

Fall 2006

# The development and utilization of EST (Expressed Sequence Tag) resources in the diploid strawberry model system

Robin Leigh Brese

*University of New Hampshire, Durham*

Follow this and additional works at: <https://scholars.unh.edu/thesis>

---

## Recommended Citation

Brese, Robin Leigh, "The development and utilization of EST (Expressed Sequence Tag) resources in the diploid strawberry model system" (2006). *Master's Theses and Capstones*. 192.  
<https://scholars.unh.edu/thesis/192>

This Thesis is brought to you for free and open access by the Student Scholarship at University of New Hampshire Scholars' Repository. It has been accepted for inclusion in Master's Theses and Capstones by an authorized administrator of University of New Hampshire Scholars' Repository. For more information, please contact [nicole.hentz@unh.edu](mailto:nicole.hentz@unh.edu).

THE DEVELOPMENT AND UTILIZATION OF EST (EXPRESSED SEQUENCE  
TAG) RESOURCES IN THE DIPLOID STRAWBERRY MODEL SYSTEM

BY

Robin Leigh Brese  
BA, Wheaton College, 2003

THESIS

Submitted to the University of New Hampshire  
in Partial Fulfillment of  
the Requirements for the Degree of

Master of Science

in

Plant Biology

September, 2006

UMI Number: 1437618

### INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

**UMI**<sup>®</sup>

---

UMI Microform 1437618

Copyright 2006 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

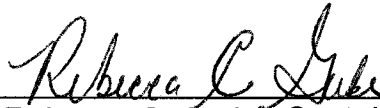
ProQuest Information and Learning Company  
300 North Zeeb Road  
P.O. Box 1346  
Ann Arbor, MI 48106-1346

This thesis has been examined and approved.



---

Thesis Director, Thomas M. Davis,  
Professor of Plant Biology/Genetics



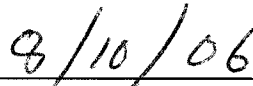
---

Rebecca C. Grubé, Sustainable  
Horticultural Crop Specialist



---

John J. Collins, Professor of  
Biochemistry and Molecular Biology



---

Date

## TABLE OF CONTENTS

LIST OF FIGURES.....	iii
LIST OF TABLES.....	iv
ABSTRACT.....	v

CHAPTER		PAGE
	OVERVIEW AND OBJECTIVES.....	1
I.	LITERATURE REVIEW.....	5
	The Cultivated Strawberry.....	5
	Polyploid Origin.....	9
	<i>Fragaria vesca</i> Genomic Resources.....	11
	Complementary DNA (cDNA) Libraries.....	15
	Polymerase Chain Reaction (PCR) Based Mapping.....	19
	Cleaved Amplified Polymorphic Sequence (CAPS).....	21
	Ribosomal Protein Genes.....	23
II.	METHODS.....	27
	Plant Material.....	27
	Total RNA Isolation.....	28
	Poly(A) mRNA Isolation.....	29
	cDNA Library Construction.....	29
	Clone Evaluation.....	33
	Sequencing and Bioinformatics.....	33
	DNA Isolation.....	34
	Gene Selection.....	36

	Polymorphism Detection.....	36
	PCR Conditions.....	37
	CAPS Conditions.....	39
	Gel Electrophoresis.....	39
III.	RESULTS.....	40
	cDNA Library Construction.....	40
	Clone Picking and Evaluation.....	43
	Bioinformatic Analyses.....	46
	Selection of Ribosomal Protein Genes.....	52
	Gene Amplification.....	57
	PCR and CAPS Polymorphisms.....	57
IV.	DISCUSSION.....	71
	RNA Quality.....	71
	Library Redundancy.....	72
	Homology Searches.....	74
	Expressed Sequence Tag Submission.....	76
	Library Utility.....	76
	Ribosomal Protein Gene Markers.....	78
	Polymorphism Detection.....	82
	Mapping Populations.....	84
	Summary and Completion of Objectives.....	86
V.	LITERATURE CITED.....	89

## LIST OF FIGURES

FIGURE		PAGE
1	Map of global <i>Fragaria</i> distribution.....	7
2	Intraspecific <i>Fragaria</i> linkage map.....	17
3	Interspecific <i>Fragaria</i> linkage map.....	18
4	Ribosomal protein genes mapped in <i>Arabidopsis</i> ...	26
5	CLONTECH cDNA synthesis procedure.....	30, 31
6	pDNR-LIB vector.....	32
7	Total RNA and mRNA isolations.....	42
8	PCR evaluation of random clones.....	45
9	Monomorphic and polymorphic allele example.....	59
10	RPL37aB PCR and digestion results.....	65
11	RPL32A PCR and digestion results.....	66
12	RPL10B PCR and digestion results.....	67
13	RPL27aC PCR and digestion results.....	68
14	RPS14 PCR and digestion results.....	69
15	RPL23B PCR and digestion results.....	70

## LIST OF TABLES

FIGURE		PAGE
1	<i>Fragaria</i> species, ploidies, and distribution.....	8
2	Eppendorf PCR kit reagents.....	38
3	Expressed Sequence Tag database matches.....	49
4	Most redundant clones in <i>F. vesca</i> cDNA library.....	50
5	<i>F. vesca</i> unigene matches to databases.....	51
6	60S ribosomal protein genes used.....	53, 54
7	40S ribosomal protein genes used.....	55, 56
8	Estimated and observed PCR product sizes.....	63
9	Polymorphisms observed.....	64



## ABSTRACT

### THE DEVELOPMENT AND UTILIZATION OF EST (EXPRESSED SEQUENCE TAG) RESOURCES IN THE DIPLOID STRAWBERRY MODEL SYSTEM

By

Robin Leigh Brese

University of New Hampshire, September, 2006

Enhancement of genomic resources is needed for the cultivated strawberry (*Fragaria ×ananassa*), a member of the Rosaceae family. However, the octoploid genome composition of *F. ×ananassa* ( $2n = 8x = 56$ ) presents practical difficulties associated with allelic complexity and has precluded the development of genetically homogeneous, inbred lines. A diploid ( $2n=2x=14$ ) relative, *Fragaria vesca*, had been used as a model system for strawberry genomics because of its small (~200 Mb) basic genome size, ancestry to the octoploid strawberries, and availability of inbred lines, among other favorable features.

In this study, a cDNA library was constructed using RNA isolated from developing flower buds of *Fragaria vesca* ssp. *vesca* cv. 'Yellow Wonder', an inbred, everbearing variety that is also being used as a mapping parent. A cDNA library constructed using the CLONTECH SMART cDNA Library Construction kit and consisting of 7680 clones was picked (2688 by hand and 4992 robotically)

and was spotted onto high-density filters. Sampling indicated that the library has low redundancy with average insert size between 1-2 kb. Sequencing was performed for 3298 clones at The Hubbard Center for Genome Studies (HCGS) at the University of New Hampshire, The Genome Database for Rosaceae (GDR) at Clemson University Genomics Institute, and Virginia Bioinformatics Institute (VBI) at Virginia Tech. Bioinformatic analyses were performed at Clemson University on the sequences from all three locations, resulting in 2717 quality Expressed Sequence Tags (ESTs) and 1910 unigenes. ESTs appear in GenBank as accession numbers DV438013 – DV440729.

Singlets (individual unique genes) and contigs (assemblies of redundant genes) were used in sequence similarity searches against NCBI Plant proteins, mapped peach ESTs, *Arabidopsis* ESTs, *Populus* ESTs, Rosaceae ESTs, TAIR *Arabidopsis* proteins, and SwissProt databases. Approximately 600 sequences from the *F. vesca* cDNA library did not have homology to any sequence in any of the databases searched. This investigation contributes new library and EST resources for strawberry and Rosaceae genomics.

To utilize some of the ESTs produced by the construction of the cDNA library, a number of genes were selected to be screened for polymorphism in two mapping populations: the intraspecific *Fragaria vesca* ssp. *vesca* cv. 'Yellow Wonder' x *Fragaria vesca* ssp. *vesca* cv. 'Pawtuckaway' and the interspecific *F. vesca* ssp. *vesca* accession FDP815 x *F. nubicola* accession FDP601. Ribosomal protein genes have been chosen for mapping due to several potentially interesting characteristics. The remarkable evolutionary conservation

of ribosomal protein genes has made it possible to use them for complex phylogenetic studies, which can show relationships between plant, animal, and fungal species. A growing interest in ribosomal protein genes in plant genetics has already produced significant data for the model species *Arabidopsis thaliana*, in which 249 ribosomal proteins have been identified and mapped. This provides a useful database for comparison with *F. vesca* ribosomal protein genes.

Genes were selected for evaluation based on several characteristics including the number and size of introns, number of family members in gene family, and sizes of 5' and 3' UTRs (untranslated regions). Primer pair design was accomplished by first comparing each gene sequence to known, orthologous *Arabidopsis* ribosomal protein sequences that are catalogued in Genbank. From these comparisons the likely location of introns were deduced. Once the putative intron sites were located, PCR primer pairs were designed to flank introns in order to amplify them using the Polymerase Chain Reaction (PCR). These PCR products were used to determine if there were allelic differences between parental plants for each respective mapping population. Allelic variants were detected electrophoretically in one of three forms: presence/absence polymorphisms, band mobility polymorphisms, and cleaved amplified polymorphic sequence (CAPS) markers using *AluI*, *BfaI*, and *HaeIII* restriction enzymes. Thirty primer pairs were designed to amplify twenty eight different ribosomal protein genes, resulting in six that could be mapped: RPL23B, RPL32A, RPL37aB, RPL10B, RPL27aC, and RPS14 (three were polymorphic between the intraspecific mapping parents and four were polymorphic between

the interspecific parents). These results suggest that despite the potentially useful features of ribosomal protein genes, they may prove to be unsuitable as mapping markers due to their highly conserved nature.

## OVERVIEW AND OBJECTIVES

Strawberry (*Fragaria*) belongs to one of the world's most economically important plant families: the Rosaceae. The Rosaceae family includes a number of fruit crops such as apple, pear, cherry, almond, apricot, plum, and raspberry, as well as ornamentals such as rose. The commercial importance of the Rosaceae family has contributed to increasing interest in its study, and has prompted numerous genetic investigations. Peach has been advocated as a model organism for Rosaceae family genomics, largely due to its relatively small genome size of ~300 Mb (Jung et al, 2004). However, the diploid strawberry species *Fragaria vesca* has an even smaller genome size of ~200 Mb (extrapolated from Akiyama et al, 2001), and may prove to be an invaluable tool in Rosaceae genomics.

Genetic studies of strawberry are complicated by a variety of factors. Some of these factors include varying numbers of chromosome sets (ploidy level) among species, hybrid origins of some species including the cultivated species, *Fragaria ×ananassa*, and combinations of continuous and discontinuous inheritance patterns (Galletta and Maas, 1990). *F. ×ananassa* is an octoploid that arose in the mid-1700s by hybridization between the wild octoploid species *F. chiloensis* and *F. virginiana* (Hancock, 1990). Diploid ancestors have been valuable in understanding other polyploid species of economic importance such as alfalfa, potato, and wheat (da Silva, 1996). *F. vesca* ( $2n = 2x = 14$ ) is a

presumed wild diploid ancestor of *F. ×ananassa* ( $2n = 8x = 56$ ) (Hancock, 1990), and as such is an appropriate diploid model species for strawberry genomics.

The overall goal of this research is to contribute to development of genomic tools for *Fragaria*, and to further develop the strawberry linkage map using the *F. vesca* model system. This project had the following objectives:

**Project objectives:**

I. Develop an Expressed Sequence Tag (EST) set of 500 or more unigene (non-redundant) sequences.

1) Construct a cDNA library from *F. vesca* flower bud tissue.

2) Generate and analyze EST sequences.

II. Develop selected ESTs as gene-based markers for comparative linkage mapping

3) Identify and evaluate ribosomal protein genes as candidate mapping markers.

4) Compare the extent of marker polymorphism in two diploid mapping populations.

**Rationale**

Expressed Sequence Tags (ESTs) are an important resource in the development of genetic maps. ESTs publically available on GenBank have thus far been insufficient for producing genetic maps in *Fragaria* (Lewers et al, 2005). At the time this project was initiated, fewer than 1500 ESTs were available for *F. vesca* in this database. This lack of EST resources has hindered the development of genetic linkage maps and other genomic tools such as

microarrays. The cDNA library produced from flower bud tissue was expected to contain a diverse array of gene sequences due to the complexities of the tissue and developmental stage, thus making it a good source for EST development and unigene (non-redundant sequence) discovery. As an inbred diploid ancestor of the cultivated strawberry, *F. vesca* is an appropriate and useful model species for elucidating the octoploid genome. It can also serve well as a model plant for the Rosaceae family.

Comparative maps within *Fragaria* will provide a framework within which to better understand the genome of the cultivated octoploid. It is likely that the markers developed from *F. vesca* will be transferable across species as well as across genera of the Rosaceae family, as has already been shown in other studies (Lewers et al, 2005). This investigation contributes new library, EST, and mapping resources that further our understanding of both the strawberry and Rosaceae genomes, and of genes expressed during flower and fruit development. It will also help to assess which of the two available diploid mapping populations is most suitable as the focus for future mapping efforts by revealing in which population polymorphisms are most easily discovered.

The COS (Conserved Orthologue Set) is a set of highly conserved, low or single copy genes that are being used for the development of comparative mapping tools. The SOL Genomics Network at Cornell University created a COS based on genes that are present in both tomato and *Arabidopsis*. Approximately 1000 markers have been developed as comparative mapping tools from COS genes, as described by the SOL Genomics Network

(<http://www.sgn.cornell.edu/index.pl>). The CAPS markers developed in this study that are part of the Conserved Ortholog Set (COS) may also be used not only in the Rosaceae family, but also in more distantly related species, especially due to the high conservation of the ribosomal protein genes that became the primary focus of this investigation.

Ribosomal protein genes have been chosen for mapping due to several potentially interesting characteristics. There are somewhere between 60-80 different proteins that make up the cytosolic ribosome along with four ribosomal RNA molecules (Moran, 2000). These ribosomal proteins can be distinguished from each other by their amino acid sequences as deduced from the sequences of cDNAs or PCR analysis (Wu et al, 1995). The remarkable evolutionary conservation of orthologous ribosomal protein genes across wide taxonomic distances (e.g., between *Arabidopsis* and mouse) has made it possible to use them for complex phylogenetic studies, which can show relationships between plant, animal, and fungal species (Barakat et al, 2001). This conservation also makes ribosomal protein genes excellent candidates for membership in the Conserved Ortholog Set (COS). The strawberry ribosomal protein gene sequences identified in the present study are expected to be broadly useful for mapping in the *Rosaceae* family. A growing interest in ribosomal protein genes in plant genetics has already produced significant data for the model species *Arabidopsis thaliana*, in which 249 ribosomal proteins have been identified and mapped (Barakat et al, 2001). This provides a useful database to which *F. vesca* ribosomal protein genes can be compared as a means of identification.



## CHAPTER I

### LITERATURE REVIEW

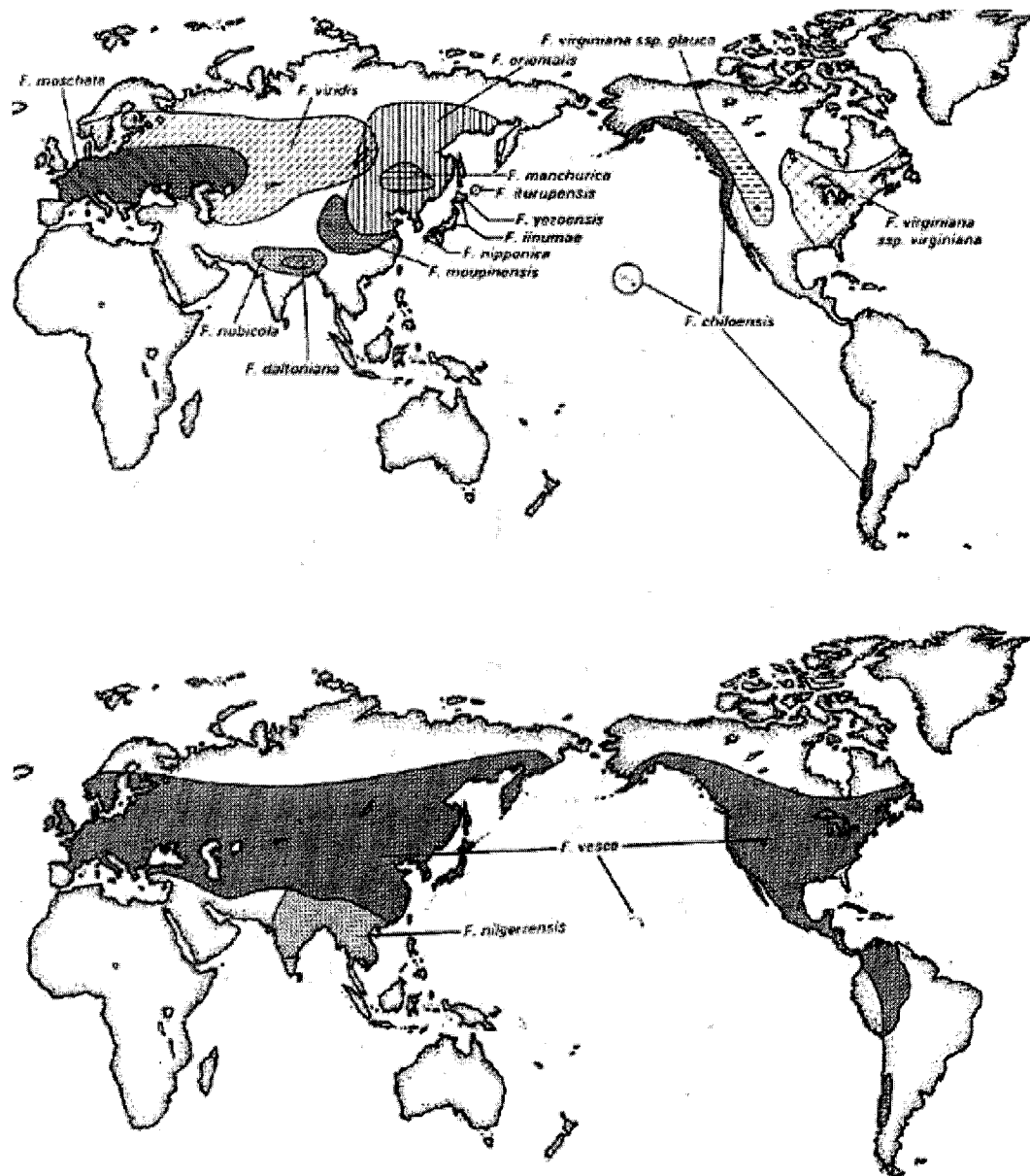
#### The Cultivated Strawberry

Strawberry is a non-climacteric fruit belonging to the genus *Fragaria*, which is part of the commercially important plant family the Rosaceae. The Rosaceae family includes a number of fruit crops such as apple, pear, cherry, almond, apricot, plum, and raspberry, as well as ornamentals such as rose, flowering cherry, crab-apple, and quince (Jung et al, 2004). The United States leads the world in strawberry production, with more than 80% coming from California (Hancock and Luby 1993). While apple grosses the highest fruit crop income, followed by grape and citrus crops, the cultivated strawberry, *Fragaria ×ananassa*, ranks fourth on the list of United States fruit production values (USDA, 2001). In recent years, the total value of US production has exceeded 1 billion dollars (ERS, 2005). During the 2004 production year, 20,880 Ha of land in the US were harvested for strawberries, yielding 1,004,110 Mt of berries (FAO, 2005). This makes the strawberry of considerable interest to growers and breeders alike. The strawberry breeding program initiated in 1920 by George M. Darrow is the longest continuous program of the United States Department of Agriculture (Lawrence et al, 1990).

