5-2007

BIOINFORMATICS TOOL DEVELOPMENT AND SEQUENCE ANALYSIS OF ROSACEAE FAMILY EXPRESSED SEQUENCE TAGS

Margaret Staton
Clemson University, mestato@yahoo.com

BIOINFORMATICS TOOL DEVELOPMENT AND SEQUENCE ANALYSIS OF ROSACEAE FAMILY EXPRESSED SEQUENCE TAGS

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Plant and Environmental Science

---

by
Margaret E. Staton
May 2007

Accepted by:
Dr. Doreen Main, Committee Chair
Dr. Bert Abbott
Dr. Vance Baird
Dr. Halina Knap

ABSTRACT

BACKGROUND: An international community of researchers has generated a significant number of Expressed Sequence Tags (ESTs) for the Rosaceae, an economically important plant family that includes most temperate fruits such as apple, cherry, peach, and strawberry as well as other commercially valuable members. ESTs are fragments of expressed genes that can be used for gene discovery, developing markers for mapping and cultivar improvement via marker assisted selection.

DESCRIPTION: The Genome Database for Rosaceae (GDR) was initiated to provide a curated and integrated web-based relational database for this family. I developed a key component of GDR to assemble and annotate the publicly available ESTs from the four main genera of the family (*Prunus*, *Malus*, *Fragaria*, *Rosa*). I created both genera and family level unigenes using the software CAP3 after extensive filtering, trimming and assembly. Further analysis includes marker mining for single nucleotide polymorphisms (SNPs) and simple sequence repeast (SSRs) with putative primer identification, and oligo identification for potential microarray development. Functional genomics efforts are supported with sequence similarity searching against major protein and nucleotide databases, gene product ontology assignment, and protein motif identification. I deployed the entire project on the GDR with all data available for browsing, searching, and downloading.

CONCLUSIONS: The GDR and its associated EST unigene project are meeting a major need for timely annotation and curation of sequence data for the Rosaceae community. The results of my analysis highlight major genes and pathways of interest including ripening, disease resistance, and transcription factors. The easily accessible pool of annotated coding sequences should further both functional and structural genomics characterization in Rosaceae. The unigene elucidates the levels of sequence similarity shared across different plant species and the implications for resource sharing across the family. GDR can be accessed at http://www.rosaceae.org/.

DEDICATION

I dedicate this work to my parents. Their love and support have made this work as well as all my success and happiness possible.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Doreen Main, for all of her guidance and patience. I am grateful to the other committee members: Dr. Bert Abbott, Dr. Vance Baird, and Dr. Halina Knap for their support and help as mentors. I am grateful to Dr. Kevin Folta and Dr. Bryon Sosinski for providing me with research opportunities and engaging discussions.

TABLE OF CONTENTS

Table of Contents (Continued)

LIST OF TABLES

List of Tables (Continued)

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Rosaceae Genomics

In temperate regions of the world, Rosaceae is one of the most economically important plant families. Fruit such as apple, apricot, blackberry, cherry, pear, peach, plum, raspberry, and strawberry are the major products from this family. Another edible member is the almond. The total value from the food production of members of this family is estimated at over $8 billion dollars in the United States in 2006 (National Agricultural Statistics Service, 2006). These crops were grown on over 1.5 million acres, and the most valuable members are almonds, apples, strawberries, and peaches (Table 1.1) This major crop family contributes to a nutritious and diverse diet by adding vitamins, minerals, dietary fiber, and antioxidants. The Rosaceae encompasses other commercially valuable members such as lumber (black cherry), and ornamentals (roses, flowering cherry, crabapple, quince and pear). Sales of these plants in nurseries contribute even more to the domestic and international value of this family.

Table 1.1: Rosaceae statistics from the United States Agricultrual Statistics Service, estimates for 2006

Crop (Location)	Bearing Acreage in US in 2006	Value of Production in 2006 ($1000s of dollars)
Almonds (CA)	580,000	2,198,215
Apples	377,490	2,099,129
Apricot	15,540	29,580
Blackberries (OR)	6,900	35,380
Boysenberries	920	7,128
Loganberries (OR)	60	100
Raspberries, Black (OR)	1,500	9,780
Raspberries, Red	11,500	25,346
Raspberries, All (CA)	4,300	249,615
Cherries, Sweet	81,300	487,482
Cherries, Tart	37,200	53,453
Nectarines	36,900	124,200
Peaches	134,460	513,438
Pears	59,850	324,885
Plums (CA)	36,000	110,217
Prunes (CA)	67,000	240,784
Prunes & Plums	3,400	8,763
Strawberries	53,280	1,514,998
TOTAL	1,507,600	8,032,493

The Rosaceae family has a worldwide distribution and encompasses over 3000 species. This family is part of the order Rosales, found within the eurosid I clade of flowering plants (Figure 1.1). It is estimated that this group diverged from other Rosales around 76 Mya (Wikstrom et al., 2001). Four subfamilies are generally recognized based on fruit classification (e.g., Schulze-Menz 1964). The Rosoideae, containing strawberries, roses, blackberries and raspberries bear

indehiscent fruit. A group of ornamental shrubs and other species with dry dehiscent fruit constitute the Spiroideae. The Prunoideae include the species producing fleshy one-seeded fruits such as cherries, almonds, peaches, plums and apricots. Apples and pears are pome fruits typical of those that fall into the last subfamily, Maloideae.

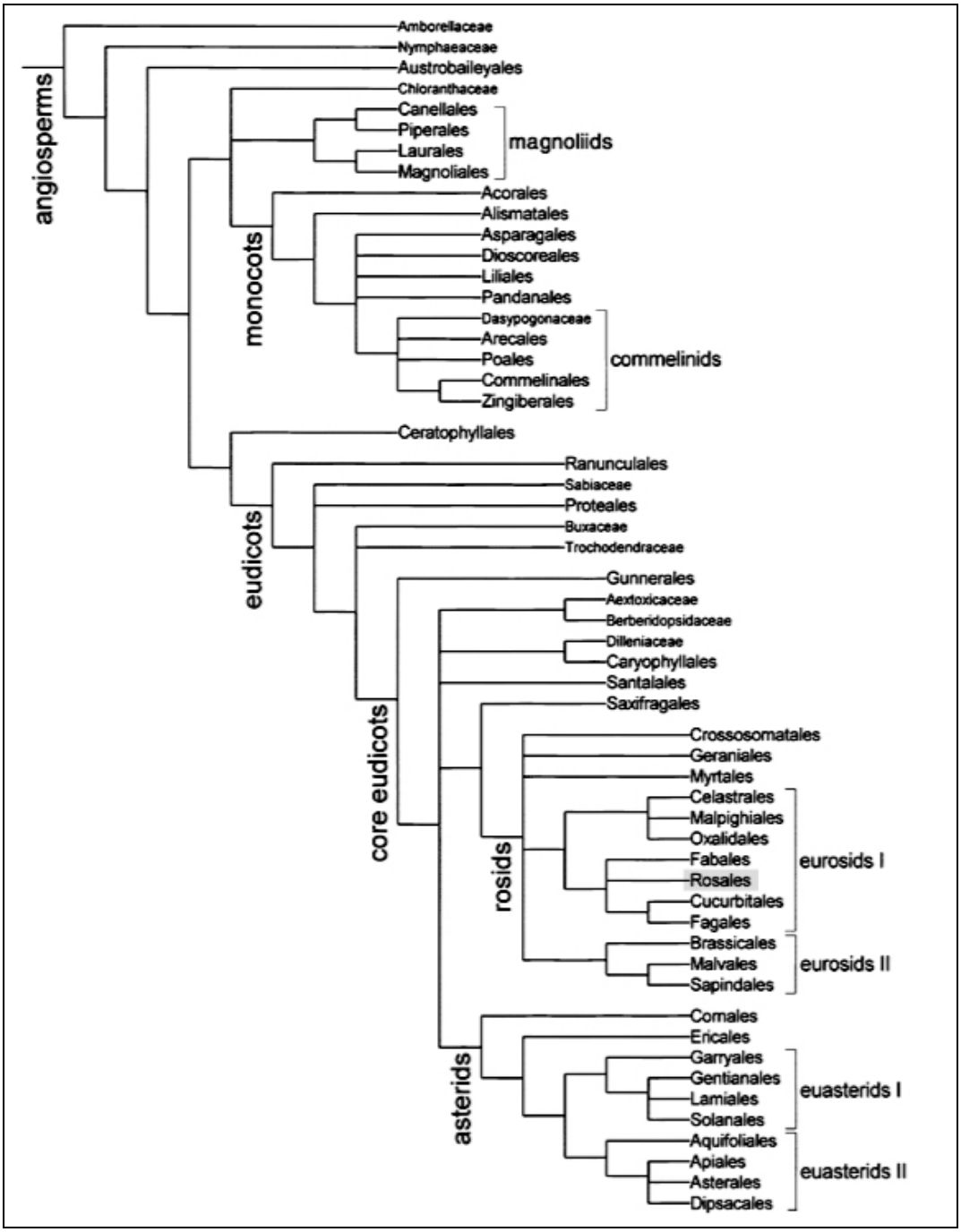


Figure 1.1: Most recent classification of orders and families of flowering plants from theAngiosperm Phylogeny Group. Interrelationships are supported by jackknife or bootstrap frequencies above 50% in large-scale analyses of angiosperms. (Figure from Angiosperm Phylogeny Group, 2003)

Despite the anatomical evidence that the Rosaceae should be divided into these four groups, recent phylogenies created from molecular information have not upheld them as accurate evolutionary divisions (Morgans et al, 1994; Evans et al, 2000). The most recent and comprehensive examination of molecular data from Potter et al., 2002 utilized parsimony analysis of sequence data from the *mat*K and the *trn*L-*trn*F region of the chloroplast genome. Three main clades were indentified: Rosoideae *sensu stricto*, actinorhizal Rosaceae, and the rest of the family (Figure 1.2). Basic subfamilies of Maloideae and Rosoideae were upheld with some modifications. A somewhat clearer picture of the evolution of the family emerged from this study but the positions of many subgroups remain unresolved.

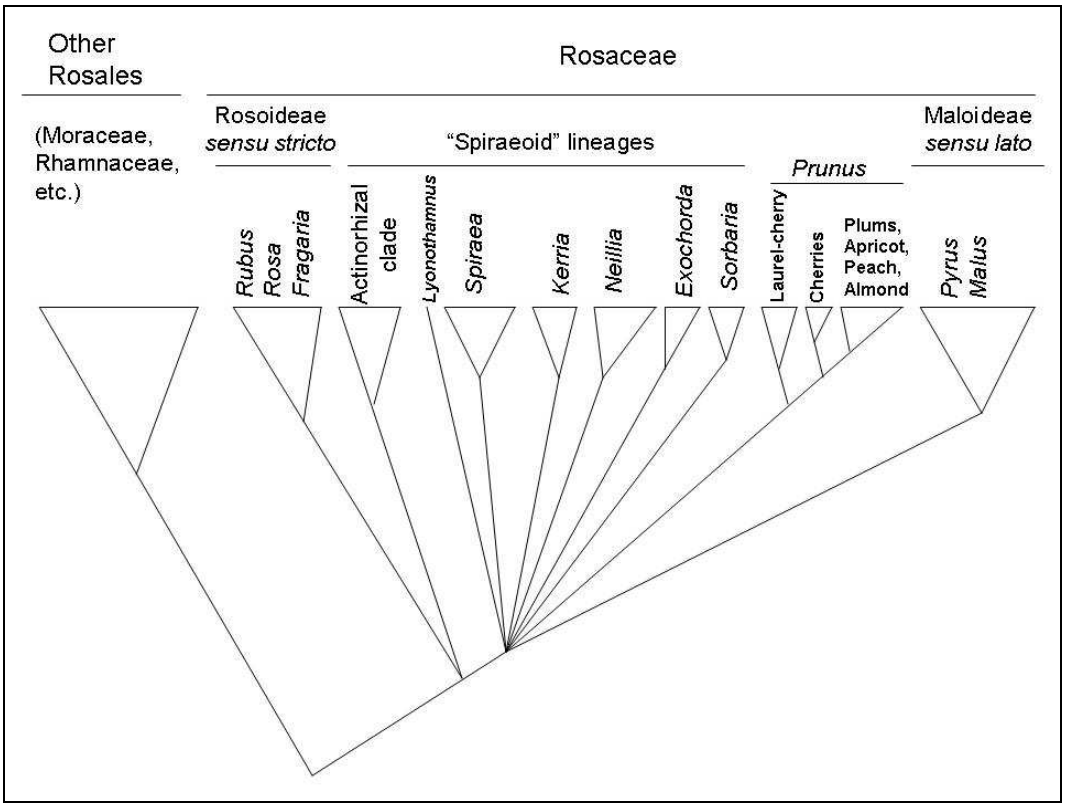


Figure 1.2: Rosaceae lineages (courtesy of Dan Potter, 2002)

The Rosaceous crop industry is facing multiple challenges to efficient and profitable production. The industry has identified key areas of crop improvement that needs to be addressed (US Rosaceae Genomics, Genetics and Breeding Consortium, 2006). A major emphasis is to improve key qualities of fresh and processed fruit such as taste, aroma, color, and freedom from defects. Post-harvest quality is another pressing issue due to the highly perishable nature of fresh fruit and thier susceptibility to aging, decay, and chilling injury.

Rosaceous crops are susceptible to multiple different types of pests and diseases that can cause high economic losses. This susceptibility has led farmers to depend on high levels of chemical pesticides that are expensive and potentially

damaging to the environment (Janick and Moore, 1996). More resistant crop varieties would help alleviate both of these issues by preventing the initial pest and disease problems. Growers also desire varieties that are resistant to common abiotic stresses such as drought and cold (Janick and Moore, 1996).

Previous efforts to improve varieties of Rosaceous species have largely depended on traditional breeding techniques. This has proved difficult in many of the species due to long generation times, high space requirements, and polyploid genomes. Apples planted from seed go through a juvenile phase when they do not produce flowers; this phase may last from three to ten or more years. Certain peach varieties take up to five years to begin fruiting (Janick and Moore, 1996a). Strawberries are vegetative, and their breeding has yielded many cultivars over a short time as new generations can be produced each year (Janick and Moore, 1996b). More advanced breeding technologies that utilize genomic tools could make significant gains in many of these crops, especially the woody ones, over a shorter time period. Understanding the genes in fruits and how they interact to produce desired phenotypes would allow selection of varieties with the most favorable combinations of gene variants. This can be accomplished through marker-assisted breeding techniques, manipulation of gene expression, or inclusion of new genes in the peach genome. Of all of these techniques, a marker-assisted selection program for breeders may be the most important because it can be implemented immediately to yield important results (Dirlewanger et al., 2004b). Marker assisted selection allows the breeder to select a subset of seedlings with known desirable traits and grow only this subset for

further evaluation. A small tissue sample can be taken from all seedlings, and DNA analysis of known loci will reveal the combination of alleles that are encoded in each of the progeny. This efficiency of time and space will allow diverse germplasm with specific desirable traits to be included in breeding populations. Markers for the common traits found in wild types such as small fruit size or low quality could be used to screen progeny at the seedling stage. New alleles from these new genotypes could provide better disease resistance and other phenotypes not yet exploited in commercial cultivars.

Utilizing input from industries, scientists, and government agencies, a White Paper for the US Rosaceae Genomics, Genetics and Breeding Initiative has been released (US Rosaceae Genomics Genetics and Breeding Consortium, 2006). An international vision for increasing and integrating Rosaceae improvement and research is due to be published in 2007. Overall, these initiatives have concluded that Rosaceous genomes must be analyzed and exploited, genomic database resources for the community must be enhanced, and breeding programs must be revitalized (US Rosaceae Genomics, Genetics and Breeding Consortium, 2006). To date, three species, peach, apple, and strawberry, are the primary focus of most of the genomics efforts. These represent diverse subfamilies and some of the most economically important crops worldwide.

While many structural and functional genomic resources are already available for *Prunus*, *Malus* and *Fragaria* species, more research and funding is needed. The community has a centralized data repository, the Genome Database

for Rosaceae (GDR), to disseminate the publicly available genomic data for this family (Jung et al, 2004). Initial studies suggest that there is a significant sequence synteny within *Prunus* and across the Rosaceae (Dirlewanger et al., 2004a; Dirlewanger et al., 2004b). Efforts are being focused on integrating Rosaceous genomic information from individual species across the family using comparative genomics. Breeding programs can accelerate the use of marker assisted selection and other molecular techniques to incorporate the current genomics information in new varieties.

Structural Resources

Structural genomics refers to the physical structure and organization of a genome. Knowledge of structural genomics is necessary for manipulating genes and DNA segments in genomic studies. The basic haploid chromosome number of the economically important Rosaceae species is known. The *Prunus* genus, which encompasses peach, plum, almond, apricot and cherry, has 8 chromosomes (2n=2x=16) (Jelenkovic and Harrington, 1972). The strawberries grown commercially, *Fragaria x ananassa*, are an octoploid member of the *Fragaria* genus. Diploid strawberry species such as *F. vesca* and *F. nubicola* are generally accepted as primary candidates for physical genomic research due to the relative simplicity of developing and interpreting diploid maps (Sargent et al., 2004). These *Fragaria* species as well as *F. x ananassa* have a haploid chromosome number of 7 (Jelenkovic and Harrington, 1972). Apple and pear are from a lineage that is assumed to have undergone whole genome duplication since divergence from the other Rosaceae members and has a haploid chromosome number of 17 (Lespinasse et al., 1976).

A basic resource for many genomic studies is a genetic linkage map where genetic distance is measured in centimorgans. These maps are used to translate genomic information into molecular markers for breeding. An integrated reference map for *Prunus* has been adopted and contains 562 markers (Joobeur et al., 1998; Aranzana et al., 2003; Dirlewanger et al., 2004a). This map was created from an interspecific cross of peach and almond, and it currently spans 519 cM. Other *Prunus* maps are also available, including peach, apricot, sweet cherry, sour cherry, Myrobalam plum, and almond. A map spanning 424 cM with 182 markers is available for diploid strawberry (Sargent et al. 2006). Strawberry represents an attractive mapping system due to its self-compatibility, small genome, and short generation time, but it has only a medium level of marker saturation thus far. Several apple maps have been developed, and the most comprehensive includes over 800 markers (Liebhard et al., 2003). Mapping in apple can be time consuming due to its 6 to 10 year or more generation time (Janick and Moore, 1996). Other Rosaceous species with genetic maps include rose, pear, and raspberry. A list of the major available genetic maps for Rosaceous species is outlined in Table 1.2.

Table 1.2: Genetic Linkage Maps Available for Rosaceae Species

The full citations for these maps can be found in the "Map References" section following the "References" section of this chapter.

Species	Reference	Markers	Map Size
Peach	Chaparro et al., 1994	83	396 cM
	Dirlewanger et al., 1998	249	712 cM
	Lu et al., 1998	153	1297 cM
	Dirlewanger et al., 1999	249	712 cM
	Shimada et al., 2000	87	1000 cM
	Verde et al., 2005	216	665 cM
Peach X Almond	Foolad et al., 1995	107	800 cM
	Joobeur et al., 1998	246	491 cM
	Dettori et al, 2001	109	521 cM
	Bliss et al., 2002	161	1144 cM
	Aranzana et al., 2003	342	522 cM
	Dirlewanger et al., 2004a	166	716 cM
	Howad et al., 2005	264	68 bins
Myrobalan plum	Dirlewanger et al., 2004a	93	525 cM
Sour Cherry	Wang et al., 1998	126	462 cM
		95	279 cM
Sweet Cherry	Stockinger et al., 1996	89	503 cM
Apricot	Hurtado et al., 2002	132	511 cM
		80	467 cM
	Vilanova et al., 2003	211	602 cM
Almond	Viruel et al., 1995	93	393 cM
		69	394 cM
	Joobeur et al., 2000	126	415 cM
		99	416 cM

Table 1.2: Genetic Linkage Maps Available for Rosaceae Species (Continued)

Species	Reference	Markers	Map Size
Apple	Hemmat et al., 1994	360	1120 cM
	Conner et al., 1997	238	1206 cM
		110	
		183	
	Maliepaard et al., 1998	194	842 cM
		163	984 cM
	Liebhard et al., 2003	840	1140 cM
			1450 cM
Diploid Strawberry	Lerceteau-Kohler et al., 2003	235	1604 cM
		280	1496 cM
	Sargent et al., 2004	76	448 cM
	Sargent et al., 2006	182	424 cM
Rose	Mattiesch and Debener, 1999	278	326 cM
			370 cM
	Rajapakse et al., 2001	171	902 cM
		167	682 cM
	Crespel et al., 2002	68	238 cM
		108	287 cM
	Yan et al., 2005	520	487 cM
			490 cM
	Dugo et al., 2005	133	388 cM
			260 cM
Pear	Yamamoto et al., 2002	226	949 cM
		154	926 cM
	Pierantoni et al., 2004	41	
		31	
Raspberry	Graham et al., 2004	273	789 cM

A physical map is another invaluable structural genomic resource that maps the genome in physical distances (base pairs) instead of centiMorgans. A peach framework physical map anchored on the general *Prunus* genetic map is under development (Horn et al., 2005). It currently has 1,899 BAC contigs spanning an estimated 279 Mb of the genome, and it is due to be completed in 2007. Twenty eight trait loci corresponding to agronomic characters from the general *Prunus* genetic map have already been anchored on this physical map (Dirlewanger et al., 2004b). A transcriptome map is being developed and currently has the positions of 1258 ESTs identified by hybridization to ordered BACs (Horn et al., 2005). Two complementary BAC libaries have been constructed for the apple, one of *Malus floribunda* 821 'Florina' (Xu et al, 2001) and one of the cultivar 'GoldRush' (Xu et al., 2002). A physical map is now available with 2702 contigs that span an estimated 927 Mb (Han et al., 2007). The diploid strawberry genome has been integrated into an 8x Fosmid library (Davis, 2006) but no physical map is yet available.

Despite the availability of genetic linkage maps and physical maps, further research is needed in the area of structural genomics. Neutral molecular markers still need to be linked to loci controlling traits of interest and mapping data needs to be integrated into overall reference maps that can be used for anchoring the physical maps. Further research will need to be conducted in comparative genomics to utilize the high levels of synteny expected between family members and facilitate the discovery of transferable markers. More single nucleotide polymorphisms (SNPs) will be required to fully saturate the available maps.

Ultimately, the Rosaceae community needs a fully sequenced genome. *Arabidopsis* with a haploid genome size of only 115 Mb (*Arabidopsis* Genome Initiative, 2000) is the closest plant relative with a fully sequenced and annotated genome but has many limitations. The fruit type, growth habit, and life history of *Arabidopsis* are very different from the members of the Rosaceae. *Populus trichocarpa* is a closer relative and now has a draft sequenced genome with 7.5X coverage, but it has many of the same limitations as *Arabidopsis* (Tuskan et al., 2006). Fortunately, the Rosaceae have comparatively small genome sizes that facilitate a relatively inexpensive genome sequence. The peach genome is 290 Mb (Baird et al., 1994) and strawberry is 164 Mb (Akiyama et al., 2001). Apple is somewhat larger at 769 Mb (Patocchi et al., 1999) but still a relatively small genome when compared to the majority of other crop species.

In January, the Department of Energy's Joint Genome Institute announced they will sequence the peach genome by 2008. This will involve an 8X coverage of the peach double haploid 'Lowell', the same cultivar as the physical map. At the same time an Italian group from the Istituto Agrario San Michele all'Adige announced that they will complete a 4X coverage of the apple cultivar 'golden delicious' by the end of 2007. Both of these genome sequences will be made publicly available. A whole genome sequence will promote genomics research in a profusion of areas including gene and cis-element discovery, transcriptome analysis, epigenetic studies, high-density genotyping, and polymorphism discovery.

Functional Resources

Functional genomics plays an important role in the future of Rosaceae research and improvement. This area focuses on gene expression, gene function, protein structure and interactions, and metabolic network structure. An important goal of the Rosaceae community is to identify and characterize genes controlling or impacting important phenotypic traits (US Rosaceae Genomics Genetics and Breeding Consortium, 2006). The current focus is mostly on fruit characters as fruit is the most prominent source of economic value. Traits of interest include sugar and acid levels, color, firmness and fruit size, self-incompatibility, and biotic and abiotic stress resistance (DeCroocq et al., 2005; Dirlewanger et al., 2004a; Dirlewanger et al., 2004b; Liebhard et al, 2003, Wunsch and Hormaza, 2004). Finding the genes impacting these characters may lead to direct manipulation of crop genetics and improved fruit quality. Rosaceae are mostly perennial species with long maturation times, which makes traditional breeding more difficult. However, if the genes of interest are tagged with molecular markers, they can be used in marker-assisted selection, thus allowing only the seedlings containing the tagged genes of interest to go forward for further field evalaution (Dirlewanger et al., 2004b).

Expressed sequences tags (ESTs) are one of the most valuable functional genomic resources for studying gene expression. By sampling mRNA, these short sequences of expressed genes are able to give researchers a snapshot of the genes being expressed in a particular tissue, at a given time for a particular Rosaceae variety All publicly released ESTs are stored in NCBI's dbEST. This database is a division of GenBank that contains EST library and sequence information

(Boguski et al., 1993). The number of sequences being added to the database has been growing exponentially since the inception of dbEST. As of August 18, 2006 there were 38,266,600 sequences in dbEST, of which more than 10 million represented viridiplantae (green plant) species (http://www.ncbi.nlm.nih.gov /dbEST/dbEST_summary.html). Many communities studying different plant species have produced and utilized ESTs; 35 different plant species have more than 50,000 ESTs currently in dbEST (http://www.ncbi.nlm.nih.gov /dbEST/dbEST_summary.html).

A total of 374,654 ESTs from Rosaceae species were available from GenBank's dbEST on August 21, 2006. These ESTs represent 18 different species spread across 6 genera (Table 1.3) and 156 different libraries. Twenty-two different tissues have been characterized in these libraries (Table 1.4). ESTs can be utilized for identifying candidate genes for different traits, mining for molecular markers such as SSRs and SNPs, and finding relative abundances of genes being expressed in different tissues and development stages.

Table 1.3: Rosaceae ESTs available at NCBI by Species as of 08/21/2006

Genus	Number of ESTs
Malus	259088
x domestica	253660
x domestica X sieversii	3944
sieboldii	1163
hybrid rootstock	321
Prunus	86583
persica	66249
armeniaca	15105
dulcis	3864
cerasus	1255
cerasus X avium X canescens	89
avium	21
Fragaria	19038
x ananassa	5376
vesca	13662
Rosa	9289
hybrid cultivar	5563
wichurana	1932
chinensis	1794
Pyrus	335
communis	238
communis X ussuriensis	82
pyrifolia	15
Rubus ideaus	322
TOTAL	374655

Table 1.4: Rosaceae ESTs available at NCBI by Library Tissue as of 08/21/2006

Tissue	Number of ESTs
Carpel	47
Flower	23161
Fruit Mesocarp	33206
Fruit	78562
Fruit Endocarp	5072
Fruit Epicarp	7749
Fruit Epicarp & Mesocarp	12029
Fruit Mesocarp	34457
Gynoecium	1006
Inflorescence Meristem	8743
Leaf	56457
Petal	5305
Phloem	9376
Receptacle	23
Receptacle & Achenes	35
Root	11251
Seed	8258
Shoot	18696
Unspecified	4978
Vegetative Meristem	33833
Whole Plant	17350
Xylem	5061

Microarray technology provides high throughput detection of gene expression levels (Richmond and Somerville, 2000). Schena et al., 1995 first developed the cDNA microarray technology which has since been widely used (Duggan et al., 1999). Specifically, these microchips have been used in plants to identify particular gene functions (Aharoni et al., 2000; Gutierrez et al., 2002), evaluate transcriptional response to physiological and environmental conditions

(Reymond et al., 2000; Van Hal et al., 2000; Lee et al., 2002; Oztur et al., 2002; Potokina et al., 2002; Zhu et al., 2003), and evaluate transcript profiles between genetically modified and control species (Val Hal et al., 2000).

Currently in Rosaceae, only apple has a publicly available microarray chip. The Plant Genome Program Award #0420394 included the development of a NimbleGen oligonucleotide array with 390,000 spots representing 55,000 sequences developed from publicly available apple EST data (McNellis et al, 2007). Data from this chip has yet to be published. Future development of either family-wide or individual species chips could promote functional genomic studies in the Rosaceae by allowing researchers a relatively economical and empirically proven means of identifying differentially expressed genes.

Transgenics, which includes the introduction of new genes and knocking out expression of genes, is a powerful method for elucidating gene function. Apple transformation has been achieved with an *Agrobacterium*-based approach (Defilippi et al, 2004; Szankowski et al, 2003; Markwick et al, 2003). Transgenic apple lines are currently available with resistance to apple scab and fire blight (Bolar et al, 2001) and suppressed ethylene and volatile esters in the fruit (Defilippi et al, 2004). A number of research groups have reported success with transforming strawberry using *Agrobacterium* (Folta et al, 2006; Oosumi et al, 2006; Lunkenbein et al, 2006). To date, there have been no reproducible studies reported for peach transformation, although several groups are currently working on this problem.

Functional genomics can start with the trait or with a candidate sequence, but the ultimate goal is to find the specific sequence, discern its expression patterns and understand the metabolic roles of the resulting protein. Generation and mapping of ESTs, especially from under-represented species, will continue to be an integral part of the functional genomics area. The EST sets contain redundancies that need to be filtered into a more useable unigene set in which each gene is, theoretically, represented only once. The unigenes need to be mapped onto the available physical and genetic maps. These new ESTs will help researchers to discover and utilize allelic diversity, find single nucleotide polymorphisms (SNPs), and develop microarray technology. Adding QTLs to the various genetic maps will also be important. A greater understanding of functional genomics in crop species will ultimately lead to better varieties.

Twenty eight traits of economic importance have been mapped to the general *Prunus* map and tightly linked markers have been identified. These represent excellent candidates for marker assisted selection techniques. A table from Dirlewanger et al., 2004b is reprinted in Table 1.5. These traits are highly representative of many of the qualities important to breeders, growers and consumers.

Table 1.5: Description of 28 major traits controlling morphological or agronomic characters in different *Prunus* crops that can be located on the reference map (From Dirlewanger et al., 2004b)

Characters	Species	Symbol
Fruit flesh color (white/yellow)	Peach	Y
Sharka resistance	Apricot	Sharka
Evergrowing	Peach	Evg
Flower color	Almond x peach	B
Root-knot nematode resistance	Peach	Mi
Shell hardness	Almond	D
Broomy (or pillar) growth habitat	Peach	Br
Double flower	Peach	Dl
Flesh color around the stone	Peach	Cs
Anther color (yellow/anthocyanic)	Almond x peach	Ag
Polycarpel	Peach	Pcp
Flower color	Peach	Fc
Blooming time	Almond	Lb
Flesh adhesion (clingstone/freestone)	Peach	F
Non-acid fruit	Peach	D
Kernel taste (bitter/sweet)	Almond	Sk
Skin hariness (nectarine/peach)	Peach	G
Leaf shape (narrow/wide)	Peach	Nl
Plant height (normal/dwarf)	Peach	Dw
Male sterility	Peach	Ps
Fruit shape (flat/round)	Peach	S*
Self-incompatibility	Almond	S
Self-incompatibility	Apricot	S
Fruit skin color	Peach	Sc
Leaf color (red/green)	Peach	Gr
Root-knot nematode resistance	Myrobalan plum	Ma
Resistance to powdery mildew	Peach	Sf
Leaf gland (reniform/globose)	Peach	E

Other Resources

Although investigations of Rosaceae genetics and genomics are relatively new compared with other major crop families, the Rosaceae community is well-organized with elected steering committees at both the national and international level. The current U.S White paper is designed to integrate research in Rosaceae family members and advocate more funding for these important studies. A unified community-based approach will hopefully accelerate the discovery process and reduce redundancy of effort.

With expanding sequence resources in models such as apple and peach, the application of comparative genomics can be increasingly used to transfer information between species in Rosaceae, especially those with fewer sequence resources such as cherry, pear, raspberry and rose. By understanding the level of similarity and difference between the various genera and species, researchers can assess the usefulness of applying tools from one species to others. This can help eliminate duplication of research effort for each crop of interest, reducing time, and costs. Currently, maps from *Prunus* show a high degree of co-linearity among component species (Dirlewanger et al., 2004b). A preliminary analysis of the apple and *Prunus* genome demonstrated a high level of synteny between these two genomes. A recent study indicates marker transferability of primers flanking coding regions from *Fragaria* to *Prunus* and *Malus* (Sargent et al., 2007).

The Rosaceae community has a central data repository and online website for information exchange and community news, the Genome Database for Rosaceae (GDR, Jung et al., 2004). Funded by the NSF Plant Genome Program Award #0320544, GDR was initiated to integrate the available structural and

functional genomics data for peach. The GDR has since expanded its aims to incorporate all publicly available genomics data for the family while also providing online analysis tools and services (Jung et al, 2004). Future modules that are being developed will include information on genes, alleles, traits, segregation data, and germplasm resources. Long term maintenance and enhancement of this resource will be required for efficient data dissemination and analysis in the community.

Expressed Sequence Tags as Research Tools

Expressed sequence tags (ESTs) are partial sequences of expressed genes randomly picked from a cDNA library. Usually a single-pass read of approximately 200 to 600 base pairs is produced from the 3' and/or the 5' end of the cDNA clone. Pioneered in 1991 by Adams et al. and utilized extensively in the human genome project (Davies, 1993), ESTs have since become an essential tool for gene discovery and mapping in many different organisms; they reveal not only which genes are being expressed in a tissue but also relative levels of expression. Tissues from different developmental stages or produced from different conditions may be compared to determine differential gene expression.

The production of ESTs (Baxevanis and Ouellette, 2001) begins by isolating RNA from the tissue of interest and selecting the mRNA with an oligo(dT) primer that recognizes the polyA tails. These mRNAs are reverse transcribed into cDNA and directionally cloned into vectors to make a cDNA library. Individual clones are then picked and sequenced, providing 200 to 600 or more high quality bases. The resulting sequences may contain untranslated regions or other undesired sequence artifacts. Contamination, including vector,

mitochondrial, and bacterial sequences, must be trimmed or removed from the set. Publicly available ESTs are submitted to one of three international sequence databases: GenBank, EMBL, and DDBJ, each of which is updated on a nightly basis to ensure uniformity of sequences across all three repositories.

Researchers can choose to sequence from either the 3' or the 5' end of the sequence. The 3' generally yields part of the polyA tail in the sequence, allowing the end of the sequence to be easily identified. Identifiying the end may enable the triplet codon frame to be established and make functional identification somewhat easier. However extra bases or other sequencing errors may still shift or interfere with the codon frame. Sequencing from the 5' end maximizes the coding region obtained by not sequencing the polyA tail and following untranslated region, but the ends of the sequences will vary. Sequencing from both ends is generally the most useful technique but many researchers choose not to incur the extra expense and time investment. If only one end is sequenced, the 5' end is generally preferred for its maximization of coding base pairs.

There are some drawbacks to using the technique of EST sequencing. ESTs are generally produced with a single read of the sequencer, leading to lower quality sequence. The mRNA is also very unstable prior to being reverse transcribed into cDNA, and it often undergoes substitutions, insertions, and deletions. It is not uncommon to find chimeras in an EST library where the 5' end and 3' end are actually different genes joined together. The sequences must be trimmed of low quality bases and screened for obvious problems before a sequence analysis can be performed.

