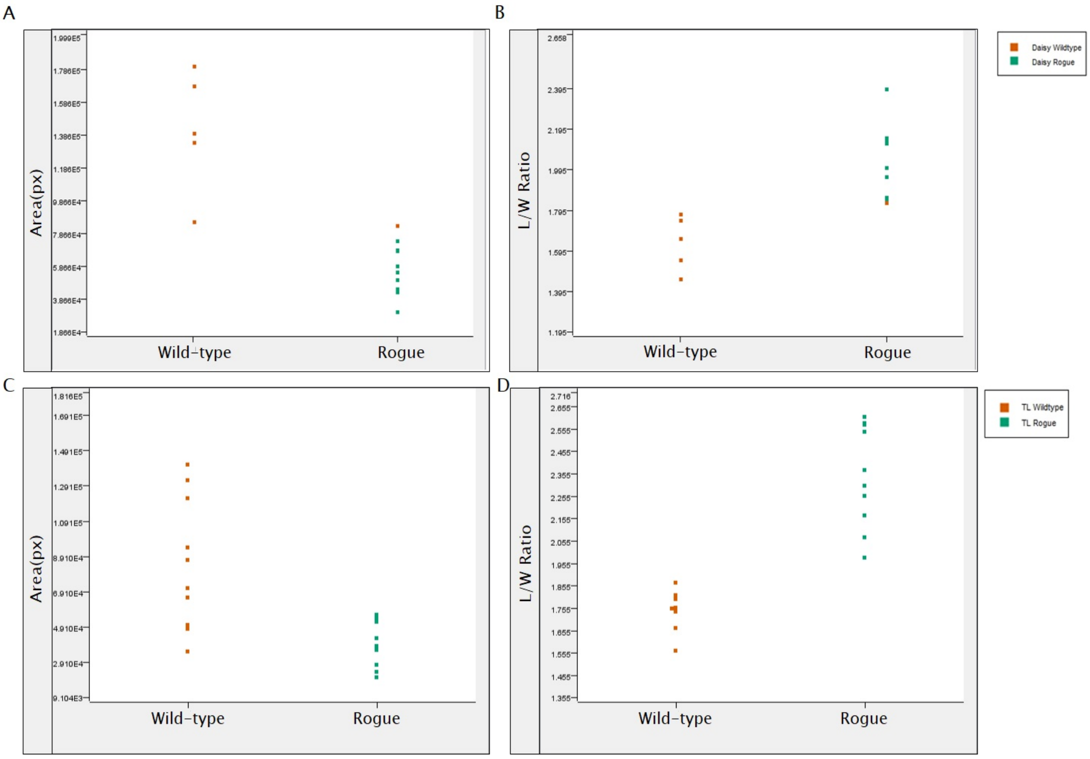


FIGURE 3.6: **Phenomic variables accurately classify rogues and wild types.** Whole stipule phenomic classification of P. sativum cultivars Daisy (A, B) and Thomas Laxton (C, D) using the KNIME decision tree machine learning model (80% of data used for model training, remaining 20% as test set for both Daisy and Thomas Laxton). (Daisy stipules total = 78: Rogue = 48 (3 individuals), wild type = 30 (2 individuals); Thomas Laxton stipules total = 96, Rogue = 48 (3 individuals), wild type = 48 (3 individuals)).

TABLE 3.1: **Phenomic classification results**

| **Cultivar** | **Correct Classifications** | **Misclassified** | **Model accuracy** | **Error rate** | **Cohen's kappa** |
|---|---|---|---|---|---|
| Daisy | 15 | 1 | 93.75% | 6.25% | 0.86 |
| Thomas Laxton | 20 | 0 | 100% | 0% | 1 |

Overall the models for each cultivar were effective in classifying rogue and wild type peas by the phenotype of their stipules. This approach provides the means for faster phenomic classification and provides an objective method for epigenotype classification.

Once I had determined the model could accurately predict rogue and wild type epigenotypes based on stipule phenomic properties of l/w ratio and area, I trained the model using all rogue and wild type parent data for each cultivar and classified F1 and F2 hybrid crosses at the whole-plant level. As previously stated, whole hybrid plants were classified as either rogue or wild type based on the classification given to the majority of stipules for said individual (see methods).

All Thomas Laxton F1 and F2 hybrid individuals were classified as rogue using the re-trained decision tree learner (Figure 3.7 A, B). Similar to the results observed from phenomic analysis in Figure 3.1 C & D, the first 3 nodes were classified as wild type in most cases, though almost all subsequent nodes classified as rogue (Figure 3.7 B). Therefore, the plant-level classification for all 24 hybrid individuals of Thomas Laxton was rogue, based on the classification for the majority of stipules belonging to each plant (Figure 3.7 A).

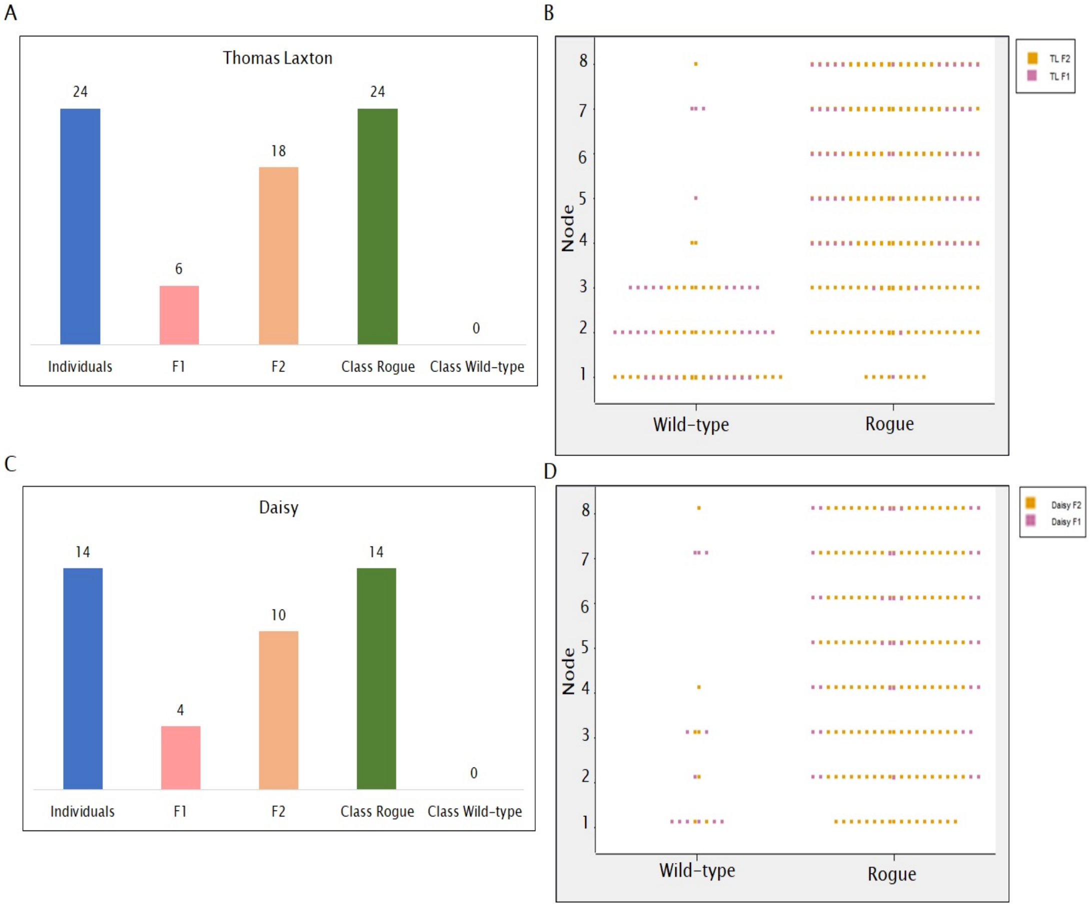


Figure 3.7: **Hybrid progeny classify as rogues.** F1 and F2 generation hybrids of Rogue + wild type crosses for cultivars Thomas Laxton (A, B) and Daisy (C, D) classify as rogues based on stipule l/w ratio and area. Stipule phenomics data for parent lines of each cultivar were used to train the decision tree learner (Thomas Laxton; wild type = 3 (48 stipules), Rogue = 3 (48 stipules), F1 = 6 (96 stipules), F2 = 18 (288 stipules), Daisy; wild type = 2 (30 stipules), Rogue = 3 (48 stipules), F1 = 4 (64 stipules), F2 = 10 (160 stipules).

Similarly, all 14 Daisy F1 and F2 hybrids classified at a whole-plant level as rogues, based on the classification of the majority of stipule samples beyond node 3 (Figure 3.7 C). Utilising rogue and wild type parent phenomics data to classify hybrid offspring was largely successful. Whilst the decision tree learner model is simplistic, it is also proves to be an effective method of classifying plants based on phenomic descriptor variables.

## 3.4 Conclusion

F1 hybrid crosses between rogues and wild types of Thomas Laxton and Daisy cultivars displayed the expected intermediate phenotype described in literature, where initial leaflet node phenomic parameters are similar to wild type parents, but later leaf nodes differentiate towards the rogue phenotype. In my assessed cultivars, this shift in stipule phenotype was primarily observed from node 4 (Figure 3.1 C, D; Figure 3.3 C, D). Interestingly, while morphometric properties of the F2 generation of both cultivars are more rogue-like compared with the F1 generation, Thomas Laxton F2s display a degree of variation in their transition to the rogue phenotype. This is based on mean l/w ratios and area, which are somewhere between a rogue and an F1 hybrid. A probable conclusion may be that the epimutation does not completely establish in the first two generations of hybrid crosses in Thomas Laxton, though further molecular investigation is required to confirm this mechanism. This phenotype in Thomas Laxton hybrids is consistent with observations from early studies, where varied levels of 'rogueness' were recorded in hybrids [6].

The leaf morphometric analysis overall was successful in investigating the inheritance of the rogue phenotype in hybrid crosses, and in characterising the phenotypic properties for each individual cultivar. Rogues and wild types can be differentiated by l/w ratio and area, and these phenomic variables were able to automate accurate and unbiased epigenotype classification through the decision tree learner method. While this was a good first attempt at automated classification, there are improvements which could be made in the future to strengthen the analysis. First, as phenotype shifts are most prevalent from node 3-4, data could be subsetted to only include nodes beyond this point. This would increase accuracy and precision of machine learning approaches, and also remove leaf morphology variability introduced by leaf age. Combining multiple populations of rogue/wild type parents and hybrids rather than singular populations used in this study would also increase the power of analysis. Lastly, including more pheontypic data on other components of rogue phenotypes will offer further variables to accurately differentiate between rogues and wild types through automated classification. It was noted by Bateson and Pellew (1915) [6] that rogue epimutants of the Gradus cultivar were found to have a pea pod phenotype, where the shape and curvature of the pea pods differentiated from wild types. Other rogue phenotypic traits are likely present in my selected cultivars, but it was not feasible to investigate these due to time limitations of the project. Future work should involve whole-plant phenotype investigation, and integrate new-found traits into the modelling for epigenotype characterisation.

# Chapter 4

# Identification of candidate rogue loci through short- and long-read transcriptomics

## 4.1 Summary and Objectives

There has been little investigation of the causal locus underlying the rogue phenotype at a genomic and transcriptomic level to-date. As discussed in Chapter 1, the study conducted by Santo et al. (2017) [40] provided little progress towards identifying the paramutant rogue locus and its inheritance mechanism. It is therefore necessary to use other well characterised paramutant examples like maize, as models for paramutation mechanisms in *P. sativum*. All current characterised examples of paramutation in eukaryotes involve the heritable and recurrent transcriptional silencing of a gene *via* trans-homolog interactions. I therefore hypothesised the epimutated rogue locus behaves similarly to other eukaryotic examples, and is transcriptionally silenced or repressed after trans-homolog interactions have taken place. I also hypothesised that the rogue epimutation is conserved across rogue cultivars of *P. sativum*, where the same phenotypic rogue traits affecting leaf morphology have been observed. Prior research to identify the rogue locus was limited due to the lack of a publicly available reference genome and transcriptome, though this has recently changed. A new *P. sativum* reference published by Kreplak et al. (2019)[11] now enables the use of genome-wide transcriptomic assays. In my study, I used long and short-read RNA-sequencing to identify candidates for the rogue locus.

## 4.2 Enhancing the pea transcriptome annotation with long-read sequencing

Preliminary analyses of short read RNA-sequencing data generated from whole stipules of cultivars Thomas Laxton, Daisy and Black-Eyed Susan revealed that transcriptome-wide analyses might be limited by two factors: an incomplete genome sequence and an incomplete transcriptome annotation.

The pea reference genome was initially constructed from short Illumina reads using the Caméor cultivar [11]. The assembly covers 88% (3.92Gb) of the estimated pea genome size (4̃.45Gb). Kreplak et al. (2019)[11] reported that remaining errors consist of collapsed repeat regions, at centromeres and telomeric regions, and comment that long-read sequencing approaches would provide the coverage necessary to rectify these gaps. I determined how well my short read RNA-seq data could be aligned to the original pea reference genome, and found a range of 64–84% of reads were successfully mapped to *P. sativum* genome reference v1a. These results were lower than expected, as typical alignment rates for other organisms such as *A. thaliana* exceed 90% [60] (Table 4.1). This confirmed that gaps and errors remain in the v1a genome assembly, which could impact my ability to identify the rogue locus.

Table 4.1: Comparison of alignment rates for *A. thaliana* vs *P. sativum* preliminary analysis

| | ***A. thaliana*** | ***P. sativum* (preliminary analysis)** |
|---|---|---|
| **Alignment rate** | 92.4% - 99.5% | 63.82% - 83.98% |

RNA-seq read assignment rates, defined by the amount of reads aligning with genes or other genomic features, were also lower than in *A. thaliana*. When visualising the mapped reads, I found that many reads mapped to regions missing annotation, indicating improvement of the reference transcriptome annotation was required (Figure 4.1). Transcriptomic studies are largely limited by the availability and quality of the reference, and it is possible the rogue locus could be missed in my analysis due to gaps in the transcriptome annotation.

Paramutant loci have previously been found in repeat and centromeric regions. Examples include the tandem repeat sequences neighbouring the *b1* locus in maize and the pericentromeric *SLTAB2* locus in tomato [18, 24]. This indicates the importance of having a well annotated reference.

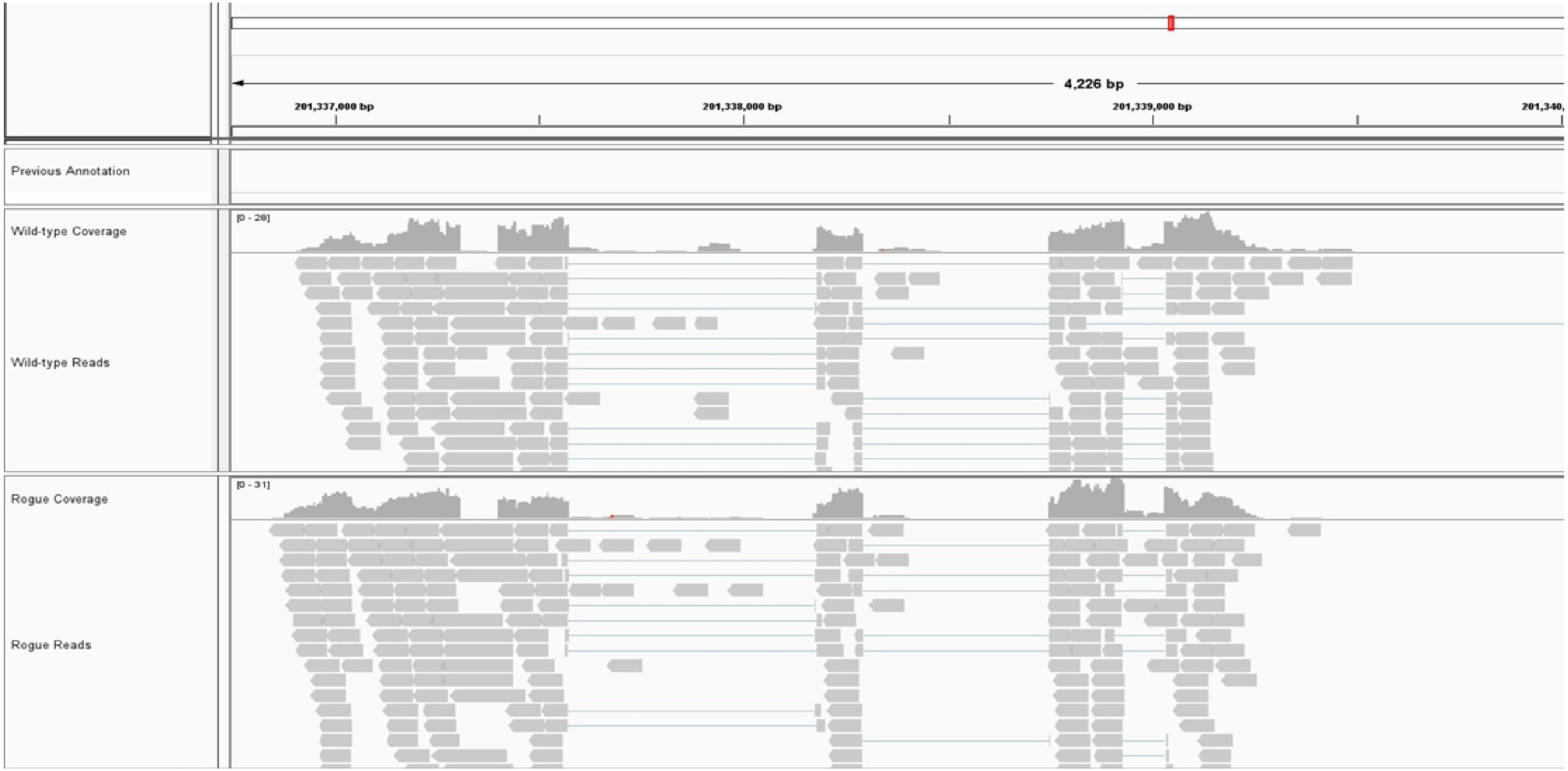


Figure 4.1: **Reads map to un-annotated regions.** An example of multiple short mRNA reads for Thomas Laxton wild type and rogue replicates mapping to un-annotated regions in the current transcriptome annotation for *P. sativum*. BAM files viewed in Integrated Genomics viewer (IGV) software.

I decided to improve the current transcriptome annotation in order to maximise the sensitivity of my later differential expression analyses. To do so I undertook long-read nanopore sequencing of the transcriptomes of two *P. sativum* cultivars, Thomas Laxton and Daisy. This included 3 replicates both rogue and wild type epigenotypes (Figure 4.2. The sequencing run produced a total of 23.85 M reads, with 19.85 M reads passing filter (Figure 4.3 A).

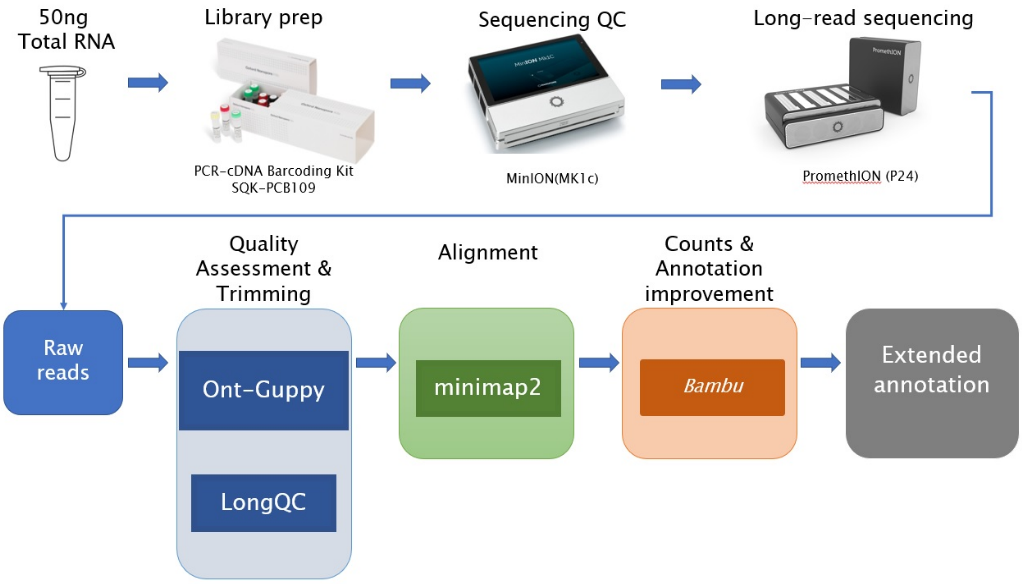


FIGURE 4.2: **Long-read transcriptome sequencing pipeline.** Pipeline used to generate long-read data from total RNA of 2 *P. sativum* cultivars. Refer to Method section for more detail (Chapter 2).

The distribution of read lengths (with outliers discarded) ranged from 280 bp to 2.5 kb (Figure 4.3 B), with an estimated N50 of 910 bp, meaning that half of the data is contained in reads longer than 910 bp. These lengths are typical for nanopore transcriptome sequencing. After trimming, the mean read length was 784 bp and N50 903 bp (Figure 4.3 C). The coverage over certain read lengths showed fluctuations (Figure 4.3 D), and per read coverage distributions showed relatively low coverage values (Figure 4.3 E). Here, coverage is defined by read overlaps in LongQC. The averaged quality value (QV) score for normal reads is acceptable - a QV score ranging between 8-10 is considered a sufficient quality value to proceed with analysis (Figure 4.3 F) [53]. While the read coverage results were not to a high standard, they were nonetheless adequate to make improvements to the reference transcriptome annotation.