There are 23 species of *Fragaria* that have been recognized through the careful analysis of over 500 accessions growing in the wild and/or documented in

herbariums throughout the world (Staudt, 1989; Hancock, 1999). *Fragaria* inhabits a diverse array of environments ranging from the entire northern hemisphere (*F. vesca*), Europe and central Asia (*F. viridis*), the Himalaya Mountain region (*F. nubicola*), eastern Tibet and southwestern China (*F. moupinensis*), Japan (*F. iinumae*) and Europe (*F. moschata*) (Evans, 1982) (Figure 1, Table 1). The 13 diploid species are the most diverse and occupy the greatest scope of niches (Hancock and Luby, 1993). The ability of these species to colonize such an array of climates and environments has been attributed to their characteristic of being “specialists” that are complex and well adapted to particular environments (Hancock and Bringhurst, 1978). These diploid species are of much interest to the strawberry community because some of them, including *Fragaria vesca*, are believed to be the wild ancestors of the octoploid, cultivated species.

Cultivated strawberry, *F. ×ananassa*, was first described by French botanist Duchesne in 1766 during his work at the Royal Gardens of Versailles (Johnson, 1990). The plant arose as an accidental hybridization between *F. chiloensis* and *F. virginiana*, which had found their ways to France by various routes more than 250 years ago (Hancock and Luby, 1993). Both species evolved as ancient polyploids through events of natural selection (Bringhurst, 1990). The principal cultivated species, *F. ×ananassa*, is an octoploid resulting at least indirectly from human intervention.



**Figure 1.** Hancock and Luby (1993) map showing the global distribution of *Fragaria* species.

**Table 1.** The *Fragaria* species, their ploidies and geographic distributions (Folta and Davis, 2006).

<b>Species</b>	<b>Ploidy</b>	<b>Primary Natural Geographic Range</b>
<i>F. vesca</i>	2x	Northern hemisphere
<i>ssp. vesca</i>		Europe to Siberia as far as Lake Baikal
<i>ssp. bracteata</i>		North America, Rocky Mountains westward
<i>ssp. americana</i>		North America, mainly east of Rocky Mountains
<i>ssp. californica</i>		California, Southern Oregon coast
<i>F. viridis</i>	2x	Europe to Siberia as far as Lake Baikal
<i>F. yezoensis</i>	2x	Japan
<i>F. nipponica</i>	2x	Japan
<i>F. nubicola</i>	2x	Eastern Himalayan region
<i>F. bucharica</i>	2x	Western Himalayan region
<i>F. daltoniana</i>	2x	Eastern Himalayan region
<i>F. nilgerrensis</i>	2x	Central Asia into China
<i>F. mandshurica</i>	2x	Northeastern Asia
<i>F. pentaphylla</i>	2x	Southwest china, Himalayan region
<i>F. gracilis</i>	2x	Northwest China
<i>F. iinumae</i>	2x	southern and central Sahalin, Russia. Japan,
<i>F. ×bifera</i>	2x, 3x	Europe, hybrid between <i>F. vesca</i> and <i>F. viridis</i>
<i>F. corymbosa</i>	4x	Northern China
<i>F. moupinensis</i>	4x	Southwest China
<i>F. orientalis</i>	4x	Northeastern Asia
<i>F. tibetica</i>	4x	Eastern Himalayan region
<i>F. moschata</i>	6x	Europe, eastward to Ural Mountains to Lake Baikal
<i>F. chiloensis</i>	8x	Western N. and S. America, Hawaii
<i>F. virginiana</i>	8x	North America
<i>F. iturupensis</i>	8x	Iturup Island (Kurile Islands)
<i>F. ×ananassa</i>	8x	Widely cultivated
<i>F. ×bringhurstii</i>	5x, 6x, 9x	California coast, hybrid between <i>F. vesca</i> and <i>F. chiloensis</i>

## Polyploid Origin

Polyploid plant species, having more than two genome sets, have been important to the evolution and hardiness of many species by stabilizing hybrid genome composition (de Wet, 1971). It is believed that recent and/or ancient polyploidy may be a feature of over 70% of angiosperm species, even those with very small genomes such as *Arabidopsis thaliana* (Gottlieb, 2003). There are many cultivated polyploid species, including sugarcane (8x-18x), potato (4x), alfalfa (4x), banana (3x), apple (3x), and many ornamentals (Paterson, 2005).

There have been a number of hypotheses developed as to the constitution of the octoploid strawberry genome. The first was put forth by Federova (1946) giving the formula AABB<sup>4</sup>BCC. This was accepted for some time until cytological evidence from studies by Senanayake and Bringhurst (1967) revised it to AAA'A'BBBB. The former model differs from the latter in that the C genome would have been contributed from a distinctly different *Fragaria* species, such that no meiotic pairing would have been expected between the C genome chromosomes and those of the B or A genomes. In contrast, replacing C with A' implies that the A and A' genomes are from two very closely related sources, and that some degree of meiotic pairing between them might be possible. Both models postulate that there are four copies of the B genome, thereby implying the genes in this genome would be inherited in a tetrasomic (as opposed to disomic) manner. The current genome constitution model was published by Bringhurst (1990), in which B versus B' genome types are distinguished, suggesting a fully diploidized AAA'A'BBB'B' structure. This model was based on

cytological evidence of bivalent chromosome pairing, and genetic evidence only of disomic inheritance patterns (Bringhurst, 1990). The early cytological studies showing exclusive bivalent chromosome pairing in the octoploid, cultivated octoploid were interpreted as an indication of diploidization (Byrne and Jelenkovic, 1976). However more recent studies have demonstrated that there may be mixed patterns of inheritance, both poly- and disomic, which further complicates the issue by suggesting that the octoploid genome may not be as highly diploidized as proposed by Bringhurst's genome model (Lerceteau-Köhler et al, 2003).

Polyploidy is the result of chromosome doubling in somatic cells or the fusion of unreduced gametes (de Wet, 1971). This chromosome doubling can result in allopolyploidy, which involves two or more chromosome sets that are genetically distinct, or autopolyploidy, in which multiple chromosome sets from the same species are involved. The octoploid genome strawberry would likely have arisen through multiple polyploidy events, with ancestors existing at least transiently at intermediate ploidy levels.

*Fragaria* chromosomes are small in size, and their numbers occur in a polyploidy series, based on  $x = 7$ , of  $2n = 14, 28, 35, 42,$  and  $56$  (Iwatsubo and Naruhashi, 1989). The intermediate polyploids may provide an important evolutionary link between the diploids and octoploids, as noted by Bringhurst and Gill (1970). Many of these polyploids, including pentaploid hybrids, are aggressive and hardy. This hardiness was also mentioned in earlier literature by Darrow (1950) who pointed out that the hexaploid was stronger and sturdier than

the diploid, and the octoploid even more so than the hexaploid. This gave rise to the possibility of breeding desirable diploid traits into polyploidy intermediates to produce octoploids with higher vigor, greater drought and heat resistance, sweeter more fragrant fruit, and other beneficial qualities (Darrow, 1950). A study by Hancock and Bringhurst (1981) showed that this was in fact the case and that octoploids were found occurring over broader ranges of climate, soil types, soil pH, salinity, organic carbon content, and texture. Desirable diploid traits could be bred into more vigorous octoploids through the use of synthetic octoploids produced by breeding hybrid hexaploid x diploid crosses with octoploid cultivars or through the combination of two diploids and a tetraploid crossed with an octoploid (Evans, 1982; Darrow, 1966).

### ***Fragaria vesca* Genomic Resources**

Despite the debate over genome constitution, it is generally accepted that diploid ancestors must have contributed genomes to the octoploid. This was first suggested in 1926 by Longley who also proposed that *F. vesca* ( $2n = 2x = 14$ ) was the most primitive of these original diploids (Bringhurst and Khan, 1963). *F. vesca*, the common woodland strawberry, stands erect, can be runnering or non-runnering, and has a "soft, aromatic, hemispherical receptacle with raised seeds" (Darrow, 1966). There are four subspecies of *F. vesca* (ssp. *vesca*, *americana*, *californica*, and *bracteata*), and this species is found throughout the northern hemisphere (Staudt, 1989) in a variety of climates including temperate, grasslands, Mediterranean, and subtropical (Hancock, 1990).

Genetic studies of strawberry have been hampered by several complicating factors: varying ploidy levels, hybrid origins, combinations of inheritance patterns that are continuous (discrete or Mendelian) and discontinuous (quantitative or multifactorial) within the one individual (Galleta and Maas, 1990). Many of these problems can be avoided by the use of the morphologically diverse, highly interfertile diploid species (Sargent et al, 2004). *F. vesca* provides an excellent resource for insight into the genome of the cultivated octoploid. Diploid ancestors have been valuable in understanding other polyploid species of economic importance such as alfalfa, potato, and wheat (da Silva, 1996).

Diploid strawberry may serve not only as a model system for the octoploid strawberry, but also more broadly for the Rosaceae family (Sargent et al, 2004). This is partly due to its small genome size of approximately 200 Mb (extrapolated from Akiyama et al, 2001), only about 25% larger than that of plant model species *Arabidopsis thaliana*. Peach, having a genome size about twice that of *Arabidopsis*, near 300 Mb, has also been advocated as a Rosaceae model species (Jung et al, 2004). There have also been studies, including that done by Lewers et al (2005) which demonstrate that peach molecular markers may not be transferable to other genera, such as *Fragaria* and *Rubus*, as originally hoped for in this proposed model species. In the same study, it was found that markers designed in one *Fragaria* species were generally able to amplify DNA from all *Fragaria* species tested as well as DNA in the *Rubus* genus (blackberry and raspberry) (Lewers et al, 2005). Like *Fragaria*, *Rubus* belongs to the Rosaceae



subfamily Rosoideae. Such findings give weight to the idea that more effort and resources should be devoted to the development of genomic resources for *Fragaria*.

While structural genomics resources available for peach are extensive, there has been little public sequence information in databases for strawberry. This has remained true despite the development of the Genome Database for Rosaceae (GDR) at Clemson University, which has created a comprehensive web-based resource for rosaceous species (Folta et al, 2005). The lack of available sequence data has slowed progress for projects such as gene expression surveys and marker development in *Fragaria* (Folta et al, 2005). Prior to the resources contributed by this thesis investigation, only about 1500 Expressed Sequence Tags (ESTs) were available for *F. vesca* in Genbank (NCBI, 2005). The number of sequences available, such as those with Simple Sequence Repeats (SSRs), has been insufficient for producing well-saturated genetic maps in *Fragaria* (Lewers et al, 2005).

It has been difficult to identify genetic linkages in the cultivated strawberry due to its octoploid nature (Hadonou et al, 2004). This difficulty is apparent in all polyploids as they have a larger number of segregating genotypes, electrophoretic co-migration of DNA fragments produced by restriction digestion and/or PCR, and complicated, sometimes poorly characterized, genomes (da Silva, 1996). Linkage maps are important in genomic study of a species, as they provide a framework for genetic description (Davis and Yu, 1997). Genetic maps in crop species can form the “roadmap” for navigating the genome and provide a

vast amount of resources to breeders (Allen, 1994). Diploid linkage relationships may be important in resolution of the octoploid map in *Fragaria*.

There are currently linkage maps available for *F. vesca*, though considerable work still needs to be done with them. The first map, produced by Davis and Yu (1997), was developed primarily using Randomly Amplified Polymorphic DNA (RAPD) markers with an intraspecific *F. vesca* population (Figure 2). However, there are problems with RAPD marker-based maps, because RAPD markers typically have a dominant character, banding patterns can be difficult to interpret and reproduce, and RAPD markers cannot always be transferred between species or mapping populations (Hadonou et al, 2004). A second generation diploid map was produced by Sargent and colleagues (2004) at East Malling Research (UK) using Simple Sequence Repeats (SSRs) genotyped in an interspecific population from *F. vesca* and *F. nubicola* (Figure 3). However this map also has some drawbacks. Significant segregation distortion was observed, with high gene clustering at a high proportion of loci, likely due to the interspecific progeny used (Sargent et al, 2004). In order to resolve the difficulties with the current linkage maps, markers will need to be developed for both maps using different mapping techniques to produce a map with easily interpretable, reproducible markers that do not cluster but are well-distributed on the map.

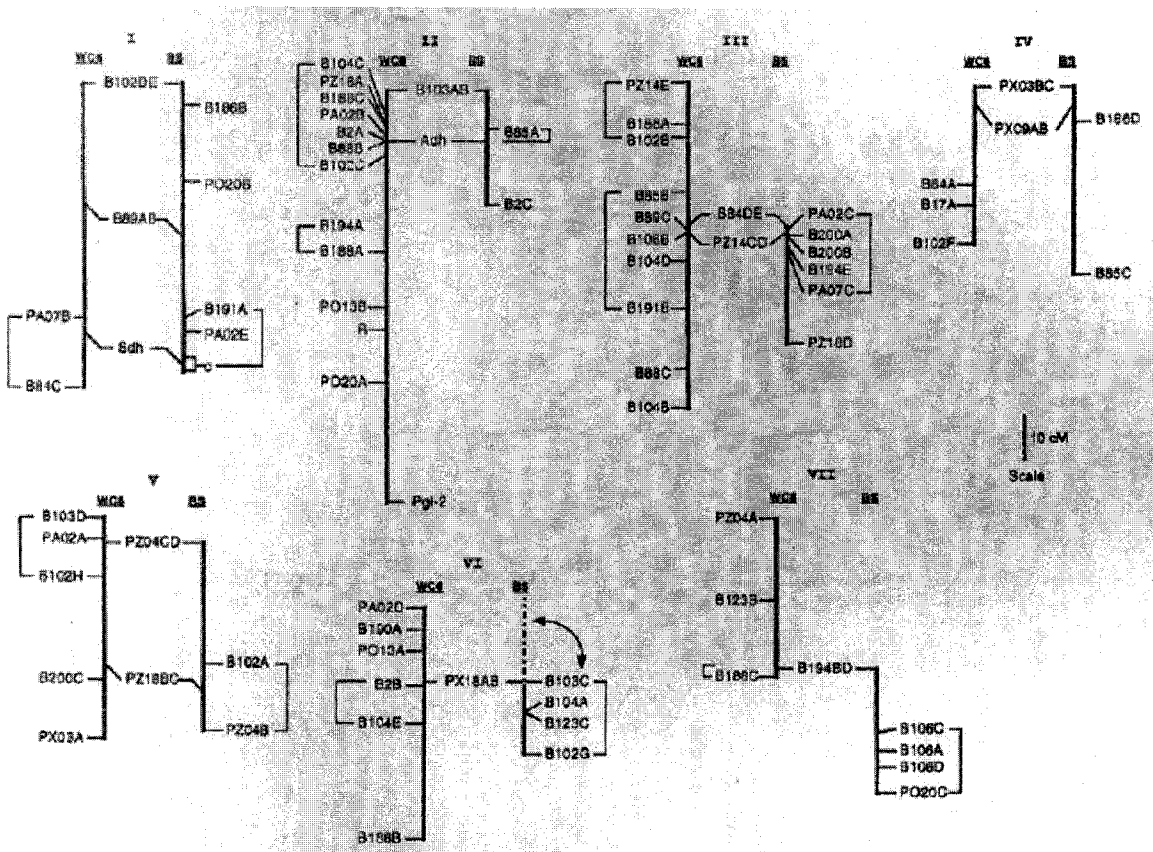
## Complementary DNA (cDNA) Libraries

Complementary DNA, or cDNA, libraries are a useful resource for genomic analysis as they are constructed from messenger RNA (mRNA) and therefore reflect the gene expression in the tissue from which the library was constructed, in contrast to genomic libraries which contain genetic material that may or may not be expressed (Ying, 2004). The spectrum of gene expression in a tissue and/or developmental stage provides insight into regulation of genes during various developmental processes as well as what DNA sequences are in fact functional at all. These sequences are important because those that are being expressed give characteristics to the organism (Sasaki et al, 1994).

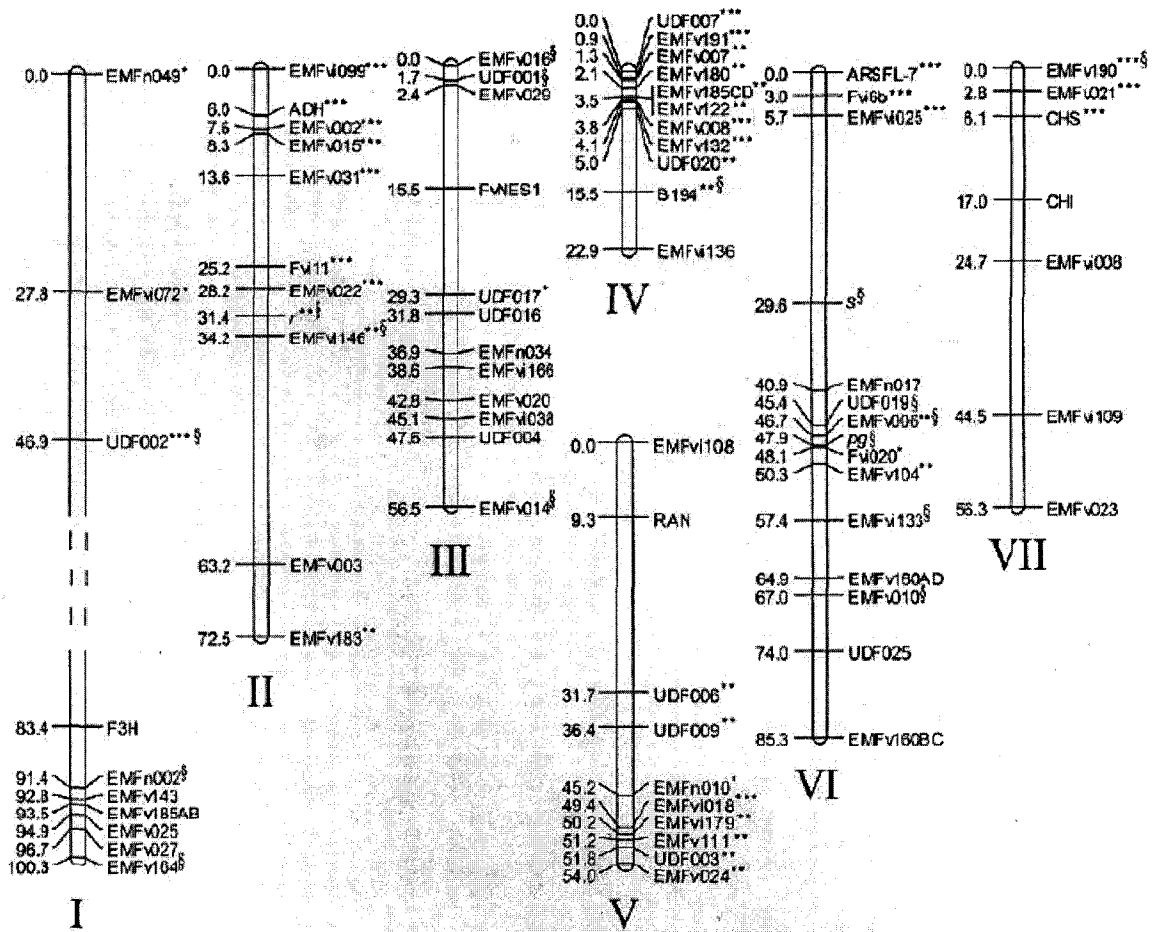
Diploid strawberry (*F. vesca*) has been an attractive model system for cDNA generation due to its small genome, short reproductive cycle, and ability to be transformed (Nam et al, 1999). However, difficulty has been encountered in cDNA library construction due to high levels of phenolic compounds and polysaccharides in strawberry tissues (Manning, 1998). While some studies have successfully produced strawberry Expressed Sequence Tags (ESTs), there are still relatively few available in public databases and none have been produced using developing flower buds as the source of genetic material. The generation of full-length cDNAs for specific gene sequences is essential for the molecular study of any organism, including monitoring large numbers of genes, function of over-expressed genes, altered gene expression in response to various cues, and other applications (Ying, 2004). Libraries made from cDNA

are also invaluable tools in the generation of large numbers of novel genes (Sasaki et al, 1994), which is much needed in the field of *Fragaria* genomics.