EST libraries are less likely to contain genes that are very rarely transcribed. To try to maximize the likelihood of sequencing these genes, the technique of subtractive hybrization, or normalization, is used (Bonaldo et al, 1996; Soares et al, 1994). A pool of RNA is removed from the library of interest by letting it hybridize to other RNAs. These other RNAs can be obtained from a seperate sample, thereby reducing commonly expressed genes. They may be taken from a different tissue, leaving only the sequences unique to the original library sample.

An EST is usually compared at the sequence level to known proteins in public databanks to identify potential homologs and infer possible function. The Basic Local Alignment Search Tool (BLAST) and the Fast-All (FASTA) are the two main sequence programs used for similarity searching (Altschul et al., 1990; Pearson and Lipman, 1988). Significant sequence similarity of an unknown sequence to a characterized protein sequence allows researchers to identify genes of particular interest and find candidate genes (Hatey et al., 1998). Disease resistance in particular is an example of a trait of interest across multiple species; ESTs have helped researchers to find these genes in many species such as *Arabidopsis thaliana* (Meyers et al. 2002), potato (Ronning et al., 2003), rice (Jantasuriyarat et al., 2005), soybean (Tian et al., 2004) and many others. Known resistance genes can now be used as databases in the BLAST or FASTA searches to reveal other potential resistance genes in uncharacterized ESTs.

By randomly choosing the cDNA inserts to sequence, the number of copies of each gene sequenced can be compared and used to infer relative

expression levels for those genes. By comparing ESTs produced from different tissues, different species, or different environmental conditions/stresses important gene expression data can be gathered. The availability of 152,635 ESTs for tomato made analysis of library expression levels possible and resulted in the identification of transcription factors associated specifically with ripening. However, this analysis can be error-prone, and requires careful statistical consideration as it relies heavily on the availability of the sequence data (Wang et al., 2004). ESTs from normalized libraries should not be used in this type of frequency analysis.

ESTs are also utilized in gene family and gene evolution studies. Constructing long sequences from overlapping ESTs of sufficient quality can allow homologs to be identified. The sequences may elucidate paralogs, genes separated by a gene duplication event, from orthologs, genes separated by a speciation event. Different copies of genes present in highly similar gene families can be compared across taxa and utilized in evolutionary studies (Cooke et al., 1997; Epple et al., 1997). Researchers are also beginning to use ESTs from divergent plant species to compare rates of evolution for different genes (Van der Hoeven et al., 2002). By noting the number of substitutions in the nucleotide sequence, it may be possible to infer evolutionary pressure on each gene in question. Substitution rates and gene family information can also be used to build phylogenies and ascertain the evolutionary relationship of plant species.

Genetic mapping is an important tool for plant genomics that requires the generation of molecular markers. ESTs are a rich resource for simple sequence

repeats (SSRs), also know as microsatellites. These short repetitve sequences are useful for comparative mapping because of their high polymorphism and transportability. Candidate SSRs in ESTs are easily located using computer algorthims and can then be screened for polymorphism against a DNA panel (Cardle et al., 2000; Jung et al., 2005; La Rota et al., 2005). Single nucleotide polymorphisms (SNPs) can also be mined from ESTs (Garg et al., 1999). These markers are highly abundant but require multiple high-copy reads of the same mRNA to identify. Molecular markers derived from ESTs have the advantage of being located in a coding region and thus being directly correlated to a specific locus in the genome.

ESTs can be useful in genetic mapping in other ways as well. They can be used to help order the BAC clones on a physical map and then anchor the physical map onto the genetic map. This approach has been used for linkage maps in rice and maize (Harushima et al., 1998; Davis et al., 1999) as well as a physical map of rice (Kurata et al., 1997).

Despite the abundance of gene prediction algorithms available, identifying coding regions in genomic sequence is still considered an imperfect science at the present time. An increasingly important use of ESTs involves aligning ESTs with genomic sequence to help validate predicted open reading frames. ESTs can complement predictive algorithms by revealing alternative splicing and transcription start/stop sites. *Arabidopsis* chromosome 2 (Lin et al, 1999) and 4 (Mayer et al, 1999) were annotated with ESTs by estimating how often genes along each chromosome were being expressed in different tissues.

EST Unigenes

Unigenes strive to define a single sequence for each genomic locus that results in an mRNA transcript. A common method of creating a putative unigene for an organism is clustering/assembling ESTs that come from the same transcript. This makes EST resources more useful by reducing their inherent redundancy and through aligning sequences to find longer consensus sequences increases the probability of finding homolog matches. As more ESTs are sequenced and added to the public domain, the unigene can be refined and become more accurate.

Indexing EST data in this manner has become a major effort for many large online databases including NCBI's UniGene (Pontius et al., 2003; Wheeler et al., 2003), the TIGR Gene Index (Quackenbush et al., 2000), the Sequence Tag Alignment and Consensus Knowledgebase (STACK) (Christoffels et al., 2001), and PlantGDB (Dong et al., 2004). Each database uses different data sources and algorithms. Unigenes are routinely created for one or multiple cDNA libraries in many individual species, for example wheat (Lazo et al., 2004), barley (Michalek et al., 2004), and soybean (Tian et al., 2004).

The results from any clustering algorithm are limited by the sequencing and sampling error of the data. Genes of low copy number are often not found in EST libraries and will not be represented in the unigene. The quality of the data can also be an issue; very high error rates will make assembling transcripts much more difficult.

The accuracy of a unigene is also dependent on the bioinformatics methods used to perform the clustering. Two types of error can occur during

unigene production, commonly referred to as Type I and Type II errors (Burke et al., 1999; Wang et al., 2003). Type I error refers to ESTs from the same gene being falsely separated into two or more clusters or singletons. Type II error is the opposite, when two or more ESTs from different genes are placed in the same cluster, also referred to as a contig. These errors tend to be correlated; reducing one will inflate the other (Wang et al., 2003). Ideally, the assembly algorithm should be stringent enough to separate paralogs but also capable of tolerating sequencing errors.

In general, a unigene set can be expected to overestimate the number of genes in the EST libraries (Vodkin et al., 2004). Type I errors occur if two reads do not overlap at all or do not overlap enough to be identified as the same read. However, Type II errors are also problematic and are characterized by sequences identified as the same gene that are actually from different loci in the genome. Type II errors occur in most data sets due to the presence of gene families. Often, genes in the same family have regions of very similar sequence, which can lead to "over-assembly". Genomic sequencing in *Arabidopsis* indicated that 80% of proteins are encoded by families (AGI, 2000), making Type II error a potentially difficult problem for plant assemblies.

The CAP3 assembling program is preferred for the creation of a Rosaceae unigene because it is efficient, reliable and more stringent than other BLAST-based approaches. It was shown by Liang et al., 2000 to be superior to the TIGR assembler (Sutton et al., 1995) and Phrap (Ewing and Green, 1998) in its ability to distinguish gene family members. The main stringency parameter of concern for

the CAP3 program is the "p" value, the percentage identity in the overlap region (Wang et al., 2003). However, as the quality and quantity of EST data varies greatly for each species, it is impossible for the program to perform well with default settings every time it is used; often different levels of stringency need to be tested to find the optimal parameter settings for a particular data set.

Research Question

The Rosaceae family of plants is a biologically diverse group with high economical and nutritional value. Increasing the available genomic and genetic resources for this family will ultimately result in better varieties as well as increase our overall understanding of the biology and genetics of fruits and trees. A large set of ESTs is available for the family and includes multiple species, tissues, development stages, and conditions. These libraries have not been data-mined to extract the maximum amount of useful information across species and genera. Many of the libraries have been analyzed within the context of the species (e.g. Newcombe et al, 2006; Park et al, 2006; Horn et al, 2005) or for particular candidate genes (e.g. Lalli et al, 2005; Silva et al, 2005; Beuning et al, 2004), but a genus-wide or family-wide examination may yield more useful information. The question asked in the course of this research is how bioinformatics can be used to analyze EST datasets, yield maximum knowledge for each sequence, and further genomic research in the Rosaceae community via online resources.

The research elucidates what genes are being expressed in the species analyzed and how we can use these genes for eventual crop improvement. Effective dissemination of this data to the community is accomplished through

GDR. The EST data and corresponding unigenes can be used for developing better genetic maps through marker mining, and the sequences can be used to anchor these genes to physical maps. The unigenes can be used to analyze gene families, gene copy numbers, levels of sequence divergence, and evolutionary relationships within Rosaceae. Comparing cDNA libraries can yield interesting candidate genes involved in traits of interest to consumers, growers, and researchers that may be useful in multiple Rosaceae species. Questions can ultimately be answered about the genes being expressed in different tissues and stages that impact important agricultural qualities and how these genes are related across the family.

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, et al. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252(5013):1651-6.

Aharoni A, Keizer LCP, Bouwmeester HJ, Sun Z, Alvarez-Huerta M, Verhoeven HA, Blaas J, van Houwelingen AMML, De Vos RCH, van der Voet H, Jansen RC, Guis M, Mol J, Davis RW, Schena M, van Tunen AJ andO'Connell AP. 2000. Identification of the SAAT gene involved in strawberry flavor biogenesis by use of DNA microarrays. *Plant Cell* 12: 647–661.

Akiyama Y, Yamamoto Y, Ohmido N, Oshima M and Fukui K. 2001. Estimation of the nuclear DNA content of strawberries (*Fragaria spp.*) compared with *Arabidopsis thaliana* by using dual-stem flow cytometry. *Cytologia* 66:431-436.

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215(3):403-410.

The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815.

Aranzana MJ, Pineada A, Cosson P, Dirlewanger E, Ascasibar J, Cipriani G, Ryder CD, Testolin R, Abbott A, King GJ, Iezzoni AF and Arús P. 2003. A set of simple-sequence repeat (SSR) markers covering the *Prunus* genome. *Theor Appl Genet* 106(5):819–825.

Baird WV, Estager AS and Wells J. 1994. Estimating nuclear DNA content in peach and related diploid species using laser flow cytometry and DNA hybridization. *J Amer Soc Hort Sci* 119:1312-1316.

Baxevanis AD and Ouellette BF. (eds). 2005. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 3rd Edition. Hobokin, NJ: Wiley, John & Sons, Inc.

Bennett MD and Leitch IJ. 2005. Nuclear DNA amounts in angiosperms: progress, problems and prospects. Ann Bot (Lond). 95(1):45-90.

Beuning L, Bowen J, Persson H, Barraclough D, Bulley S and Macrae E. 2004. Characterisation of Mal d 1-related genes in *Malus*. Plant Mol Biol. 55(3):369-88.

Boguski MS, Lowe TM and Tolstoshev CM. 1993. dbEST - Database for "Expressed Sequence Tags". *Nat Genet.* Aug;4(4):332-3.

Bolar JP, Norelli JL, Harman GE, Brown SK and Aldwinckle HS. 2001. Synergistic activity of endochitinase and exochitinase from Trichoderma atroviride (T. harzianum) against the pathogenic fungus (Venturia inaequalis) in transgenic apple plants. Transgenic Res. 10(6):533-43.

Bonaldo MF, Lennon G and Soares MB. 1996. Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res. 6(9):791-806.

Burke J., Wang H., Hide W. and Davison,D. 1998. Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res*., 8, 276–290.

Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D and Waugh R. 2000. Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* 156(2):847-54.

Christoffels A, van Gelder A, Greyling G, Miller R, Hide T and Hide W. 2001. STACK: Sequence Tag Alignment and Consensus Knowledgebase. *Nucleic Acids Res*. 29(1):234-8.

Cooke R, Raynal M, Laudie M and 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):1127-40.

Dandekar A, Iezzoni A, Abbott A, Sosinski B, Peace C, Crisosto C., Finn C., Dardick C., Simon C, Potter D, Swietlik D, Byrne D, Okie D, Main D, Stover E, Ogundiwin E, Volk G, Fazio G, Aldwinckle H, Slovin J, Norelli J, Hancock J, McFerson J, Olmstead J, Goffreda J, Postman J, Folta K, Hummer K, Lewers K, Pritts M, Forsline P, Scorza R, Bell R, Korban S, van Nocker S, Brown S, Davis T, Gradziel T, Sjulin T, Shulaev V and Loescher W. 2006. The U.S. Rosaceae Genomics, Genetics, and Breeding Initiative. Retrieved on Aug. 27, 2006 from http://www.bioinfo.wsu.edu/gdr/community/rosexec/RosWP_March_2006.doc

Davies, K. 1993. The EST express gathers speed. *Nature* 364(6437):554.

Davis GL, McMullen MD, Baysdorfer C, Musket T, Grant D, Staebell M, Xu G, Polacco M, Koster L, Melia-Hancock S, Houchins K, Chao S and Coe EH (1999) A maize map standard with sequenced core markers, grass genome reference points and 932 expressed sequence tagged site ESTs in a 1,736-locus map. *Genetics* 152:1137–1172

Davis T. 2006. A Diploid Platform for Strawberry Genetics [abstract]. In: Plant and Animal Genomes XIII Conference; Jan 15-19 2006; San Diego, CA. Available from: http://www.intl-pag.org/pag/13/abstracts/PAG13_W125.html.

Decroocq V, Foulongne M, Lambert P, Gall OL, Mantin C, Pascal T, Schurdi-Levraud V, and Kervella J. 2005. Analogues of virus resistance genes map to QTLs for resistance to sharka disease in *Prunus davidiana*. *Mol Genet Genomics* 272(6):680-9. Epub 2005 Jan 22.

Defilippi BG, Dandekar AM, Kader AA. 2004. Impact of suppression of ethylene action or biosynthesis on flavor metabolites in apple (*Malus domestica* Borkh) fruits. J Agric Food Chem. 52(18):5694-701.

Dirlewanger E, Cosson P, Howad W, Capdeville G, Bosselut N, Claverie M, Voisin R, Poizat C, Lafargue B, Baron O, Laigret F, Kleinhentz M, Arus P and Esmenjaud D. 2004a. Microsatellite genetic linkage maps of myrobalan plum and an almond-peach hybrid--location of root-knot nematode resistance genes. *Theor Appl Genet.* 109(4):827-38. Epub 2004 Jul 6.

Dirlewanger E, Graziano E, Joobeur T, Garriga-Caldere F, Cosson P, Howad W and Arus P. 2004b. Comparative mapping and marker-assisted selection in Rosaceae fruit crops. *Proc Natl Acad Sci U S A* 101(26):9891-6. Epub 2004 May 24.

Dong Q, Schlueter SD and Brendel V. (2004) PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res*. 32(Database issue):D354-D359.

Duggan DJ, Bittner M, Chen Y, Meltzer P and Trent JM. 1999. Expression profiling using cDNA microarrays. Nat Genet. 21(1 Suppl):10-4. Review.

Epple P, Apel K and Bohlmann H. 1997. ESTs reveal a multigene family for plant defensins in *Arabidopsis thaliana*. *FEBS Lett*. 400(2):168-72.

Evans RC, Alice LA, Campbell CS, Kellogg EA, Dickinson TA. 2000. The granule-bound starch synthase (GBSSI) gene in the Rosaceae: multiple loci and phylogenetic utility. Mol. Phylogenet. Evol. 17: 388-400.

Ewing B and Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8:186–94.

Fei Z, Tang X, Alba RM, White JA, Ronning CM, Martin GB, Tanksley SD and Giovannoni JJ. 2004. Comprehensive EST analysis of tomato and comparative genomics of fruit ripening. *Plant J*. 40(1):47-59.

Folta KM, Dhingra A, Howard L, Stewart PJ and Chandler CK. 2006. Characterization of LF9, an octoploid strawberry genotype selected for rapid regeneration and transformation. Planta. 2006 Apr 14; [Epub ahead of print]

Garg K, Green P and Nickerson DA. 1999. Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. *Genome Res*. 9(11):1087-92.

Gutierrez RA, Ewing RM, Cherry JM and Green PJ. 2002. Identification of unstable transcripts in *Arabidopsis* by cDNA microarray analysis: rapid decay is associated with a group of touch and specific clock-controlled genes. *Proc. Natl. Acad. Sci. USA* 99(17): 11513–11518.

Han Y, Gasic K, Marron B, Beever JE and Korban SS. 2007. A BAC-based physical map of the apple geome. *Genomics*. Epub Jan 30.

Harushima Y, Yano M, Shomura A, Sato M, Shimano T, Kuboki Y, Yamamoto T, Lin S, Antonio BA, Parco A, Kajiya H, Huang N, Yamamoto K, Nagamura Y, Kurata N, Khush GS and Sasaki T. 1998. A high-density rice genetic linkage map with 2275 markers using a single F2 population. *Genetics* 148:479–494.

Hatey F, Tosser-Klopp G, Clouscard-Martinato C, Mulsant P and Gasser F. 1998. Expressed sequence tags for genes: a review. *Genet Sel Evol* 30:521–541

Horn R, Lecouls AC, Callahan A, Dandekar A, Garay L, McCord P, Howad W, Chan H, Verde I, Main D, Jung S, Georgi L, Forrest S, Mook J, Zhebentyayeva T, Yu Y, Kim HR, Jesudurai C, Sosinski B, Arus P, Baird V, Parfitt D, Reighard G, Scorza R, Tomkins J, Wing R and Abbott AG. 2005. Candidate gene database and transcript map for peach, a model species for fruit trees. *Theor Appl Genet*. 2005 Apr 22; [Epub ahead of print]

Janick J and Moore JN, eds. 1996a. *Fruit Breeding: Tree and Tropical Fruits*. Vol I. New York: John Wiley & Sons.

Janick J and Moore JN, eds. 1996b. *Fruit Breeding: Vine and Small Fruits*. Vol II. New York: John Wiley & Sons.

Jantasuriyarat C, Gowda M, Haller K, Hatfield J, Lu G, Stahlberg E, Zhou B, Li H, Kim H, Yu Y, Dean RA, Wing RA, Soderlund C and Wang GL. 2005. Large-scale identification of expressed sequence tags involved in rice and rice blast fungus interaction. *Plant Physiol.* 138(1):105-15.

Jelenkovic G and Harrington E (1972) Morphology of the pachytene chromosomes in *Prunus persica*. Can J Genet Cytol 14: 317-24.

Joobeur T, Viruel MA, de Vicente MC, Jauregui B, Ballester J, Dettori MT, Verde I, Truco MJ, Messeguer R, Balle I, Quarta R, Dirlewanger E and Arús P. 1998. Construction of a saturated linkage map for *Prunus* using an almond × peach F2 progeny. *Theor Appl Genet* 97(7):1034–1041.

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(1):130.

Jung S, Abbott A, Jesudurai C, Tomkins J and Main D. 2005. Frequency, type, distribution and annotation of simple sequence repeats in Rosaceae ESTs. *Funct Integr Genomics* 5(3):136-43. Epub 2005 Mar 11.

Kurata N, Umehara Y, Tanoue H and Sasaki T. 1997. Physical mapping of the rice genome with YAC clones. *Plant Mol Biol.* 35:101–113.

La Rota M, Kantety RV, Yu JK and Sorrells ME. 2005. Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *BMC Genomics* 6(1):23.

Lalli DA, Decroocq V, Blenda AV, Schurdi-Levraud V, Garay L, Le Gall O, Damsteegt V, Reighard GL, Abbott AG. 2005. Identification and mapping of resistance gene analogs (RGAs) in *Prunus*: a resistance map for *Prunus*. Theor Appl Genet. 111(8):1504-13. Epub 2005 Nov 10.

Lazo GR, Chao S, Hummel DD, Edwards H, Crossman CC, Lui N, Matthews DE, Carollo VL, Hane DL, You FM, Butler GE, Miller RE, Close TJ, Peng JH, Lapitan NL, Gustafson JP, Qi LL, Echalier B, Gill BS, Dilbirligi M, Randhawa HS, Gill KS, Greene RA, Sorrells ME, Akhunov ED, Dvorak J, Linkiewicz AM, Dubcovsky J, Hossain KG, Kalavacharla V, Kianian SF, Mahmoud AA, Miftahudin, Ma XF, Conley EJ, Anderson JA, Pathan MS, Nguyen HT, McGuire PE, Qualset CO and Anderson OD. 2004. Development of an expressed sequence tag (EST) resource for wheat (Triticum aestivum L.): EST generation, unigene analysis, probe selection and bioinformatics for a 16,000-locus bin-delineated map. *Genetics*. 168(2):585-93.

Lee JM, Williams ME, Tingey SV and Rafalski JA. 2002. DNA array profiling of gene expression changes during maize embryo development. *Funct. Integr. Genomics* 2(1–2): 13–27.

Lespinasse Y, Alston FH and Watkins R. 1976. Cytological techniques for use in apple breeding. *Ann Appl Biol*. 82:349-353.

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL and Quackenbush J. 2000. An optimized protocol for analysis of EST sequences. *Nucleic Acids Res*. 28(18):3657-65.

Liebhard R, Kellerhals M, Pfammatter W, Jertmini M and Gessler C. 2003. Mapping quantitative physiological traits in apple (*Malus x domestica* Borkh.). *Plant Mol Biol.* 52(3):511-26.

Lin X, Kaul S, Rounsley S, Shea TP, Benito MI, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M, Feldblyum TV, Buell CR, Ketchum KA, Lee J, Ronning CM, Koo HL, Moffat KS, Cronin LA, Shen M, Pai G, Van Aken S, Umayam L, Tallon LJ, Gill JE, Adams MD, Carrera AJ, Creasy TH, Goodman HM, Somerville CR, Copenhaver GP, Preuss D, Nierman WC, White O, Eisen JA, Salzberg SL, Fraser CM, Venter JC. 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature. 402(6763):761-8.

Lunkenbein S, Salentijn EM, Coiner HA, Boone MJ, Krens FA and Schwab W. 2006. Up- and down-regulation of *Fragaria x ananassa* O-methyltransferase: impacts on furanone and phenylpropanoid metabolism. J Exp Bot. 57(10):2445-53. Epub 2006 Jun 23.

Markwick NP, Docherty LC, Phung MM, Lester MT, Murray C, Yao JL, Mitra DS, Cohen D, Beuning LL, Kutty-Amma S, Christeller JT. 2003. Transgenic tobacco and apple plants expressing biotin-binding proteins are resistant to two cosmopolitan insect pests, potato tuber moth and lightbrown apple moth, respectively. Transgenic Res. 12(6):671-81.

Mayer K, Schuller C, Wambutt R, Murphy G, Volckaert G, Pohl T, Dusterhoft A, Stiekema W, Entian KD, Terryn N, Harris B, Ansorge W, Brandt P, Grivell L, Rieger M, Weichselgartner M, de Simone V, Obermaier B, Mache R, Muller M, Kreis M, Delseny M, Puigdomenech P, Watson M, Schmidtheini T, Reichert B, Portatelle D, Perez-Alonso M, Boutry M, Bancroft I, Vos P, Hoheisel J, Zimmermann W, Wedler H, Ridley P, Langham SA, McCullagh B, Bilham L, Robben J, Van der Schueren J, Grymonprez B, Chuang YJ, Vandenbussche F, Braeken M, Weltjens I, Voet M, Bastiaens I, Aert R, Defoor E, Weitzenegger T, Bothe G, Ramsperger U, Hilbert H, Braun M, Holzer E, Brandt A, Peters S, van Staveren M, Dirske W, Mooijman P, Klein Lankhorst R, Rose M, Hauf J, Kotter P, Berneiser S, Hempel S, Feldpausch M, Lamberth S, Van den Daele H, De Keyser A, Buysshaert C, Gielen J, Villarroel R, De Clercq R, Van Montagu M, Rogers J, Cronin A, Quail M, Bray-Allen S, Clark L, Doggett J, Hall S, Kay M, Lennard N, McLay K, Mayes R, Pettett A, Rajandream MA, Lyne M, Benes V, Rechmann S, Borkova D, Blocker H, Scharfe M, Grimm M, Lohnert TH, Dose S, de Haan M, Maarse A, Schafer M, Muller-Auer S, Gabel C, Fuchs M, Fartmann B, Granderath K, Dauner D, Herzl A, Neumann S, Argiriou A, Vitale D, Liguori R, Piravandi E, Massenet O, Quigley F, Clabauld G, Mundlein A, Felber R, Schnabl S, Hiller R, Schmidt W, Lecharny A, Aubourg S, Chefdor F, Cooke R, Berger C, Montfort A, Casacuberta E, Gibbons T, Weber N, Vandenbol M, Bargues M, Terol J, Torres A, Perez-Perez A, Purnelle B, Bent E, Johnson S, Tacon D, Jesse T, Heijnen L, Schwarz S, Scholler P, Heber S, Francs P, Bielke C, Frishman D, Haase D, Lemcke K, Mewes HW, Stocker S, Zaccaria P, Bevan M, Wilson RK, de la Bastide M, Habermann K, Parnell L, Dedhia N, Gnoj L, Schutz K, Huang E, Spiegel L, Sehkon M, Murray J, Sheet P, Cordes M, Abu-Threideh J, Stoneking T, Kalicki J, Graves T, Harmon G, Edwards J, Latreille P, Courtney L, Cloud J, Abbott A, Scott K, Johnson D, Minx P, Bentley D, Fulton B, Miller N, Greco T, Kemp K, Kramer J, Fulton L, Mardis E, Dante M, Pepin K, Hillier L, Nelson J, Spieth J, Ryan E, Andrews S, Geisel C, Layman D, Du H, Ali J, Berghoff A, Jones K, Drone K, Cotton M, Joshu C, Antonoiu B, Zidanic M, Strong C, Sun H, Lamar B, Yordan C, Ma P, Zhong J, Preston R, Vil D, Shekher M, Matero A, Shah R, Swaby IK, O'Shaughnessy A, Rodriguez M, Hoffmann J, Till S, Granat S, Shohdy N, Hasegawa A, Hameed A, Lodhi M, Johnson A, Chen E, Marra M, Martienssen R, McCombie WR. 1999. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. Nature. 402(6763):769-77.

McNellis T, Jensen PJ, Crassweller R, Maximova S, Travis JW, Altman N, Makalowksa I, Praul C and Fazio G. 2007. Development of Gene Expression Phenotypic Markers for Apple [abstract]. In: Plant and Animal Genomes XV Conference; Jan 13-17 2007; San Diego, CA. Available from: http://www.intl-pag.org/pag/15/abstracts/PAG15_W21_152.html.

Meyers BC, Morgante M and Michelmore RW. 2002. TIR-X and TIR-NBS proteins: two new families related to disease resistance TIR-NBS-LRR proteins encoded in *Arabidopsis* and other plant genomes. *Plant J*. 32(1):77-92.

Michalek W, Weschke W, Pleissner KP, Graner A. 2002. EST analysis in barley defines a unigene set comprising 4,000 genes. *Theor Appl Genet*. 104(1):97-103.

Morgan DR, Soltis DE, Robertson KR. 1994. Systematic and evolutionary implications of rbcL sequence variation in Rosaceae. Amer. J. Bot. 81: 890-903.

National Agricultural Statistics Service. 2007. Noncitrus Fruits and Nuts, 2006 Preliminary Summary. Downloaded on March 23, 2007 from http://usda.mannlib.cornell.edu/MannUsda/viewDocumentInfo.do?documentID=1113

Newcomb RD, Crowhurst RN, Gleave AP, Rikkerink EH, Allan AC, Beuning LL, Bowen JH, Gera E, Jamieson KR, Janssen BJ, Laing WA, McArtney S, Nain B, Ross GS, Snowden KC, Souleyre EJ, Walton EF, Yauk YK. 2006. Analyses of expressed sequence tags from apple. Plant Physiol. 141(1):147-66. Epub 2006 Mar 10.

Oosumi T, Gruszewski HA, Blischak LA, Baxter AJ, Wadl PA, Shuman JL, Veilleux RE, Shulaev V. 2006. High-efficiency transformation of the diploid strawberry (*Fragaria vesca*) for functional genomics. Planta. 223(6):1219-30. Epub 2005 Dec 1.

Oztur ZN, Talame V, Deyholos M, Michalowski CB, Galbraith DW, Gozukirmizi N, Tuberosa R and Bohnert HJ. 2002. Monitoring large-scale changes in transcript abundance in drought- and salt-stressed barley. *Plant Mol. Biol*. 48(5–6): 551–573.

Park S, Sugimoto N, Larson MD, Beaudry R, van Nocker S. 2006. Identification of genes with potential roles in apple fruit development and biochemistry through large-scale statistical analysis of expressed sequence tags. Plant Physiol. 141(3):811-24.

Patocchi A, Gianfranceschi L and Gessler C. 1999. Towards the map-based clone of *Vf:* fine and physical mapping of the *Vf* region. *Theor. Appl. Genet.* 99:1012-1017.

Pearson WR and Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 85(8):2444-2448.

Pontius JU, Wagner L and Schuler GD. 2003. UniGene: a unified view of the transcriptome. In: The NCBI Handbook. Bethesda (MD): National Center for Biotechnology Information.

Potokina E, Sreenivasulu N, Altschmied L, Michalek W and Graner A. 2002. Differential gene expression during seed germination in barley (*Hordeum vulgare* L.). *Funct. Integr. Genomics* 2(1–2): 28–39.

Potter D, Gao F, Bortiri PE, Oh S, Baggett S. 2002. Phylogenetic relationships in Rosaceae inferred from chloroplast matK and trnL-trnF nucleotide sequence data. Pl. Syst. Evol. 231: 77-89.

Quackenbush, J., Liang, F., Holt, I., Pertea, G. and Upton, J. 2000. The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Research* 28:141-145.

Reymond P, Weber H, Damond M and Farmer EE. 2000. Differential gene expression in response to mechanical wounding and insect feeding in *Arabidopsis*. *Plant Cell* 12: 707–719.

Richmond T and Somerville S. 2000. Chasing the dream: plant EST microarrays. *Curr. Opin. Plant Biol*. 3: 108–116.

Ronning CM, Stegalkina SS, Ascenzi RA, Bougri O, Hart AL, Utterbach TR, Vanaken SE, Riedmuller SB, White JA, Cho J, Pertea GM, Lee Y, Karamycheva S, Sultana R, Tsai J, Quackenbush J, Griffiths HM, Restrepo S, Smart CD, Fry WE, Van Der Hoeven R, Tanksley S, Zhang P, Jin H, Yamamoto ML, Baker BJ and Buell CR. 2003. Comparative analyses of potato expressed sequence tag libraries. *Plant Physiol.* 131(2):419-29.

Sargent DJ, Davis TM, Tobutt KR, Wilkinson MJ, Battey NH and Simpson DW. A genetic linkage map of microsatellite, gene-specific and morphological markers in diploid *Fragaria*. *Theor Appl Genet*. 2004 Nov;109(7):1385-91. Epub 2004 Jul 29.

Sargent DJ, Clarke J, Simpson DW, Tobutt KR, Arus P, Monfort A, Vilanova S, Denoyes-Rothan B, Rousseau M, Folta KM, Bassil NV, Battey NH. 2006. An enhanced microsatellite map of diploid *Fragaria*. Theor Appl Genet. 112(7):1349-59. Epub 2006 Feb 28.

Sargent DJ, Rys A, Nier S, Simpson DW and Tobutt KR. 2007. The development and mapping of functional markers in *Fragaria* and their transferability and potential for mapping in other genera. *Theor Appl Genet*. 114:373-384.

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270: 467–470.

Schulze-Menz G. K. 1964. Rosaceae. Pp. 209-218 in: Melchior, H., ed., Engler's Syllabus der Pflanzenfamilien II (12th ed.) Gebrüder Borntraeger, Berlin.

Silva C, Garcia-Mas J, Sanchez AM, Arus P, Oliveira MM. 2005. Looking into flowering time in almond (*Prunus dulcis* (Mill) D. A. Webb): the candidate gene approach. Theor Appl Genet. 110(5):959-68. Epub 2005 Feb 8.

Soares MB, Bonaldo MF, Jelene P, Su L, Lawton L, Efstratiadis A. 1994. Construction and characterization of a normalized cDNA library. Proc Natl Acad Sci U S A. 91(20):9228-32.

Sutton GG, White O, Adams MD and Kerlavage AR. 1995. TIGR Assembler: a new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*. 1:9-19.

Szankowski I, Briviba K, Fleschhut J, Schonherr J, Jacobsen HJ, Kiesecker H. 2003. Transformation of apple (*Malus domestica* Borkh.) with the stilbene synthase gene from grapevine (*Vitis vinifera* L.) and a PGIP gene from kiwi (*Actinidia deliciosa*). Plant Cell Rep. 22(2):141-9. Epub 2003 Jul 9.

Tian AG, Wang J, Cui P, Han YJ, Xu H, Cong LJ, Huang XG, Wang XL, Jiao YZ, Wang BJ, Wang YJ, Zhang JS and Chen SY. 2004. Characterization of soybean genomic features by analysis of its expressed sequence tags. *Theor Appl Genet*. 108(5):903-13. Epub 2003 Nov 18.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Dejardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjarvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leple JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouze P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K,Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y and Rokhsar D. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*. 313(5793):1596-604.

The U.S. Rosaceae Genomics, Genetics, and Breeding Initiative. (March 2006). Retreived August 23, 2006 from http://www.mainlab.clemson.edu/gdr/community/funding/

Van der Hoeven R, Ronning C, Giovannoni J, Martin G and Tanksley S. 2002. Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *Plant Cell* 14(7):1441-56.

Van Hal NLW, Vorst O, van Houwelingen AMML, Kok EJ, Peijnenburg A, Aharoni A, van Tunen AJ and Keijer J. 2000. The application of microarrays in gene expression analysis. *J. Biotech.* 78: 271–280.

Vodkin LO, Khanna A, Shealy R, Clough SJ, Gonzalez DO, Philip R, Zabala G, Thibaud-Nissen F, Sidarous M, Stromvik MV, Shoop E, Schmidt C, Retzel E, Erpelding J, Shoemaker RC, Rodriguez-Huete AM, Polacco JC, Coryell V, Keim P, Gong G, Liu L, Pardinas J, Schweitzer P. 2004. Microarrays for global expression constructed with a low redundancy set of 27,500 sequenced cDNAs representing an array of developmental stages and physiological conditions of the soybean plant. BMC Genomics. 2004 5(1):73.

Wang JP, Lindsay BG, Leebens-Mack J, Cui L, Wall K, Miller WC and dePamphilis CW. 2004. EST clustering error evaluation and correction. *Bioinformatics* 20(17):2973-84. Epub 2004 Jun 9.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JU, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L and Yaschenko E. 2005. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2005 Jan 1;33(Database issue):D39-45.

Wikstrom N, Savolainen V and Chase MW. 2001. Evolution of the angiosperms: calibrating the family tree. *Proc Biol Sci.* 268(1482):2211-20.

Wunsch A and Hormaza JI. 2004. Cloning and characterization of genomic DNA sequences of four self-incompatibility alleles in sweet cherry (*Prunus avium* L.). *Theor Appl Genet.* 108(2):299-305. Epub 2003 Sep 4.

Xu M, Korban SS, Song J and Jiang J. 2002. Constructing a bacterial artificial chromosome library of the apple cultivar GoldRush. *Acta Horticult* 595:103-112.

Xu M, Song J, Cheng Z, Jiang J and Korban SS. 2001. A bacterial artificial chromosome (BAC) library of *Malus floribunda* 821 and contig construction for positional cloning of the apple scab resistance gene *Vf*. *Genome* 44(2001): 1104-1113.