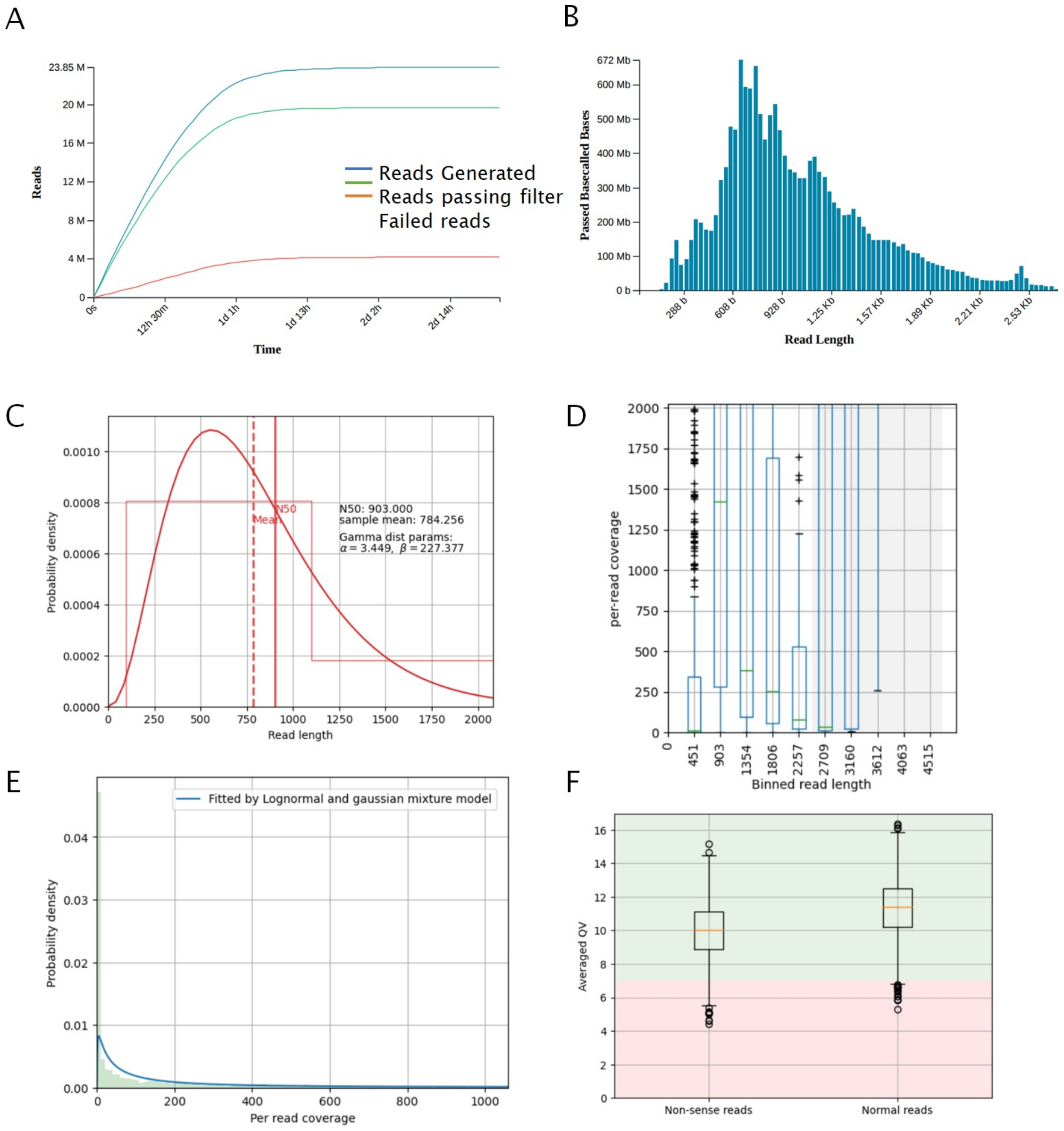


FIGURE 4.3: **Long-read transcriptomics quality assessment.** Sequencing run information shows number of reads generated (A) and read length distribution with outliers discarded (B). After quality trimming long-reads, LongQC was used to assess mean read length and quality (C), read coverage and coverage quality (D, E) and overall read quality values (F).

Satisfied with the quality of the long-read transcriptome data, I used *Bambu* to improve transcript annotation and quantify expression. *Bambu* is specifically designed for novel transcript discovery and quantification using long-read data, and integrates newly discovered transcripts and genes to an improved annotation. Clustering and PCA analysis of gene and transcript expression through *Bambu* showed that Thomas Laxton and Daisy cultivars separated, indicating transcriptomic variation between the two different cultivars

(Figure 4.4). Within each cultivar, rogue and wild type epigenotypes did not form distinct clusters, suggesting that transcriptome differences may be subtle. Of note, low per-sample sequencing depth limits the power of differential expression analyses, so I decided to focus on transcriptome annotation improvement with the long-read transcriptome data, and to perform differential expression analysis using short-read RNA-seq.

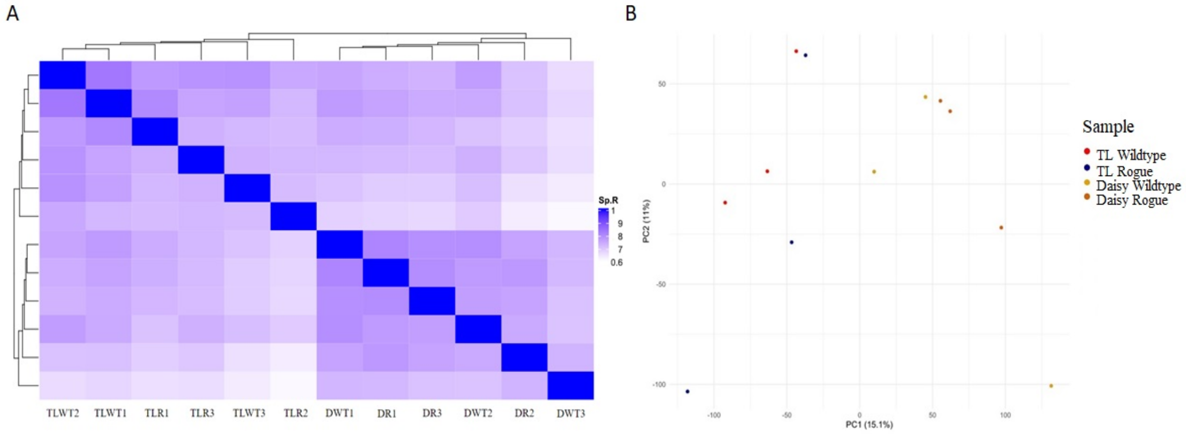


FIGURE 4.4: **Long-read sample clustering.** Cultivars of the same genotype cluster together by gene and transcript counts from long-read samples, but epigenotypes are less uniformly clustered (A). Plotting sample data in a PCA plot (B) shows similar results, with low amounts of variation in the data explained by PC1 (15.1%) and PC2 (11%)

I next investigated the extent of improvements made by *BAMBU* to the transcriptome annotation by comparing the amount of transcriptomic features and meta-features contained in the original and improved annotations. This was done using the FeatureCounts function from the Rsubread package. In this case, features correspond to exon regions while meta-features refer to entire genes. I found an increase in features (exons) over the previous annotation by 22.8% (83,856 new features), and an increase in meta-features (genes) by 32.8% (21,868 new meta-features) (Figure 4.5).

The increase in annotated features and meta-features resulted in an increase in read assignment rate for all mRNA data (Figure 4.5). Read assignment increased by 11.9% on average, and in some samples an improvement of >15% assigned reads could be observed (Table A.1). A visual inspection of the annotation identified several instances of improvement throughout. For example, the annotation for *Psat4g069560* was extended by 2 exons, well-supported by long-read data (Figure 4.6).

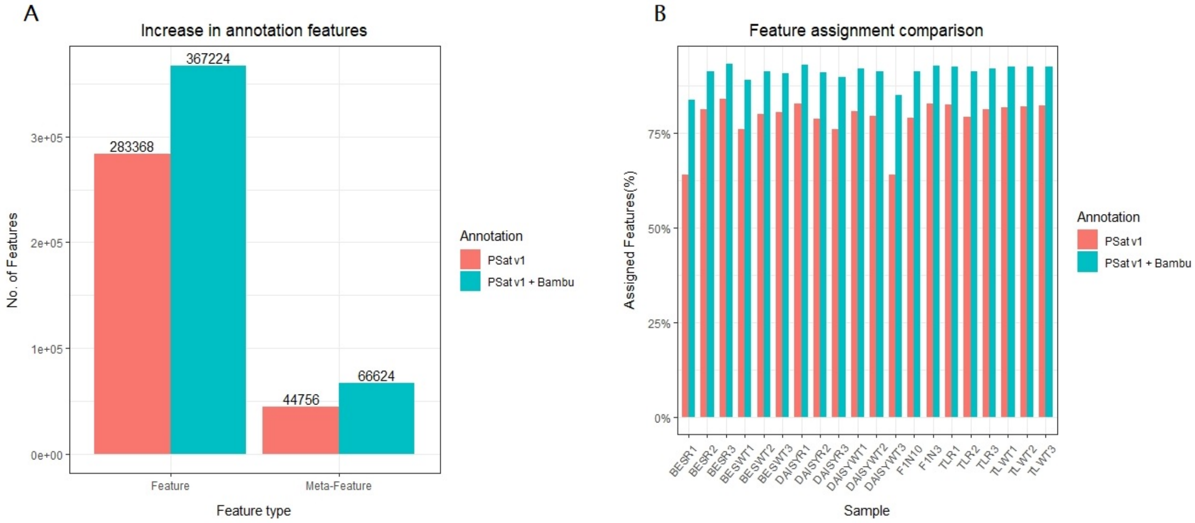


FIGURE 4.5: **Increased features and feature assignment.** Long-read sequencing data and *Bambu* improved the previous transcriptome annotation for *P. sativum* through the addition of 83,856 features and 21,868 meta-features (A). From additions to the annotation, an increase in reads assigning to features and meta-features for all samples was observed (B).

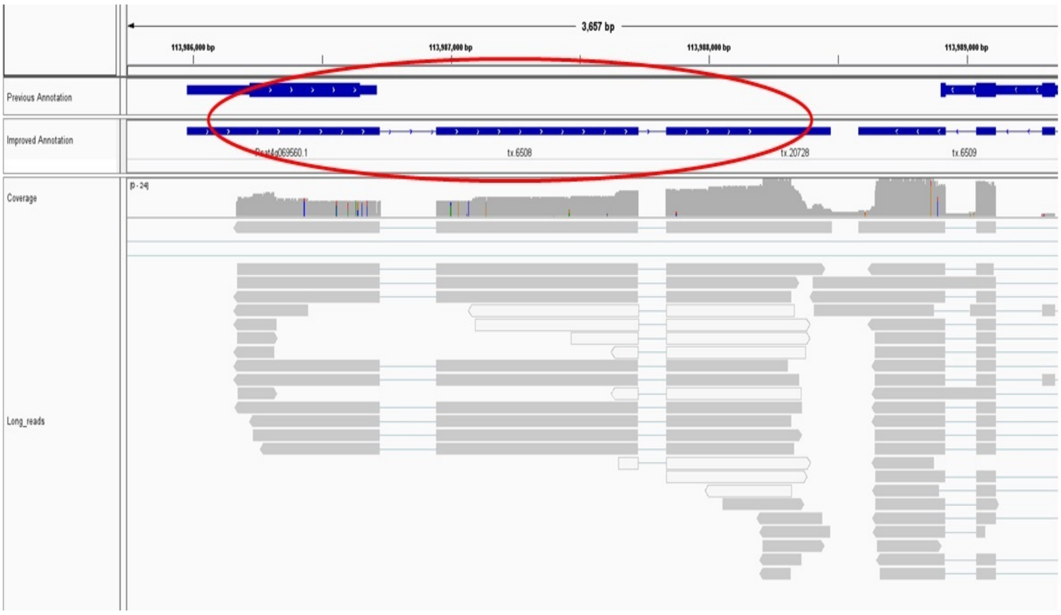


FIGURE 4.6: **Improved annotation of existing genes.** Long read coverage displayed in IGV supports additional exon and intron regions for Psat4g069560 - an example of alterations made to the transcriptome annotation. This is depicted by dark blue tracks displaying the gene model for both previous and improved annotations. The grey track for reads indicates good quality and accurate mapping for the supporting reads, and tracks with a white fill indicate secondary alignment, where the given read could align well to another region.

There were also regions where entirely new genes were added to the reference annotation (Figure 4.7), supported by uniquely mapped long reads. These were but two examples of annotation improvements which have been made across the entire reference annotation. Having a more complete transcriptome annotation increases the sensitivity of differential analysis and the probability of identifying strong candidate genes for the rogue locus.

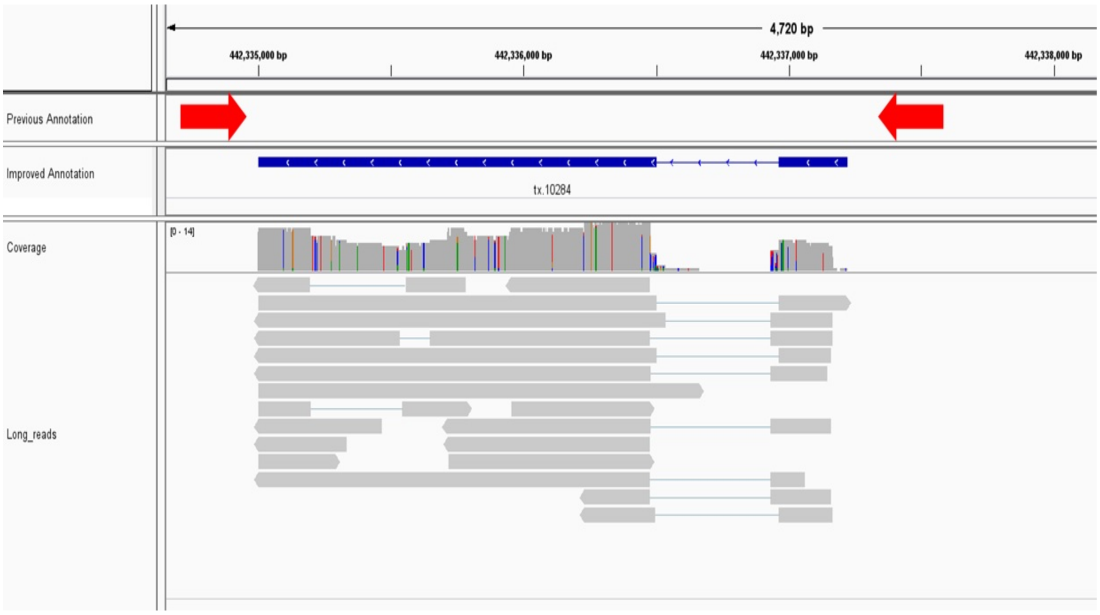


FIGURE 4.7: **Newly annotated genes.** Long-read transcriptome data supports addition of a new, active gene to the annotation depicted by the blue track in the improved annotation, and the absence of a gene model in the previous annotation. Coverage of all long read RNA-sequencing samples support 2 exons and an intron in this new gene's structure. The coverage value of 14 indicates 14 primary aligned long reads map to this region - primary alignments are depicted by grey read tracks. The coloured bars in the coverage track indicate potential SNPs, which are made visible if the given nucleotide varies from the reference in more than 20% of mapped reads.

The annotation improvements made through long-read sequencing and *Bambu* were overall successful, though there are some caveats of the analysis which were later identified. In the later stages of analysis I found that *Bambu* did not consistently handle feature annotations in the new GTF file. As a result, the number of meta-features reported by FeatureCounts includes all previously annotated transcripts and all the newly annotated genes, but not all **newly annotated transcripts**. This means some reads which map to newly discovered transcripts are not taken into account in subsequent DE analysis of my mRNA data. While the following DE analysis is less comprehensive than it could be, the validity of results presented hereafter remain unaffected by this small issue. Future work will involve rectifying the handling of the GTF annotation by *Bambu* to make future analysis as exhaustive and extensive as possible.

## 4.3 Genetic and transcriptomic differences between rogues and wild types are identified by short-read mRNA sequencing

Based on classical models of paramutation involving heritably repressed or silenced genes, I expected to find repressed genes in epimutant rogue peas relative to wild types. The repressed rogue locus may be involved in developmental pathways, in this case impacting leaf morphology and development. To investigate transcriptomic variation between rogues and wild type peas, and to identify candidate genes for the rogue locus, I conducted short-read mRNA sequencing on whole stipule tissue from 3 cultivars of *P. sativum* (Thomas Laxton, Daisy and Black-Eyed Susan) with 3 replicates for wild types and rogues. I used stipule tissue as the rogue phenotype is most obvious in this tissue. I also analysed stipules from node 3 and node 10 from an F1 hybrid cross, as the phenotype shift between lower and upper nodes in hybrids may be accompanied by gene expression shifts at the causal locus. Comparison of gene expression in these tissues may identify the causal locus of the rogue phenotype. To identify candidate genes for the locus, I conducted differential expression (DE) analysis of this short-read sequencing data against the improved transcriptome annotation generated with long reads.

I first analysed all cultivars together, in an attempt to identify a common down-regulated gene across all rogue cultivars. Multi-dimensional scaling (MDS) of gene counts showed that samples first separated by cultivar in dimensions 1 & 2 (Figure 4.8 A, B). The F1 hybrid samples also separated from other cultivars, likely due to genetic differences as the parent pair for this cross were Sutton's Hundredfold wild type and Sutton's Early Giant rogue. This pair of parents were selected as they were the only cultivars available at the time - ideally both parents would originate from the same cultivar. Black-Eyed Susan rogues and wild types separated completely from dimensions 1-3, indicating substantial transcriptomic variation may be present (Figure 4.8 A, C). Rogues and wild types from Thomas Laxton and Daisy were more closely clustered, indicating transcriptomic variation is more subtle. While F1 hybrid samples didn't perfectly overlap, the variation observed is more likely due to random or technical variation rather than transcriptomic variation.

Using the more comprehensive transcript annotation in this analysis did not have a dramatic effect on sample clustering on the MDS plot in comparison to preliminary analysis with the

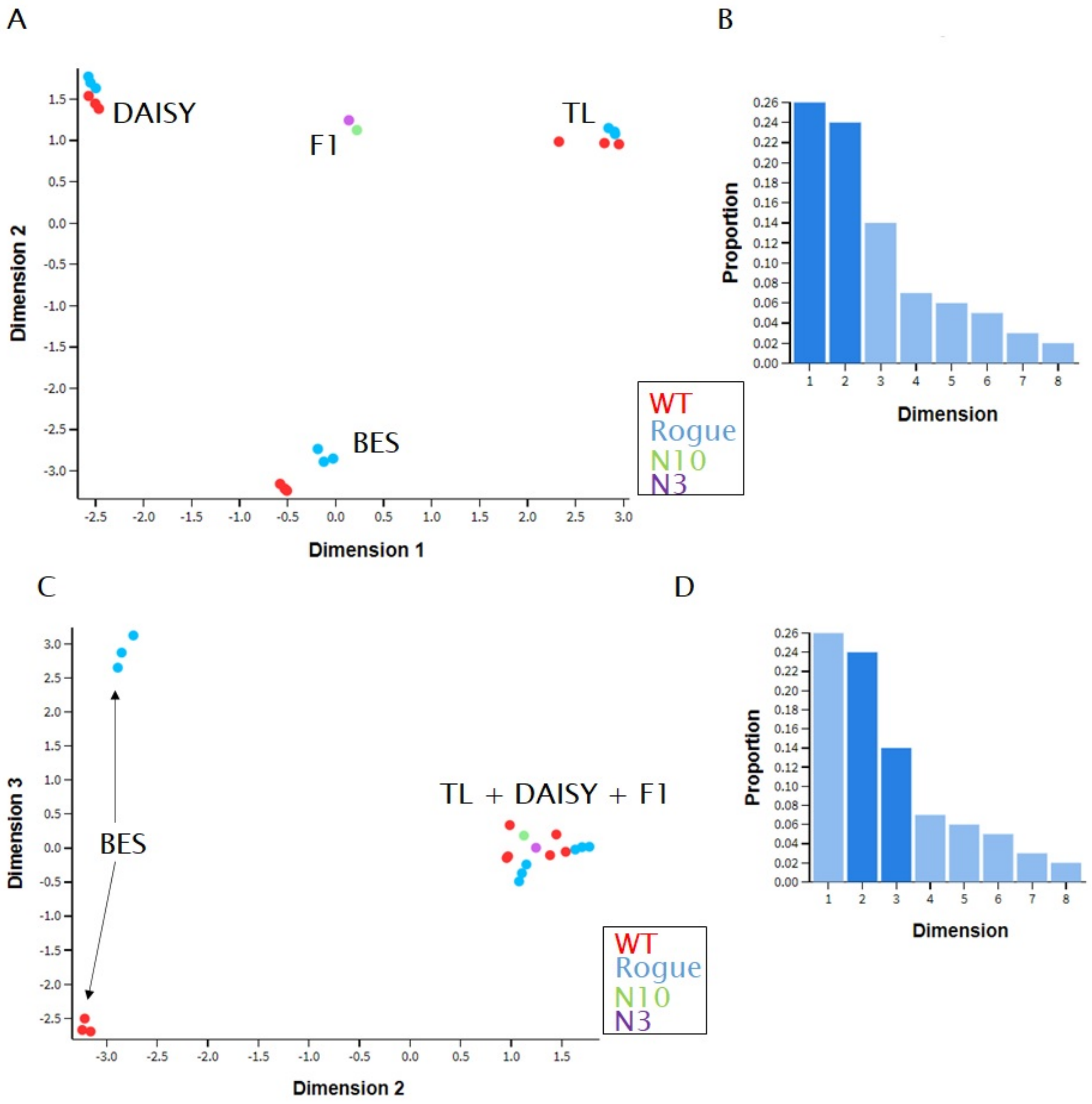


FIGURE 4.8: **Between-cultivar differences dominate transcriptomic variation.** MDS plot of short-read mRNA-seq read count data. (A) transcriptomic variation on the MDS plot is depicted by the distance of one data point to another. (B) Dimensions 1 & 2 separate cultivars best and explain 50% of the overall variances observed in the data-set (y axis = proportion of variance explained, x axis = dimensions). (C) MDS plot set with dimensions 2 % 3 displays large transcriptomic variation within the Black-Eyed Susan cultivar, and genetic differences to other assessed cultivars. (D) Dimension 3 can explain 38% of the overall variation in the data set. DAISY = Daisy, TL = Thomas Laxton, BES = Black-Eyed Susan, WT = Wild type, N3 = F1 leaflet node 3, N10 = F1 leaflet node 10.

previous annotation (Figure A.1). This suggested that the added features were not the main drivers of the sample separation.

