USING *FRAGARIA* AS A MODEL SYSTEM FOR THE STUDY OF SUBGENOME DOMINANCE AND ADAPTATION IN CROPS

By

Elizabeth Alger

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Plant Breeding, Genetics and Biotechnology - Horticulture – Doctor of Philosophy

2021

ABSTRACT

USING *FRAGARIA* AS A MODEL SYSTEM FOR THE STUDY OF SUBGENOME DOMINANCE AND ADAPTATION IN CROPS

By

Elizabeth Alger

Polyploidy, or the presence of three or more complete genomes in a single organism, has occurred frequently in plants, especially in the angiosperm lineage. Allopolyploids, or polyploids resulting from the merging of different genomes in an interspecific hybrid, have often been shown to experience subgenome dominance. Subgenome dominance is the phenomenon where there is bias in the gene loss and expression between the different genomes in a polyploid, known as subgenomes. Despite the prevalence of polyploids and subgenome dominance, little is known about the factors and mechanisms that influence this process. Strawberry (*Fragaria sp.*) is emerging as a powerful model system to investigate polyploid subgenome dominance evolution due to the recent identification of the four extant diploid progenitor species of the cultivated octoploid strawberry (*Fragaria* x *ananassa*). Having the diploid progenitors in hand allows us to identify differences between the dominant subgenome, *F. vesca*, and the other three progenitors that may have an impact of subgenome dominance. One possible factor is transposable element (TE) abundance, as low TE density has been consistently associated with the dominant subgenome in allopolyploids. Epigenetic silencing of TEs by DNA methylation to suppress TE activity has been shown to result in decreased expression of neighboring genes and this lowered gene expression may affect the establishment of subgenome dominance. *F. vesca* will be used as a diploid model for the study of subgenome dominance in strawberry where I can examine how TE abundance and other factors influence gene expression in a single accession and in hybrid crosses between different accessions. Tracking changes in gene expression in the

hybrids will allow us to examine how genomes with difference sizes and genomic factors interact. The results and insights observed from this study can then be applied to subgenome dominance research in octoploid strawberry. In addition to the germplasm and genomic resources, strawberries are also a high value crop and the loss of their production due to (a)biotic stressors results in the loss of millions of United States dollars annually. Using a population of octoploid strawberries segregating for salt tolerance, I will identify candidate genes related to salt tolerance. Together this work will identify factors and mechanisms related to subgenome dominance and use genotypic data in a practical breeding context.

ACKNOWLEDGEMENTS

I would like to start by giving a special thank you to my PhD advisor, Dr. Patrick Edger, for his support, encouragement, patience, and kindness throughout my degree. His mentoring kept me enthusiastic about my research even in my final year when I think all graduate students feel the research fatigue. Furthermore, during the uncertain and turbulent times of the COVID-19 pandemic, he prioritized the safety and well-being of our lab members while doing all he could to ensure our research could move forward. I could not imagine having a better mentor.

I want to thank my family for being so supportive throughout my education, from preschool through PhD. I never doubted my ability to earn my doctorate thanks to your encouragement and motivation. Mom, I want to thank you for telling me to "go for the B" whenever I stressed myself too much during high school. I know I never listened, but you'll never know how much pressure that took off me. Dad, I want to thank you for all our long talks about being open to the unforeseen and unfamiliar paths of life. Earning a Ph.D. in horticulture is not something my younger self ever would have predicted, but I am so grateful I took some unexpected opportunities and ended up here. And Chase, you're the best brother I could have asked for, and I will treasure that speech you gave about me for the rest of my life. It is on every electronic I own and Google Drive, for good measure.

I want to thank my partner Mauricio for surprise snacks and flowers, for cooking meals and composing music for me, and for nights out and quiet nights in. You were there for the most stressful moments of my degree and to help me celebrate every accomplishment. You have always been a source of comfort and strength for me, and I feel so lucky to have you in my life.

Finally, I would like to thank my dogs, Azlan and Elena, who reminded me to get outside for a walk every day under the threat of annoyance and destruction. An adorable source of unconditional love is something every graduate student should have.

TABLE OF CONTENTS

LIST OF TABLES vii

LIST OF FIGURES viii

CHAPTER 1 1
One subgenome to rule them all: underlying mechanisms of subgenome dominance 1
Abstract 2

CHAPTER 2 3
Chromosome-Scale Genome for a Red-Fruited, Perpetual Flowering and Runnerless Woodland Strawberry (*Fragaria vesca*) 3
Abstract 4

CHAPTER 3 5
Methylation and transposable elements shape the transcriptional landscape in woodland strawberry 5
Abstract 6
Introduction 6
Results and Discussion 8
Do methylated TEs impact expression of neighboring genes? 8
How does gene expression compare between two *F. vesca* accessions? 14
Are there differences between biased and unbiased genes? 19
Summary 26
Materials & Methods 27
TE abundance, methylation, and expression in 'Hawaii 4' 27
Comparisons between *F. vesca* accessions 28
REFERENCES 29

CHAPTER 4 33
Beach strawberry (*Fragaria chiloensis*) genome provides insights into high salinity tolerance 33
Abstract 34
Introduction 34
Results 37
'Del Norte' genome assembly and annotation 37
'Camarosa' and 'Del Norte' genetic maps 41
QTL mapping for salinity tolerance 42
Genomic prediction and heritability estimation 44
RNAseq analysis and differential expression analysis 44
Discussion 47
'Del Norte' genome assembly and annotation 47
QTL mapping for salinity tolerance 48

Genomic prediction and heritability estimation 49
RNAseq analysis and differential expression analysis 49
Materials and Methods 51
PacBio sequencing 51
Illumina sequencing for the assembly and annotation of 'Del Norte' genome 52
PacBio assembly 53
'Del Norte' assembly correction and annotation 54
'Camarosa' x 'Del Norte' F1 population development and genotyping 55
'Camarosa' and 'Del Norte' genetic map construction 56
Screen progeny of 'Camarosa' x 'Del Norte' F1 for salinity tolerance 57
QTL mapping for salt tolerance 58
Genomic prediction and heritability estimation 59
RNA-seq and differential expression analysis 59
REFERENCES 61

CHAPTER 5 69
Concluding Remarks 69
Expanding genomic resources for the *Fragaria sp.* model system 70
Subgenome dominance in F1 hybrids 70
Improvement of agronomic traits in strawberry 71
REFERENCES 73

LIST OF TABLES

Table 3.1 Summarized transposable element (TE) composition for the *F. vesca* accessions 10

Table 3.2 Genome annotation statistics for *F. vesca* accessions 17

Table 4.1 Statistics for the final genome assembly for the 'Camarosa' and 'Del Norte' genomes 38

Table 4.2 Gene features statistics and BUSCO scores for the 'Camarosa' and 'Del Norte' genome annotations 39

Table 4.3 Summarized transposable element (TE) composition for the *F. chiloensis* 'Del Norte' genome 39

Table 4.4 Gene count and syntelogs for 'Camarosa' and 'Del Norte' 41

Table 4.5 Metrics for the 'Camarosa' and 'Del Norte' genetics maps.....42

Table 4.6 Gene markers for the QTL peaks 44

Table 4.7 'Camarosa' top 10 GO biological processes 45

Table 4.8 'Del Norte' top 10 GO biological processes 46

LIST OF FIGURES

Figure 3.1 Average TE density of all ‘Hawaii 4’ genes as a function of window size and location 9

Figure 3.2 The average gene expression levels at different TE density levels 11

Figure 3.3 Average methylation level for each TE density window 13

Figure 3.4 Average expression at different gene body methylation levels 14

Figure 3.5 TE density, methylation, and expression model 16

Figure 3.6 *In Silico* Hybrid 18

Figure 3.7 2339 x 562 and 562 x 2339 hybrids 19

Figure 3.8 Upstream TE density for biased and unbiased genes 20

Figure 3.9 Expression of 2339 biased genes and 562 biased genes 21

Figure 3.10 Methylation averages upstream, downstream, and within genes 23

Figure 3.11 Expression-Methylation comparison plot 25

Figure 4.1 ‘Del Norte’ collection site 37

Figure 4.2 Synteny between *F.* × *ananassa* cv. ‘Camarosa’ and *F. chiloenesis* ‘Del Norte’ 40

Figure 4.3 QTL graphs 43

Figure 4.4 Phenotypes of individuals segregating in the F1 mapping population 58

CHAPTER 1

One subgenome to rule them all: underlying mechanisms of subgenome dominance

The work presented in this chapter is part of the final publication:

Alger EI, Edger PP. 2020. One subgenome to rule them all: underlying mechanisms of subgenome dominance. *Curr Opin Plant Biol* 54: 108-113.

Abstract

Allopolyploids, which are formed from the hybridization of two or more diploid progenitor species, often experience subgenome dominance, where one of the parental genomes (subgenomes) has higher levels of gene expression and ultimately greater gene retention compared to the other subgenomes. Low transposable element (TE) abundance near genes has been associated with the dominant subgenome in several allopolyploids, but TEs are unlikely to be the only causal factor responsible for subgenome expression dominance. In this review, we will examine the role of TEs in subgenome dominance as well as discuss how genetic incompatibilities among subgenomes likely contributes to the rapid emergence of subgenome dominance. Lastly, we highlight several burning questions about subgenome dominance that remain largely unanswered.

Full text of this work: https://www.sciencedirect.com/science/article/pii/S1369526620300340

CHAPTER 2

Chromosome-Scale Genome for a Red-Fruited, Perpetual Flowering and Runnerless Woodland Strawberry (*Fragaria vesca*)

The work presented in this chapter is part of the final publication:

Alger EI, Platts AE, Deb SK, Luo X, Ou S, Cao Y, Hummer KE, Xiong Z, Knapp SJ, Liu Z, McKain MR and Edger PP. 2021. Chromosome-Scale Genome for a Red-Fruited, Perpetual Flowering and Runnerless Woodland Strawberry (*Fragaria vesca*). *Front Genet* 12:671371.

Abstract

Although a high-quality reference genome is available for the diploid woodland strawberry (*Fragaria vesca*), it is for the 'Hawaii 4' accession which produces runners and yellow fruit. A reference genome that produces red fruit and is runnerless, key variants for two important traits for the commercial strawberry, is not publicly available. Here we report a near-complete genome of Fragaria vesca 'CFRA 2339' using Oxford Nanopore long read sequencing. The ‘CRFA 2339’ genotype produces red fruit, perpetually flowers, and is completely runnerless. The final assembly spanned 229.5Mb and has a contig N50 length of ~24.3 million base pairs (Mb). Three of the chromosomes are captured by a single contig, another three chromosomes are split into two contigs, and the one remaining chromosome is split into three contigs. These contigs were anchored to 7 pseudomolecules using comparisons to the 'Hawaii-4' genome yielding a final scaffold N50 length of ~29.6Mb. We also produced an annotation with a BUSCO score of 96%. Furthermore, comparative analyses uncovered previously identified mutations associated with fruit color, runnerless and perpetual flowering phenotypes. We anticipate that this new genomic resource will be utilized to uncover the underlying genetics of many important traits for strawberry.

Full text of this work: https://www.frontiersin.org/articles/10.3389/fgene.2021.671371/full

CHAPTER 3

Methylation and transposable elements shape the transcriptional landscape in woodland strawberry

Abstract

Transposable elements (TE) are a major driving force in shaping the evolution of eukaryotic genomes. The majority of observed genome size variation across land plants is due to the proliferation of various TE families. Novel TE insertions may have deleterious effects on the host genome, and genomes have evolved various mechanisms, including RNA-directed DNA methylation, to suppress the activity of TEs. The epigenetic silencing of TEs by DNA methylation has also been shown to result in decreased expression of neighboring genes. Thus, the genome must balance the 'trade-off' of silencing TEs with negatively impacting the expression of its genes. Here, we examined the 'trade-off' hypothesis in woodland strawberry (*Fragaria vesca),* including the assembly of a new chromosome reference genome for a new accession ('CFRA 562') to investigate gene expression bias in *in silico* and *in vivo* intraspecific hybrids between 'CFRA 2339' and 'CFRA 562'. Our analyses revealed that gene expression is negatively correlated with DNA methylation and surrounding TE density. Gene expression differences in the parents result in gene expression differences among parental genomes in our intraspecific hybrids. Finally, genes with expression biased toward one parent show different TE density and methylation patterns, than unbiased genes. Taken together, our results demonstrate a relationship between TE density, methylation, and gene expression that influences parental expression dominance in intraspecific hybrids.

Introduction

Transposable elements (TEs) are the single most abundant feature in many eukaryotic genomes [1]. TE abundance is highly variable and has largely contributed to the observed ~2,400 and ~66,000 fold-range in genome size across angiosperms and eukaryotes, respectively [2].

Transposable elements were originally considered to be purely "selfish" parasitic features of the genome [3,4]. However, recent studies have shown that TEs have also played a major role in facilitating adaptation to new environments [5,6]. Transposable elements can directly influence genes in the genome and ultimately phenotypic traits by inserting into genes [7], creating new promoters [8] or impacting the expression of nearby genes [9,10]. For example, the insertion of a TE into the *cortex* gene gave rise to the black morphotype that permitted the peppered moth (*Biston betularia*) to adapt during the Industrial Revolution [11].

Because of their ability to rapidly proliferate and potentially be destructive, TEs are often silenced throughout the genome, by mechanisms such as DNA methylation [12]. However, the methylation of TEs can also lower expression of nearby genes by 'spilling over' onto the gene itself or by altering chromatin states that prevents the transcriptional machinery from binding to the promoter region [13]. This creates an evolutionary 'trade-off' where the benefits of suppressing the TE must be weighed against the cost of reduced gene expression [14]. For genes whose expression levels are important for improved fitness or even survival [15,16], the unmethylated state of a TE may be beneficial and thus left free to proliferate.

We utilized the woodland strawberry (*Fragaria vesca*) to further examine the trade-off hypothesis and evolutionary role of TEs. With its short generation time, available genomic resources and wide natural diversity, *F. vesca* is an ideal model species to investigate the evolutionary role of TEs in genome evolution. *Fragaria vesca* is distributed across most of the northern hemisphere and adapted to environments ranging from desert grasslands to moist forests to high-elevation alpine habitats [17]. In addition, *F. vesca* has excellent genomic resources with two publicly available chromosome-scale reference genomes, 'Hawaii 4' [18] and 'CFRA 2339' [19], and here we present the genome for a new accession 'CFRA 562'.

Here, we leveraged the new reference genomes to examine the relationship between TE abundance, DNA methylation levels, and gene expression in *F. vesca*. We had two major goals for this study: i) determine if TE density and DNA methylation levels influenced the expression of genes in *F. vesca* 'Hawaii 4' and ii) verify whether observed differences in gene expression abundance between accessions are passed on to their intraspecific hybrid offspring, resulting in differences between the subgenomes.

Results and Discussion

Do methylated TEs impact expression of neighboring genes?

First, we investigated if TEs influence gene expression across the *F. vesca* genome and whether this may have played a role in *F. vesca*. We used the *F. vesca* 'Hawaii 4' genome [18] to examine the relationship between TEs, DNA methylation, and gene expression in diploid strawberry. First, we examined types of TEs found near genes and intragenic using a cumulative window approach with windows from 500bp to 10kb increasing size in 500bp increments both upstream and downstream. We examined total TE density and also split TEs into 4 types: Long Interspaced Nuclear Elements retrotransposons (LINEs), Long Terminal Repeats retrotransposons (LTRs), DNA transposons (DNA1k), and unknown TEs (**Figure 3.1**). Using EDTA to annotate TEs in the *F. vesca* 'Hawaii 4' genome revealed that LTRs are the most prevalent TEs in the genome, composing ~15% of the genome (**Table 3.1**). Our data revealed that LTRs tend to accumulate further away from genes (**Figure 3.1**), whereas the density (the fraction of TE bases divided by the relevant window at hand) of the other TEs types examined remained at a relatively constant level from 2kb to 10kb windows away from the gene.