Zhao Y, Liu Q and Davis RE. 2004. Transgene expression in strawberries driven by a heterologous phloem-specific promoter. *Plant Cell Rep*. 23(4):224-30. Epub 2004 Jul 2.

Zhu T, Budworth P, Chen W, Provart N, Chang HS, Guimil S, Su W, Ester B, Zou GZ and Wang X. 2003. Transcriptional control of nutrient partitioning during rice grain filling. *Plant Biotechnol. J*. 1: 59–70.

Map References

Aranzana MJ, Pineada A, Cosson P, Dirlewanger E, Ascasibar J, Cipriani G, Ryder CD, Testolin R, Abbott A, King GJ, Iezzoni AF and Arús P. 2003. A set of simple-sequence repeat (SSR) markers covering the *Prunus* genome. *Theor Appl Genet* 106(5):819–825.

Bliss FA, Arulsekar S, Foolad MR, Becerra V, Gillen AM, Warburton ML, Dandekar AM, Kocsisne GM and Mydin KK. 2002. An expanded genetic linkage map of *Prunus* based on an interspecific cross between almond and peach. *Genome* 45:520–529

Chaparro JX, Werner DJ, OrsquoMalley D and Sederoff RR. 1994. Targeted-mapping and linkage analysis of morphological, isozyme, and RAPD markers in peach. *Theor Appl Genet* 87 (6-7):805–815.

Conner JP, Brown SK and Weeden NF (1997) Randomly amplified polymorphic DNA-based genetic linkage maps of three apple cultivars. *J Am Soc Hort Sci* 122:350–359.

Crespel L, Chirollet M, Durel E, Zhang D, Meynet J and Gudin S. 2002. Mapping of qualitative and quantitative phenotypic traits in *Rosa* using AFLP markers. *Theor Appl Genet*. 105(8):1207-1214. Epub 2002 Oct 11.

Dettori MT, Quarta R and Verde I. 2001. A peach linkage map integrating RFLPs, SSRs, RAPDs, and morphological markers. *Genome* 44:783–790.

Dirlewanger E and Bodo C. 1994. Molecular genetic mapping of peach. *Euphytica* 77:101–103.

Dirlewanger E, Cosson P, Howad W, Capdeville G, Bosselut N, Claverie M, Voisin R, Poizat C, Lafargue B, Baron O, Laigret F, Kleinhentz M, Arus P and Esmenjaud D. 2004a. Microsatellite genetic linkage maps of myrobalan plum and an almond-peach hybrid--location of root-knot nematode resistance genes. *Theor Appl Genet.* 109(4):827-38. Epub 2004 Jul 6.

Dirlewanger E, Moing A, Rothan C, Svanelle L, Pronier V, Guye A, Plomion C and Monet R. 1999. Mapping QTLs controlling fruit quality in peach [*Prunus persica* (L.) Batsch]. *Theor Appl Genet* 98(1):18–31.

Dirlewanger E, Pronier V, Parvey C, Rothan C, Guye A and Monet R. 1998. Genetic linkage map of peach [*Prunus persica* (L.) Batsch] using morphological and molecular markers. *Theor Appl Genet* 97(5-6):888–895.

Dugo ML, Satovic Z, Millan T, Cubero JI, Rubiales D, Cabrera A and Torres AM. 2005. Genetic mapping of QTLs controlling horticultural traits in diploid roses. *Theor Appl Genet*. [Epub ahead of print]

Foolad MR, Arulsekar S, Becerra V and Bliss FA (1995) A genetic linkage map of *Prunus* based on an interspecific cross between peach and almond. *Theor Appl Genet* 91(2):262–269.

Graham J, Smith K, MacKenzie K, Jorgenson L, Hackett C and Powell W. 2004. The construction of a genetic linkage map of red raspberry (*Rubus idaeus* subsp. idaeus) based on AFLPs, genomic-SSR and EST-SSR markers. *Theor Appl Genet*. 109(4):740-9. Epub 2004 May 4.

Hemmat M, Weeden NF, Manganaris AG and Lawson DM. Molecular marker linkage map for apple. *J Hered*. 1994 Jan-Feb;85(1):4-11.

Howad W, Yamamoto T, Dirlewanger E, Testolin R, Cosson P, Cipriani G, Monforte AJ, Georgi L, Abbott AG and Arus P. (2005). Genetics 171(3):1305-9. Epub 2005 Aug 22.

Hurtado MA, Romero C, Vilanova S, Abbott AG, Llacer G and Badenes M. 2002. Genetic linkage maps of two apricot cultivars (*Prunus armeniaca* L.) and mapping of PPV (Sharka) resistance. *Theor Appl Genet* 105(2-3):182–191.

Joobeur T, Periam N, de Vicente MC, King GJ and Arús P (2000) Development of a second generation linkage map for almond using RAPD and SSR markers. *Genome* 43:649–655.

Joobeur T, Viruel MA, de Vicente MC, Jauregui B, Ballester J, Dettori MT, Verde I, Truco MJ, Messeguer R, Balle I, Quarta R, Dirlewanger E and Arús P. 1998. Construction of a saturated linkage map for *Prunus* using an almond × peach F2 progeny. *Theor Appl Genet* 97(7):1034–1041.

Lerceteau-Kohler E, Guerin G, Laigret F andDenoyes-Rothan B. 2003. Characterization of mixed disomic and polysomic inheritance in the octoploid strawberry (*Fragaria* × *ananassa*) using AFLP mapping. *Theor Appl Genet* 107(4):619–628.

Liebhard R, Koller B, Gianfranceschi L and Gessler C. 2003. Creating a saturated reference map for the apple (*Malus x domestica* Borkh.) genome. *Theor Appl Genet.* 106(8):1497-508. Epub 2003 Apr 2.

Lu ZX, Sosinski B, Reighard G, Baird WV and Abbott AG. 1998. Construction of a genetic linkage map and identification of AFLP markers for resistance to root-knot nematodes in peach rootstocks. *Genome* 41:199–207.

Maliepaard C, Alston FH, Van Arkel G, Brown LM, Chevreau E, Dunemann F, Evans KM, Gardiner S, Guilford P, van Heusden AW, Janse J, Laurens F, Lynn JR, Manganaris AG, Den Nijs APM, Periam N, Rikkerink E, Roche P, Ryder C, Sansavini S, Schmidt H, Tartarini S, Verhaegh JJ, Vrielink-Van Ginkel M and King GJ. 1998. Aligning male and female linkage maps of apple (*Malus pumila* Mill.) using multi-allelic markers. *Theor Appl Genet* 97(1-2):60–73.

Mattiesch L and Debener T. 1999. Construction of a genetic linkage map for roses using RAPD and AFLP markers *Theor Appl Genet* 99(5)891-899.

Pierantoni L, Cho KH, Shin IS, Chiodini R, Tartarini S, Dondini L, Kang SJ and Sansavini S. 2004. Characterisation and transferability of apple SSRs to two European pear F1 populations. *Theor Appl Genet.* 109(7):1519-24. Epub 2004 Aug 31.

Rajapakse S, Byrne DH, Zhang L, Anderson N, Arumuganathan K and Ballard RE. 2001. Two genetic linkage maps of tetraploid roses *Theor Appl Genet* 103(4):575-583.

Sargent DJ, Clarke J, Simpson DW, Tobutt KR, Arus P, Monfort A, Vilanova S, Denoyes-Rothan B, Rousseau M, Folta KM, Bassil NV, Battey NH. 2006. An enhanced microsatellite map of diploid *Fragaria*. Theor Appl Genet. 112(7):1349-59. Epub 2006 Feb 28.

Sargent DJ, Davis TM, Tobutt KR, Wilkinson MJ, Battey NH and Simpson DW. A genetic linkage map of microsatellite, gene-specific and morphological markers in diploid *Fragaria*. *Theor Appl Genet*. 2004 Nov;109(7):1385-91. Epub 2004 Jul 29.

Stockinger EJ, Mulinix CA, Long CM, Brettin TS and Iezzoni AF. A linkage map of sweet cherry based on RAPD analysis of a microspore-derived callus culture population. *J Hered*. 1996 May-Jun;87(3):214-8.

Shimada T, Yamamoto T, Hayama H, Yamaguchi M and Hayashi T. 2000. A genetic linkage map constructed by using an intraspecific cross between peach cultivars. *J Jpn Soc Hortic Sci* 69:536–542.

Verde I, Lauria M, Dettori MT, Vendramin E, Balconi C, Micali S, Wang Y, Marrazzo MT, Cipriani G, Hartings H, Testolin R, Abbott AG, Motto M and Quarta R. (2005) Microsatellite and AFLP markers in the *Prunus persica* [L. (Batsch)]xP. ferganensis BC(1) linkage map: saturation and coverage improvement. Theor Appl Genet 111(6):1013-21.

Vilanova S, Romero C, Abbott AG, Llacer G and Badenes ML. An apricot (*Prunus armeniaca* L.) F2 progeny linkage map based on SSR and AFLP markers, mapping plum pox virus resistance and self-incompatibility traits. *Theor Appl Genet.* 2003 Jul;107(2):239-47. Epub 2003 Mar 14.

Viruel MA, Messeguer R, de Vicente MC, Garcia-Mas J, Puidomenech P, Vargas F and Arús P. 1995. A linkage map with RFLP and isozyme markers for almond. *Theor Appl Genet* 91(6-7):964–971.

Wang D, Karle R, Brettin TS and Iezzoni AF. 1998. Genetic linkage map in sour cherry using RFLP markers. *Theor Appl Genet* 97(8):1217–1224.

Yamamoto T, Kimura T, Shoda M, Imai T, Saito T, Sawamura Y, Kotobuki K, Hayashi T and Matsuta N. 2002. Genetic linkage maps constructed by using an interspecific cross between Japanese and European pears. *Theor Appl Genet.* 106(1):9-18. Epub 2002 Jun 19.

Yan Z, Denneboom C, Hattendorf A, Dolstra O, Debener T, Stam P and Visser PB. 2005. Construction of an integrated map of rose with AFLP, SSR, PK, RGA, RFLP, SCAR and morphological markers. *Theor Appl Genet.* 110(4):766-77. Epub 2005 Jan 26.

CHAPTER 2

SMALL EST LIBRARY ANALYSIS*

*This work was originally published in *BMC Plant Biology*:
Folta KM, Staton M, Stewart PJ, Jung S, Bies DH, Jesdurai C and Main D. 2005. Expressed sequence tags (ESTs) and simple sequence repeat markers from octoploid strawberry (Fragaria x ananassa). *BMC Plant Biol* 5:12.

Background

Commercial strawberry has a value of 1.4 billion dollars in the United States and represents a significant regional crop throughout the world (National Agricultural Statistical Services, 2006). *Fragaria x ananassa*, the commercially grown strawberry species, has an octoploid genome. The potential challenge of working with an octoploid species may have lead to limited molecular study and a resulting information discrepancy between strawberry and other common fruits. In early 2004, only 279 sequences existed in public databanks for octoploid strawberry, and only about 200 more for other *Fragaria* species. This gene deficit for strawberry represents a barrier to meaningful study of functional genomics, genetic mechanisms, as well as the molecular-systematic relationships between the octoploid strawberry, the *Rosaceae*, and other species. Basic sequence information would promote the development of transgenic technologies that would advance molecular-physiological studies and potentially benefit the grower and consumer.

To remedy this scarcity of sequence data, approximately1800 expressed sequence tags (ESTs) were sequenced from a whole-plant cDNA library derived

from various tissues of the Strawberry Festival cultivar by Kevin M. Folta of the University of Florida. This cultivar was chosen because of its east-coast and west-coast lineage as well as its range of favorable horticultural attributes. Strawberry Festival produces large, uniform, firm fruit, and is resistant to *Botrytis cinera*, the causative agent behind gray mold (Chandler et al., 2000). It is the predominant cultivar grown in Florida, and has been well studied in many reports of fungicide use, disease resistance, and post-harvest fruit quality.

Strawberry has significant potential as a research model and tool, and the lack of molecular markers for breeding makes sequence examination especially timely. ESTs are a valuable source for microsatellite markers, also known as simple sequence repeats (SSRs). SSRs are useful for plant genetic mapping and breeding because of their high reproducibility, multiallelic nature, codominant inheritance, and relative abundance (Powell et al., 1996). Information gained from the octoploid *Fragaria* species will also translate to defining molecular markers to facilitate mapping in both the diploid species (e.g. *Fragaria vesca*) and octoploid cultivars. Numerous researchers have utilized SSRs derived from EST sequence information to create or expand genetic map. This includes such plant species as cotton (Park et al, 2005), ryegrass (Favill et al, 2004), and red raspberry (Graham et al, 2004), a close relative of strawberry.

A comprehensive sequence database is the cornerstone of functional genomics studies, and this information will aid development of genetic tools in *Fragaria* and in the *Rosaceae* in general. Examination of expressed gene sequence variation in the octoploid may aid in the understanding of polyploidy

evolution and the progenitor diploid species contributing the octoploid genome. Sequence information constitutes a basis for eventual reverse-genetic and activation-tag studies. Both the diploid and octoploid species are excellent candidates for such studies as they are efficiently transformed and regenerated (Alsheikh et al., 2002; Passey et al., 2003; Rugini and Orlando, 1992), possess a diploid genome estimated at 164 Mb (Akiyama et al, 2001), just slightly larger than that of *Arabidopsis thaliana*, and can be rapidly propagated from seed (3-5 months) or runners (Darrow, 1966). A sampling of the strawberry transcriptome facilitates the initiation of such studies.

In this study over 1300 unique transcripts were assembled from 1,847 ESTs derived from whole-plant vegetative tissues 24 hours after salicylic acid treatment. The cDNA library was prepared from total RNA pooled from roots, petioles, stolons, leaves and meristems to generate a diverse set of transcripts with limited redundancy. Multiple analyses, such as developing a unigene set, annotation with putative function and identification of SSRs, opens additional paths that will speed research into strawberry physiology, evolution, genetics and genomics.

Despite the relatively small size of the EST data set, much useful information can be obtained from it. Many EST libraries of this size are being developed within the *Rosaceae* family as well as other species (Albert et al., 2005; Guterman et al., 2002; Jung et al., 2004). A thorough analysis will allow a maximum amount of information to be extracted from the sequences. However, no standardized protocol exists for the bioinformatics for small EST libraries.

The analysis presented here represents a first step toward a standardized pipeline for efficient and comprehensive analysis of small EST datasets.

Materials and Methods

Sequence Processing

A total of 1847 EST clones were sequenced at the University of Florida ICBR Core Facility using ET Terminator (Amersham Inc, Schaumburg, IL) from the 3' end. I processed the sequences using publicly available software incorporated in a fully automated in house script (ProcEST.pl). I converted sequence trace files into FASTA formatted sequence and quality score files using the PHRED (Ewing et al., 1998) base-calling program. I identified and masked vector and host contamination using the sequence comparison program CROSS_MATCH (Gordon et al., 1998). Vector trimming excised the longest non vector sequence and further trimming removed low quality bases (less than phred score 20) at both ends of a read. I discarded sequences if they had greater than 5% ambiguous bases, more than 40 PolyA or Poly T bases or less than 100 high quality bases (minimum phred score of 20). Using this protocol, 81% of the sequences (1505) were considered high quality and submitted to the NCBI public EST repository. To reduce redundancy and increase transcript length I assembled the high quality sequences using the contig assembly program CAP3 (Huang and Madan, 1999). I performed various assemblies using different CAP3 parameters to identify the build that produced the most effective assembly requiring the least manual editing. I selected more stringent parameters (- p 90 -d 60) to prevent over assembly and help identify potential paralogs. I refined the assembly where possible using homology to the SWISS-PROT database to indicate contig

accuracy. I determined likely homology by comparing the contigs and clones against the SWISS-PROT database (Boeckmann et al., 2003) using the FASTX3.4 algorithm with an expectation value cut-off < 1e-6 (Pearson and Lipman, 1988). I deconstructed contigs whose clones showed difference in homology and joined contigs with the same sequence similarity matches to other contigs using default CAP3 parameters. I derived the unigene data set by combining the contig and singleton data sets.

Functional Characterization

I performed functional characterization of the unigene data set that consisted of pairwise comparison of both the high quality clones and the contig consensus sequences against the NCBI nr (Wheeler et al., 2005), SWISS-PROT (Boeckmann et al., 2003), and the *Arabidopsis* protein (Rhee et al., 2003) databases using the FASTX3.4 algorithm (Pearson and Lipman, 1988). The most significant matches (EXP < 1e -7 for NCBI nr and EXP <1e-6 for the SWISS-PROT and *Arabidopsis* protein searches) for each contig and individual clones in the library were recorded. I further classified the SWISS-PROT matches via the Gene Ontology tool (Harris et al., 2004).

I characterized the unigene sequences by comparison with the GenBank Rosaceae EST dataset (225741 as of 022805) and 256 peach mapped ESTs (Joobeur et al., 1998), downloaded from GDR. Using the BLASTN algorithm (Altschul et al., 1990), sequences with > 85% similarity over an alignment length of 100 bp were considered significant matches.

Open Reading Frame and Microsatellite Analysis

Open reading frames (ORFs) were identified in the ESTs using the software program FLIP (Bossard, 1997) and the longest ORF was recorded as the putative coding region. Simple Sequence Repeats (SSRs) were identified in the unigene data set using a modified version (CUGISSR, Jung et al., 2005) of a perl script SSRIT (Temnykh et al., 2001). I recorded SSRs for the final dataset of dimers with at least 5 repeats, trimers with at least 4 repeats, tetramers with at least 3 repeats, and pentamers with at least 3 repeats. Using the FLIP output, CUGISSR reports the location of SSRs and primers in the relation to the putative coding region. I used Primer3 (Rozen and Skaletsky, 2000) to attempt to generate primers for the SSRs using the default software parameters.

Data Storage and Web Interface

I uploaded all sequence, assembly, homology, ORF and SSR data as well as library, protocol, contact and publication information to the GDR. I developed GDR scripts (described in detail in Chapter 4) to allow users to browse, query or download all the project data.

Public Access and Dissemination

I developed a number of different EST project sections on the GDR including the *Fragaria* EST dataset detailed here. These web pages are extensively linked such that users can easily access data of interest regardless of the navigation entry point. To access the project pages for this EST project, users can go to the project page, which can be found by a link in the "projects" drop down menu in the top navigation bar. The resulting project page links this project: "Folta - University of Florida" (http://www.bioinfo.wsu.edu/gdr/projects

/fragaria/folta/FA_SEa/index.shtml). The sidebar for this project allows the user to view the project description, the library details, the processing protocol, a report on the successful clones, unigene details, gene homology pages, microsatellite analysis, contact information, and associated publication information. The cDNA phage library and individual clones generated in this study are available upon request from the Folta laboratory.

For members of the Rosaceae community who are interested in searching the dataset, the EST search page allows users to go directly to the *Fragaria* page (www.bioinfo.wsu.edu/cgi-bin/gdr/newFragariaSearch_ChooseForm.cgi?lib_name=FA_SEa). The ESTs and the unigene can be searched by clone name or accession number, by homology, and by features such as presence of a microsatellite or component of a contig. Once an EST or contig has been selected, the sidebar allows users to view all information relating to the sequence (or consensus sequence), the library details, the assembly information, the open reading frame and microsatellites, homology, and for contigs, the component ESTs.

Results

Sequence Processing

A total of 1847 ESTs were sequenced, resulting in 1505 high-quality trimmed sequences that were submitted to GenBank on August 6, 2004. Of the 342 sequences that failed to meet high-quality standards, one failed for having more than 5% N's and 341 failed for having less than 100 high quality bases. Representing a success rate of 81.5%, the resulting submitted sequences have an average length of 613 bp and an average PHRED value of 35. PHRED values are

a quality score assigned to each base in a sequence and range from 4 to 60 with higher values corresponding to higher quality. These scores are associated with error probability based on a logarithmic distribution. The cut-off of 35 was chosen to maximize high quality bases; 35 represents the likelihood of error as less than .01%. The submitted sequences have an average of 478 high quality bases per sequence, and an average length of contiguous high quality bases of 267.

Functional Characterization

The primary method of inferring sequence function is to computationally examine levels of similarity to experimentally verified proteins or putative proteins. I employed both FASTA and BLAST software to compare the unigene developed from the Folta *Fragaria* EST library against known databases. In order to gain as much information as possible, I chose to use multiple databases ranging from verified amino acid sequences to putative nucleotide sequences.

I used the FASTX3.4 algorithm to compare the unigene sequences against three protein databases. NCBI's nr database represents the most comprehensive protein database available, including all publicly-available putative amino acid sequences. I downloaded the database from NCBI on February 15, 2005, and it contained 2,321,663 proteins. The FASTX algorithm with a cutoff of E < 1e-7 yielded matches for 1068 of the total unigene set, or 81.9% (Table 2.1). An E value reflects the degree of statistical confidence a researcher may have in a given alignment; it incorporates information on the length of the alignment, the percent of identity within the alignment and the size of the database. An E value of less than 1e-7 suggests high confidence in the alignment being significant.

I also performed a comparison against SWISS-PROT that yielded a lower number of significant matches. The SWISS-PROT database version 46.0 contains 172,233 sequences and was downloaded on March 2, 2005. This database is a curated and highly-annotated database, and all the proteins have experimentally demonstrated function. Of the unigenes, 720 (55.2%) had results with a significant cut-off value of E < 1e-6 (Table 2.1).

The third protein database used was the *Arabidopsis* proteins developed by TAIR from the sequenced *Arabidopsis* genome. Chosen as the model dicot sequenced genome most closely related to strawberry; this database contained 29,161 proteins and was downloaded on February 28, 2005. Using an E-value cut-off of <1e-6, 1080 unigenes (82.8%) were found to have significant matches (Table 2.1).

Only 194 unigenes (14.9%) were found to have no significant matches to any of the three protein databases utilized in the functional characterization. These sequences may represent long untranslated regions, structural RNAs, or *bona fide* proteins without characterization in the protein databases used.

I compared the *Fragaria* unigenes to publicly-available Rosaceae ESTs in order to assess how *Fragaria* relates to the rest of the *Rosaceae* family at the gene sequence and content levels. I employed the BLASTN algorithm for the nucleotide homology searches. I downloaded the publicly available *Rosaceae* ESTs on February 28, 2005, including 225,741 ESTs from five different genera (*Fragaria*, *Prunus*, *Rosa*, *Malus* and *Pyrus*). Using a stringency requirement of greater than or equal to 85% identity over at least 100 base pairs, I found 965

unigenes (74.0%) to have matches (Table 2.1). Since this dataset is composed of public ESTs, it contains a large amount of redundancy. The majority of public ESTs have been sequenced from the 5' end, so ESTs generated from the 3' end in this case may be less likely to find homologs in a search against public ESTs. Still, of the 194 unigenes that did not show significant homology with the protein database searches, 64 had homologs represented in the *Rosaceae* EST set. This leaves 130 transcripts without any functional annotation.

In a final attempt to characterize the transcripts with no information, the 130 sequences were run against the InterPro suite of databases using InterProScan. InterPro is a composite database that incorporates information from multiple protein family, domain, and functional site databases (Mulder et al, 2005). The InterProScan tool searches all of these databases in an attempt to find regions of similarity in the query sequence (Quevillon et al, 2005). The search of the 130 *Fragaria* sequences yielded no functional matches or new information.

Linkage relationships have been identified for many peach ESTs and have facilitated placement on the peach genetic map. A total of 295 peach ESTs have been conclusively anchored to the genetic maps by sharing BACs with genetic markers previously used for BAC hybridization (Horn et al, 2005). Comparison of the strawberry unigene to this set of peach ESTs presents a basis for developing linkage relationships between the established peach (Dirlewanger et al, 2004a) and growing *Fragaria* linkage maps (Sargent et al, 2004). Of the 1304 unigenes, 22 had significant (>= 85% identity over at least 100 base pairs) matches to the mapped peach ESTs (Table 2.1).

Table 2.1: Sequence similarity search results for the Fragaria unigene sequences.

	NCBI's nr	SWISS-PROT	TAIR's Arabidopsis proteins	Rosaceae ESTs	Mapped Peach ESTs
Algorithm	FASTX3.4	FASTX3.4	FASTX3.4	BLASTN	BLASTN
Database Size	2,321,556	172,233	29,161	225,741	295
Number of Sequences with Matches	1068	720	1080	965	22
Sequences with Matches	81.9%	55.2%	82.8%	74.0%	1.7%

Open Reading Frames and Microsatellite Analysis

I identified simple sequence repeats (SSRs) in the strawberry unigene set. 206 unigenes were found to contain 241 total SSRs with trinucleotides being the most common motif length (Table 2.2). The motifs found were grouped into categories with AG/GA/CT/TC being the most common (Table 2.3). To examine the distribution of SSRs in the putative coding region and the UTR, I detected open reading frames in the unigenes using the FLIP program. FLIP was able to identify a potential ORF in 1297 of the 1304 strawberry unigenes (99.5%). Based on the longest predicted ORF for each unigene, 160 (66.4%) of the SSRs are located in putative coding regions. Putative primers were successfully predicted by primer3 in a total of 199 SSRs in 171 different unigenes. 140 of these are located in ORFs.

Table 2.2: Motif lengths for SSRs with putative primer sequences.

Motif Length	In an ORF	NOT in an ORF	Total
2 bp	42	23	65
3 bp	87	16	103
4 bp	10	15	25
5 bp	1	5	6
All	140	59	199

Table 2.3: Most common motifs for SSRs with putative primer sequences.

Motif	Number of Microsatellites
AT|TA	15
AG|GA|CT|TC	68
AC|CA|TG|GT	11
GC|CG	0
AAT|ATA|TAA|ATT|TTA|TAT	2
AAG|AGA|GAA|CTT|TTC|TCT	40
AAC|ACA|CAA|GTT|TTG|TGT	4
ATG|TGA|GAT|CAT|ATC|TCA	6
AGT|GTA|TAG|ACT|CTA|TAC	0
AGG|GGA|GAG|CCT|CTC|TCC	22
AGC|GCA|CAG|GCT|CTG|TGC	13
ACG|CGA|GAC|CGT|GTC|TCG	7
ACC|CCA|CAC|GGT|GTG|TGG	10
GGC|GCG|CGG|GCC|CCG|CGC	7

Discussion

Fragaria x ananassa is complex polyploid, with evidence suggesting it arose from a spontaneous cross between *Fragaria virginiana* and *Fragaria chilioensis*. The genome contains contributions from at least three diploid species (Bringhurst, 1990; Senanayake and Bringhurst, 1967). Over the past century cultivation of octoploid strawberry has progressed solely on the careful efforts of breeders, physiologists and biochemists. This complex genome and coincidental difficult genetics has slowed the development of molecular markers and other tools that would benefit breeding efforts and understanding of strawberry genomics. This project marks a starting point to advance the traditional strawberry research avenues using modern molecular tools in structural and functional genomics studies. It demonstrates that computational tools may be used to mine diverse types of useful data from a single cDNA library. As these tools become available as web-based applications, small-scale sequencing efforts may extract valuable information that will shape research questions in under-represented crops like strawberry.

The transcripts characterized from this project will allow development of genomics resources for the study of other important physiological responses. A subset of these ESTs is shown in Table 2.4, and the full set of homology matches leading to the assignment of function can be found in Appendix A. These ESTs are relevant to the strawberry industry and may represent important molecular tools to researchers. The first set represents a series of ESTs with sequence homology to genes associated with the photoperiodic control of flowering. These include a close homolog to CONSTANS (CO), a likely transcription factor that

induces specific meristem identity genes under the appropriate photoperiod (Putterill et al., 1995; Valverde et al., 2004). A homolog of a critical regulator of meristem identity AGL20/SUPPRESSOR OF CO OVEREXPRESSION was also identified. This gene encodes a MADS-box transcription factor that likely functions downstream of CO in conferring light signals to the promoters of meristem identity genes (Onouchi et al., 2000). An EST representing VERNALIZATION INSENSITIVE 3 also was identified in this library. VIN3 is a protein shown to function downstream of CO in regulating seasonal flowering responses (Sung and Amasino, 2004). VIN3 is a chromatin-remodeling protein that represses FLC, a protein that negatively-regulates CO function (Michaels and Amasino, 2001) allowing the plant to appropriately time flowering relative to seasonal chilling.

Table 2.4: Unigenes putatively coding for genes involved in important physiological processes.

EST	Homolog	E Value
Photoperiodic Control of Flowering Time		
FA_SEa0007C05	B-box, zinc-finger protein CONSTANS	2.40E-21
FA_SEa0016A05	MADS box protein AGL20/SUPPRESSOR OF CONSTANS	4.90E-15
FA_SEa0002H08	VIN3 – Vernalization insensitive 3 protein	9.80E-34
Disease Resistance		
FA_SEa0004D05	Disease resistance protein (TIR-NBS-LRR class)	2.20E-22
FA_SEa0006F10	Enhanced Disease Susceptibility protein EDS5	7.10E-58
FA_SEa0007F04	Plant defensin PDF2.2	3.10E-22
FA_SEa0010B10	Pathogenesis-related thaumatin (PR5)	2.30E-53
FA_SEa0014H12	Putative thaumatin (PR5)	2.40E-21
FA_SEa0015A01	Harpin-induced protein	2.40E-27
FA_SEa0015D01	NDR1 family protein	7.00E-69
FA_SEa0017F09	Disease resistance protein (CC-NBS-LRR class)	5.40E-29
FA_SEa0020H01	Harpin-induced protein	2.50E-29
FA_SEa0010F01	glycosyl hydrolase family 17 p (PR2)	2.60E-12
FA_SEa0017H06	Osmotin-like protein (PR5)	3.40E-16
FA_SEa0001D03*	Peroxidase PRXR1 (PR9)	8.90E-54
FA_SEa0019D07	Bet v 1 (PR10)	2.30E-39
FA_SEa0012C06	Lipid transfer protein LPT4 (PR14)	1.40E-27
Photomorphogenesis		
FA_SEa0004E09	B-zip transcription factor HY5	4.60E-37
FA_SEa0001C09*	NON-PHOTOTROPIC HYPOCOTYL 3	3.20E-31
FA_SEa0006H04*	Far-red impaired / FAR1	3.30E-29

*SSR detected in this sequence

Analysis of this dataset revealed a suite of likely homologs to pathogenesis-related (PR) genes, such as thionins, Ndr1, 1-3-glucanase and chitinases, and LRR proteins. The prevalence of this family of proteins was not surprising as the plants were treated with salicylic acid 24 h before RNA harvest to enrich for PR genes in the library. These genes are of particular interest to plant scientists because of their potential to help define the mechanism(s) of disease resistance and susceptibility. It is possible that these genes may be especially useful targets for antisense or overexpression in unveiling these agriculturally-important traits, or possibly in the design of transgenic plants with heightened resistance to common plant pathogens. All of these facets are important, as strawberry cultivation requires copious application of fungicides and/or bacteriostatic compounds to ensure proper fruit set.

The information distilled from all of these analyses can now be used to design strawberry-specific probes to assess gene expression patterns and develop transgenics to directly test gene function. These important studies are underway and will facilitate comparisons between the biological sensory/response mechanisms in strawberry to those of model systems.

The apparent sequence conservation between *Fragaria* and other rosaceous tree crops suggests that cross-species microarray studies may be productive within the *Rosaceae*. This study demonstrates that less than 10% of the ESTs are unique to strawberry. This value is likely inflated, as ESTs by nature contain variable untranslated regions and other features that may preclude efficient identification of homologs. Of the 1305 ESTs, 965 have strong sequence

similarity with other Rosaceae ESTs. Those featuring at least 85% homology over 100 bases have an average identity of 88.6%. Considering only the best match found for each unigene, the rate of similarity between the unigenes and the ESTs was 90.7%. The high degree of similarity may be a useful platform for comparisons between molecular-mechanistic differences exhibited between diverse species with little sequence variation. Here, the diversity within the Rosaceae is likely due to variation in gene expression as well as sequence error. EST data and microarray technologies are an ideal platform to study these patterns.

SSRs derived from ESTs provide a basis to assign linkage relationships to known gene products and such studies have been initiated in diploid strawberry (Sargent et al., 2004). In the EST collection presented here, a number of SSRs are present in transcripts correlating to putative allergens, regulators of the circadian clock, and general housekeeping genes. These transcripts can now be readily mapped in the diploid using existing populations, and such studies are currently underway. Furthermore, specific genes of interest can be studied for variation within diploid species or for intron-specific polymorphisms that will allow their assignment to the diploid strawberry linkage map. These studies will ultimately facilitate the generation of molecular markers to follow traits/genes of interest in the commercial cultivars, adding the resolution of molecular tools to complement conventional breeding strategies.

The general proportions of the different functional groups (Figure 1) reflect well the expected state of the mature plant transcriptome as reported in

previous studies. Transcripts encoding enzymes associated with the cell cycle, cytoskeleton or cell walls are not abundant as mature plants are less reliant on processes governing greater cell number or cell size. Approximately half of the transcripts associated with photosynthesis are members of the chlorophyll a/b binding protein family; the other half typically contains plastid-encoded transcripts. As expected, the majority of transcripts detected represent enzymes of general metabolism.

Conclusion

Although a small EST set, the complete suite of analyses performed demonstrates that a finite transcriptome snapshot may provide ample resources to seed additional study. Here a relatively small number of ESTs has provided sufficient information to engage in further molecular, physiological and genetic studies. For instance, the pretreatment with salicylate likely enriched the expression of pathogenesis-related transcripts that can now be used to study disease progression in specific strawberry cultivars with large variations in sensitivity and resistance. Clearly, the development of a comprehensive SSR catalog allows characterization of these potential genetic markers in the progeny of polymorphic cultivars, in an important crop species virtually devoid of linkage associations. Unlike other markers, EST-derived SSRs by definition originate from a sequence that is expressed, adding functional resolution to linkage groups built on structural polymorphisms. More importantly, the same suite of tools used to perform these analyses will be made available through a public interface at the GDR, making comparable analyses possible. These applications are an important rationale for sequencing and analysis of a limited EST set, as even a small

research program may find sufficient resources to initiate molecular-genetic study of an under-represented crop species.

References

Akiyama Y, Yamamoto Y, Ohmido N, Oshima M and Fukui K. 2001. Estimation of the nuclear DNA content of strawberries (*Fragaria spp.*) compared with *Arabidopsis thaliana* by using dual-stem flow cytometry. *Cytologia* 66:431-436.

Albert VA, Soltis DE, Carlson JE, Farmerie WG, Wall PK, Ilut DC, Solow TM, Mueller LA, Landherr LL, Hu Y, Buzgo M, Kim S, Yoo MJ, Frohlich MW, Perl-Treves R, Schlarbaum SE, Bliss BJ, Zhang X, Tanksley SD, Oppenheimer DG, Soltis PS, Ma H, DePamphilis CW and Leebens-Mack JH. 2005. Floral gene resources from basal angiosperms for comparative genomics research. *BMC Plant Biol.* 30;5(1):5.

Alsheikh MK, Suso HP, Robson M, Battey NH and Wetten A. 2002. Appropriate choice of antibiotic and Agrobacterium strain improves transformation of anti biotic-sensitive *Fragaria vesca* and F-v. semperflorens. *Plant Cell Reports* 20(12):1173-1180.

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215(3):403-410.

Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S and Schneider M. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31:365-370.