**Figure 2.** *F. vesca* linkage map for intraspecific mapping population 'Yellow Wonder' x 'Pawtuckaway' produced using RAPD markers (Davis and Yu, 1997)



**Figure 3.** Linkage map for interspecific *F. vesca* x *F. nubicola* hybrid progeny produced using SSR markers in FDP815 x FDP601 mapping population (Sargent et al, 2004).



## **Polymerase Chain Reaction (PCR) Based Mapping**

Since its discovery in the early 80s by Kary Mullis, the Polymerase Chain Reaction (PCR) has become one of the most widely applied molecular techniques in the world. While PCR has been adapted for many uses, its most basic function is the exponential accumulation of a target DNA sequence through thermal cycles of denaturation, hybridization, and polymerase extension (Mullis et al, 1986). The resulting product for one double-stranded DNA molecule will be approximately  $2^n$ , where n equals the number of cycles used (Saiki et al, 1988). DNA is replicated by the incorporation of synthetic oligonucleotides that prime polymerase extension products by hybridizing to template DNA (Mullis et al, 1986). The reaction is stabilized through the utilization of a thermostable DNA polymerase (Taq) isolated from *Thermus aquaticus*, allowing for improved specificity and yield by increased temperatures (Saiki et al, 1988). PCR revolutionized the field of molecular biology and paved the way for modern DNA technology.

One requirement of the PCR reaction is to know the target area of DNA to be synthesized in order to construct oligonucleotides that will hybridize to template strands (Mullis and Faloona, 1987). This has become increasingly easy due to the availability of DNA sequences in public databases. The national database GenBank has a large number of Expressed Sequence Tags (ESTs) generated from cDNAs whose putative functions can be elucidated using the Basic Local Alignment Search Tool (BLAST) (Chee et al, 2004). In conjunction with the use of these accessible sequences, techniques based on PCR are now

being used in addition to or in place of traditional genetic analysis because PCR markers can be scored using a small sample without relying on costly or labor-intensive methods (Konieczny and Ausubel, 1993). Markers produced from PCR have proven to be use useful in qualitative and quantitative trait mapping, identification of germplasm, and in marker-assisted selection (Frary et al, 2005).

There are various PCR-based techniques used for mapping including the following: Randomly Amplified Polymorphic DNA (RAPD), Amplified Fragment Length Polymorphism (AFLP), and Sequence-Tagged Site (STS) such as Simple Sequence Repeat (SSR), (Chee et al, 2004). Methods such as RAPD and AFLP do not require prior knowledge of DNA sequence and can be employed with with any DNA sample, while STS methods target indels (insertions or deletions) found in known intron length polymorphisms. These different applications make it possible to rapidly screen libraries and allows for recovery of full length DNA from cDNA libraries (You and Scholl, 1998). Markers generated in this way have the potential to be used across species as well, since recent studies have found high levels of amino acid similarity between many genes, with conserved intron position and size (Chee et al, 2004). A recent study found that a high percentage of PCR-based markers for comparative mapping in wheat could also be used in rice (Yu et al, 2004). This lends promise to the possibility that developing EST-derived PCR-based mapping markers in model ancestral species may prove to be a valuable resource among plant families.



## **Cleaved Amplified Polymorphic Sequence (CAPS)**

Cleaved Amplified Polymorphic Sequence (CAPS) analysis is a PCR-based version of the Restriction Fragment Length Polymorphism (RFLP) technique that can be used to identify Single Nucleotide Polymorphisms (SNPs) or insertion/deletion (indel) polymorphisms within a specifically targeted DNA fragment by using gene-specific DNA primers for PCR followed by restriction enzyme digestion (Neff et al, 1998). Digestion allows for the identification of the gain or loss of a restriction site within the fragment due to the presence of a SNP, or to an alteration in fragment size due to an indel. CAPS has become the most widely used approach for detection of SNPs because previous methods (such as RAPD, AFLP, and SSR) cannot be utilized at low cost during smaller investigations and have some problems with reproducibility, and thus use has been somewhat restricted (Konovalov et al, 2005). Researchers previously had to choose methodology for SNP detection based on many factors including the number of samples, the cost to analyze each sample, and the amount of automation to be employed (Morales et al, 2004).

CAPS, on the other hand, has proved to rapidly and reliably detect SNPs and indels in a cost-effective and easy to use manner (Micheals and Amasino, 1998) without the use of time-consuming blotting procedures or the use of radioactivity (Jarvis et al, 1994). Studies have shown that CAPS markers are reproducible using different DNA isolation methods, different plant organs, when performed by different researchers, and when samples are of varying origins (Kunihisa et al, 2005). CAPS markers are also desirable as candidates for use

across species. This has been accomplished already in a study using pine species from different plant families for the timber industry in Japan (Matsumoto, 2004). Studies using CAPS techniques have also been performed using pea (Konovalov et al, 2005), *Arabidopsis* (Jarvis et al, 1994), melon (Morales et al, 2003) and octoploid strawberry (Kunihisa et al, 2005), among other plant species. If the respective PCR primers prove to be transferable across taxonomic borders, they will be even more useful by cutting out the need for other repetitive screening processes such a library construction, cloning and sequencing, primer design, etc. (Matsumoto, 2004).

Marker methods designed for SNP identification have been most successful when targeted to Untranslated Regions (UTRs) or locations where introns are known or assumed to be, as these areas may be more highly polymorphic than coding regions (Morales et al, 2004). The ability to map these sites is another benefit in using CAPS markers because previous methods may have been targeting less desirable areas of the genome, such as SSRs in repetitive regions located near centromeres and originating due to replication slippage and unequal crossing over during meiosis (Frary et al, 2005). Reliance on SSRs can result in distorted maps with high amounts of gene clustering, whereas this is generally not observed using CAPS methods.

Thus far, CAPS markers have been used in octoploid strawberry (*Fragaria ×ananassa*) for the use of cultivar identification in the marketplace (Kunihisa et al, 2005). There is little, if any, research done investigating the use of CAPS for gene mapping in more primitive diploid species, such as *Fragaria vesca*.

Development of markers based on diploid species may be of particular importance because it has been difficult to produce genome-specific markers in *F. xananassa* due to continued debate over its genome constitution and polyploid origin (Kunihisa et al, 2005).

## **Ribosomal Protein Genes**

Ribosomal protein genes are required for the biosynthesis of ribosomes in eukaryotic organisms (Wu et al, 1995). Ribosomes play an essential role in protein synthesis and contribute to control of the cell cycle (Barakat et al, 2001). In eukaryotes, it has been estimated that each cytosolic ribosome is constituted of somewhere between 60 and 80 different small, positively charged ribosomal proteins, along with four ribosomal RNAs (Moran, 2000). There is evidence that indicates a coordinated synthesis of the ribosomal RNAs and the ribosomal proteins, seen most notably in developing tissues where ribosomes are utilized in large numbers for protein synthesis (Beltrán-Peña et al, 1995). Ribosomal proteins have been estimated to represent approximately 15% of total cellular proteins (Joanin et al, 1993), and their synthesis is regulated at both the transcriptional and post-transcriptional levels (Kim et al, 1990).

Plants contain three distinct types of ribosomes: those that are located within the cytosol, the mitochondria, and the plastids (Gnatt and Thompson, 1990). Though there is little similarity between these divergent ribosome types, homology across taxonomic groups has been identified for all plant cytosolic ribosomal proteins studied (Gualerzi et al, 1974). They have been isolated and

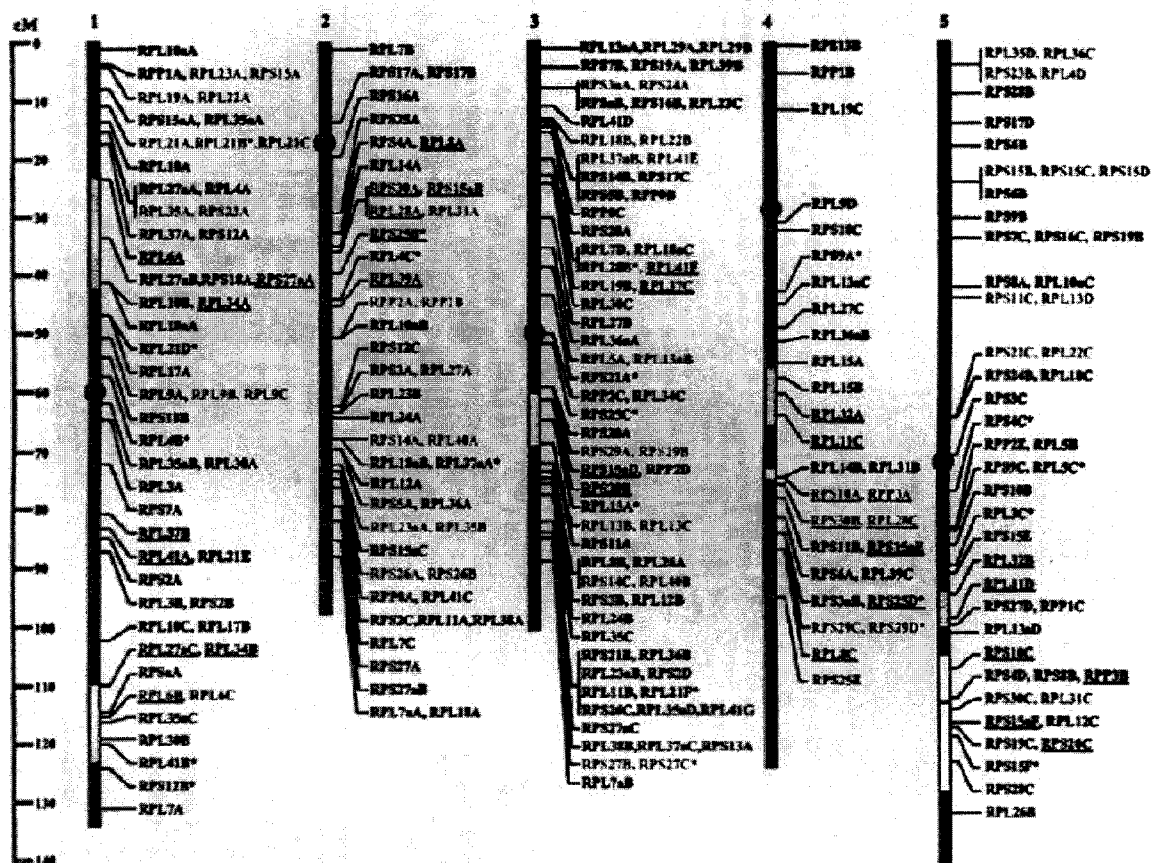
identified using deduced amino acid sequences from the cloning and PCR analysis of cDNAs (Wu et al, 1995). Cytosolic ribosomal proteins are of particular interest because they are remarkably well conserved between plant, animal, and fungal species, allowing them to be used in complex phylogenetic studies (Barakat et al, 2001).

Another interesting characteristic of these proteins that they are almost all found in two or more copies within gene families (Moran, 2000). Ribosomal proteins are designated RPL or RPS depending on whether they correspond to the large or small ribosomal subunit respectively and are given a number. If a particular ribosomal protein gene exists in a gene family, members are distinguished from one another by a capital letter. For example RPL12B would be ribosomal protein gene 12 from the large subunit and it would be family member "B". While more than three-quarters are believed to have more than one copy (Barakat et al, 2001), it appears as though only one copy of each gene is present per ribosome, suggesting that family members may form functionally distinct ribosomes (Kim et al, 1990). It has been discovered that members of a ribosomal protein gene family have nearly identical amino acid sequences, but in contrast, the untranslated 5' and 3' mRNA regions (UTRs) are highly divergent (Cooke et al, 1997). This makes ribosomal protein genes attractive for mapping, as polymorphisms in UTRs or introns could be utilized as targets for gene-specific primers, and the high conservation of amino acid coding regions as potential PCR primer sites may contribute to broad usage across species. These attributes also make ribosomal protein genes candidates for the Conserved

Ortholog Set (COS) with the SOL Genome Network at Cornell University. As initially conceived and defined, COS genes are low copy number genes that are present in both *Arabidopsis* and tomato and may prove to be useful markers for comparative mapping projects. The conserved ortholog concept is now more broadly defined and understood as describing single or low copy number genes for which the coding regions are well-conserved across broad taxonomic/phylogenetic distances.

A growing interest in ribosomal protein genes in plant genetics has already produced significant data for the model species *Arabidopsis thaliana*, in which 249 ribosomal proteins have been identified and mapped (Barakat et al, 2001) (Figure 4). This has provided an excellent resource for comparative studies and the development of tools such as PCR primers in other species. Recent studies have shown *Arabidopsis* RPL3 to have a 73% similarity to yeast RPL3 (Kim et al, 1990), pea RPL9 to have an 80% similarity to *Arabidopsis* RPL9 and 76% similarity to rice RPL9 (Moran, 2000), and RPS13 to have 73% similarity to rat RPS13 (Joanin et al, 1993). These similarities are encouraging in the pursuit to utilize ribosomal protein genes for mapping, as well as other resource development. There is still much to learn about them, their function, synthesis, distribution, regulation, and gene family structures, as relatively little is known about them in comparison to other well characterized plant gene categories.

**Figure 4.** The 249 ribosomal protein genes mapped in *Arabidopsis thaliana* (Barakat et al, 2001)



## CHAPTER II

### METHODS

#### Plant Material

The plant tissue used for cDNA library construction was taken from *Fragaria vesca* ssp. *vesca* cv. 'Yellow Wonder' (YW). Seeds of this inbred, ever-bearing variety were obtained by Scott Williamson (1995) from W. Atlee Burpee and Company (Warminster, PA). This cultivar was crossed with *F. vesca* ssp. *vesca* cv. 'Pawtuckaway' (PAWT) to produce an F<sub>1</sub> population and subsequent F<sub>2</sub> mapping population of 113 plants. The original Pawtuckaway plant was collected in the wild from Pawtuckaway State Park in Nottingham, NH by Tom Davis and Scott Williamson (1995). Plants have been maintained in greenhouses at the University of New Hampshire.

A second mapping population was also used in this study. *F. vesca* spp. *vesca* accession FDP815 and *F. nubicola* accession FDP601 were crossed, resulting in an F<sub>1</sub> population which was then used to produce an F<sub>2</sub> segregating population of 94 plants (Sargent et al, 2004). These populations were created and maintained by Dan Sargent at East Malling Research, UK. DNA was isolated at East Malling Research and shipped to the University of New Hampshire for use in this study.

## Total RNA Isolation

Developing flower buds were harvested from YW plants several times a week for a five month period. Buds were collected at their fullest size, just before opening, and were kept on ice while they were being worked with to slow potential RNase activity. Sepals were dissected away from inner tissue to remove any potential contaminants (fungi, bacteria, mites or other insects, etc). The buds, with sepals removed, were then weighed as a group, frozen in liquid nitrogen, and stored at -80°C.

Total RNA was isolated using an Ambion RNAqueous Midi Kit (Austin, TX) as described by the manufacturer. With this kit, RNA was extracted by forcing homogenized tissue, in a lysis/binding solution containing chaotropic salts, through glass fiber filters using a syringe. Total RNA was bound to the filters and then washed several times with wash solutions. An elution solution (hot distilled water with EDTA to chelate heavy metals) was used to remove RNA from the filter. For each isolation, approximately 0.25 grams of frozen bud tissue was used. This procedure was repeated until RNA had been extracted from a total of 7.0 grams of tissue. Elutions resulted in final volumes of 1.5 mL, and then liquid was evaporated off using a speed vacuum leaving a dried RNA pellet. The dried RNA pellet from the first sample tube was resuspended in 250  $\mu$ l RNase-free H<sub>2</sub>O. This same 250  $\mu$ l H<sub>2</sub>O was then pipetted into the second tube and the second RNA pellet was resuspended. This process was continued until all the RNA pellets from all samples had been suspended in the same 250  $\mu$ l. This concentrated sample was then used for polyadenylated (polyA) mRNA isolation.



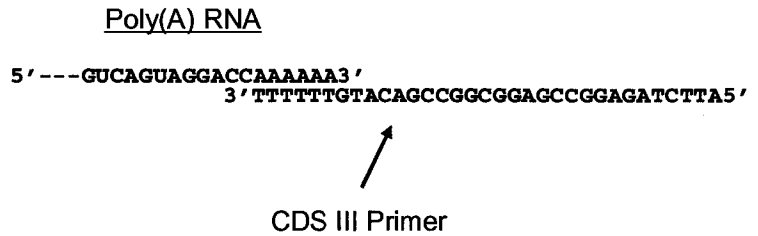
## **Poly(A) mRNA Isolation**

A MicroPoly(A) Purist kit, also from Ambion (Austin, TX), was used to extract mRNA from the concentrated total RNA sample. This kit works by targeting the poly(A) tail on the mRNA with Oligo(dT) Cellulose. Once bound to the Oligo(dT) Cellulose, mRNA was treated with wash solutions, transferred to a spin column, and eluted in an RNA storage solution. The entire 250  $\mu$ l sample of total RNA was used in the MicroPoly(A) Purist kit, resulting in 200  $\mu$ l of aqueous mRNA. An overnight ethanol precipitation was performed, and mRNA was resuspended in 40  $\mu$ l of RNase-free H<sub>2</sub>O. This sample was used to synthesize cDNA and construct the library.