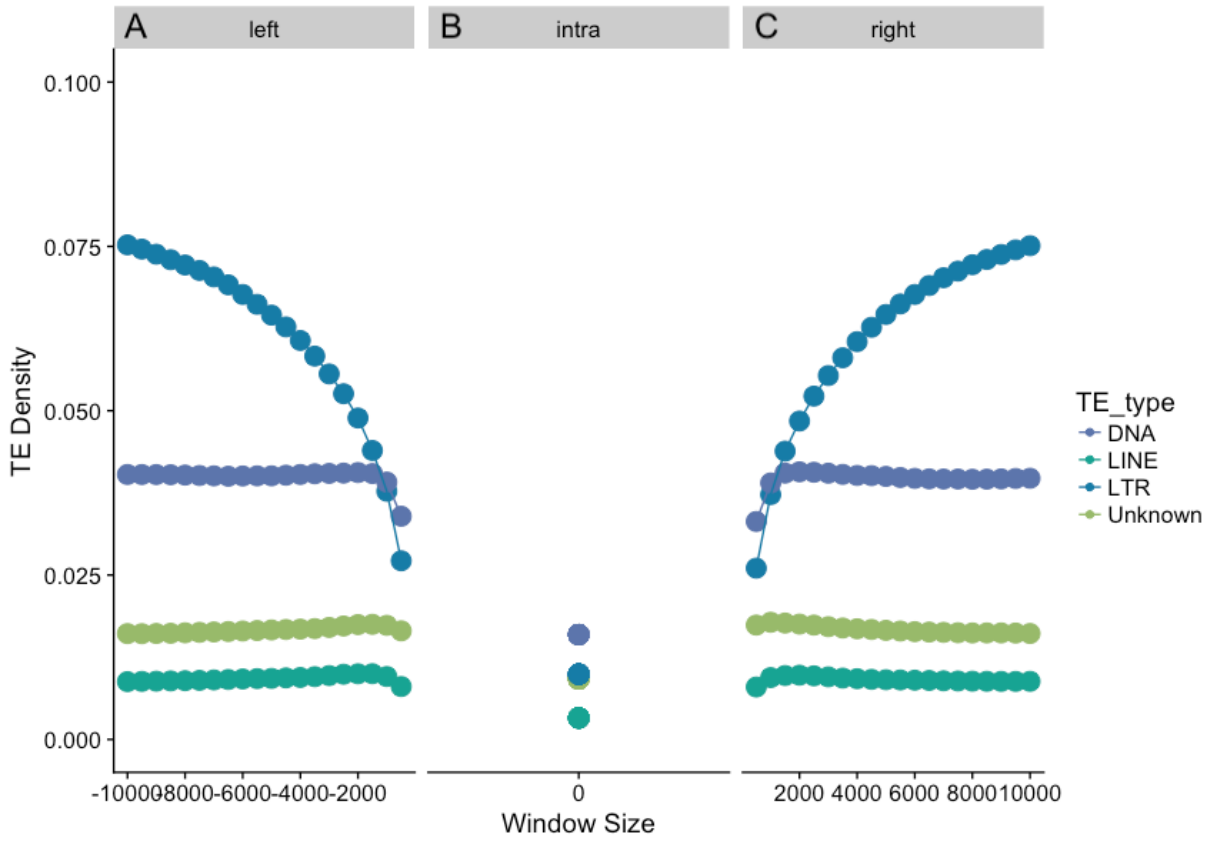

Figure 3.1 Average TE density of all 'Hawaii 4' genes as a function of window size and location

Average TE density of all genes with window sizes ranging from 0.5kb-10kb by 500 bp increments for the 'Hawaii 4' genome. TEs are also split into 4 types: types: Unknown TEs (green), DNA transposons (purple), Long Interspersed Nuclear Elements (LINE; teal), and Long Terminal Repeats (LTR; dark blue). Panel A represents the TE density values to the left of genes, panel B represents the intragenic TE density of genes, and panel C represents the TE density values to the right of genes.

TE Super-family		Count	Coverage (Mb)	Fraction of genome		
		'562'	'562'	'2339'	'H4'	'562'
Class I	LTR/*Copia*	10227	7.44	3.16%	3.34%	3.03%
	LTR/*Gypsy*	12861	14.71	6.13%	6.43%	5.99%
	LTR/Unknown	16507	11.15	5.01%	5.10%	4.54%
	Total Class I	**39595**	**33.30**	**14.30%**	**14.87%**	**13.56%**
Class II	CACTA	11655	5.90	2.62%	2.75%	2.40%
	Mutator	13021	4.63	1.90%	1.92%	1.89%
	PIF/Harbinger	2257	0.80	0.28%	0.22%	0.33%
	Tc1/mariner	144	0.09	0.03%	0.06%	0.04%
	hAT	5149	1.94	0.89%	0.84%	0.74%
	Helitron	12451	5.13	1.77%	1.36%	2.09%
	Total Class II	**44641**	**18.50**	**7.49%**	**7.15%**	**7.54%**
Repeat region		71524	25.38	10.54%	10.90%	10.34%
Total TEs		**155760**	**77.19**	**32.33%**	**32.92%**	**31.44%**

Table 3.1 Summarized transposable element (TE) composition for the *F. vesca* accessions
Summarized transposable element (TE) composition for the *F. vesca* 'CFRA 562' genome, including the fraction of the genome percentage for CFRA '2339' and 'Hawaii 4' for comparison.

Second, we calculated the average expression value of genes within each of these LTR-TE density windows. Results show that gene expression abundance is negatively correlated with TE density (**Figure 3.2**). This negative correlation between TE density and gene expression abundance is consistent with previous reports in *Arabidopsis* [14], monkeyflower [20], octoploid strawberry [21] and other species.

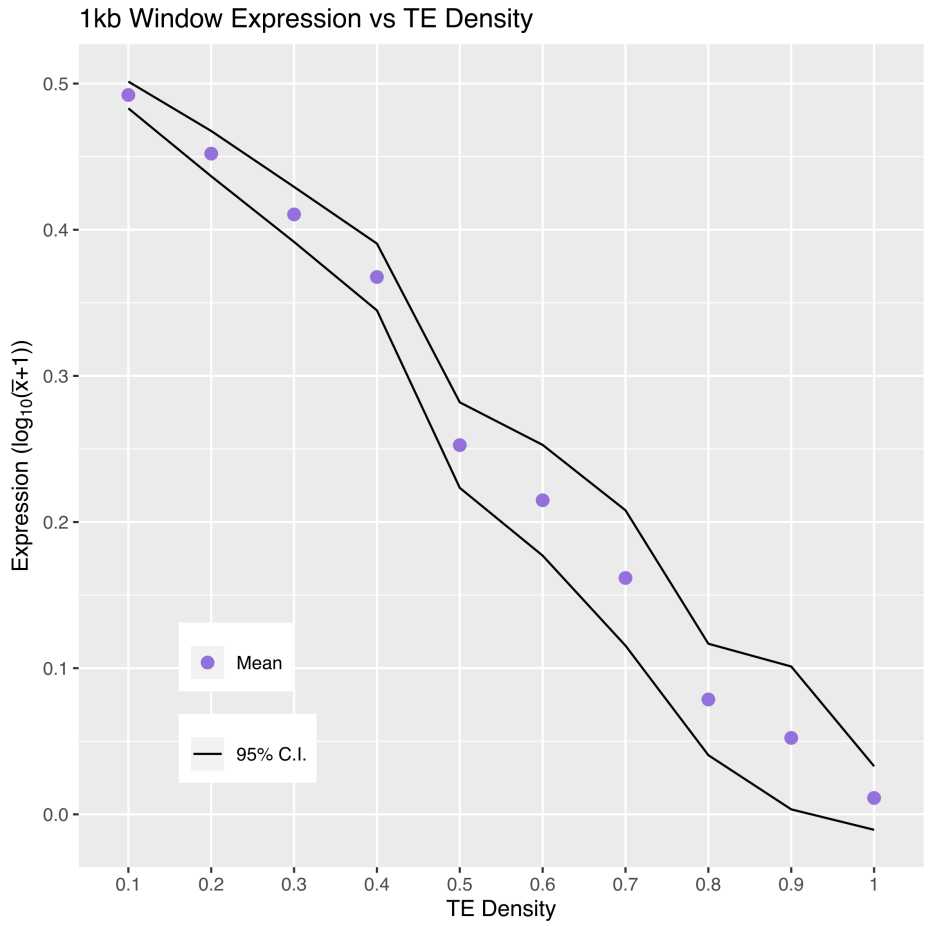

Figure 3.2 The average gene expression levels at different TE density levels
The average gene expression levels at different TE density levels using a window size of 2kb from the gene. The x-axis is TE density grouped into bins of 0-10%, >10%-20%, >20%-30%, etc. The y-axis shows average expression in FPKM.

Third, we used a 1kb window upstream and downstream from a gene to assign TE density to areas surrounding genes (**Figure 3.3**). TE density was grouped into 10% bins (0, >0-10%, >10%-20%, >20%-30%, etc.).

Fourth, we examined CG (mCG), CHG (mCHG), and CHH (mCHH) methylation (H = A, T or C) sites in genes and TEs. The mCG, mCHG, and mCHH level for each gene was defined as the ratio of mCG, mCHG, mCHH reads within a gene body to unmethylated CG, CHG, and CHH sites within a gene body, respectively. Average DNA methylation levels for mCG, mCHG, and mCHH sites (H = A, T or C) were analyzed at each TE density level. We

identified a positive relationship between TE density and for mCG, mCHG, and mCHH surrounding the gene (**Figure 3.3A**). CHH methylation was the lowest, with the highest average CHH methylation level being ~5%, while the highest average methylation for both CG and CHG was above 75% (**Figure 3.3)**.

mCG, mCHG, and mCHH were then grouped in the same 10% bins used for TE density. We plotted the number of gene windows at each TE density level colored with the number of gene windows for each methylation bin (**Figure 3.3B**). We found most of the gene windows for all methylation types had a methylation level of below 10% except mCG. mCG had the highest number of gene windows with methylation levels above 10% (53% of all genes), with a higher proportion of gene windows with >10% mCG as TE density increased. The vast majority of mCHG was ~10%, but the proportion of gene windows with >10% mCHG did increase with TE density (**Figure 3.3B**). For CHH methylation, very few gene windows had methylation above >10%, with the highest methylation level being only 50% (**Figure 3.3B**). Taken together, these patterns suggest that there are few windows with a high TE density, but these windows are more likely to have higher methylation than windows with low TE density.

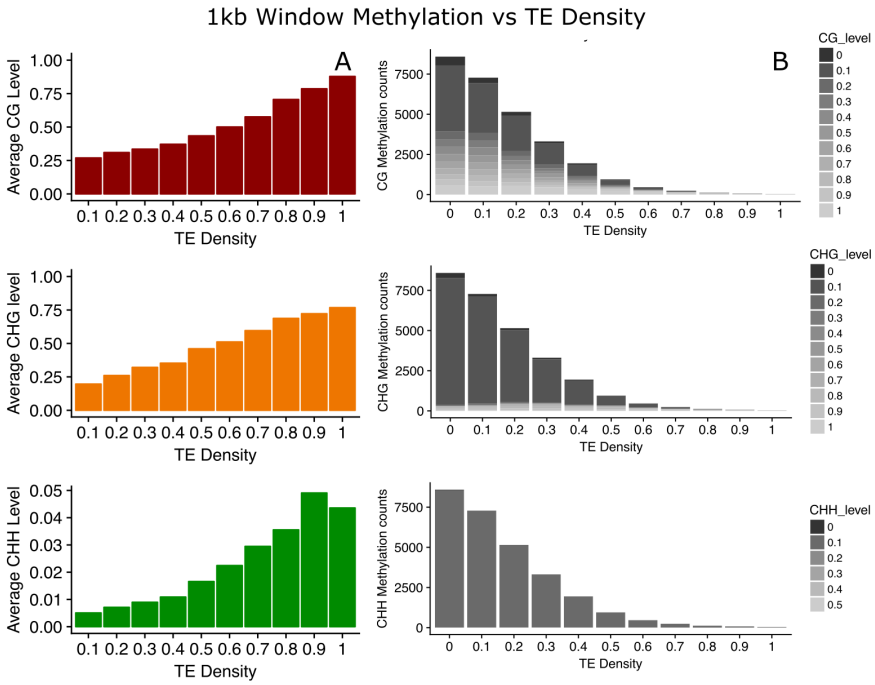

Figure 3.3 Average methylation level for each TE density window
Average methylation level for each TE density window and the number of gene windows at each TE density level colored with methylation for the 1kb window size. Panel **A** shows TE density and methylation surrounding genes. The x-axis is TE density grouped into 10% and the y-axis is the averaged amount of methylation for each TE density bin. Panel **B** shows the number of genes at each TE density bin for 1kb surrounding genes. Methylation is also grouped into 10% bins and the TE density counts are colored with the number of gene windows for each methylation bin.

Finally, we found the average gene expression for gene body mCG, mCHG and mCHH levels grouped together in 10% intervals. Because DNA methylation generally functions to suppress transcription [22], we expected that gene expression levels would decrease as DNA methylation increased. This was supported by our data for mCHG and mCHH, which was clearly negatively correlated with gene expression abundance (**Figure 3.4**). Increasing mCG levels did not appear to have a consistent relationship with average gene expression. Thus, our results show that gene body mCHG and mCHH within genes are both associated with lower gene expression levels in *F. vesca* 'Hawaii 4' (**Figure 3.4**). Because we observed the same negative correlation between TE density and gene expression (**Figure 3.2**) and showed that higher TE density results

in higher mCG, mCHG, and mCHH (**Figure 3.3**), mCHG and mCHH of TEs near genes may be associated with lowered gene expression. This influence on gene expression may explain why we observed a general trend of LTRs being preferentially inserted and/or maintained away from genes in the *F. vesca* 'Hawaii 4' genome (**Figure 3.1**).

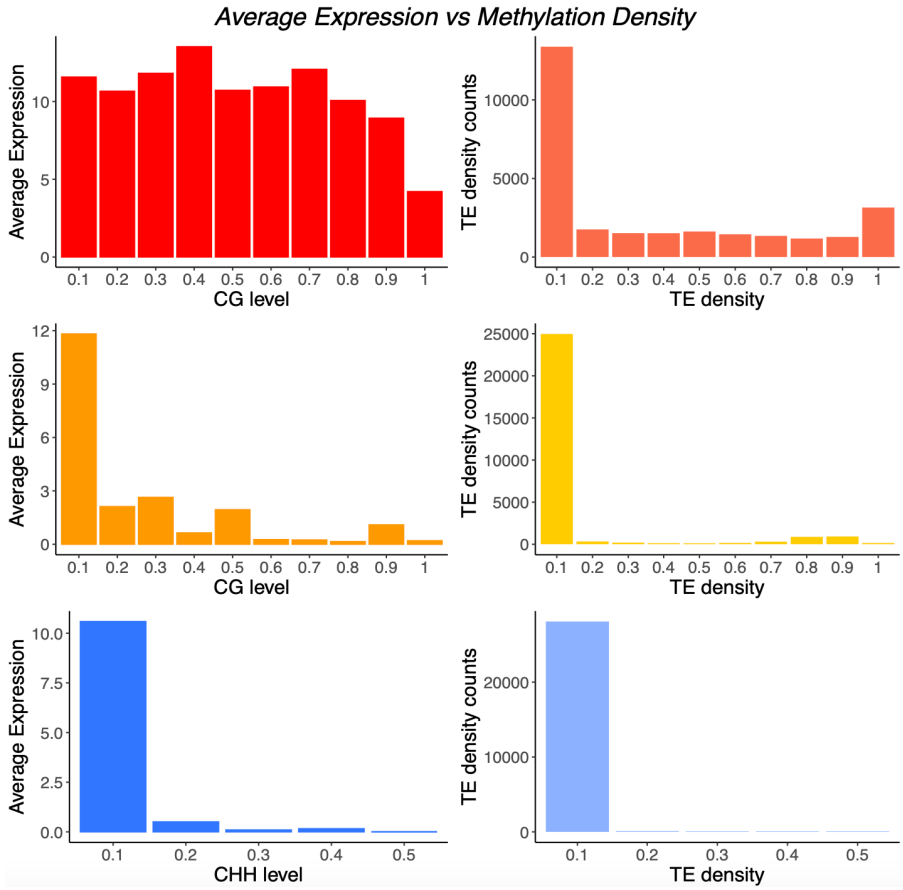

Figure 3.4 Average expression at different gene body methylation levels
The x-axis shows methylation levels grouped into 10% interval bins. Average amount of expression (in FPKM) in all gene widows with the same methylation level is shown on the y-axis

How does gene expression compare between two F. vesca accessions?

Based on our observations in 'Hawaii 4', which showed a decrease in expression as methylation and TE abundance increases, we propose a model where the methylation of TEs influences the expression of nearby genes rather than the TEs themselves (**Figure 3.5**). Applying

this idea to hybrid crosses between different *F. vesca* accessions, we expect the same gene in each parent to have similar expression if the gene is surrounded by a similar amount of methylated TEs in both parents or if the TEs near the gene are unmethylated in both parents (**Figure 3.5**). In hybrids where the amount of methylated TE surrounding the gene is different in each parent, we expect the parent with less methylated TEs to have higher gene expression. Even if TE levels are the same, if the TEs in one parent are methylated while TEs in the other parent are unmethylated, we expect the gene in the parent with unmethylated TEs to have higher expression (**Figure 3.5**).