The gene expression variation between Black-Eyed Susan rogues and wild types was much larger than between the rogues and wild types of either Daisy or Thomas Laxton cultivars (Figure 4.8). I investigated whether genetic differences might contribute to this variation, in addition to presumed epigenetic differences. I analysed single nucleotide polymorphisms (SNPs) in my stipule transcriptome data set (referred to as mRNA1 hereafter) and a second dataset that was generated prior to my study (referred to as mRNA2, whole leaves). All pea lines in mRNA1 (Daisy, Thomas Laxton, Black-Eyed Susan) were imported from the John Innes Centre (Norwich, UK) whereas in mRNA2 (Thomas Laxton, Black-Eyed Susan, Sutton's Hundredfold) the rogues came from the John Innes Centre and the wild types from the Australian Grains Genebank. I found Black-Eyed Susan samples in both mRNA1 and mRNA2 formed an outgroup relative to the other cultivars when they were clustered based upon SNPs found in commonly expressed genes (Figure 4.9). Divergence between rogues and wild types was greatest in the Black-Eyed Susan cultivar with an individual dissimilarity (ID) value of 0.4, which might explain the increased separation on the MDS plot. Importantly, there was clear evidence of genetic drift between the rogue and wild type lines within each cultivar. This could be expected as rogue and wild type lines have been kept separate for up to 100 years. Therefore gene expression differences between rogues and wild types within a cultivar have both genetic and epigenetic contributions.

The SNP analysis also showed that mRNA1 was the better data set to inform my work, due to the rogue and wild type pea lines in this data set originating from the same accession (John Innes Centre, Norwich, UK). Using material from the same accession reduces the potential for genetic or epigenetic noise confounding analysis, which may have occurred in the material in mRNA2 where rogues and wild types were maintained in different geographic regions through different suppliers. This was illustrated in Thomas Laxton wild types from the John Innes Centre which were more closely related to the corresponding rogues from the same collection (ID of 0.1) than to the Thomas Laxton wild types from the Australian Grains Genebank (Figure 4.9). This example showed that mRNA2 would have introduced more noise in analysis, and confirmed that the newer mRNA1 data set was indeed better controlled.

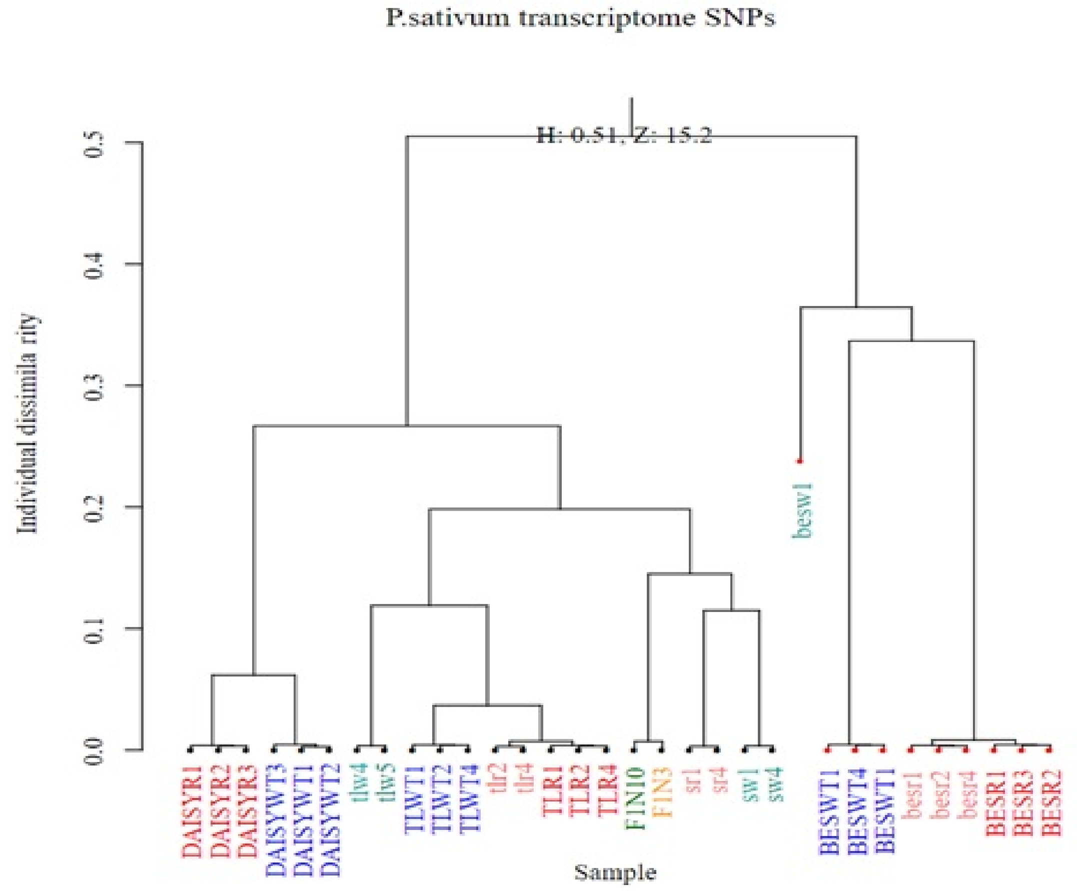


FIGURE 4.9: **Genetic variation between cultivars and epigenotypes.** Clustering of cultivars and epigenotypes from each data set determined by SNPs, and compared through an individual dissimilarity matrix to identify potential genetic variations. Individual dissimilarity (ID) scores range from 0 (all group members genetically identical) to 1 (group members are all genetically unique). Black-Eyed Susan appears as a strong outlier based on SNP content relative to other assessed cultivars from both mRNA1 (uppercase names) and mRNA2 (lowercase names) data sets. R = Rogue, WT = wild type, TL = Thomas Laxton, BES = Black-Eyed Susan, S = Sutton's Hundredfold, F1 = Sutton's Hundredfold wild type X Sutton's Early Giant rogue, N10 = leaflet node 10, N3 = leaflet node 3.

I next analysed genes DE between rogues and wild types, reasoning that one or more of these genes may be candidates for the locus that defines the rogue phenotype. I identified 221 significantly down-regulated genes and 142 up-regulated genes in rogues relative to wild types ($p \leq 0.1$, logFC=1.5) in a combined analysis of all cultivars in the stipule transcriptome (mRNA1) data set (Table A.2). By comparison, I identified only 185 down-regulated and 96 up-regulated genes in rogues when using the previous version of the pea transcriptome annotation. This indicates that the improvements made to the annotation increased the amount of identifiable DE genes in combined cultivar analysis. However, combined DE analysis of all cultivars revealed no obvious candidate gene for the rogue locus. Even among the top-ranked DE genes, up or down-regulation was not consistent between cultivars. One explanation for this is that there are multiple rogue loci, differing from cultivar to cultivar.

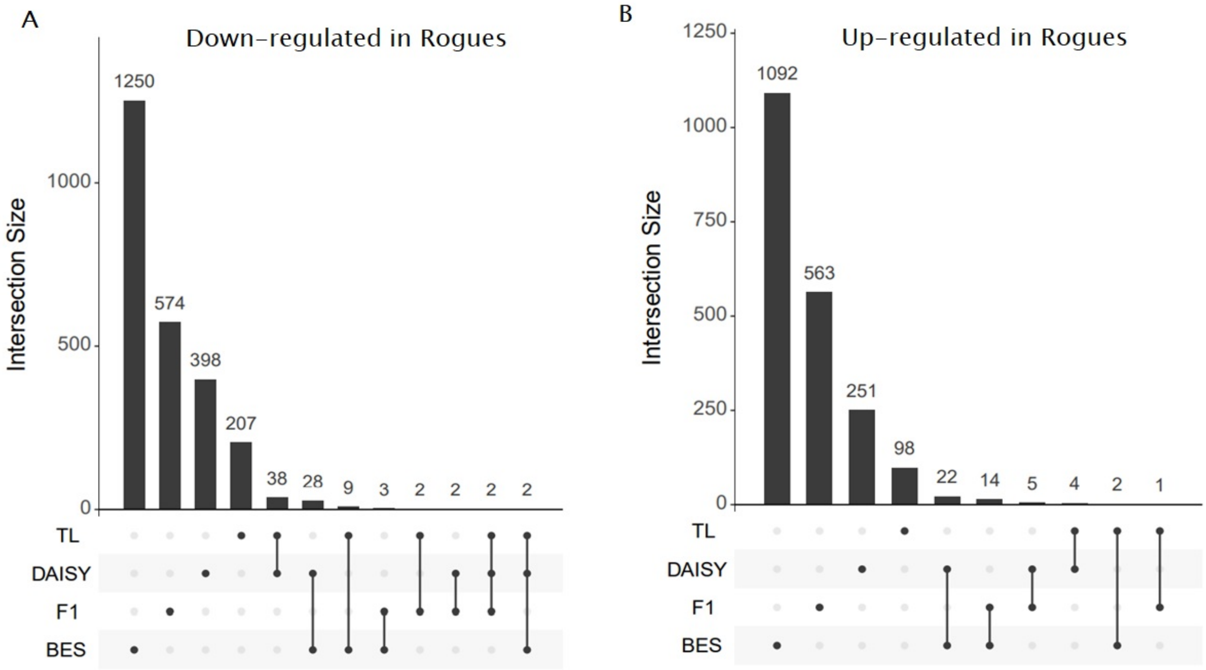


Figure 4.10: **Limited overlaps in rogues vs wild types DE across cultivars.** UpSet plot shows commonality in DE down-regulated (A) and up-regulated (B) genes in rogue peas. Connected lines indicate a common DE gene between the given cultivars. Each category is mutually exclusive. TL = Thomas Laxton, DAISY = Daisy, BES = Black-Eyed Susan, F1 = first generation hybrid cross.

I performed a new DE analysis with each cultivar taken individually, attempting to identify cultivar-specific candidates for the rogue locus/loci. Consistent with the level of separation on the MDS plot and the level of genetic drift between rogues and wild types, the number of DE genes within a cultivar was highest for Black-Eyed Susan, with 1292 significantly down-regulated genes and 1130 significantly up-regulated genes in rogues relative to wild types (Figure 4.10 A, B). Thomas Laxton and Daisy as well as Black-Eyed Susan and Daisy share the largest amount of commonly down-regulated genes in rogues - 42 and 28 genes respectively. A total of 4 genes were commonly down-regulated in rogues across 3 different cultivars and/or the F1 hybrid samples. These genes are consequently strong candidates for the rogue locus. Fewer genes were commonly up-regulated in rogues than were down-regulated, with a total of 48 genes shared between at least 2 cultivars and/or

F1 hybrid samples (Figure 4.10 B). Complete gene lists from individual DE analysis can be viewed in Table A.2. While I was more interested in down-regulated genes, as this fits the paramutation model, it is possible the rogue locus works contrary to expectations and involves an up-regulated locus. Consequently these genes were also considered as candidates.

I then visualised the read alignments at top DE genes to further validate and visualise that each gene was indeed DE between rogues and wild types. At tx.2324, a transcript down-regulated in rogues of both Thomas Laxton and Black-Eyed Susan cultivars that overlaps with the *Psat2g037120* gene from the existing annotation, reads aligned well for Thomas Laxton though some variations were obvious for Black-Eyed Susan (Figure 4.11 A, B). The read coverage differences between rogues and wild types support the DE call for both cultivars - reads increase 5-fold in Thomas Laxton wild types relative to rogues, and increase 3-fold in Black-Eyed Susan wild types (Figure 4.11 A, B). While both of these examples do not display complete silencing in rogue samples, aligned reads are significantly reduced in both cultivars at this example of a strongly DE locus. When inspecting other candidates, similar patterns were observed where some alignments showed variations to the annotation, mostly owing to genetic variations between cultivars and in Black-Eyed Susan in particular. For the most part, DE gene alignments were accurate and supported the DE call for candidate rogue loci.

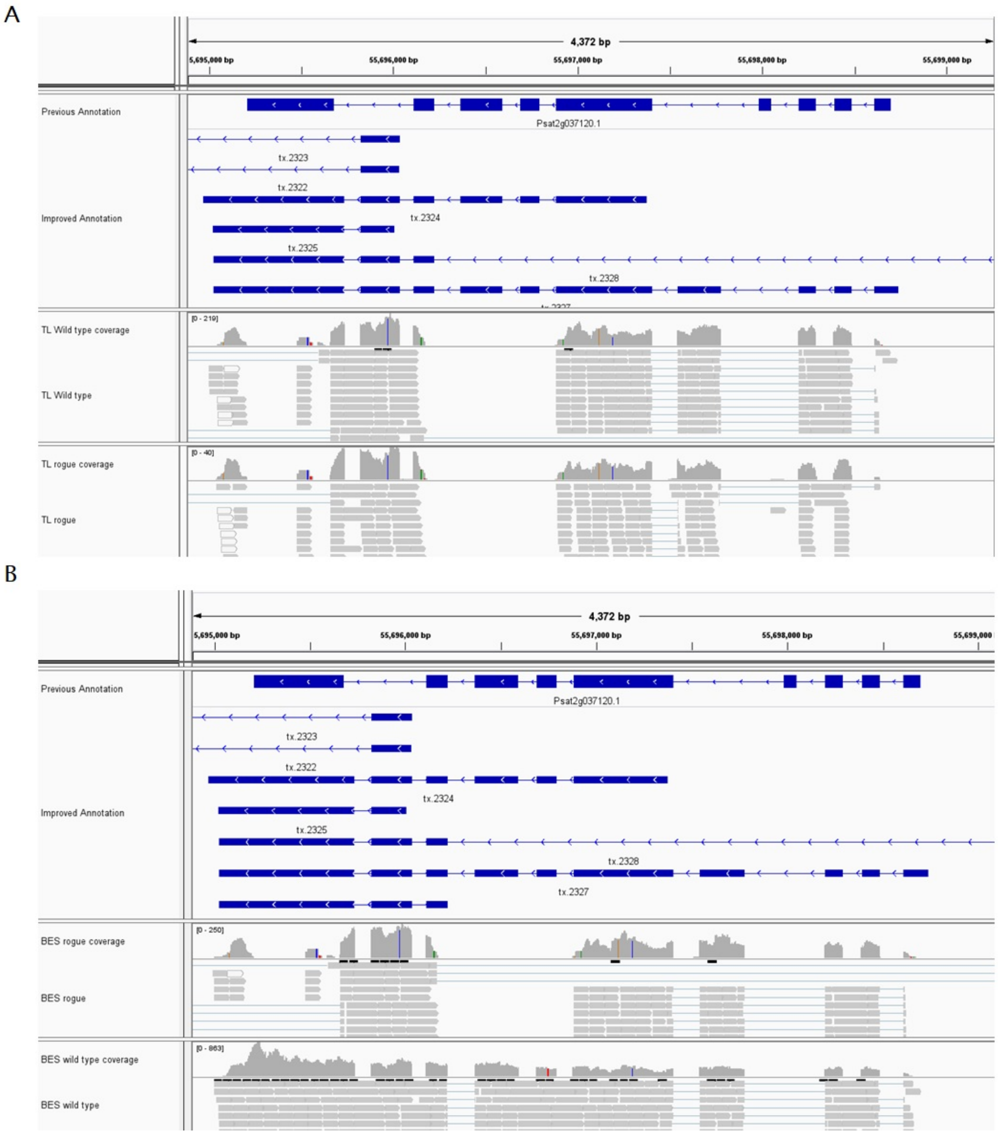


FIGURE 4.11: **Aligned reads support top DE gene.** Read alignments positioned at top DE gene (tx.2324) for Thomas Laxton and Black-Eyed Susan show reduced reads in rogue samples relative to wild type samples, supporting the DE call in analysis. Comparison made in IGV genome browser.

## 4.4 Conclusions

This study provides the first description of transcriptome-wide differences between rogue and wild type peas. The pea genome assembly and transcriptome annotation are still imperfect, and long-read transcriptome sequencing was able to enrich the annotation with revised gene structures and the addition of new transcripts. The increase in assignable features and meta-features improved my ability to call larger numbers of DE genes relative to my preliminary analysis, and increased the likelihood of identifying the rogue locus.

My combined analysis of all cultivars together identified 221 significantly down-regulated genes in rogues relative to wild types, though individual analysis was required to draw comparisons between cultivars and identify genes involved in important plant developmental pathways. Within each individual cultivar there was a large number of DE genes between rogues and wild types, but overlap between cultivars was limited. This is surprising as the similar phenotypic modifications exhibited in rogues from different cultivar suggested that a common pathway would be affected across cultivars. However a small number of common DE genes remains compatible with my hypothesis that the causal locus of the rogue phenotype is conserved across the pea cultivars.

DE and SNP analysis revealed the genetic distance between Black-Eyed Susan rogues and wild types was large, and this cultivar was likely impacted the most by genetic drift. The number of down-regulated DE genes between Black-Eyed Susan rogues and wild types (1292) was much larger by comparison to other assessed cultivars. This likely effect of genetic drift makes identification of the causal rogue locus most difficult in Black-Eyed Susan, due to the resulting genetic background noise. However, some level of drift is expected in all rogue cultivars as the epigenotypes have been kept separate since their discovery. For all cultivars, it was clear that varying levels of genetic drift between rogues and wild types had a confounding effect on gene expression differences, making it more difficult to isolate the differential expression that is due to rogue-specific epigenetic differences.

Prior to my study, no attempt had been made to identify the rogue locus through transcriptomic means aside from targeted expression analysis by Santo et al. (2017) [40] of a small number of candidates. I have conducted the first comparative analysis of the entire rogue transcriptome in multiple cultivars of *P. sativum.* The DE analysis suggested strong candidate genes for the rogue locus, but did not provide definitive evidence on its own. It also brings little information about the mechanisms of rogue paramutation. A feature of the RdDM model of paramutation is the involvement of small-RNAs, in particular 24-nt sRNAs. To provide an orthogonal line of evidence in the identification of strong candidate loci as well as to investigate the mechanisms of paramutation in pea, I next focused on the small RNAomes of wild type and rogue peas.

# Chapter 5

# The small-RNA profile of rogue and wild type peas

## 5.1 Summary and Objectives

In the preceding chapter I built a solid foundation for analysis of the mechanisms driving rogue epimutation by determining for the first time the profile of DE genes between rogue and wild type peas. My next objective was to assess the association of small RNAs (sRNAs) with rogue paramutation by investigating the sRNA profile comparing rogues and wild types, and identify sRNA clusters neighbouring DE genes which may be implicated in transcriptional repression of the nearby locus.

sRNAs are a known constant in the paramutation systems characterised thus far. These are essential molecules involved in transcriptional regulation pathways like RNA-directed DNA Methylation (RdDM). If rogue paramutation in *P. sativum* shares a similar trans-homolog silencing mechanism with other paramutant examples, a comprehensive comparison of the sRNA profile in rogues and wild types could be combined with my differential gene expression analysis to produce a strong case for candidate genes of the rogue locus.

I began my investigation by extracting and sequencing the sRNAs from the stipules of rogues and wild types of three epimutant cultivars (Thomas Laxton, Daisy, Sutton's Hundredfold). I then analysed the data to identify regions of differential sRNA abundance between rogue and wild type peas. From what we understand of paramutation, I expected to find differential sRNA abundance within regulatory regions of the rogue locus, possibly at the promoter, which would let us infer a transcriptionally repressive interaction. However, it was also possible I would discover differential sRNA abundances at enhancers located hundreds of kilobases away from the rogue locus.

I found that 24-nt sRNA clusters were highly abundant in *P. sativum*, and were DE between rogue epimutants and their wild type counterparts within the cultivars I assessed. I confirmed that sRNA clusters were within close proximity to, or overlapped, annotated genes and that a number of these clusters were commonly up-regulated in rogues relative to wild types across multiple cultivars.

## 5.2 Cluster Analysis

I began with combined analysis of all samples in my sRNA-seq data by investigating the clustering and size distributions of the sRNA reads using ShortStack, to ensure I had captured a profile typical of Angiosperms. In general 24-nt sRNAs are the most abundant class of siRNAs, followed by 21-nt and 22-nt sRNAs [61, 62]. I applied ShortStack, which integrated all samples to make a complete map of sRNA clusters throughout the pea genome then classified the most abundant length of sRNAs observed within each cluster. I was interested to see the sRNA diversity within my samples, as the known transcriptional regulatory pathways involved in paramutation require 24-nt or 21-nt sized sRNAs to form and direct transcriptional regulatory complexes. Clusters within my samples were predominantly composed of 24-nt sRNAs, and therefore most clusters were classified as 24-nt by DicerCall (Figure 5.1). This presents a general overview of the dominant sRNA sizes within clusters in my data set, which fit the expected sRNA profile typical of Angiosperms.