Boguski MS, Lowe TM and Tolstoshev CM. 1993. dbEST - database for "expressed sequence tags". *Nat Genet.* 4(4):332-2.

Bringhurst RS. 1990. Cytogenetics and Evolution in American *Fragaria*. *Hortscience* 25(8):879-881.

Bossard N. 1997. FLIP: a Unix program used to find/translate ORFs. In.: Bionet Software.

Chandler C, Legard D, Dunigan D, Crocker T and Sims T. 2000. 'Strawberry Festival' strawberry. *Hortscience* 35:1366-1367.

Darrow G. 1966. The Strawberry. New York: Holt, Rinehart and Winston.

Dirlewanger E, Cosson P, Howad W, Capdeville G, Bosselut N, Claverie M, Voisin R, Poizat C, Lafargue B, Baron O, Laigret F, Kleinhentz M, Arus P and Esmenjaud D. 2004a. Microsatellite genetic linkage maps of myrobalan plum and an almond-peach hybrid--location of root-knot nematode resistance genes. *Theor Appl Genet.* 109(4):827-38. Epub 2004 Jul 6.

Ewing B, Hillier L, Wendl MC and Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8(3):175-185.

Faville MJ, Vecchies AC, Schreiber M, Drayton MC, Hughes LJ, Jones ES, Guthridge KM, Smith KF, Spangenberg GC, Bryan GT and Forster JW. 2004. Functionally associated molecular genetic marker map construction in perennial ryegrass (Lolim perrenne L.). Theor Appl Genet. 100(1):12-32.

Gordon D, Abajian C and Green P. 1998. Consed: a graphical tool for sequence finishing. *Genome Res.* 8(3):195-202.

Graham J, Smith K, MacKenzie K, Jorgenson L, Hackett C and Powell W. 2004. The construction of a genetic linkage map of red raspberry (Rubus idaeus subsp idaeus) based on AFLPs, genomic-SSR and EST-SSR markers. Theor Appl Genet. 109(4):740-9.

Guterman I, Shalit M, Menda N, Piestun D, Dafny-Yelin M, Shalev G, Bar E, Davydov O, Ovadis M, Emanuel M, Wang J, Adam Z, Pichersky E, Lewinsohn E, Zamir D, Vainstein A and Weiss D. 2002. Rose scent: genomics approach to discovering novel floral fragrance-related genes. *Plant Cell.* Oct;14(10):2325-38.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R; Gene Ontology Consortium. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 32(Database issue):D258-261.

Horn R, Lecouls AC, Callahan A, Dandekar A, Garay L, McCord P, Howad W, Chan H, Verde I, Main D, Jung S, Georgi L, Forrest S, Mook J, Zhebentyayeva T, Yu Y, Kim HR, Jesudurai C, Sosinski B, Arus P, Baird V, Parfitt D, Reighard G, Scorza R, Tomkins J, Wing R and Abbott AG. 2005. Candidate gene database and transcript map for peach, a model species for fruit trees. *Theor Appl Genet*. 2005 Apr 22; [Epub ahead of print]

Huang X and Madan A. 1999. CAP3: A DNA sequence assembly program. *Genome Res.* 9(9):868-877.

T Joobeur, MA Viruel, MC de Vicente, B Jáuregui, J Ballester, MT Dettori, I Verde, MJ Truco, R Messeguer, I Batlle, R Quarta, E Dirlewanger, and P Arús. 1998. Construction of a saturated linkage map for *Prunus* using an almond x peach F-2 progeny. *Theoretical and Applied Genetics* 97(7):1034-1041.

Jung S, Jesudurai C, Staton M, Du Z, Ficklin S, Cho I, Abbott A, Tomkins J and Main D. 2004. GDR (Genome Database for Rosaceae): integrated web resources for Rosaceae genomics and genetics research. *BMC Bioinformatics* 5(1):130.

Michaels SD and Amasino RM. 2001. Loss of FLOWERING LOCUS C activity eliminates the late-flowering phenotype of FRIGIDA and autonomous pathway mutations but not responsiveness to vernalization. *Plant Cell* 13(4):935-941.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJ, Silventoinen V, Studholme DJ, Vaughan R and Wu CH. 2005. InterPro, progress and status in 2005. Nucleic Acids Res. 33 (Database Issue):D201-5.

National Agricultural Statistical Services. 2006. *United States Department of Agriculture. Noncitrus Fruits and Nuts: 2005.* Retreived Aug 23 2006 from http://jan.mannlib.cornell.edu/reports/nassr/fruit/pnf-bb/

Onouchi H, Igeno MI, Perilleux C, Graves K and Coupland G. 2000. Mutagenesis of plants overexpressing CONSTANS demonstrates novel interactions among *Arabidopsis* flowering-time genes. *Plant Cell* 12(6):885-900.

Park YH, Alabady MS, Ulloa M, Sickler B, Wilkins TA, Yu J, Stelly DM, Kohel RJ, el-Shihy OM and Cantrell RG. 2005. Genetic mapping of new cotton fiber loci using EST-derived microsatellites in an interspecific recombinant inbred line cotton population. Mol Genet Genomics 274(4):428-41.

Passey AJ, Barrett KJ and James DJ. 2003. Adventitious shoot regeneration from seven commercial strawberry cultivars (*Fragaria x ananassa* Duch.) using a range of explant types. *Plant Cell Reports* 21(5):397-401.

Pearson WR and Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* 85(8):2444-2448.

Powell W, Machray GC and Provan J. 1996. Polymorphism revealed by simple sequence repeats. Trends Plant Sci. 1, 215-222.

Putterill J, Robson F, Lee K, Simon R and Coupland G. 1995. The CONSTANS gene of *Arabidopsis* promotes flowering and encodes a protein showing similarities to zinc finger transcription factors. *Cell* 80(6):847-857.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R and Lopez R. 2005. InterProScan: protein domains identifier. Nucleic Acids Research 33: W116-W120

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, Miller N, Mueller LA, Mundodi S, Reiser L, Tacklind J, Weems DC, Wu Y, Xu I, Yoo D, Yoon J and Zhang P. 2003. The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Research* 31(1):224

Rozen S and Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol*. 132:365-86.

Rugini E and Orlando R. 1992. High-Efficiency Shoot Regeneration from Calluses of Strawberry (*Fragaria X Ananassa*-Duch) Stipules of Invitro Shoot Cultures. *J Hort Sci.* 67(4):577-582.

Sargent DJ, Davis TM, Tobutt KR, Wilkinson MJ, Battey NH and Simpson DW. 2004. A genetic linkage map of microsatellite, gene-specific and morphological markers in diploid *Fragaria*. *Theor Appl Genet*. 109(7):1385-91.

Senanayake YDA and Bringhurst RS. 1967. Origin of *Fragaria* Polyploids .I. Cytological Analysis. *Am J Bot* 54(2):221-&.

Sung S and Amasino RM. 2004. Vernalization in *Arabidopsis thaliana* is mediated by the PHD finger protein VIN3. *Nature* 427(6970):159-164.

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S and McCouch S. 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): Frequency, length variation, transposon associations, and genetic marker potential. *Genome Research* 11(8):1441-1452.

Valverde F, Mouradov A, Soppe W, Ravenscroft D, Samach A and Coupland G. 2004. Photoreceptor regulation of CONSTANS protein in photoperiodic flowering. *Science* 303(5660):1003-1006.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JU, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L and Yaschenko E. 2005. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2005 Jan 1;33(Database issue):D39-45.

CHAPTER 3

ROSACEAE UNIGENE DEVELOPMENT

Introduction

An important practical outcome of understanding gene function is integration of marker technology into breeding programs to enhance cultivar improvement. Many plants are economically and nutritionally important but are still limited by a lack of genetic and genomic information. This is true of the Rosaceae plant family that includes such valuable crops as apple, peach, strawberry, pear, almond, rose, blackberry, raspberry and other ornamental species. Much funding at the federal level has been dedicated to model species such as *Arabidopsis*, rice and maize with relatively little resources directed to non-model plants such as the Rosaceae. Research funding for plants such as these is, therefore, always at a premium, and researchers can significantly benefit from sharing knowledge and resources stemming from closely related species. It benefits the Rosaceae community of researchers to understand the degree of sequence conservation across the economically important members of the family and to estimate how well molecular tools and information developed in one will be useful for others.

Gene content and gene number are unknown in most plants. Only three high coverage plant genomes are available: *Arabidopsis thaliana* (Rhee et al., 2003), *Oryza sativa* (Itoh et al., 2007), and *Poplus trichocarpa* (Tuskan et al., 2006). However, expressed sequence tags (ESTs) are available for a large

number of plant species and facilitate gene discovery and gene sequence determination within these species. As detailed in the introductory chapter, considerable resources exist in Rosaceae with respect to ESTs. Large EST sequence sets are available for species such as *Malus x domestica* (cultivated apple) and *Prunus persica* (peach) while other species have relatively small EST sets. Using these resources, I have undertaken the creation of both Rosaceae genera and family wide unigenes to evaluate the degree of conservation between the species. The advantages as well as the drawbacks to unigenes are reviewed in the first chapter.

Creation of a unigene that examines the redundancy in EST datasets on a genera as well as family level will elucidate some of the overlapping genes and allow candidate gene studies to utilize data from closely related species. The ultimate set of genes for these closely related plants are expected to contain extensive homology. Mapping in various *Prunus* species, including peach, almond, apricot, cherry, *P. davidiana*, *P. cerasifera*, and *P. ferganensis*, has led to the conclusion that all diploid *Prunus* species share the same basic genomic complement and can be mapped as a single genome (Dirlewanger et al, 2004a). While most *Prunus* species are diploid, apples are allotetraploid. Using 30 loci from the reference *Prunus* map that have homologs in a saturated apple map, a putative high level of synteny and collinearity between the two component apple genomes and the *Prunus* genome was been established (Dirlewanger et al, 2004b). An initial investigation into strawberry yielded promising synteny

results. Sargent et al., 2007 demonstrated that twenty primer pairs from *Fragaria* amplified a product of the expected size in *Malus* and *Prunus*.

The Rosaceae family represents many different fruit types such as pomes (apples), drupes (peaches), and achenes (strawberries). Comparing the different species can allow identification of genes that control different fruit ripening, quality, taste, and other traits specific to the individual species. The species may also have different responses to their varying pathogens and stressors that can be obtained from the different tissues and stressors unique to each cDNA library.

The methods of assembling a unigene for an entire family with its inherent sequence differences have not been well characterized. Species-specific unigenes have become a resource from large online databases including NCBI's UniGene (Pontius et al., 2003; Wheeler et al., 2003), the TIGR Gene Indices (Quackenbush et al., 2000), the Sequence Tag Alignment and Consensus Knowledgebase (STACK) (Christoffels et al., 2001), and PlantGDB (Dong et al., 2004). Each database uses different data sources and algorithms. Unigenes are routinely created for one or multiple cDNA libraries in many individual species, for example wheat (Lazo et al., 2004), barley (Michalek et al., 2004), and soybean (Tian et al., 2004). Here two different methods will be analyzed and compared. The methods will utilize the assembly software CAP3 (Huang and Madan, 1999).

The final unigene will be a resource for researchers from many species and genomic specialties in the family. The unigene is mined for markers such as SSRs and SNPs in order to facilitate genetic and comparative mapping. Candidate gene studies and metabolic pathway analysis will be furthered by the

functional characterization of the unigenes through comparison to other sequence and protein motif databases. The information and results from the project are deployed online for browsing, searching, and downloading by the entire community. Genes of interest from varying species and cDNA libraries are highlighted both online and in this chapter.

Materials and Methods

Sequence Processing

I downloaded all the public Rosaceae ESTs from dbEST on June 14, 2006. As dbEST has minimal quality curation for submitted sequences, low quality or contaminated sequences are routinely found in their datasets (McEntyre and Ostell, 2005). To optimally filter this data set it is beneficial to obtain the original sequence trace files and associated quality values from the submitting author. The libraries processed through GDR as part of its community service were available with quality values, and I contacted a number of other researchers who had contributed significant Rosaceae EST data sets to dbEST to request sequence and quality files. For those sequences for which we could not obtain this information, I assigned an average default quality value of 15 for each base. All ESTs were screened against NCBI's UniVec vector sequence database (Kitts et al., in preparation) downloaded on June 6, 2006 using the software package cross_match (Gordon et al., 1998). The ESTs were filtered using the BLAST algorithm (E cut-off <1e-6) against genera-specific tRNA, rRNA, and snRNA sequences downloaded from GenBank. Sequences downstream of more than 10 consecutive A's or T's were trimmed as were the quality files to match the resulting sequences. Sequences with less than 100 base pairs were excluded from

further analysis. I curated the tissue information from each of the 151 cDNA libraries to correspond to the most applicable Plant Structure Ontology term (Ilic et al., 2006).

I divided the total trimmed sequences into the five represented genera: *Malus*, *Prunus*, *Fragaria*, *Rosa*, and *Pyrus*. As only 330 ESTs were available for *Pyrus* they wer excluded from further analysis. The other four genera datasets were assembled using CAP3 (Huang and Madan, 1999) with an overlap percentage parameter of 90 (-p 90). The resulting four sets of singlets and contigs were again assembled together by CAP3 with -p 90 to produce an overall putative Rosaceae family unigene. For the purpose of comparison, I produced another Rosaceae family unigene by directly assembling all the trimmed ESTs at -p 90 with CAP3. I chose the "p" parameter value based both on values analyzed by Wang et al. (2004) and previous Mainlab experience with EST assembly. The "p" parameter specifies the minimum percent identity of each overlap created by the program during alignment.

Assembly Functional Characterization

I made a thorough effort to functionally characterize all putative transcripts by comparing the unigene consensus sequences to various sequence databases. The BLAST suite of programs (Altschul et al., 1997) was used to compare the unigenes to both protein and nucleotide sequence sets with an expectation value (E value) cutoff < 1e-6. The comprehensive protein database Uniprot, which includes SWISS-PROT and TrEMBL (Wu et al., 2005), the TAIR-predicted *Arabidopsis* protein set (Rhee et al., 2003), and the JGI-predicted *Populus* protein set (provided by DoE Joint Genome Institute and Poplar Genome

Consortium) were utilized for the first round of functional characterization using BLASTX.

The Gene Ontology Consortium provides three ontological sets for gene characterization: biological process, cellular component, and molecular function (The Gene Ontology Consortium, 2000). The SWISS-PROT group at EBI provides keywords and mappings from these keywords to GO terms (http://www.geneontology.org/external2go/spkw2go). They also created and maintain their own smaller subset of the entire GO ontologies, referred to as "GOA Slim" (http://www.ebi.ac.uk/GOA/). Using these mappings, I assigned the unigene sequences to the three GOA Slim ontologies based on their best SWISS-PROT match.

I attempted to utilize the functional results in verifying the unigene assembly. The ESTs were examined to identify whether the ESTs comprising a contig also shared significant sequence similarity to a known protein. The top ten significant results for SWISS-PROT and TrEMBL were recorded for each EST. The results for each EST comprising a contig were compared to find matches. Contigs ultimately fell into one of four categories: (1) All ESTs share at least one sequence similarity match, (2) All ESTs with matches share at least one sequence similarity match but some ESTs have no significant matches, (3) No ESTs have significant matches, and (4) The ESTs with matches do not share a common match.

I used the *Malus x domestica* (apple) unigene produced from PlantGDB (Dong et al., 2004) as a comparison for our unigene sets. Using an E value cut-

off of 1^{e-9} TBLASTX, a sensitive BLAST program that includes 6-frame translation of query and database, was used to find significant matches. Putative unique transcripts from PlantGDB for the twelve most sequenced, most important, and evolutionarily diverse sets of plants were used for further comparison. These included *Arabidopsis thaliana* (thale cress), *Glycine max* (soybean), *Gossypium* (tree and upland cotton), *Hordeum vulgare* (barley), *Lycopersicon esculentum* (tomato), *Medicago truncatula* (barrel medic), *Oryza sativa* (rice), *Pinus taeda* (loblolly pine), *Solanum tuberosum* (potato), *Triticum aestivum* (bread wheat), Vitis *vinifera* (wine grape) and *Zea mays* (maize or corn). All of these species have more than 200,000 transcripts for assembly, maximizing the number of expressed genes represented. Two other species with smaller EST sets were used for comparison to another group of fruit trees: *Citrus clementina* and *Citrus sinensis* (sweet orange). Their putative unigene was assembled from 61393 and 94289 transcripts, respectively.

Conserved sequence motifs can be used to infer information about a coding region even if a known protein does not provide a stringent sequence similarity match. I used the InterPro suite of protein family, domain, and function site databases and the corresponding InterProScan tool to analyze the sequences in the final Rosaceae unigene. The following InterPro databases were scanned with default parameters: ProDom, TIGRFAMS, TMHMM, PRINTS, PROSITE, PIRSF, Gene3d, Pfam, and SMART (Mulder et al., 2007).

Marker and Oligo Mining

I mined both SSRs and SNPs from the unigene sets. The SSRs were detected using an in-house pipeline based on a modified version of SSR-IT

(Temnykh et al., 2001). Microsatellites were extracted if they contained dinucleotide motifs occurring at least 5 times or trinucleotides 4 times to give an overall length of at least 10 base pairs. Tetranucleotide or pentanucleotide were flagged at 3 repeats, the minimum to be considered a microsatellite. The minimum microsatellite motif repeat frequency parameters were selected based on discussions with researchers who had found polymorphism in cassava at these levels (D Main, personal communication). I used the software FLIP (Brossard, 1997) to predict the open reading frame (ORF) of each unigene and used this information to determine whether the microsatellites occur in a coding region or an untranslated region (UTR). For the purposes of marker and oligo mining only, unigenes without ORFs were assumed to be coding. The longest ORF predicted was used in the case of sequences with more than one possible ORF. Primer3 (Rozen and Skaletsky, 2000) was used to generate primers for the SSRs where possible using the default software parameters. The SNPs were generated with the autoSNP (Barker et al., 2003) package using default stringencies for the genera unigene contigs.

Microarray technology has grown to be an essential tool to monitor changes in gene expression patterns for different tissues, cultivars, treatments or conditions. The unigenes created can be used to produce the gene target sequences for inclusion in a microarray. An example platform might be a NimbleGen arrays with direct synthesis of isothermal oligonucleotides on a slide of approximately 55-70 bases. An array such as this could be a standardized platform for functional genomics for all researchers within this family. In an

effort to begin this important type of research, we used an algorithm developed in-house to identify 55-70 bp isothermal oligos from the ORF sequences of the unigenes at both the genera and the family level.

Data Dissemination and Download

The unigene versions presented in this paper are accessible by direct download on the GDR website (http://www.rosaceae.org). I created comprehensive html pages that document the project and include the ability to search the public ESTs or the unigenes through name, taxonomy, putative markers, or functional characterization results. I constructed tutorials on downloading, browsing, and searching EST and unigene data that can be found at www.bioinfo.wsu.edu/gdr/tutorial/index.shtml.

Results

EST Collection and Assembly

A total of 369,106 Rosaceae ESTs were downloaded from NCBI's dbEST. Quality values were available for 196,957 of these ESTs, leaving over 46% to be assigned a default quality value. Filtering and trimming left 359,001 ESTs representing 151 cDNA libraries and 17 species (Table 3.1). Curation of the tissues to Plant Structure Ontology was completed, and 20 tissue types were represented in the set with only 1.4% unknown. Malus x domestica is the most sequenced species of the set with 68.4% of the total ESTs and fruit tissues clearly dominated with 44.1% of the total tissues.

Table 3.1: Genus, species, and tissue representation in public Rosaceae ESTs after filtering.

[DM1]

ORGANISM	NUMBER	%
Fragaria	18729	5.2
x ananassa	5276	1.5
vesca	13453	3.7
Malus	250907	69.9
hybid rootstock	320	0.1
sieboldii	1126	0.3
x domestica	245545	68.4
x domestica x sieversii	3916	1.1
Prunus	83751	23.3
armeniaca	14710	4.1
avium	21	<0.1
avium x cerasus x canescens	84	<0.1
cerasus	12	<0.1
dulcis	3776	1.1
persica	65148	18.1
Rosa	5284	1.5
chinensis	1790	0.5
hybrid cultivar	3494	1.0

TISSUE	NUMBER	%
Carpel	47	<0.1
Flower	22829	6.4
Fruit	31245	8.7
Fruit Endocarp	73633	20.5
Fruit Epicarp	7543	2.1
Fruit Epicarp & Mesocarp	11822	3.3
Fruit Mesocarp	65377	18.2
Gynoecium	924	0.3
Inflorescence Meristem	8562	2.4
Leaf	55068	15.3
Petal	5284	1.5
Phloem	9240	2.6
Receptacle	20	<0.1
Receptacle & Achenes	33	<0.1
Root	11167	3.1
Seed	8169	2.3
Shoot	14450	4.0
Unspecified	4906	1.4
Vegetative Meristem	32739	9.1
Whole Plant	17170	4.8
Xylem	4979	1.4

Clustering ESTs into a unigene set reduces their inherent redundancy and aligning sequences into longer consensus sequences facilitates more effective

homology identification. The resulting unigene has contigs, consisting of overlapping sequences, and singlets that are low-frequency transcripts or otherwise cannot be associated with a contig. Unigene sets attempt to represent each unique gene at a particular loci in a single sequence. The resulting members of the unigene are either a consensus contig sequence based on many transcripts of the same gene or a stand alone singlet sequence from a single transcript. In creating the unigenes with CAP3, we chose a high stringency to avoid over-assembly. Over-assembly generally results in gene family members or other distinct genes being assembled into a single contig.

The trimmed ESTs were separated and assembled into the 4 genera unigenes: *Fragaria*, *Malus*, *Prunus*, and *Rosa*. This first level of assembly achieved an overall reduction of 66.7% from total ESTs to unigenes (Table 3.2). While these four unigenes represent a significant decrease in redundancy, these closely related genera are expected to share many genes. A unigene for the entire family would reduce the redundancy further, facilitate comparative genomics between family members, and highlight genes that are shared between family members. Two types of unigenes were produced: one with a CAP3 assembly of all the Rosaceae ESTs and one with a CAP3 assembly of the genera unigenes including contigs and singlets (Table 3.3). Among other advantages (see Discussion), the latter allows a higher degree of compaction.

Table 3.2: Genera Unigene Statistics

Genus	Number of Sequences	Number of Singlets	Number of Contigs	Number of Unigenes	Reduction (%)
Fragaria	18729	7073	2939	10012	46.5
Malus	250907	58982	23868	82850	67.0
Prunus	83751	14903	8818	23721	71.7
Rosa	5284	2258	705	2963	43.9

Table 3.3: Rosaceae Unigene Statistics

	Number of Sequences	Number of Singlets	Number of Contigs	Number of Unigenes	Reduction (%)
Rosaceae without using prior genera assembly	359001	120389	27751	148140	58.7
Rosaceae using prior genera assembly	119546*	76573	13764	90337	74.8

* This set consists of the total contigs and singlets from the four genera unigenes.

Despite the effort to assemble transcripts across species and genera, the clones tend to cluster within the same organism (Table 3.4). A total of 11,549 (83.9%) contigs consist of all ESTs from the same genera. The theory that sequences that match the same protein were not being merged across genera was further explored by selecting contigs consisting of all *Prunus* transcripts and all *Malus* transcripts that matched the same SWISSWISS-PROT protein. ClustalW was then used to perform a multiple sequence alignment on the ESTs underlying these contigs. These alignments confirmed that sequences from *Prunus* and

Malus are diverged and tend to group together, thereby preventing assembly across the two genera. These results are found in Appendix C.

Table 3.4: The distribution of genera within overall Rosaceae contigs

Contig Description	Number of Contigs	Frequency (%)
One Genus Represented	11549	83.9
Prunus ONLY	815	
Malus ONLY	10544	
Fragaria ONLY	168	
Rosa ONLY	22	
Two Genera Represented	2132	15.5
Malus and Prunus	1431	
Three Genera Represented	78	10.4
Four Genera Represented	5	<0.1

Assembly Functional Characterization

The unigenes were examined for sequence similarity by BLAST comparison to protein sequence databases. The SWISS-PROT database version 52 with 260,175 amino acid sequences provides a curated set of proteins with high levels of annotation and low levels of redundancy. The TrEMBL database (3,874,166 seqs) is a computer-annotated supplement to SWISS-PROT which contains all other publicly available proteins. These two databases coupled with the *Arabidopsis* proteins from TAIR and the *Populus* proteins v1.1 from JGI were used to putatively identify the function of as many clones and unigenes as possible (Table 3.5).

Using the best SWISS-PROT match for the Rosaceae unigenes, 31,486 (34.9%) were assigned to a GO Slim term. However, not all of these were assigned to a GO term in all three ontologies: 17,287 unigenes had biological process annotation, 19,017 had cellular component annotation, and 24,628 had molecular function annotation. The GO Slim charts for the molecular function ontology and the biological process ontology are displayed with the corresponding numbers of Rosaceae unigenes for each term in Tables 3.7 and 3.8, respectively.

The number of ESTs and assemblies available for many plant species allows comparative analysis to be performed. Both the common genes among plants as well as their unique or fast-evolving transcripts are of interest to researchers. The PlantGDB comparisons provide illuminating results about the relationship between the gene content of the Rosaceae family in comparison to a diverse set of other plants. PlantGDB produces a set of nonredundant Unique Transcripts (PUTs) for a variety of different plant species. Their procedure uses a series of clustering steps including pre-clustering and CAP3 assembly with a stringent overlap percent (-p 95) as well as Vmatch and PaCE. The *Malus x domestica* unigene version 154 was compared to our *Malus* unigene using TBLASTX. Their unigene achieves the same reduction in redundancy of the dataset when compared to our *Malus* unigene (72.0% vs. 71.7%). Considering both sets utilize the same public EST dataset, it is not surprising that 95.9% of our unigenes share significant similarity to one of their PUTs. Our larger combined genera dataset included *Malus* ESTs from species such as *sieboldii* and hybrid

rootstocks, which could have led to our slightly higher number of unigenes. The *Prunus* and *Fragaria* unigenes show ~72% sequence similarity to the *Malus* PUTs (Table 3.6).

Table 3.5: Frequency of unigene matches with protein databases using the BLASTx algorithm

	Frequency of Matches (%)				
Database	Rosaceae Unigene	*Malus* Unigene	*Prunus* Unigene	*Fragaria* Unigene	*Rosa* Unigene
SWISS-PROT	40.2	42.5	39.9	45.5	50.3
TrEMBL	67.1	69.6	67.5	69.9	76.9
Arabidopsis Proteins from TAIR	67.6	70.0	67.2	71.3	78.8
Populus Proteins v1.1	70.2	72.6	70.1	72.8	81.1

Table 3.6: Frequency of unigene matches with PlantGDB databases using the tBLASTx algorithm

	Frequency of Matches (%)				
Database	Rosaceae Unigene	*Malus* Unigene	*Prunus* Unigene	*Fragaria* Unigene	*Rosa* Unigene
Malus x domestica PUTs	85.7	95.9	72.2	72.8	81.3
Medicago truncatula PUTs	62.5	65.1	62.4	67.8	75.6
Glycine max PUTs	64.4	67.1	64.8	69.3	75.8
Citrus clementina PUTs	55.4	56.9	53.2	59.3	66.9
Citrus sinensis PUTs	58.1	60.9	58.7	63.0	70.4
Gossypium PUTs	64.6	67.4	64.7	68.9	76.1
Arabidopsis thaliana PUTs	64.4	66.8	64.4	69.5	75.5
Lycopersicon esculentum PUTs	61.3	63.9	61.2	67.3	74.2
Solanum tuberosum PUTs	62.0	64.6	62.1	67.5	74.1
Vitis vinifera PUTs	63.2	65.0	63.0	66.6	73.1
Zea mays PUTs	60.7	62.9	61.1	65.5	71.8
Triticum aestivum PUTs	59.6	62.1	59.9	64.7	70.8
Oryza sativa PUTs	62.4	64.5	63.0	67.1	72.7
Hordeum vulgare PUTs	59.2	61.7	59.3	82.7	70.1
Pinus taeda PUTs	54.9	56.9	55.3	60.4	66.5

Table 3.7: Rosaceae unigenes mapped to the GO Slim biological process ontology

GO Category	Rosaceae Unigenes	
	Number	Frequency (%)
GO:0007275 : development	926	1.0
→GO:0030154 : cell differentiation	422	0.5
GO:0007582 : physiological process	11297	12.5
→GO:0008152 : metabolism	7447	8.2
→GO:0009056 : catabolism	75	0.1
→GO:0043170 : macromolecule metabolism	6145	6.8
→GO:0009405 : pathogenesis	16	<0.1
→GO:0046903 : secretion	308	0.3
→GO:0050875 : cellular physiological process	4084	4.5
→GO:0008151 : cell growth and/or maintenance	4084	4.5
→GO:0006810 : transport	4084	4.5
GO:0009987 : cellular process	10754	11.9
→GO:0030154 : cell differentiation	422	0.5
GO:0050789 : regulation of biological process	2466	2.7

Table 3.8: Rosaceae unigenes mapped to the GO Slim molecular function ontology

GO Category	Rosaceae Unigenes	
	Number	Frequency (%)
GO:0003774 : motor activity	147	0.2
GO:0003824 : catalytic activity	15447	17.1
→GO:0004386 : helicase activity	493	0.5
→GO:0016491 : oxidoreductase activity	3536	3.9
→GO:0016740 : transferase activity	5413	6.0
→GO:0016787 : hydrolase activity	4575	5.1
→GO:0016829 : lyase activity	978	1.1
→GO:0016853 : isomerase activity	605	0.7
→GO:0016874 : ligase activity	1048	1.2
GO:0004871 : signal transducer activity	876	1.0
→GO:0004872 : receptor activity	807	0.9
GO:0005198 : structural molecule activity	264	0.3
GO:0005215 : transporter activity	1129	1.2
→GO:0005386 : carrier activity	369	0.4
→GO:0015075 : ion transporter activity	383	0.4
→GO:0015267 : channel or pore class transporter	212	0.2
GO:0005488 : binding	15746	17.4
→GO:0005515 : protein binding	543	0.6
GO:0016209 : antioxidant activity	27	<0.1
GO:0030234 : enzyme regulator activity	172	0.2
GO:0030528 : transcription regulator activity	14	<0.1
GO:0045182 : translation regulator activity	507	0.6

The procedure to verify the contigs via sequence similarity to a known protein showed less than 3% contained conflicting ESTs. Overall, 59.1% of genera contigs and 68.6% of Rosaceae contigs contained members all with significant similarity to the same known protein. Contigs not in these two categories were confounded by ESTs without a match to SWISS-PROT or TrEMBL proteins (Table 3.9).

Table 3.9: Verification of contigs through sequence similarity to known proteins

	Genera Contigs		Rosaceae Contigs	
Total Contigs	36365		13764	
Contigs verified by homology for all ESTs	21485	59.1%	9445	68.6%
Contigs verified by homology for all ESTs with results (some no Matches)	8947	24.6%	2330	16.9%
Contigs with no homology results for any ESTS	4871	13.4%	1675	12.2%
Contigs with homology conflicts between ESTs	1062	2.9%	304	2.2%

Of the 90,337 unigenes 45.8% had at least one identifiable protein motif from InterProScan. The top ten most common motifs for the Rosaceae unigene are listed in Table 3.10. InterProScan results are available with associated GO Terms, allowing groups of proteins associated with a certain function to be easily examined. One area of particular interest to many researchers is transcription regulation. Four GO terms (GO:0006355, regulation of transcription, DNA-

dependent; GO:0045449, regulation of transcription; GO:0003700, transcription factor activity; GO:0030528, transcription regulator activity) were used to extract the top ten motifs associated with transcription regulation (Table 3.11). A total of 1,765 unigenes, about 2.0% of the overall unigene set, were found to be involved in transcription regulation. Ninety seven motifs associated with transcription regulation were found in the set at least once.

Table 3.10: Most common InterProScan motifs in the Rosaceae Unigene

IPR Entry	Num of Unigenes	Description of motif
IPR000719	7902	Protein kinase
IPR001680	6660	WD-40 repeat
IPR002048	3832	Calcium-binding EF-hand
IPR001611	3691	Leucine-rich repeat
IPR002110	2820	Ankyrin
IPR002885	2815	Pentatricopeptide repeat
IPR000504	2386	RNA-binding region RNP-1 (RNA recognition motif)
IPR000626	2219	Ubiquitin
IPR001841	1879	Zinc finger, RING-type
IPR000894	1817	Ribulose bisphosphate carboxylase, small chain

Table 3.11: Most common transcription regulation associated InterProScan motifs in the Rosaceae unigene.

IPR Entry	Num of Unigenes	Description of motif
IPR001471	1684	Pathogenesis-related transcriptional factor and ERF
IPR001356	877	Homeobox
IPR002100	638	Transcription factor, MADS-box
IPR012287	456	Homeodomain-related
IPR001092	427	Basic helix-loop-helix dimerisation region bHLH
IPR003657	388	DNA-binding WRKY
IPR004827	369	Basic-leucine zipper (bZIP) transcription factor
IPR001789	334	Response regulator receiver
IPR003441	310	No apical meristem (NAM) protein
IPR001965	296	Zinc finger, PHD-type

The Rosaceae and genera unigenes were also compared to PlantGDB PUTs from fourteen other species of diverse evolutionary distance from the Rosaceae (Figure 3.1). Counter to the original theory that longer divergence time would lead to fewer shared genes, the unigenes show remarkably stable levels of sequence similarity across all fourteen species surveyed. 62% of the Rosaceae unigene had a match to the *Medicago truncatula* PUTs, a species also in the eurosids I clade, and 54.9% with *Pinus taeda*, a much more evolutionarily distant gymnosperm (Table 3.7). Figure 3.2 shows the Rosaceae unigene and the *Prunus* unigene compared to the *Malus* PUT set and the other 14 PUT sets. The Rosaceae unigene is heavily influenced by *Malus* ESTs that account for 70% of its input sequences. However, even as a control, the *Prunus* unigene also shows a

very similar number of matches to each plant except *Malus*. To *Malus,* it is more similar than to other plants but still less conserved than the Rosaceae unigene.