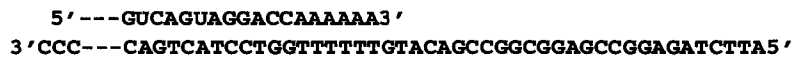
## **cDNA Library Construction**

A cDNA library was constructed by Darren Bauer at the Hubbard Center for Genome Studies (HCGS) at the University of New Hampshire using a CLONTECH SMART cDNA Library kit (Mountain View, CA). This kit was used for cDNA synthesis (Figure 5) and cloning with a pDNR-LIB vector (Figure 6). Figure 5 shows the actual primer and vector sequences along with *Sfi*I sites used for directional cloning. The resulting clones were picked either robotically or by hand. Enough clones were picked to fill thirteen 384-well plates, comprising a total of 4992 ordered clones. Those clones that were picked into plates robotically were also spotted in duplicate onto nylon hybridization filters at the HCGS.

**Figure 5.** cDNA synthesis using CLONTECH SMART cDNA Library kit. Dashes (---) are used to represent the beginning of an mRNA sequence

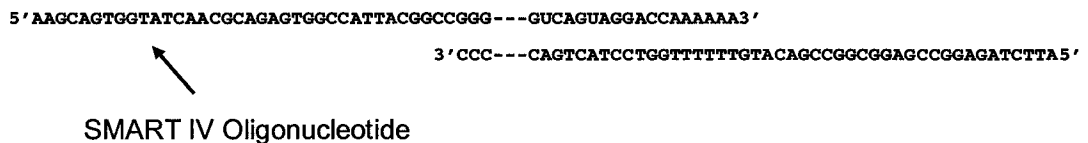


First-strand synthesis and dC tailing by PowerScript™ reverse transcriptase

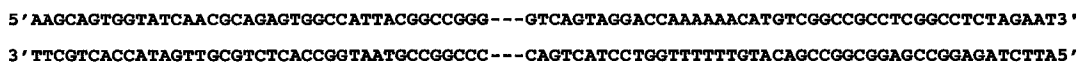


Template switching and extension by reverse transcriptase

\*mRNA will be removed during primer extension to allow second strand cDNA synthesis



Primer extension to produce double-stranded cDNA



**Figure 5. (cont.)** cDNA synthesis using CLONTECH SMART cDNA Library kit. Dashes (---) are used to represent the beginning of an mRNA sequence

Digested with *Sfi* restriction enzyme

5' CGGCCGGG---GTCAGTAGGACCAAAAACATGTCGGCCGCCT 3'  
 3' AATGCCGGCC---CAGTCATCCTGGTTTTTTGTACAGCCGGC 5'

Ligated into vector digested with *Sfi*

5' CGGCCGGG---GTCAGTAGGACCAAAAACATGTCGGCCGCCT 3'  
 3' AATGCCGGCC---CAGTCATCCTGGTTTTTTGTACAGCCGGC 5'

ACCGGACATATGCCCGGGAATTCGGCCATTA  
 TGGCCTGTATACGGCCCTTAAGCCGGT

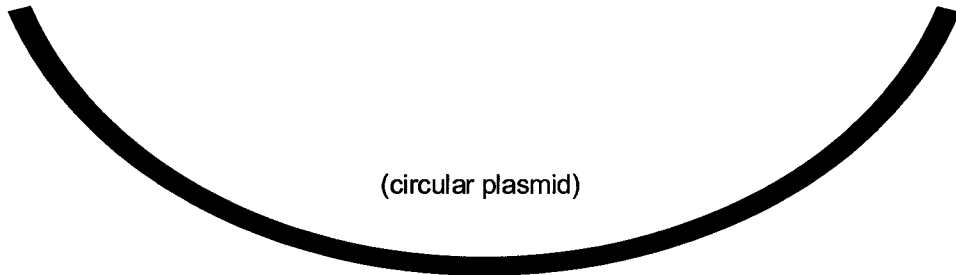
(vector)

CGGCCTGCAGGAT  
 GGAGCCGGACGTCCTA

(vector)

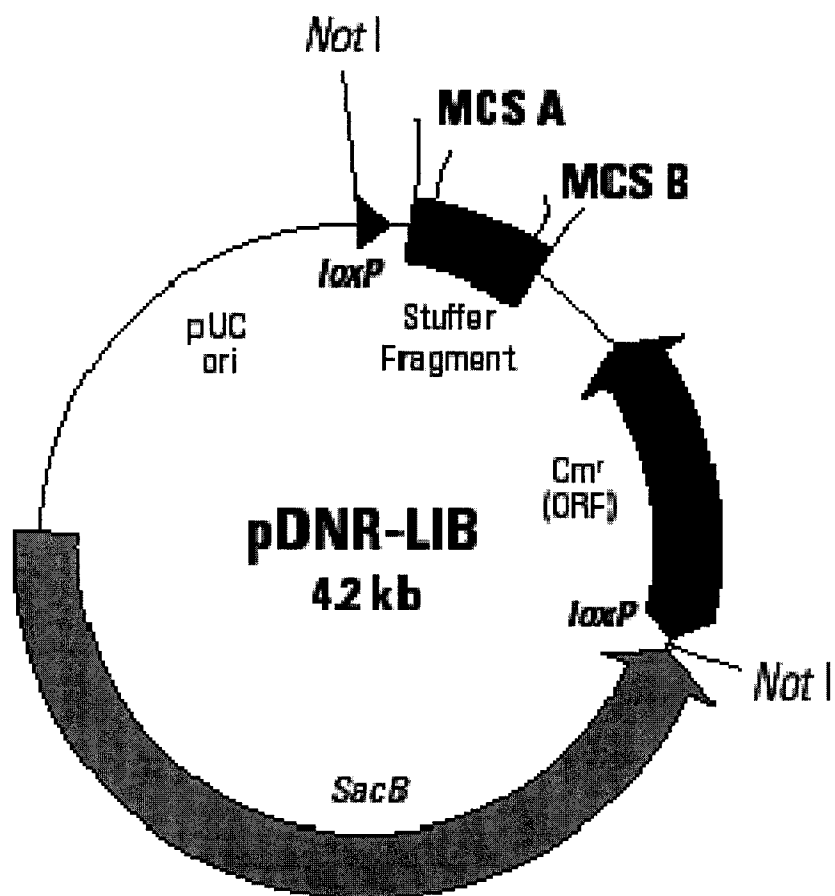


ACCGGACATATGCCCGGGAATTCGGCCATTACGGCCGGG---GTCAGTAGGACCAAAAACATGTCGGCCGCCTCGGCCTGCAGGAT  
 TGGCCTGTATACGGCCCTTAAGCCGGTAATGCCGGCC---CAGTCATCCTGGTTTTTTGTACAGCCGGCGGAGCCGGACGTCCTA



(circular plasmid)

**Figure 6.** pDNR-LIB vector used for directional cloning with CLONTECH SMART cDNA Library kit.



## **Clone Evaluation**

Vector insert sizes were evaluated using Polymerase Chain Reaction (PCR) with clone colonies as the DNA template (i.e., by "colony PCR"). An Eppendorf PCR kit (Westbury, NY) was used according to manufacturer's instructions (Table 2), in conjunction with M13 forward and reverse vector primers for PCR. PCR products were electrophoresed on a 2% standard agarose gel in 1x Tris-Borate-EDTA (TBE) buffer. Samples were electrophoresed at approximately 6-8 volts/cm (based on gel length) for 2 hours, then stained for 30 minutes with 0.5  $\mu\text{g/ml}$  EtBr (ethidium bromide), and destained for 1 hour in  $\text{H}_2\text{O}$  (changing  $\text{H}_2\text{O}$  at least twice) before being photographed under UV light using a NucleoTech GelExpert camera and software system (San Carlos, CA).

## **Sequencing and Bioinformatics**

Clones were sequenced at three locations: The Hubbard Center for Genome Studies (HCGS) at the University of New Hampshire, the Virginia Bioinformatics Institute (VBI) at Virginia Tech, and the Clemson University Genomics Institute (CUGI). The 226 clones processed at HCGS were sequenced from the PCR product produced from isolated plasmid template and purified using Millipore Columns. Sequencing of clones at VBI and CUGI utilized rolling circle amplification to produce sequencing templates directly from colonies. In total 3298 clones were sequenced unidirectionally (from the 5' end), using the M13 forward vector primer. Two 384-well plates were processed at

VBI, six 384-well plates were processed at CUGI, and 226 clones were processed at HCGS. Sequence data obtained at all three locations was sent to CUGI and combined into one data set for final bioinformatics analyses. Sequences were trimmed, assembled into singlets and contigs, examined for Simple Sequence Repeats (SSRs), and then sequence similarity searches were performed against databases from GenBank, The *Arabidopsis* Information Resource (TAIR), and SwissProt. These searches looked for nucleotide and predicted amino acid sequence homology to each of the following: all available organisms in databases, *Arabidopsis thaliana*, *Populus*, peach (*Prunus persica*), and the Rosaceae family.

## **DNA Isolation**

DNA was isolated from each of the 113 plants in the 'Yellow Wonder' x 'Pawtuckaway' mapping population, as well as from both the parents. Kevin deHaan and Ben Orcheski assisted with F<sub>2</sub> population DNA isolations. Two unexpanded leaves were picked from each plant using forceps and were placed into 1.5 ml microfuge tubes on ice. For each plant, leaves were transferred to a pre-chilled ceramic mortar and ground to a fine powder in liquid nitrogen using a chilled ceramic pestle. To the ground tissue, 1 ml grinding buffer (cetyl-trimethyl-ammonium bromide (CTAB) buffer with 4  $\mu$ l/ml  $\beta$ -mercaptoethanol) was added and grinding was continued for about 2 minutes to produce a slurry. The slurry was transferred to a 1.5 ml tube and incubated at 60°C for 1 hour.

Following incubation, samples were cooled to room temperature. Tubes were filled (approximately 500-800  $\mu$ l) with 24:1 chloroform:octanol and were then vortexed for 10 seconds to mix contents. To separate phases, tubes were placed into a microcentrifuge at 14,000 x g for 5 minutes. After centrifugation, the upper aqueous phase of each sample was transferred into a clean 1.5 ml tube. Cold 95% ethanol was added to each tube until nearly full. Tubes were then held on ice at least 15 minutes to initiate DNA precipitation, after which the aqueous and alcohol phases were mixed by gentle, repeated inversion of the tubes.

Once DNA was precipitated, samples were centrifuged for 5 minutes at 14,000 x g to pellet DNA. Supernatant was poured off and 1 ml cold 70% ethanol was added to each tube. Tubes were held on ice for at least 15 minutes. Supernatant was again removed and DNA pellets were dried for 5 minutes in a speed vacuum. To each dried pellet, 50  $\mu$ l TE was added and samples were allowed to sit at 4°C overnight to allow the DNA to re-dissolve.

RNA was eliminated from DNA samples by adding 50  $\mu$ l RNase solution (1  $\mu$ l 10 mg/ml RNase stock in 1 ml sterile H<sub>2</sub>O) to each DNA sample. Tubes were spun briefly (approximately 15 seconds in a microfuge) and then the solution was pipetted up and down several times, with a cut pipette tip, to mix. Samples were then incubated for 1 hour at 37°C to digest RNA. The DNA was quantified using a fluorometer and then diluted with sterile water to 40  $\mu$ g/ml.

## Gene Selection

Ribosomal protein gene sequences were selected from the YW EST set to be used for comparative mapping studies. Though they have the potential to be Conserved Ortholog Set genes and may have utility across species, it was unknown which subcategories or regions of these sequences might be most polymorphic. For this reason, primer pairs were designed to target different areas of the ribosomal protein gene sequences. Genes were selected that had varying numbers of introns, that had large 5' or 3' UTRs, and that represented large (60S) or small (40S) ribosomal subunits. While primer sequences were based on cDNA sequence, intron locations were predicted based on their location in putatively orthologous ribosomal protein genes in *Arabidopsis thaliana* (sequences publicly available at NCBI). Primers were designed to utilize these various characteristics by flanking one or more introns and/or by placing one primer in a UTR. This was done in order to evaluate what areas of the genes might have higher rates of polymorphism and would therefore be more easily mapped. Primers were then designed using Primer Select and EditSeq (DNA Star Package, LaserGene, Inc.). All primers were custom made by Integrated DNA Technologies (Coralville, IA).

## Polymorphism Detection

PCR amplifications using the primers described above were used to look for polymorphisms between parents in both mapping populations: the intraspecific YW (*F. vesca*) x PAWT (*F. vesca*) and the interspecific FDP815 (*F.*



*vesca*) x FDP601 (*F. nubicola*). Polymorphisms were detected after PCR as the presence/absence of DNA bands on a gel, or as mobility differences between bands. If one of these types of polymorphisms was visible between two parents, it would then be possible to screen the corresponding F<sub>2</sub> population in a similar manner to obtain the segregation data needed to map the gene. If no electrophoretically detectable polymorphism was observed using PCR alone, the CAPS technique was employed in an attempt to find restriction digest polymorphisms.

## PCR Conditions

PCR was performed using PCR kits purchased from Eppendorf (Westbury, NY) and reactions were completed as directed by the manufacturer (Table 2). Some reactions were doubled for a final volume of 50  $\mu$ l so that they could be used for several CAPS digests, but concentrations remained the same for each reagent.

A thermocycler was used for temperature cycling using the following program:

1. 94°C for 2 mins
2. 94°C for 30 sec
3. Primer specific annealing temperature (see Tables 6 and 7) for 30 sec
4. 72°C for 1 min
5. 29 times to step 2
6. 72°C for 5 mins
7. 4°C hold

**Table 2.** Reagents used in Eppendorf PCR kit (Eppendorf, 2005).

<b>Component</b>	<b>Volume</b>
dH <sub>2</sub> O	14.3 $\mu$ l
5x Taq Master	5.0 $\mu$ l
10x Taq buffer with Mg <sup>2+</sup> (25 mM)	0.55 $\mu$ l
dNTP Mix (10 mM)	1.1 $\mu$ l
Forward Primer (20 $\mu$ M)	0.55 $\mu$ l
Reverse Primer (20 $\mu$ M)	0.55 $\mu$ l
Template DNA (40 ng/ $\mu$ l)	2.5 $\mu$ l
Taq DNA Polymerase (5 U/ $\mu$ l)	0.2 $\mu$ l
<b>Total Volume</b>	<b>25 <math>\mu</math>l</b>

## **CAPS Conditions**

The CAPS technique was used when no polymorphisms were observed from PCR alone. For each digest, the following were added to a 0.5 ml tube: 8.0  $\mu$ l PCR product, 0.5  $\mu$ l restriction enzyme, 2.0  $\mu$ l 10x restriction enzyme buffer, and 9.5  $\mu$ l H<sub>2</sub>O. Tubes were flicked gently to mix and then contents were spun down by allowing centrifuge to just reach 14,000 x g before stopping. Samples were then incubated at 37°C overnight. Three enzymes were used to look for polymorphisms for each gene. These were *Alu*I, *Bfa*I, and *Hae*III. Digests were then electrophoresed to look for restriction digest polymorphisms.

## **Gel Electrophoresis**

PCR and CAPS results were observed on 2% agarose gels composed of 1% standard agarose and 1% low melt agarose (SeaPlaque or Nusieve from Cambrex (East Rutherford, NJ)) and 1x TBE buffer. Samples were electrophoresed at 6-8 volts/cm (based on gel length) in 1x TBE buffer for 2 hours and then stained for 30 minutes with 0.5  $\mu$ g/ml EtBr. Gels were then destained for 1 hour in H<sub>2</sub>O (changing water at least twice) and then photographed using the NucleoTech GelExpert system (San Carlos, CA).

## CHAPTER III

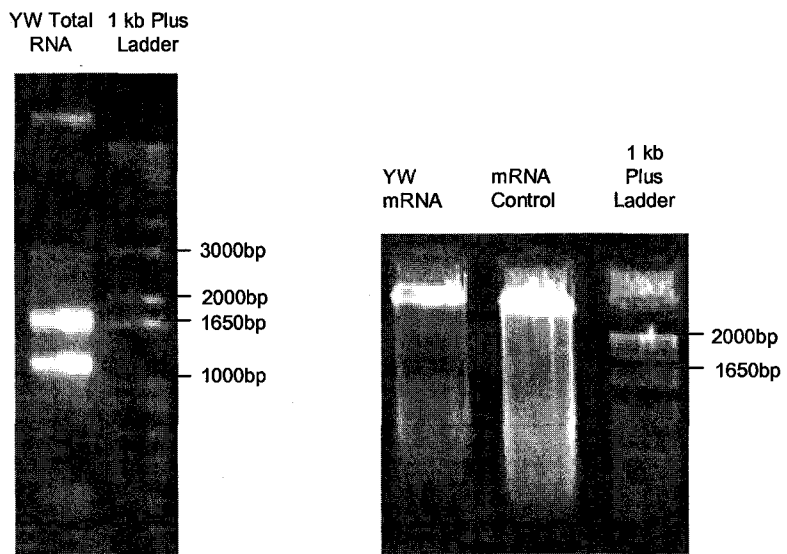
### RESULTS

#### **cDNA Library Construction**

Objectives 1 and 2 of this project were accomplished by the construction of a cDNA library and generation of Expressed Sequence Tags (ESTs). Total RNA isolated from approximately 7.0 grams of developing 'Yellow Wonder' (YW) flower buds was used as the mRNA source for library construction. When the quality of total RNA was checked by running out a 1  $\mu$ l sample on a native agarose gel with a 1kb+ DNA ladder, the expected ribosomal bands were seen at around 1200 bp and 1700 bp (Figure 7). These isolations were then used for poly(A) isolations. The recommended amount of mRNA for library construction was 1  $\mu$ g suspended in 1  $\mu$ l solution. However even after concentrating mRNA isolations from 7.0 grams of tissue, the resulting poly(A) RNA concentration was still considerably lower than desired, as seen by the lesser band brightness as compared to the 1  $\mu$ g/ $\mu$ l human placenta poly(A) control (Figure 7). By comparison with the standard, it was estimated that the YW Poly(A) RNA concentration was between 0.5-0.75  $\mu$ g/ $\mu$ l, which was lower than the preferable concentration. Nevertheless, because of concern that efforts to further concentrate this mRNA would result in a loss of total mRNA yield or a reduction

in quality due to shearing or degradation, this mRNA sample was used to produce cDNA and to successfully construct a library using the CLONTECH SMART cDNA Library Construction kit.

**Figure 7.** Total RNA and mRNA isolations from developing YW flower buds. Human Placenta mRNA (1  $\mu\text{g}/\mu\text{l}$ ) was used as a control for mRNA concentration.