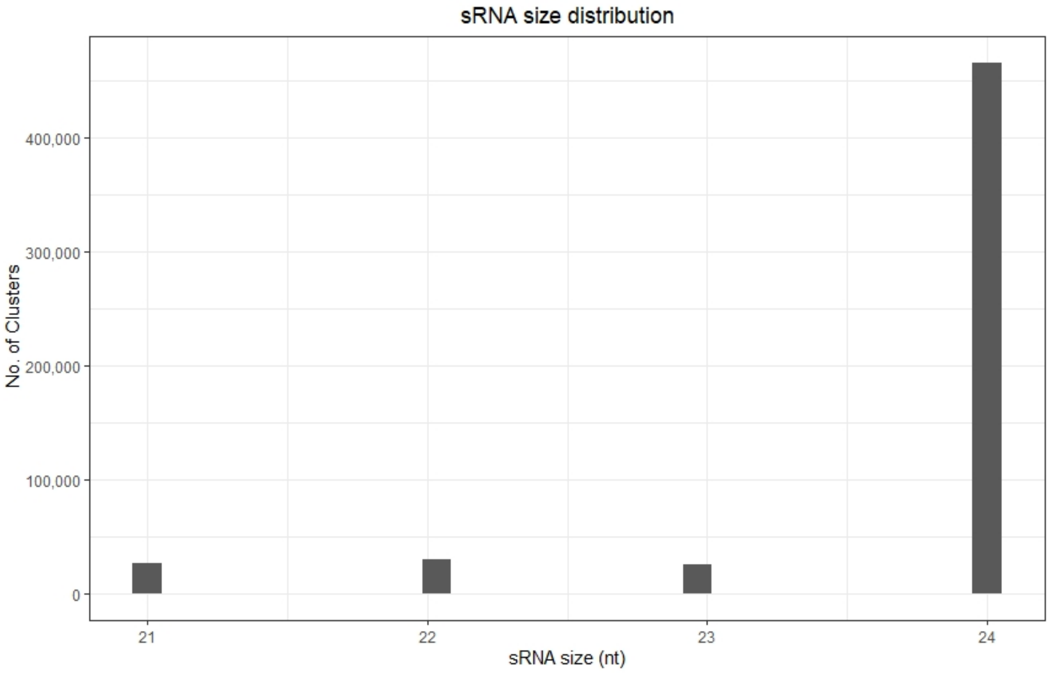


Figure 5.1: **Clusters predominantly composed of 24-nt sRNAs are the most abundant.** A module of ShortStack cluster analysis, DicerCall, summarises the most frequent sRNA length within each analysed cluster. All sample replicates and epigenotypes are used to define sRNA clusters in pea.

I then determined the distribution of sRNA abundance across all annotated clusters, ensuring substantial counts for each size category were present, and to further validate that my libraries were typical of Angiosperms. The distributions of abundance for both 21-nt and 22-nt clusters were similar (Figure 5.2): a large dynamic range was observed from as little as 10 sRNA-seq read counts per cluster to over 1 million. By comparison, the read count distribution for 23-nt clusters was generally lower, although there were again outliers with very high sRNA counts. A comparatively larger proportion of the 24-nt class clusters had intermediate read counts (100–1000). These results confirmed libraries contained well represented sRNAs at each cluster, and were in expected norms for Angiosperms.

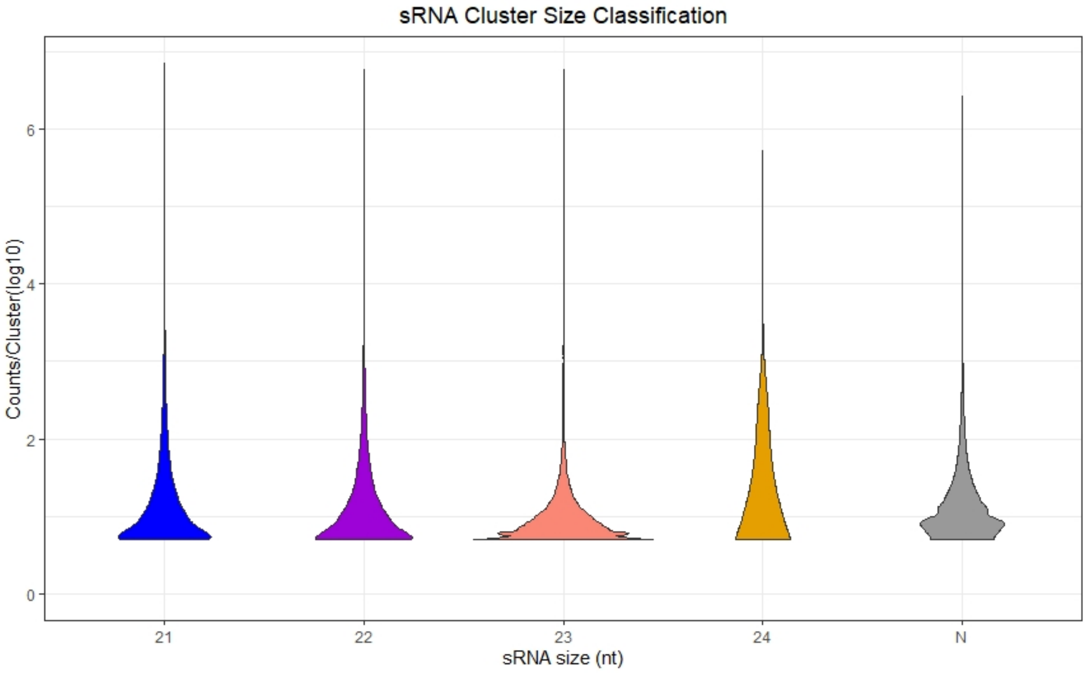


Figure 5.2: **sRNA counts per cluster.** The log base 10 of read counts per cluster plotted against DicerCall sRNA size (nt) classifications. Large dynamic ranges of counts for 21, 22 and 23-nt sRNAs were present, ranging from 10 counts per cluster to over 1 million in some cases. A larger density of read counts per cluster for 24-nt sRNAs was observed relative to other size groups. The category "N" indicates RNAs which are not related to an RNAi process, and are most often fragmented products of abundant RNAs. The counts per cluster of N values is largely reduced relative to sRNAs, where most counts per cluster are $< 10$. This means fragmented products weren't abundant relative to actual sRNAs in my data set.

Next, I wanted to identify any potential global variation in sRNA populations between cultivars, and compare rogues and wild types within individual cultivars that might be associated with the rogue phenotype. Slight variation was observed in the relative proportions of 21-nt and 24-nt sRNAs between individual samples within each cultivar (Figure 5.3). In the Daisy cultivar 24-nt sRNAs were most abundant, with proportions ranging from 25-40% of all reads across all samples (Figure 5.3 A). While there was some variation between rogue and wild type samples, no meaningful observations can be made to distinguish epigenotypes here. Reads for 21-nt sRNAs in Daisy are the second most abundant, ranging from 15-20% for all reads across samples. Daisy wild type samples display slight proportional increases 22-nt sRNAs relative to rogues. The remaining reads were made up of 20, 23 and 25-nt sRNAs, accounting for less than 30% of total reads in this case. The relative sRNA size distributions between Sutton's Hundredfold epigenotypes were different from those of Daisy, with higher 21-nt sRNA read counts in rogues relative to wild types, and higher 24-nt read counts in wild types relative to rogues (Figure 5.3 B). There was no distinction between epigenotypes of Thomas Laxton (Figure 5.3 C). Overall there was no consistent trend between rogues and wild types. This absence of a global shift in sRNA profiles is expected if a single locus, or a small number of loci, is responsible for the rogue phenotype.

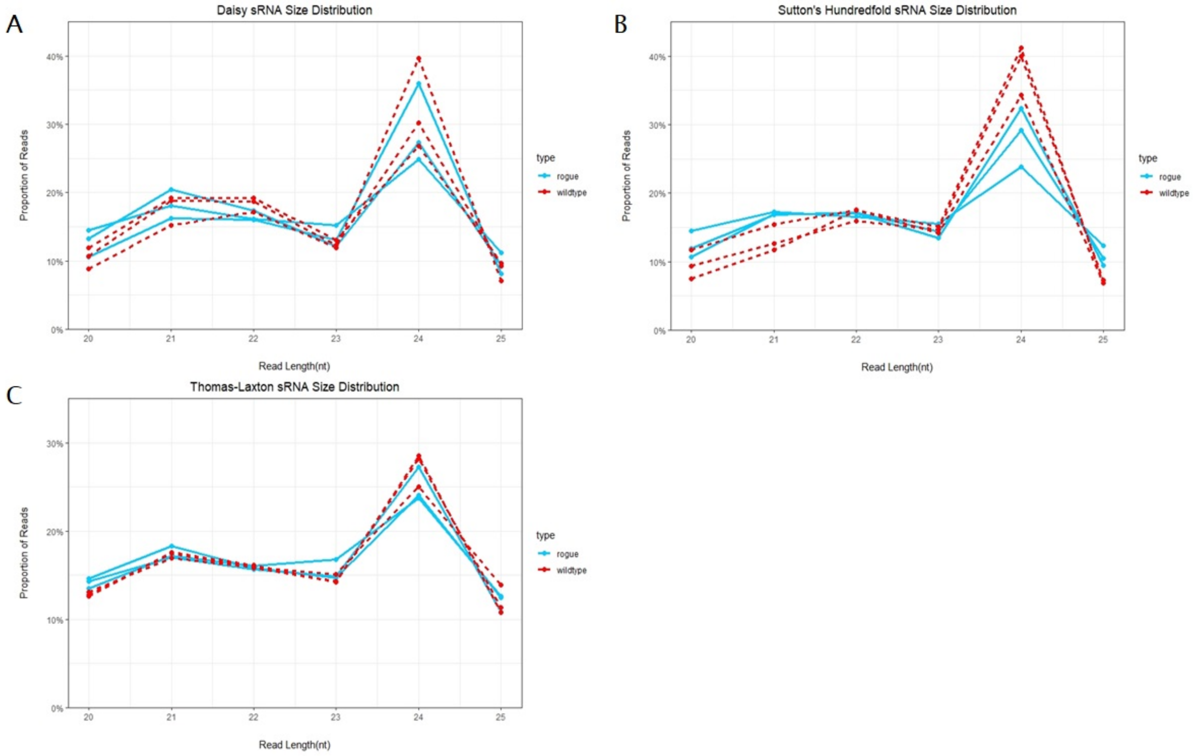


Figure 5.3: **sRNA size distribution in rogues and wild types of each cultivar.** Read size distribution as a proportion of total number of sRNA reads per sample, separated by cultivar (A = Daisy, B = Sutton's Hundredfold, C = Thomas Laxton) and epigenotype (Rogue = Blue, wild type = Red). The size distribution was dominated by 24-nt sRNAs within each cultivar. In most cases the size distribution of rogues did not differ from wild types, though Sutton's Hundredfold rogues displayed slightly elevated levels of 21-nt sRNAs - a siRNA size class involved in post-transcriptional silencing.

I next investigated the proximity of sRNA clusters to nearby genes, to gain a global view of associations between sRNAs and genes in *P. sativum*. Plotting the sRNA cluster density against the log10 base distance either side of the nearest gene identified many sRNA clusters close enough to potentially affect gene expression (Figure 5.4). I found that approximately 90% of sRNA clusters were within 100 kb distance of the nearest gene, and 23% actually overlapped annotated genes. The results coincide with sRNA profiles of other plant species, where typically 24-nt sRNAs are found in close proximity to expressed genes, usually to regulate transcription depending on type of sRNA and associated gene [63]. High levels of sRNAs overlapping genes means clusters can be flagged as potential regulators of the nearby gene, some of which may be candidates for the rogue locus. I also observe a smaller group of clusters ranging 60-100 kb in distance to their nearest gene, which is somewhat expected due to the size of the pea genome, and its complex non-coding repeat regions. *P. sativum* shares a similarly repeat-rich genome to a close relative, *M. truncatula*, where repeat regions are regulated by sRNA-mediated silencing to avoid over-expression and detrimental autoimmune responses [64]. Similar transcriptional regulation likely occurs in *P. sativum*, and can explain the large amounts of sRNA clusters overlapping genes. These results showed that high levels of global gene regulation are likely driven by regulatory sRNAs in *P. sativum*. A finer-scale analysis might reveal that candidate genes for the rogue locus are associated with DE sRNAs regulating their gene expression.

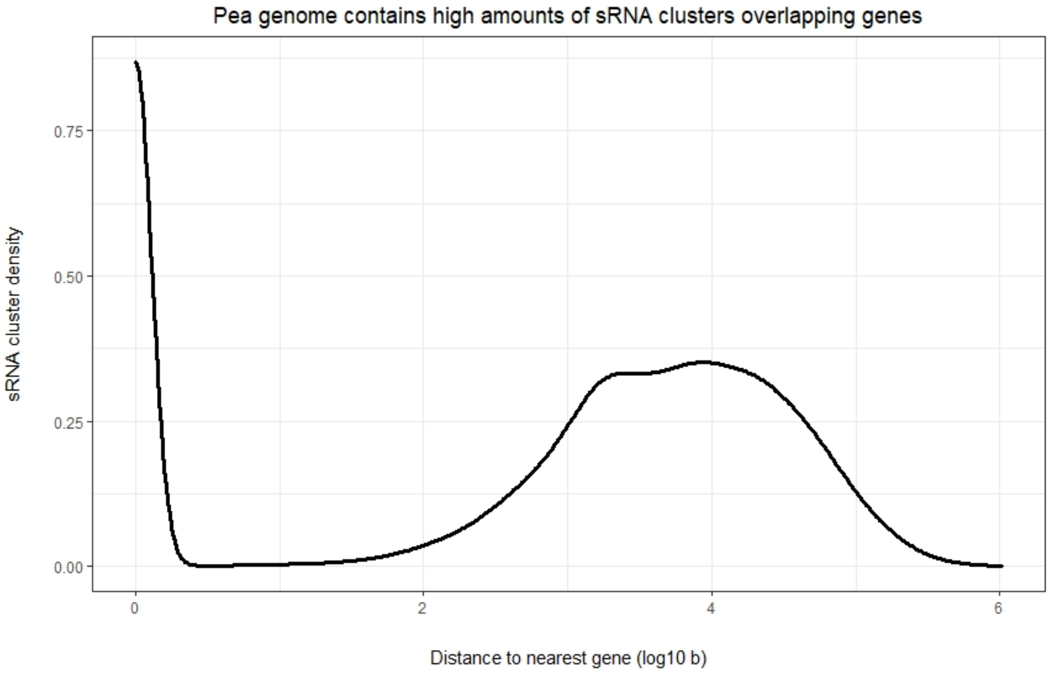


FIGURE 5.4: **sRNA cluster distance to nearest gene.** Distribution of sRNA cluster distances to their nearest gene (up or downstream). Many clusters are within 100 kb of the nearest gene. A smaller group of clusters are observed between 60–100 kb from nearest gene, and most likely are situated between large non-coding or repeat regions of the pea genome.

Having explored the global sRNA populations within my data set, their abundances and proximity to annotated genomic regions, the next step forward was to investigate potential variation in sRNA expression between epigenotypes.

## 5.3 Differentially Expressed sRNA Clusters

Should paramutation in the pea occur through similar mechanisms to other known examples, detecting variation in sRNA expression between epigenotypes would further support the identification of strong candidates for the rogue locus. I continued my sRNA analysis by first plotting combined read counts for all identified clusters, separated by cultivar on an MDS plot. The purpose was to investigate if variation in sRNA expression was present between rogues and wild types. I observed that within each cultivar, rogue and wild type samples separated by sRNA expression, as each epigenotype forms its own unique grouping across the first dimension (Figure 5.5). Dimension 1 accounted for 30–50% of the overall variance within each cultivar. As I established in the short-read mRNA-seq and SNP analysis, both genetic and epigenetic differences likely contribute to the variation in gene expression between rogues and wild types. Therefore, the variation observed in sRNA expression may infer transcriptional regulation through sRNAs varies between rogues and wild types.

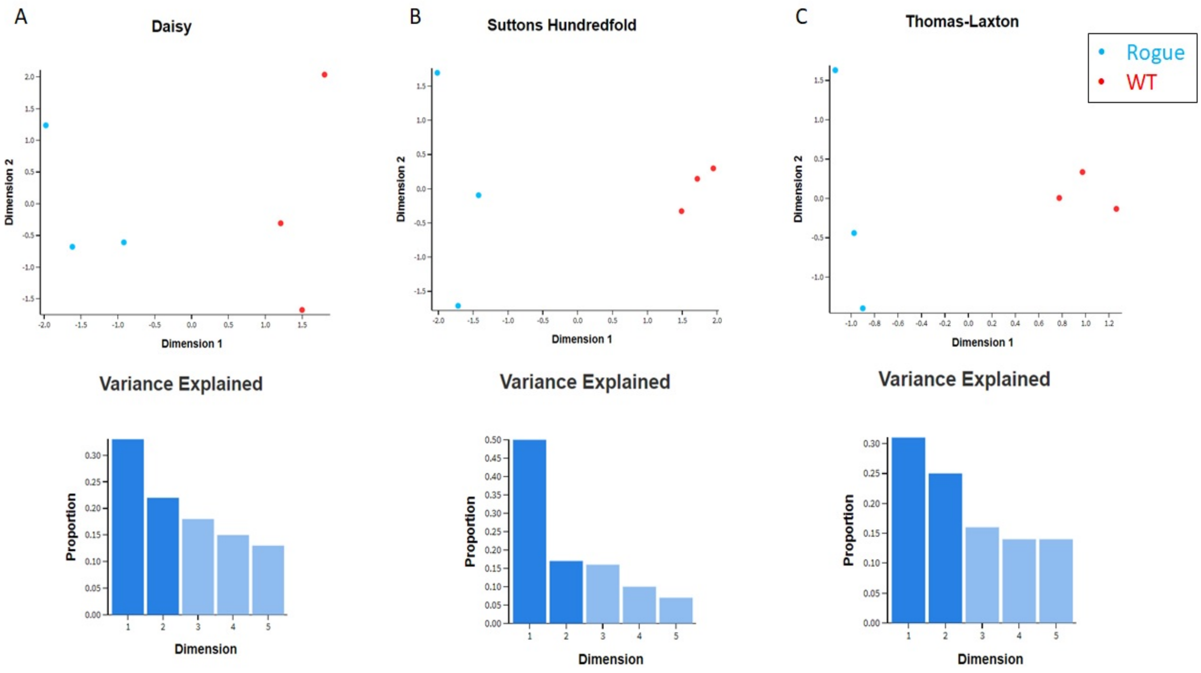


FIGURE 5.5: **Variation in sRNA expression between rogues and wild types.** MDS plots depict variation in global sRNA expression between epigenotypes. Each point represents all sRNA read counts from all clusters in each biologically independent replicate from sRNA-seq analysis (Rogue = blue, wild type = red). All assessed cultivars (Daisy = A, Sutton's Hundredfold = B, Thomas Laxton = C) display variation in sRNA expression between rogue and wild type epigenotypes. Variance explained by dimensional planes 1 & 2 collectively exceed 50% of the overall variance within the read data.

To close in on potential differences in sRNA expression between rogues and wild types associated with the rogue locus, I continued with DE analysis to identify commonly DE sRNA clusters in rogues. I suspect a sRNA locus involved in the rogue phenotype is common across cultivars. I found 1343 up-regulated and 2042 down-regulated sRNA clusters in rogues relative to wild types ($p < 0.1$, logFC=1.5) from a total population of 586,209 sRNA clusters (Table A.2). A much greater number of sRNA clusters were DE between rogues and wild types in Sutton's Hundredfold relative to the other cultivars. I searched for up-regulated DE clusters shared by rogues, as I expect sRNAs affecting gene expression of the rogue locus are up-regulated in rogues relative to wild types. I plotted the top up-regulated sRNA clusters in each cultivar on an UpSet plot (Figure 5.6). There were 223 up-regulated clusters common between at least 2 cultivars, and 2 sRNA clusters common to all three cultivars. As was observed in DE analysis of mRNA data, sRNA expression overall appears unique to individual cultivars. However, the short list of shared DE sRNA clusters are candidates for association with the rogue locus that can be further examined.

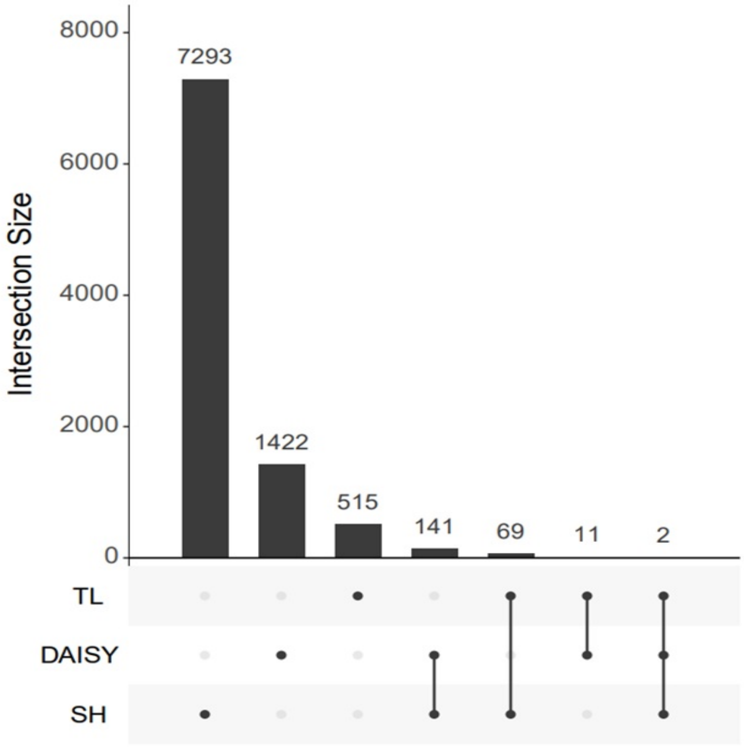


FIGURE 5.6: **Up-regulated sRNA rogue clusters common between cultivars.** UpSet plot displays the commonly up-regulated sRNA clusters between rogues for each cultivar. Daisy and Sutton's Hundredfold share the largest set of up-regulated clusters (141), followed by Thomas Laxton and Sutton's Hundredfold (69 common clusters). Thomas Laxton and Daisy share the least up-regulated sRNA clusters with only 11 identified. Clusters common to all 3 cultivars are also present (2 common sRNA clusters). In respect of individual cultivars, Sutton's Hundredfold has large numbers of unique DE sRNA clusters, with 7505 up-regulated in rogues.