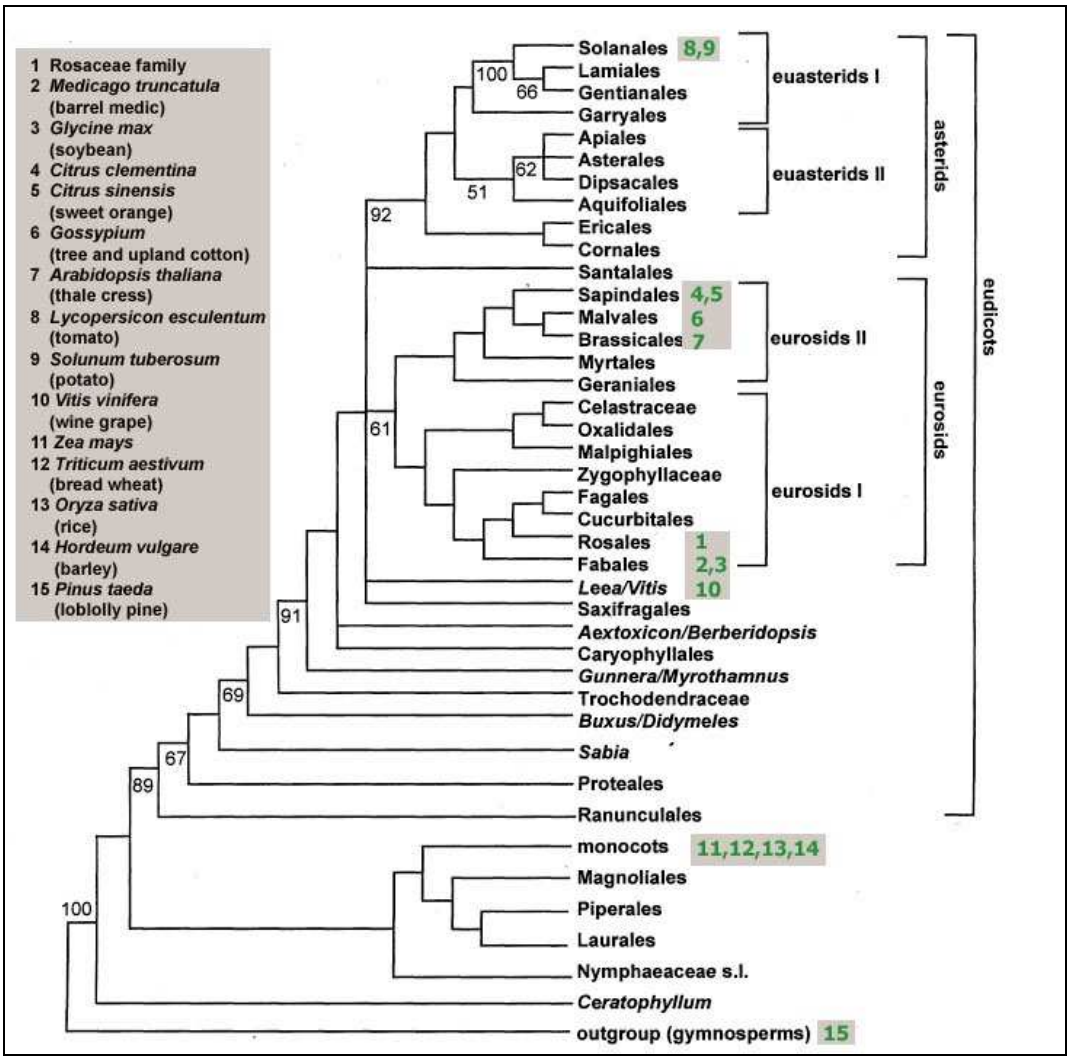


Figure 3.1: Picture adapted from Savolainen et al., 2000, Figure 4

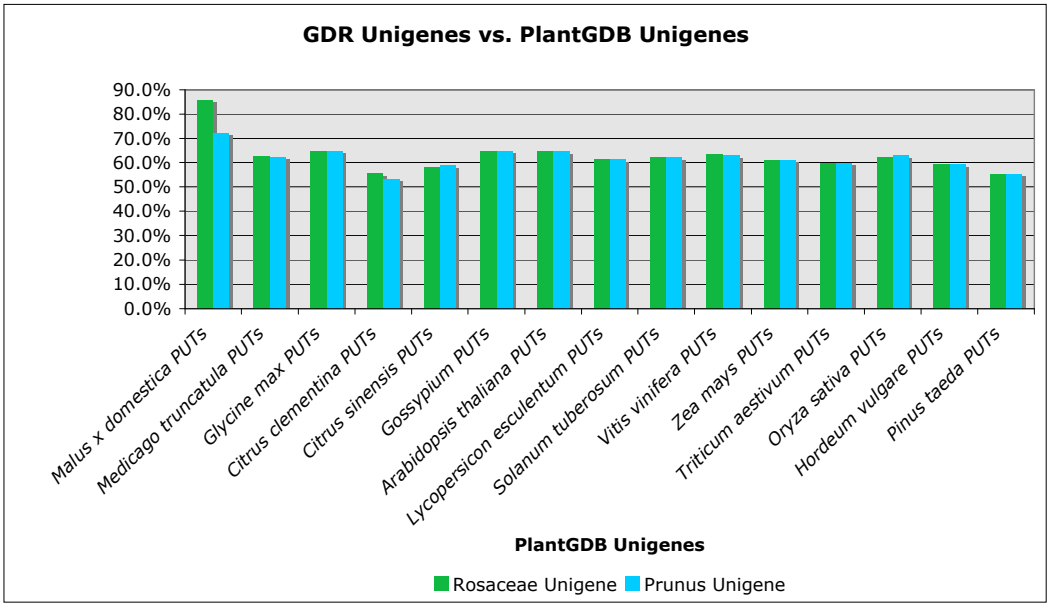


Figure 3.2: The percent of unigenes with significant similarity to various plant assemblies from PlantGDB.

I attempted to further characterize the relationship between these plant unigene sets by examining the Rosaceae unigenes and how many homologs were identifiable in other plants. The two *Citrus* unigenes were excluded from this analysis. They have considerably less EST data than the other groups for comparison. Their unigene is likely to be missing more of the genes from the genome and would influence the results. Interestingly, the unigenes tended to either match all the other plants in the group or to match none of them (Figure 3.3). The same chart was created for the *Prunus* unigene, and the percentages stayed within 1% of agreement with those listed below (data not shown).

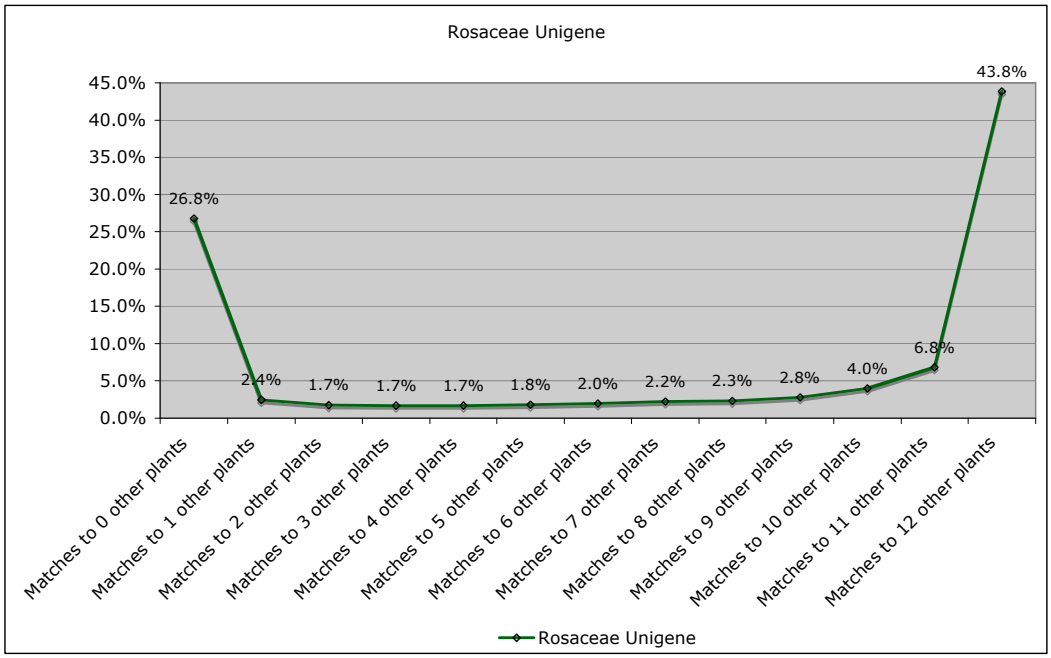


Figure 3.3: The percentage of Rosaceae unigenes that show sequence similarity to other plant unigenes from PlantGDB.

To examine members of a truly conserved set, I selected all the Rosaceae unigenes (total of 90337) with an E value match of less than 1e-50 to a member of all twelve other plant unique transcript sets. The most common SWISS-PROT matches from this set were used to infer protein function and are in the Table 3.12. These proteins are grouped by their SWISS-PROT name regardless of originating species. In a similar manner to the SWISS-PROT results, subsets of unigenes were extracted from the overall InterProScan results to examine further. The set of unigenes that stringently matched the twelve other plantGDB sets had a match rate of 91.5% with InterPro. The top ten most common motifs from this set are all found in the top 25 most common motifs from the overall unigene (Table 3.13).

Table 3.12: The most common Uniprot matches to Rosaceae unigenes with sequence similarity value of E<1e-50 to 14 other plant species.

Match	Number of Unigenes	Description
EF1A	77	Elongation factor 1-alpha (EF-1-alpha)
UBIQ	64	3-demethylubiquinone-9 3-methyltransferase (EC 2.1.1.64)
SRK6	55	Putative serine/threonine-protein kinase receptor [Precursor]. EC 2.7.11.1.
BAK1	47	BRASSINOSTEROID INSENSITIVE 1-associated receptor kinase 1 [Precursor]. EC 2.7.11.1
CB21	36	Chlorophyll a-b binding protein 151, chloroplast [Precursor]. LHCII type II CAB-151. LHCP
DFRA	34	Dihydroflavonol-4-reductase. EC 1.1.1.219. DFR. Dihydrokaempferol 4-reductase
RBS	33	Ribulose bisphosphate carboxylase small chain, chloroplast [Precursor]. EC 4.1.1.39
G3PC	32	Glyceraldehyde-3-phosphate dehydrogenase, cytosolic. EC 1.2.1.12. GAPC
PBS1	30	Serine/threonine-protein kinase PBS1. EC 2.7.11.1. AvrPphB susceptible protein 1
H3	30	Histone H3. HHT1
UBC4	28	Ubiquitin-conjugating enzyme E2-21 kDa 1. EC 6.3.2.19
TSJT1	28	Stem-specific protein TSJT1
TT12	28	TRANSPARENT TESTA 12 protein
ASO	26	L-ascorbate oxidase homolog [Precursor]. EC 1.10.3.3. Ascorbase.
CB26	25	Chlorophyll a-b binding protein CP26, chloroplast [Precursor]
ATG8	21	Autophagy-related protein 8 [Precursor]. Autophagy-related ubiquitin-like modifier ATG8.
PTR2	20	Peptide transporter PTR2. Histidine-transporting protein
MYB4	20	Transcription repressor MYB4. Myb-related protein 4

Table 3.13[DM2]: The most common InterPro matches to Rosaceae unigenes with sequence similarity of E<1e-50 to 14 other plant species.

IPR Entry	Number of (Conserved) Unigenes	Description of motif	Rank in Overall Unigene
IPR000719	2978	Protein kinase	1
IPR001680	1520	WD-40 repeat	2
IPR002048	1306	Calcium-binding EF-hand	3
IPR000626	1275	Ubiquitin	8
IPR000608	1246	Ubiquitin-conjugating enzyme, E2	11
IPR001806	1197	Ras GTPase	17
IPR000425	1083	Major intrinsic protein	16
IPR000504	953	RNA-binding region RNP-1 (RNA recognition motif)	7
IPR002016	801	Haem peroxidase, plant/fungal/bacterial	21
IPR001245	775	Tyrosine protein kinase	13
IPR000795	728	Protein synthesis factor, GTP-binding	23

Another interesting set of unigenes, those with matches to none of the other twelve plants, may represent sequencing errors, unfiltered contamination, or Rosaceae-specific genes. This set comprises of 24181 Rosaceae unigenes, of which 1391 are contigs (5.8%). Because protein to protein sequence comparisons are more accurate and likely to detect homology than nucleotide to nucleotide comparisons, we examined the matches to the SWISS-PROT and TrEMBL databases for this set manually. Only 862 had matches with less than 279 being matches to other plants. Most of the others were from bacteria or viruses that were presumably missed in quality filtering despite scanning with the UniVec

database. The remaining showed very specific categories of genes including those shown in Table 3.14. Further information including the matches and E values can be found in Appendix B. This set of unigenes without matches to the other 12 plants shows a very low percentage of InterProScan matches (4.8%). This is only slightly higher than the percentage with plant SWISS-PROT results (3.6%). The results do not correspond with the top results in the overall unigene; only three appear in the top 25 of the overall unigene motifs (Table 3.15). The set of transcriptional regulation associated motifs in the unigenes without matches included 27 motifs corresponding to 152 unigenes (Table 3.16).

Table 3.14: Categories of Uniprot matches to Rosaceae unigenes that do not match other plant transcripts.[DM3]

Gene Category	Number of Rosaceae Unigenes	*Malus**	*Prunus**	*Fragaria**	*Malus* and *Prunus***
Allergens	6	5	1	0	0
DNA Binding	18	14	4	0	0
Nucleic Acid Binding	7	6	1	0	0
Resistance Proteins	38	35	2	1	0
Ripening Related	22	14	7	1	0
Self-Incompatibility	5	3	1	1	0
Stress Response	26	16	8	2	0
Transcription Factors	22	16	2	4	0
Other Transcription/ Translation Regulation	5	2	3	0	0
Transposable Element Related	54	33	16	4	1

* Refers to either singlets of this genus or contigs with all transcripts coming from this genus. **Refers to a contig with member ESTs from both genera.

Table 3.15: The most common InterProScan matches to Rosaceae unigenes with no sequence similarity to 14 other plant species.

IPR Entry	Number of (Conserved) Unigenes	Description of motif	Rank in Overall Unigene
IPR010916	88	TonB box, N-terminal	49
IPR001810	85	Cyclin-like F-box	44
IPR001878	68	Zinc finger, CCHC-type	27
IPR002048	55	Calcium-binding EF-hand	3
IPR000583	54	Glutamine amidotransferase, class-II	112
IPR013032	53	EGF-like region	71
IPR007087	50	Zinc finger, C2H2-type	18
IPR003006	46	Immunoglobulin/major histocompatibility complex	95
IPR002052	40	N-6 Adenine-specific DNA methylase	175
IPR000719	32	Protein kinase	1

Table 3.16: The most common transcription regulation associated InterProScan matches to Rosaceae unigenes with no sequence similarity to 14 other plant species.

IPR Entry	Number of (Conserved) Unigenes	Description of motif	Rank in Overall Unigene
IPR003340	20	Transcriptional factor B3	15
IPR001647	16	Bacterial regulatory protein, TetR	40
IPR000847	15	Bacterial regulatory protein, LysR	34
IPR000005	15	Helix-turn-helix, AraC type	30
IPR003441	10	No apical meristem (NAM) protein	9
IPR002197	9	Helix-turn-helix, Fis-type	21
IPR000524	8	Bacterial regulatory protein GntR, HTH	55
IPR012287	8	Homeodomain-related	4
IPR001867	6	Transcriptional regulatory protein, C-terminal	49
IPR000418	6	Ets	37

Marker and Oligo Mining

An abundance of potential SSR markers were discovered in the EST and unigene data. An average of 21% of unigenes yielded a repeat with the *Malus* unigene showing the lowest relative amount (17.8%) and *Prunus* unigenes having the highest (24.8%) (Table 3.17). More than 33,000 SSRs were mined from genera contigs and 27,260 were found in Rosaceae contigs, however, the Rosaceae repeats are expected to be represented in the genera datasets. Around 80% of the *in silico* microsatellites yielded putative primers via the Primer3 package.

Table 3.17: SSRs mined from Rosaceae Unigene and Genera Unigene sets

Unigene Set	Number of SSRs	Frequency of SSRs with primers (%)	Number of SSRs outside of putative ORF	Number of SSRs with primers and outside ORF	Frequency of Unigenes with an SSR (%)
Rosaceae	27260	82.7	44.4	33.0	23.9
Malus	21465	82.4	40.3	29.5	17.8
Prunus	8320	81.3	49.7	36.8	24.8
Fragaria	2897	78.4	34.2	21.9	21.4
Rosa	760	79.7	40.6	66.67	19.5

The genera show differing distributions of 2, 3, 4 and 5 base pair motifs (Figure 3.4). The *Malus* and *Prunus* unigenes have a higher occurrence of 2 bp motifs than 3 bp motifs while *Fragaria* and *Rosa* unigenes were the opposite. All are similar in having a higher percentage of 3 base pair motifs within putative open reading frames as would be expected to conserve the triplet codon reading frame. The dinucleotide motifs exhibit a marked bias toward AG/GA/CT/TC (from 68% in *Prunus* to 82% in *Rosa*) and against CG/GC (<1% for all sets) (Figure 3.5). This is expected and has been noted in other studies with apple (Newcomb et al, 2006), *Prunus* species (Jung et al, 2005), and other plants (For example, Kumpatla & Mukhopadhyay, 2005). The other two categories of dinucleotides, AT/TA and AC/CA/TG/GT, were more variable in number between datasets.

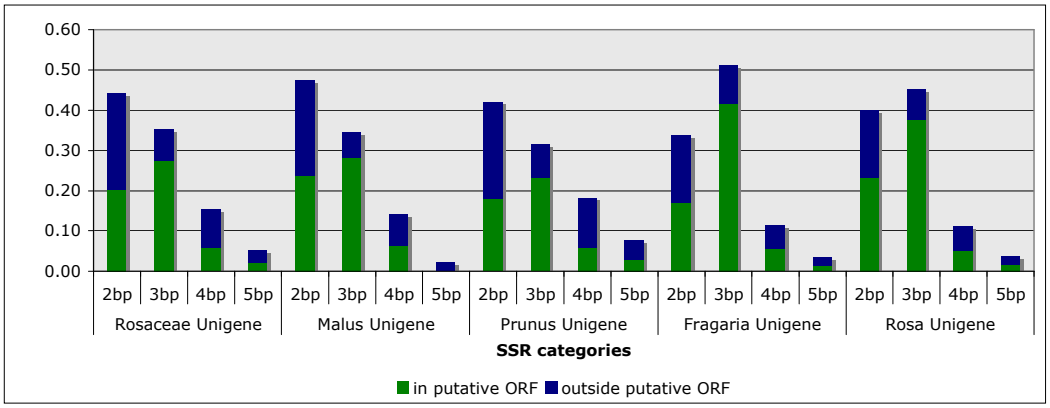


Figure 3.4: Motif length and ORF position of *in silico* mined SSRs.

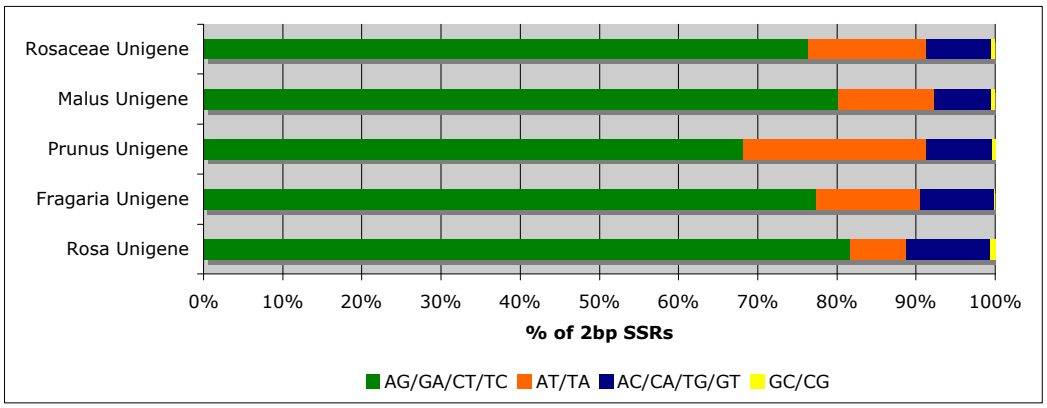


Figure 3.5: Dinucleotide motif frequency in in silico mined SSRs

SNPs were mined from the genera unigene contigs using autoSNP. A total of 20,244 SNPs were found from 31.5Mb of aligned sequence. The *Malus*, *Prunus*, and *Rosa* contigs showed the same frequency of SNPs (.07 per 100 bp) and similar frequency of transitions (~45%), transversions (~26%) and indels (25%). *Fragaria* differed with only .01 SNPs per 100 bp and a far higher

percentage of indels (37%). Transitions were more common than transversions or indels across all the genera (Table 3.18). SNPs were not mined from the Rosaceae contigs due to the higher than expected sequence divergence.

Table 3.18: Frequency of *in silico* mined SNPs across unigenes

Contig Set	Number of SNPs	Number of Transitions	Number of Transversions	Number of Indels	Frequency per 100 bp
Malus	14298	7060 (49.4%)	3836 (26.8%)	3402 (23.8%)	0.07
Prunus	5284	2345 (44.4%)	1353 (25.6%)	1586 (30.0%)	0.07
Fragaria	342	132 (38.6%)	83(24.3%)	127 (37.1%)	0.01
Rosa	320	140 (43.8%)	85 (26.6%)	95 (29.7%)	0.07

An in-house perl script was used to to select 55-70-mer oligos from the Rosaceae unigene that may be used to create a microarray chip. The script has numerous design features that contribute to quality unique oligo selection including filtering out matches to non-target sequences and setting ideal oligo length, GC content, melting temperature and salt concentration. For the Rosaceae unigene, the script was able to generate 54,750 oligos representing 20,675 of the unigene sequences (22.9%).

Discussion

EST Collection and Assembly

Two family-wide unigenes were constructed with CAP3 for comparison; one assembly uses all the ESTs, the other uses the contigs and singlets from the

previously generated genera unigenes. The latter was ultimately selected as the more useful. By utilizing the previous assemblies, multiple advantages were obtained. First, a higher confidence in the starting sequences is possible. Many miscalled bases of ESTs that have been assembled into contigs will be filtered out of the consensus sequence. Also, allelic variation between cultivars or species may be filtered into a single allele, allowing easier assembly between genera. The double-assembly method also results in a higher degree of compaction from ESTs to unigenes. The direct assembly yields 148,140 unigenes (58.7% reduction) while using the previous assembly and assembling again results in 90,337 unigenes (74.8% reduction). Plants in the same family are expected to share a highly similar gene compliment, and fewer unigenes is more likely to reflect the true relationship of genes between the Rosaceae. Finally, the assemblies allow researchers to start with a unigene from a genera assembly and find putative homologs in other Rosaceae species via the Rosaceae assembly.

Out of a total 13,764 family contigs, only 5 contigs from the Rosaceae unigene include ESTs from all four genera, and only 78 span three genera. This is quite unexpected as most genes should be duplicated in all the different species analyzed. Allelic variation and sequence divergence are the expected reasons for this. The stringent level of assembly may not allow as many homologs to be identified, but does lend confidence to those that are. As the EST database grows, future versions of this unigene may be able to merge the clones between genera more effectively and provide more useful information for comparative genetics. The lack of overlap significantly impacts the probability of transferring molecular

tools from one genus to another. Despite the expected gene overlap, it may be difficult to detect on a large scale due to higher sequence divergence than originally suspected.

To assess if transcripts matching the same genes are not being assembled I found the best Uniprot match for each Rosaceae unigene. I evaluated how often more than one unigene matches the same Uniprot protein. A total of 41,887 unigenes had a best match to the Uniprot dataset; 19247 proteins from Uniprot were a best match to a Rosacea unigene. 9,680 unigene sequences (10.7%) have a unique match while 80,657 unigenes (89.3%) share matches to a set of 9,567 Uniprot proteins. These "duplicated" matches range from 2 unigenes matching the same protein up to a maximum of 36 unigenes matching one protein. This highlights that many of the unigenes are probably expressing the same protein or a protein within the same family but due to sequence divergence they are not being assembled into one consensus sequence.

The ability to accurately assemble transcripts to a defined "set of genes" for an organism is known to be quite difficult. For example, the *Arabidopsis thaliana* PUTs from PlantGDB number 144,280 despite a well-curated set of genes from the genomic sequence of about 27,000. It is nearly impossible to adequately filter out all contamination such as untranslated regions and chimeras from an EST dataset, thus expanding the number of estimated unigenes. Low quality and short sequences can impact the ability of alignment algorithms to find significant overlap between sequences and can lead to further underassembly. We

expect our unigenes to be overestimates of the number of actual genes due to these factors.

Other unigenes for apple sequences have been published in the past. The UniGene pipeline from NCBI uses public ESTs (187969), mRNAs (386) and HTCs (9) from *Malus x domestica* to create clusters of nonredundant putative genes. One of the main requirements for cluster formation is a recognizable polyadenylation signal or tail. This reduces their final set of clusters to 14,626, an incomplete, but likely very accurate set for a large sample of genes. Our unigene attempts to be more comprehensive by not requiring 3' identification and publishing a putative consensus sequence. NCBI unigenes may be more useful for researchers that want to examine full-length genes only.

The HortResearch apple ESTs (151,687) have also been published as a unigene produced from the TIGR gene indices clustering tools including CAP3 (Newcomb et al, 2006). Their version has fewer contigs and singlets than our *Malus* unigene, uses a slightly higher threshold and achieves slightly more reduction. This can be attributed to our much larger starting dataset (250,907 ESTs) with more cDNA libraries of diverse tissues and development stages, which likely represents a larger pool of genes. The HortResearch unigene may be more useful for examining certain gene families with very similar members of more than 95% expected sequence similarity. Similar to our analysis, they performed SSR and SNP mining.

PlantGDB also produces a set of nonredundant *Malus x domestica* PUTs (PlantGDB-assembled Unique Transcripts). Their unigene has the advantage of

being comparable with numerous other species assembled in the same manner. It has been compared to our *Malus* assembly above. Over 96% of our *Malus* unigene has a significantly similar match to their dataset. The PlantGDB transcript set is limited to only *Malus x domestica* sequences while we utilize the small number of transcripts from other *Malus* species. This extra data could account for the 4% difference in non-matching sequences.

Assembly Functional Characterization

The group of Rosaceae unigenes with a match in all 12 other PlantGDB sets appears to represent a group of conserved proteins present throughout the plant kingdom. These conserved proteins may be essential to cellular functions and therefore unlikely to diverge or be lost over time. The most common motifs found by InterProScan in the overall set of unigenes match the common motifs in this set. This indicates that conserved gene motifs are found in many proteins throughout the plant kingdom.

In contrast to the similarity between the overall unigene and putatively conserved genes, the Rosaceae unigenes without matches to other plant sequences show fewer common motifs and demonstrate unique categories of gene functions. The idea that evolution of species depends on rapid changes in regulatory genes instead of the metabolic proteins themselves has been noted in other plants (Frary et al., 2000; Wang et al., 1999; Van der Hoeven et al., 2002). This may account for the large number of DNA binding, nucleic acid binding, transcription factors, and other transcription/translation regulation genes. They may have evolved significantly enough that direct sequence similarity with another nucleotide sequence is too slight to identify, but a protein sequence is more likely to retain

enough similarity at the amino acid level to infer homology. The putative fast-evolving genes showed no correlation to a particular species or tissue type.

A subset of the InterProScan results for the set of unigenes with no match to the other plants, the Pfam database results, was analyzed manually for interesting categories of gene function. The motifs verify many of the same categories found in the SWISS-PROT results. F-boxes and leucine-rich repeats represent the most common motifs from this group with 39 and 26 instances, respectively. Six B3 DNA binding domains indicate transcription factors, and 10 other motifs were linked to transcription factor activity. Transposable elements were represented by 5 retrotransposon gag protein motifs and 6 zinc knuckles, mostly found from retroviral gag proteins. Four dehydrin matches and two heat shock match indicate heat stress response.

Marker and Oligo Mining

SSRs mined *in silico* are valuable markers for mapping with an estimated 60 to 90% amplification success reported in other studies (Varshney et al., 2005). SSRs mined from the untranslated region of an EST are more polymorphic than those in a coding region due to lower selection pressure, and may be more likely to provide useful markers for mapping. The location of the unigene microsatellites in relation to putative open reading frames was assessed using the ORF-finding software FLIP. The percentage of markers outside a coding region varied from a low of 34.2% in the *Rosa* unigene to a high of 49.7% in the *Prunus* unigene. When narrowing down the SSRs to only those outside the putative ORF with primer prediction from Primer3, *Malus* has a set of 6330, *Prunus* has 3063,

Fragaria has 634, *Rosa* has 164, and the overall Rosaceae overall unigene has 9,006.

SSRs are valued in comparative mapping because of their high polymorphism, codominance, and high transportability between species. SSRs have been transferred from apple to pear (Pierantoni et al., 2004) and *Prunus* to *Malus* (Dirlewanger et al., 2004b). Twenty primer pairs flanking polymorphic regions of *Fragaria* were demonstrated to amplify a product of the expected size in *Prunus* and *Malus* (Sargent et al., 2007) suggesting SSRs may also be transferable between these genomes. SSRs that occur within an open reading frame are less polymorphic and more conserved between species, making them especially useful for comparative mapping. As *Malus* and *Prunus* have the most available sequences, they also share the most contigs. There are 640 SSRs in the overall Rosaceae contigs containing *Malus* and *Prunus* ESTs. These represent a starting point of sequences that must be tested in the lab for amplification and polymorphism.

The SSRs reported in the study by Jung et al., 2005 and other studies can now be reexamined against this larger dataset. The *Fragaria* dataset reported earlier contained 1505 of the 18729 ESTs available for this version of the *Fragaria* unigene, however, that library was from an octoploid species while the majority of the ESTs reported here are from the diploid *Fragaria vesca*. The same assembly algorithm and definition of an SSR were used for both. The results from that study seem to match the results of this one quite well. The octoploid set yielded 15.8% of sequences with an SSR as opposed to the 21.4%

reported here. Trinucleotides were the most common in both sets and the percent identified as inside a coding region remained virtually identical. Using the same parameters may have caused a strong correlation in the data but it could also indicate that a small sample of ESTs can be used to predict overall microsatellite trends for the species.

Jung et al., 2005 examined SSR rates from a putative peach unigene of 4539 sequences and found only 4% of sequences contained SSRs, but they used a more stringent definition of microsatellite that required at least 18 bp. A similar rate was found in their almond unigene of 933 sequences. Reducing our set of SSRs to at least that length leaves 5.6% of sequences with a microsatellite. Newcomb et al., 2006 mined a large set of apple ESTs for SSRs using a similar definition of an SSR but a nonredundant set. They found 17.8% of ESTs with a putative SNP while we found 17% of our unigenes contained at least one SNP.

Similarly to the SSRs, the SNPs are a resource that can be utilized for mapping. SNPs were mined from the genera unigene contigs but not the Rosaceae unigene due to an expected higher level of sequence divergence and lack of transferability across species. However, SNPs can make an excellent marker for fine scale mapping and saturation within a species by developing primers that correspond to the differing nucleotide sequence and its surrounding sequence or by developing CAPs (Cleaved Amplified Polymorphic Sequence). These markers use primers outside the SNP to obtain a PCR product and then restriction enzymes are used to find the SNP.

The *Fragaria* contigs showed much lower levels of polymorphism from the SNP analysis. An estimate of 1 SNP per 100 base pairs in *Fragaria* differed from the 7 per 100 base pairs found in the other three sets of unigenes. This may not be statistically significant as there were fewer contigs to examine as compared with the much larger *Prunus* and *Malus* contig sets. Also, the low percentage could be due to a higher amount of inbreeding in strawberries when compared to the other crops.

Microarrays can elucidate the differences in expression levels from mRNA samples from a variety of conditions including different tissues, development stages, or environmental conditions. This technology has helped researchers find genes involved in seed germination (Duque et al., 2003), maturing stems (Casu et al., 2004) and leaf senescence (Lin and Wu, 2004). Microchips have been used in plants to evaluate transcriptional response to biotic (Narusaka et al., 2003; Reymond et al., 2000; Whitham et al., 2004) and abiotic (Oztur et al., 2002) stresses. The citrus family (Forment et al., 2005) and *Arabidopsis* researchers (Horvath et al., 2003) have already proven the usefulness of an array that contains sequences from multiple related species. Also, the rise in availability of EST sequences and the complete *Arabidopsis* and rice genomes indicates many coding sequences of plants are highly conserved, especially those with core biological functions (Munkvold et al., 2004; Van der Hoeven et al., 2002). This assures that not only will the target sequences be useful across species, but also that they can be functionally characterized through sequence similarity searching.

A microarray for the Rosaceae will become a standardized platform for functional genomics for all researchers within this family. The main problems with microarray research currently include difficulties with normalization and analysis of the data (Quackenbush, 2001) as well as comparing and reproducing results from different microarray platforms (Kothapelli et al., 2002). By providing a flexible yet standardized microarray chip for the entire community, the results produced should be more comparable and have less overall variation. The GDR will be able to provide a repository for the raw data giving researchers the option to analyze the data with different statistical techniques and reanalyze old data as new software packages and statistics are developed in this growing area.

Conclusion

Other EST unigenes based on a certain species of the Rosaceae family have been created, but the unigene I created investigates the overlap of ESTs across the whole Rosaceae family of species. With many plants of economic importance and limited economic funding, it is fitting for genomic researchers to investigate the amount of overlap between the species and genera and to estimate how molecular tools from one species may be applied to another. Our unigene was not able to assemble transcripts between genera very effectively indicating that sequence divergence is an issue that will have to be addressed via better bioinformatics assembly methods. The overall assembly does elucidate many useful genomic features such as markers and candidate genes that researchers can access online quickly and easily. Regular updates of this unigene by means of the GDR team will continuously improve the information available. New data will be

incorporated into the original sets of sequences and new, more effective bioinformatics tools may be added to the processing pipeline. This is just one of the ways that GDR will fulfill its function to add value to the genomic data for Rosaceae and disseminate it effectively.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.

Barker G, Batley J, O' Sullivan H, Edwards KJ and Edwards D. 2003. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics* 12(19(3)):421-422.

Biswas M, O'Rourke JF, Camon E, Fraser G, Kanapin A, Karavidopoulou Y, Kersey P, Kriventseva E, Mittard V, Mulder N, Phan I, Servant F and Apweiler R. 2002. Applications of InterPro in protein annotation and genome analysis. Brief Bioinform. 3(3):285-95.

Brossard N. 1997. FLIP: a Unix Program used to find/translate orfs. Bionet.software<Message-ID:347B3A1B.794BDF32@bch.umontreal.ca>

Casu RE, Dimmock CM, Chapman SC, Grof CP, McIntyre CL, Bonnett GD and Manners JM. 2004. Identification of differentially expressed transcripts from maturing stem of sugarcane by in silico analysis of stem expressed sequence tags and gene expression profiling. *Plant Mol Biol.* 54(4):503-17.

Dirlewanger E, Graziano E, Joobeur T, Garriga-Caldere F, Cosson P, Howad W and Arus P. 2004b. Comparative mapping and marker-assisted selection in Rosaceae fruit crops. *Proc Natl Acad Sci U S A* 101(26):9891-6. Epub 2004 May 24.

Dong Q, Schlueter SD and Brendel V. (2004) PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res*. 32(Database issue):D354-D359.

Duque P and Chua NH. 2003. IMB1, a bromodomain protein induced during seed imbibition, regulates ABA- and phyA-mediated responses of germination in *Arabidopsis*. *Plant J*. 35(6):787-99.

Forment J, Gadea J, Huerta L, Abizanda L, Agusti J, Alamar S, Alos E, Andres F, Arribas R, Beltran JP, Berbel A, Blazquez MA, Brumos J, Canas LA, Cercos M, Colmenero-Flores JM, Conesa A, Estables B, Gandia M, Garcia-Martinez JL, Gimeno J, Gisbert A, Gomez G, Gonzalez-Candelas L, Granell A, Guerri J, Lafuente MT, Madueno F, Marcos JF, Marques MC, Martinez F, Martinez-Godoy MA, Miralles S, Moreno P, Navarro L, Pallas V, Perez-Amador MA, Perez-Valle J, Pons C, Rodrigo I, Rodriguez PL, Royo C, Serrano R, Soler G, Tadeo F, Talon M, Terol J, Trenor M, Vaello L, Vicente O, Vidal Ch, Zacarias L and Conejero V. 2005. Development of a citrus genome-wide EST collection and cDNA microarray as resources for genomic studies. *Plant Mol Biol*. 2005 Feb;57(3):375-91.