## Clone Picking and Evaluation

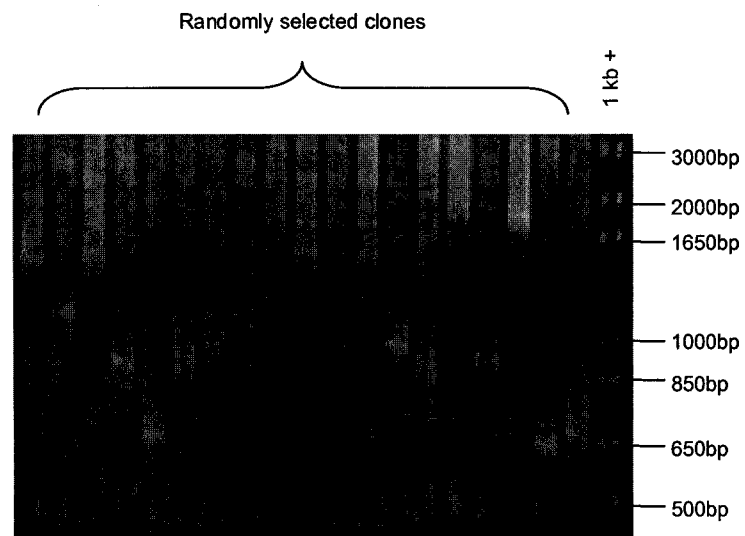
In total, 7680 clones were picked from the completed library; 2688 by hand and 4992 robotically. These clones contained the CLONTECH pDNR-LIB vector with directionally inserted cDNA sequences from YW mRNA. Those clones that were picked by hand were transferred into seven 384-well plates and stored at  $-80^{\circ}\text{C}$ . Robotically selected clones were also picked into 384-well plates, labeled with HCGS bar code numbering, and stored at  $-80^{\circ}\text{C}$ , but were done so in duplicate and also spotted onto high density filters. Membranes were spotted using a grid so that they could be probed and individual clones identified if desired, though no filter probing was done in the context of this project.

Before bioinformatics was performed, clone insert sizes were evaluated to determine if inserts were long enough to potentially contain full-length or near full-length coding sequences. Variability in insert sizes was also examined, because if all clones were of similar size, it could be possible that they represented redundant clones or that something occurred during the process of making cDNA that inadvertently selected for clones of a certain size (for example, sequences of larger than some specific size could have been broken during the speed vacuuming process). These evaluations were accomplished using PCR with M13 forward and reverse primers in PCR from plasmids. It was found in initial screens of randomly selected clones that insert sizes varied greatly, with a general range between 1-2 kb (Figure 8). Based on insert sizes, it was concluded that the library likely contained various full-length, or at least relatively long, sequences that could be used for further investigation. Several

plates were subjected to high throughput sequencing, and for subsequent analysis of the sequences.



**Figure 8.** PCR of clones using M13 primers to target vector inserts for relative sizes.



## Bioinformatic Analyses

Sequencing and bioinformatics were performed at three locations: The Hubbard Center for Genome Studies (HCGS) at the University of New Hampshire, The Genome Database for Rosaceae (GDR) at Clemson University Genomics Institute, and Virginia Bioinformatics Institute (VBI) at Virginia Tech. HCGS sequenced 226 clones. VBI sequenced two of three 384-well plates sent to them (HCGS-YW-cDNA-011604, plates 1,2,3) containing 768 clones. At GDR, six 384-well plates (HCGS-YW-cDNA-011604, plates 4 through 9) containing 2304 clones were sequenced. The total number of clones sequenced was 3298, which is approximately 42.9% of the clones in the library. There are still 4443 clones that could be sequenced and analyzed. The resulting sequences from all three genomics facilities were combined into one data set for bioinformatic studies at GDR. It was determined that 2717 of the 3298, or 82.4%, were considered of high enough quality to be used in homology searches. The remaining 581 clones were not used for any further investigations. The latter sequences were considered to have failed because they had greater than 5% ambiguous bases (represented by "N" in the sequence) or they had less than one hundred high quality bases, as assessed by a scoring system by Ewing et al (1998). The 2717 Expressed Sequence Tags (ESTs) were submitted to GenBank's EST database and cataloged under accession numbers DV438013 - DV440729. These sequences comprised 63.5% of *Fragaria vesca* sequences and 25% of all *Fragaria* species sequences in GenBank at the time of their submission.

Homology searches were accomplished by the GDR with the 2717 ESTs using several different databases with the fastx3.4 algorithm devised by Pearson and Lipman, 1988 (Table 3). The data set had 1918 clones, or 70.6%, with matches to the NCBI Plant Protein database. There were only 25 (0.9%) clones with matches to mapped peach ESTs. The TAIR *Arabidopsis* database was searched and yielded 1869 matching clones, which was 68.8% of the 2717 EST clones. SwissProt matched 1276 clones, or 47.0% of those used in the search. A number of EST databases were also searched including *Arabidopsis*, Rosaceae, and *Populus*, which resulted in 391 (14.4%), 1892 (69.6%), and 784 (28.9%) matches, respectively. The criterion for matches to the protein databases was an e value of  $< 1e-6$ . EST database matches were based on sequences having  $> 85\%$  amino acid identity and overlap of  $>100$  bp. These numbers give a rough estimate of the number of clones with homology to those in existing databases. They do not however account for sequence redundancy, as constituted by sequence duplication or repetition within the EST data set.

Though the library has a comparatively low level of redundancy (in comparison to other data sets archived by CUGI), there were 807 redundant sequences. Redundancy varies greatly in libraries based on the tissue used to make the library as well as the developmental stage of the tissue. The most redundant clones in this library were proteins that had homology to metallothionein, which accounted for more than 100 of the redundant clones (Table 4). To account for repeated sequence in the library, non-redundant sequences were designated as “unigenes” which includes singlets (sequences

appearing only once in the EST data set) and contigs (assemblies of repetitive sequences). Contigs were configured using an assembly program called CAP3 (Huang and Madan, 1999). There were 1910 unigenes in the 2717 sequences analyzed; 1581 were from singlets and 329 were from contigs. Unigenes (singlets and contigs) were also searched for homology against several databases. This resulted in homology of 1310 unigenes (68.6%) to the NCBI NR database, 1235 unigenes (64.7%) to Rosaceae ESTs, and 1213 (63.5%) to the TAIR *Arabidopsis* database (Table 5). This means that approximately 600 unigenes, 1/3 of those in the library, have no homologous sequence in any of the databases searched.

**Table 3.** Number and percent of 2717 clones submitted to Genbank with matches in selected databases

<b>Database</b>	<b>Number of clones in EST data set with match</b>	<b>Percent of clones in EST data set with match</b>
NCBI Plant proteins	1918	70.6
Mapped Peach ESTs	25	0.9
Arabidopsis ESTs	391	14.4
Populus ESTs	784	28.9
Rosaceae ESTS	1892	69.6
TAIR Arabidopsis Proteins	1868	68.8
SwissProt	1276	47.0

**Table 4.** Most redundant sequences in library with the number of clones appearing in each contig

<b>Contig</b>	<b># Clones</b>	<b>Gene Identity</b>
Contig197	55	metallothionein-like protein
Contig7	50	Metallothionein-like protein type 2 MET1
Contig86	34	unnamed protein product
Contig43	29	pollen allergen-like protein
Contig20	19	metallothionein-like protein
Contig200	18	unnamed protein product
Contig128	17	unnamed protein product
Contig74	15	Gty37 protein
Contig10	14	calmodulin
Contig132	14	pathogenesis-related protein

**Table 5.** Number and percent of 1910 unigenes from cDNA library with matches in selected databases

<b>Database</b>	<b>Number of unigenes with match</b>	<b>Percent of unigenes with match</b>
NCBI NR	1310	68.6
Rosaceae ESTS	1235	64.7
TAIR Arabidopsis Proteins	1312	68.7

## **Selection of Ribosomal Protein Genes**

The second part of this investigation aimed to identify and evaluate conserved ortholog set (COS) genes as candidate mapping markers and then use these markers to compare the extent of polymorphism between two mapping populations. Ribosomal protein genes became the focus in development of a *Fragaria* COS marker set due to their potential utility across species. Many ribosomal protein genes were represented in the library. Tables 6 and 7 show the selected ribosomal protein genes, the primer sequences used to target them, and the primer locations including any introns or UTR that may be included in the product. Genes and primer positions were selected to encompass a wide variety of possibilities with respect to the number and size of intron, exon, and UTR sequences contained within the respective PCR products. The locations of these features were deduced by comparison to orthologous *Arabidopsis* sequences available in GenBank. Due to the high level of conservation among ribosomal genes, even between plant and animal species, the size and location of introns, exons, and UTRs was expected to be similar in *Arabidopsis* and *Fragaria*. Primers were then designed from YW cDNA sequence from the library. A total of 30 primer pairs (there were two cases in which two primer pairs were targeted to different regions of the same gene: RPS14 and RPS29B) were selected to amplify segments of 28 different ribosomal protein genes.



**Table 6.** Primers designed to target nuclear-encoded, 60S ribosomal protein large-subunit (RPL) genes

Primer Pair Name	Gene	<i>F. vesca</i> GenBank EST	Best <i>Arabidopsis</i> Match	Primer Sequence	Primer Location		# Introns Expected in Product	Annealing Temp C
					Forward	Reverse		
7F20	RPP0B	DV439140	At3g09200	F - GAG AAG ACC CGT AAC GAT GCT AT R - ACA GGG GTT ATG ATT TCC ACA GTT	Exon 1	Exon 4	3	55.2
5A17	RPL3A	DV438378	At1g43170	F - TTC CAG AAA GAT GAG ATG ATT G R - CGA TAC TTG AAA CAG CGA TAA	Exon 5	Exon 6	1	53.8
2-4E12	RPL7B	DV440228	At2g01250	F - ATC TAC AAT AAA GCC AAG CAG T R - GTT TCC AGC ATC TCC ACC T	Exon 1	Exon 6	5	53
Contig239	RPL10B	DV439852	At1g26910	F - GGT GTT GAT GAG TTC CCC TTC TG R - ACA TCA ATA TTA CTG GCG TTC ACA	Exon 2	3'UTR	1	55.8
2-4E03	RPL10aA	DV440222	At1g08360	F - CCA AGG AGA AGA ATC GGA AGT R - ACG TTT TGC CAA TTC TTT TTC A	Exon 2	Exon 4	2	53.8
6N19	RPL11D	DV438974	At5g45775	F - AGA AGA TTG CGT GCT ATG TGA CTG R - CAT CCT TTG TAA CTC TGT GCT GAA	Exon 5	Exon 6	1	55.3
Contig150	RPL13aD	DV438936	At5g48760	F - GAA CAC CAA GCC CTC TCA R - GAC CTT TCC TTC CTC TTC TT	Exon 1	Exon 4	3	54.4
7P09	RPL18B	DV439337	At3g05590	F - TGT TCG CAT CGC TGA GG R - ACC ACG GCT GTT TTT C	Exon 4	Exon 5	1	56.6

**Table 6 (cont).** Primers designed to target nuclear-encoded, 60S ribosomal protein large-subunit (RPL) genes

Primer Pair Name	Gene	<i>F. vesca</i> GenBank EST	Best <i>Arabidopsis</i> Match	Primer Sequence	Primer Location		# Introns Expected in Product	Annealing Temp C
					Forward	Reverse		
Contig214	RPL22C	DV439600	At5g27770	F - GGC TTT TAC CAT TGA TTG R - CAT TTC GGT CCT TGT TG	Exon 1	Exon 2	1	51
9K05	RPL23B	DV439867	At2g33370	F - GAA CCT TTA CAT CAT CTC CGT CAA R - TTA GCA GCA CTG GCA ATC CT	Exon 2	Exon 4	2	55.6
6O13	RPL27aC	DV438988	At1g70600	F - GCA CCA CCA CCG CAT CCT CTT R - CCG CCG GCT TCC TTA ATC TTC TT	Exon 1	3'UTR	0	60.4
Contig132	RPL32A	DV438772	At4g18100	F - ACA TTG GTT ACG GTT CTG ACA R - CTC TCA ACA ATA TCC TTC CTC T	Exon 2	Exon 2	0	50.2
Contig90	RPL34A	DV438433	At1g26880	F - CAC CAA GTC CAA CCA GCA CAG R - GAC CAA AAA GGC ACG GAT GA	Exon 1	Exon 4	3	56.2
Contig169	RPL37aB	DV439395	At3g10950	F - TGC TGT GAA GAG GAA GG R - GCA CTA GCA ACC AAA TAA A	Exon 4	3'UTR	0	51
Contig113	RPL39C	DV440390	At4g31985	F - CAA GAA GAA GCT GGG GAA GAA GAT R - CCC GAT TGA CCA AAC ATA AAA GAT	Exon 2	3'UTR	1	55.9

**Table 7.** Primers designed to target nuclear-encoded, 40S ribosomal protein small-subunit (RPS) genes

Primer Pair Name	Gene	<i>F. vesca</i> GenBank EST	Best <i>Arabidopsis</i> Match	Primer Sequence	Primer Location		# Introns Expected in Product	Annealing Temp C
					Forward	Reverse		
5M01	RPS5A	DV438602	At2g37270	F - CCT TGG AGA TTA CAT TGG AGT GAC R - GAA CCT TTG GCA GCA TTG ATA A	Exon 2	Exon 3	1	55.2
2-6E01	RPS7A	DV438774	At1g48830	F - TCC TCC ACC CGA AAG A R - CAG AAC ATG GCA GAA CAG T	Exon 5	3'UTR	0	49.5
t-6G14	RPS9C	DV438826	At5g39850	F - TTT CAA GAA GCC AAG AC R - TAC ACA ACC AAA GGG ATA	Exon 2	Exon 3	1	52.1
Contig295	RPS13A	DV440558	At4g00100	F - GAC TCT CAT GGG ATT GCT CAG G R - TCA TAA AAC GAA AAC GAC ACA TCA	Exon 3	3'UTR	2	55.1
5B20#1	RPS14	DV438397	At3g11510	F - CTC CGC TTT AGT GTT TTA R - AGT TTC CCT TCC AGA GA	5'UTR	Exon 3	2	50.8
5B20#2	RPS14	DV438397	At3g11510	F - GAA ACT TTG GTC CGC ATC AC R - TCT TCT ACC ACC CTT TCT ACG AGT	Exon 3	Exon 6	3	55.3
9B20	RPS15D	DV439697	At5g09510	F - AGC TCT GAC CTT TTC TCT R - CCA TGC TAA TAC CTT GTT T	5'UTR	3'UTR	3	51.1
Contig135	RPS20B	DV439200	At3g47370	F - TCC GCC GTC ACC TAG AGT AAA A R - TCG ACA CCA GGT TCA ATA GTA ATG	5'UTR	Exon 3	2	55.9

**Table 7 (cont).** Primers designed to target nuclear-encoded, 40S ribosomal protein small-subunit (RPS) genes

Primer Pair Name	Gene	<i>F. vesca</i> GenBank EST	Best <i>Arabidopsis</i> Match	Primer Sequence	Primer Location		# Introns Expected In Product	Annealing Temp C
					Forward	Reverse		
2-4H01	RPS23B	DV440248	At5g02960	F - CAA GCT GAA GAA CCA CCG TAG AAG R - GTC CAA ACC CAG CAA TCA ACA C	Exon 2	Exon 4	2	55.1
4A04	RPS24A	DV438016	At3g04920	F - CCG ATA ACA AAG CCG TGA CCG R - TTG CTG CAG AAA CAT GGA AAC TC	Exon 1	3'UTR	4	55.9
2-5H04	RPS25	DV440329	At4g34555	F - AGC AAA GGC AAG CAG AAG G R - GCG TGA GCC GAG ACC AT	Exon 3	Exon 4	1	55.6
2-5D02	RPS29B	DV440291	At3g44010	F - CTC GCA TCC CAA GTC R - TGC AGC ACA TCA GTC C	Exon 1	Exon 2	1	50
t-6B07	RPS29B	DV440566	At3g44010	F - AGT AAT TGC GGG GTC CTG R - TTT TGC ACG CGA AGA GAA	Exon 2	3'UTR	1	52.6
Contig159	RPS30A	DV439026	At2g19750	F - TCT ACC GGC GAG CGA CTA C R - GCA CAT ATC CAT AAA ATC TCA CA	5'UTR	3'UTR	2	54.2
Contig211	RPS30C	DV439558	At5g56670	F - GCA ACT GCG CTC TAC TCT CCA R - TGC CAA AGC CAA CCA CAG	5'UTR	Exon 2	1	56.4

## Gene Amplification

Of the 30 primer pairs used in this study, 27 produced PCR products from genomic DNA isolates of at least one of the mapping parents from each set (YW x PAWT and FDP815 x FDP601). Primers for ribosomal protein genes RPS29B, RPS20B, RPS9C did not generate any product, leaving 27 primer pairs (for 26 genes) that did generate products. In most instances, a single PCR product band was visible on the gel; however, there were some cases where faint secondary bands were also present and these instances are discussed in the following chapter. Product sizes (Table 8) ranged from around 150 bp (RPS7A) to approximately 2000 bp (RPL23B). Although a few PCR products that were close or indistinguishable in size to those predicted from *Arabidopsis* (e.g., RPL3A and RPL32A), most PCR product sizes were larger or smaller than predicted, with the maximum deviation from predicted size exceeding 1600 bp (RPL23B) (Table 8).

## PCR and CAPS Polymorphisms

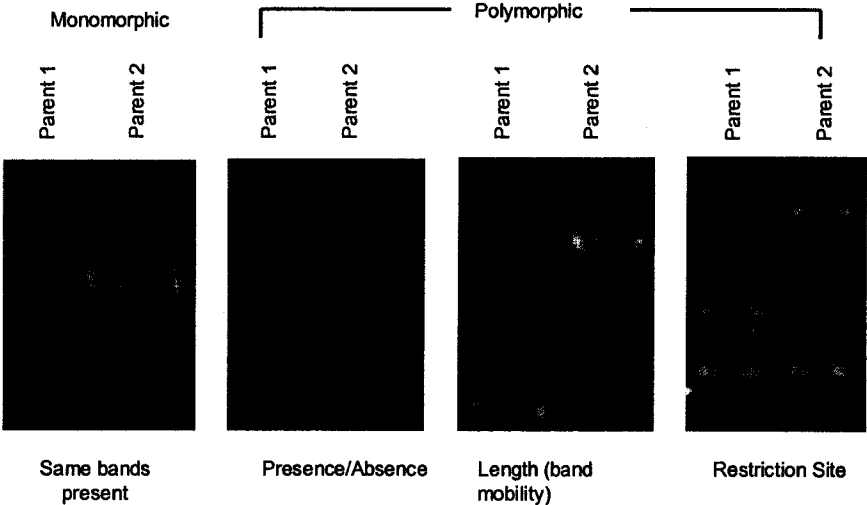
PCR products alone, or after digestion with either *AluI*, *BfaI*, or *HaeIII* restriction enzymes, were used to look for polymorphisms in both sets of mapping parents. Polymorphisms between YW and FDP815, both of which belong to *F. vesca* ssp. *vesca*, were also noted. With PCR products alone, three possible outcomes were anticipated, as illustrated in Figure 9: i) monomorphic bands in which both parents had the same bands; ii) presence/absence polymorphism where one parent generated a DNA fragment that the other did

not; and iii) PCR product length polymorphism, in which a band or bands were present in both parents but migrated to different gel positions due to PCR product length differences. Restriction digestion of PCR products provided an additional opportunity to detect differences between the parents, in which digest banding patterns differed due to gain or loss of a restriction site from a SNP or to one or more indels within the amplified region (Figure 9).