Finally, I investigated the alignment of some interesting DE sRNA clusters to annotated genes and transcripts in the improved reference transcriptome. The purpose was to visualise differential sRNA coverage, and to investigate where clusters are situated in relation to genes or newly identified transcripts. As an example of the differentiation in sRNA coverage between rogues and wild types, Figure 5.7 displays the cluster coverage for the Daisy cultivar in relation to Psat6g175480. The merged reads of Daisy rogue replicates showed an increase in 24-nt sRNA clustering within the promoter region of this gene, relative to wild types which displayed reduced levels of sRNAs in this region (Figure 5.7). Similar results on a larger scale can be observed in Figure 5.8, where all combined reads for each epigenotype within each assessed cultivar have been plotted against a newly discovered transcript (tx. 23431) in the improved annotation. Coverage of 21-nt and 24-nt sRNAs for the rogue epigenotype in Thomas Laxton, Sutton's Hundredfold and Daisy all exceed that of wild types for each respective cultivar (Figure 5.8). These two examples display ideal sRNA cluster profiles in relation to annotated genes and transcripts which may be implicated in regulating transcription in these regions.

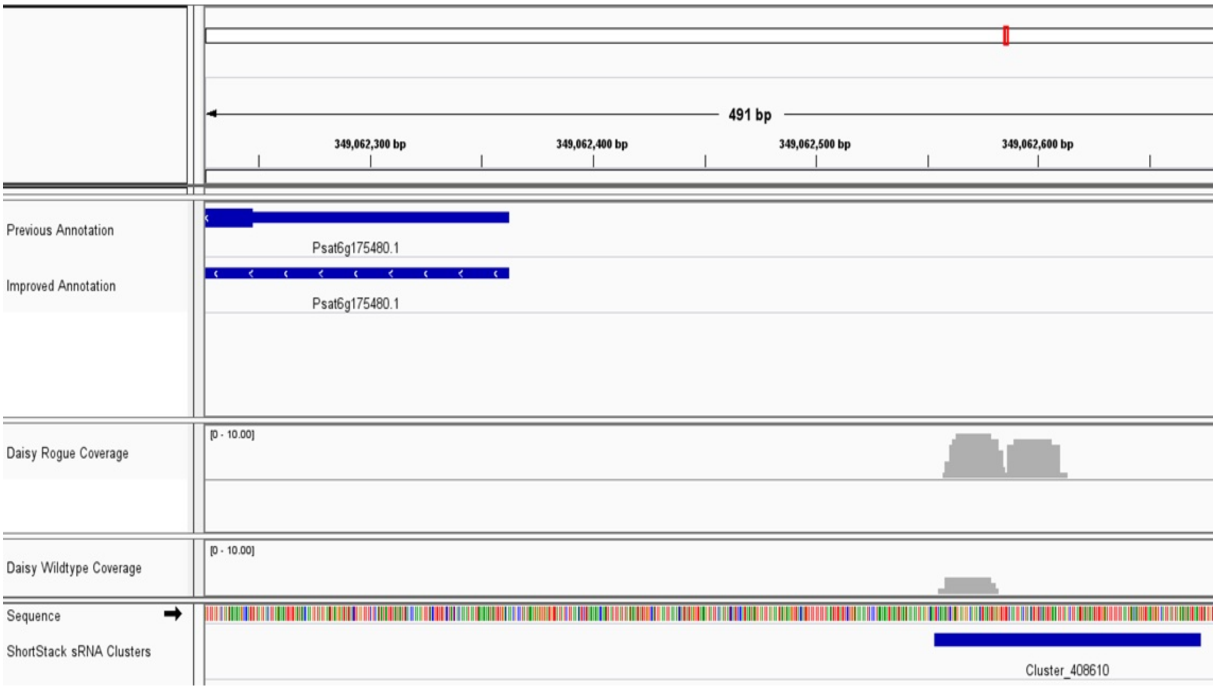


FIGURE 5.7: **sRNA abundances in the promoter region of Psat6g175480.** This 24-nt sRNA cluster maps to the promoter region (denoted by blue arrow orientation on the annotation tracks in IGV of the previous annotation and improved annotation) of candidate gene Psat6g175480. Rogue replicate clusters display higher levels of coverage in the Daisy cultivar relative to wild types.

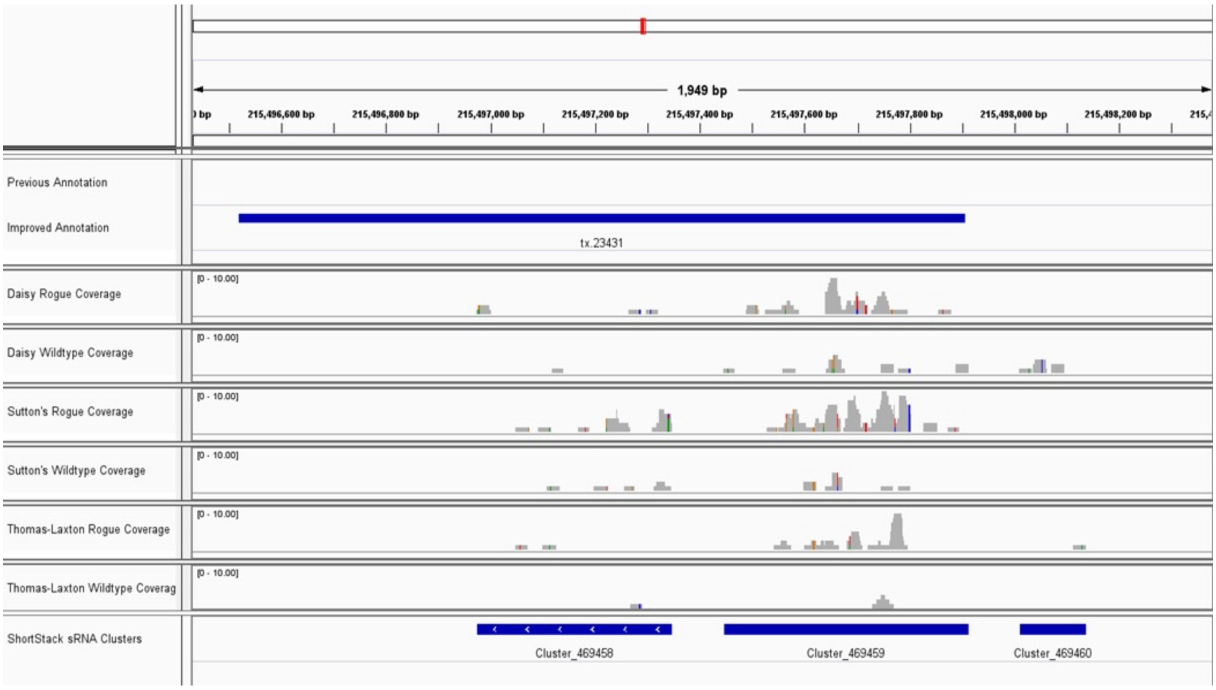


Figure 5.8: **Differential sRNA cluster coverage at transcript tx.23431.** Rogue replicates for all cultivars in this example display increased coverage for two clusters of sRNAs mapping to tx.23431. The reads in this case are dominated by 24-nt sRNAs, though 21-nt sRNAs are also present in smaller counts, indicating a potential site for transcriptional repression through DNA methylation or post-transcriptional silencing.

## 5.4 Conclusions

Here I have presented insights into the sRNA profiles of rogue and wild type peas, obtained through sRNA sequencing and data analysis. The global abundances of 24-nt sRNAs (Figure 5.1 & 5.2) may imply that the pea transcriptome contains a number of loci which are regulated by sRNAs and associated complexes. Rogues and wild types shared similar distributions of sRNA size classes, although specific sRNA clusters accumulated different abundances of sRNAs between epigenotypes. Many clusters were situated in close proximity to, or overlapping, genes. These included examples where sRNA abundances were much higher in rogues than in wild types (Figure 5.7 & 5.8). Considered with the preceding chapters, my study confirms that the pea genome and transcriptome contain many of the constituents required for paramutation; abundant 21-nt and 24-nt sRNAs, DE genes between epigenotypes, and large repeat regions scattered throughout the *P. sativum* genome.

Identifying the key constituents required of paramutation individually is an important step for investigating paramutant-like trans-homolog interactions in pea, though correlation is not causation. However, integrating multiple lines of evidence through the -omics approaches I utilised in this study can narrow down the list of candidates for the rogue loci. I decided to combine findings from all my sequencing analysis with preliminary data sets, to build a matrix which ranks candidate genes based on their fulfilment of selection criteria. The selection process ranks genes which fit my hypothesis surrounding paramutation in the pea, which are discussed in the next chapter.

# Chapter 6

# Integrative ’omics analyses to refine the list of rogue locus candidate genes

## 6.1 Summary and Objectives

In the two preceding chapters I described my investigations into how sRNA and mRNA transcriptomes of rogue and wild type peas differed. My next objective was to integrate these complementary ’omics data sets to build a robust list of candidate genes for the rogue locus. Candidate genes were selected based upon being repressed in rogues relative to wild types. I also included preliminary differential methylation data between a Sutton’s Hundredfold rogue and Caméor wild type generated previously by my supervisor, as DNA methylation is an expected feature of paramutation. Candidate genes were then assigned ranking scores according to several features, including:

1. The number of cultivars in which candidate genes were repressed in rogues relative to wild types.
2. The presence of sRNA clusters within a 10 kb range to the given gene.
3. The presence of DE sRNA clusters between rogues and wild types.

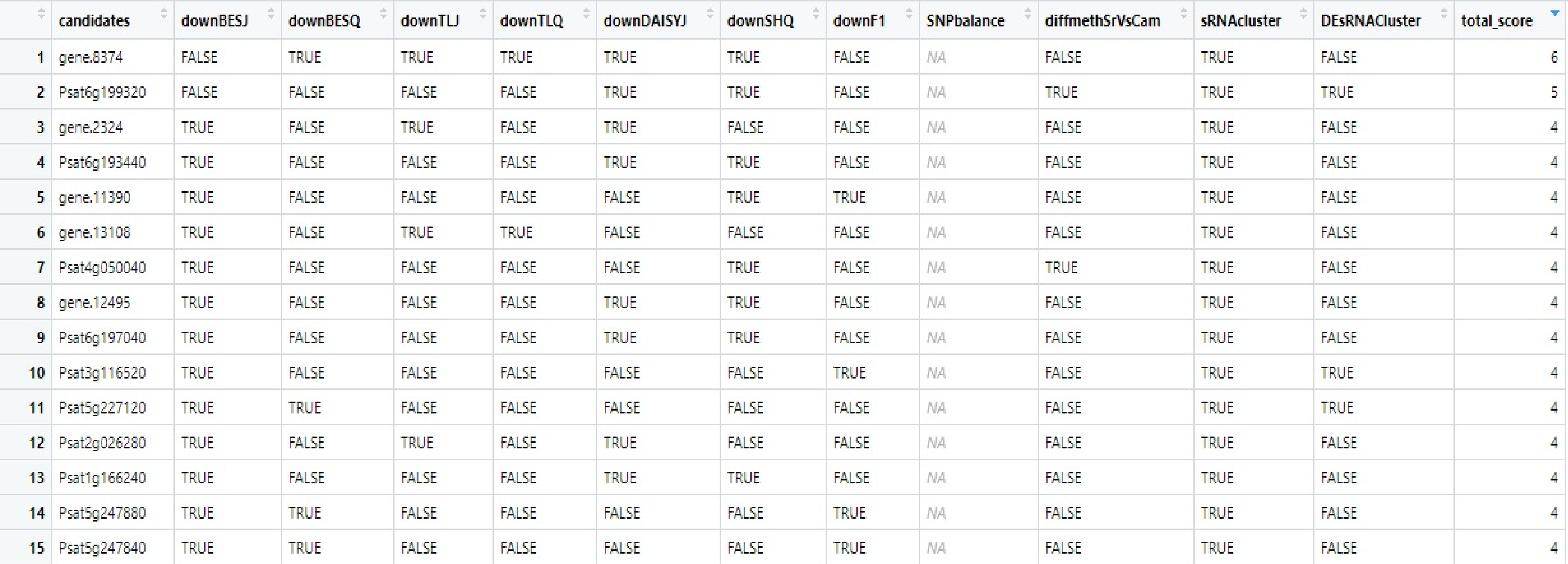
4. The presence of differential methylation at candidate loci from the preliminary methylome data set

Through integration of multiple transcriptomic and methylome datasets, I aimed to identify genes which display conserved epigenetic patterns across rogue epimutants of multiple pea cultivars.

## 6.2 Integration of different 'omics datasets highlights candidate genes for the rogue locus

As previous models of paramutation systems involve consistent mechanisms, such as sRNAs and differentially methylated genomic regions, integration of multiple -omics datasets would improve my ability to identify the rogue locus in *P. sativum.* I designed a systematic ranking system to score candidate genes, where candidates are ranked based upon the number of transcriptomic characteristics they fulfil. The Methods chapter gives a full description of the ranking metric.

Candidates *Psat6g199320* and *gene.8374* ranked highest, with scores of 5 and 6 respectively (Figure 6.1, while another 42 genes received a score $\geq 4$ (Table A.2). I selected a score $\geq 4$ as a reasonable cut-off point to merit further investigation into gene function. Most genes with this ranking were down-regulated in rogues across 2 or more cultivars, and either involved nearby sRNA clusters or potentially differentially methylated regions from the preliminary methlyation data set. Of the candidates with rankings 4 or above, 39% were involved in plant developmental and regulatory pathways, such as auxin transport and cell wall biosynthesis, which could potentially support a role in altering leaf morphology in rogue peas. The newly annotated Gene.8374 received the highest ranking, but I do not present further insights here due to previously mentioned discrepancies in the improved transcriptome annotation requiring further analysis.



| | candidates | downBESJ | downBESQ | downTLJ | downTLQ | downDAISYJ | downSHQ | downF1 | SNPbalance | diffmethSrVsCam | sRNAcluster | DEsRNACluster | total_score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | gene.8374 | FALSE | TRUE | TRUE | TRUE | TRUE | TRUE | FALSE | NA | FALSE | TRUE | FALSE | 6 |
| 2 | Psat6g199320 | FALSE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | NA | TRUE | TRUE | TRUE | 5 |
| 3 | gene.2324 | TRUE | FALSE | TRUE | FALSE | TRUE | FALSE | FALSE | NA | FALSE | TRUE | FALSE | 4 |
| 4 | Psat6g193440 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | NA | FALSE | TRUE | FALSE | 4 |
| 5 | gene.11390 | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | TRUE | NA | FALSE | TRUE | FALSE | 4 |
| 6 | gene.13108 | TRUE | FALSE | TRUE | TRUE | FALSE | FALSE | FALSE | NA | FALSE | TRUE | FALSE | 4 |
| 7 | Psat4g050040 | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | NA | TRUE | TRUE | FALSE | 4 |
| 8 | gene.12495 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | NA | FALSE | TRUE | FALSE | 4 |
| 9 | Psat6g197040 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | NA | FALSE | TRUE | FALSE | 4 |
| 10 | Psat3g116520 | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | NA | FALSE | TRUE | TRUE | 4 |
| 11 | Psat5g227120 | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | NA | FALSE | TRUE | TRUE | 4 |
| 12 | Psat2g026280 | TRUE | FALSE | TRUE | FALSE | TRUE | FALSE | FALSE | NA | FALSE | TRUE | FALSE | 4 |
| 13 | Psat1g166240 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | NA | FALSE | TRUE | FALSE | 4 |
| 14 | Psat5g247880 | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | NA | FALSE | TRUE | FALSE | 4 |
| 15 | Psat5g247840 | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | NA | FALSE | TRUE | FALSE | 4 |

FIGURE 6.1: **Candidate gene matrix.** Top-ranked candidate genes for the rogue locus in *P. sativum*. Fulfilment of each criterion added a value of 1 to the overall score of a candidate gene, which was then used to rank all candidates. Gene 8374 had the highest ranking due to consistent repression of gene expression in rogue epimutants across most cultivars in both mRNA sequencing data sets, while also having sRNA clusters within 10 kb of the TSS. However, discrepancies in the improved annotation must be resolved to validate this new gene as a top candidate.

I selected a subset of top candidate genes to interrogate in greater detail, based on their ranking scores and potential gene function. These were *Psat6g199320*, *Psat5g220240*, *Psat7g141280* and *Psat7g260040*. I have hypothesised that the rogue locus is transcriptionally repressed and this criterion contributed to selection of candidate genes. I next assessed in detail the extent to which the top candidates were repressed to better understand how they differed between rogues and wild types. I plotted the log CPM values of candidate genes (Figure 6.2). Expression of these 4 candidates was substantially reduced in rogues relative to wild types in at least 2 cultivars, or the F1 hybrid. The logFC at each candidate loci averaged $\geq 2$ across cultivars with $p \leq 0.1$ (Table A.2). While expression was not completely silenced at these loci, substantial repression was present in most cultivars. *Psat5g220240* was the most consistently repressed, with expression down-regulated in rogues of all cultivars and F1 hybrid samples for the mRNA1 dataset (Figure 6.2 B).

To further validate the expression profile of the top candidates, I verified that the mRNA1 and 2 short read RNA-seq read alignment at each locus matched the annotated gene structure and was consistent with the DE results. This was valuable as the reference transcriptome is imperfect and requires multiple points of validation. The following alignment examples for candidate genes were selected based on cultivars which had the highest DE between rogues and wild types. *Psat6g99320* was strongly repressed in Daisy rogues relative to wild types. Visualising aligned Daisy reads in the Integrated Genomics Viewer (IGV), I found a substantially greater number of aligned reads was present for wild type replicates in relation to rogue replicates (Figure 6.3 A). Read alignment also supported the alterations made to the structure of this gene in my improved annotation. *Psat5g220240* also met expectations, with large reductions in the number of aligned reads for the Daisy rogue samples relative to wild types (Figure 6.3 B). Results were similar for *Psat7g141280* in Thomas Laxton and Black-Eyed Susan cultivars (Figure 6.3 C, D). While transcription was not completely absent in each example, the difference in expression between epigenotypes was notably large and potentially sufficient to have phenotypic impacts depending upon the gene function of each candidate.

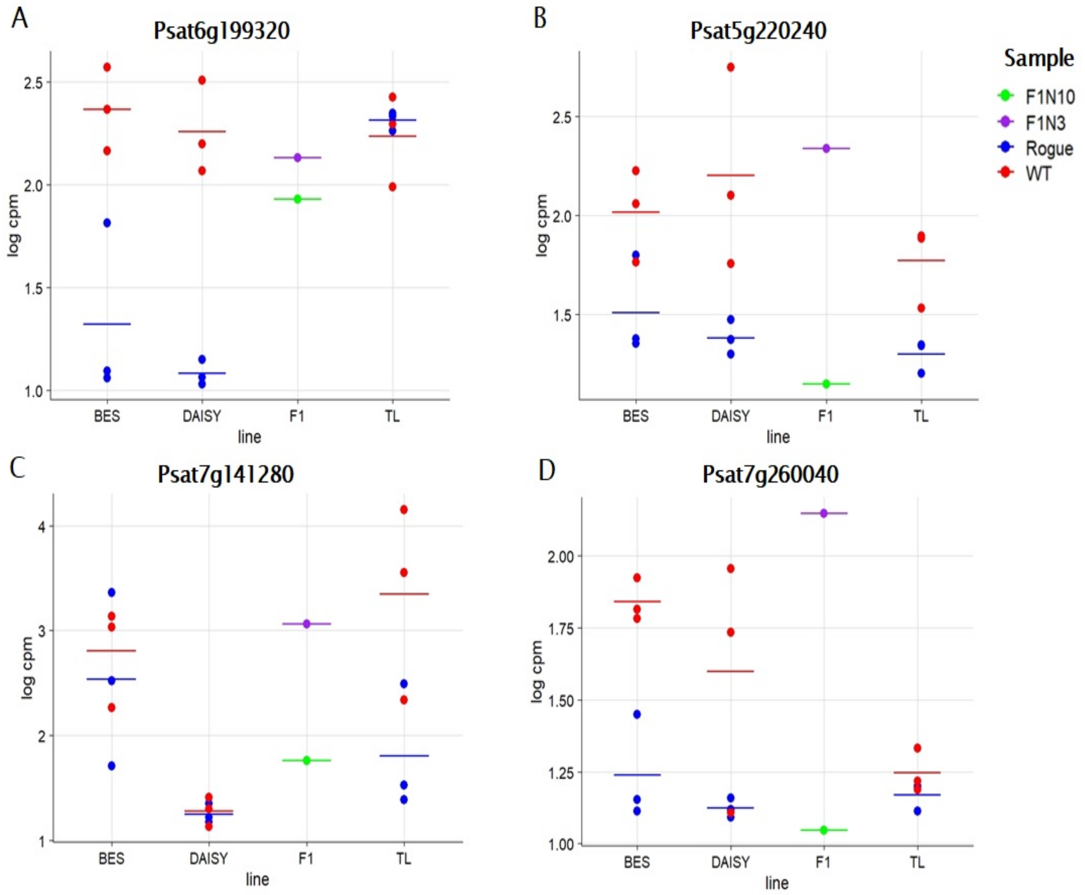


Figure 6.2: **Candidate genes repressed in rogues.** Log CPM read values confirm gene expression differs between epigenotypes for candidate genes *Psat6g199320* (A), *Psat5g220240* (B), *Psat7g141280* (C), *Psat7g260040* (D). Plots A, B and C show differences in expression between at least 2 cultivars or the F1 hybrid 10th node and 3rd node. *Psat5g220240* (B) is more strongly DE, being repressed in rogues of all cultivars as well as the F1 hybrid. Coloured lines on each plot represent mean expression for replicates in each epigenotype.