Frary A, Nesbitt TC, Grandillo S, Knaap E, Cong B, Liu J, Meller J, Elber R, Alpert KB and Tanksley SD. 2000. Fw2.2: A quantitative trait locus key to the evolution of tomato fruit size. Science 289, 85-88.

The Gene Ontology Consortium. 2000. Gene Ontology: tool for the unification of biology. Nature Genet. 25:25-29.

Gordon D, Abajian C and Green P. 1998. Consed: a graphical tool for sequence finishing. *Genome Res.* 8(3):195-202.

Horvath, D.P., Schaffer, R., West, M. and Wisman, E. 2003. *Arabidopsis* microarrays identify conserved and differentially expressed genes involved in shoot growth and development from distantly related plant species. *Plant J*. 34: 125–134.

Huang X and Madan A. 1999. CAP3: A DNA sequence assembly program. *Genome.* Res. 9(9):868-877.

Ilic K, Kellogg EA, Jaiswal P, Zapata F, Stevens PF, Vincent LP, Avraham S, Reiser L, Pujar A, Sachs MM, Whitman NT, McCouch SR, Schaeffer ML, Ware DH, Stein LD and Rhee SY. 2006. Plant Structure Ontology. Unified Vocabulary of Anatomy and Morphology of a Flowering Plant. *Plant Physiol.* Dec 1 [Epub ahead of print].

Itoh T, Tanaka T, Barrero RA, Yamasaki C, Fujii Y, Hilton PB, Antonio BA, Aono H, Apweiler R, Bruskiewich R, Bureau T, Burr F, de Oliveira AC, Fuks G, Habara T, Haberer G, Han B, Harada E, Hiraki AT, Hirochika H, Hoen D, Hokari H, Hosokawa S, Hsing Y, Ikawa H, Ikeo K, Imanishi T, Ito Y, Jaiswal P, Kanno M, Kawahara Y, Kawamura T, Kawashima H, Khurana JP, Kikuchi S, Komatsu S, Koyanagi KO, Kubooka H, Lieberherr D, Lin Y, Lonsdale D, Matsumoto T, Matsuya A, McCombie WR, Messing J, Miyao A, Mulder N, Nagamura Y, Nam J, Namiki N, Numa H, Nurimoto S, O'Donovan C, Ohyanagi H, Okido T, Oota S, Osato N, Palmer LE, Quetier F, Raghuvanshi S, Saichi N, Sakai H, Sakai Y, Sakata K, Sakurai T, Sato F, Sato Y, Schoof H, Seki M, Shibata M, Shimizu Y, Shinozaki K, Shinso Y, Singh KN, Smith-White B, Takeda J, Tanino M, Tatusova T, Thongjuea S, Todokoro F, Tsugane M, Tyagi AK, Vanavichit A, Wang A, Wing RA, Yamaguchi K, Yamamoto M, Yamamoto N, Yu Y, Zhang H, Zhao Q, Higo K, Burr B, Gojobori T & Sasaki T. (2007) Curated genome annotation of Oryza sativa ssp. japonica and comparative genome analysis with *Arabidopsis* thaliana. *Genome Research* 17: 175-183.

Jung S, Abbott A, Jesudurai C, Tompkins J and Main D. 2005. Frequency, type, distribution and annotation of simple sequence repeats in Rosaceae ESTs. *Funct Integr Genomics*. 5(3):136-43. Epub 2005 Mar 11.

Jung S, Jesudurai C, Staton M, Du Z, Ficklin S, Cho I, Abbott A, Tomkins J and Main D. 2004. GDR (Genome Database for Rosaceae): integrated web resources for Rosaceae genomics and genetic research. *BMC Bioinformatics* 5:130.

Kitts PA, Madden TL, Sicotte H, and Ostell JA - Manuscript in preparation. The UniVec website can be accessed at http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html

Kumpatla SP and Mukhopadhyay S. 2005. Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. *Genome* 48:985-998.

Lin JF and Wu SH. 2004. Molecular events in senescing *Arabidopsis* leaves. *Plant J.* 39(4):612-28.

McEntyre J and Ostell J, eds. 2005. The NCBI Handbook. Bethesday(MD):National Library of Medicine (US), NCBI. Article : GenBank: The Nucleotide Sequence Database by Ilene Mizrachi updated July 27th, 2004

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJ, Thomas PD, Valentin F, Wilson D, Wu CH and Yeats C. 2007. New developments in the InterPro database. Nucleic Acids Res. 35(Database issue):D224-8.

Munkvold JD, Greene RA, Bermudez-Kandianis CE, La Rota CM, Edwards H, Sorrells SF, Dake T, Benscher D, Kantety R, Linkiewicz AM, Dubcovsky J, Akhunov ED, Dvorak J, Miftahudin, Gustafson JP, Pathan MS, Nguyen HT, Matthews DE, Chao S, Lazo GR, Hummel DD, Anderson OD, Anderson JA, Gonzalez-Hernandez JL, Peng JH, Lapitan N, Qi LL, Echalier B, Gill BS, Hossain KG, Kalavacharla V, Kianian SF, Sandhu D, Erayman M, Gill KS, McGuire PE, Qualset CO and Sorrells ME. 2004. Group 3 chromosome bin maps of wheat and their relationship to rice chromosome 1. *Genetics*. 168(2):639-50.

Narusaka Y, Narusaka M, Seki M, Ishida J, Nakashima M, Kamiya A, Enju A, Sakurai T, Satoh M, Kobayashi M, Tosa Y, Park P and Shinozaki K. 2003. The cDNA microarray analysis using an *Arabidopsis* pad3 mutant reveals the expression profiles and classification of genes induced by Alternaria brassicicola attack. *Plant Cell Physiol*. 2003 Apr;44(4):377-87.

Newcomb RD, Crowhurst RN, Gleave AP, Rikkerrink EHA, Allan AC, Beuning LL, Bowen JH, Gera E, Jamieson KR, Janssen BJ, Laing WA, McArtney S, Nain B, Ross GS, Snowden KC, Souleyre EJF, Walton EF, and Yauk Y-K. 2006. Analyses of expressed sequence tags from apple. *Plant Physiology* 141:147-166.

Oztur ZN, Talame V, Deyholos M, Michalowski CB, Galbraith DW, Gozukirmizi N, Tuberosa R and Bohnert HJ. 2002. Monitoring large-scale changes in transcript abundance in drought- and salt-stressed barley. *Plant Mol. Biol*. 48(5–6): 551–573.

Pierantoni L, Cho KH, Shin IS, Chiodini R, Tartarini S, Dondini L, Kang SJ and Sansavini S. 2004. Characterisation and transferability of apple SSRs to two European pear F1 populations. *Theor Appl Genet*. 109(7): 1519-24.

Reymond P, Weber H, Damond M and Farmer EE. 2000. Differential gene expression in response to mechanical wounding and insect feeding in *Arabidopsis*. *Plant Cell* 12: 707–719.

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, Miller N, Mueller LA, Mundodi S, Reiser L, Tacklind J, Weems DC, Wu Y, Xu I, Yoo D, Yoon J and Zhang P. 2003. The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Research* 31(1):224

Rozen S and Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol*. 132:365-86.

Savolainen V, Chase MW, Hoot SB, Morton CM, Soltis DE, Bayer C, Fay MF, Anette YdB, Sullivan S and Qiu Y-L. 2000. *Systematic Biology* 49(2):306-362.

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S and McCouch S. 2001. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res*. 11:1441-1452.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Dejardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjarvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leple JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouze P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K,Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y and Rokhsar D. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*. 313(5793):1596-604.

Van der Hoeven R, Ronning C, Giovannoni J, Martin G and Tanksley S. 2002. Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *The Plant Cell* 14, 1441-1456.

Varshney RK, Graner A and Sorrells ME. 2005. Genic microsatellite markers in plants: features and applications. *Trends in Biotechnology* 23(1):48-55 . Available online 25 Nov 2004.

Wang R, Stec A, Hey J, Lukens L, and Doebley J. 1999. The limits of selection during maize domestication. *Nature* 398, 236-239.

Whitham SA, Quan S, Chang HS, Cooper B, Estes B, Zhu T, Wang X and Hou YM. 2003. Diverse RNA viruses elicit the expression of common sets of genes in susceptible *Arabidopsis thaliana* plants. *Plant J*. 33(2):271-83.

Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N and Suzek B. 2005. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*. 34(Database issue):D187-91.

CHAPTER 4

THE GENOME DATABASE FOR ROSACEAE

Introduction

The NSF funded the Genome Database for Rosaceae (GDR) in 2003 through award #0320544. The GDR was originally focused on peach genomic resources but has since grown to incorporate all the public structural and functional data for the Rosaceae family (Jung et al., 2004). The use of a curated and integrated relational database of sequence information coupled with an online interface is one of the most important genomic tools for researchers focusing on a species or a group of species. This type of website provides researchers with a comprehensive view of the data being generated in their area of interest and functions as a clearinghouse of news and information. These sites often have important roles in annotation, curation, and permanent data storage. Scientists from all over the world are able to access, analyze, integrate, and apply the data to their own research in a timely manner. The usefulness of this type of database has been repeatedly proven in other species. Examples of effective and fruitful plant genomics databases with comparative mapping data include TAIR for *Arabidopsis* (Rhee et al., 2003), the Sol Genomics Network for Solanaceous species (Mueller et al., 2005), and Gramene for the grasses (Jaiswal et al., 2006).

The large amount of data and dispersed worldwide community of researchers for Rosaceous species necessitates a properly curated and centralized database. The investigators of this project outlined three main goals for the

website: (1) to develop an organized and integrated web resource for peach genomics data to facilitate gene discovery in other member species by a comparative mapping approach, (2) to collect and integrate all Rosaceae genomics data, and (3) to develop online tools and resources for the Rosaceae community.

The Rosaceae research community has responded enthusiastically to this resource. Between July 2005 and June 2006 GDR had 262,284 hits by researchers from 44 countries. The community has elected a steering committee that published a White Paper outlining the future goals and needs for genomic research in this family (US Rosaceae Genomics, Genetics and Breeding Consortium, 2006). This paper calls for "enhanced Rosaceae genomics database resources." It recognizes the indispensable contribution of the GDR to the community and specifically calls for continued funding and expanded resources to manage the next wave of microarray and genomic sequence data.

The GDR currently contains all the publicly available Rosaceae sequences. ESTs are updated nightly from the dbEST at NCBI (McEntyre and Ostell, 2005). Regular annotation of the ESTs includes unigene creation, marker mining, and assignment of function through sequence similarity searches. Controlled vocabularies such as Gene Ontology (Harris et al., 2004) and Plant Structure Ontology (Ilic et al., 2006) are utilized in this process. Sequenced BACs, proteins, and organelle genomes are also available for download. GDR offers free EST library analysis for any researcher in the community and places

all results online for searching and downloading. Libraries analyzed by the GDR team currently span almond, peach, strawberry and raspberry.

GDR houses and maintains extensive mapping resources. CMap (Fang et al., 2003), a comparative map viewer, allows researchers to view numerous Rosaceae genetic maps and the peach transcriptome map (Horn et al., 2005). CMap includes the ability to display multiple maps simultaneously, to find maps with a certain feature or to find the number of contact points or features in common between two maps. Currently 37 maps spanning apple, pear, *Prunus*, almond, apricot, cherry, peach, raspberry, rose and strawberry are available for comparison. The genetically anchored peach physical map is available in WebFPC or WebChrom. These software packages are both downloaded from an Arizona Institue of Genomics website (http://www.genome.arizona.edu/software/fpc/download_web/) and allow viewing of contig and marker alignments. The TxE genera *Prunus* map is available as interactive linkage groups with anchored BACs and ESTs.

GDR provides bioinformatic tools for researchers such as dedicated sequence similarity servers. Users can run BLAST (Altschul et al., 1997) or FASTA (Pearson and Lipman, 1998) to compare their sequence of interest to Rosaceae ESTs, Rosaceae unigenes, the *Arabidopsis* genome, and other sequence datasets. The downloadable results are accessible via a web page which the user is directed to via email on completion of the search. The results include the name, description, and organism of the matching sequence as well as the expectation value, the beginning and ending indices of the overlap in the query and the match,

and percent identity across the overlap. Where a specific GDR dataset match is found, a hyperlink directs the user to all the associated information for that sequence in GDR. The database also provides a tool for searching for microsatellites within sequences and generating primer pairs for those SSRs. Users can assemble groups of sequences using CAP3 (Huang and Madan, 1999). All the tools allow the user maximum ability to change default parameters and customize their results.

GDR functions as a central and public communications hub for the community, providing mailing lists, message boards, and archives for the community and certain subgroups. Conference announcements, abstracts, and reports are also available. A page is devoted to the elected executive committee for the family and provides meeting announcements, minutes and publications. Researchers can browse the funding sources for the community and view abstract and progress for many projects. Quarterly newsletters keep members of the mailing lists abreast of GDR developments as well as overall community news, which can be submitted by any of the component mailing list members. Relevant publications from Pubmed (www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed) and Agricola (agricola.nal.usda.gov) are downloaded weekly and can be searched or browsed by author, title or keyword.

Ultimately, the GDR hopes to play an important role in providing added value to the Rosaceae genomic data being produced worldwide. Both automated and curated analysis is needed to link the sequence and biological information for this family of plants. Improved community integration and communication is

central for using Rosaceae synteny relationships to rapidly apply data from each species to other members of the family. This chapter will focus on my development of the functional genomics resources within GDR.

Infrastructure

The GDR is based upon an underlying relational database implemented using the Oracle Database Management System version 9.2.0. The database currently has a total of 57 tables to represent the various types of data in GDR and their properties. I created the tables relevant to the rest of the discussion, some of which are briefly outlined in Figure 4.1. GO term tables and tables not relevant to the main functional part of the database have been excluded.

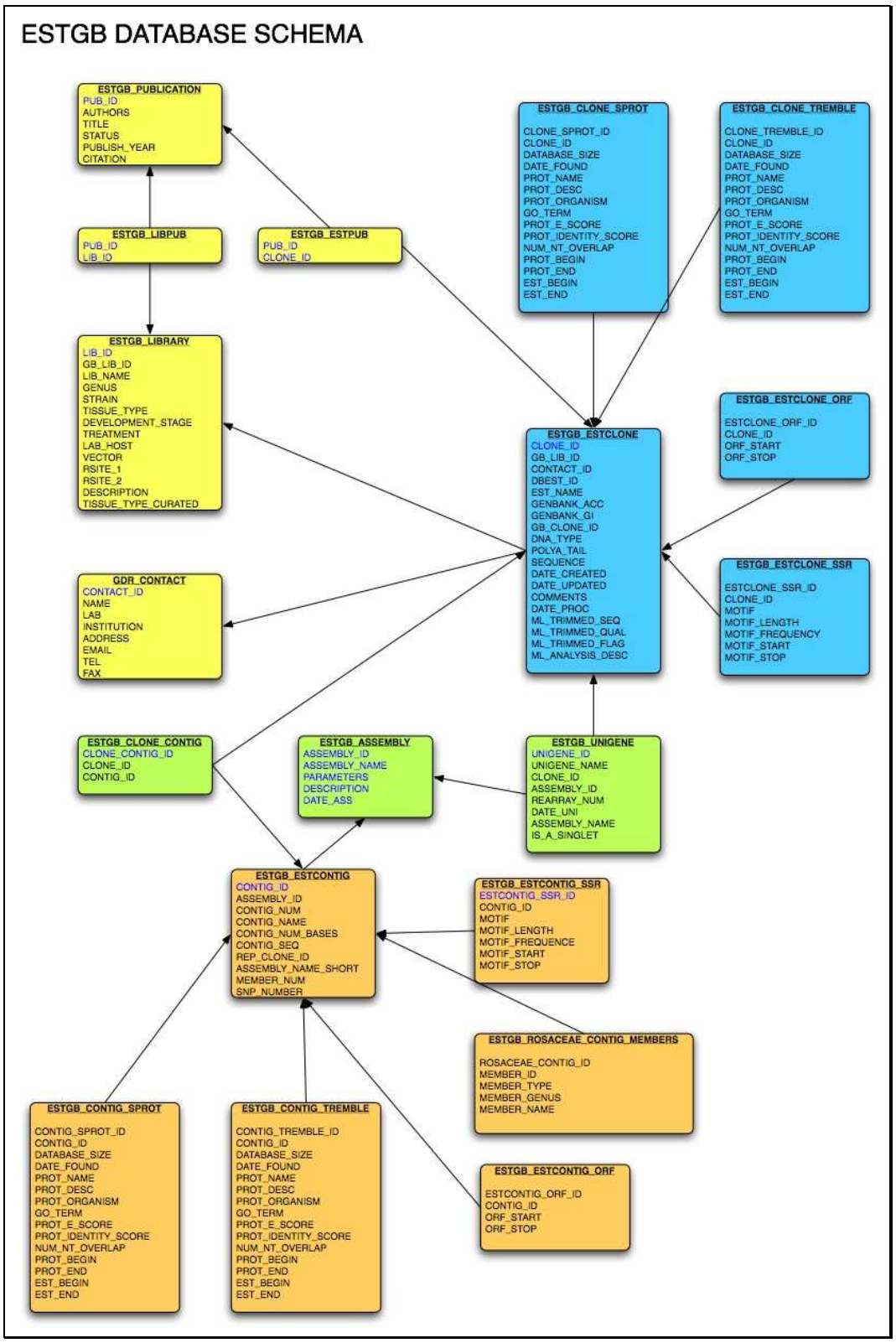


Figure 4.1: GDR Schema – Functional Genomics Tables

I accomplished data processing, annotation, and uploading with scripts written in the Perl programming language. Perl has the advantages of built-in, powerful string processing features via regular expressions and a large base of bioinformatics source code from the publicly available bioPerl packages. I developed perl code to handle the download of public EST data, automatic annotation and subsequent upload to the GDR database and also wrote the processing pipelines for EST data and unigenes in Perl. I continued this trend with web interfaces using Perl code embedded in Common Gateway Interface (CGI) scripts. These scripts can move data between the web server and the database, allowing the development of pages to view, query, and download data.

Navigation

The homepage for the GDR (Figure 4.2) was designed to allow users to easily find the data they are looking for as well as stay abreast of current developments in GDR and the community at large. The main navigation bar with drop down menus remains the same throughout all of the GDR pages to enable users to move seemlessly between sections of the database. The main page includes points of interest that list updates to the website including new data or tools and important community information.

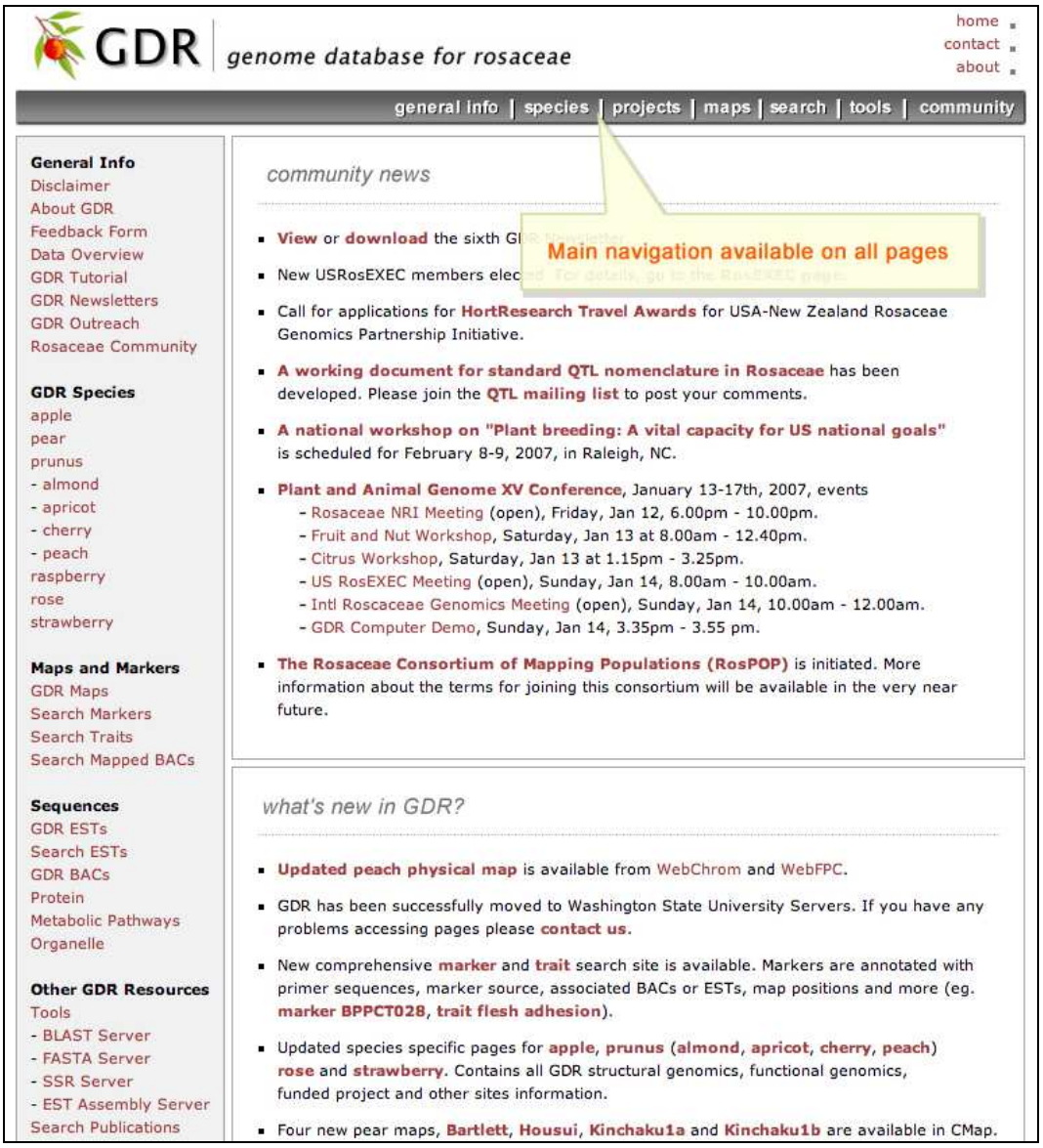


Figure 4.2: GDR Home Page

An important part of the international Rosaceae genomics effort is the ongoing development of extensive EST resources from a variety of different tissues and species. EST bioinformatics analysis has become a standardized process that is offered as a free service to all Rosaceae researchers. Currently, 25,752 ESTs have been analyzed by GDR and uploaded to GenBank. I

performed the analysis for 11,307 these. Another 7,085 are in process and expected to be publicly available soon. The full analysis includes trace file processing to obtain a high-quality clone library, assembly of the sequences to produce longer transcripts and reduce redundancy, and sequence annotation. The BLAST sequence similarity analysis tool (Altschul et al., 1990) is used to assign putative function to the ESTs and unigenes by comparison with the SWISS-PROT and TrEMBL databases (Wu et al., 2005). Researchers can also request comparison to additional databases such as other Rosaceae EST sequences, *Arabidopsis* proteins, predicted *Populus* proteins, etc. These matches are used to assign GO terms that facilitate searching sequences by keywords and grouping sequences by similar function. Marker mining is performed that identifies simple sequence repeats (SSRs or microsatellites) in all ESTs and unigene contigs. An example of this type of analysis was presented in Chapter 2. All of this data is made available publicly to researchers through the GDR both in html format and as downloadable files on a ftp website.

The cDNA libraries analyzed and publicly submitted by the GDR team represent a small portion of the public Rosaceae ESTs available. I designed a script that downloads and enters into the database all publicly available Rosaceae ESTs in dbEST on a nightly basis. I perform annotation of these sequences that is similar to the analysis of individual libraries. This includes genera and family-wide levels of unigene assembly with CAP3. Assignment of putative function by means of sequence similarity searching is performed and the SWISS-PROT and TrEMBL results are available online for viewing and searching. The assignment

of GO terms facilitates keyword searching and allows users to find genes with functions or in pathways of interest easily. The community is interested in genetic as well as comparative mapping in all the Rosaceous species. I mine the unigenes and ESTs for microsatellites to promote these mapping projects. Putative primers for the microsatellites are extracted from sequences using the software Primer3. The microsatellite files for download include useful information for researchers such as optimal marker characteristics and putative open reading frame (ORF) location. The most recent addition to the annotation pipelines is mining simple nucleotide polymorphisms (SNPs) from contigs. This will allow researchers to begin finer scale mapping once the reference maps for the varying species are saturated with microsatellites. I have ensured backward compatibility for GDR by offering full access to versions 2 and 3 of the unigenes as well as downloadable data for version 1. I created tutorials that are available on the website to guide users through accessing and searching all of the EST data and bioinformatic features.

Unigene Project Viewing and Access

The GDR provides not just access and storage of data but also conducts important data analysis. For expressed sequence tag data reducing redundancy and assigning putative function are necessary initial steps to utilize the data in further studies. The first project to be explored is the Rosaceae version 3 unigene. The creation and analysis performed for this project are covered in Chapter 3.

The main data overview page is a convenient place to start for finding and viewing EST projects (Figure 4.3). This page can be reached from the homepage by clicking the “GDR Data” link in the “General Info” drop down menu in the top

navigation bar. A chart at the bottom of this page lists the relevant EST data projects including libraries with in-house analysis and overall assemblies of public data. The Rosaceae v3 unigene can be reached by clicking on "Rosaceae Assembly".

Rosaceae EST Data:

- **SNP Analysis:** The publicly available autoSNP software was used to analyze the most recent versions of the Rosaceae unigene and the individual genera unigenes. A summary of the results and downloadable data can be found here.
- **EST Assemblies:** Assemblies of publicly available ESTs are available for the family Rosaceae and the genera Malus, Prunus, Fragaria, Rosa and Pyrus. Homology data for each assembly with SWISS-PROT is available to browse and download. Homology with the NCBI nr protein database and other rosaceae ESTs will be available soon.
- **EST Analysis:** Full bioinformatics analysis, including quality trimming, assembly, homology searching, microsatellite discovery and putative unigene development, is currently available for Clemson Peach and Almond , UF Octoploid Strawberry, and CRA ISF Peach ESTs. Rose and blackberry projects will be available soon.

Organism	Project	EST Number
Rosaceae	Rosaceae Assembly	359001
Fragaria	Fragaria Assembly	18729
x ananassa (octoploid strawberry)		
whole plant:	Kevin Folta - University of Florida	1505
vesca (diploid strawberry)		
unopened flower buds:	Tom Davis - University of New Hampshire	2717
Malus	Malus Assembly	250907
Prunus	Prunus Assembly	83751
amygdalus (almond)		
developing seed:	Albert Abbott - Clemson University	2794
persica (peach)		
fruit mesocarp:	Albert Abbott - Clemson University	9984
root:	Albert Abbott - Clemson University	Available Soon
shoot:	Albert Abbott - Clemson University	7085
fruit mesocarp:	Elisa Vendramin - CRA ISF	1667
Rosa	Rosa Assembly	5284

Figure 4.3: GDR Data Overview Page

The main page of the Rosaceae assembly version 3 project overviews the aims and basic results of the project (Figure 4.4). Users can navigate the rest of the project by clicking the links in the grey bar to the right of the main description. This lists the pages of data that can be accessed, including searching

the ESTs, library details, protocols and downloads, putative homology, microsatellite analysis, contact and publication information, and gene ontology (GO) classification. The genera unigene assemblies are available from the original data overview page and also include a page in their sidebar linking to SNP analysis. This is not available for the Rosaceae unigene due to the higher amount of sequence divergence. The pages from each of these links can be found in Appendix B. The content of each page is explained in Table 4.1.

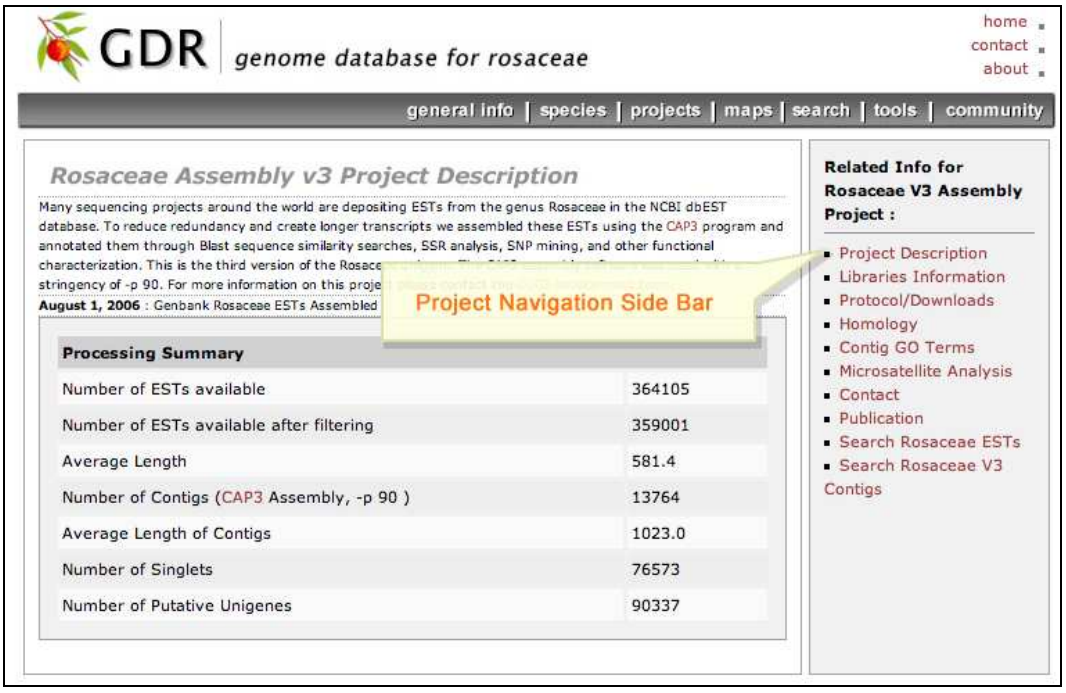


Figure 4.4: Rosaceae Unigene Version 3 EST Project Home Page

Table 4.1: Overview of Unigene Project Pages

Page	Description
Project Description	Overviews the main aims and techniques of the project. The chart gives overall statistics.
Libraries Information	Includes charts with the number of separate libraries, species, tissues, and development stages included in the project as well as a list of the number of ESTs in each species. A link opens a separate page with the details of each library individually.
Protocol & Downloads	Specifies the bioinformatic software applications and methods utilized in the project with links to references. The text is followed by links to downloadable files including sequences of clones and unigenes, BLAST results, and Excel spreadsheets detailing ORFs, SSRs, and primers.
Homology	Describes the BLAST searches against the Uniprot databases used for functional annotation and links to Excel spreadsheets containing this data. Links also take users to a search page.
GO Terms	An expandable tree is available for each of the three GO term ontologies: biological process, cellular component, and molecular function. Clicking on the term will return all the unigenes mapped to this term via their SWISS-PROT sequence similarity results.
Microsatellite Analysis	The number of SSRs and motif statistics for the project are covered. Also, links for downloading Excel sheets with ORFs, SSRs and primers are available.
SNP Analysis	Includes an overview of the SNPs found in the data as well as links to contigs with SNPs and a search page where SNPs can be selected as a criterion.
Contact	Includes name, email and other information for contacting the creators of the project.
Publication	Includes the publication information.
Search Rosaceae ESTs	Links to a search page for clones of Rosaceae species. This is covered in the next section of the chapter.
Search Rosaceae V3 Contigs	Links to a search page for contigs of Rosaceae species. This is covered in the next section of the chapter.

In house EST libraries and resulting assemblies have similar features to the larger assemblies of public data. They contain more processing information such as quality values, successful clone reports, and plate reports. These projects can be found on the original Data Overview page and are listed under the originating lab and primary investigator. The home page for the project discussed in Chapter 2 can be seen in Figure 4.5.

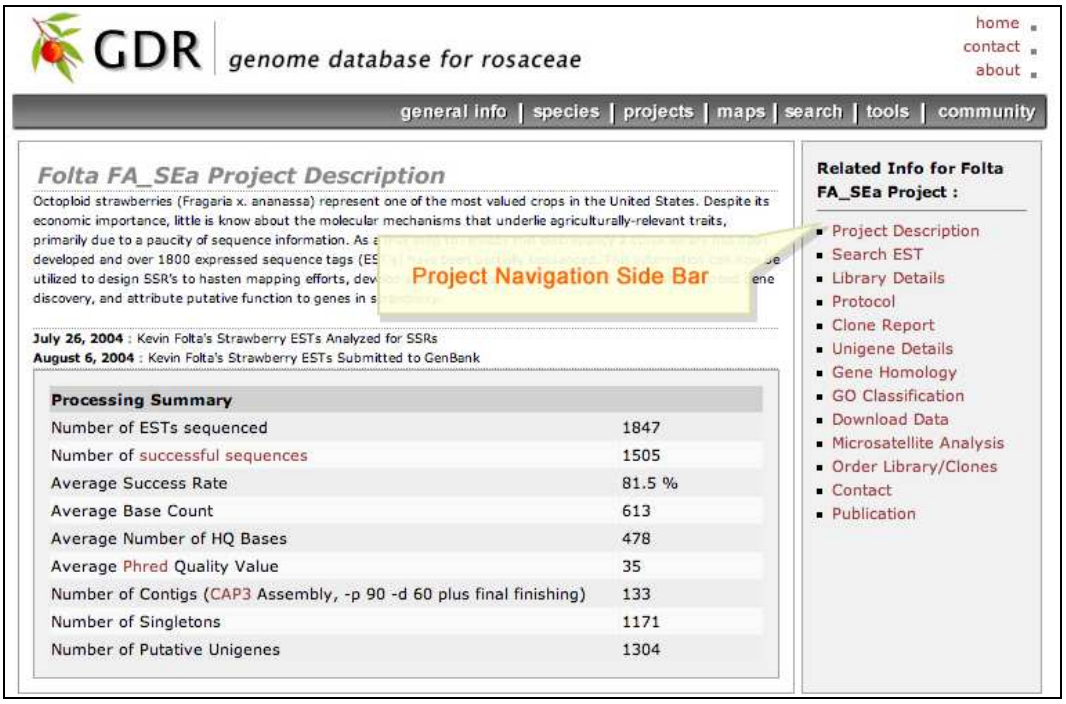


Figure 4.5: Kevin Folta EST Project Home Page

Searching

I have developed extensive search pages to allow maximum customization of results for researchers. The EST and unigene gene search page can be accessed through the "EST" link under the "Search" menu on the top navigation bar. The

initial page allows searching of all Rosaceae clones and unigenes and links to a contig search page (Figure 4.6). Users can enter the name of the clone either as an accession number or the EST name associated with the clone in dbEST. Users can also upload a file of names to be returned. Features may be used to limit searches so that the results returned have SSRs or represent a unigene sequence from a particular unigene assembly. This feature will return the unigene singlets and a representative EST from each contig. To maintain backward compatibility, version 2 unigenes from 2005 are available as well as the most recent version 3. Users can also link directly to a contig search page. The search page also allows users to limit the search to particular genera or species via the drop down box in the "Taxonomy" section. The "Tissue" section also has a drop down box that lists all the tissues assigned from the Plant Ontology terms to the various cDNA libraries. The final "Putative Function" section allows researchers to enter descriptions, source organisms, or GO terms relating to a gene of interest. The search will find genes with SWISS-PROT matches to any of these keywords. All of these search features can be combined in any way to produce highly tailored results for the user.