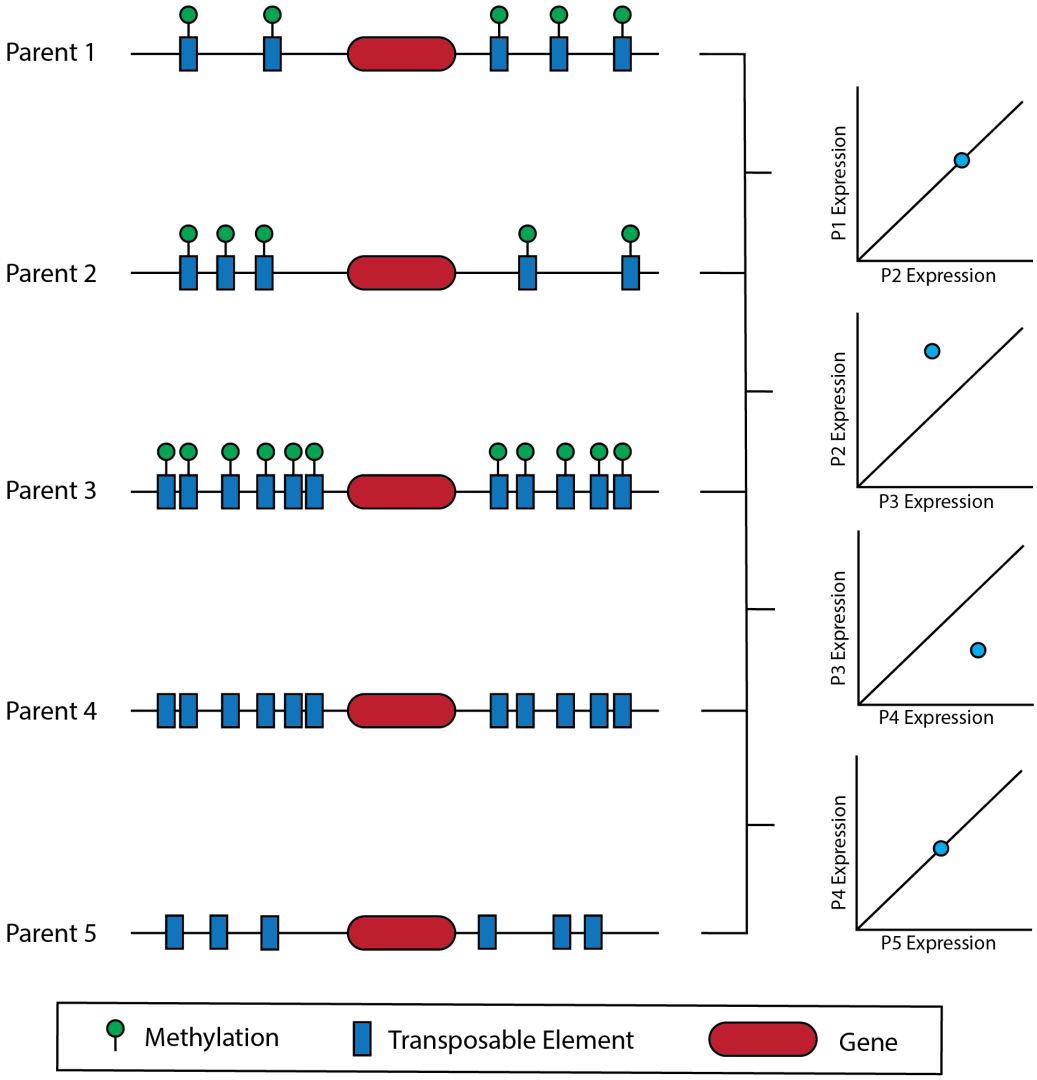

Figure 3.5 TE density, methylation, and expression model
Examples of expected expression comparison graphs for a single gene given various TE densities and methylation levels. Between parents where the gene has the same TE density and methylation levels, the parents are expected to have similar expression levels. Between parents where the TEs near the gene are methylated and one parent has more TEs than the other, the expression is expected to be higher in the parent with less TEs. Between parents where the gene has the same TE density in both, but TEs are only methylated in one parent, expression is expected to be higher in the parent without methylation. Between parents where the TEs near the gene are unmethylated, but each parent has different TE densities, the parents are expected to have similar expression levels due to the lack of methylation.

To examine expression differences between accessions, we compared differences in TE density, methylation, and expression between the recently published 'CFRA 2339' genome [19] and a new *F. vesca* genome, 'CFRA 562' (PI 551890), published here. This new genome

annotation has a BUSCO score of 95.2% and is comparable to the high quality *F. vesca* 'Hawaii 4' genome released in 2018 [23] with a BUSCO score of 96.4% as well as the 'CFRA 2339' genome with a BUSCO score of 96.0% (**Table 3.2**).

Genome Annotation Statistics			
	'Hawaii 4'	'CFRA 2339'	'CFRA 562'
Number of genes	28,588	30,349	32,242
Mean length of genomic loci	3,213	3,297	3,092
Mean exon number	5.5	5.8	5.3
Complete BUSCOs	96.4%	96.0%	95.2%
Fragmented BUSCOs	0.817%	1.33%	1.42%
Missing BUSCOs	2.75%	2.66%	3.44%

Table 3.2 Genome annotation statistics for *F. vesca* accessions
Gene features numbers and length statistics and BUSCO scores for the final genome annotation for the *F. vesca* 'Hawaii 4' *F. vesca* 'CFRA 2339' and *F. vesca* 'CFRA 562' genomes.

RNA isolated from young leaves of 'CFRA 2339' and 'CFRA 562' was mapped to each respective reference genome using STAR and quantified using StringTie. Syntelogs were identified between the genomes using CoGe for comparison [24].Each syntelog was then plotted using the average FPKM from 'CFRA 2339' on the x-axis and 'CFRA 562' on the y-axis as an *in silico* hybrid to determine if expression was skewed toward one of the genomes. We found overall gene expression was skewed toward 'CFRA 2339' (**Figure 3.6**).

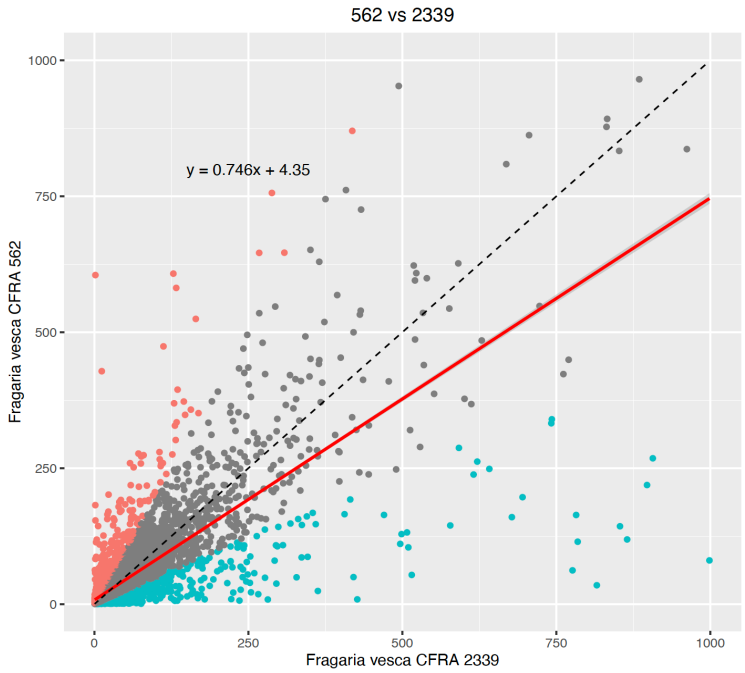

Figure 3.6 *In Silico* Hybrid
Overall expression comparison between *F. vesca* 'CFRA 2339' and *F. vesca* 'CFRA 562'. Each dot represents a single gene found in both accessions plotted using the expression level from the two accessions. Red dots are syntelogs with 2-fold higher expression in 'CFRA 562' and blue dots are syntelogs with 2-fold higher expression in 'CFRA 2339'.

We then crossed 'CFRA 2339' and 'CFRA 562' to create *in vivo* hybrids. The cross was made with 'CFRA 2339' as the maternal parent and 'CFRA 562' as the paternal parent, and vice versa. RNA was again extracted from young leaves and the reads were mapped to the 'CFRA 2339' genome as well as the 'CFRA 562' reference genome to identify expression differences between the two genomes within the hybrids. As observed in the *in silico* hybrid, both the 2339 ♀ x 562 ♂ (**Figure 3.7A**) and 562 ♀ x 2339 ♂ (**Figure 3.7B**) cross showed gene expression skewed toward 'CFRA 2339'. This agrees with our hypothesis that gene expression differences in the parents influence expression differences between genes in a hybrid and may determine which subgenome becomes dominant. While these are both F1 crosses, studies in allopolyploids have shown gene expression differences in early hybrids are maintained, with the subgenome with the higher expression becoming the subgenome [25].

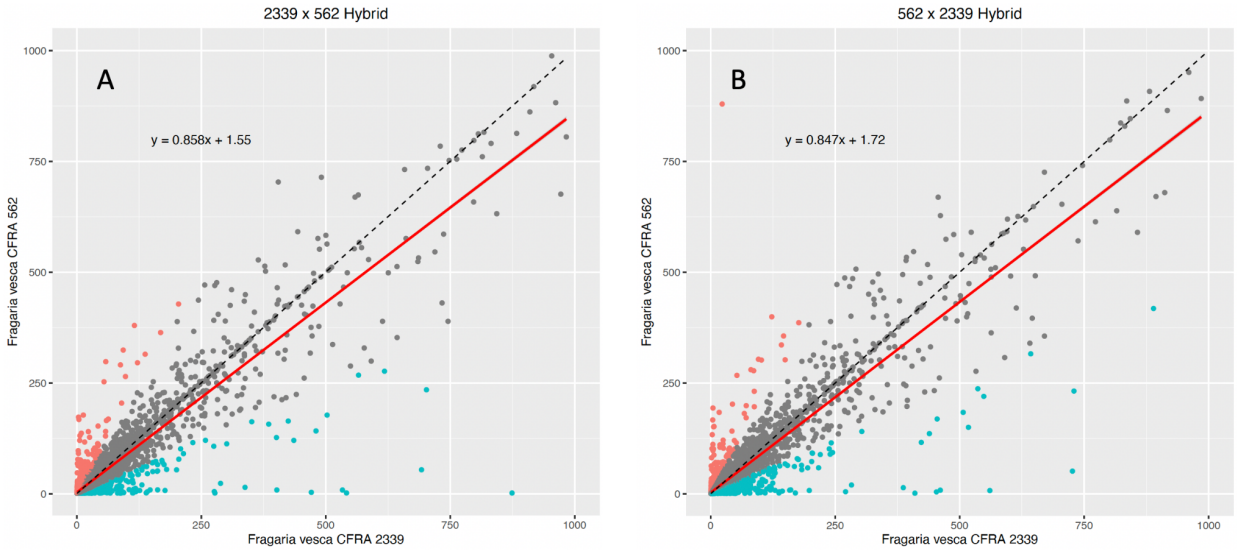

Figure 3.7 2339 x 562 and 562 x 2339 hybrids
Young leaf RNA from the hybrids was mapped to the 'CFRA 2339' and 'CFRA 562' reference genomes to determine the gene expression for each subgenome separately to identify differences in gene expression between subgenome. Panel **A** displays the 2339 ♀ x 562 ♂ hybrid and **B** displays the 562 ♀ x 2339 ♂ hybrid. Red dots are syntelogs with 2-fold higher expression in 'CFRA 562' and blue dots are syntelogs with 2-fold higher expression in 'CFRA 2339'.

Are there differences between biased and unbiased genes?

After identifying that 'CFRA 2339' had higher overall expression, we compared the TE density 2kb upstream and 2kb downstream from unbiased syntelogs, syntelogs that showed a 2-fold bias toward 'CFRA 2339' and syntelogs that showed a 2-fold bias toward 'CFRA 562' (**Figure 3.8**). For all three graphs, we can see that 'CFRA 562' has greater TE density near genes than 'CFRA 2339', though the difference is smaller for unbiased genes (**Figure 3.8B**) than biased genes. This trend is expected for 'CFRA 2339' biased genes (**Figure 3.8A**), as we hypothesized methylation of nearby TEs may result in lower expression in 'CFRA 562', and for unbiased genes, as we predicted unbiased genes would have more similar surrounding TE densities than biased genes. However, following the same logic, we would expect 'CFRA 2339' to have higher overall TE density for 'CFRA 562' biased genes (**Figure 3.8C**). This may be related to 'CFRA 562' biased genes having a less extreme overall gene expression bias compared

to ‘CFRA 2339’, shown in the slopes of the linear regression model where ‘CFRA 562’ biased genes slope (**Figure 3.9B**) is 4x steeper than the ‘CFRA 2339’ biased genes (**Figure 3.9A**), and therefore having smaller differences between TE densities surrounding genes as well. Overall, surrounding TE densities indicated that ‘CFRA 2339’, which shows higher expression, has lower overall TE content surrounding biased genes and unbiased genes show more similar TE density between the accessions.

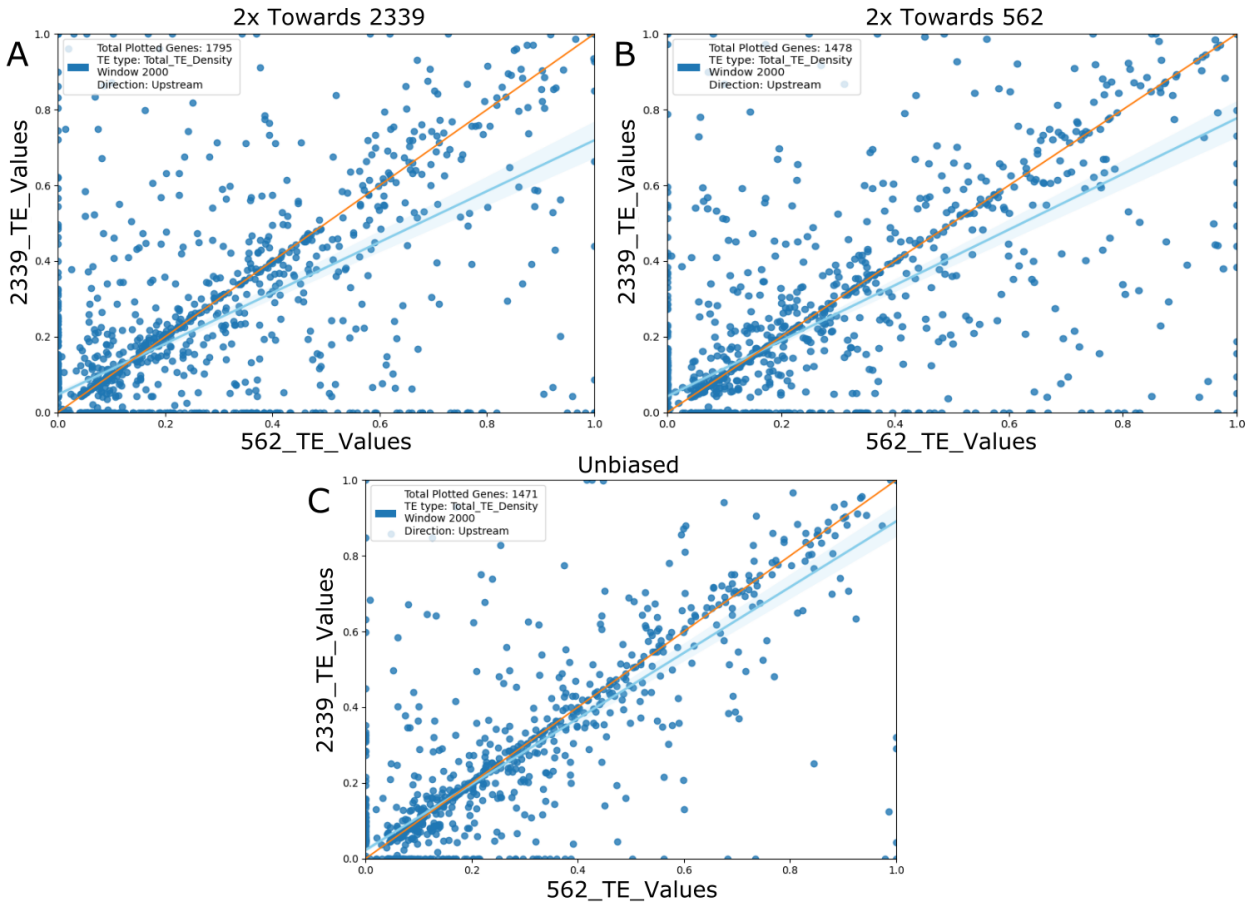

Figure 3.8 Upstream TE density for biased and unbiased genes
Panel **A** shows 2339 biased genes plotted using the 562 and 2339 TE density for each gene, panel **B** shows 562 biased genes plotted using the 562 and 2339 TE density for each gene, and panel **C** shows unbiased genes plotted using the 562 and 2339 TE density for each gene.

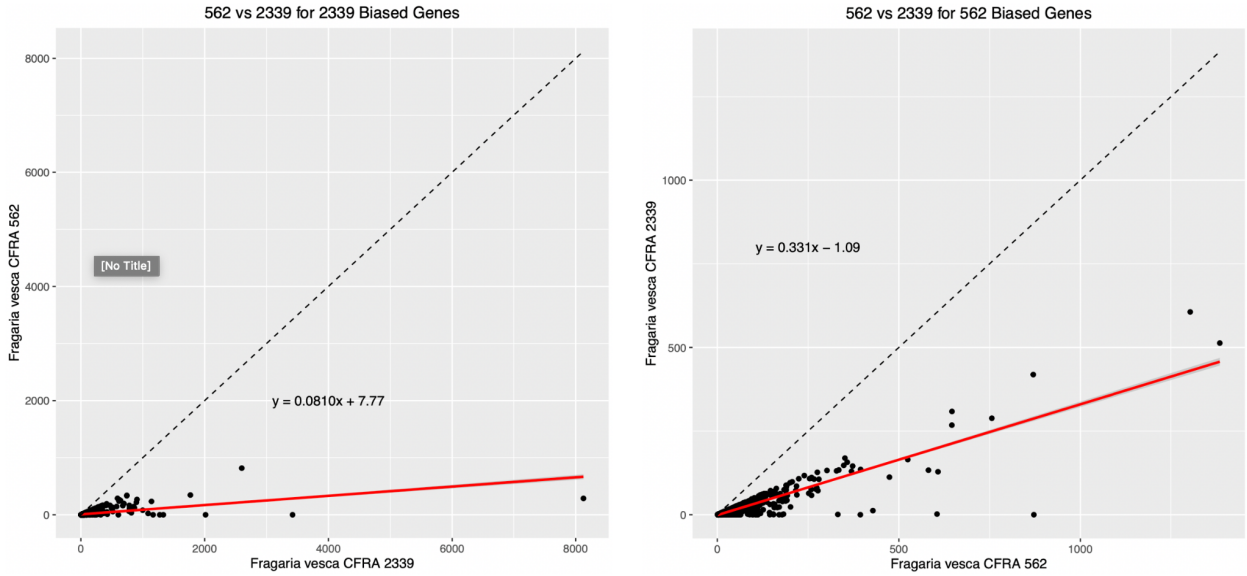

Figure 3.9 Expression of 2339 biased genes and 562 biased genes
Panel **A** shows 'CFRA 2339' biased genes plotted using 'CFRA 2339' and 'CFRA 562' expression data and **B** shows 'CFRA 2339' biased genes plotted using 'CFRA 2339' and 562 expression data.