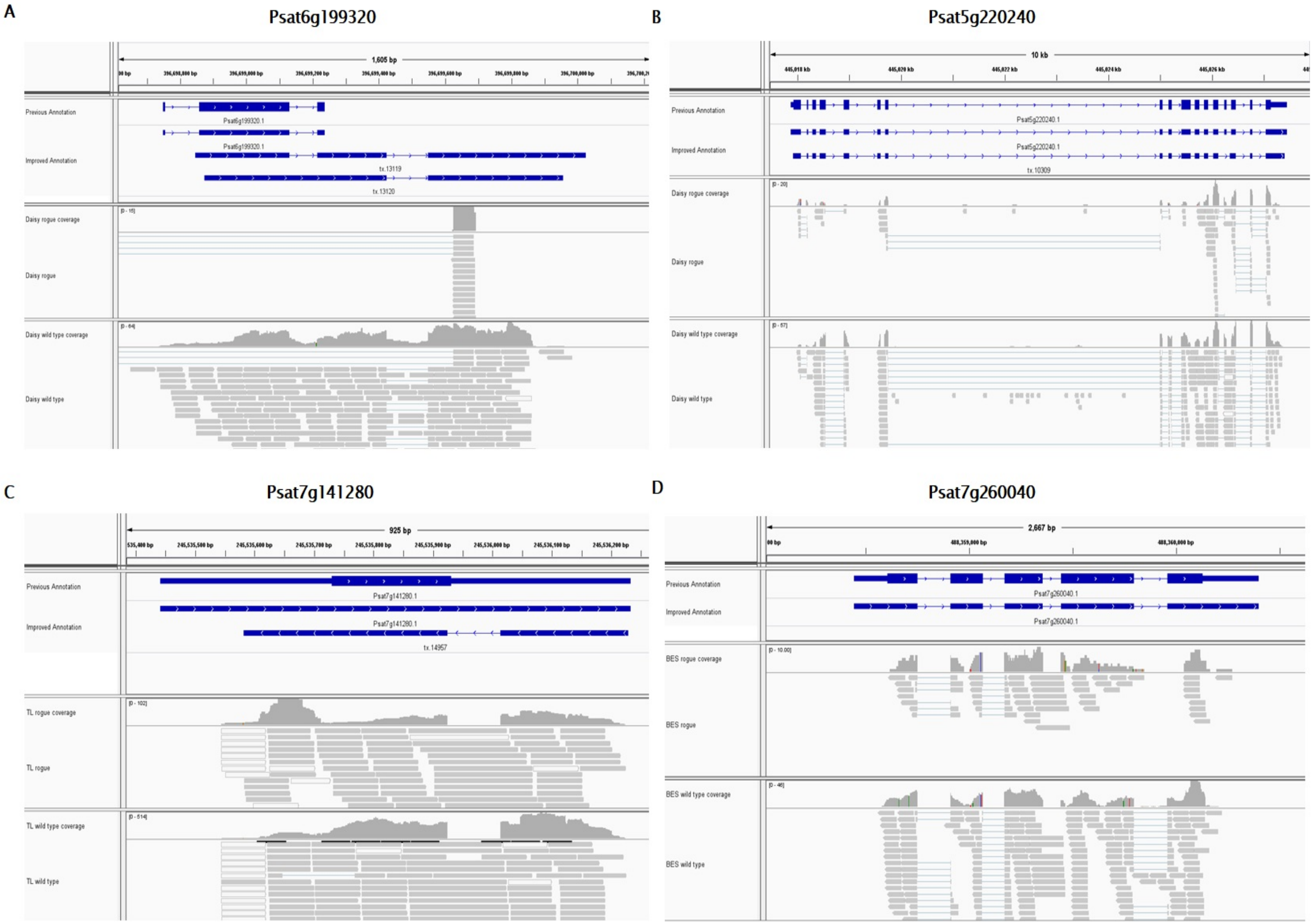


FIGURE 6.3: **Read alignments support top candidate rogue loci.** Read alignments for the cultivar with the largest DE for 4 candidate genes (A = *Psat6g199320*, B = *Psat5g220240*, C = *Psat7g141280*, D = *Psat7g260040*) for the rogue locus using the previous (original) and improved annotation. Grey alignments indicate primary alignment in IGV. Note some read tracks are compressed to better display the amount of reads aligned to the given locus.

I next assessed in detail the sRNA clusters neighbouring my 4 top candidate loci. Most examples of trans-homolog silencing in paramutation involve sRNA clusters proximal to the paramutant locus, which then cause repression of transcription through DNA methylation pathways. I had used the presence/absence of sRNA clusters within 10 kb of my top candidate loci, and their DE status, as a component of my metric to score candidates for the rogue locus. Here I present sRNA cluster coverage of the top 4 candidates in cultivars Thomas Laxton and Daisy alone to match mRNA examples, as Black-Eyed Susan sRNAs were not sequenced. While sRNA clusters aligned within 10 kb either side of my 4 top candidate loci, sRNA read abundance was increased in wild types relative to rogue samples in Daisy and Thomas Laxton cultivars (Figure 6.4). The observation of fewer sRNAs at candidate loci in rogues suggests that short-range effects of sRNAs may not be involved in the repression of these loci. If any of these are indeed the rogue locus then long-range epigenetic mechanisms may instead be involved. Such long-range epigenetic mechanisms are involved in regulation of paramutant maize loci [65].

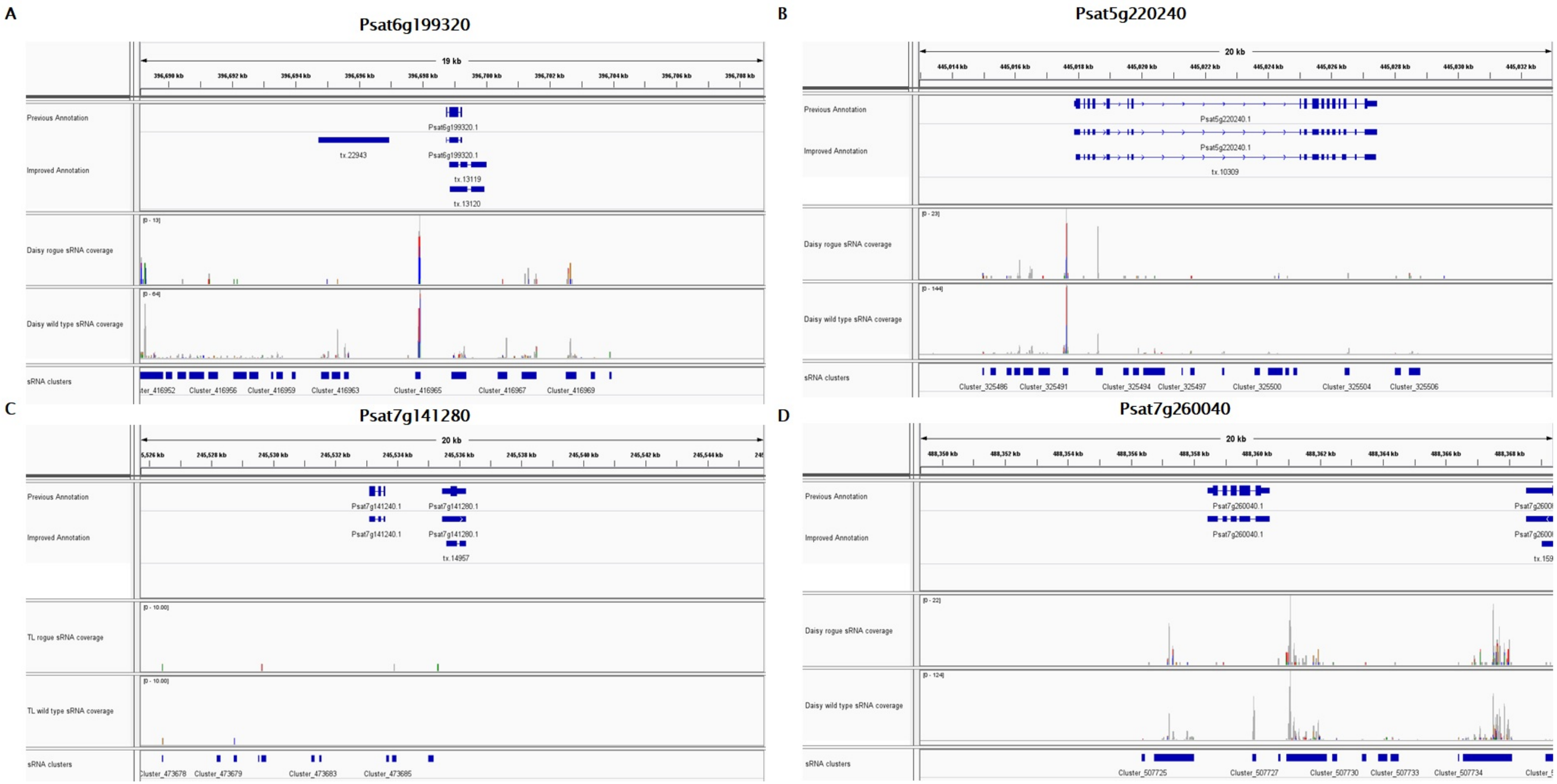
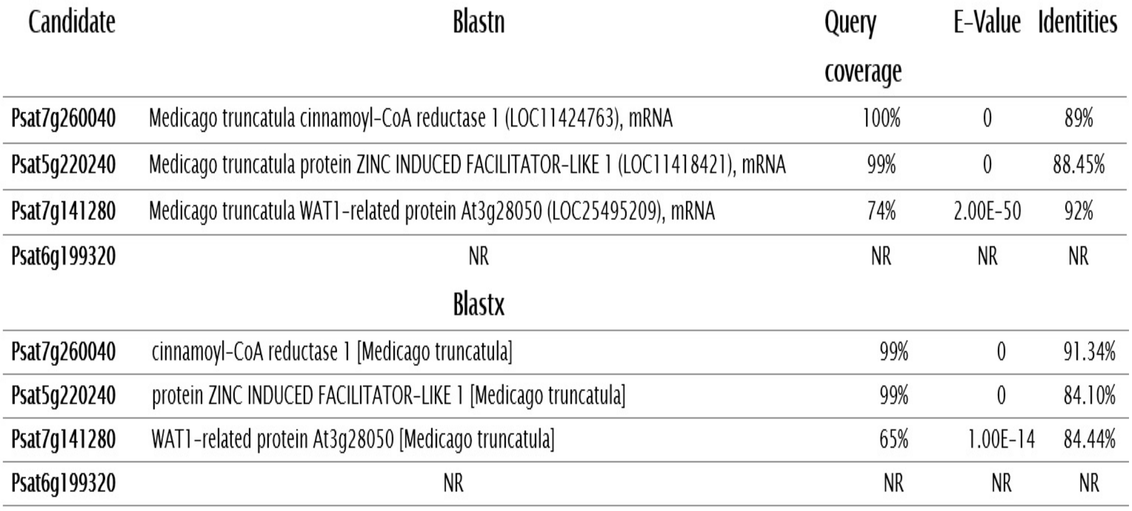


FIGURE 6.4: **sRNA clusters within 10 kb of candidates.** Alignments for sRNA reads proximal to candidates *Psat6g199320* (A), *Psat5g220240* (B), *Psat7g141280* (C), *Psat7g260040* (D) show similarities between epigenotypes for 10 kb range up and downstream of the given candidate loci. Note that for *Psat7g260040* sRNA reads from Daisy are shown instead of Black-Eyed Susan, because the sRNA population was not assessed in Black-Eyed Susan.

Having established the transcriptomic and epigenomic characteristics of my 4 top candidate loci, I investigated their known or predicted gene functions to gain insight into how they might potentially be involved in the rogue paramutation phenotype. I identified sequence similarities to genes in other angiosperms using NCBI's BLASTn and BLASTx queries. Three out of four candidate sequences matched translated proteins in relatives of *P. sativum*, such as *M. truncatula* (Table 6.1). The proteins encoded by these matching genes were involved in plant developmental pathways. *Psat7g260040* had sequence similarity with *M. truncatula*'s Cinnamoyal-CoA reductase 1, whose orthologue in *A. thaliana* functions in the later stages of lignin biosynthesis [66]. The arabidopsis knockdown mutants of this gene had altered leaf morpohology. Similarly, *Psat5g220240* had sequence similarity with a gene involved in plant development in *M. truncatula*. This encoded ZINC INDUCED FACILITATOR-LIKE 1 (ZIFL1), which has a role in transport of the hormone auxin, which is a regulator of development [67]. *Psat7g141280* had high similarity to NODULIN21 in *M. truncatula*, which is homologous to walls-are-thin 1 (WAT1) in *A. thaliana* (Table 6.1). WAT1 is primarily involved in secondary cell wall synthesis, is expressed in leaves, and is involved in the auxin-activated signalling pathway [68]. No genes with sequence similarity were identified for *Psat6g199320*, so at this stage I cannot infer its function.

Table 6.1: **Candidate gene sequence similarity matches known translated proteins.**



| Candidate | Blastn | Query coverage | E-Value | Identities |
|---|---|---|---|---|
| **Psat7g260040** | Medicago truncatula cinnamoyl-CoA reductase 1 (LOC11424763), mRNA | 100% | 0 | 89% |
| **Psat5g220240** | Medicago truncatula protein ZINC INDUCED FACILITATOR-LIKE 1 (LOC11418421), mRNA | 99% | 0 | 88.45% |
| **Psat7g141280** | Medicago truncatula WAT1-related protein At3g28050 (LOC25495209), mRNA | 74% | 2.00E-50 | 92% |
| **Psat6g199320** | NR | NR | NR | NR |
| | **Blastx** | | | |
| **Psat7g260040** | cinnamoyl-CoA reductase 1 [Medicago truncatula] | 99% | 0 | 91.34% |
| **Psat5g220240** | protein ZINC INDUCED FACILITATOR-LIKE 1 [Medicago truncatula] | 99% | 0 | 84.10% |
| **Psat7g141280** | WAT1-related protein At3g28050 [Medicago truncatula] | 65% | 1.00E-14 | 84.44% |
| **Psat6g199320** | NR | NR | NR | NR |

## 6.3 Conclusions

I have conducted an integrated analysis of the transcriptomic and epigenomic characteristics of rogue *P. sativum*, and believe no comparable study currently exists. Through this I have identified strong candidates for the rogue locus. DE mRNAs and sRNA clusters share commonalities across multiple epimutant cultivars. However, contrary to my initial expectations I was unable to identify a single candidate gene down-regulated across all cultivars. A possible explanation is that different cultivars have different rogue loci. Alternatively, the rogue gene may be absent from the current genome assembly or transcriptome annotation, or variation between experiments may have led to confounding DE of the rogue gene. Nevertheless, I successfully identified candidate genes which are DE in multiple cultivars and also show signs of known paramutation mechanisms such as the presence of nearby sRNA clusters and potential differential methylation. Three of four top candidate genes were also found to have sequence homology with genes involved in developmental pathways in other plant species, as were many other candidates. This is consistent with my hypothesis that the locus/loci inducing the rogue phenotype is linked to developmental pathways which can drive alterations in leaf morphology.

Future work should aim to examine the functions of candidate genes through *in vivo* assays, such as targeted genetic knock-down. This may confirm a link between the candidates and the rogue phenotype. Further research to evaluate the mechanisms of the rogue phenotype might include targeted bisulfite sequencing, which would read out the DNA methylation status of the loci. This could confirm differential DNA methylation between rogues and wild types, which was indicated by the preliminary DNA methylation data I used in candidate gene rankings. The limitation of the preliminary DNA methylation data was that it consisted of different cultivars from those in my transcriptomic assays, and only contained 1 biological replicate for each epigenotype. Confidence in genes involved in rogue epimutation would be increased by conducting a study of DNA methylation in matched cultivars with increased replication.

While there is still much work to be done to fully understand paramutation in *P. sativum*, my study has identified rogue candidate genes for the first time, through incorporating multiple -omics techniques on a transcriptome-wide scale.

# Chapter 7

# Discussion

The main objective of this thesis was to identify candidate genes for the rogue locus. The recently published *P. satium* reference genome sequence made a transcriptome-wide scale analysis possible, whereas prior to its release such approaches were difficult and contributed to the lack of progress in understanding rogue paramutation for the past century. I integrated data generated using multiple 'omics techniques in my attempt to identify the causal locus, as observations of paramutation in other species implicate multiple regulatory components. My resulting candidate gene list now provides a basis for further research into gene function and the mechanisms behind paramutation in the rogue pea.

The rate at which phenotypic data can be processed and evaluated is a limiting factor when studying the genetic basis of crop traits [69]. I developed an automated classification model that increased the throughput of phenomics analysis in order to tackle this challenge and assist in characterising the rogue phenotype. The model could accurately and non-arbitrarily classify individuals into their correct epigenotype based on leaf morphometric properties. It currently functions at a cultivar-specific level, though further improvements could potentially enable the model to encompass all cultivars in analysis.

In this chapter, I discuss the implications of my findings about the rogue pea and draw comparisons with well-studied paramutant phenomena in maize and other species. I explore the strengths and shortcomings of my automated classification, and draw conclusions of rogue inheritance in my selected cultivars. Finally, I discuss the implications of my research in a broader context concerning paramutation and epigenetic inheritance, and its potential to positively impact the agricultural industry.

## 7.1 Rogue trait inheritance and classification through machine learning

Phenomic analysis confirmed that rogue leaf morphology is inherited in hybrids between rogue and wild type individuals of cultivars Thomas Laxton and Daisy. Both cultivars displayed the characteristic intermediate rogue phenotype in F1 individuals, as described in the early literature on rogue leaf morphology and inheritance [6, 7]. This phenotype was primarily observed from the $4^{\text{th}}$ node onwards, with lower nodes appearing wild type. There was a complete shift to the rogue phenotype in the F2 generation of the Daisy cultivar, whereas leaf morphometric properties of Thomas Laxton F2s were between mean F1 and rogue values. I also attempted to assess the inheritance of the rogue phenotype in the cultivar Black-Eyed Susan, but an unknown plant pathogen severely distorted development of plants in that experiment so the data were uninterpretable and were not included in this thesis. A repeat experiment on Black-Eyed Susan plants without pathogen infection would be needed to confirm rogue trait inheritance in this cultivar. Overall, rogue leaf morphology was found to be inherited in rogue/wild type hybrid crosses of two cultivars, in agreement with founding studies of rogue peas.

The inheritance of rogue leaf morphology in this study is consistent with examples of paramutant loci in other species [23]. Inheritance of the paramutant loci implicated in anthocyanin biosynthesis in maize is similar, particularly in F1 hybrids which display intermediate phenotypes with incomplete or varied transitions to the full paramutant phenotype observed in subsequent generations [65, 70]. The variation observed in Thomas Laxton F2s leaf morphology is not completely unexpected, as early studies record similar variances in F2 hybrid phenotypes and used categories for ’rogueness’ based on the degree to which leaf morphology traits resembled rogue parents [6].

An explanation for varied hybrid phenotypes I observed is that the rogue locus can exist in multiple regulatory states, and therefore create variations in rogue leaf morphology as a direct result. The occurrence of multiple paramutant states has been observed before in the meta-stable *Pl1-Rh* locus in maize, conferring variation in anther pigmentation [70]. Complete paramutation reversions have been observed at the *Pl1-Rh* locus in maize, where individuals displaying intermediate phenotypes either produce progeny with stronger paramutant traits or completely revert to a wild type like state [70]. Similar reversions can occur at paramutant-like transgene insertions of the maize *b1* locus, which are unstable relative to endogenous *b1* paramutation [65]. My crossing study and phenomics analysis did not identify the occurrence of any phenotype reversions in the F2 generation. However, I did not examine subsequent generations nor the potential for rogue alleles to revert back to a wild type state if rogues were back-crossed to parent lines. A large scale back-crossing study would be ideal to examine reversion in rogue pea cultivars. Such studies were conducted during the initial discovery of the rogue pea phenotype, but examined different cultivars. However, examination of reversion in a broader range of pea cultivars would provide broader, more generalisable understanding of the phenotype.

Plant phenomics was once a limiting factor for biology, but advances in machine learning and automated phenomics are now closing the gap between genotype, epigenotype and phenotype in speed and feasibility of research [71–73]. Developing a quantitative morphometric description of rogue pea cultivars to systematically test the inheritance of the rogue phenotype in hybrid crosses was important, due to the current lack of genetic markers to distinguish between epigenotypes. Using a decision tree learner under the supervised learning model was sufficient to distinguish epigenotypes within my cultivars based on two phenomic properties of l/w ratio and area. The strengths of my automated analysis were its accuracy and precision, ease of use and non-biased method to assign epigenotypes based on phenomic properties. Assessing and characterising the inheritance of rogue phenotype traits across multiple cultivars at a larger scale can now be conducted more quickly and in a non-biased manner using the morphometric and automated classification systems that I developed in this study.

One of the larger limitations I found in my study was the acquisition of phenomic data which was a bottleneck to analysis. This requires improvement to increase throughput for any future large-scale assays. Image scanning and manual segmentation of stipule samples, while accurate, was slow and laborious. Future studies could employ automated 3D imaging techniques to greatly increase the throughput of data acquisition, through the use of multiple camera angles and conveyor belt systems to capture large plant populations [74]. While some of these systems can be costly, the information which can be acquired through 3D imaging could be beneficial to identify other potential components of the rogue phenotype simultaneously. The stipule segmentation process could also be improved to a non-intrusive method through the use of tailored segmentation algorithms which can detect and capture the shape of specific plant organs [75, 76]. While most algorithms have been developed using *A. thaliana*, they could be adapted for other plant species like *P. sativum*. Additionally, throughput could be improved by enabling simultaneous analysis of many cultivars.

My automated analysis could also be enhanced by identifying and using additional phenomic variables unique to the rogue phenotype. The l/w ratio and area of rogue stipules are the most obvious phenotypic change, but other features exist. For example, a rogue phenotype affecting pod morphology was noted in the Gradus cultivar [6]. While it was not feasible to investigate the presence of further phenotypes in my study, future work could do so and then use these for in-depth characterisation and improved phenomics approaches. Should resources permit, a deep-learning approach could be used to identify distinctive phenotypic properties yet to be discovered. Deep-learning approaches have been used in wheat ( *Triticum aestivum L.*) to classify individuals based on shoot and root phenomic properties, with accuracies exceeding 97% [77]. Accurately classifying and confirming inheritance patterns in rogue peas with increased throughput will aid in the future when examining the link between the rogue phenotype and epigenotype in *P. sativum*. Nonetheless, my approaches were effective in developing the first automated system to accurately classify rogue peas based on known phenomic properties.