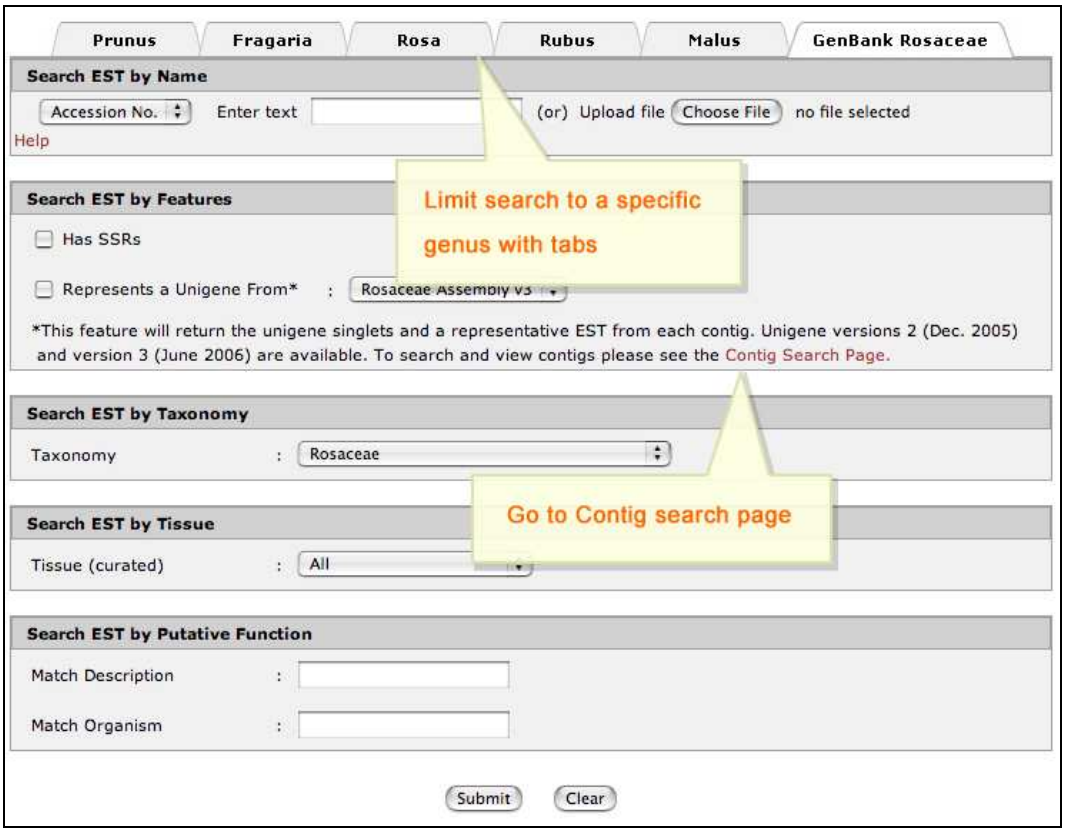


Figure 4.6: Main Rosaceae EST Search Page

The results page lists the ESTs as clickable links to more information such as sequence, library details, reference and contact information, sequence homology, unigene information, SSR and ORF information, and mapping information where available (Figure 4.7). Another link will also use InterProScan via web services to return InterProScan results. Besides clicking on each individual clone, the results can be downloaded as a fasta-formatted file or as a tab-delimited file with SWISS-PROT homology results.

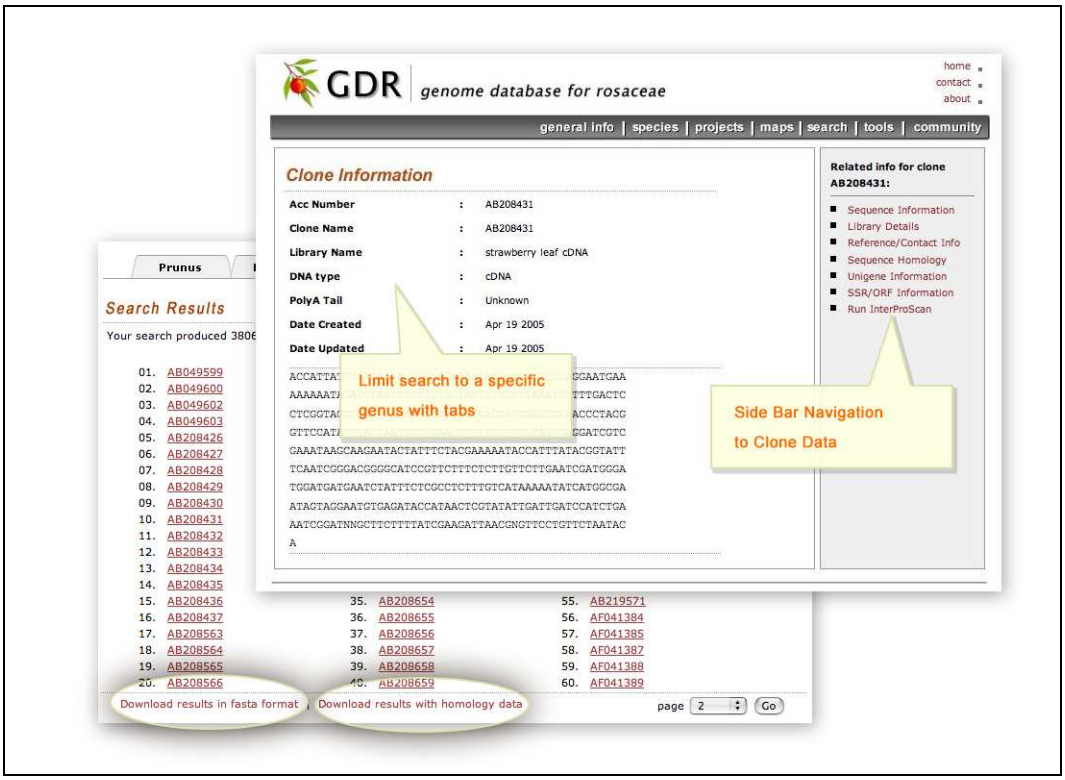


Figure 4.7: Results of EST search

I created a contig search page that provides a similar set of choices for searching with the additional feature of containing a SNP. The contig viewing pages include consensus sequence information, comprising ESTs with library information, sequence homology, SSR and ORF information, and autoSNP output. The InterProScan web services feature is also available.

Conclusions

The GDR fulfills a specific need for an integrated genomics clearing house for the Rosaceae family, which contains numerous species of economic importance. The data includes maps, markers, and publically available sequences such as ESTs. The database and website provide users with access to this data via direct download, querying, or browsing. GDR increases the value of the public

data by adding extensive annotation and curation that is tailored directly to the interests of the Rosaceae researchers. Further online tools and community resources allow researchers to exchange results and ideas in an internationally available forum. The database is an essential part of the community for Rosaceae research and can become ever more important as genomic sequences and expression data becomes available.

References

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215(3):403-410.

Fang Z, Polacco M, Chen S, Schroeder S, Hancock D, Sanchez H and Coe E. 2003. cMap: the comparative genetic map viewer. *Bioinformatics* 19(3):416-7.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R; Gene Ontology Consortium. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 32(Database issue):D258-261.

Horn R, Lecouls AC, Callahan A, Dandekar A, Garay L, McCord P, Howad W, Chan H, Verde I, Main D, Jung S, Georgi L, Forrest S, Mook J, Zhebentyayeva T, Yu Y, Kim HR, Jesudurai C, Sosinski B, Arus P, Baird V, Parfitt D, Reighard G, Scorza R, Tomkins J, Wing R and Abbott AG. 2005. Candidate gene database and transcript map for peach, a model species for fruit trees. *Theor Appl Genet*. 2005 Apr 22; [Epub ahead of print]

Huang X and Madan A. 1999. CAP3: A DNA sequence assembly program. *Genome Res.* 9(9):868-877.

Ilic K, Kellogg EA, Jaiswal P, Zapata F, Stevens PF, Vincent LP, Avraham S, Reiser L, Pujar A, Sachs MM, Whitman NT, McCouch SR, Schaeffer ML, Ware DH, Stein LD and Rhee SY. 2006. Plant Structure Ontology. Unified Vocabulary of Anatomy and Morphology of a Flowering Plant. *Plant Physiol.* Dec 1 [Epub ahead of print].

Jaiswal P, Ni J, Yap I, Ware D, Spooner W, Youens-Clark K, Ren L, Liang C, Zhao W, Ratnapu K, Faga B, Canaran P, Fogleman M, Hebbard C, Avraham S, Schmidt S, Casstevens TM, Buckler ES, Stein L and McCouch S. 2006. Gramene: a bird's eye view of the cereal genomes. Nucleic Acids Res. 34(Database issue):D717-23.

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(1):130.

McEntyre J and Ostell J, eds. 2005. The NCBI Handbook. Bethesday(MD):National Library of Medicine (US), NCBI. Article : GenBank: The Nucleotide Sequence Database by Ilene Mizrachi updated July 27th, 2004

Mueller LA, Solow TH, Taylor N, Skwarecki B, Buels R, Binns J, Lin C, Wright MH, Ahrens R, Wang Y, Herbst EV, Keyder ER, Menda N, Zamir D and Tanksley SD. 2006. The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond. *Plant Physiol.* 138(3):1310-7.

Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*. 1988;85:2444–2448.

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, Miller N, Mueller LA, Mundodi S, Reiser L, Tacklind J, Weems DC, Wu Y, Xu I, Yoo D, Yoon J and Zhang P. 2003. The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Research* 31(1):224

The U.S. Rosaceae Genomics, Genetics, and Breeding Initiative. (March 2006). Retreived August 23, 2006 from http://www.mainlab.clemson.edu/gdr/community/funding/

Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N and Suzek B. 2005. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*. 34(Database issue):D187-91.

CHAPTER 5

DISCUSSION AND CONCLUSIONS

The research presented in this document attempts to explore ways that bioinformatics methods have been used to add value to expressed sequence tags (ESTs) from the Rosaceae family and how this information can be disseminated in a timely, effective and efficient way to the community of genomic researchers. The main methods of interest are assembly, marker mining, function characterization via sequence similarity searching, and oligo development. I have used these methods on datasets of varying sizes and across many species with results that shed light on the genomics of the entire family and plants in general.

The initial genomic investigation into a species that is not viewed as a critical research area by the government is often a small project, especially small EST datasets. With scarce funding researchers may use small sequence samples to target their area of interest. This was true of the Rosaceae family in the years 1998 to 2003 in which 25,209 ESTs were deposited in Genbank. Despite their focused nature small datasets have extensive information that can be mined with comprehensive efforts. The example of the strawberry library discussed in Chapter 2 illustrates this premise. While the initial impetus for creation of the cDNA library was to discover the genes associated with stress, many other types of data were discovered from analysis with bioinformatics tools. Over 290 potential SSR markers were found to spur physical genomics, and functional genomic knowledge was increased with the identification of genes in multiple

pathways such as photoperiodic control of flowering time and photomorphogenesis. The overall data provided an initial glimpse of the octoploid strawberry gene set.

Over the last few years the prices of sequencing have dropped and the funding for the Rosaceae crops has grown. Gene sequence data is now being produced at a rapid rate for many plants including those in the Rosaceae family. Since September of 2003 when this research began, the number of ESTs has increased by over 1400%, from 25,209 to 380,687. This data represents an extensive sampling of genes from diverse species, tissues, development stages and conditions. However, the public sequence database dbEST does not add value to this data through annotation and assembly. For each researcher to download and attempt to assemble and characterize this data would be a major redundancy of effort. Thus GDR plays a central role in annotating this data and presenting the findings online. Researchers for the Rosaceae can easily see what has been done in closely related species, find data that can help accelerate their research or justify a grant and find collaborators.

One of the ultimate outcomes of this research is a software pipeline that can be used to regularly analyze and update new sequences. The creation of family wide and genera unigenes can be performed regularly and the data can be deployed online in a timely manner. This continual updating is a necessary feature of any sequence database as new information becomes available. Sequences previously characterized must also be analyzed continually as protein

databases are growing at an ever-increasing rate. This new protein information can provide new annotation for sequences sampled previously.

A major roadblock to the current EST project is the lack of accuracy in algorithms to assemble gene sequences. The need for assembly programs that are both effective and efficient is growing. CAP3, while the most commonly used software for assembly, has limitations. These are not just due to major sequence artifacts but also to an ineffective balance between Type I and Type II errors, discussed in the first chapter. As an example, *Arabidopsis thaliana* has the most extensively annotated genome for plants, resulting in a confident estimate of actual gene content of about 25,000. PlantGDB uses preclustering along with CAP3 to assembly the *Arabidopsis* ESTs, numbering 808,214 in the current version. This large number should preclude major sequencing gaps. Using a preclustering method and CAP3, the plantGDB assembly contains 150,533 sequences. This is an overestimate of more than 500%. This example illustrates the need for better and more accurate assembly methods.

The advantages of a more sensitive and accurate unigene are numerous. Researchers are interested in not only the number genes in an organism but also the number of genes in a family and accurately distinguishing paralogs and orthologs for evolutionary studies. With two sequenced and annotated plant genomes, another genome (poplar) awaiting annotation and several more underway, information is now becoming available to allow us to begin improving assembly methods. It is now possible to take the sequence data, assemble it, and compare the results to the annotated genome to find where errors occurred and

why. In this way the assembly programs can be modified to be more accurate or new assembly algorithms can be created.

An increase in the amount of information that can be included in an assembly may be necessary to achieve a more powerful result. The ploidy level of a given species can have a large impact on the levels of divergence between gene copies and gene family members. By entering this as a parameter in the software, more accurate results could be obtained. The elucidation of well-conserved genes in plant genomes may also prove an important source of information for better assemblies. Contigs can be seeded with these known protein sequences and motifs to find the correct reading frame and thus increase assembly despite sequencing error. Ultimately, hidden markov models may be the most effective statistical algorithms to use. The ability of Clustal to perform multiple sequence alignments is unparalleled but its algorithm is too computationally intensive to be used on the massive scale required for EST data. However, ever increasing computational power and the availability of grids with many processors will reduce the cost of computational power in the same way the cost of sequencing has been reduced. This will open up the possibility of using more complex statistical algorithms on a larger scale.

The key to providing the best set of bioinformatics tools and data to the Rosaceae community is not only staying up to date with increases in EST sequence data but also anticipating and handling new types of data and new software tools. JGI has announced an 8x sequence of the peach genome and a 4x sequence of the apple genome will be available soon from the

Istituto Agrario San Michele all'Adige. A genomic sequence module is needed for GDR to house and add value to this dataset. A centralized repository linked to the other community resources will be necessary to link the current genetic and physical maps to the genomic sequence. The genomic sequence can be used to refine the Rosaceae unigene much as it has elucidated the genes in *Arabidopsis*. However, the EST data can also be used to improve the annotation of the genomic sequence itself by providing evidence of events such as exons being alternatively spliced.

GDR, as a family specific database, can provide resources that focus on the specific needs of the community. As the resources grow so must the bioinformatics grow to add value to the sequences that is meaningful and useful to researchers. For example, Rosaceae researchers are interested in specific traits for crop improvement such as fruit taste, aroma, softening, and other ripening processes as well as both abiotic and biotic stress responses. One of the goals of GDR is to connect the EST data directly to metabolic pathways through sequence similarity results and then directly to characteristics of interest. This would create a module that would allow researchers to enter at any point (EST, enzyme, trait) and be able to explore the available annotation. The ESTs or proteins could be downloaded by the researcher along with any associated markers and primers. This information can feed directly into many scientific studies of interest including genetic or transcriptome mapping and marker assisted selection.

We have already begun development of trait and metabolic pathway modules. Besides flavor and ripening pathways such as ethylene and sugars,

other pathways or partial pathways should be included such as resistance and cell wall metabolism. The GDR advisory board, representing a cross section of interests across the community, would be a good resource for deciding which genes and pathways would be the most important to highlight. However, some manual curation would be involved for this type of module.

One of the aims of the community of genetics researchers studying the Rosaceae family of plants is to further the functional genetics information available. Microarray technology has grown to be popular for this type of study within many plant species to monitor changes in gene expression patterns for different treatments or conditions. The main problems with microarray research currently include difficulties with normalization and analysis of the data as well as comparing and reproducing results from different microarray platforms. The GDR will be able to provide a repository for the raw data giving researchers the option to analyze and compare data with different techniques and reanalyze old data as new software packages and statistics are developed in this growing area. Access to the Rosetta Resolver Gene Analysis System via GDR provides Rosaceae researchers with a unique opportunity to pertform meta-analyses across species and experimental conditions. This preminent array analysis package automatically links out to sequence and biochemical pathway databases.

Bioinformatics is a necessary computational tool for mining useful information from biological data and solving biological problems. The inherent complexity of biological data requires wielding software and algorithms in a statistically proven way to derive useful information. As the data increases, the

tools used in bioinformatics will need further refinement and new tools will need to be developed as well. The overall challenge is to continue to develop ever more sophisticated analysis tools to fully extract maximum knowledge from the wealth of genomic data that will increasingly be available as sequencing costs significantly reduce. The bioinformatics outlined in this paper have elucidated much useful information about the genomes in the Rosaceae family but further refinement is needed. The introduction of new types of data from the research community will need to be analyzed and connected to the current data to provide a consolidated, integrated view for the researcher.

APPENDICES

Appendix A

Unigenes putatively coding for genes involved in important physiological processes

The evidence for each putative gene assignment is give below. The database from the hit is listed first (NR = NCBI's nr; TA = TAIR's *Arabidopsis* proteins; SP = SWISS-PROT), then the name of the protein, the protein description, the protein organism, and the E value of the match.

FA_Sea0007C05 - B-box, zinc-finger protein CONSTANS

NR	AAC99309.1	CONSTANS-like protein	*Malus x domestica*	2.40E-31
TA	At5g24930.1	zinc finger (B-box type) family protein, similar to CONSTANS-like protein 1 GI:4091804 from (Malus x domestica)	*Arabidopsis thaliana*	7.30E-21
SP	COL4_ARATH	Zinc finger protein CONSTANS-LIKE 4	*Arabidopsis thaliana*	3.60E-20

FA_Sea0016A05 - MADS box protein AGL20/SUPPRESSOR OF CONSTANS

TA	At2g45660.1	MADS-box protein (AGL20)	*Arabidopsis thaliana*	1.90E-15
NR	AAO22989.1	MADS-box transcription factor CDM36	*Chrysanthemum x morifolium*	4.90E-15
SP	AGL19_ARATH	Agamous-like MADS box protein AGL19	*Arabidopsis thaliana*	1.30E-08

FA_Sea0002H08 - VIN3 – Vernalization insensitive 3 protein

SP	VIN3_ARATH	VERNALIZATION-INSENSITIVE protein 3	*Arabidopsis thaliana*	1.30E-34
TA	At5g57380	vernalization insensitive 3 (VIN3)	*Arabidopsis thaliana*	9.80E-34

FA_Sea0004D05 - Disease resistance protein (TIR-NBS-LRR class)

TA	At4g16890.1	disease resistance protein (TIR-NBS-LRR class), putative, domain signature TIR-NBS-LRR exists, suggestive of a disease resistance protein.	*Arabidopsis thaliana*	6.80E-18
NR	AAG48132.1	putative resistance protein	*Glycine max*	2.20E-22
SP	TMVRN_NICGU	TMV resistance protein N	*Nicotiana glutinosa* (Tobacco)	3.10E-21

FA_Sea0006F10 - Enhanced Disease Susceptibility protein EDS5

TA	At2g21340.2	enhanced disease susceptibility protein, putative / salicylic acid induction deficient protein, putative	*Arabidopsis thaliana*	7.10E-58
NR	AAL27003.1	enhanced disease susceptibility 5	*Arabidopsis thaliana*	7.90E-44
SP	EDS5_ARATH	Enhanced disease susceptibility 5	*Arabidopsis thaliana*	6.10E-47

FA_Sea0007F04 - Plant defensin PDF2.2

TA	At2g02100.1	plant defensin-fusion protein, putative (PDF2.2), plant defensin protein family member	*Arabidopsis thaliana*	3.10E-22
NR	CAH58740	Defensin	*Plantago major*	4.10E-18
SP	DEF1_CAPAN	Defensin J1-1 precursor	*Capsicum annuum* (Bell pepper)	1.40E-12

FA_Sea0010B10 - Pathogenesis-related thaumatin (PR5)

TA	At1g19320.1	pathogenesis-related thaumatin family protein, Pathogenesis-related protein 5 precursor (PR-5)	*Arabidopsis thaliana*	1.20E-21
NR	AAO12209.1	thaumatin-like cytokinin-binding protein	*Brassica oleracea*	2.30E-53
SP	TLPH_ARATH	Thaumatin-like protein [Precursor]	*Arabidopsis thaliana*	1.20E-27

FA_Sea0014H12 - Putative thaumatin (PR5)

TA	At1g18250.1	thaumatin, putative, identical to SP|P50699 Thaumatin-like protein precursor	*Arabidopsis thaliana*	7.10E-16
NR	NP_173261.1	putative thaumatin	*Arabidopsis thaliana*	7.20E-20
SP	TLPH_ARATH	Thaumatin-like protein [Precursor]	*Arabidopsis thaliana*	2.40E-21

FA_Sea0015A01 - Harpin-induced protein

TA	At3g11660.1	harpin-induced family protein / HIN1 family protein / harpin-responsive family protein	*Arabidopsis thaliana*	3.80E-26
NR	AAM67015.1	putative harpin-induced protein	*Arabidopsis thaliana*	2.40E-27

FA_Sea0015D01 - NDR1 family protein

TA	At5g11790.1	Ndr family protein, similar to SP|O23969 Pollen specific protein SF21 {Helianthus annuus}; contains Pfam profile PF03096: Ndr family	*Arabidopsis thaliana*	7.00E-69
NR	NDRG1_HUMAN	N-myc downstream regulated gene 1 protein	*Homo sapiens*	2.30E-17

FA_Sea0017F09 - Disease resistance protein (CC-NBS-LRR class)

TA	At5g66910.1	disease resistance protein (CC-NBS-LRR class), putative, domain signature CC-NBS-LRR exists	*Arabidopsis thaliana*	2.10E-23
NR	BAB08633.1	disease resistance protein-like	*Arabidopsis thaliana*	5.40E-29
SP	DRL43_ARATH	Probable disease resistance protein At5g66910	*Arabidopsis thaliana*	3.10E-28

FA_SEa0020H01 - Harpin-induced protein

TA	At2g35980.1	harpin-induced family protein (YLS9) / HIN1 family protein / identical to cDNA YLS9 mRNA for hin1 homolog GI:13122295	*Arabidopsis thaliana*	3.40E-18
NR	BAD22533.1	harpin inducing protein 1-like 9	*Nicotiana tabacum*	2.50E-29

FA_Sea0010F01 - glycosyl hydrolase family 17 p (PR2)

TA	At3g57270.1	glycosyl hydrolase family 17 protein, similar to beta-1,3-glucanase GI:16903144 from (*Prunus persica*)	*Arabidopsis thaliana*	2.60E-12
NR	CAB91554.1	beta 1-3 glucanase	*Vitis vinifera*	5.40E-15
SP	E13A_SOYBN	Glucan endo-1,3-beta-glucosidase protein	*Glycine max* (Soybean)	1.40E-12

FA_Sea0017H06 - Osmotin-like protein (PR5)

TA	At2g28790.1	osmotin-like protein, putative, similar to SP|Q41350 Osmotin-like protein precursor	*Arabidopsis thaliana*	5.00E-13
NR	AAB41124.1	osmotin-like protein	*Lycopersicon esculentum*	1.80E-15
SP	OLP1_LYCES	Osmotin-like protein [Precursor]	*Lycopersicon esculentum*	3.40E-16

FA_SEa0001D03 - Peroxidase PRXR1 (PR9)

TA	At4g21960.1	peroxidase 42 (PER42) (P42) (PRXR1)	*Arabidopsis thaliana*	8.20E-51
NR	CAB79151.1	peroxidase prxr1	*Arabidopsis thaliana*	2.50E-53
SP	PER42-ARATH	Peroxidase 42 [Precursor]	*Arabidopsis thaliana*	8.90E-54

FA_SEa0019D07 - Bet v 1 (PR10)

TA	At1g24020.1	Bet v I allergen family protein, contains Pfam profile PF00407: Pathogenesis-related protein Bet v I family	*Arabidopsis thaliana*	2.10E-35
NR	AAM65899.1	pollen allergen-like protein	*Arabidopsis thaliana*	2.30E-39

FA_SEa0012C06 - Lipid transfer protein LPT4 (PR14)

TA	At5g59310.1	lipid transfer protein 4 (LTP4), identical to lipid transfer protein 4 from Arabidopsis thaliana	*Arabidopsis thaliana*	1.70E-19
NR	CAA65477.1	lipid transfer protein	*Prunus dulcis*	1.90E-25
SP	NLTP3_PRUDU	Nonspecific lipid-transfer protein 3 [Precursor]	*Prunus dulcis*	1.40E-27

FA_Sea0004E09 - B-zip transcription factor HY5

TA	At5g11260.1	bZIP protein HY5 (HY5), identical to HY5 protein GI:2251085	*Arabidopsis thaliana*	4.60E-37
NR	BAC20320.1	bZIP with a Ring-finger motif	*Lotus corniculatus var. japonicus*	2.00E-31

FA_Sea0001C09 - NON-PHOTOTROPIC HYPOCOTYL 3

TA	At1g67900.2	phototropic-responsive NPH3 family protein, contains NPH3 family domain, Pfam:PF03000	*Arabidopsis thaliana*	3.20E-31
NR	AAP68226.1	At1g67900	*Arabidopsis thaliana*	9.00E-28

FA_SEa0006H04 - Far-red impaired / FAR1

TA	At4g12850.1	far-red impaired responsive family protein; contains Pfam profile PF03101: FAR1 family	*Arabidopsis thaliana*	3.30E-29
NR	AAS88777.1	At4g12850	*Arabidopsis thaliana*	9.30E-27

Appendix B

Putative Unique Rosaceae Unigenes

We analyzed a dataset of rosaceae unigenes without matches to twelve sets of plantGDB transcripts. This set, putatively representing Rosaceae-specific transcripts, numbers 24181. We examined this set for significant matches to the SWISS-PROT and TrEMBL databases to attempt to infer homology. BLAST with a cut-off of E<1e-9 results were considered. Only 862 unigenes had matches and 279 match other plants. Most of the others were from bacteria or viruses that were presumably missed in quality filtering. The rest showed very specific categories of genes including those shown in the following table. Many of these categories are known to be faster evolving than other sets of genes.

ALLERGENS				
Unigene Name	SWISS-PROT Match	SWISS-PROT Description	Organism	E Value
Contig7345	NLTP_ MALDO	Nonspecific lipid-transfer protein precursor (LTP) (Allergen Mal d 3)	*Malus domestica*	1.00E-20
Malus_ AT000341	Q5J026_ MALDO	Lipid transfer protein precursor (Major allergen and lipid transfer protein Mal d 3)	*Malus domestica*	4.00E-15
Malus_ AT000352	Q5J026_ MALDO	Lipid transfer protein precursor (Major allergen and lipid transfer protein Mal d 3)	*Malus domestica*	1.00E-14
Malus_ CV082042	Q5VJR1_ MALDO	Mal d 1-like (Major allergen Mal d 1.03E)	*Malus domestica*	3.00E-11
Malus_ CV997766	MAL11_ MALDO	Major allergen Mal d 1 (Mal d I)	*Malus domestica*	4.00E-14
Prunus_ DY635989	Q2I6V8_ PRUPE	Major allergen Pru p 1	*Prunus persica*	3.00E-11

DNA BINDING				
Unigene Name	SWISS-PROT Match	SWISS-PROT Description	Organism	E Value
Malus_ Contig13125	NAC67_ORYSA	NAC domain-containing protein 67 (ONAC067)	*Oryza sativa*	8.00E-11
Prunus_ DW343405	Q6K777_ ORYSA	Putative DNA polymerase epsilon catalytic subunit protein isoform b	*Oryza sativa*	2.00E-21
Prunus_ DW343875	Q1RZ64_ MEDTR	DNA-directed DNA polymerase B	*Medicago truncatula*	4.00E-13
Prunus_ DW346806	Q8HD74_ BRANA	Orf6 protein	*Brassica napus*	9.00E-12
Prunus_ DY654082	Q9LM82_ ARATH	F2D10.21	*Arabidopsis thaliana*	1.00E-12
Contig10503	Q6Z4I5_ ORYSA	DNA binding protein-like	*Oryza sativa*	4.00E-09
Contig2568	Q7XU13_ ORYSA	OSJNBa0091D06.8 protein	*Oryza sativa*	1.00E-08
Contig7085	NAC61_ ARATH	Putative NAC domain-containing protein 61 (ANAC061)	*Arabidopsis thaliana*	6.00E-08
Contig7460	Q7XU13_ ORYSA	OSJNBa0091D06.8 protein	*Oryza sativa*	1.00E-07
Malus_ CN872432	Q7XU13_ ORYSA	OSJNBa0091D06.8 protein	*Oryza sativa*	2.00E-13
Malus_ CN909378	Q1SEE1_ MEDTR	Integrase, catalytic region; Zinc finger, CCHC-type; Peptidase aspartic, catalytic	*Medicago truncatula*	2.00E-14
Malus_ CN922172	Q4ABN6_ BRARP	01P13-1	*Brassica rapa* subsp. pekinensis	2.00E-15
Malus_ CN924648	Q1SJP8_ MEDTR	D-galactoside/L-rhamnose binding SUEL lectin; Integrase, catalytic region; Galactose-binding like; Peptidase aspartic, catalytic	*Medicago truncatula*	9.00E-10
Malus_ CN948632	Q4ABX5_ BRARP	4D11_12	*Brassica rapa subsp. pekinensis*	2.00E-10
Malus_ CO903900	Q5DW96_ PRUYE	Plastid DNA-binding protein (Fragment)	*Prunus yedoensis*	7.00E-55
Malus_ EB125280	Q2HVX8_ MEDTR	Helix-loop-helix DNA-binding	*Medicago truncatula*	2.00E-12
Malus_ EB147616	Q1SP86_ MEDTR	Helix-loop-helix DNA-binding	*Medicago truncatula*	2.00E-10
Malus_ CN870361	KNAP2_ MALDO	Homeobox protein knotted-1-like 2 (KNAP2)	*Malus domestica*	3.00E-11

NUCLEIC ACID BINDING				
Unigene Name	SWISS-PROT Match	SWISS-PROT Description	Organism	E Value
Contig9890	Q9LTU1_ ARATH	Replication protein A1-like	*Arabidopsis thaliana*	4.00E-09
Malus_ CN911169	Q1RZD5_ MEDTR	Zinc finger, C2H2-type	*Medicago truncatula*	6.00E-21
Malus_ CN928149	Q9LIR7_ ARATH	Arabidopsis thaliana genomic DNA, chromosome 3, BAC clone:F14O13	*Arabidopsis thaliana*	1.00E-11
Malus_ CN941602	Q1S3P4_ MEDTR	Zinc finger, CCCH-type	*Medicago truncatula*	5.00E-11
Malus_ CN869844	Q9LN78_ ARATH	T12C24.22	*Arabidopsis thaliana*	5.00E-10
Malus_ CN889149	Q9FNQ1_ ARATH	RNA helicase	*Arabidopsis thaliana*	5.00E-59
Prunus_ BU043054	Q9LQE5_ ARATH	F15O4.40	*Arabidopsis thaliana*	5.00E-14

RESISTANCE				
Unigene Name	SWISS-PROT Match	SWISS-PROT Description	Organism	E Value
Contig10855	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	3.00E-11
Contig12209	Q1SGR7_ MEDTR	TIR; Disease resistance protein; AAA ATPase	*Medicago truncatula*	1.00E-07
Contig417	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	2.00E-57
Contig7471	Q6URA1_ 9ROSA	Putative TIR-NBS type R protein 4	Rosales	8.00E-11
Contig7540	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	4.00E-23
Contig7631	Q9SHI3_ ARATH	Similar to disease resistance proteins	*Arabidopsis thaliana*	1.00E-08
Contig8995	Q19PN0_ POPTR	TIR-NBS-LRR-TIR type disease resistance protein (Fragment)	*Populus trichocarpa*	6.00E-12
Fragaria_ DY668873	Q9FKE5_ ARATH	Disease resistance protein RPS4	*Arabidopsis thaliana*	5.00E-12
Malus_ CN491689	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	4.00E-11
Malus_ CN493446	Q6QT45_ QUESU	Resistance protein (Fragment)	*Quercus suber*	2.00E-15
Malus_ CN495384	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	7.00E-52

RESISTANCE (Continued)				
Unigene Name	SWISS-PROT Match	SWISS-PROT Description	Organism	E Value
Malus_ CN909990	Q69L15_ ORYSA	Putative Avr9/Cf-9 rapidly elicited protein 141	*Oryza sativa*	1.00E-12
Malus_ CN918149	Q2L361_ MALDO	Putative CC-NBS-LRR resistance protein	*Malus domestica*	3.00E-16
Malus_ CN996310	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	4.00E-30
Malus_ CN996566	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	6.00E-17
Malus_ CO754758	Q2L359_ MALDO	Putative CC-NBS-LRR resistance protein	*Malus domestica*	2.00E-29
Malus_ CO867486	Q2L361_ MALDO	Putative CC-NBS-LRR resistance protein	*Malus domestica*	5.00E-21
Malus_ Contig12623	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	4.00E-41
Malus_ Contig14559	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	1.00E-61
Malus_ Contig14645	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	2.00E-10
Malus_ Contig15163	Q2L360_ MALDO	Putative CC-NBS-LRR resistance protein	*Malus domestica*	5.00E-23
Malus_ Contig18684	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	8.00E-11
Malus_ Contig23745	Q2L361_ MALDO	Putative CC-NBS-LRR resistance protein	*Malus domestica*	4.00E-18
Malus_ Contig3333	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	2.00E-16
Malus_ CV883412	RGA4_ SOLBU	Putative disease resistance protein RGA4 (RGA4-blb)	*Solanum bulbocastanum*	6.00E-13
Malus_ DR033890	Q6UJ68_ MALDO	NBS-LRR resistance gene-like protein ARGH17 (Fragment)	*Malus domestica*	5.00E-16
Malus_ DR991235	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	6.00E-22
Malus_ DT040468	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	7.00E-33
Malus_ DT042200	RGA2_ SOLBU	Disease resistance protein RGA2 (RGA2-blb) (Blight resistance protein RPI)	*Solanum bulbocastanum*	9.00E-10
Malus_ DT043243	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	3.00E-61
Malus_ DY255684	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	4.00E-34
Malus_ EB110576	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	2.00E-15