The majority of the amplification targets were found to be monomorphic in this investigation. Of the 26 genes that yielded PCR products, only six displayed polymorphisms of any kind. Of these, five were ribosomal large (60S) subunit protein (RPL) genes: RPL23B, RPL32A, RPL37aB, RPL10B, and RPL27aC. The only small (40S) ribosomal subunit (RPS) gene found to be polymorphic was RPS14. The results for each of these six genes are described below (as compiled in Table 9), followed by a summary of results in terms of type of polymorphism and differences between pairs of mapping parents. Technical issues, including faint polymorphic bands that appeared in most of the samples, will be addressed in the discussion chapter.

Amplification and digestion of the RPL37aB ribosomal protein PCR product yielded the most straight forward results (Figure 10). This primer pair targeted a region of exon 4 beginning in coding sequence and ending in 3'UTR, and that did not include any introns. The CAPS technique with *AluI* revealed a restriction site polymorphism between YW and PAWT. Undigested PCR products and digestion products from *Bfal* and *HaeIII* were found to be monomorphic.

**Figure 9.** Possible results for monomorphic and polymorphic alleles in the mapping population. Presence/absence and length (band mobility) polymorphisms may be observed after PCR, while restriction site polymorphisms are revealed by restriction enzyme digestion.



For FDP815 and FDP601, PCR products and restriction digestion products were monomorphic for each set of digests (*AluI*, *BfaI*, and *HaeIII*) (Results not shown). YW/PAWT parents and FDP815/FDP601 parents shared the same size DNA fragments for each respective digest, except for the polymorphism observed with PAWT digested with *AluI*. With this digest, YW, FDP815, and FDP601 had the same bands present.

For gene RPL32A, a primer pair was designed to amplify a region of the gene within exon 2, and did not span an intron. A PCR product presence/absence polymorphism was observed between FDP815 and FDP601, while the PCR products were monomorphic between YW and PAWT, both with and without restriction digestion (Figure 11). The presence/absence polymorphism between FDP815 and FDP601 was a band of around 220 bp that was present only in FDP601, while both parents shared a band at 270 bp. YW and PAWT *F. vesca* mapping parents shared the same single band as *F. vesca* FDP815. There were also some other faint bands present in both FDP815 and FDP601.

The primer pair for gene RPL10B was used to amplify region starting in exon 2 coding sequence, ending in the 3'UTR region of exon 3, and including one intron. The PCR products were found to be monomorphic among the mapping parents, but digestion with *HaeIII* detected a restriction site polymorphism between the YW and PAWT mapping parents (Figure 12). It was also noticed that with PCR alone, hazy bands were present around 1000 bp and 500 bp. Though there were no other polymorphisms observed between paired



mapping parents, YW and PAWT differed from FDP815 and FDP601 with respect to PCR product sizes and all digestion patterns (results not shown).

For RPL27aC the gene region without any introns, beginning in exon 1 and ending in the 3' UTR, was targeted. The PCR and restriction digestion products resulted in monomorphic DNA bands between the YW and PAWT parents (Figure 13). A presence/absence polymorphism could be detected between the FDP815 and FDP601 parents using PCR alone. A restriction site polymorphism (CAPS) was also revealed between these parents using *HaeIII*. FDP815 also had some large very faint bands present in the PCR product and restriction digests that were not present in FDP601. Aside from the polymorphisms mentioned, all other bands were monomorphic between YW, PAWT, FDP815, and FDP601 for PCR and each digest.

The region of RPS14 targeted by primer pair RPS14-#2 was not found to be polymorphic between the YW and PAWT parents by PCR alone or after digestion. The primers for this gene were located in exon 3 and exon 6 and amplified a region containing 3 introns. This region was polymorphic between the FDP815 and FDP601 mapping parents using PCR alone and after digestion with *AluI* and *HaeIII* (Figure 14). As with many other of the PCR and restriction digests products, some faint secondary bands of varying sizes were observed between the mapping parents.

RPL23B was the most highly polymorphic gene in this study (Figure 15). It was the only gene that was polymorphic in both the intraspecific and interspecific mapping parents. The targeted region of the gene was between

exons 2 and 4 and included two introns. Polymorphism could be detected between YW and PAWT using two different restriction enzymes (*Alul* and *Bfal*), and between FDP815 and FDP601 with PCR alone or with each of the three enzymes tested (*Alul*, *Bfal*, and *HaeIII*). For PCR and digestion with *HaeIII*, YW, PAWT, and FDP815 have the same bands present. All four parents have different DNA fragments present with *Alul* and *Bfal* digestion.

Considering the results for all six genes, there were four instances in which polymorphisms were detected using PCR alone and ten that were detected using the CAPS technique; four with *Alul*, two with *Bfal*, and four with *HaeIII* (Table 9). The four polymorphisms that could be detected by PCR alone (without digestion) were all discovered between the interspecific parents, and all but one of these also showed CAPS polymorphism. Using the CAPS technique, four instances of polymorphisms were identified between the intraspecific parents along with six between the interspecific parents.

It was noted that in both sets of mapping parents, there were instances where faint, secondary bands were present that were sometimes polymorphic between the parents. These bands were hazy in appearance and in some cases it could not be determined whether one or more bands were present. In most instances these bands were reproducible, though their intensities could vary making them difficult to see on some gels. Due to these extra bands, restriction digestion product sizes frequently did not add up to the undigested DNA fragment size. Some potential explanations for this phenomenon are given in the discussion section.

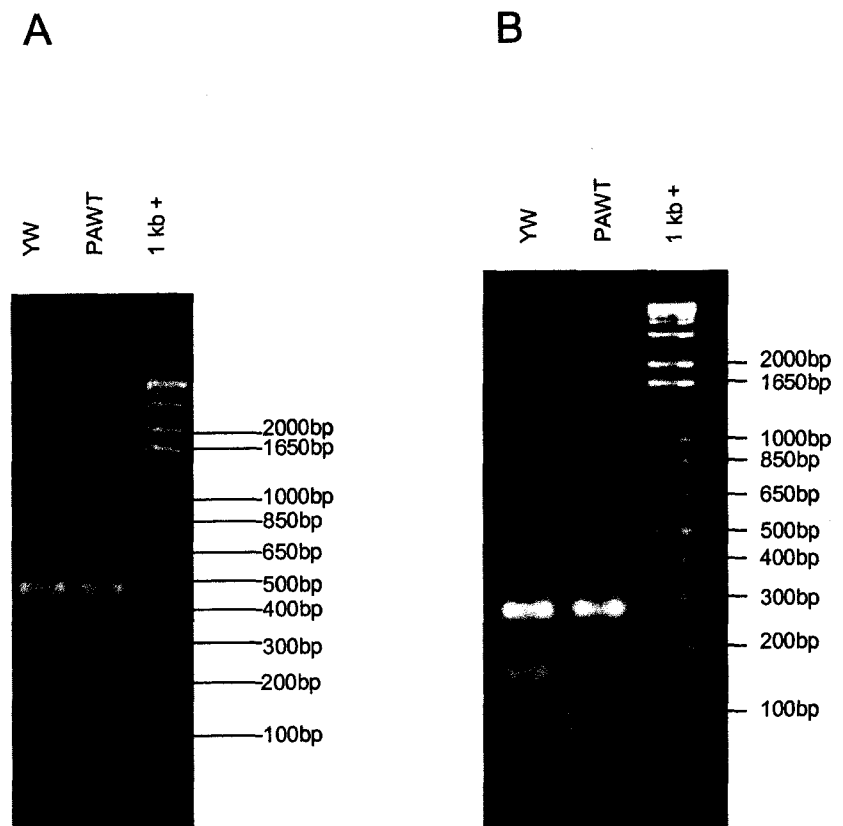
**Table 8.** Estimated and observed PCR product size of ribosomal protein genes. Approximate *Fragaria* exon size was based on cDNA clone sequence from Yellow Wonder library. Expected PCR product size was taken from *Arabidopsis* sequence on GenBank and includes introns and UTRs. RPL10B had one observed product in YW/PAWT parents around 1100bp, but an additional smaller band around 660 was also observed in FDP815/FDP601 parents.

Gene	Approx. <i>Fragaria</i> exon size (~bp)	Expected <i>Arabidopsis</i> PCR product size (~bp)	Observed <i>Fragaria</i> PCR product size (~bp)	Difference (~bp)
RPP0B	272	875	850	-25
RPL3A	640	700	700	0
RPL7B	557	1150	1350	200
RPL10B	677	800	660/1100	-140/+300
RPL10aA	541	950	1400	450
RPL11D	285	640	850	210
RPL13aD	306	900	150	-750
RPL18B	289	350	1000	650
RPL22C	272	780	950	170
RPL23B	306	350	2000	1650
RPL27aC	306	400	300	-100
RPL32A	172	172	175	3
RPL34A	250	780	850	70
RPL37aB	323	660	425	-235
RPL39C	228	390	350	-40
RPS5A	453	510	550	40
RPS7A	165	275	150	-125
RPS13A	449	775	1500	725
RPS14#1	232	530	700	230
RPS14#2	291	500	1000	500
RPS15D	732	1250	1700	450
RPS23B	284	465	480	15
RPS24A	442	1000	1600	600
RPS25	206	550	180	-370
RPS29B	95	785	650	-135
RPS30A	365	875	575	-300
RPS30C	204	325	460	135

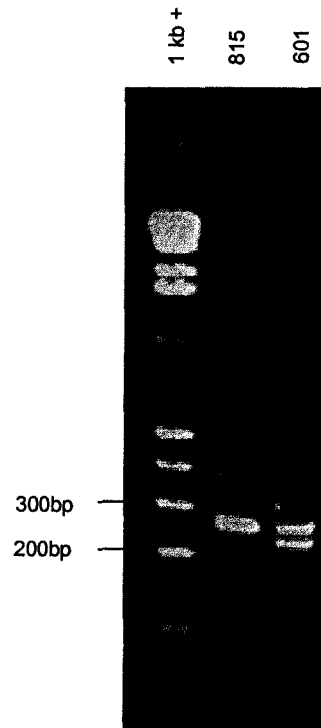
**Table 9.** Polymorphisms observed in YW x PAWT and FDP815 x FDP601 mapping populations. The first symbol represents the YW x PAWT parents and the second symbol represents the FDP815 x FDP601 parents. Polymorphisms are indicated with a + , while no polymorphism is marked by a -.

<b>Gene</b>	<b>PCR</b>	<b><i>AluI</i></b>	<b><i>Bfal</i></b>	<b><i>HaeIII</i></b>
RPS14#2	- / +	- / +	- / -	- / +
RPL23B	- / +	+ / +	+ / +	- / +
RPL32A	- / +	- / -	- / -	- / -
RPL37aB	- / -	+ / -	- / -	- / -
RPL10B	- / -	- / -	- / -	+ / -
RPL27aC	- / +	- / -	- / -	- / +

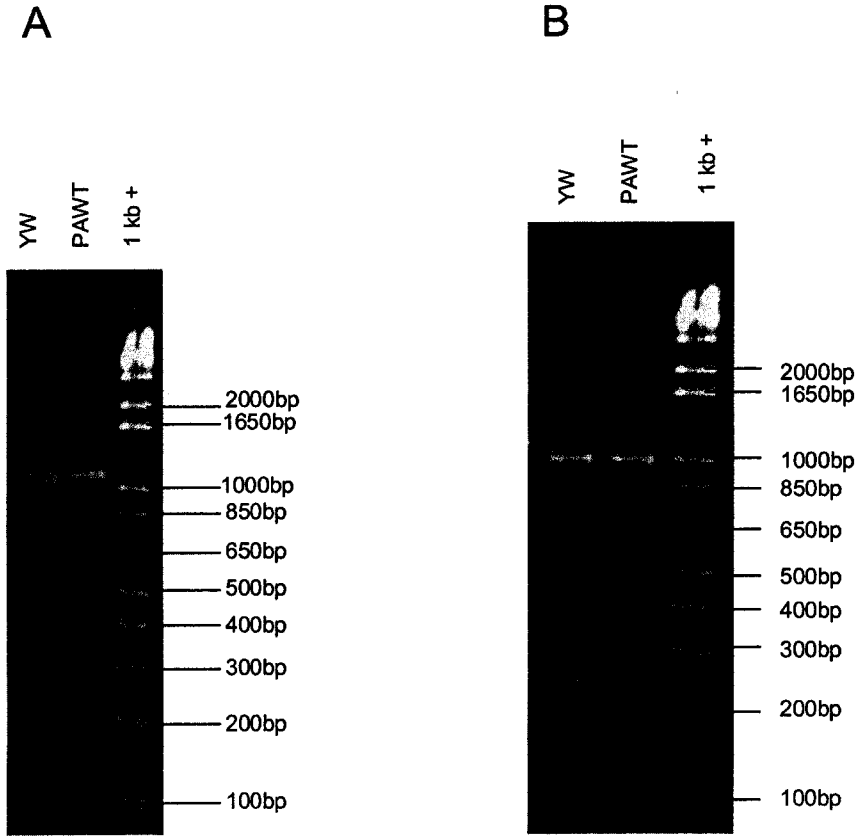
**Figure 10.** PCR of RPL37aB (A) followed by digestion with *AluI* (B). Polymorphic bands were seen between YW and PAWT mapping parents.



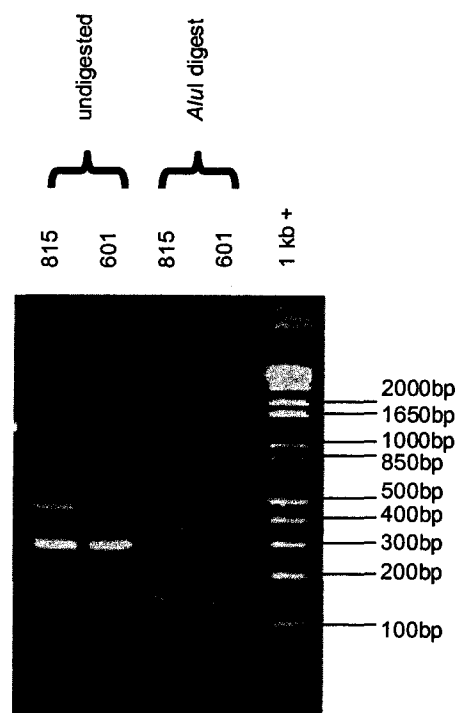
**Figure 11.** PCR of RPL32A showing polymorphic bands between FDP815 and FDP601 mapping parents.



**Figure 12.** PCR of RPL10B (A) followed by digestion with *Hae*III (B). Polymorphic bands were seen between YW and PAWT mapping parents.

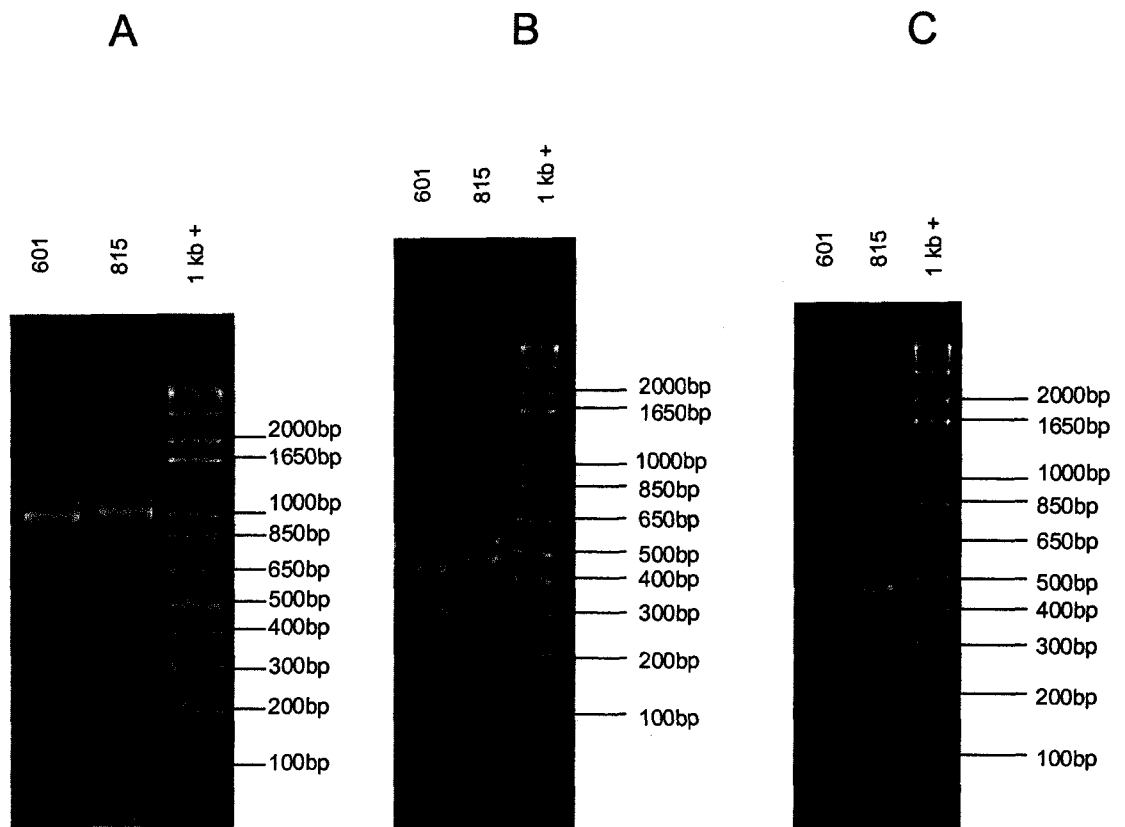


**Figure 13.** PCR of RPL27aC followed by digestion with *Hae*III. Polymorphic bands were seen between FDP815 and FDP601 mapping parents both with and without digestion.