We also examined mCG, mCHG, and mCHH methylation 2kb upstream from the transcriptional start site (TSS), 2kb downstream from the transcription termination site (TTS), and within the gene body (between TSS and TTS) for unbiased syntelogs and syntelogs with a 2-fold gene expression difference towards 'CFRA 2339' ('CFRA 2339' biased) or towards 'CFRA 562' ('CFRA 562' biased) (**Figure 3.10**). This analysis clearly showed a difference in methylation patterns between biased genes and unbiased genes, with unbiased genes showing higher methylation upstream from the gene body and lower methylation downstream for all methylation types. For gene body methylation, unbiased genes had higher mCG than biased genes, but lower mCHG and mCHH.

Based on our model (**Figure 3.5**), we would expect the 'CFRA 2339' genome to have lower overall methylation levels than the 'CFRA 562' genome because 'CFRA 2339' has higher gene expression (**Figure 3.7 & 3.8**). We hypothesized we would observe this in the biased genes as well, with 'CFRA 2339' biased genes having lower methylation than 'CFRA 562' biased

genes. However, with the exception of gene body mCG, methylation patterns for ‘CFRA 2339’ biased genes and ‘CFRA 562’ biased genes could not be differentiated from one another. For gene body mCG, ‘CFRA 2339’ biased genes showed higher mCG than ‘CFRA 562’ biased genes (**Figure 10**), which could be related to the higher gene expression observed in ‘CFRA 2339’ as gene body mCG is associated with increased gene expression [26–28].

The peak seen upstream of the TSS for mCHH is likely the result of mCHH islands, which are peaks in mCHH usually observed upstream from a gene and may act as a barrier between heterochromatin and euchromatin [29,30]. These mCHH islands are often, but not always, associated with TEs near genes as a way to silence TEs without negatively impacting the gene expression [29–31]. This peak is more pronounced in unbiased genes than biased genes, suggesting mCHH islands protecting gene expression are more often found upstream unbiased than biased genes.

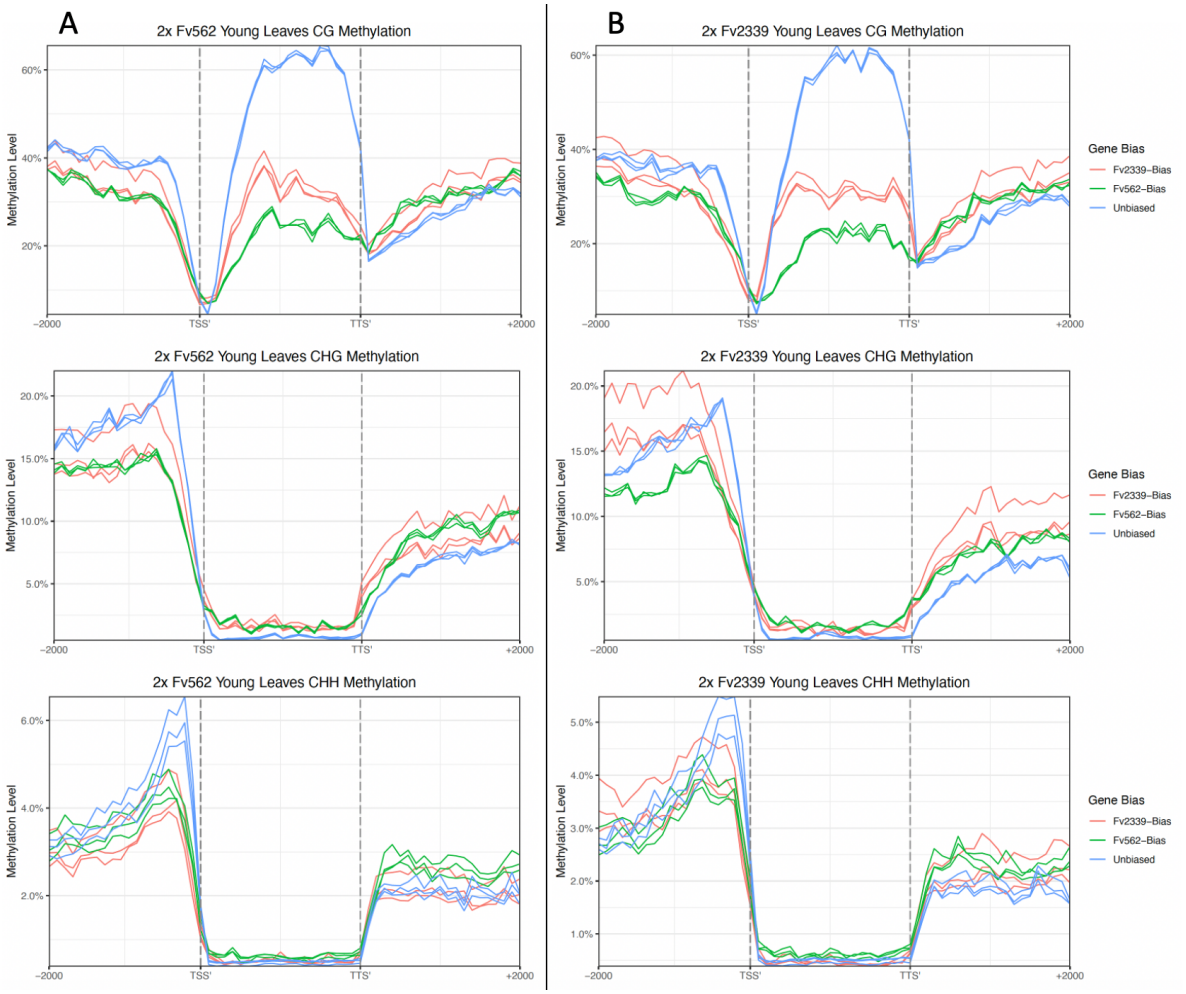

Figure 3.10 Methylation averages upstream, downstream, and within genes
All graphs show the average amount of methylation 2 kb upstream and downstream from all gene, as well as methylation averages within genes. Panel **A** has 'CFRA 562' biased genes (green), 'CFRA 2339' biased genes (red), and unbiased genes (blue) mapped to the 'CFRA 562' genome. Panel **B** shows the same genes mapped to the 'CFRA 2339' genome.

We then compared gene expression ratios between 'CFRA 2339' and 'CFRA 562' syntelogs to methylation ratios ('CFRA 2339'/'CFRA 562') for 1kb upstream from the TSS site. Again, we observed differences between biased and unbiased genes, with unbiased genes having more similar levels of mCG, mCHG, and mCHH between syntelogs than biased genes (**Figure 3.11**). We expected to see methylation differences between 'CFRA 2339' and 'CFRA 562' biased genes to reflect differences in gene expression as described in **Figure 3.5**, but no clear differences were observed. mCG and mCHG did not show bias toward either accession, but

mCHH was shifted towards ‘CFRA 562’ for both ‘CFRA 2339’ and ‘CFRA 562’ biased genes (**Figure 3.11),** suggesting the ‘CFRA 562’ syntelogs have more mCHH than the ‘CFRA 2339’ syntelogs. Relating methylation to TE content, this does track with the observation that ‘CFRA 562’ syntelogs also have higher TE content near genes for both ‘CFRA 2339’ and ‘CFRA 562’ biased genes (**Figure 3.8**) and are therefore more likely to have mCHH islands [29,30].

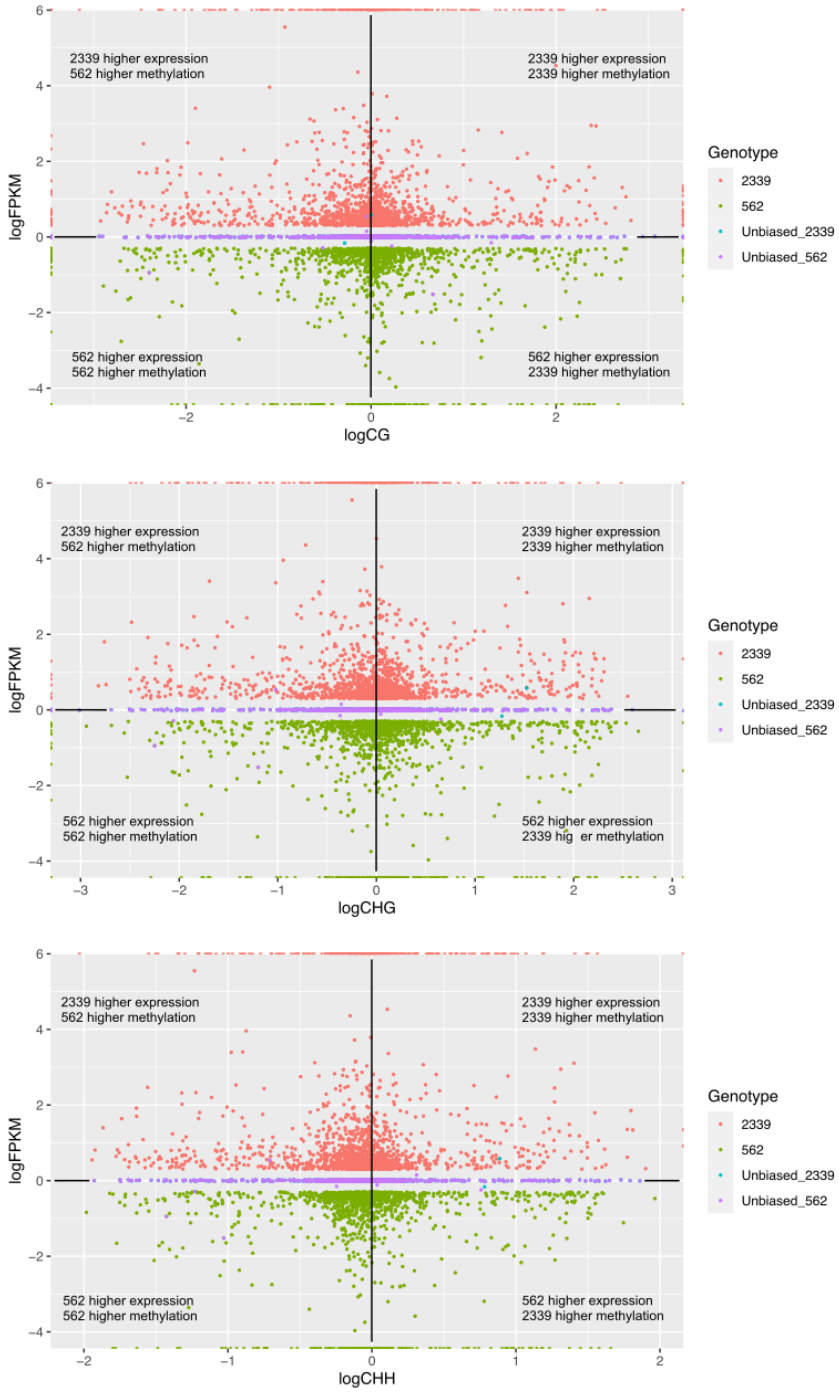

Figure 3.11 Expression-Methylation comparison plot
'CFRA 2339' gene expression (FPKM) and methylation were divided by 'CFRA 562' gene expression and methylation. The log transformed expression (x-axis) and methylation (y-axis) ratios were plotted for 'CFRA 2339' biased genes (red), 'CFRA 562' biased genes (blue), and unbiased genes (purple & blue). The graphs are divided into four quadrants labeled with the genome showing higher gene expression and methylation.

Summary

Our studies in *F. vesca* acc. 'Hawaii 4', which has a ~96.4% complete genome with gene and TE annotations [18], revealed that gene expression decreases as the TE density around genes increases. We also observed an increase in methylation as TE density increased and a decrease in gene expression as gene body CHG and CHH methylation levels increase. Using these results, we propose a model where methylation of TEs surrounding genes, not the TEs themselves, suppresses gene expression in nearby genes (**Figure 3.5**). When genomes with TE content differences and methylation are combined in a hybrid, gene expression will be skewed toward the genome with lower methylated TE content surrounding genes.

We also observed expression differences between two other *F. vesca* genomes, 'CFRA 2339' and 'CFRA 562'. Expression analysis revealed that overall gene expression is skewed toward 'CFRA 2339' in the in-silico hybrid. We hypothesized that in a hybrid cross between these two accessions, differences in gene expression in the parents would result in the 'CFRA 2339' subgenome having higher overall expression than the 'CFRA 562' subgenome. This was confirmed in 2339 ♀ x 562 ♂ and 562 ♀ x 2339 ♂ hybrid crosses (**Figure 3.7**), suggesting expression differences in parents can be passed down to hybrids and influence which subgenome will be dominant.

TE density and methylation in unbiased genes was compared to TE density and methylation in 'CFRA 2339' and 'CFRA 562' biased genes to identify differences between biased and unbiased genes. 'CFRA 562' and 'CFRA 2339' TE density was more similar in unbiased genes compared to biased genes (**Figure 3.9**). **Figure 3.10** and **Figure 3.11** shows that unbiased and biased genes also have different CG, CHG, and CHH methylation patterns, while 'CFRA 2339' and 'CFRA 562' biased genes have similar methylation patterns. Taken together

this confirms that differences in gene expression are related to surrounding methylation and TE density.

This study examined the relationship between TE density, methylation, and gene expression in *F. vesca* and found results suggesting: i) methylated TEs near and within genes lower gene expression, ii) gene expression differences in the parents can be used to identify which subgenome will be dominant in a hybrid, and iii) genes that show biased expression between parents have different TE densities and methylation patterns compare to unbiased genes. These findings demonstrate the importance of genomic features in the determination of subgenomic gene expression differences in F1 hybrids.

Materials & Methods

TE abundance, methylation, and expression in 'Hawaii 4'

Genome-wide associations were conducted between expression, methylation, and TE abundance. TEs were annotated using EDTA [32]. The percentage of TEs found in windows from 500bp-10kb in 500 base pair increments upstream and downstream from a gene and TE density within the gene was calculated to find using the TE_density tool (https://github.com/sjteresi/TE_Density). Three zones were defined for each gene: upstream, intragenic, and downstream. Chromosome identity, TE identity and TE length, and gene name and gene length were used to calculate TE density and were extracted from the 'Hawaii 4' annotation data. Gene length is defined as the number of base pairs between the first exon base pair and the last exon base pair. Python code was developed to calculate TE density for every gene in the genome where TE density is defined as: ∑ TE base pairs / Window Size upstream and downstream of the gene, respectively, and intergenic TE density is defined as: ∑ TE base

pairs / Gene Length. TE abundance calculations were then compared with gene expression and methylation data collected during the *F. vesca* f. *alba* 'Hawaii 4' genome annotation [18] as well as comparisons between gene expression and methylation data.

Comparisons between F. vesca accessions

The *F. vesca* 'CFRA 562' genome was sequenced, assembled, and annotated using the same methods described in Alger et al. [19] for *F. vesca* 'CFRA 2339'. For gene expression comparison, young leaves from *F. vesca* 'CFRA 2339' and 'CFRA 562' were collected at 12pm with 3 replicates each. Paired-end RNA-seq libraries were sequenced using Illumina HiSeq4000 platform in the Genomics Core at Michigan State University. The resulting RNA-seq reads were mapped to the 'Hawaii 4' reference genome using STAR (v. 2.5.3) [33]. Gene expression normalized for gene length and sequencing depth (FPKM) for uniquely mapped reads in each accession was completed using StringTie with default setting [34]. The values were plotted in R using ggplot() and the linear regression for each graph was found using the lm() function. TE density and plots for biased and unbiased genes were generated using the TE_Density tool (https://github.com/sjteresi/TE_Density) as described for 'Hawaii 4' with details in the CFRA_Syntelog_TE_Differences repository (https://github.com/sjteresi/CFRA_Syntelog_TE_Differences).

REFERENCES

REFERENCES

1. Feschotte, C., Jiang, N. & Wessler, S. R. Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.* **3**, 329–341 (2002).

2. Pellicer, J., Fay, M. F. & Leitch, I. J. The largest eukaryotic genome of them all? *Bot. J. Linn. Soc.* **164**, 10–15 (2010).

3. Doolittle, W. F. & Sapienza, C. Selfish genes, the phenotype paradigm, and genome evolution. *Nature* **284**, 601–603 (1980).

4. Goodier, J. L. & Kazazian, H. H., Jr. Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell* **135**, 23–35 (2008).

5. Fedoroff, N. V. Presidential address. Transposable elements, epigenetics, and genome evolution. *Science* **338**, 758–767 (2012).

6. Quadrana, L., Etcheverry, M., Gilly, A. & Caillieux, E. Transposon accumulation lines uncover histone H2A. Z-driven integration bias towards environmentally responsive genes. *bioRxiv* (2018).

7. McCLINTOCK, B. The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. U. S. A.* **36**, 344–355 (1950).

8. Lynch, V. J. *et al.* Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Rep.* **10**, 551–561 (2015).

9. Hollister, J. D. & Gaut, B. S. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.* **19**, 1419–1428 (2009).

10. Ichino, L. *et al.* MBD5 and MBD6 couple DNA methylation to gene silencing through the J-domain protein SILENZIO. *Science* (2021) doi:10.1126/science.abg6130.

11. Van't Hof, A. E. *et al.* The industrial melanism mutation in British peppered moths is a transposable element. *Nature* **534**, 102–105 (2016).

12. Matzke, M. A. & Mosher, R. A. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* **15**, 394–408 (2014).

13. Hirsch, C. D. & Springer, N. M. Transposable element influences on gene expression in plants. *Biochim. Biophys. Acta Gene Regul. Mech.* **1860**, 157–165 (2017).

14. Hollister, J. D. & Gaut, B. S. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.* **19**, 1419–1428 (2009).

15. Birchler, J. A. & Veitia, R. A. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 14746–14753 (2012).

16. Birchler, J. A. Genomic Balance Plays Out in Evolution. *The Plant Cell* tpc.00329.2019 (2019) doi:10.1105/tpc.19.00329.