RESISTANCE (Continued)				
Unigene Name	SWISS-PROT Match	SWISS-PROT Description	Organism	E Value
Malus_ EB114350	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	2.00E-15
Malus_ EB151511	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	3.00E-29
Malus_ EB151787	Q2HUU3_ MEDTR	Disease resistance protein; Calcium-binding EF-hand; AAA ATPase	*Medicago truncatula*	5.00E-17
Prunus_ AJ823882	Q6URA2_ 9ROSA	TIR-NBS-LRR type R protein 7	Rosales	3.00E-21
Prunus_ DY646807	MRP9_ ARATH	Multidrug resistance-associated protein 9 (EC 3.6.3.44) (Glutathione S-conjugate transporting ATPase 9) (ATP-energized glutathione S-conjugate pump 9)	*Arabidopsis thaliana*	8.00E-22
Contig2989	CYTM_ SOLTU	Multicystatin (MC)	*Solanum tuberosum*	2.00E-08

RIPENING				
Unigene Name	SWISS-PROT Match	SWISS-PROT Description	Organism	E Value
Malus_ CO414873	Q9FHX3_ ARATH	Receptor lectin kinase-like protein (Lectin protein kinase family protein)	*Arabidopsis thaliana*	4.00E-18
Malus_ CO415575	Q45TX6_ MALDO	Starch branching enzyme I	*Malus domestica*	3.00E-30
Malus_ DT040152	Q84L65_ PYRCO	Xyloglucan endotransglycosylase	*Pyrus communis*	6.00E-14
Malus_ EB114578	Q5J3N9_ MALDO	Sucrose phosphate phosphatase (EC 3.1.3.24)	*Malus domestica*	8.00E-16
Malus_ EB146797	Q9FHX3_ ARATH	Receptor lectin kinase-like protein (Lectin protein kinase family protein)	*Arabidopsis thaliana*	7.00E-24
Malus_ CN934037	LE14B_ PRUAR	LEC14B homolog	*Prunus armeniaca*	5.00E-17
Contig10296	Q68UW1_ PYRCO	Polygalacturonase	*Pyrus communis*	6.00E-07
Contig8761	Q6YYW5_ ORYSA	Putative expansin 11	*Oryza sativa*	2.00E-08
Malus_ CN944300	Q1W5D1_ HEVBR	Solanesyl diphosphate synthase	*Hevea brasiliensis*	1.00E-10
Malus_ CV082424	O48629_ PRUAR	Putative auxin-repressed protein	*Prunus armeniaca*	2.00E-10
Malus_ DT002539	Q8LSK7_ 9ROSI	Auxin-regulated protein	rosids	6.00E-13

RIPENING (Continued)				
Unigene Name	SWISS-PROT Match	SWISS-PROT Description	Organism	E Value
Malus_ EB116462	Q93WZ6_ PRUPE	Abscisic stress ripening-like protein	*Prunus persica*	1.00E-14
Malus_ EB117867	Q93WZ6_ PRUPE	Abscisic stress ripening-like protein	*Prunus persica*	1.00E-11
Malus_ EB120299	Q5S004_ CUCSA	Ethylene response factor 3	*Cucumis sativus*	5.00E-11
Prunus_ DN676698	O48629_ PRUAR	Putative auxin-repressed protein	*Prunus armeniaca*	2.00E-13
Prunus_ DT454892	Q93WZ6_P RUPE	Abscisic stress ripening-like protein	*Prunus persica*	5.00E-25
Prunus_ DT454987	Q93WZ6_ PRUPE	Abscisic stress ripening-like protein	*Prunus persica*	2.00E-43
Prunus_ DY635974	Q93WZ6_ PRUPE	Abscisic stress ripening-like protein	*Prunus persica*	5.00E-16
Prunus_ DY652441	Q2Z1Y3_ PRUMU	Expansin	*Prunus mume*	4.00E-11
Prunus_ DY652857	O50000_ PRUAR	Abscisic stress ripening protein homolog	*Prunus armeniaca*	1.00E-11
Prunus_ DY653665	Q93WZ6_ PRUPE	Abscisic stress ripening-like protein	*Prunus persica*	9.00E-28
Fragaria_ Contig381	Q9FVF1_ FRAAN	Alcohol acyltransferase	*Fragaria ananassa*	1.00E-17

SELF-INCOMPATIBILITY				
Unigene Name	SWISS-PROT Match	SWISS-PROT Description	Organism	E Value
Malus_ CO417763	Q9XIR5_ ARATH	Similar to translational activator	*Arabidopsis thaliana*	7.00E-15
Malus_ CO417854	Q3EA10_ ARATH	Protein At4g16195	*Arabidopsis thaliana*	2.00E-17
Fragaria_ DV440585	Q852Q3_ PRUMU	S7-RNase	*Prunus mume*	2.00E-19
Malus_ CN994087	Q9MB59_ MALDO	Se-RNase	*Malus domestica*	9.00E-28
Prunus_ AJ873095	Q84KJ9_ PRUDU	S locus F-box protein c	*Prunus dulcis*	1.00E-120

STRESS RESPONSE				
Unigene Name	SWISS-PROT Match	SWISS-PROT Description	Organism	E Value
Malus_ CN869627	Q9LUX4_ PYRPY	Glycine rich protein (Fragment)	*Pyrus pyrifolia*	9.00E-10
Prunus_ DY646882	Q9ZW93_ ARATH	F5A8.5 protein	*Arabidopsis thaliana*	1.00E-18
Malus_ CV883152	Q9SW89_ PRUDU	Abscisic acid response protein	*Prunus dulcis*	5.00E-14
Malus_ CN861106	Q9SW89_ PRUDU	Abscisic acid response protein	*Prunus dulcis*	1.00E-17
Fragaria_ CO817582	Q40968_ PRUPE	Dehydrin	*Prunus persica*	5.00E-11
Fragaria_ DY669955	Q8W3I7_ ORYSA	Putative DnaJ domain containg protein, 3'-partial (Fragment)	*Oryza sativa*	3.00E-23
Malus_ CN496472	O04648_ ARATH	A_TM021B04.9 protein	*Arabidopsis thaliana*	2.00E-20
Malus_ CN863502	Q5QIC0_ PRUPE	Dehydrin 2	*Prunus persica*	9.00E-17
Malus_ CN869457	Q30E95_ PRUPE	Type II SK2 dehydrin (Fragment)	*Prunus persica*	1.00E-10
Malus_ CN903463	Q40968_ PRUPE	Dehydrin	*Prunus persica*	2.00E-13
Malus_ CN914194	Q1SIX7_ MEDTR	Forkhead-associated; Tyrosyl-DNA phosphodiesterase	*Medicago truncatula*	3.00E-14
Malus_ CN921728	Q1SJW3_ MEDTR	UspA	*Medicago truncatula*	2.00E-17
Malus_ CN940093	MSH7_ ARATH	DNA mismatch repair protein MSH6-2 (AtMsh6-2) (MutS homolog 7)	*Arabidopsis thaliana*	3.00E-10
Malus_ CO753477	Q30E95_ PRUPE	Type II SK2 dehydrin (Fragment)	*Prunus persica*	9.00E-14
Malus_ Contig3796	Q1T0V8_ MEDTR	Heat shock protein Hsp20	*Medicago truncatula*	6.00E-10
Malus_ DT043057	Q84UH1_ PRUPE	Defensin protein 1	*Prunus persica*	7.00E-12
Malus_ EB115083	Q30E95_ PRUPE	Type II SK2 dehydrin (Fragment)	*Prunus persica*	4.00E-16
Malus_ EB119260	Q30E95_ PRUPE	Type II SK2 dehydrin (Fragment)	*Prunus persica*	4.00E-11
Malus_ EB145819	Q40968_ PRUPE	Dehydrin	*Prunus persica*	1.00E-13
Prunus_ AJ631390	Q30E95_ PRUPE	Type II SK2 dehydrin (Fragment)	*Prunus persica*	3.00E-21
Prunus_ AJ631394	Q30E95_ PRUPE	Type II SK2 dehydrin (Fragment)	*Prunus persica*	2.00E-15

STRESS RESPONSE (Continued)				
Unigene Name	SWISS-PROT Match	SWISS-PROT Description	Organism	E Value
Prunus_ DY636011	Q30E95_ PRUPE	Type II SK2 dehydrin (Fragment)	*Prunus persica*	1.00E-10
Prunus_ DY646022	Q30E95_ PRUPE	Type II SK2 dehydrin (Fragment)	*Prunus persica*	4.00E-23
Prunus_ DY646050	Q30E95_ PRUPE	Type II SK2 dehydrin (Fragment)	*Prunus persica*	2.00E-13
Prunus_ DY647472	Q30E95_ PRUPE	Type II SK2 dehydrin (Fragment)	*Prunus persica*	1.00E-20
Prunus_ DY652978	Q30E95_ PRUPE	Type II SK2 dehydrin (Fragment)	*Prunus persica*	5.00E-11

TRANSCRIPTION FACTORS				
Unigene Name	SWISS-PROT Match	SWISS-PROT Description	Organism	E Value
Contig10829	Q7X9I6_ MALDO	MADS box protein	*Malus domestica*	4.00E-20
Malus_ CN494770	Q9M7S0_ MALDO	Homeodomain protein	*Malus domestica*	1.00E-17
Malus_ Contig3003	KNAP1_ MALDO	Homeobox protein knotted-1-like 1 (KNAP1)	*Malus domestica*	6.00E-59
Malus_ EB128043	Q9FQ01_ 9ROSI	Basic leucine zipper transcription factor	rosids	4.00E-11
Malus_ EB136357	KNAP2_ MALDO	Homeobox protein knotted-1-like 2 (KNAP2)	*Malus domestica*	4.00E-22
Prunus_ Contig1282	Q9XH73_ PRUAR	Homeobox leucine zipper protein	*Prunus armeniaca*	1.00E-17
Prunus_ DY635995	O81365_ PRUAR	AP2 domain containing protein (Fragment)	*Prunus armeniaca*	2.00E-14
Prunus_ DY653634	O81365_ PRUAR	AP2 domain containing protein (Fragment)	*Prunus armeniaca*	2.00E-16
Contig12628	Q84WX1_ BRANA	BHLH transcription factor	*Brassica napus*	5.00E-11
Contig12777	Q2LMF1_ MALDO	MYB6	*Malus domestica*	8.00E-15
Fragaria_ Contig1577	Q6RF31_ 9ROSI	MADS box transcription factor	rosids	2.00E-10
Fragaria_ DY670036	Q9LD95_ ARATH	Sigma factor-like protein (SigF) (Putative RNA polymerase sigma-70 factor protein)	*Arabidopsis thaliana*	9.00E-11
Malus_ CN861043	Q8VWW8_ MALDO	Transcription factor AHAP2	*Malus domestica*	4.00E-23
Malus_ CO865898	Q2LME2_ MALDO	MYB22	*Malus domestica*	3.00E-14

TRANSCRIPTION FACTORS (Continued)				
Unigene Name	SWISS-PROT Match	SWISS-PROT Description	Organism	E Value
Malus_ CO867282	Q2LME2_ MALDO	MYB22	*Malus domestica*	2.00E-35
Malus_ CO903858	Q2LME9_ MALDO	MYB11	*Malus domestica*	4.00E-20
Malus_ Contig16722	Q1SG54_ MEDTR	GRAS transcription factor	*Medicago truncatula*	2.00E-13
Malus_ DR998738	Q2LME9_ MALDO	MYB11	*Malus domestica*	1.00E-56
Malus_ EB139466	Q2LMD8_ MALDO	MYB92	*Malus domestica*	5.00E-14
Malus_ EB154218	Q6RF31_ 9ROSI	MADS box transcription factor	rosids	2.00E-11
Malus_ EB175541	Q2LMF0_ MALDO	MYB7	*Malus domestica*	2.00E-10
Prunus_ CV051106	Q9SR27_ ARATH	Putative transcription factor	*Arabidopsis thaliana*	2.00E-16

OTHER TRANSCRIPTION/TRANSLATION REGULATION				
Unigene Name	SWISS-PROT Match	SWISS-PROT Description	Organism	E Value
Prunus_ Contig3020	Q94BW3_ CINCA	Type 2 ribosome-inactivating protein cinnamomin III precursor	*Cinnamomum camphora*	8.00E-32
Prunus_ DW343155	Q1RZJ4_ MEDTR	Aldo/keto reductase; Sigma-54 factor, interaction region	*Medicago truncatula*	1.00E-106
Malus_ CV880602	Q9FH01_ ARATH	Similarity to CHP-rich zinc finger protein	*Arabidopsis thaliana*	5.00E-15
Malus_ CO052477	Q6YNS0_ PRUAV	Putative translation-initiation factor 3 subunit	*Prunus avium*	4.00E-25
Prunus_ DY640722	IF2C_ PHAVU	Translation initiation factor IF-2, chloroplast precursor (PvIF2cp)	*Phaseolus vulgaris*	6.00E-16

TRANSPOSABLE ELEMENT RELATED				
Unigene Name	SWISS-PROT Match	SWISS-PROT Description	Organism	E Value
Malus_ Contig17295	O81630_ ARATH	F8M12.22 protein (RNA-directed DNA polymerase activity)	*Arabidopsis thaliana*	5.00E-10
Malus_ Contig17477	Q1S9K1_ MEDTR	Integrase, catalytic region; Zinc finger, CCHC-type; Peptidase aspartic, catalytic	*Medicago truncatula*	7.00E-10
Prunus_ AJ872476	Q1SCY9_ MEDTR	Integrase, catalytic region; Zinc finger, CCHC-type	*Medicago truncatula*	1.00E-12

TRANSPOSABLE ELEMENT RELATED (Continued)				
Unigene Name	SWISS-PROT Match	SWISS-PROT Description	Organism	E Value
Prunus_ AJ873120	Q1SEE1_ MEDTR	Integrase, catalytic region; Zinc finger, CCHC-type; Peptidase aspartic, catalytic	*Medicago truncatula*	2.00E-11
Prunus_ Contig4976	Q9SA04_ ARATH	F28K20.4 protein (RNA-directed DNA polymerase activity)	*Arabidopsis thaliana*	3.00E-10
Prunus_ DW343206	Q5GIS9_ CUCME	Ulp1-like peptidase (Ulp1 peptidase-like)	*Cucumis melo*	1.00E-24
Prunus_ DW344028	Q1SD84_ MEDTR	Integrase, catalytic region	*Medicago truncatula*	5.00E-13
Prunus_ DW344244	Q6QZP0_ DAUCA	DNA-directed RNA polymerase	*Daucus carota*	3.00E-20
Prunus_ DW344563	Q5GIS9_ CUCME	Ulp1-like peptidase (Ulp1 peptidase-like)	*Cucumis melo*	5.00E-12
Prunus_ DW345563	Q9XG91_ PHAVU	Tpv2-1c protein (Fragment)	*Phaseolus vulgaris*	2.00E-10
Prunus_ DW346268	Q8H6Q8_ PONTR	CTV.20	*Poncirus trifoliata*	5.00E-14
Prunus_ DW347461	Q1SCY9_ MEDTR	Integrase, catalytic region; Zinc finger, CCHC-type	*Medicago truncatula*	5.00E-11
Contig4317	Q1SS89_ MEDTR	Integrase, catalytic region	*Medicago truncatula*	1.00E-07
Contig4323	Q1S3K7_ MEDTR	RNA-directed DNA polymerase (Reverse transcriptase); Zinc finger, CCHC-type; Endonuclease/exonuclease/phosphatase	*Medicago truncatula*	1.00E-06
Fragaria_ DV439596	Q1S3K7_ MEDTR	RNA-directed DNA polymerase (Reverse transcriptase); Zinc finger, CCHC-type; Endonuclease/exonuclease/phosphatase	*Medicago truncatula*	2.00E-13
Fragaria_ DY670243	Q1S3K7_ MEDTR	RNA-directed DNA polymerase (Reverse transcriptase); Zinc finger, CCHC-type; Endonuclease/exonuclease/phosphatase	*Medicago truncatula*	6.00E-13
Malus_ CN859348	Q1S8I3_ MEDTR	RNA-directed DNA polymerase (Reverse transcriptase); Expansin/Lol pI	*Medicago truncatula*	2.00E-11
Malus_ CN860026	Q75L75_ ORYSA	Putative reverse transcriptase	*Oryza sativa*	1.00E-13
Malus_ CN868023	Q2AA00_ ASPOF	Reverse transcriptase family protein	*Asparagus officinalis*	1.00E-10
Malus_ CN924484	Q1SKR2_ MEDTR	Integrase, catalytic region; Zinc finger, CCHC-type	*Medicago truncatula*	4.00E-23
Malus_ CN924937	Q204I7_ MALDO	Reverse transcriptase (Fragment)	*Malus domestica*	3.00E-11

TRANSPOSABLE ELEMENT RELATED (Continued)				
Unigene Name	SWISS-PROT Match	SWISS-PROT Description	Organism	E Value
Malus_ CN924979	Q204I7_ MALDO	Reverse transcriptase (Fragment)	*Malus domestica*	9.00E-10
Malus_ CO905370	Q1SEE1_ MEDTR	Integrase, catalytic region; Zinc finger, CCHC-type; Peptidase aspartic, catalytic	*Medicago truncatula*	2.00E-19
Malus_ CV656145	Q1SCY9_ MEDTR	Integrase, catalytic region; Zinc finger, CCHC-type	*Medicago truncatula*	3.00E-13
Prunus_ AJ873519	Q6L975_ VITVI	GAG-POL	*Vitis vinifera*	2.00E-18
Prunus_ DW341050	Q1S8I3_ MEDTR	RNA-directed DNA polymerase (Reverse transcriptase); Expansin/Lol pI	*Medicago truncatula*	8.00E-16
Prunus_ DW345168	Q1SS87_ MEDTR	Gag-pol polyprotein-related	*Medicago truncatula*	4.00E-11
Contig2543	Q9ZS84_ LYCES	Polyprotein	*Lycopersicon esculentum*	5.00E-08
Malus_ Contig8242	Q1T2D5_ MEDTR	Chromo	*Medicago truncatula*	2.00E-14
Contig11260	Q949J4_ LYCES	Putative copia-like polyprotein	*Lycopersicon esculentum*	9.00E-07
Contig11350	Q9ZQK0_ ARATH	Putative retroelement pol polyprotein	*Arabidopsis thaliana*	3.00E-11
Contig11628	Q2R6F2_ ORYSA	Retrotransposon protein, putative, unclassified	*Oryza sativa*	8.00E-07
Contig1434	Q2AA50_ ASPOF	Retrotransposon gag protein	*Asparagus officinalis*	5.00E-09
Contig8743	Q1SJ04_ MEDTR	RNA-directed DNA polymerase (Reverse transcriptase); Integrase, catalytic region; Ribonuclease H; Retrotransposon gag protein; Retrovirus capsid, C-terminal; Peptidase aspartic, catalytic	*Medicago truncatula*	6.00E-12
Contig8909	Q2QS85_ ORYSA	Retrotransposon protein, putative, Ty3-gypsy subclass	*Oryza sativa*	9.00E-08
Fragaria_ Contig1244	Q9LH75_ ARATH	Ac transposase-like protein (Hypothetical protein At3g14800)	*Arabidopsis thaliana*	1.00E-09
Fragaria_ DY670637	Q1S8I5_ MEDTR	Probable Ta11-like non-LTR retroelement protein [imported]- Arabidopsis thaliana	*Medicago truncatula*	2.00E-11
Malus_ CN492003	Q9XE43_ ARATH	Putative non-LTR retrolelement reverse transcriptase	*Arabidopsis thaliana*	5.00E-11
Malus_ CN854949	Q53NY9_ ORYSA	Retrotransposon protein, putative, Ty3-gypsy sub-class	*Oryza sativa*	6.00E-11
Malus_ CN856635	Q60DB1_ ORYSA	Retrotransposon protein, putative, unclassified	*Oryza sativa*	7.00E-13

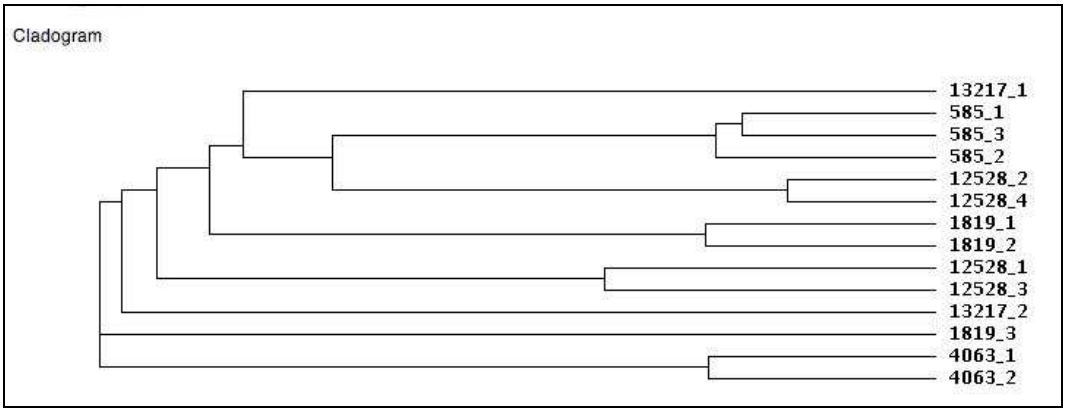
TRANSPOSABLE ELEMENT RELATED (Continued)				
Unigene Name	SWISS-PROT Match	SWISS-PROT Description	Organism	E Value
Malus_ CN858499	Q5GIT1_ CUCME	MuDRA-like transposase (MuDRA transposase-like)	*Cucumis melo*	1.00E-14
Malus_ CN860215	Q1RWL1_ MEDTR	Zinc finger, CCHC-type; Retrotransposon gag protein; Polynucleotidyl transferase, Ribonuclease H fold	*Medicago truncatula*	2.00E-18
Malus_ CN867476	Q2R4N2_ ORYSA	Retrotransposon protein, putative, unclassified	*Oryza sativa*	8.00E-12
Malus_ CN878460	O22148_ ARATH	Putative non-LTR retroelement reverse transcriptase	*Arabidopsis thaliana*	2.00E-10
Malus_ CN900895	Q2QVF5_ ORYSA	Transposon protein, putative, mariner sub-class	*Oryza sativa*	9.00E-11
Malus_ CN921366	Q949J4_ LYCES	Putative copia-like polyprotein	*Lycopersicon esculentum*	6.00E-17
Malus_ CN924951	Q2QWF9_ ORYSA	Retrotransposon protein, putative, unclassified	*Oryza sativa*	7.00E-11
Malus_ CN932984	POLX_ TOBAC	Retrovirus-related Pol polyprotein from transposon TNT 1-94 [Includes: Protease (EC 3.4.23.-); Reverse transcriptase (EC 2.7.7.49); Endonuclease]	*Nicotiana tabacum*	5.00E-13
Malus_ Contig17460	Q1SJ04_ MEDTR	RNA-directed DNA polymerase (Reverse transcriptase); Integrase, catalytic region; Ribonuclease H; Retrotransposon gag protein; Retrovirus capsid, C-terminal; Peptidase aspartic, catalytic	*Medicago truncatula*	4.00E-11
Malus_ Contig18758	Q9FWZ5_ ARATH	Putative retroelement polyprotein	*Arabidopsis thaliana*	9.00E-24
Malus_ Contig8348	Q6JJ56_ IPOTF	Putative copia-like polyprotein	*Ipomoea trifida*	6.00E-10
Malus_ CV794276	Q2AA50_ ASPOF	Retrotransposon gag protein	*Asparagus officinalis*	2.00E-13
Prunus_ DW342097	Q1SJ04_ MEDTR	RNA-directed DNA polymerase (Reverse transcriptase); Integrase, catalytic region; Ribonuclease H; Retrotransposon gag protein; Retrovirus capsid, C-terminal; Peptidase aspartic, catalytic	*Medicago truncatula*	1.00E-16
Prunus_ DW342250	Q9SJP0_ ARATH	Putative retroelement pol polyprotein	*Arabidopsis thaliana*	5.00E-14
Prunus_ DW346992	Q2QQV8_ ORYSA	Retrotransposon protein, putative, unclassified	*Oryza sativa*	7.00E-10

Appendix C

Multiple Sequence Alignments of ESTs in Rosaceae Unigene Contigs

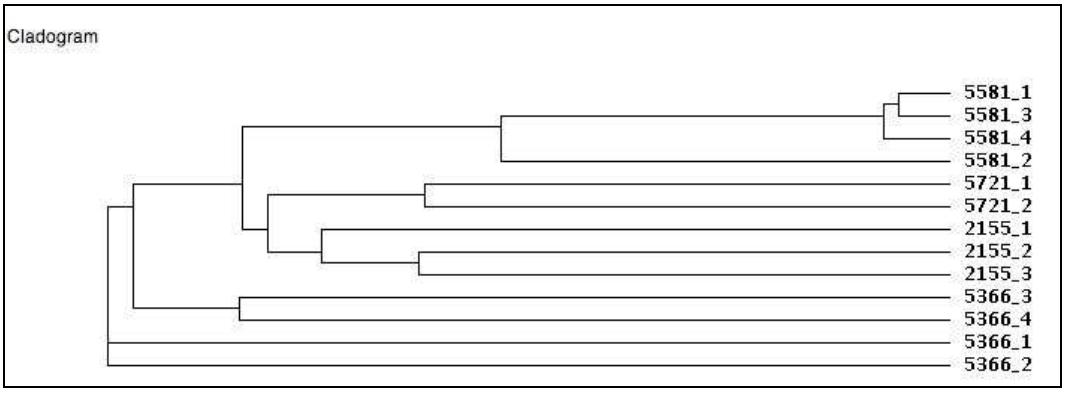
Example 1. Five contigs from the Rosaceae unigene match the SWISS-PROT protein 5NG4_PINTA. This protein, from Pinus taeda, is an auxin-induced protein. One of the matching contigs consists of all Prunus sequences; the other four consist of all Malus sequences. The ESTs that are part of these contigs were named with the format of the contig number followed by a unique number, and all of these were entered into ClustalW for assembly. This is one of the few examples where the Prunus unigenes are not grouped together in the cladogram. This contig may be an example of some other type of sequencing or assembly error other than evolutionary divergence between Prunus and Malus.

Contig Name	Source of ESTs	E-Value
Rosaceae_Contig13217	*Prunus*	2.00E-44
Rosaceae_Contig585	*Malus*	9.00E-56
Rosaceae_Contig1819	*Malus*	8.00E-62
Rosaceae_Contig4063	*Malus*	1.00E-29
Rosaceae_Contig12528	*Malus*	4.00E-23



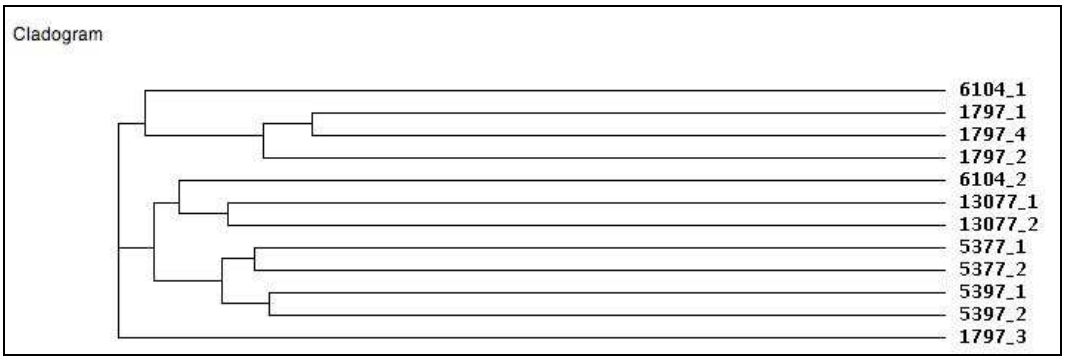
Example 2. Four contigs from the Rosaceae unigene match the SWISS-PROT protein 12KD_FRAAN. This protein, from Fragaria x ananassa, is an auxin repressed 12.5 kDa protein. Two of the matching contigs consists of all Prunus sequences; the other two consist of all Malus sequences. The ESTs that are part of these contigs were named with the format of the contig number followed by a unique number, and all of these were entered into ClustalW for assembly. The cladogram result confirms that the sequences grouped together do have more bases in common, and the two Prunus contigs are more closely related than the two Malus contigs.

Contig Name	Source of ESTs	E-Value
Rosaceae_Contig5581	*Prunus*	2.00E-13
Rosaceae_Contig5721	*Prunus*	2.00E-35
Rosaceae_Contig2155	*Malus*	2.00E-43
Rosaceae_Contig5366	*Malus*	2.00E-39



Example 3. Five contigs from the Rosaceae unigene match the SWISS-PROT protein ELI_PEA. This protein, from Pisum sativum, is an early light-induced protein. Two of the matching contigs consists of all Prunus sequences; the other three consist of all Malus sequences. The ESTs that are part of these contigs were named with the format of the contig number followed by a unique number, and all of these were entered into ClustalW for assembly. The cladogram result confirms that three of the four Prunus sequences group together with this model.

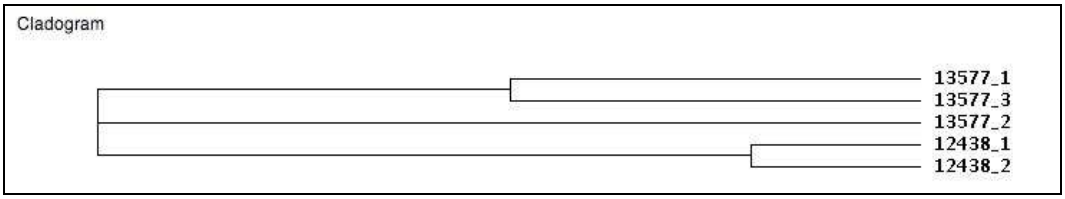
Contig Name	Source of ESTs	E-Value
Rosaceae_Contig6104	*Prunus*	1.00E-49
Rosaceae_Contig13077	*Prunus*	5.00E-40
Rosaceae_Contig1797	*Malus*	2.00E-39
Rosaceae_Contig5377	*Malus*	9.00E-34
Rosaceae_Contig5397	*Malus*	6.00E-40



Example 4. Two contigs from the Rosaceae unigene match the SWISS-PROT protein SUSY_SOYBN. This protein, from Glycine max, is a sucrose synthase protein. One of the matching contigs consists of all Prunus sequences; the other consists of all Malus sequences. The ESTs that are part of these contigs were

named with the format of the contig number followed by a unique number, and all of these were entered into ClustalW for assembly. The cladogram result confirms that the sequences grouped together do have more bases in common, and the Prunus sequences are more closely related than the Malus sequences.

Contig Name	Source of ESTs	E-Value
Rosaceae_Contig13577	*Prunus*	0
Rosaceae_Contig12438	*Malus*	1.00E-143



Appendix D

Bioinformatic Software Utilized in Research Efforts

Numerous bioinformatics software packages and databases were used in this research. They are listed in alphabetical order below.

SOFTWARE:

- autoSNP
    - Used to generate SNPs from unigene contigs
    - Barker G, Batley J, O' Sullivan H, Edwards KJ and Edwards D. 2003. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics* 12(19(3)):421-422.
- BLAST
    - Used to find sequence similarities between ESTs and either protein or nucleotide sequence databases.
    - Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215(3):403-410.
- CAP3
    - Used to align groups of ESTs into nonredundant consensus sequences (contigs) and singlets. These two groups comprise a unigene set.

- Huang X and Madan A. 1999. CAP3: A DNA sequence assembly program. *Genome Res.* 9(9):868-877.

- CROSS_MATCH
  - Used to mask vector regions from EST sequences.
  - Gordon D, Abajian C and Green P. 1998. Consed: a graphical tool for sequence finishing. *Genome Res.* 8(3):195-202.

- FASTX3.4
  - Used to find sequence similarities between ESTs and either protein or nucleotide sequence databases.
  - Pearson WR and Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A.* 1988;85:2444–2448.

- FLIP
  - Used to find putative ORFs in ESTs.
  - Brossard N. 1997. FLIP: a Unix Program used to find/translate orfs. Bionet.software<Message-ID:347B3A1B.794BDF32@bch.umontreal.ca>

- InterProScan
  - Used to find protein families, domains and functional sites in ESTs
  - Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A,

Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJ, Thomas PD, Valentin F, Wilson D, Wu CH and Yeats C. 2007. New developments in the InterPro database. *Nucleic Acids Res.* 35(Database issue):D224-8.

- PHRED
    - Used to base-call chromatograms and produce sequence and quality files.
    - Ewing B and Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8:186–94.
- Primer3
    - Used to generate primers for putative microsatellite sequences from ESTs.
    - Rozen S and Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol.* 132:365-86.
- SSRIT
    - Used to mine putative microsatellites from EST sequences.
    - Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S and McCouch S. 2001. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length

variation, transposon associations, and genetic marker potential. *Genome Res.* 11:1441-1452.

DATABASES:

- *Arabidopsis* proteins
    - Amino acid sequences from the Arabidopsis genome as curated by TAIR.
    - Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, Miller N, Mueller LA, Mundodi S, Reiser L, Tacklind J, Weems DC, Wu Y, Xu I, Yoo D, Yoon J and Zhang P. 2003. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Research* 31(1):224
- Gene Ontology
    - Controlled vocabularies for biological process, cellular component and molecular function of proteins.
    - Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash

RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R; Gene Ontology Consortium. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 32(Database issue):D258-261.

- GOA Slim
    - A smaller subset of the overall Gene Ontologies as generated by EMBL's EBI.
    - http://www.ebi.ac.uk/GOA/
- Mapping from keywords to GO Terms for SWISS-PROT
    - A mapping of the SWISS-PROT keywords to the Gene Ontology terms.
    - http://www.geneontology.org/external2go/spkw2go
- NCBI dbEST
    - A database of all the publicly available ESTs.
    - McEntyre J and Ostell J, eds. 2005. The NCBI Handbook. Bethesday(MD):National Library of Medicine (US), NCBI. Article : GenBank: The Nucleotide Sequence Database by Ilene Mizrachi updated July 27th, 2004

- NCBI nr
    - A comprehensive, nonredundant database of public protein sequences.
    - Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JU, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L and Yaschenko E. 2005. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2005 Jan 1;33(Database issue):D39-45.
- NCBI UniVec
    - A database of vector and oligonucleotide sequences. Used to screen contamination from ESTs.
    - Kitts PA, Madden TL, Sicotte H, and Ostell JA - Manuscript in preparation. The UniVec website can be accessed at http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html
- PlantGDB
    - A database of tentative unique genes for multiple plant species.
    - Dong Q, Schlueter SD and Brendel V. (2004) PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res.* 32(Database issue):D354-D359.
- Plant Structure Ontology

- A controlled vocabulary for plant structures. Used for tissue types of cDNA libraries.
- Ilic K, Kellogg EA, Jaiswal P, Zapata F, Stevens PF, Vincent LP, Avraham S, Reiser L, Pujar A, Sachs MM, Whitman NT, McCouch SR, Schaeffer ML, Ware DH, Stein LD and Rhee SY. 2006. Plant Structure Ontology. Unified Vocabulary of Anatomy and Morphology of a Flowering Plant. *Plant Physiol.* Dec 1 [Epub ahead of print].

- *Populus* proteins
  - Amino acid sequences from the Populus genome as curated by JGI.
  - Provided by DoE Joint Genome Institute and Poplar Genome Consortium at http://genome.jgi-psf.org/Poptr1/Poptr1.home.html
- SWISS-PROT
  - Curated protein database with extensive annotation.
  - Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S and Schneider M. 2003. The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31:365-370.
- TrEMBL
  - Computationally curated addition to SWISS-PROT. Includes translations of all nucleotide sequences from EMBL.

o Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N and Suzek B. 2005. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* 34(Database issue):D187-91.