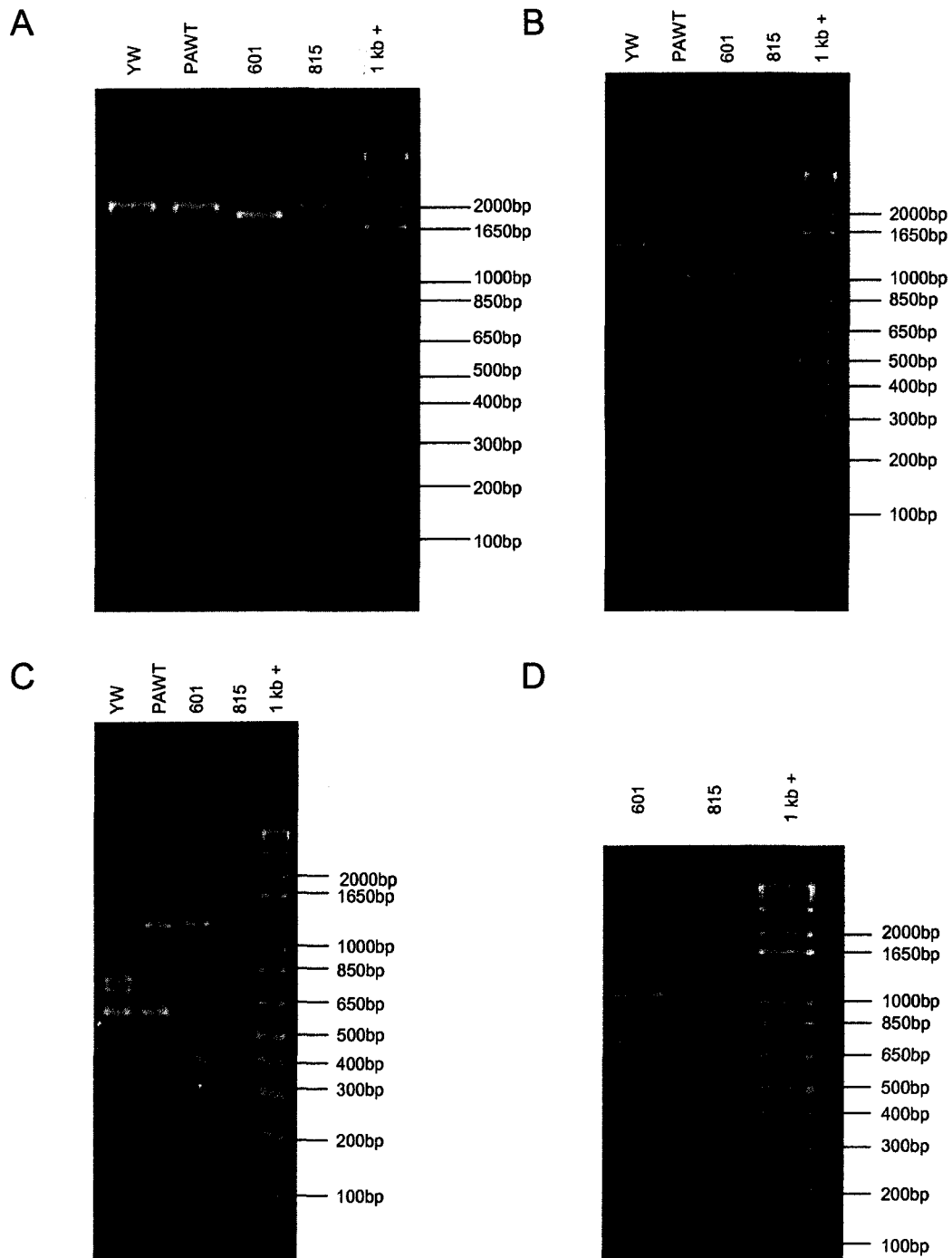




**Figure 14.** PCR of RPS14 (primer pair #2) (A) followed by digestion with *AluI* (B) and *HaeIII* (C). Polymorphic bands were seen between FDP815 and FDP601 mapping parents.



**Figure 15.** PCR of RPL23B (A) followed by digestion with *AluI* (B), *BfaI* (C), and *HaeIII* (D). Polymorphic bands were seen between both the YW/PAWT mapping parents (with *AluI* and *BfaI*) and the FDP815/FDP601 mapping parents (with PCR, *AluI*, *BfaI*, and *HaeIII*).



## CHAPTER IV

### DISCUSSION

#### RNA Quality

The quality of a cDNA library can be affected by several factors, but the starting quantity and condition of the mRNA used are among the most significant. After substantial effort was invested in collecting and dissecting 7 grams of flower bud tissue, and in isolating a concentrated sample of total RNA, mRNA isolated from this total RNA was of a lower concentration than recommended in the CLONTECH library construction kit. It is unclear what caused the reduced RNA yield in comparison to the manufacturer's predicted yields but this could have been remedied by additional poly(A) isolations, utilizing more flower bud tissue. However, this mRNA sample proved sufficient for the construction of a cDNA library consisting of 7680 clones, which was more than enough to fulfill the objectives of the project. Moreover, based on the sequencing results from the library, it is apparent that despite its low concentration the mRNA was of sufficiently high quality to yield many cDNAs that appear to be full length with reads starting in 5' UTR and ending in the 3' UTR (i.e., having PolyA tails).

All cDNA sequencing was done directionally from the 5' end. Most of the resulting EST sequences that began before the start codon (i.e., within the 5' UTR) and extended past the stop codon (into the 3' UTR), suggesting that the entire coding region was present. It is possible that some distal parts of the 5' or

3' UTR sequences may have been lost or excluded, but this did not pose a problem in the gene mapping investigation. All of the ribosomal protein genes selected as mapping candidates contained at least some UTR sequence, and many of these were utilized as target sites for primer design.

## **Library Redundancy**

The 2717 member 'Yellow Wonder' EST set contained 807 redundant sequences. This number constituted less than 30% of the Expressed Sequence Tags produced. Redundant clones are assembled into groups called contigs, and each contig (regardless of its number of members) is considered to be one unigene in bioinformatic analyses. Non-redundant clones, or singlets, are also unigenes. The unigene set, therefore, includes non-redundant singlets and contigs.

With around 70% of the clones being non-redundant unigenes, it would be worthwhile to continue sequencing the remaining 4443 clones in the library until the discovery rate of new clones dropped to a level where it was no longer fruitful to do so. However, at a prevailing cost of about \$4 per sequence, the project budget did not allow for the sequencing of these remaining clones at the present time. If more clones were to be sequenced and analyzed, it is to be expected that the unigene discovery rate would gradually go down and the number of contigs would likely increase. Many of the existing singlets would probably gain partners for assembly into contigs upon reanalyzing the data. However additional sequencing would also yield many hundreds or thousands of new

unique sequences and may increase discovery of gene sequences that are less highly transcribed.

The redundancy that appears in a given library is in part the result of the tissue used and the developmental stage it was in when harvested. The cDNAs that were present in particularly large numbers represented genes encoding proteins that were related to stress, such as metallothionein-like proteins, calmodulin, and pathogenesis-related proteins. It would make sense that developing tissues would be allocating resources to binding potentially harmful molecules and protecting themselves against pathogens. These may have also been in response to stresses that existed in the greenhouse where they were growing, such as heat, drought, or chemical treatments. Such stresses may have stimulated expression of these genes.

Another factor that accounts for EST redundancy is the stringency of contig assembly. Parameters for how much identity between sequences was necessary to make them a contig could be made less stringent by requiring sequences to have less bases in common. It is possible that genes would then be grouped that shared sequence similarity but were unique in function. As an example in which this could be a problem, the ribosomal protein genes belong to small gene families where members differ only by slight variations in nucleotide sequence. If these slight variations were not accounted for or not present in the read (cDNA was not full length), it could be impossible to differentiate between family members in the library. On the other hand, redundancy would be decreased if parameters for contig assembly were made more stringent. But this

could create other problems - for instance sequences from the same gene that do belong to one continuous sequence, but cannot be assembled into a single contig due to insufficient overlap of sequence (as could occur if the coding region was very long, and not fully sampled by any one cDNA clone or sequence). The CAP3 program (Huang and Madan, 1999) was used to assemble contigs with the hope of preventing over-assembly, while still grouping cDNAs that were truly redundant (i.e., transcripts of the same gene or allele).

## **Homology Searches**

Similarly to the assembly of contigs, the success of homology searches can be varied greatly based on the stringency used in performing the searches. For the bioinformatic analyses of this library, protein database queries were considered matches only if they had e values of  $<1e-6$ . For EST matches, query sequences required to have  $>85\%$  identity to match sequence and needed to have  $>100\text{bp}$  overlap. Lower stringency would result in homology to more database sequences. However, these matches may not be reliable and may be matching sequences with short highly similar or identical homologous reads that are not from homologous genes. Conversely, more strict guidelines for matches would result in fewer matches and the dataset would not account for many matches that are valid. This would give the false impression that the library had many more unique sequences than it actually did. The intent with the employed set of parameters for homology searches is that genes are being matched to sequences in the database that share a significant amount of similarity, even if

they do not have 100% of their bases or predicted amino acids in common. This balance was particularly important for the selection of ribosomal protein genes, where family members were distinguished from one another based on their small differentiating portions of sequence.

The level of stringency used for the homology searches in this cDNA library resulted in no match for around 600 of the sequenced clones. Their lack of homology to the databases could be explained in a number of ways. It is possible that these are genes that have not before been identified. These novel genes would probably exist in other species, but have either not been discovered in those species or not submitted to the databases. It is also possible that they have not been identified previously because they are in fact unique to strawberry. This set of strawberry-specific genes may not exist in any of the databases searched. Another explanation is that these genes are either in databases that were not searched, or they do exist in the databases searched, but gene regions appearing in the databases were different than the regions of the genes that were used for the query. This should not be true for many of the sequences since most are full length and should have some homologous region to anything in the database for that gene. It is unlikely that any one of these explanations could account for the 600 match-less clone sequences. More probably, they can be explained by a combination of these factors.

## **Expressed Sequence Tag Submission**

The 2717 ESTs deposited to GenBank was a significant contribution to the field of *Fragaria* genomics, as well as that of the Rosaceae family. At the time of their submission, this was the largest single submission by any lab for *Fragaria vesca*. At the time of submission it accounted for 63.5% of the sequences for this species, more than doubling what was previously available. This submission also accounted for 25.4% of those sequences available for the entire *Fragaria* genus at that time. This will undoubtedly be useful to the continued efforts to resolve the octoploid genome constitution and will provide useful sequences to the field of Rosaceae genomics. Many of these ESTs will also be useful to researchers working with other organisms as well. Sequences were submitted as ESTs rather than annotated NR sequences partly for this reason. It is an important feature of the EST database that contig members are not grouped and similar sequences can be submitted individually. This allows researchers to utilize them as they see fit in a way that is applicable to their own investigation and set guidelines which may contain different assembly parameters.

## **Library Utility**

The ESTs generated by the production of the *F. vesca* cDNA library are an important resource to the genomics community because the submission set contained thousands of different sequences that could potentially be useful to investigators throughout the world. Sequences have already been utilized for other scientific investigations within the lab of Tom Davis at the University of New



Hampshire, such as the ribosomal protein genes that were used within this investigation for mapping studies. Additionally, three of these YW cDNA sequences have been converted into probes used to extract corresponding clones from a *Fragaria vesca* genomic library constructed in a fosmid vector. These three cDNA sequences included: a CC-NBS-LRR disease resistance gene (CAE46486), the granule-bound starch synthase gene (CAA06958), and the *pistillata* floral development gene (CAC28022). The three respective fosmid clones have now been completely sequenced, providing a by-product the complete genomic sequences corresponding to the three cDNA clones in question, as well as the identities of their immediate genomic neighbors. ESTs are also providing assistance towards the annotation of homologous genomic sequence from the sequenced fosmid clones since ESTs by definition are expressed sequences.

This library is also interesting because it contains the first set of *Fragaria* genes from developing flower buds. Existing GenBank submissions were the result of cloning and sequencing seedlings, leaflets, or fruit from *Fragaria* species. This is significant because cDNA libraries are tissue specific and can be used to compare gene expression patterns between tissues. There are not really enough genes sequenced in this library, or any other existing library, to do comparative studies at this point. However, it is possible that the unique sequences present in this library are the result of some genes which are flower bud specific or that are up-regulated during bud development making them more likely to be cloned from this tissue type. More complete sequencing projects

would be needed to further elucidate the relationships between different libraries and tissue specific gene expression in *Fragaria*.

## **Ribosomal Protein Gene Markers**

Ribosomal protein genes were selected for mapping marker development for several reasons. Because of their highly conserved coding sequences, PCR primers developed for strawberry may be transferable to other species within the Rosaceae, enhancing their value as genomic tools. The entire complement of ribosomal protein genes has been characterized in *Arabidopsis*, providing an invaluable comparator against which the corresponding *Fragaria* orthologs can be identified. Homology searches resulted in approximately 70 different ribosomal protein genes being identified within the cDNA library. The presence of these genes may be indicative of their up-regulation in developing tissues, where it has been proposed that the synthesis of ribosomal protein genes is coordinated with synthesis of ribosomal RNAs (Beltrán-Peña et al, 1995). There were relatively equivalent numbers of 40S small subunit and 60S large subunit ribosomal protein genes. Most genes represented single members of gene families. However there were several gene families in which more than one member was present in the library. An example is RPL30 which has three family members in *Arabidopsis* designated "A", "B", and "C". Members "A" and "B" were represented in the *F. vesca* library. In this study, it was not attempted to distinguish individual family members from one another with the primer pairs used to target ribosomal protein genes so there was a possibility that more than

one product would result from PCR. A BLASTn search was performed with the ribosomal protein gene forward primer sequences against the EST database. There was only one instance in which a search returned more than one hit for a singlet sequence (contigs would be expected to have more than one match in GenBank since they are assemblies of redundant sequences). The forward primer 6N19, which amplifies ribosomal protein RPL11D, had a match for another *F. vesca* EST that has homology to RPL11A. This sequence was not polymorphic between either set of mapping parents, so it was not investigated any further.

Primers were designed to target UTRs and/or to flank the locations of introns based their location of these features in publicly available *Arabidopsis* sequences from GenBank. Due to the high conservation of ribosomal protein gene coding sequences, it was of interest to know whether intron positions and sizes would be similar in *F. vesca*. Most of the genes screened had larger PCR products than could be accounted for by the predicted exon size (based on YW cDNA sequence which lacked introns), suggesting that introns were represented in the respective products (Table 8). There were only six examples (RPP0B, RPS23B, RPS5A, RPL34A, RPL39C, and RPL3A) where the approximate exon size and the observed PCR product size were roughly the same (<100 bp difference) when a larger product with the inclusion of an intron was predicted. Even when no intron was included (RPL27aC, RPL32A, Contig169, and RPS7A), observed PCR product size still deviated from the predicted size (by ~100 bp, 3 bp, 235 bp, 125 bp respectively). This suggests that for these particular genes

intron number or position is not conserved between *Fragaria* and *Arabidopsis*. It is also possible that gene family members that were not the target (i.e., paralogs rather than orthologs) were amplified by the primers used. This could result in the difference between expected and observed PCR product size since intron number, length, and position can differ between members of the same gene family.

For the remaining 21 primer pairs that produced amplification products, the larger than expected product sizes may have been the result of large introns in the sequence. While the location of introns was evidently conserved, their apparent sizes in comparison to those in *Arabidopsis* almost never was. There were only six instances where the difference between predicted PCR product size based on *Arabidopsis* and observed PCR product sizes was less than 100bp. Based on the apparent conservation between intron location, it seems as though *Arabidopsis* sequences can be used effectively to target regions of ribosomal protein genes in other species, but not to predict the size of those regions. Clearly intron size is not as conserved among species as are exon amino acid sequences in ribosomal protein genes. This is not a surprising result, as similar findings have been reported by other researchers.

Despite the potential advantage of genes with highly conserved coding sequences as sites for primer design, this attribute may also correlate with higher conservation of intron sequences within species - a disadvantage in the search for easily genotyped polymorphisms needed for marker development and mapping. Studies of other gene categories in *Fragaria* have resulted in a much

higher percent of polymorphic genes than was found with this ribosomal protein gene set. For instance, in a study of fruit color and anthocyanins pathway genes in *Fragaria vesca* by Deng and Davis (2001), each of the six genes examined was found to have intron length polymorphisms between *F. vesca* and *F. nubicola* mapping parents and could be used for mapping. These parents represent the same two species used for the interspecific FDP815 x FDP601 mapping population. This may suggest that ribosomal protein genes may be a more difficult gene category for mapping, at least via the PCR-based methods employed here. Of the 26 genes that amplified using PCR in this investigation, only 6 were found to be polymorphic in at least one pair of mapping parents, some only after the application of CAPS techniques. The genes that were polymorphic between mapping parents did not have many features in common regarding number of introns and their size. It was noticed however that for three of the six polymorphic genes the reverse primer was located in the 3' UTR. Another trend was that five of the six were from the large ribosomal subunit (out of 15 large subunit genes examined). Only one polymorphic gene was uncovered from the small subunit (out of 13 small subunit genes examined). It is possible that these characteristics may prove to be useful in selecting candidate mapping genes for future studies, but the sample size is too small at this point to determine if there is a definitive correlation.

## **Polymorphism Detection**

The small proportion of ribosomal protein gene polymorphisms discovered in this study may be indicative of high sequence conservation, even in non-coding regions, and that these genes would be very difficult to map in the populations examined using the types of techniques employed. It is also possible that SNPs exist but have not been identified by the enzymes used in this study or by using the techniques employed. The likelihood of discovering SNPs to be used for genotyping a mapping population could be greatly increased by refining the methods used in this study. If the current methods were continued for reasons of time or cost effectiveness, a larger number of genes would have to be screened using a larger number of enzymes with CAPS. This would increase the chances of a polymorphism being detected.

A more direct approach would be to sequence each PCR product being examined for each mapping parent. The PCR products might have to be cloned to facilitate clean sequencing. The resulting sequences could then be searched and compared for polymorphic restriction enzyme sites. Then only a single, appropriate enzyme would be needed, upon confirmation of a respective CAPS polymorphism between the parents. As long as the polymorphism was seen in the CAPS product, the enzyme could then be used to genotype the F<sub>2</sub> segregating population. This would take more effort in the way of background work, as well as create a larger financial burden, but it would streamline the polymorphism discovery process. It would almost guarantee that SNPs would be detected if present in any single gene and once identified would provide an easy

way to screen for them. This is in contrast to the “shot in the dark” method of just trying a number of enzymes and hoping to catch a SNP by chance.

It was noticed that in many of the PCRs and restriction digests, there were faint bands present that were polymorphic between the mapping parents. The presence of these bands complicated the effort to distinguish true polymorphisms and may account for some instances of failure of restriction fragment sizes from adding up to the undigested DNA fragment size from PCR. This is noticeable for example with RPL23B PCR product digest with *AluI*, *Bfal*, and *HaeIII* (Figure 15). There are several possible reasons why this might have happened. Firstly, digestion may have been incomplete. Some of the enzymes used were several years old and therefore may not have been working optimally. This would have resulted in the larger faint bands that appeared on the gels, such as with RPL27aC (Figure 13). It would have been useful to digest DNA of a known sequence containing restriction sites for the enzymes used so that the DNA fragments produced could be examined and used as a control.

Another set of factors that may have prevented enzymes from properly cutting were the conditions of the CAPS reactions. Star activity, or relaxed substrate specificity, can occur when enzyme concentrations are too great, pH is changed, organic solvents are present, glycerol concentrations exceed 5%, or digestion time is too long (NEB, 2006). Since CAPS uses PCR products that are not cleaned up, it is possible that reaction conditions caused nonspecific enzyme cutting, or that impeded the enzyme's functions. This would result in smaller than

expected fragment sizes. Again, DNA of a known sequence and digestion pattern could have been used as a control to eliminate this possibility.

It is difficult to determine which of these reasons, or combination of them, may be the cause of the faint bands because there was no basis to know the expected fragment sizes for these digests. It was unknown if the enzymes used would even cut the DNA fragments used for PCR since it was unknown if they contain the corresponding restriction sites. To alleviate these possibilities, the fragments could be sequenced and searched for sites as described above so that it could be determined what the expected band sizes should be. For PCR products, more than one band may have been present due again to potential contamination of DNA or reagents used by multiple researchers within the lab. It is also possible that more than one gene was being amplified by the primer pairs. This is a distinct possibility with ribosomal protein genes since they occur in multigene families. Genes from the same family may be amplified by the same primer pair and could co-migrate on a gel, but then digest differently. Sequencing of all PCR products seen on a gel could help elucidate this situation.