17. Hilmarsson, H. S. *et al.* Population genetic analysis of a global collection of Fragaria vesca using microsatellite markers. *PLoS One* **12**, e0183384 (2017).

18. Edger, P. P. *et al.* Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (Fragaria vesca) with chromosome-scale contiguity. *Gigascience* **7**, 1–7 (2018).

19. Alger, E. I. *et al.* Chromosome-Scale Genome for a Red-Fruited, Perpetual Flowering and Runnerless Woodland Strawberry (). *Front. Genet.* **12**, 671371 (2021).

20. Edger, P. P. *et al.* Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower. *Plant Cell* **29**, 2150–2167 (2017).

21. Edger, P. P. *et al.* Origin and evolution of the octoploid strawberry genome. *Nat. Genet.* **51**, 541–547 (2019).

22. Niederhuth, C. E. *et al.* Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* **17**, 194 (2016).

23. Edger, P. P. *et al.* Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (Fragaria vesca) with chromosome-scale contiguity. *Gigascience* **7**, 1–7 (2018).

24. CoGe. *Comparative Genomics* https://genomevolution.org/coge/.

25. Flagel, L., Udall, J., Nettleton, D. & Wendel, J. Duplicate gene expression in allopolyploid Gossypium reveals two temporally distinct phases of expression evolution. *BMC Biol.* **6**, 16 (2008).

26. Tran, R. K. *et al.* DNA methylation profiling identifies CG methylation clusters in Arabidopsis genes. *Curr. Biol.* **15**, 154–159 (2005).

27. Zhang, X. *et al.* Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell* **126**, 1189–1201 (2006).

28. Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T. & Henikoff, S. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat. Genet.* **39**, 61–69 (2007).

29. Martin, G. T., Seymour, D. K. & Gaut, B. S. CHH Methylation Islands: A Nonconserved Feature of Grass Genomes That Is Positively Associated with Transposable Elements but Negatively Associated with Gene-Body Methylation. *Genome Biology and Evolution* vol. 13 (2021).

30. Gent, J. I. *et al.* CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. *Genome Res.* **23**, 628–637 (2013).

31. Harris, C. J. *et al.* A DNA methylation reader complex that enhances gene transcription. *Science* **362**, 1182–1186 (2018).

32. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).

33. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2012).

34. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. (2019) doi:10.1101/694554.

CHAPTER 4

Beach strawberry (*Fragaria chiloensis*) genome provides insights into high salinity tolerance

Abstract

Cultivated strawberry (*Fragaria* × *ananassa*) is an interspecific hybrid among two wild octoploid species: the Virginia strawberry (*F. virginiana*) and beach strawberry (*F. chiloensis*). Cultivars are highly sensitive to salinity, and seawater intrusion into aquifers is impacting the productivity of this crop in several important agricultural areas. Here we report a chromosome-scale assembly for *F. chiloensis* 'Del Norte' that is adapted to a high salinity environment – a lighthouse island off the coast of California. We identified two major quantitative trait loci (QTL) that encodes for high salinity tolerance using a genetic mapping population generated from crossing 'Del Norte' with a historically important cultivar 'Camarosa'. Interestingly, both QTL were contributed by the same diploid progenitor subgenome. A comparison of these regions between the 'Camarosa' and 'Del Norte' genomes revealed both gene copy number and presence-absence variation. Gene functional classification analyses, combined with differential gene expression analysis following a salt treatment, identified several candidate genes previously characterized in other species to be involved in salinity tolerance. These findings and the new reference genome, combined with the strawberry 50K genotyping array, were sufficient to identify markers that could be used to guide molecular breeding efforts to develop new cultivars with superior salinity tolerance.

Introduction

The cultivated strawberry (*F.* × *ananassa*) was developed at the beginning of the eighteenth century from an interspecific crossing between two North American wild octoploid strawberries, *F. virginiana* and *F. chiloensis* [1–3]. The male progenitor for cultivated strawberry was *F. virginiana,* known as the Virginia strawberry, which is native to the United States and

Canada [3,4]. The female progenitor was *F. chiloensis*, known as the beach strawberry, which is located in Chile and along the coast from California to the Aleutian Islands, and in the mountainous regions of Hawaii and Maui [1,5–7]. The first *F.* × *ananassa* plants were created from a spontaneous cross between these North American species in France in the early 1700s [2,3]. While *F.* × *ananassa* has superior fruit quality compared to its wild progenitors, cultivars often lack important biotic and abiotic resistance traits displayed by wild *F. virginiana* and *F. chiloensis*. *F. chiloensis*, for example, has ecotypes well adapted to grow in high levels of salinity.

Environmental stresses can greatly impact the growth and productivity of crops. Among these, salinity is one of the largest threats to agriculture worldwide and is a major escalating factor due to climate change [8]. Globally, 1 billion hectares of land is negatively affected by salinity and about 20% (63 million ha) of all irrigated arable land is salt affected [9,10]. California is one of the most important agricultural regions in the world and is well-known to suffer from salt-induced land degradation [11]. California supplies 91% of the total volume of strawberries in the U.S. and 73% of the planted area [12]. The strawberry industry is primarily located along the California Central Coast, which is experiencing significant seawater intrusion in several areas [13,14]. Therefore, it is very important to provide breeders with reliable sources of resistance and molecular tools, such as genetic markers, to accelerate the process to deliver salt-resistant cultivars.

A crucial part of salt tolerance in plants is the abscisic acid (ABA) dependent signaling pathway. ABA signaling induces regulatory factors, stress tolerance genes, and enzymes related to phospholipid signaling and represses genes related to plant growth and development [15,16]. Transcription factors are promising candidate genes for increasing salinity tolerance due to their

ability to target multiple genes related to salt tolerance and their overlap with other stress responses, especially with drought stress [17]. The NAC, MYB, WRKY, and bZIP are the four main transcription factors involved in stress response and ABA biosynthesis and have been characterized in *F. vesca* and *F.* × *ananassa* [15,17–23]. The ABA pathway also includes protein kinases, like mitogen activated protein kinases (MAPK) and calcineurin B-like protein (CBL)-interacting protein kinases (CIPK), which are induced by both salt and ABA treatment and act as another potential source of candidate genes for increased salinity tolerance [24,25].

Here, we present the genome of the highly salt tolerant *F. chiloensis* ecotype 'Del Norte', which was collected from a lighthouse island off the coast of California **(Figure 4.1)**. We utilized a genetic mapping population derived from a cross between 'Del Norte' and cultivar 'Camarosa' to identify potential salinity tolerance genes. The genome of 'Camarosa' was previously published [26]. Thus, we have chromosome-scale reference genomes for both parents of this genetic mapping population. Furthermore, the 'Del Norte' genome will serve as the first publicly available genome for *Fragaria chiloensis* - an important new resource for the strawberry breeding, genetics, and genomics communities. Our primary goal of this study was to identify the underlying genetics encoding salinity tolerance in cultivated strawberry and to develop potential markers to guide future breeding efforts.

Figure 4.1 'Del Norte' collection site
A map of California showing the area in Del Norte County where the high salinity tolerant 'Del Norte' ecotype was collected, with an aerial shot of the collection site on a lighthouse island off the coast.

Results

'Del Norte' genome assembly and annotation

Long read PacBio sequencing was combined with short read Illumina sequencing to produce a high-quality genome assembly for the wild octoploid strawberry *F. chiloensis* 'Del

Norte'. PacBio sequencing generated 5.3 M CLR sequences with an average size of 22.2 kb, and containing 117 Gb of total sequence, providing over 140x coverage for the 'Del Norte' genome. The reads were first assembled using FALCON and corrected using MiniMap/RagTag to align the genome to the *F.* × *ananassa* cv. 'Camarosa' for scaffolding and Pilon using Illumina data. The final assembly spans 815.5 MB across 697 scaffolds with an N50 27.9 Mb and contains 28 pseudomolecules (**Table 4.1**).

Final Assembly Scaffold Statistics		
	'Camarosa'	'Del Norte'
Total Sequences	28	697
Assembly size	805.5 Mb	815.5 Mb
Gaps %	0.65%	0.012%
N50	24.1 Mb	27.9 Mb
Max	43.6 Mb	41.3 Mb

Table 4.1 Statistics for the final genome assembly for the 'Camarosa' and 'Del Norte' genomes

MAKER was used to annotate the genome and identified 90,862 genes with a mean length of 3,274bp. We used Benchmarking Universal Single-Copy Orthologs (BUSCO) to estimate the completeness of the genome annotation and found the genome contained 97.5% of the core genes from the BUSCO eudicot_odb10 database (**Table 4.2**). The Extensive *de novo* TE Annotator (EDTA) was used to annotate transposable elements. TEs were found to make up 42.28% of the 'Del Norte' genome, with long-terminal-repeat retrotransposons being the most abundant and composing 25.75% of the genome (**Table 4.3**).

Final Genome Annotation Statistics		
	‘Camarosa’	‘Del Norte’
Number of genes	108,087	90,862
Mean length of genomic loci	3,158	3,274
Mean exon number	5.4	6.2
Complete BUSCOs	98.9%	97.5%
Fragmented BUSCOs	0.10%	0.30%
Missing BUSCOs	1.00%	2.20%

Table 4.2 Gene features statistics and BUSCO scores for the ‘Camarosa’ and ‘Del Norte’ genome annotations

TE Super-family		‘Del Norte’ Count	‘Del Norte’ Coverage (Mb)	‘Del Norte’ Fraction of genome (%)	‘Camarosa’ Fraction of genome (%)
Class I	LTR/*Copia*	61767	43.24	5.30%	4.70%
	LTR/*Gypsy*	86447	90.36	11.08%	10.14%
	LTR/Unknown	144680	76.41	9.37%	8.98%
	Total Class I	**292894**	**210.0**	**25.75%**	**23.82%**
Class II	CACTA	83218	34.11	4.18%	6.20%
	Mutator	103709	32.57	3.99%	5.41%
	PIF/Harbinger	22779	8.132	1.00%	1.40%
	Tc1/mariner	1824	0.4717	0.06%	0.14%
	hAT	38388	15.32	1.88%	2.64%
	Helitron	87761	44.16	5.42%	4.36%
	Total Class II	**337679**	**134.8**	**16.53%**	**20.42%**
Total TEs		**630573**	**344.8**	**42.28%**	**44.24%**

Table 4.3 Summarized transposable element (TE) composition for the *F. chiloensis* ‘Del Norte’ genome

The assembled genome of 'Del Norte' is slightly larger than that of 'Camarosa' (**Table 4.1**), possibly reflecting an actual difference in genome sizes. Flow cytometry for 'Del Norte' estimated the genome size at 828.1Mb, which is larger than the 813.4Mb previously estimated for 'Camarosa' [26]. The 28 pseudomolecules between the two genomes are largely collinear with a few notable structural variants (**Figure 4.2**). These variants were screened but need to be more closely re-evaluated in follow-up studies with either chromosome conformation capture (Hi-C) or with additional long read data. The total number of annotated genes and transposable elements in the 'Del Norte' genome, compared to the 'Camarosa' genome, is roughly 15.9% fewer and 41.0% fewer (**Table 4.3**), respectively.

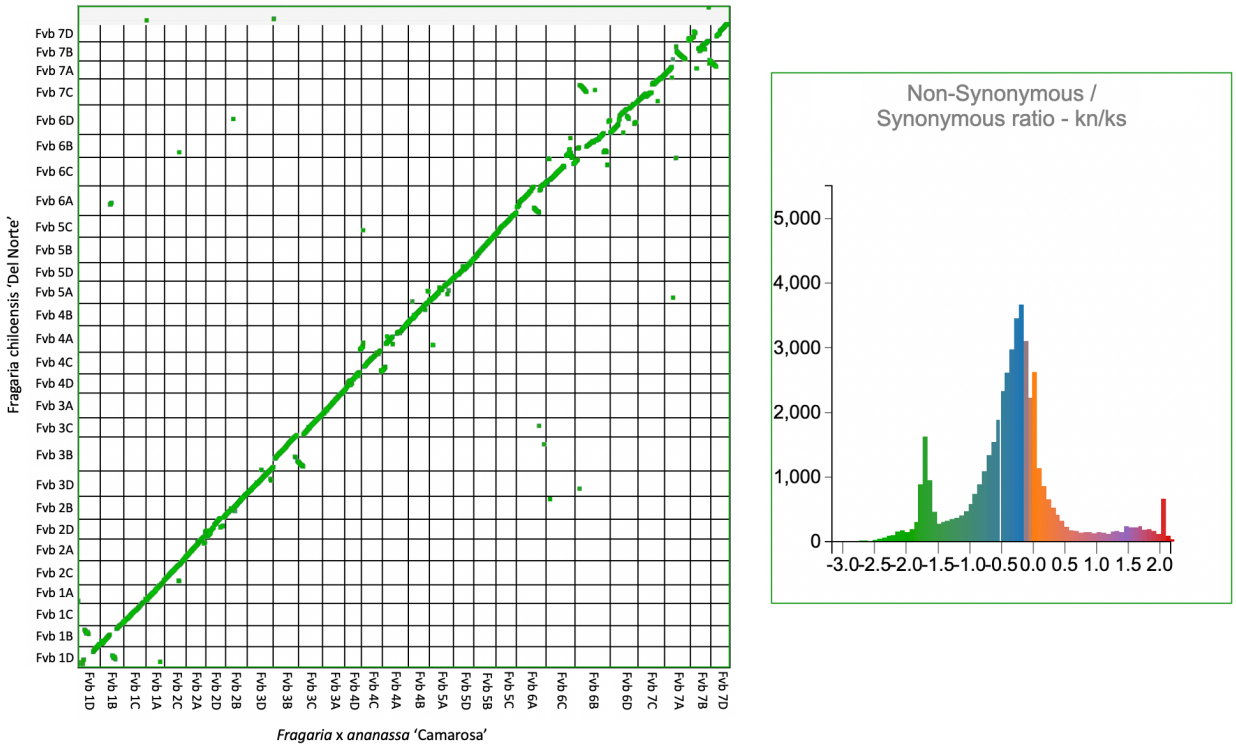

Figure 4.2 Synteny between *F.* × *ananassa* cv. 'Camarosa' and *F. chiloenesis* 'Del Norte'
Synteny analysis for *F.* × *ananassa* cv. 'Camarosa' (x-axis) and *F. chiloenesis* 'Del Norte' (y-axis) was performed with CoGe (Regenerate here: https://genomevolution.org/coge/) [27] with coloring depicting Kn/Ks ratios and syntenic orthologs shown on the diagonal line.

Comparative genetics (**Figure 4.2)** was used to identify the number of syntenic genes between the 'Camarosa' and 'Del Norte' genome. In whole genome comparisons, we found that 45.6% (49,291 genes) of 'Camarosa' genes had 'Del Norte' syntelogs (**Table 4.4**). For 'Camarosa' chromosomes 3B and 4B, 46.3% (1885 genes) and 50.1% (1497 genes) had 'Camarosa' syntelogs, respectively (**Table 4.4**). Within the QTL interval on 3B, 80.6% (50 genes) of the 62 'Camarosa' genes had 'Del Norte syntelogs, while 67.8% (181 genes) of the 267 'Camarosa' genes on 4B QTL had 'Del Norte' syntelogs (**Table 4.4**).

'Camarosa'	Number of 'Del Norte' syntelogs	Number of unique 'Camarosa' genes	Total number of 'Camarosa' genes
Whole genome	49291	57760	108087
Chromosome 3B	1885	2186	4071
Chromosome 4B	1497	1489	2986
3B QTL	50	12	62
4B QTL	181	86	267

Table 4.4 Gene count and syntelogs for 'Camarosa' and 'Del Norte'
Comparisons of 'Camarosa' and 'Del Norte' genomes to identify the number of 'Del Norte' syntelogs in the 'Camarosa' genome, in the 'Camarosa' 3B and 4B chromosomes, and within the 'Camarosa' 3B and 4B QTLs.

'Camarosa' and 'Del Norte' genetic maps

The number of segregating SNP markers was five-fold greater in the *F.* × *ananassa* parent (6,975) than the *F. chiloensis* parent (1,406) (**Table 4.5**). These assembled into 2,231 co-segregating bins in the female-specific ('Camarosa') subset and 1,039 co-segregating bins in the male-specific ('Del Norte') subset (**Table 4.5**). A single TAG-SNP was selected within each bin for locus grouping and ordering.

Metric	Female	Male
Number of Segregating SNP Markers	6,975	1,406
Number of TAG-SNP Marker Loci Mapped	2,231	1,039
Number of Linkage Groups	33	28
Genetic Map Length (cM)	2,473	2,241
Locus Spacing (cM)	0.4	1.6

Table 4.5 Metrics for the 'Camarosa' and 'Del Norte' genetics maps
Metrics used for QTL analysis of the F1 population. 'Camarosa' is the female parent and 'Del Norte' is the male parent.

QTL mapping for salinity tolerance

QTL mapping using the salt tolerance phenotype data was used to identify possible QTLs in both the 'Camarosa' and the 'Del Norte' genetics maps (**Figure 4.3**). Analysis using the single QTL model identified no significant 'Del Norte' QTL but did identify 2 'Camarosa' QTLs with significant LOD scores on linkage 3B (LG 3B) at ~112cm **(Figure 4.3B)** and linkage group 4B (LG 4B) at ~15 cm (**Figure 4.3C**) with all other linkage groups having LOD values far below the significance threshold (**Figure 4.3D**). The predicted QTL interval on 4B encompassed most of the linkage group, from 11 cM to 71 cM. In addition to the significant peak at 15 cM from LG 4B, the LOD score approaches the significance threshold at 27 cM and 34.11 cM. On LG 3B, the QTL interval was predicted to be from 99 cM to 115.39 cM. SNP markers that border the three peaks on LG 4B and the single peak on LG 3B can be found in **Table 4.6**. Single QTL analysis supports that there are salinity QTLs on both LG 4B and LG 3B, given the low LOD values on all other linkage groups, but both intervals are too large to identify individual candidate genes with this information alone.