## 7.2 Rogue candidate genes potentially implicated in developmental pathways

Here I have reported that global transcriptomic differences exist between pea epigenotypes, with hundreds of DE genes between rogues and wild types of cultivars Thomas Laxton, Daisy, Black-Eyed Susan and Sutton's Hundredfold. I found the sensitivity of DE analysis was increased by making improvements to the existing transcriptome annotation for *P. sativum* through long-read sequencing and novel transcript discovery analysis. Some genes down-regulated in rogues were shared between cultivars, making a plausible case for a common locus driving the rogue phenotype, but only 2 genes were commonly down-regulated in rogues of all 3 assessed cultivars. These are strong candidates for the rogue locus but require further analysis and functional validation. Overall, the transcriptomic variation between rogues and wild types was mostly cultivar-specific and could be mainly due to the effects of genetic and epigenetic drift owing to the separation of lines for over 100-years.

This increases the difficulty of identifying the rogue locus, which highlights the necessity of a multi-omics approach to define paramutant characteristics in *P. satiuvm*.

Integrating data from other 'omics techniques, such as sRNA-seq and preliminary DNA methylation data, enabled me to identify candidate genes with higher confidence. sRNAs and DNA methylation are central to paramutation mechanisms across known paramutant angiosperms and other eukarytoic organisms, making them a strong line of investigation. Using these data and a score-based approach I identified a collection of candidate rogue genes belonging to many regulatory pathways involved in plant development. Some of the candidates have orthologues in other species that function in plant development, whilst others are yet to be functionally examined. *Psat6g199320* and *gene.8374* were the top scoring candidates, having neighbouring sRNA clusters, down-regulation in rogue epigenotypes across multiple cultivars and, for *Psat6g199320*, potentially differential methylation. However, functional annotations could not be inferred for these genes when orthology was examined.

The three remaining candidates that I examined were well characterised in *A. thaliana* and have also been identified in *M. truncatula*, a close relative of *P. sativum*. Orthologues of *Psat7g260040* in *A. thaliana* encode cinnamoyl-CoA reductase 1 which has been found to play a major role in secondary cell wall synthesis and cell wall lignification [78, 79]. The resulting phenotypes in knock-down mutants of *A. thaliana* and *M. truncatula* consist mainly of stunted growth and in some cases variations in cell structure and xylem collapse [80, 81]. Preliminary studies by Pyke and Hedley (1984) [44] noted no visually-obvious variations in size or structure of cells of rogue leaves, rather there were fewer cells. However, the cell composition and structure of rogue peas is understudied, so other cellular changes may remain undetected. Another top candidate, *Psat7g141280*, also likely functions in secondary cell wall synthesis, as seen in orthologue WAT-1 [68]. Interestingly, knockdown mutants of this locus induce similar phenotypes as cinnamoyl-CoA reductase 1 - inducing stunted growth and affecting cell elongation [68]. Consequently, the functions of orthologues of *Psat7g2600240* and *Psat7g141280* could be considered consistent with the effects of the rogue phenotype. Their repression in rogue epimutants could potentially also be linked with secondary phenotypes less obvious than the reduced stipules and leaflets of the rogue phenotype. This is an attractive avenue of future study.

Candidate *Psat5g220240* encodes a probable auxin exporter belonging to the Major facilitator superfamily. The protein product of its orthologue in *A. thaliana*, *ZIFL1*, is primarily localised in root and leaf cells [67]. Auxin spatial distribution in leaves is known to impact leaf shape development post-initiation stage, where its distribution is modulated by *PIN1*, an auxin efflux facilitator modulated by ZIFL1 [82, 83]. Therefore, if *Psat5g220240* shares similar functionality to *ZIFL1* in *P. sativum*, it is plausible the inactivation or down-regulation of this locus could drive the leaf shape exhibited by rogue plants.

Detection of rogue locus candidates was made difficult due to the imperfect genome and transcriptome references and, as a result, the causal locus may still remain unidentified. This highlights the need for further improvement of the references through additional deep-sequencing studies and bioinformatic analysis. Considering the transcriptome annotation that I performed, I must refine the application of *Bambu* in the future to ensure all novel transcripts are included, as previously discussed. Future work might also focus on repeat regions, which are highly associated with trans-homolog interactions and are known to impact the degree of paramutagenicity of some loci [17, 65]. The pea genome is repeat-rich and sequencing repeat regions is challenging [84]. Repeat sequences are also often not handled well by bioinformatics packages, resulting in their assembly to collapsed contigs [84]. Long-read and ultra long-read sequencing can rectify many of the issues surrounding assembly of repeat regions, and progress is being made to improve the handling of repeat regions in analysis [85].

Another limitation in my study was the lack of whole-methylome analysis specifically of the cultivars that I used in my other analyses. The preliminary DNA methylation data available to me used different cultivars and only 1 replicate. This limited the utlility of these data. I recommend conducting more extensive DNA methylation analyses in future, since DNA methylation changes are a consistent feature of paramutation.

It is also important to consider that mechanisms driving rogue inheritance in peas may deviate from other paramutant examples. Whilst it might appear improbable, it is possible that the locus is an up-regulated gene or does not require the DNA methylation and RNA-mediated silencing regulatory pathways. In that case sRNA and methylome analysis would not be informative. It is also possible that the inheritance mechanism is cultivar specific as I identified few rogue/wild type common DE genes across cultivars, though I consider the limitations of the reference genome and transcriptome to be more likely causes of this.

There are consequently many considerations for future studies to build on my insights into rogue transcriptomics.

Functional validation of candidate genes is the next step in elucidating the location and mechanism of the rogue locus. Validating rogue candidates with knockdowns or knock-out assays in pea is somewhat challenging. *P. sativum* is amenable to agrobacterium-mediated transformation, though the method is laborious, with very low efficiencies and regeneration rates than compared to model organisms such as *A. thaliana* [86–88]. Virus induced gene silencing (VIGS) is perhaps a more viable option that enables transient knockdown assays; the pea early browning virus (PEBV) has been used as a vector with some success [89, 90]. However, virus-based assays can cause disease symptoms that mask the resulting phenotype from gene knockdown. VIGS has been known to affect leaf morphology depending on the selected virus, which is not ideal for investigating candidate genes for the rogue locus [91]. Targeted mutagenesis using CRISPR is not a well developed method in pea, but can be delivered biolistically [92]. Alternatively, future studies might consider orthogonal approaches making use of model organisms to build evidence for the rogue locus or loci.

## 7.3 *P. sativum* and implications of paramutation in a global context

Identifying the rogue locus and associated inheritance mechanism is important to epigenomics, and how we interpret and understand inheritance. Plant hybrids are highly desired in agriculture, due to their potential to out-perform parental lines in what is referred to as hybrid vigour [93, 94]. Our current understanding of hybrid vigour is primarily genetic: heterozygosity at many loci can compensate for slightly deleterious mutations that each parent possesses. Loss of hybrid vigour in subsequent generations is attributed to the segregation of alleles and loss of heterozygosity [93, 94]. However, epigenetic interactions between alleles also occur in hybrids, with similarities to paramutation and involving similar epigenetic mechanisms [22]. It is still not clear to what extent these epiallele interactions contribute to hybrid vigour or its breakdown. Through studying paramutation in hybrids of epimutant species like maize, and newly arising examples like *P. sativum*, we will better understand the epigenetic component to hybrid vigour. Biotech research could potentially use the fixative nature of paramutant epialleles to ‘lock-in’ desirable traits in crops. The fixation of paramutant states has been confirmed at the *b1* locus, where no observations of reversion for *B’* have been recorded [95]. Alternatively, inhibiting paramutation might increase epigenetic heterozygosity which could benefit hybrid vigour. While manipulation epialelle interactions is a challenging goal, advances in biotechnology paired with deeper understanding of the mechanisms involved may make such applications a reality.

Research into paramutation in *P. sativum* is still in its early days. Modern genomics and epigenomics techniques can now be applied due to the development of quality reference genome and transcriptome sequences. My study demonstrates that new information can be gained about pea traits by application of modern transcriptomics-based approaches. After 100 years of little progress, I have generated the first candidate list of genes for the rogue locus. I drew evidence from the small-RNAome and methylome, and identified potential epigenetic components involved at these loci. My findings represent the foundation for future genetic and epigenetic research in *P. sativum* and will aid in identifying the rogue locus through validation studies, and in revealing the epigenetic components driving the non-Mendelian inheritance of the rogue phenotype.

# Appendix A

# Appendix I

Table A.1: **Comparison of read assignment rates for previous and improved transcriptome annotation**

**Previous transcriptome annotation**

| Sample | Aligned reads | Assigned reads | Proportion Assigned |
|---|---|---|---|
| BESR1 | 23649351 | 15092881 | 63.82 |
| BESR2 | 26663593 | 21661867 | 81.24 |
| BESR3 | 28078262 | 23580558 | 83.98 |
| BESWT1 | 31820837 | 24174139 | 75.97 |
| BESWT2 | 29000023 | 23176891 | 79.92 |
| BESWT3 | 31018982 | 24951341 | 80.44 |
| DAISYR1 | 23261564 | 19228780 | 82.66 |
| DAISYR2 | 23069807 | 18143870 | 78.65 |
| DAISYR3 | 26771014 | 20286605 | 75.78 |
| DAISYWT1 | 29344032 | 23670604 | 80.67 |
| DAISYWT2 | 33024732 | 26183320 | 79.28 |
| DAISYWT3 | 20453757 | 13079884 | 63.95 |
| TLR1 | 29045810 | 23932311 | 82.4 |
| TLR2 | 30974748 | 24540103 | 79.23 |
| TLR3 | 33423330 | 27155744 | 81.25 |
| TLWT1 | 37541638 | 30685663 | 81.74 |
| TLWT2 | 34812884 | 28483497 | 81.82 |
| TLWT3 | 27985475 | 22978390 | 82.11 |
| F1N3 | 28492889 | 23535605 | 82.6 |
| F1N10 | 26018883 | 20512372 | 78.84 |

**Improved transcriptome annotation**

| Sample | Aligned reads | Assigned reads | Proportion Assigned |
|---|---|---|---|
| BESR1 | 23649351 | 19775290 | 83.62 |
| BESR2 | 26663593 | 24314777 | 91.19 |
| BESR3 | 28078262 | 26164982 | 93.19 |
| BESWT1 | 31820837 | 28279664 | 88.87 |
| BESWT2 | 29000023 | 26404846 | 91.05 |
| BESWT3 | 31018982 | 28145602 | 90.74 |
| DAISYR1 | 23261564 | 21587207 | 92.8 |
| DAISYR2 | 23069807 | 20954393 | 90.83 |
| DAISYR3 | 26771014 | 23980372 | 89.58 |
| DAISYWT1 | 29344032 | 26927800 | 91.77 |
| DAISYWT2 | 33024732 | 30121204 | 91.21 |
| DAISYWT3 | 20453757 | 17347435 | 84.81 |
| TLR1 | 29045810 | 26852925 | 92.45 |
| TLR2 | 30974748 | 28223096 | 91.12 |
| TLR3 | 33423330 | 30681937 | 91.8 |
| TLWT1 | 37541638 | 34656630 | 92.32 |
| TLWT2 | 34812884 | 32182782 | 92.45 |
| TLWT3 | 27985475 | 25876942 | 92.47 |
| F1N3 | 28492889 | 26376368 | 92.57 |
| F1N10 | 26018883 | 23679003 | 91.01 |

Table A.2: **DE analysis data sets and candidate gene list**

| **Data** | **Type** | **URL** |
|---|---|---|
| Combined analysis DE gene list | mRNA, upregulated | figshare.com/s/d3d6317cd03ceaf83390 |
| Combined analysis DE gene list | mRNA, downregulated | figshare.com/s/64c4eae2ebc3cab57db9 |
| Daisy DE gene list | mRNA, upregulated | figshare.com/s/002752fa2e8e2632bf72 |
| Daisy DE gene list | mRNA, downregulated | figshare.com/s/996f6c53b209267658b0 |
| Thomas Laxton DE gene list | mRNA, upregulated | figshare.com/s/68ed5f5169e325662eb9 |
| Thomas Laxton DE gene list | mRNA, downregulated | figshare.com/s/29f0b5842c974c568a65 |
| Black-Eyed Susan DE gene list | mRNA, upregulated | figshare.com/s/0028581687c277d48943 |
| Black-Eyed Susan DE gene list | mRNA, downregulated | figshare.com/s/fcd33c62bcc637faa39a |
| F1 hybrid DE gene list | mRNA, upregulated | figshare.com/s/ae4d0e3eddd2211b29bd |
| F1 hybrid DE gene list | mRNA, downregulated | figshare.com/s/47fb1d362e9b39225e7d |
| Combined sRNA cluster DE list | sRNA, upregulated | figshare.com/s/0e5030b0628e9487825f |
| Thomas Laxton sRNA DE list | sRNA, upregulated | figshare.com/s/438f25ca670661006f33 |
| Daisy sRNA DE list | sRNA, upregulated | figshare.com/s/e02c3524e639c3f1c04b |
| Sutton's Hundredfold sRNA DE list | sRNA, upregulated | figshare.com/s/4c7ec067b190dee0edd0 |
| sRNA distance to nearest gene | sRNA, upregulated | figshare.com/s/777bcdaa4367ea703932 |
| Candidate genes | Combined transcriptomics data | figshare.com/s/b6a6d15e28f15489ad27 |

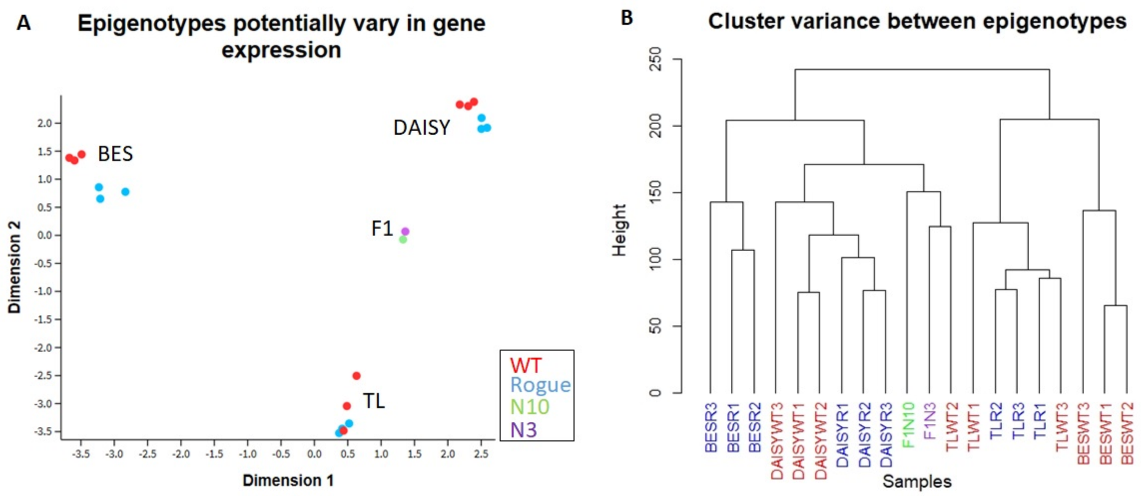


Figure A.1: **Preliminary DE analysis.** MDS plot (A) and Cluster Dendrogram (B) display separation of clusters, suggesting transcriptomic variation between epigenotypes of P. sativum cultivars Black-eyed Susan (BES), Thomas-Laxton (TL), Daisy and F1 hybrid node 3 (N3) and node 10 (N10) samples. Interestingly, a TL rogue sample is shown to cluster with wild-type samples

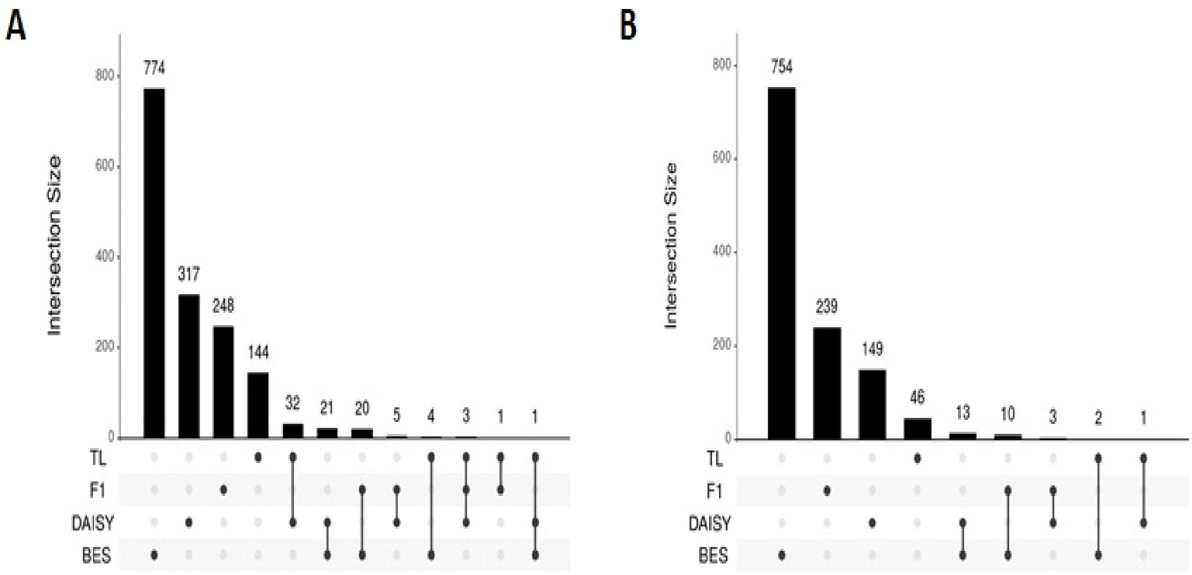


Figure A.2: **Common DE genes identified using previous annotation.** Relative to the improved annotation, DE genes identified using the previous annotation are largely reduced. DE genes are still shared between cultivars, and Black-Eyed Susan displays increased DE genes relative to other cultivars - there are just less by comparison to analysis using the improved annotation

Table A.3: **Daisy phenomics summary stats**

| Daisy | L/W Ratio | | | Area | | |
|---|---|---|---|---|---|---|
| | Mean | Std.dev | Std.error | Mean | Std.dev | Std.error |
| Wild type | 1.63 | 0.20 | 0.04 | 121201.27 | 44463.49 | 8117.89 |
| Rogue | 2.04 | 0.22 | 0.03 | 54828.95 | 19228.01 | 2775.32 |
| F1 Hybrid | 1.87 | 0.27 | 0.04 | 59158.87 | 20387.91 | 2749.10 |
| F2 Hybrid | 1.95 | 0.23 | 0.02 | 43521.39 | 18257.35 | 1447.90 |

Table A.4: **Thomas-Laxton phenomics summary stats**

| Thomas-Laxton | L/W Ratio | | | Area | | |
|---|---|---|---|---|---|---|
| | Mean | Std.dev | Std.error | Mean | Std.dev | Std.error |
| Wild type | 1.71 | 0.11 | 0.02 | 111932.53 | 47327.67 | 6831.16 |
| Rogue | 2.30 | 0.22 | 0.03 | 39590.51 | 15276.52 | 2204.98 |
| F1 Hybrid | 1.99 | 0.27 | 0.03 | 49865.05 | 16915.32 | 1726.41 |
| F2 Hybrid | 2.15 | 0.25 | 0.01 | 41987.03 | 15111.13 | 890.43 |

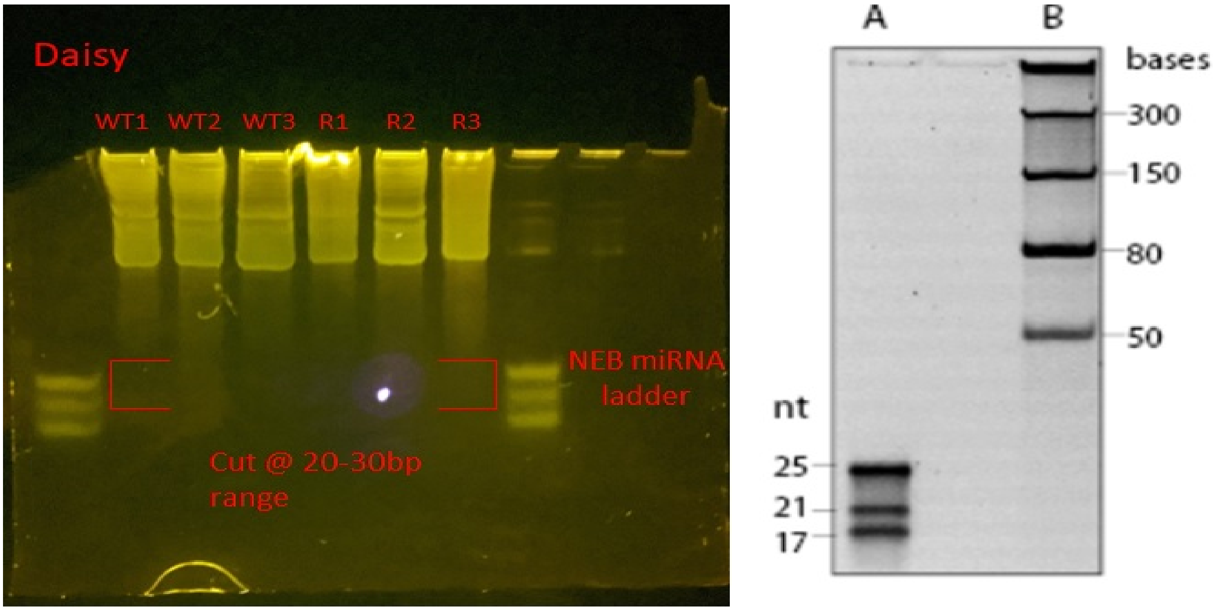


Figure A.3: **sRNA size selection gel.** Denaturing polyacrylamide gel stained with SYBR gold over an UV transilluminator. RNA samples from Daisy total RNA are pictured. Gel bands were excised between 20-30bp marks using the NEB miRNA ladder as reference.