## **Mapping Populations**

Two pairs of mapping parents were compared in this investigation to evaluate the relative usefulness of the respective intraspecific versus interspecific mapping populations. While linkage maps already exist for both populations, the more convenient population for use by the Davis lab is the intraspecific *F. vesca* (YW) x *F. vesca* (PAWT) population. This is partially due to the fact that this



intraspecific population was developed and is maintained at the University of New Hampshire by Tom Davis and his lab. This population is also advantageous because of the close relationship of the parents, though this makes them less likely to have polymorphisms between them. The wider *F. vesca* (FDP815) x *F. nubicola* (FDP601) interspecific cross was expected to have a higher extent of polymorphism because the parents were more distantly related. This expectation was not entirely supported by the data though. In addition to the inconvenience of possessing only parental DNA samples but not the mapping population plants themselves, there was concern that possible chromosomal rearrangements between the two respective species might introduce discrepancies into the resulting linkage map.

It was found that by using the methods employed in this investigation, neither population had a high number of polymorphic genes, nor was there a great difference between the number discovered in each. Screening of the intraspecific parents yielded three genes that could be mapped, whereas four were found using the interspecific parents. The main distinction between the two pairs of mapping parents was the ease in which polymorphisms were uncovered. All three of the polymorphic genes in the intraspecific parents required the use of CAPS. None of them were identified using PCR alone. On the other hand, all four of the polymorphic genes in the interspecific population could be seen after PCR alone, and three could be detected with CAPS as well. This suggests that the genes that are polymorphic within the intraspecific population have fewer SNPs and indels between the two parents' sequence. This would be different in

the interspecific population, where length polymorphisms and SNPs can be uncovered using several methods, therefore showing that more than one exists. The sequence is more polymorphic between the interspecific parents. This is not surprising though since they are different species.

In conclusion, similar numbers of ribosomal protein gene markers are available for addition to each of the respective linkage maps; however, this could be done more easily in the interspecific population, perhaps using PCR alone for marker genotyping.

The markers produced in this investigation will be a useful addition to the both linkage maps by providing them with gene-based anchors that are part of a Conserved Ortholog Set. If modifications were made to the methods used allowing for easier CAPS enzyme selection, it maybe be just as easy to use the intraspecific population. A larger sample size would need to be used to determine the utility of these two populations in comparison to one another.

## **Summary and Completion of Objectives**

This project was intended to develop and evaluate some much needed genomic resources for *Fragaria vesca*. These resources will be useful not only to investigators studying *F. vesca*, but will also enhance its value as a model diploid system for the cultivated strawberry and the Rosaceae family. The objectives set forth at the beginning of this project were divided into two categories: I. The development of ESTs using a cDNA library; and II. Utilization of ESTs as gene-based markers for linkage maps.

Part I was fulfilled by the construction of a cDNA library and generation of expressed sequence tags. The high quality cDNA library was produced from mRNA isolated from 'Yellow Wonder' flower buds, producing 2717 expressed sequence tags that were submitted to GenBank, of which 1910 were unigenes. At the time, this submission was the largest number of *F. vesca* sequences contributed by one lab to the database. The library could be mined for more data and analyzed for content not explored in this study. There are also 4443 clones left that could be sequenced, potentially adding thousands more sequences to the dataset. The available sequences will undoubtedly be useful to continued research on this project and to projects in other laboratories.

Part II was completed by identifying and evaluating ribosomal protein genes as candidate mapping markers and by comparing the extent of polymorphism in two mapping populations. Primer pairs were designed to target various regions of 28 ribosomal protein genes in two mapping populations. Six genes were found to be polymorphic and are now available for genotyping and mapping in the respective segregating populations. Three polymorphic genes were identified in the intraspecific population using CAPS methods and four were identified in the interspecific population using PCR alone and the CAPS technique. The results of this investigation provide new, gene-based markers for the linkage maps of both mapping populations. Though some ribosomal protein genes useful for mapping were identified, they were not as easily identified as originally hoped. As previously mentioned, genes of different ontology categories have been mapped with greater ease. This was seen with metabolic genes

involved with strawberry fruit color in a study by Deng and Davis (2001). In their study, all of the six genes examined were able to be mapped, whereas in this investigation only six primer pairs out of thirty resulted in an observed polymorphism that could be genotyped in the available mapping populations. This demonstrates that the highly conserved nature of ribosomal protein genes may make them more difficult to map than genes in other ontology categories within the conserved ortholog set.

## LITERATURE CITED

- Akiyama, Y., Yamamoto, Y., Ohmido, N., Ohshima, M., & Fukui, K. (2001). Estimation of the nuclear DNA content of strawberries (*Fragaria* spp.) compared with *Arabidopsis thaliana* by using dual-step flow cytometry. *Cytologia*, 66(4), 431-436.
- Allen, F.L. (1994). Usefulness of plant genome mapping to plant breeding. In: *Plant Genome Analysis* (P.M. Gresshoff, editor). CRC Press, Ann Arbor. 11-18.
- Barakat, A., Szick-Miranda, K., Chang, I. F., Guyot, R., Blanc, G., Cooke, R., Delseny, M., & Bailey-Serres, J. (2001). The organization of cytoplasmic ribosomal protein genes in the *Arabidopsis* genome. *Plant Physiol*, 127(2), 398-415.
- Beltran-Pena, E., Ortiz-Lopez, A., & Sanchez de Jimenez, E. (1995). Synthesis of ribosomal proteins from stored mRNAs early in seed germination. *Plant Mol Biol*, 28(2), 327-336.
- Bringhurst, R.S. (1990). Cytogenetics and evolution in American *Fragaria*. *Hortscience*, 25(8), 879-881.
- Bringhurst, R.S. and D. A. Khan, (1963). Natural pentaploid *Fragaria chiloensis*-*F. vesca* hybrids in coastal California and their significance in polyploid *Fragaria* evolution. *Amer. J. Biol.* 50: 658-661.
- Bringhurst, R. S. and Tarlock Gill. (1970). Origin of *fragaria* polyploids. II. Unreduced and double-unreduced gametes. *Amer. J. Bot.* 57(8), 969-976.
- Byrne, D., & Jelenkovic, G. (1976). Cytological diploidization in the cultivated octoploid strawberry *Fragaria X ananassa*. *Can J Genet Cytol*, 18(4), 653-659.
- Chee, P. W., Rong, J., Williams-Coplin, D., Schulze, S. R., & Paterson, A. H. (2004). EST derived PCR-based markers for functional gene homologues in cotton. *Genome*, 47(3), 449-462.
- Cooke, R., Raynal, M., Laudie, M., & Delseny, M. (1997). Identification of members of gene families in *Arabidopsis thaliana* by contig construction from partial cDNA sequences: 106 genes encoding 50 cytoplasmic ribosomal proteins. *Plant J*, 11(5), 1127-1140.

- Darrow, G. M. (1950). Polyploidy in fruit improvement. *The Scientific Monthly*, April: 211-219.
- Darrow, G. M. (1966). *The Strawberry*. Holt, Reinhart and Winston Press, NY.
- Davis, T. M., & Yu, H. (1997). A linkage map of the diploid strawberry, *Fragaria vesca*. *Journal of Heredity*, 88(3), 215-221.
- da Silva, J. A. G. and M. E. Sorrells. (1996). Linkage analysis in polyploids using molecular markers. In: *Methods of Genome Analysis in Plants* (Prem P. Jauhar, editor). CRC Press, NY. 211-228.
- de Wet, J.M.J. (1971). Polyploid and evolution in plants. *Taxon*, 20(1):29-35.
- Deng, C. and T.M. Davis. (2001). Molecular identification of the yellow fruit color (c) locus in diploid strawberry: a candidate gene approach. *Theor Appl Genet*, 103:316-322.
- Economic Research Service (ERS). (2005).  
<http://www.ers.usda.gov/Browse/Crops/FruitTreeNuts.htm>
- Evans, W. D. (1982). The production of multispecific octoploids from *Fragaria* species and the cultivated strawberry. *Euphytica*, 31(3), 901-907.
- Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*, 8(3), 175-185.
- Federova, N. J. (1946). Crossability and phylogenetic relations in the main european species of *Fragaria*. *Compt. Rend. (Doklady). Acad. Sci. U.S.S.R.* 52: 545-547.
- Folta K.M. and Davis T.M. 2006. Strawberry genes and genomes. *Curr. Opinions Plant Biol.* Accepted.
- Folta, K. M., Staton, M., Stewart, P. J., Jung, S., Bies, D. H., Jesdurai, C., & Main, D. (2005). Expressed sequence tags (ESTs) and simple sequence repeat (SSR) markers from octoploid strawberry (*Fragaria x ananassa*). *BMC Plant Biol*, 5, 12.
- Food and Agriculture Organization (FAO). (2005).  
<http://faostat.fao.org/faostat/collections?version=ext&hasbulk=0&subset=agriculture>

- Frary, A., Xu, Y., Liu, J., Mitchell, S., Tedeschi, E., & Tanksley, S. (2005). Development of a set of PCR-based anchor markers encompassing the tomato genome and evaluation of their usefulness for genetics and breeding experiments. *Theor Appl Genet*, 111(2), 291-312.
- Galletta, G. J., & Maas, J. L. (1990). Strawberry Genetics. *Hortscience*, 25(8), 871-879.
- Gantt, J. S. and M. D. Thompson. (1990). Plant cytosolic ribosomal protein S11 and chloroplast ribosomal protein CS17. *Journal of Biological Chemistry*. 265(5): 2763-2767.
- Gottlieb, L. D. (2003). Plant polyploidy: gene expression and genetic redundancy. *Heredity*, 91(2), 91-92.
- Gualerzi, C., Janda, H. G., Passow, H., & Stoffler, G. (1974). Studies on the protein moiety of plant ribosomes. Enumeration of the proteins of the ribosomal subunits and determination of the degree of evolutionary conservation by electrophoretic and immunochemical methods. *J Biol Chem*, 249(11), 3347-3355.
- Hadonou, A. M., Sargent, D. J., Wilson, F., James, C. M., & Simpson, D. W. (2004). Development of microsatellite markers in *Fragaria*, their use in genetic diversity analysis, and their potential for genetic linkage mapping. *Genome*, 47(3), 429-438.
- Hancock, J. F. (1990). Ecological genetics of natural strawberry species. *Hortscience*, 25(8), 869-871.
- Hancock, J. F. (1999). *Strawberries*. CABI Publ., Oxon.
- Hancock, J. F., & Bringhurst, R. S. (1978). Ecological differentiation in perennial, octoploid species of *Fragaria*. *Genetics*, 88(4), S35-S36.
- Hancock, J. F., & Bringhurst, R. S. (1981). Evolution in California populations of diploid and octoploid-*Fragaria* (Rosaceae) - a comparison. *American Journal of Botany*, 68(1), 1-5.
- Hancock, J. F., & Luby, J. J. (1993). Genetic resources at our doorstep - the wild strawberries. *Bioscience*, 43(3), 141-147.

- Huang, X., & Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Res*, 9(9), 868-877.
- Iwatsubo, Y., & Naruhashi, N. (1989). Karyotypes of 3 Species of *Fragaria* (Rosaceae). *Cytologia*, 54(3), 493-497.
- Jarvis, P., Lister, C., Szabo, V., & Dean, C. (1994). Integration of CAPS markers into the RFLP map generated using recombinant inbred lines of *Arabidopsis thaliana*. *Plant Mol Biol*, 24(4), 685-687.
- Joanin, P., Gigot, C., & Philipps, G. (1993). cDNA nucleotide sequence and expression of a maize cytoplasmic ribosomal protein S13 gene. *Plant Mol Biol*, 21(4), 701-704.
- Johnson, Jr., Harold A. (1990). The contributions of private strawberry breeders. *HortScience*, 25(8), 879-999.
- Jung, S., Jesudurai, C., Staton, M., Du, Z., Ficklin, S., Cho, I., Abbott, A., Tomkins, J., & Main, D. (2004). GDR (Genome Database for Rosaceae): integrated web resources for Rosaceae genomics and genetics research. *BMC Bioinformatics*, 5, 130.
- Kim, Y., Zhang, H., & Scholl, R. L. (1990). Two evolutionarily divergent genes encode a cytoplasmic ribosomal protein of *Arabidopsis thaliana*. *Gene*, 93(2), 177-182.
- Konieczny, A., & Ausubel, F. M. (1993). A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. *Plant J*, 4(2), 403-410.
- Konovalov, F., Toshchakova, E., & Gostimsky, S. (2005). A CAPS marker set for mapping in linkage group III of pea (*Pisum sativum* L.). *Cell Mol Biol Lett*, 10(1), 163-171.
- Kunihisa, M., Fukino, N., & Matsumoto, S. (2005). CAPS markers improved by cluster-specific amplification for identification of octoploid strawberry (*Fragaria x ananassa* Duch.) cultivars, and their disomic inheritance. *Theor Appl Genet*, 110(8), 1410-1418.
- Lawrence, F. J., Galletta, G. J., & Scott, D. H. (1990). Strawberry breeding work of the United States Department of Agriculture. *Hortscience*, 25(8), 895-896.



- Lerceteau-Kohler, E., Guerin, G., Laigret, F., & Denoyes-Rothan, B. (2003). Characterization of mixed disomic and polysomic inheritance in the octoploid strawberry (*Fragaria x ananassa*) using AFLP mapping. *Theor Appl Genet*, 107(4), 619-628.
- Lewers, K. S., Styan, S. M. N., Hokanson, S. C., & Bassil, N. V. (2005). Strawberry GenBank-derived and genomic simple sequence repeat (SSR) markers and their utility with strawberry, blackberry, and red and black raspberry. *Journal of the American Society for Horticultural Science*, 130(1), 102-115.
- Longley, A. E. (1926). Chromosomes and their significance in strawberry classification. *Jour. Agric. Res.* 32: 559-568.
- Manning, K. (1998). Isolation of a set of ripening-related genes from strawberry: their identification and possible relationship to fruit quality traits. *Planta*, 205(4), 622-631.
- Matsumoto, A. and Y. Tsumura. (2004). Evaluation of cleaved amplified polymorphic sequence markers for *Chamaecyparis obtusa* based on expressed sequence tag information from *Cryptomeria japonica*. *Theor Appl Genet*, 110:80-91.
- Michaels, S. D., & Amasino, R. M. (1998). A robust method for detecting single-nucleotide changes as polymorphic markers by PCR. *Plant J*, 14(3), 381-385.
- Morales, M., Roig, E., Monforte, A. J., Arus, P., & Garcia-Mas, J. (2004). Single-nucleotide polymorphisms detected in expressed sequence tags of melon (*Cucumis melo* L.). *Genome*, 47(2), 352-360.
- Moran, D. L. (2000). Characterization of the structure and expression of a highly conserved ribosomal protein gene, L9, from pea. *Gene*, 253(1), 19-29.
- Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol*, 51 Pt 1, 263-273.
- Mullis, K. B., & Faloona, F. A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol*, 155, 335-350.
- National Center for Biotechnology Information (NCBI). (2005). [www.ncbi.nih.gov](http://www.ncbi.nih.gov)

- Nam, Y. W., Tichit, L., Leperlier, M., Cuerq, B., Marty, I., & Lelievre, J. M. (1999). Isolation and characterization of mRNAs differentially expressed during ripening of wild strawberry (*Fragaria vesca* L.) fruits. *Plant Mol Biol*, 39(3), 629-636.
- Neff, M. M., Neff, J. D., Chory, J., & Pepper, A. E. (1998). dCAPS, a simple technique for the genetic analysis of single nucleotide polymorphisms: experimental applications in *Arabidopsis thaliana* genetics. *Plant J*, 14(3), 387-392.
- New England Biolabs Inc. (NEB). (2006).  
[www.neb.com/nebecomm/tech\\_reference/restriction\\_enzymes/star\\_activity.asp](http://www.neb.com/nebecomm/tech_reference/restriction_enzymes/star_activity.asp)
- Patterson, J. T., Larson, S. R., & Johnson, P. G. (2005). Genome relationships in polyploid *Poa pratensis* and other *Poa* species inferred from phylogenetic analysis of nuclear and chloroplast DNA sequences. *Genome*, 48(1), 76-87.
- Pearson, W. R., & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8), 2444-2448.
- Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., & Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239(4839), 487-491.
- Senanayake, Y. D. A. and R. S. Bringham. (1967). Origin of polyploids. I. cytological analysis. *Amer. J. Bot.* 54(2): 221-228.
- Sargent, D. J., Davis, T. M., Tobutt, K. R., Wilkinson, M. J., Battey, N. H., & Simpson, D. W. (2004). A genetic linkage map of microsatellite, gene-specific and morphological markers in diploid *Fragaria*. *Theor Appl Genet*, 109(7), 1385-1391.
- Sargent, D.J., J. Clarke, D.W. Simpson, K.R. Tobutt, P. Arus, A. Monfort, S. Vilanova, B. Denoyes-Rothan, M. Rousseau, K.M. Folta, N.V. Bassil, and N.H. Battey. (2006). An enhanced microsatellite map of diploid *Fragaria*. *Theor Appl Genet*. In press.
- Sasaki, T., Song, J., Koga-ban, Y., Matsui, E., Fang, F., Higo, H., Nagasaki, H., Hori, M., Miya, M., & Murayama-Kayano, E. (1994). Toward cataloguing all rice genes: large-scale sequencing of randomly chosen rice cDNAs from a callus cDNA library. *Plant J*, 6(4), 615-624.

- SOL Genomics Network (SOL). (2005). <http://www.sgn.cornell.edu/index.pl>
- Staudt, G. (1989). The species of *Fragaria*, their taxonomy and geographical distribution. *Acta Horticulturae*, 265(1): 23-33.
- United States Department of Agriculture (USDA). (2001). [www.fas.usda.gov/htp/horticulture/Apples/apple01.pdf](http://www.fas.usda.gov/htp/horticulture/Apples/apple01.pdf)
- Ying, S. Y. (2004). Complementary DNA libraries - An overview. *Molecular Biotechnology*, 27(3), 245-252.
- You, T. H., & Scholl, R. L. (1998). PCR amplification of cDNA libraries for cloning and screening. *Biotechniques*, 24(4), 574-575.
- Yu, J. K., Dake, T. M., Singh, S., Benscher, D., Li, W., Gill, B., & Sorrells, M. E. (2004). Development and mapping of EST-derived simple sequence repeat markers for hexaploid wheat. *Genome*, 47(5), 805-818.
- Williamson, S. C. (1993). Germplasm development and assessment in the diploid strawberry species *Fragaria vesca* L. (Master's thesis). Durham, NH: University of New Hampshire.
- Wu, J., Matsui, E., Yamamoto, K., Nagamura, Y., Kurata, N., Takuji, S., & Minobe, Y. (1995). Genomic organization of 57 ribosomal protein genes in rice (*Oryza sativa* L.) through RFLP mapping. *Genome*, 38(6), 1189-1200.