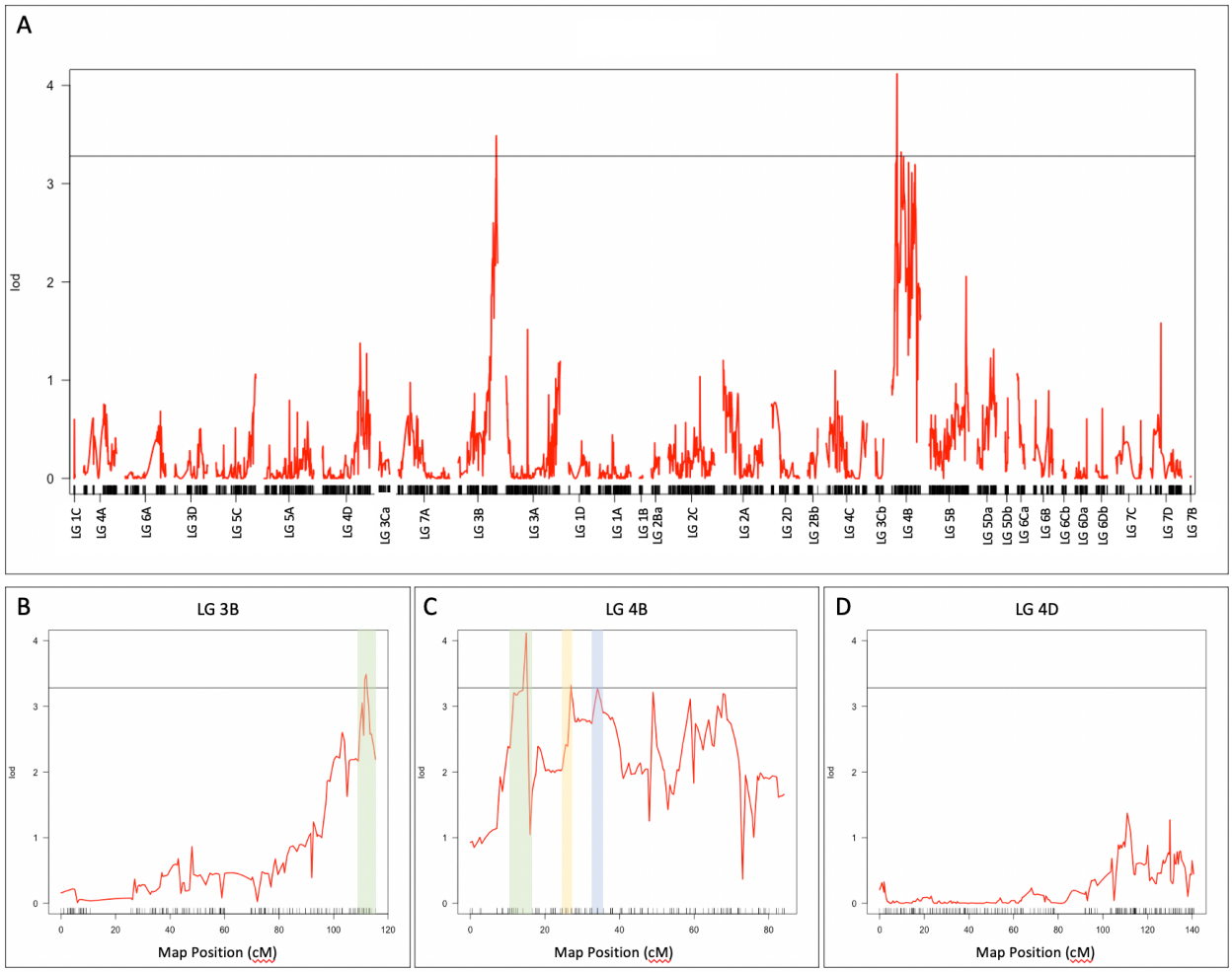

Figure 4.3 QTL graphs
QTL graphs from R/QTL showing the LOD scores on positions along the 'Camarosa' genetic map and the LOD significance threshold. (A) shows the LOD scores across all the 'Camarosa' chromosomes, with peaks above the LOD threshold in LG 3B and 4B. (B) and (C) show the LOD graphs for the chromosomes 3B and 4B, respectively. The significant peak on 3B is highlighted in green and the three significant peaks on 4B are highlighted in green, yellow, and blue. (D) shows LG 4D, which has no LOD scores near the significance threshold and is representative of the graphs from the remaining 'Camarosa' chromosomes.

QTL Marker Interval for 3B 111.94cM peak		QTL Marker Interval for 4B 27 cM peak	
AX-184187019	109.06 cM	AX-184565228	24.63 cM
AX-184168109	111.94 cM	cLG4B.loc27	27.00 cM
AX-184841526	115.39 cM	AX-184587388	28.07 cM
QTL Marker Interval for 4B 15 cM peak		QTL Marker Interval for 4B 34.11 cM peak	
AX-184072895	11.14 cM	AX-89848723	32.60 cM
cLG4B.loc15	15.00 cM	AX-184284533	34.11 cM
AX-166505789	16.61 cM	AX-184179461	35.61cM

Table 4.6 Gene markers for the QTL peaks
Gene markers for the QTL peaks and the intervals for 'Camarosa' chromosomes 3B and 4B.

Genomic prediction and heritability estimation

Our analyses suggest that tolerance to salt is heritable and, due to its complexity, genomic prediction (GP) has the opportunity to improve genetic gains for salt tolerance given a more diverse and appropriately connected training population. In the population that was evaluated in this study, narrow sense heritability of salt tolerance was low ($h^2 = 0.17$) and the broad sense heritability on an entry mean basis was moderate ($H^2 = 0.32$). Cross validated (k=1,000) genomic prediction yielded prediction accuracies of $r_{GS} = 0.1898$ ($r_E = 0.45$) from Bayesian Lasso and $r_{GS} = 0.1933$ ($r_E = 0.46$) from Bayesian Ridge Regression. This suggests that nearly 50% of the total genetic variance is constituted by non-additive genetic factors, e.g., dominance and epistasis and that markers are only able to explain 45% of the additive genetic variance.

RNAseq analysis and differential expression analysis

Differential expression analysis using RNAseq data from control and salt treated 'Camarosa' and 'Del Norte' plants was performed to identify genes differentially expressed during salt treatment. In total, 4,074 genes in 'Camarosa' and 234 genes in 'Del Norte' showed

significant differential expression between the control and salt-treated plants. OrthoFinder [28] was used to assign *Arabidopsis thaliana* orthologs to these genes for functional enrichment analysis using STRING [29]. The top 10 enriched GO Biological Processes for 'Camarosa' (**Table 4.7**) includes 'response to stress' and 'response to abiotic stimulus', both GO terms encompassing salinity response. The top 10 terms for 'Del Norte' (**Table 4.8**), however, did not include terms related to salinity response.

'Camarosa' GO Biological Process Functional Enrichment Top 10					
#term ID	**term description**	**observed gene count**	**background gene count**	**strength**	**false discovery rate**
GO:0042221	response to chemical	289	2654	0.26	3.18E-18
GO:1901700	response to oxygen-containing compound	184	1398	0.34	4.2E-18
GO:0050896	response to stimulus	458	5064	0.18	1.14E-16
GO:0006950	response to stress	280	2932	0.2	2.83E-11
GO:0001101	response to acid chemical	131	1058	0.32	8.25E-11
GO:0042493	response to drug	83	533	0.41	8.42E-11
GO:0010033	response to organic substance	184	1786	0.24	2.72E-09
GO:0010200	response to chitin	32	113	0.67	8.17E-09
GO:0009628	response to abiotic stimulus	170	1699	0.22	1.13E-07
GO:0010035	response to inorganic substance	95	795	0.3	5.4E-07

Table 4.7 'Camarosa' top 10 GO biological processes
Top 10 GO biological processes found to be enriched for the differentially expressed genes in 'Camarosa' between the control and salt-related plants. Functional enrichment analysis was completed with STRING. Response to chemical, response to stimulus, response to stress, and response to abiotic stimulus biological processes all include salt response.

'Del Norte' GO Biological Process Functional Enrichment Top 10					
#term ID	**term description**	**observed gene count**	**background gene count**	**strength**	**false discovery rate**
GO:0006357	regulation of transcription by RNA polymerase II	83	361	1.13	8.62E-59
GO:0030154	cell differentiation	87	680	0.87	1.45E-43
GO:0048869	cellular developmental process	92	814	0.82	2.18E-42
GO:0009751	response to salicylic acid	37	167	1.11	8.35E-25
GO:0046677	response to antibiotic	40	253	0.97	3.76E-22
GO:0014070	response to organic cyclic compound	44	331	0.89	7.58E-22
GO:0032502	developmental process	115	2492	0.43	3.72E-20
GO:1901700	response to oxygen-containing compound	83	1398	0.54	3.84E-20
GO:0006351	transcription, DNA-templated	98	1957	0.47	3.3E-19
GO:0006355	regulation of transcription, DNA-templated	104	2167	0.45	3.3E-19

Table 4.8 'Del Norte' top 10 GO biological processes
Top 10 GO biological processes found to be enriched for the differentially expressed genes in 'Del Norte' between the control and salt-related plants. Functional enrichment analysis was completed with STRING.

The differentially expressed genes were then combined with the QTL results to identify potential candidate genes on LG 3B and LG 4B in 'Camarosa'. LG 3B contained 157 differentially expressed genes and LG 4B contained 111 differentially expressed genes. Based on annotations for the assigned *A. thaliana* orthologs, 27 of these differentially expressed genes are related to salt stress response for LG 3B and 17 for LG 4B. Genes that also fall within the 112 cM peak on LG 3B and within the 14 cM, 27 cM, and the 34.11 cM peaks on LG 4B were also identified.

Only 12 differentially expressed genes were found on 'Del Norte' on LG 3B and only 8 on 'Del Norte' 4B. Of these genes, a gene encoding a Na+/Ca 2+ exchanger-like protein on was found on both 'Del Norte' LG 3B and 'Camarosa' LG 3B and a gene encoding a SWEET sucrose efflux transporter family protein can be found on 'Del Norte' LG 4B and 'Camarosa' LG 4B. 'Del Norte' 3B also contained a high-affinity potassium transporter (HKT) gene involved in sodium ion transport known to be involved in salinity tolerance in strawberry not found on 'Camarosa' 3B [30].

Discussion

'Del Norte' genome assembly and annotation

The 'Del Norte' genome assembly and annotation presented here will be the first published genome for *F. chiloensis* or any wild octoploid strawberry. The final assembly has a length of 815.5 Mb and is anchored to the anticipated 28 pseudomolecules (**Table 4.1**). The Benchmarking Universal Single-Copy Orthologs (BUSCO) [31] was used to estimate the completeness and quality of the genome assembly and annotation. With 97.5% of the core genes in the BUSCO eudicots dataset identified, this genome is comparable to the high-quality genome for *F.* × *ananassa* 'Camarosa', which has a BUSCO score of 98.9% (**Table 4.2**). Genome comparison found that 45.6% of 'Camarosa' genes had 'Del Norte' syntelogs. Synteny between 'Camarosa' and 'Del Norte' genes was increased within the QTLs, with 80.6% synteny for the 3B QTL and 67.8% synteny for 4B QTL (**Table 4.4)**. As a highly salt tolerant ecotype and the female progenitor for cultivated strawberry, this high quality 'Del Norte' genome will provide an excellent resource for the study of salinity tolerance among other important agronomic traits in

strawberry. The 'Del Norte' genome and findings presented here should be useful to guide breeding efforts to improve salinity tolerance in future cultivars.

QTL mapping for salinity tolerance

Linkage groups in the 'Camarosa' genome have been partitioned into distinct subgenomes for the four diploid progenitors that make up allo-octoploid strawberry: *F. vesca*, *F. iinumae*, *F. viridis*, and *F. nipponica*, with *F. vesca* being the dominant subgenome [26]. As the dominant subgenome, *F. vesca* has more highly expressed homoeologs and greater gene content resulting from higher gene retention of homoeologs and retention of tandem duplicates [26]. Because of the increased expression and gene content, the dominant subgenome also experiences stronger selective pressures in order to preserve vital genes and pathways compared to the three submissive subgenomes. The reduction in selective pressure on the submissive subgenomes may increase evolvability and promote adaptation to environmental stresses (e.g., soil salinity levels) [32]. The salinity QTLs identified in this study, LG 3B and LG 4B, are both on chromosomes assigned to the submissive *F. iinumae*-like subgenome. Our analyses here suggest that the *F. iinumae*-like subgenome may encode for salinity tolerance, whereas the *F. vesca*-like subgenome in 'Camarosa' was previously shown to control various fruit quality traits [26].

While QTLs were identified in 'Camarosa', none were identified in the 'Del Norte' ecotype, despite it being highly salt tolerant. The 'Del Norte' genetic map has a lower marker density compared to 'Camarosa', making QTL detection less sensitive in 'Del Norte'. However, the observed differences may be due to gene presence-absence variation between the two genomes (**Figure 4.2**).

Genomic prediction and heritability estimation

For this population, the narrow sense heritability of salt tolerance was low ($h^2 = 0.17$), and the broad sense heritability was moderate ($H^2 = 0.32$). Despite the parents of the evaluated population being divergent (one CA elite and one *F. Chiloensis*) [33], it is a single bi-parental population and the genomic relatedness among the full-sibs is fairly high and uniform, which is not ideal for GP and more often than not, GP is better used for predicting between, rather than within, family variation, e.g., family means [34–36]. Increasing the number of unique parents with greater half-sib/full-sib structure or incorporating a GWAS diversity panel can lead to higher heritability and predictive ability [37–41].

RNAseq analysis and differential expression analysis

Differential expression analysis between control and salt treated 'Camarosa' and 'Del Norte' plants identified 4,074 differentially expressed genes in 'Camarosa' and 229 in 'Del Norte'. Functional enrichment using *A. thaliana* orthologs found GO terms relating to stress response in the top 10 enriched GO Biological Processes for 'Camarosa', but none for 'Del Norte'. The QTL results and the *Arabidopsis* ortholog annotations were then combined with this expression data to identify potential candidate genes on the two linkage groups containing QTLs. Of the 4,074 differentially expressed genes (DEGs) in 'Camarosa', 157 mapped to LG 3B with 27 relating to salt response and 110 mapped to LG 4B with 17 related to salt response, therefore reducing our list from 4,074 to 44 candidate genes. Of these 44 genes, 20 are related to stress response signaling and ABA regulation, including transcription factors and protein kinases. For LG 3B, 5 of the 27 genes related to salt were in the QTL peak region at 112 cM. These five genes include: a member of the GRAS transcription factor family [42], a MYB family transcription factor [19], a kinase related to salt response [43], a glutathione transferase [44], and a ABA-induced

RDUF gene that positively regulates salt response [45]. For LG 4B, one salt response gene was in the QTL peak region at 15 cM, three were in the QTL peak region at 27 cM and 1 was in the QTL peak region at 34 cM. The gene underlying the QTL at 15 cM was an alternative oxidase (*AOX2*) gene related to salt response and regulated by the mitogen-activated protein kinase kinase 9-mitogen-activated protein kinase 3 (MKK9-MAPK3) [46,47]. The three genes found in the QTL region at 27 cM were: an F-box family protein [46], a glutathione transferase [44], and a phytocystatin gene [48,49]. Finally, the gene located in the QTL region at 34 cM was a xyloglucan endotransglucosylase - a gene product previously shown to be down regulated following salt treatment in *Arabidopsis* [50].

The aforementioned list of genes are strong candidates for encoding salinity tolerance because they impact the expression of many downstream genes relating to salinity tolerance. Of particular interest are the Na+/Ca+ exchanger-like protein gene on LG 3B (maker-Fvb3-2-snap-gene-115.56-mRNA-1 in 'Camarosa' and maker-Fvb3-2_RagTag-snap-gene-205.16-mRNA-1 in 'Del Norte') and the *Sugars Will Eventually Be Exported Transporters 15* (*SWEET15*) sucrose efflux transporter gene member on LG 4B (maker-Fvb4-4-snap-gene-75.62-mRNA-1 on 'Camarosa' and maker-Fvb4-4_RagTag-augustus-gene-67.73-mRNA-1 on 'Del Norte') as these were the only two differentially expressed salt response genes identified in both 'Camarosa' and 'Del Norte' and are confirmed syntelogs. The *SWEET15* gene is upregulated in *A. thaliana* and is associated with cell viability under stress conditions, with overexpression resulting in salt hypersensitivity and repression associated with increased tolerance [51]. The *SWEET15* gene was upregulated in both 'Camarosa' and 'Del Norte', consistent with the findings in *A. thaliana*. The expression of the Na+/Ca+ exchanger-like (NCL) protein gene was increased in 'Camarosa' and decreased in 'Del Norte'. In *Arabidopsis*, NCL mutants showed increased salinity tolerance

while overexpression increased salinity sensitivity [52]. Therefore, the differences in expression seen between 'Camarosa' and 'Del Norte' may be related to the differences in salt tolerance between these two species.

We have described a number of new resources for the wild octoploid *F. chiloensis*, including a high-quality genome, the first made available for this species. This ecotype provides a new source for genetic variation for cultivated strawberry. With this new genome and the genotyped F1 'Camarosa' x 'Del Norte' population, we were able to combine QTL results with differential gene expression analysis to identify potential markers and candidate genes with only one year of phenotype data. This material provided here is sufficient guidance for future breeding efforts using this population.