.

# Bibliography

[1] Smýkal P. Pea (pisum sativum l.) in biology prior and after mendel's discovery. *Genet. Plant Breed.*, 50:52–64, 2014.

[2] Mendel G. Experiments in plant hybridization. *Verhandlungen des naturforschenden Vereines in Brünn*, 5:3–47, 1865.

[3] Johzuka-Hisatomi Y., Noguchi H., and Iida S. The molecular basis of incomplete dominance at the a locus of chs-d in the common morning glory, ipomoea purpurea. *J Plant Res*, 124(2):299–304, 2011.

[4] Filatov D. A. and Charlesworth D. Substitution rates in the x- and y-linked genes of the plants, silene latifolia and s. dioica. *Mol Biol Evol*, 19(6):898–907, 2002.

[5] Nazeer W., Ali Z., Ali A., and Hussain T. Genetic behaviour for some polygenic yield contributing traits in wheat (triticum aestivum l.). *Journal of Agricultural Research*, 48(3):267–277, 2010.

[6] Bateson W. and Pellew C. On the genetics of "rogues " among culinary peas (pisum sativum). *Journal of Genetics*, 5:13–27, 1915.

[7] Bateson W. and Pellew C. The genetics of " rogues " among culinary peas (pisum sativum). *Proceedings of the Royal Society of London Series B-Containing Papers of a Biological Character*, 91:186–195, 1920.

[8] Brotherton W. Further studies of the inheritance of "rogue" type in garden peas (pisum sativum l.). *Journal of Agricultural Research*, 24:0815–0852, 1923.

[9] Cesarino I., Dello Ioio R., Kirschner G. K., Ogden M. S., Picard K. L., Rast-Somssich M. I., and Somssich M. Plant science's next top models. *Ann Bot*, 126(1):1–23, 2020.

[10] Macas J., Neumann P., and Navratilova A. Repetitive dna in the pea (pisum sativum l.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and medicago truncatula. *BMC Genomics*, 8:427, 2007.

[11] Kreplak J., Madoui M. A., Capal P., Novak P., Labadie K., Aubert G., Bayer P. E., Gali K. K., Syme R. A., Main D., Klein A., Berard A., Vrbova I., Fournier C., D'Agata L., Belser C., Berrabah W., Toegelova H., Milec Z., Vrana J., Lee H., Kougbeadjo A., Terezol M., Huneau C., Turo C. J., Mohellibi N., Neumann P., Falque M., Gallardo K., McGee R., Tar'an B., Bendahmane A., Aury J. M., Batley J., Le Paslier M. C., Ellis N., Warkentin T. D., Coyne C. J., Salse J., Edwards D., Lichtenzveig J., Macas J., Dolezel J., Wincker P., and Burstin J. A reference genome for pea provides insight into legume genome evolution. *Nat Genet*, 51(9):1411–1422, 2019.

[12] Brink R. A. Paramutation at the r locus in maize. *Cold Spring Harb Symp Quant Biol*, 23:379–91, 1958.

[13] Meyer P., Heidmann I., and Niedenhof I. Differences in dna-methylation are associated with a paramutation phenomenon in transgenic petunia. *Plant J*, 4(1):89–100, 1993.

[14] Blevins T., Podicheti R., Mishra V., Marasco M., Wang J., Rusch D., Tang H., and Pikaard C. S. Identification of pol iv and rdr2-dependent precursors of 24 nt sirnas guiding de novo dna methylation in arabidopsis. *Elife*, 4:e09591, 2015.

[15] Rassoulzadegan M., Grandjean V., Gounon P., Vincent S., Gillot I., and Cuzin F. Rna-mediated non-mendelian inheritance of an epigenetic change in the mouse. *Nature*, 441 (7092):469–74, 2006.

[16] Luteijn M. J., Van Bergeijk P., Kaaij L. J., Almeida M. V., Roovers E. F., Berezikov E., and Ketting R. F. Extremely stable piwi-induced gene silencing in caenorhabditis elegans. *EMBO J*, 31(16):3422–30, 2012.

[17] De Vanssay A., Bouge A. L., Boivin A., Hermant C., Teysset L., Delmarre V., Antoniewski C., and Ronsseray S. Paramutation in drosophila linked to emergence of a pirna-producing locus. *Nature*, 490(7418):112–5, 2012.

[18] Gouil Q., Novak O., and Baulcombe D. C. Sltab2 is the paramutated sulfurea locus in tomato. *J Exp Bot*, 67(9):2655–64, 2016.

[19] Hale C. J., Stonaker J. L., Gross S. M., and Hollick J. B. A novel snf2 protein maintains trans-generational regulatory states established by paramutation in maize. *PLoS Biol*, 5(10):e275, 2007.

[20] Hollick J. B. and Chandler V. L. Epigenetic allelic states of a maize transcriptional regulatory locus exhibit overdominant gene action. *Genetics*, 150(2):891–7, 1998.

[21] Groszmann M., Greaves I. K., Albertyn Z. I., Scofield G. N., Peacock W. J., and Dennis E. S. Changes in 24-nt sirna levels in arabidopsis hybrids suggest an epigenetic contribution to hybrid vigor. *Proc Natl Acad Sci U S A*, 108(6):2617–22, 2011.

[22] Greaves I. K., Groszmann M., Wang A., Peacock W. J., and Dennis E. S. Inheritance of trans chromosomal methylation patterns from arabidopsis f1 hybrids. *Proc Natl Acad Sci U S A*, 111(5):2017–22, 2014.

[23] Hollick J. B. Paramutation and related phenomena in diverse species. *Nat Rev Genet*, 18(1):5–23, 2017.

[24] Alleman M., Sidorenko L., McGinnis K., Seshadri V., Dorweiler J. E., White J., Sikkink K., and Chandler V. L. An rna-dependent rna polymerase is required for paramutation in maize. *Nature*, 442(7100):295–8, 2006.

[25] Jr. Erhard K. F., Stonaker J. L., Parkinson S. E., Lim J. P., Hale C. J., and Hollick J. B. Rna polymerase iv functions in paramutation in zea mays. *Science*, 323(5918): 1201–5, 2009.

[26] Sidorenko L., Dorweiler J. E., Cigan A. M., Arteaga-Vazquez M., Vyas M., Kermicle J., Jurcin D., Brzeski J., Cai Y., and Chandler V. L. A dominant mutation in mediator of paramutation2, one of three second-largest subunits of a plant-specific rna polymerase, disrupts multiple sirna silencing processes. *PLoS Genet*, 5(11):e1000725, 2009.

[27] Haag J. R., Ream T. S., Marasco M., Nicora C. D., Norbeck A. D., Pasa-Tolic L., and Pikaard C. S. In vitro transcription activities of pol iv, pol v, and rdr2 reveal coupling of pol iv and rdr2 for dsrna synthesis in plant rna silencing. *Mol Cell*, 48(5):811–8, 2012.

[28] Matzke M. A., Kanno T., and Matzke A. J. Rna-directed dna methylation: The evolution of a complex epigenetic pathway in flowering plants. *Annu Rev Plant Biol*, 66:243–67, 2015.

[29] Li Q., Gent J. I., Zynda G., Song J., Makarevitch I., Hirsch C. D., Hirsch C. N., Dawe R. K., Madzima T. F., McGinnis K. M., Lisch D., Schmitz R. J., Vaughn M. W., and Springer N. M. Rna-directed dna methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. *Proc Natl Acad Sci U S A*, 112(47):14728–33, 2015.

[30] Zhai J., Bischof S., Wang H., Feng S., Lee T. F., Teng C., Chen X., Park S. Y., Liu L., Gallego-Bartolome J., Liu W., Henderson I. R., Meyers B. C., Ausin I., and Jacobsen S. E. A one precursor one sirna model for pol iv-dependent sirna biogenesis. *Cell*, 163(2): 445–55, 2015.

[31] Kasschau K. D., Fahlgren N., Chapman E. J., Sullivan C. M., Cumbie J. S., Givan S. A., and Carrington J. C. Genome-wide profiling and analysis of arabidopsis sirnas. *PLoS Biol*, 5(3):e57, 2007.

[32] Wierzbicki A. T., Haag J. R., and Pikaard C. S. Noncoding transcription by rna polymerase pol ivb/pol v mediates transcriptional silencing of overlapping and adjacent genes. *Cell*, 135(4):635–48, 2008.

[33] Barbour J. E., Liao I. T., Stonaker J. L., Lim J. P., Lee C. C., Parkinson S. E., Kermicle J., Simon S. A., Meyers B. C., Williams-Carrier R., Barkan A., and Hollick J. B. required to maintain repression2 is a novel protein that facilitates locus-specific paramutation in maize. *Plant Cell*, 24(5):1761–75, 2012.

[34] Hollick J. B., Kermicle J. L., and Parkinson S. E. Rmr6 maintains meiotic inheritance of paramutant states in zea mays. *Genetics*, 171(2):725–40, 2005.

[35] Gent J. I., Madzima T. F., Bader R., Kent M. R., Zhang X., Stam M., McGinnis K. M., and Dawe R. K. Accessible dna and relative depletion of h3k9me2 at maize loci undergoing rna-directed dna methylation. *Plant Cell*, 26(12):4903–17, 2014.

[36] Li Q., Eichten S. R., Hermanson P. J., Zaunbrecher V. M., Song J., Wendt J., Rosenbaum H., Madzima T. F., Sloan A. E., Huang J., Burgess D. L., Richmond T. A., McGinnis K. M., Meeley R. B., Danilevskaya O. N., Vaughn M. W., Kaeppler S. M., Jeddeloh J. A., and Springer N. M. Genetic perturbation of the maize methylome. *Plant Cell*, 26(12):4602–16, 2014.

[37] Ehlert B., Schottler M. A., Tischendorf G., Ludwig-Muller J., and Bock R. The paramutated sulfurea locus of tomato is involved in auxin biosynthesis. *J Exp Bot*, 59(13): 3635–47, 2008.

[38] Ashe A., Sapetschnig A., Weick E. M., Mitchell J., Bagijn M. P., Cording A. C., Doebley A. L., Goldstein L. D., Lehrbach N. J., Le Pen J., Pintacuda G., Sakaguchi A., Sarkies P., Ahmed S., and Miska E. A. pirnas can trigger a multigenerational epigenetic memory in the germline of c. elegans. *Cell*, 150(1):88–99, 2012.

[39] Sapetschnig A., Sarkies P., Lehrbach N. J., and Miska E. A. Tertiary sirnas mediate paramutation in c. elegans. *PLoS Genet*, 11(3):e1005078, 2015.

[40] Santo T., Pereira R., and Leitão J. The pea (pisum sativum l.) rogue paramutation is accompanied by alterations in the methylation pattern of specific genome sequences. *Epigenomes*, 1:1–11, 2017.

[41] Pollard M. O., Gurdasani D., Mentzer A. J., Porter T., and Sandhu M. S. Long reads: their purpose and place. *Hum Mol Genet*, 27(R2):R234–R241, 2018.

[42] Burgess D. J. Genomics: Next regeneration sequencing for reference genomes. *Nat Rev Genet*, 19(3):125, 2018.

[43] Amarasinghe S. L., Su S., Dong X., Zappia L., Ritchie M. E., and Gouil Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol*, 21(1):30, 2020.

[44] Pyke K.A. and Hedley C.L. Aspects of the rogue phenotype of peas. *Plant Sci Letters*, 35:97–90, 1984.

[45] Bonhomme V. amd Picq S., Gaucherel C., and Claude J. Momocs: Outline analysis using r. *Journal of Stat Software*, 56(13):1–24, 2014.

[46] Rohland N. and Reich D. Cost-effective, high-throughput dna sequencing libraries for multiplexed target capture. *Genome Res*, 22(5):939–46, 2012.

[47] Andrews S. Fastqc: A quality control tool for high throughput sequence data. 2010.

[48] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, 17:10–12, 2011.

[49] Kim D., Paggi J. M., Park C., Bennett C., and Salzberg S. L. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. *Nat Biotechnol*, 37(8):907–915, 2019.

[50] Liao Y., Smyth G. K., and Shi W. The r package rsubread is easier, faster, cheaper and better for alignment and quantification of rna sequencing reads. *Nucleic Acids Res*, 47 (8):e47, 2019.

[51] Robinson M.D., McCarthy D.J., and Smyth G.K. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1): 139–140, 2010.

[52] Robinson M. D. and Oshlack A. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biol*, 11(3):R25, 2010.

[53] Fukasawa Y., Ermini L., Wang H., Carty K., and Cheung M. S. Longqc: A quality control tool for third generation sequencing long read data. *G3 (Bethesda)*, 10(4): 1193–1196, 2020.

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

[55] Danecek P., Bonfield J. K., Liddle J., Marshall J., Ohan V., Pollard M. O., Whitwham A., Keane T., McCarthy S. A., Davies R. M., and Li H. Twelve years of samtools and bcftools. *Gigascience*, 10(2), 2021.

[56] Axtell M. J. Shortstack: comprehensive annotation and quantification of small rna genes. *RNA*, 19(6):740–51, 2013.

[57] Langmead B. Aligning short sequencing reads with bowtie. *Curr Protoc Bioinformatics*, Chapter 11:Unit 11 7, 2010.

[58] Lee S., Cook D., and Lawrence M. plyranges: a grammar of genomic data transformation. *Genome Biol*, 20(1):4, 2019.

[59] Simpson J. T., Workman R. E., Zuzarte P. C., David M., Dursi L. J., and Timp W. Detecting dna cytosine methylation using nanopore sequencing. *Nat Methods*, 14(4): 407–410, 2017.

[60] Schaarschmidt S., Fischer A., Zuther E., and Hincha D. K. Evaluation of seven different rna-seq alignment tools based on experimental data from the model plant arabidopsis thaliana. *Int J Mol Sci*, 21(5), 2020.

[61] Rajagopalan R., Vaucheret H., Trejo J., and Bartel D. P. A diverse and evolutionarily fluid set of micrornas in arabidopsis thaliana. *Genes Dev*, 20(24):3407–25, 2006.

[62] Kallman T., Chen J., Gyllenstrand N., and Lagercrantz U. A significant fraction of 21-nucleotide small rna originates from phased degradation of resistance genes in several perennial species. *Plant Physiol*, 162(2):741–54, 2013.

[63] Lunardon A., Johnson N. R., Hagerott E., Phifer T., Polydore S., Coruh C., and Axtell M. J. Integrated annotations and analyses of small rna-producing loci from 47 diverse plants. *Genome Res*, 30(3):497–513, 2020.

[64] Zhai J., Jeong D. H., De Paoli E., Park S., Rosen B. D., Li Y., Gonzalez A. J., Yan Z., Kitto S. L., Grusak M. A., Jackson S. A., Stacey G., Cook D. R., Green P. J., Sherrier D. J., and Meyers B. C. Micrornas as master regulators of the plant nb-lrr defense gene family via the production of phased, trans-acting sirnas. *Genes Dev*, 25 (23):2540–53, 2011.

[65] Belele C. L., Sidorenko L., Stam M., Bader R., Arteaga-Vazquez M. A., and Chandler V. L. Specific tandem repeats are sufficient for paramutation-induced trans-generational silencing. *PLoS Genet*, 9(10):e1003773, 2013.

[66] Jones L., Ennos A. R., and Turner S. R. Cloning and characterization of irregular xylem4 (irx4): a severely lignin-deficient mutant of arabidopsis. *Plant J*, 26(2):205–16, 2001.

[67] Remy E., Cabrito T. R., Baster P., Batista R. A., Teixeira M. C., Friml J., Sa-Correia I., and Duque P. A major facilitator superfamily transporter plays a dual role in polar auxin transport and drought stress tolerance in arabidopsis. *Plant Cell*, 25(3):901–26, 2013.

[68] Ranocha P., Denance N., Vanholme R., Freydier A., Martinez Y., Hoffmann L., Kohler L., Pouzet C., Renou J. P., Sundberg B., Boerjan W., and Goffner D. Walls are thin

1 (wat1), an arabidopsis homolog of medicago truncatula nodulin21, is a tonoplast-localized protein required for secondary wall formation in fibers. *Plant J*, 63(3):469–83, 2010.

[69] Furbank R. T. and Tester M. Phenomics–technologies to relieve the phenotyping bottleneck. *Trends Plant Sci*, 16(12):635–44, 2011.

[70] Hollick J. B., Patterson G. I., Jr. Coe E. H., Cone K. C., and Chandler V. L. Allelic interactions heritably alter the activity of a metastable maize pl allele. *Genetics*, 141 (2):709–19, 1995.

[71] Houle D., Govindaraju D. R., and Omholt S. Phenomics: the next challenge. *Nat Rev Genet*, 11(12):855–66, 2010.

[72] Grosskinsky D. K., Svensgaard J., Christensen S., and Roitsch T. Plant phenomics and the need for physiological phenotyping across scales to narrow the genotype-to-phenotype knowledge gap. *J Exp Bot*, 66(18):5429–40, 2015.

[73] Ubbens J. R. and Stavness I. Deep plant phenomics: A deep learning platform for complex plant phenotyping tasks. *Front Plant Sci*, 8:1190, 2017.

[74] Pasala R. and Pandey B. B. Plant phenomics: High-throughput technology for accelerating genomics. *J Biosci*, 45, 2020.

[75] Lozano-Claros D., Meng X., Custovic E., Deng G., Berkowitz O., Whelan J., and Lewsey M. G. Developmental normalization of phenomics data generated by high throughput plant phenotyping systems. *Plant Methods*, 16:111, 2020.

[76] Dobrescu A., Giuffrida M. V., and Tsaftaris S. A. Doing more with less: A multitask deep learning approach in plant phenotyping. *Front Plant Sci*, 11:141, 2020.

[77] Pound M. P., Atkinson J. A., Townsend A. J., Wilson M. H., Griffiths M., Jackson A. S., Bulat A., Tzimiropoulos G., Wells D. M., Murchie E. H., Pridmore T. P., and French A. P. Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. *Gigascience*, 6(10):1–10, 2017.

[78] Mir Derikvand M., Sierra J. B., Ruel K., Pollet B., Do C. T., Thevenin J., Buffard D., Jouanin L., and Lapierre C. Redirection of the phenylpropanoid pathway to feruloyl malate in arabidopsis mutants deficient for cinnamoyl-coa reductase 1. *Planta*, 227(5): 943–56, 2008.

[79] Ruel K., Berrio-Sierra J., Derikvand M. M., Pollet B., Thevenin J., Lapierre C., Jouanin L., and Joseleau J. P. Impact of ccr1 silencing on the assembly of lignified secondary walls in arabidopsis thaliana. *New Phytol*, 184(1):99–113, 2009.

[80] Patten A. M., Cardenas C. L., Cochrane F. C., Laskar D. D., Bedgar D. L., Davin L. B., and Lewis N. G. Reassessment of effects on lignification and vascular development in the irx4 arabidopsis mutant. *Phytochemistry*, 66(17):2092–107, 2005.

[81] Zhou R., Jackson L., Shadle G., Nakashima J., Temple S., Chen F., and Dixon R. A. Distinct cinnamoyl coa reductases involved in parallel routes to lignin in medicago truncatula. *Proc Natl Acad Sci U S A*, 107(41):17803–8, 2010.

[82] Scarpella E., Barkoulas M., and Tsiantis M. Control of leaf and vein development by auxin. *Cold Spring Harb Perspect Biol*, 2(1):a001511, 2010.

[83] Xiong Y. and Jiao Y. The diverse roles of auxin in regulating leaf development. *Plants (Basel)*, 8(7), 2019.

[84] De Bustos A., Cuadrado A., and Jouve N. Sequencing of long stretches of repetitive dna. *Sci Rep*, 6:36665, 2016.

[85] Treangen T. J. and Salzberg S. L. Repetitive dna and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*, 13(1):36–46, 2011.

[86] Puonti-Kaerlas J., Stabel P., and Eriksson T. Transformation of pea (pisum sativum l.) byagrobacterium tumefaciens. *Plant Cell Rep*, 8(6):321–4, 1989.

[87] Nadolska-Orczyk A. and Orczyk W. Study of the factors influencing agrobacterium-mediated transformation of pea (pisum sativum l.). *Molecular Breeding*, 6:185–194, 2000.

[88] Švábová L., Smýkal P., Griga M., and Ondřej V. Agrobacterium-mediated transformation of pisum sativum in vitro and in vivo. *Biologia Plantarum*, 49(3):361–370, 2005.

[89] Constantin G. D., Krath B. N., MacFarlane S. A., Nicolaisen M., Johansen I. E., and Lund O. S. Virus-induced gene silencing as a tool for functional genomics in a legume species. *Plant J*, 40(4):622–31, 2004.

[90] Constantin G. D., Gronlund M., Johansen I. E., Stougaard J., and Lund O. S. Virus-induced gene silencing (vigs) as a reverse genetic tool to study development of symbiotic root nodules. *Mol Plant Microbe Interact*, 21(6):720–7, 2008.

[91] Burch-Smith T. M., Anderson J. C., Martin G. B., and Dinesh-Kumar S. P. Applications and advantages of virus-induced gene silencing for gene function studies in plants. *Plant J*, 39(5):734–46, 2004.

[92] Hamada H., Liu Y., Nagira Y., Miki R., Taoka N., and Imai R. Biolistic-delivery-based transient crispr/cas9 expression enables in planta genome editing in wheat. *Sci Rep*, 8 (1):14422, 2018.

[93] Groszmann M., Greaves I. K., Albert N., Fujimoto R., Helliwell C. A., Dennis E. S., and Peacock W. J. Epigenetics in plants-vernalisation and hybrid vigour. *Biochim Biophys Acta*, 1809(8):427–37, 2011.

[94] Groszmann M., Greaves I. K., Fujimoto R., Peacock W. J., and Dennis E. S. The role of epigenetics in hybrid vigour. *Trends Genet*, 29(12):684–90, 2013.

[95] Coe E. H. The properties, origin, and mechanism of conversion-type inheritance at the b locus in maize. *Genetics*, 53(6):1035–63, 1966.