Materials and Methods

PacBio sequencing

High molecular weight (HMW) DNA from a single clonal representative of wild octoploid strawberry ecotype 'Del Norte' (PI551753) was extracted from nuclei isolated from immature leaf tissue using a modified version of this protocol: dx.doi.org/10.17504/protocols.io.4vbgw2n. One gram of liquid-nitrogen frozen immature leaf tissue was homogenized into a fine powder, and then transferred to a conical tube containing 10 mL of nuclei isolation buffer. The mixture was incubated at 4 °C for 15 minutes rotating end-over-end. After incubation, the mixture was filtered into a 50 mL conical tube through 2 layers of Miracloth and then centrifuged at 4 °C for 20 minutes at 3500 x g. The supernatant was decanted, and the pellet washed with 15 mL of cold nuclei isolation buffer. The sample was again centrifuged at 4 °C for 10 minutes at 3500 x g and the supernatant decanted. The nuclei pellet

was resuspended in 1 mL cold 1X homogenization buffer, transferred to a 1.5 mL Eppendorf tube, and centrifuged at 4 °C for 5 minutes at 7000 x g. The supernatant was decanted, and the Eppendorf tube was snap frozen in liquid nitrogen prior to storage at -80 °C. DNA was extracted from the frozen nuclei pellet using the Circulomics Nanobind Plant Nuclei Big DNA kit (SKU NB-900-801-001) and HMW DNA protocol with no modifications. To eliminate any fragments <10 kb, the extracted DNA was purified using the Circulomics Short Read Eliminator kit (SKU SS-100-101-01) and protocol with no modifications. DNA quantity and quality was assessed by Genomic Tapestation, NanoDrop, and Qubit prior to submission for sequencing. DNA samples were submitted to the University of Maryland Institute for Genomic Sciences for PacBio library preparation and sequencing on the Sequel II platform on two SMRT 8M Cells to generate 5.3 M CLR sequences with an average size of 22.2 kb and containing 117 Gb of total sequence.

Illumina sequencing for the assembly and annotation of 'Del Norte' genome

DNA from a single clonal representative of wild octoploid strawberry ecotype 'Del Norte' (PI551753) was extracted from immature leaf tissue using the E-Z 96 Plant DNA Kit (Omega Bio-Tek, Norcross, GA, USA) according to the manufacturer's instructions except for the following modifications; Proteinase K was added to the initial buffer, RNase treatment was delayed until after the lysate is removed from the cellular debris, an additional spin was added, and incubation steps were heated to 65 °C during elution. 'Del Norte' DNA was sheared using the Covaris E220 and size selected for 500 bp using magnetic beads (Mag-Bind® RxnPure Plus, Omega Bio-tek). 'Del Norte' paired end 250 bp libraries were constructed and sequenced on the HiSeq2500 at the Berkeley Genomic Sequencing Laboratory.

RNA was extracted from young leaves, old leaves, dark-treated young leaves, dark-treated old leaves, methyl jasmonate-treated leaves, roots, shoots, and runners. All tissues were

collected at 12pm from ‘Del Norte’ plants grown in a growth chamber at 21 °C with 65% relative humidity with 16hr/8hr day/night cycle. RNA was extracted using MagMAX^{TM}-96 Total RNA Isolation Kit (Thermo Fisher). RNA quality and concentration were assessed by Genomic Tapestation and Qubit before submission to the Michigan State University Genomics Core for single-end 150 bp Illumina TruSeq library preparation and sequencing on the HiSeq4000.

PacBio assembly

We obtained 120X haploid genome coverage from the ‘Del Norte’ PacBio CLR subreads for assembly, reserving 50X coverage of the longest subreads (≥47-kb) as seed reads for error-correction. We performed error-correction, overlap detection, and assembly of the ‘Del Norte’ reads using the FALCON hierarchical assembly protocol. To improve subgenome-specificity and subgenome haplotype-specificity during error-correction, we imposed a maximum seed read alignment coverage of 45X, well below the expected 70X remaining subgenome-specific read coverage for the 50X seed reads. We performed consensus error-correction of the raw octoploid subreads, generating 30X pre-assembled ‘pread’ coverage of ‘Del Norte’. At this stage we reconstructed phased ‘pread’ pools by BLAST aligning the 1.9M phased ‘Del Norte’ heterozygous marker sequences from a whole-genome shotgun recombination map published by Hardigan et al. [33] to the ‘preads’ and identifying ‘preads’ with at least 20 ‘Del Norte’ marker alignments and ≥85% markers specific to either phase-1 or phase-2 haplotypes in the ‘Del Norte’ recombination map. The ‘preads’ that did not meet these criteria were assumed to represent homozygous regions of the ‘Del Norte’ genome and were included in both ‘pread’ phase pools for downstream overlap detection and assembly. The phase-1 preads were then fed into the FALCON assembler in ‘data=pread’ mode to perform the overlap detection and assembly steps using default parameters.

'Del Norte' assembly correction and annotation

In order to reduce over-assembly in which both haplotypes were output in the primary assembly, the initial 945 MB assembly (N50 of 1.6 Mb) was aligned to the 'Camarosa' octoploid reference using minimap2 [53]. Scaffolds were first binned according to the chromosome with the highest minimap2 identity with the specificity of this selection assessed as base match count on best pseudomolecule relative to second best matching pseudomolecule alignment. Specificity was typically high, greater than 10 for large contigs, falling to 5 for shorter contigs in the 10-100 kb range, suggesting there was sufficient uniqueness between the subgenomes to unambiguously assign the majority of 'Del Norte' scaffolds to their cognate 'Camarosa' pseudomolecules. Chromosome-binned sets of haplotigs were then merged used HaploMerger2 [54] to generate a revised pseudo-haploid assembly of size 814MBase, comparable to the 805 Mb of the 'Camarosa' reference, and with a 2.7Mb N50. The interim assembly was then polished twice with Pilon [55] using 83 Gb (103x) of paired-end 250 bp Illumina reads aligned with BWA-MEM [56]. Polished scaffolds were then organized according to the layout of the 'Camarosa' reference using Ragtag [57] in reference scaffolding mode to generate 28 pseudomolecules alongside 669 unplaced contigs (**Figure 4.2**).

Transposable element annotation was carried out using EDTA (v1.9.6) [58]. Gene annotation was undertaken with Maker [59] in eukaryotic mode using several EST datasets: ESTs generated from a Trinity [60] transcriptome assembly guided by the polished reference genome and informed by RNAseq data from a broad SRA *F. vesca* dataset; ESTs generated from a Trinity [60] assembly of RNAseq data generated from various tissues from *F. vesca* 'CFRA 2339' [61]; ESTs were also introduced from previous *F. vesca* 'Hawaii 4' transcript annotations including v4.0.a2 and v4.0.a1 [62,63]. The annotation also used a protein homology dataset from *F. vesca* 'Hawaii 4'

v4.0.a1 annotation and a repeat masking library also derived from v4.0.a1 [62]. Several gene prediction Markov models were also introduced including a SNAP [64] model trained essentially as described by the authors, the Genemark eukaryotic HMM [65] and AUGUSTUS [66] hmm models generated using AUGUSTUS etraining on 1,000 gene sequences derived from the *F. vesca* 'Hawaii 4' v4.0.a1 gff3 annotation together with 1kb flanking regions.

'Camarosa' x 'Del Norte' F1 population development and genotyping

Seed of a full-sib family was produced by crossing the *F.* × *ananassa* cultivar 'Camarosa' (PI670238) with the *F. chiloensis* subsp. *lucida* ecotype 'Del Norte' (PI551753). Clones of Del Norte were acquired from the USDA Agricultural Research Service, National Plant Germplasm System, National Clonal Germplasm Repository, Corvallis, OR [67]. This ecotype was originally collected from a coastal habitat north of Crescent City, California (**Figure 4.1**). Clones of the cultivar 'Camarosa' were acquired from the UC Davis Strawberry Germplasm Collection, Winters, CA. The parent plants were grown in a greenhouse at the University of California, Davis over the winter of 2016-17. The cross was made by hand emasculating and pollinating 'Camarosa' flowers with 'Del Norte' pollen. The ripe fruit were macerated in a pectinase solution (0.6 g/L) to separate achenes (seeds) from receptacles in April 2017. Seeds were scarified in a concentrated sulfuric acid solution (36 Normal) for 16 min, rinsed in water, dried on blotter paper, and germinated at room temperature (approximately 22-24 °C) in June 2017. Approximately 200 seedlings were planted in artificial media (two parts vermiculite: one part sand) and grown in a shade house in Winters, CA from June to October 2017. Seedlings were transplanted to a Winters, CA field nursery in October 2017, grown to maturity, and clonally multiplied from stolons in 2018.

DNA was isolated from the parents and full-sib offspring as described by Hardigan et al. [33]. DNA samples were submitted to Thermo Fisher, Santa Clara, CA (http:\\https://www.thermofisher.com/) for genotyping with an Axiom 50K SNP array [33]. SNP genotypes were automatically called and manually curated with the Affymetrix Axiom Analysis Suite software (Affymetrix, Santa Clara, CA). SNP markers with distinct genotypic clusters and less than 5% missing data were selected for further analysis.

'Camarosa' and 'Del Norte' genetic map construction

Parent-specific backcross-equivalent genetic maps were developed by separating SNP markers into subsets segregating in female (AB x BB) and male (BB x AB) parents, where AB is a heterozygote and BB is a homozygote. SNP markers were tested for goodness-of-fit of observed to expected segregation ratios (1 AB : 1 BB for SNP markers segregating in the female and 1 BB : 1 AB for SNP markers segregating in the male) using chi-square statistics and *p*-values estimated with R (R Core Team, 2021). SNP markers with anomalus genotypes or highly distorted segregation ratios ($p < 0.01$) were dropped from the analysis.

The probe DNA sequences for Axiom 50K SNP array markers were anchored to the 'Camarosa' V1 genome assembly [26]. 'Camarosa' V1 physical addresses were used for the SNP markers genotyped and genetically mapped in our study. Linkage groups were numbered using the chromosome nomenclature proposed by Hardigan et al. [68]. Genetic maps were constructed using a custom analysis pipeline and the R packages '*onemap*' and '*BatchMap*' [69,70]. The analysis pipeline entailed binning co-segregating SNP markers, identifying a single SNP marker within each bin (TAG-SNP markers), calculating pairwise recombination fractions between TAG-SNP markers, assigning SNP markers to linkage groups, and estimating linkage disequilibrium (LD) statistics between SNP markers to identify and eliminate false-positive

linkage group assignments. Linkage groups were initially assembled using a LOD threshold of 10.0 and a maximum recombination fraction of 0.08. These thresholds produced more linkage groups than chromosomes. Sub-linkage groups were merged by inspecting inter-group LD statistics and percent-identities of SNP probe DNA sequences against the physical reference ('Camarosa' V1). SNP marker locus orders and genetic distances were re-estimated in parallel using the RECORD algorithm as implemented in '*Batchmap'* with a window of 25 SNP markers, a window overlap of 18 SNP markers, and ripple window of six SNP markers [70,71]. The window width was incrementally reduced by 5 from 25 to 5 for smaller linkage groups to ensure that at least two overlapping windows were analyzed. The Kosambi mapping function was used to convert recombination frequencies to centimorgans [72].

Screen progeny of 'Camarosa' x 'Del Norte' F1 for salinity tolerance

The 'Camarosa' x 'Del Norte' mapping population, consisting of 192 individuals, was used for screening for salinity tolerance in a greenhouse at Michigan State University in August 2018. 'Camarosa' is one of the more salt tolerant commercial cultivars but is still sensitive to high salinity levels while the wild ecotype 'Del Norte' shows salt tolerance even at high salinity levels. The F1 mapping population was shown to segregate for salinity tolerance (**Figure 4.4A**) and was genotyped with the new strawberry 50K genotyping array [33]. Treatment plants were hand watered until pot saturation with NaCl solution and control plants were watered using a drip irrigation system. Phenotyping of the F1 population during the salinity trial was completed at week 1 and week 3 following the initial salt (NaCl) treatment. For each individual from the population, eight plants (biological replicates; bioreps) were screened, with four control plants and four treatment plants, which were treated with 200 mM NaCl solution for one week and then with 100 mM NaCl solution for an additional two weeks. Plants were given a rating of 1-3 based

on the amount of leaf scorching and death (**Figure 4.4B**). Leaf damage can be observed on the leaves of individuals with intermediate tolerance. Roughly 9% of the individuals in the population were scored as highly tolerant.

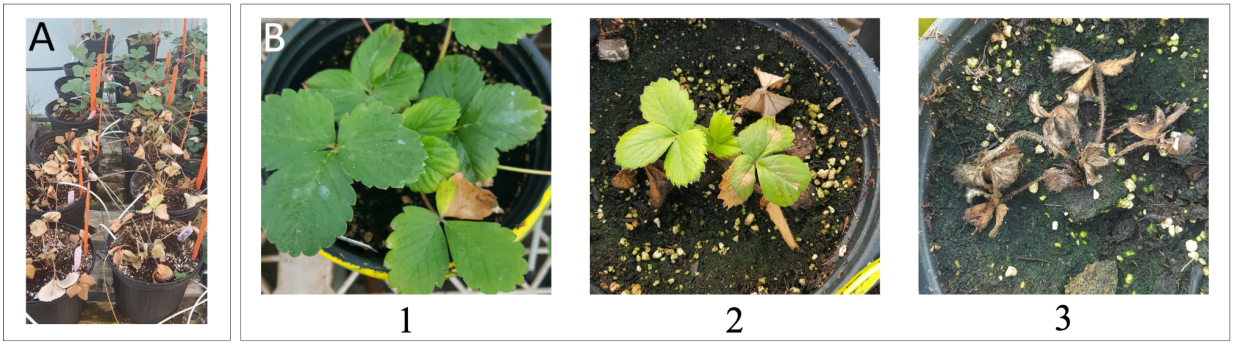

Figure 4.4 Phenotypes of individuals segregating in the F1 mapping population
Panel A - Six biological replicates of a highly sensitive individual are in the foreground and tolerant individuals are in the background. Panel B - Examples of plant appearance used for salt stress phenotyping. 1 – Large, healthy plant with no to minimal scorching; 2 – Growth inhibited plants with scorching; 3 – Dead. These pictures were taken a week after a continuous 200mM salt treatment.

QTL mapping for salt tolerance

Using the phenotypic data collected for salinity tolerance, QTL mapping was performed using R and the '*qtl'* package [73] to identify any regions associated with salt tolerance. Mapping was performed using the 'Camarosa' and 'Del Norte' genetic maps described above. The *scanone* function was used to identify QTL meeting a significance threshold determined by the permutation method (n = 10,000 permutations). After identifying significant single QTLs, the *scantwo* function was used to perform a two-QTL analysis, again using the permutation method (n = 10,000 permutations) to set a significance threshold. The Haley-Knott regression algorithm was used for both analyses [74]. The intervals for the QTLs were identified using the *lodint* function from the '*qtl*' package.

Genomic prediction and heritability estimation

Genomic-estimated breeding values (GEBVs) were estimated using the Bayesian Lasso and Bayesian Ridge Regression as implemented in the BGLR R package [75]. The input for these analyses were least square means (LSMs) for accessions. LSMs were calculated using the emmeans R package [76]. We removed SNPs with MAF < 0.05, max missing > 50%, and those that were 100% heterozygous which retained 24,745 SNPs after applying all filters. We used the rrBLUP R package to calculate the genomic relatedness matrix (GRM) [77]. Narrow-sense genomic heritability was estimated using the sommer R package using the GRM [78]. To estimate the accuracy of genomic predictions, we used Monte Carlo cross-validation (MCCV) with k = 1,000 replications generated by randomly splitting accessions into training (80%) and validation (20%) subsets. The accuracy of genomic selection (rGS) was estimated as the mean of correlations between observed phenotypes ($\bar{y}$) and GEBVs among the 10,000 replications, where the $\bar{y}$ are LSMs for accessions in the training population. We estimated the accuracy of a genomic selection relative to phenotypic selection using rE = rGS/h, where h is the square root of narrow-sense genomic heritability (h2) [79].

RNA-seq and differential expression analysis

RNA was collected from control and salt-treated mature ‘Camarosa’ and ‘Del Norte’ plants. Salt-treated and control plants were both watered to pot saturation, with the treatment plants being given 400 mM salt solution. Roots were collected 24 hours later for RNA extraction. RNA was extracted using the PureLink RNA mini kit and cleaned with the Qiagen RNeasy MinElute Cleanup kit. RNA was sequenced at the Michigan State University Genomic Core on the Illumina HiSeq4000 with 150-bp paired end reads. RNA sequencing data from ‘Camarosa’ and ‘Del Norte’ were mapped to the ‘Camarosa’ and ‘Del Norte’ reference genomes,

respectively, using the Spliced Transcript Alignment to a Reference (STAR) software [80]. HTseq [81] was then used to prepare count data for differential expression analysis using the R package DEseq2 [82]. OrthoFinder [28] was used to assign *Arabidopsis thaliana* orthologs to genes that were differentially expressed in the 'Camarosa' and the 'Del Norte' genome for functional enrichment analysis and to identify genes related to salt tolerance on 'Camarosa' linkage groups (LG) 4B and 3B.

REFERENCES

REFERENCES

1. Hancock, J. F. Ecological Genetics of Natural Strawberry Species. *HortScience* vol. 25 869–871 (1990).

2. Darrow, G. M. *The Strawberry: History, Breeding, and Physiology*. (1966).

3. Duchesne, A.-N. *Histoire naturelle des fraisiers contenant les vues d'économie réunies à la botanique, et suivie de remarques particulières sur plusieurs points qui ont rapport à l'histoire naturelle générale, par M. Duchesne fils*. (1766).

4. Harrison, R., Luby, J., Furnier, G. & Hancock, J. Morphological and molecular variation among populations of octoploid Fragaria virginiana and F. chiloensis (Rosaceae) from North America. *Am. J. Bot.* 84, 612 (1997).

5. Hancock, J. F. & Bringhurst, R. S. ECOLOGICAL DIFFERENTIATION IN PERENNIAL, OCTOPLOID SPECIES OF FRAGARIA. *American Journal of Botany* vol. 66 367–375 (1979).

6. Hultén, E. *Flora of Alaska and Neighboring Territories: A Manual of the Vascular Plants*. (Stanford University Press, 1968).

7. Staudt, G. *Systematics and Geographic Distribution of the American Strawberry Species: Taxonomic Studies in the Genus Fragaria (Rosaceae: Potentilleae)*. (Univ of California Press, 1999).

8. Hamdia, M. A. E.-S., Abd El-Samad Hamdia, M., Shaddad, M. A. K. & Doaa, M. M. Mechanisms of salt tolerance and interactive effects of Azospirillum brasilense inoculation on maize cultivars grown under salt stress conditions. *Plant Growth Regulation* vol. 44 165–174 (2004).

9. Ghassemi, F., Jakeman, A. J. & Nix, H. A. *Salinisation of Land and Water Resources: Human Causes, Extent, Management and Case Studies*. (UNSW Press, 1995).

10. Qadir, M. *et al.* Economics of salt-induced land degradation and restoration. *Natural Resources Forum* vol. 38 282–295 (2014).

11. Qadir, M. *et al.* Economics of salt-induced land degradation and restoration. *Natural Resources Forum* vol. 38 282–295 (2014).

12. United States Department of Agriculture National Agricultural Statistics Service. https://www.nass.usda.gov/Publications/Todays_Reports/reports/ncit0618.pdf.

13. Samtani, J. B. *et al.* The Status and Future of the Strawberry Industry in the United States. *HortTechnology* vol. 29 11–24 (2019).

14. United States Department of Agriculture National Agriculture Statistic Service. *Vegetables 2018 Summary.* (2019).

15. Fernando, V. C. D., Dilukshi Fernando, V. C. & Schroeder, D. F. Role of ABA in Arabidopsis Salt, Drought, and Desiccation Tolerance. *Abiotic and Biotic Stress in Plants - Recent Advances and Future Perspectives* (2016) doi:10.5772/61957.
16. Fujita, Y., Fujita, M., Shinozaki, K. & Yamaguchi-Shinozaki, K. ABA-mediated transcriptional regulation in response to osmotic stress in plants. *J. Plant Res.* 124, 509–525 (2011).

17. Motie-Noparvar, P., Varjovi, M. B., Lajayer, B. A. & Ghorbanpour, M. Engineering transcription factors: An emerging strategy for developing abiotic stress-tolerant crops. *Transcription Factors for Abiotic Stress Tolerance in Plants* 241–267 (2020) doi:10.1016/b978-0-12-819334-1.00013-7.

18. Li, C., Ng, C. K.-Y. & Fan, L.-M. MYB transcription factors, active players in abiotic stress signaling. *Environmental and Experimental Botany* vol. 114 80–91 (2015).

19. Li, H. *et al.* Genome-Wide Identification and Expression Analysis of MYB Transcription Factors and Their Responses to Abiotic Stresses in Woodland Strawberry (Fragaria vesca). *Horticulturae* vol. 7 97 (2021).

20. Wei, W. *et al.* The WRKY transcription factors in the diploid woodland strawberry Fragaria vesca: Identification and expression analysis under biotic and abiotic stresses. *Plant Physiol. Biochem.* 105, 129–144 (2016).

21. Liu, H. *et al.* Genome-wide analysis and evolution of the bZIP transcription factor gene family in six Fragaria species. *Plant Systematics and Evolution* vol. 303 1225–1237 (2017).

22. Shao, H., Wang, H. & Tang, X. NAC transcription factors in plant multiple abiotic stress responses: progress and prospects. *Front. Plant Sci.* 6, 902 (2015).

23. Moyano, E. *et al.* Genome-wide analysis of the NAC transcription factor family and their expression during the development and ripening of the Fragaria × ananassa fruits. *PLoS One* 13, e0196953 (2018).

24. Hwa, C.-M. & Yang, X.-C. The AtMKK3 pathway mediates ABA and salt signaling in Arabidopsis. *Acta Physiologiae Plantarum* vol. 30 277–286 (2008).

25. Pandey, G. K. *et al.* Calcineurin B-Like Protein-Interacting Protein Kinase CIPK21 Regulates Osmotic and Salt Stress Responses in Arabidopsis. *Plant Physiol.* 169, 780–792 (2015).

26. Edger, P. P. *et al.* Origin and evolution of the octoploid strawberry genome. *Nat. Genet.* 51, 541–547 (2019).

27. CoGe. *Comparative Genomics* https://genomevolution.org/coge/.

28. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16, 157 (2015).

29. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613 (2019).

30. Garriga, M. *et al.* Cloning and functional characterization of HKT1 and AKT1 genes of Fragaria spp.-Relationship to plant response to salt stress. *J. Plant Physiol.* 210, 9–17 (2017).

31. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol. Biol.* 1962, 227–245 (2019).

32. Hollister, J. D. Polyploidy: adaptation to the genomic environment. *New Phytol.* 205, 1034–1039 (2015).

33. Hardigan, M. A. *et al.* Genome Synteny Has Been Conserved Among the Octoploid Progenitors of Cultivated Strawberry Over Millions of Years of Evolution. *Front. Plant Sci.* 10, 1789 (2019).

34. Schopp, P., Müller, D., Wientjes, Y. C. J. & Melchinger, A. E. Genomic Prediction Within and Across Biparental Families: Means and Variances of Prediction Accuracy and Usefulness of Deterministic Equations. *G3* 7, 3571–3586 (2017).

35. Brandariz, S. P. & Bernardo, R. Small ad hoc versus large general training populations for genomewide selection in maize biparental crosses. *Theoretical and Applied Genetics* vol. 132 347–353 (2019).

36. Brauner, P. C., Müller, D., Molenaar, W. S. & Melchinger, A. E. Genomic prediction with multiple biparental families. *Theor. Appl. Genet.* 133, 133–147 (2020).

37. Gezan, S. A., Osorio, L. F., Verma, S. & Whitaker, V. M. An experimental validation of genomic selection in octoploid strawberry. *Hortic Res* 4, 16070 (2017).

38. Mangandi, J. *et al.* Pedigree-Based Analysis in a Multiparental Population of Octoploid Strawberry Reveals QTL Alleles Conferring Resistance to Phytophthora cactorum. *G3* 7, 1707–1719 (2017).

39. Norman, A., Taylor, J., Edwards, J. & Kuchel, H. Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy. *G3* 8, 2889–2899 (2018).

40. Pincot, D. D. A. *et al.* Accuracy of genomic selection and long-term genetic gain for resistance to Verticillium wilt in strawberry. *The Plant Genome* vol. 13 (2020).

41. Osorio, L. F., Gezan, S. A., Verma, S. & Whitaker, V. M. Independent Validation of Genomic Prediction in Strawberry Over Multiple Cycles. *Front. Genet.* 11, 596258 (2020).

42. Wang, T.-T. *et al.* Genome-Wide Analysis of the GRAS Gene Family and Functional Identification of in Drought and Salt Tolerance. *Front. Plant Sci.* 11, 604690 (2020).

43. Li, Z.-Y. *et al.* A mutation in Arabidopsis BSK5 encoding a brassinosteroid-signaling kinase protein affects responses to salinity and abscisic acid. *Biochem. Biophys. Res. Commun.* 426, 522–527 (2012).

44. Qi, Y. C. *et al.* Overexpression of glutathione S-transferase gene increases salt tolerance of arabidopsis. *Russian Journal of Plant Physiology* vol. 57 233–240 (2010).

45. Li, J. *et al.* The E3 ligase AtRDUF1 positively regulates salt stress responses in Arabidopsis thaliana. *PLoS One* 8, e71078 (2013).

46. Liu, J. *et al.* The F-box protein EST1 modulates salt tolerance in Arabidopsis by regulating plasma membrane Na /H antiport activity. *Journal of Plant Physiology* vol. 251 153217 (2020).

47. Song, J. B. *et al.* The F-box family genes as key elements in response to salt, heavy mental, and drought stresses in Medicago truncatula. *Functional & Integrative Genomics* vol. 15 495–507 (2015).

48. Hwang, J. E. *et al.* Distinct expression patterns of two Arabidopsis phytocystatin genes, AtCYS1 and AtCYS2, during development and abiotic stresses. *Plant Cell Rep.* 29, 905–915 (2010).

49. Zhang, X., Liu, S. & Takano, T. Two cysteine proteinase inhibitors from Arabidopsis thaliana, AtCYSa and AtCYSb, increasing the salt, drought, oxidation, and cold tolerance. *Plant Mol. Biol.* 68, 131–143 (2008).

50. Kreps, J. A. *et al.* Transcriptome changes for Arabidopsis in response to salt, osmotic, and cold stress. *Plant Physiol.* 130, 2129–2141 (2002).

51. Seo, P. J., Park, J.-M., Kang, S. K., Kim, S.-G. & Park, C.-M. An Arabidopsis senescence-associated protein SAG29 regulates cell viability under high salinity. *Planta* 233, 189–200 (2011).

52. Wang, P. *et al.* A Na /Ca2 Exchanger-like Protein (AtNCL) Involved in Salt Stress in Arabidopsis. *Journal of Biological Chemistry* vol. 287 44062–44070 (2012).

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

54. Huang, S., Kang, M. & Xu, A. HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics* 33, 2577–2579 (2017).

55. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9, e112963 (2014).

56. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).

57. Alonge, M. *et al.* RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* 20, 224 (2019).

58. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* 20, 275 (2019).

59. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12, 491 (2011).

60. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652 (2011).

61. Alger, E. I. *et al.* Chromosome-Scale Genome for a Red-Fruited, Perpetual Flowering and Runnerless Woodland Strawberry (Fragaria vesca). *Frontiers in Genetics* vol. 12 (2021).

62. Edger, P. P. *et al.* Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (Fragaria vesca) with chromosome-scale contiguity. *Gigascience* 7, 1–7 (2018).

63. Li, Y., Pi, M., Gao, Q., Liu, Z. & Kang, C. Updated annotation of the wild strawberry V4 genome. *Hortic Res* 6, 61 (2019).

64. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* 5, 59 (2004).

65. Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* 26, 1107–1115 (1998).

66. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 24, 637–644 (2008).

67. U.S. National Plant Germplasm System. https://npgsweb.ars-grin.gov/.

68. Hardigan, M. A. *et al.* Unraveling the Complex Hybrid Ancestry and Domestication History of Cultivated Strawberry. *Mol. Biol. Evol.* 38, 2285–2305 (2021).

69. Margarido, G. R. A., Souza, A. P. & Garcia, A. A. F. OneMap: software for genetic mapping in outcrossing species. *Hereditas* 144, 78–79 (2007).

70. Schiffthaler, B., Bernhardsson, C., Ingvarsson, P. K. & Street, N. R. BatchMap: A parallel implementation of the OneMap R package for fast computation of F1 linkage maps in outcrossing species. *PLoS One* 12, e0189256 (2017).

71. Van Os, H., Stam, P., Visser, R. G. F. & Van Eck, H. J. RECORD: a novel method for ordering loci on a genetic linkage map. *Theor. Appl. Genet.* 112, 30–40 (2005).

72. Kosambi, D. D. THE ESTIMATION OF MAP DISTANCES FROM RECOMBINATION VALUES. *Annals of Eugenics* vol. 12 172–175 (1943).

73. Broman, K. W., Wu, H., Sen, S. & Churchill, G. A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19, 889–890 (2003).

74. Haley, C. S. & Knott, S. A. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* 69, 315–324 (1992).

75. Pérez, P. & de los Campos, G. Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198, 483–495 (2014).

76. Russell, L. emmeans: estimated marginal means, aka least-squares means. R package version 1.4. 2. (2019).

77. Endelman, J. B. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome* vol. 4 250–255 (2011).

78. Covarrubias-Pazaran, G. Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. *PLOS ONE* vol. 11 e0156744 (2016).

79. Crossa, J. *et al.* Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity* vol. 112 48–60 (2014).

80. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* vol. 29 15–21 (2013).

81. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169 (2015).

82. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550 (2014).

CHAPTER 5

Concluding Remarks

Expanding genomic resources for the *Fragaria sp.* model system

This research improved the resources available for the *Fragaria* genus and demonstrated its strength as a model system. With the addition of two new genomes presented here in Chapter 2 and Chapter 3, *F. vesca* now has three high-quality genomes. These accessions all have short generation times, are easily crossed with one another, and have observable phenotypic differences. In addition to these diploid genomes, Chapter 4 includes the first published genome for *F. chiloensis*, the paternal wild octoploid parent of hybrid cultivated strawberry, *F. x ananassa*. The ecotype sequenced here, 'Del Norte', has a high salinity tolerance, a desirable trait for cultivated strawberry, making this genome a valuable resource to the strawberry breeding programs. With the addition of these genomes, this work has increased the resources available for this genus and demonstrated how it can be used as a model system for polyploid evolution and adaptation.

Subgenome dominance in F1 hybrids

The three *F. vesca* accessions with genomes, 'Hawaii 4' [1], 'CFRA 2339'[2], and 'CFRA 2339', were utilized to study the influence of genomic features in parental genomes on F1 hybrid gene expression bias, an indication of subgenome dominance. Based on previous research, which found that dominant subgenomes have less transposable elements (TEs) compared to recessive subgenomes [3–8], TEs were focused on in particular. Chapter 1 explained the hypothesis that the methylation used to silence TEs, not the TEs themselves, impact the expression of nearby genes, leading to the observed impact of TEs on subgenome dominance.

Analysis of the 'Hawaii 4' demonstrated that TEs and CHG and CHH methylation have a negative relationship with gene expression while there was a positive relationship between TEs

and CG, CHG, and CHH methylation. While these general trends do not prove the model presented in Chapter 1, it does confirm a relationship between TEs, methylation, and gene expression in *F. vesca*. The two new genomes, 'CFRA 2339' and 'CFRA 562' were crossed to create hybrids to examine changes in gene expression after hybridization. Comparisons between syntelogs in the two parental genomes found that 'CFRA 2339' had higher overall gene expression than 'CFRA 562'. When comparing the subgenomes in the hybrids, the same relationship was found, suggesting differences in gene expression in parents can be predictive of expression bias in hybrids.

With gene expression bias confirmed, syntelogs with the same expression in both genomes (unbiased genes) were compared to syntelogs with 2x higher gene expression in 'CFRA 2339' or 'CFRA 562' (biased genes). 'CFRA 562' had overall higher TE density near syntelogs for both 'CFRA 2339' biased genes and 'CFRA 562' biased genes. CG gene body methylation was higher for unbiased genes, while CHG and CHH methylation was lower for unbiased genes. These results establish a clear difference between unbiased genes and biased genes in the parental genomes, suggesting gene expression differences that are passed to hybrids are influenced by TE density and methylation. Next steps for this project will be to examine specific syntelogs, both biased and unbiased, to determine if methylated TEs suppress gene expression more than unmethylated TEs, as predicted.

Improvement of agronomic traits in strawberry

With increasing soil salinity threatening the production of many crops, including strawberry, improving salt tolerance is a top priority [9–11]. Toward that end, Chapter 4 presents a high-quality genome for highly salt tolerant *F. chiloensis* ecotype 'Del Norte'. A population

created from a cross between ‘Del Norte’ and *F. x ananassa* cv. ‘Camarosa’ was used to identify two QTLs associated with salinity tolerance in octoploid strawberry. These QTLs were identified on chromosomes 3B and 4B, which are both part of the *F. iinumae* subgenome. While *F. iinumae* is not the dominant subgenome, these results suggest that this genome controls salinity tolerance in octoploid strawberry. This research will provide the basis for future salinity breeding projects utilizing this ‘Del Norte’ x ‘Camarosa’ population.

REFERENCES

REFERENCES

1. Edger, P. P. *et al.* Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (Fragaria vesca) with chromosome-scale contiguity. *Gigascience* 7, 1–7 (2018).

2. Alger, E. I. *et al.* Chromosome-Scale Genome for a Red-Fruited, Perpetual Flowering and Runnerless Woodland Strawberry. *Front. Genet.* 12, 671371 (2021).

3. Schnable, J. C., Springer, N. M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U. S. A.* 108, 4069–4074 (2011).

4. Edger, P. P. *et al.* Origin and evolution of the octoploid strawberry genome. *Nat. Genet.* 51, 541–547 (2019).

5. Edger, P. P. *et al.* Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower. *The Plant Cell* vol. 29 2150–2167 (2017).

6. Wendel, J. F., Lisch, D., Hu, G. & Mason, A. S. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. *Current Opinion in Genetics & Development* vol. 49 1–7 (2018).

7. Hollister, J. D. & Gaut, B. S. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.* 19, 1419–1428 (2009).

8. Cheng, F. *et al.* Epigenetic regulation of subgenome dominance following whole genome triplication in Brassica rapa. *New Phytol.* 211, 288–299 (2016).

9. Hamdia, M. A. E.-S., Abd El-Samad Hamdia, M., Shaddad, M. A. K. & Doaa, M. M. Mechanisms of salt tolerance and interactive effects of Azospirillum brasilense inoculation on maize cultivars grown under salt stress conditions. *Plant Growth Regulation* vol. 44 165–174 (2004).

10. Ghassemi, F., Jakeman, A. J. & Nix, H. A. *Salinisation of Land and Water Resources: Human Causes, Extent, Management and Case Studies*. (UNSW Press, 1995).

11. Qadir, M. *et al.* Economics of salt-induced land degradation and restoration. *Natural Resources Forum* vol. 38 282–295 (2014).