Characterization of T-DNA integration sites within a population of insertional mutants of the diploid strawberry *Fragaria vesca* L.

by

Juan Jairo Ruiz-Rojas

Dissertation submitted to the faculty of Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Horticulture

Dr. Richard E. Veilleux- Chairman

Dr. Bingyu Zhao

Dr. Boris Vinatzer

Dr. M. A. Saghai Maroof

Dr. Vladimir Shulaev

November 1, 2010
Blacksburg, VA

Keywords: Rosaceae, T-DNA, SNPs, CAPS, dCAPS, genetic mapping, microsimilarities

Characterization of T-DNA integration sites within a population of insertional mutants of the diploid strawberry *Fragaria vesca* L.

Juan Jairo Ruiz-Rojas

Abstract

Cultivated strawberry (*Fragaria* × *ananassa*) is an octoploid (2*n*=8*x*=56) species that belongs to the Rosaceae family and the high ploidy level makes genetic and molecular studies difficult. However, its commercial success because of its unique flavor and nutritious qualities has increased interest in the development of genomic resources. *Fragaria vesca* L. is a diploid (2*n*=2*x*=14) species with a small genome size (206 Mbp), short reproductive cycle, and facile vegetative and seed propagation that make it an attractive model for genomic studies. The availability of an efficient transformation methodology for *Fragaria vesca* has facilitated the use of a T-DNA mutagenesis system to develop a collection of several hundred insertional T-DNA mutants at Virginia Tech, using either of two commercially available vectors, pCAMBIA 1302 and 1304. In this study, we have used expression of the green fluorescent protein (GFP) as a tool to identify homozygous mutant lines. Three different approaches were conducted, first we identified 11 homozygous lines by PCR, then another 55 homozygous lines by absence of segregation of GFP expression in T_2 seedlings, and finally we attempted to distinguish homozygous from hemizygous lines by relative GFP expression measured using a commercially available GFP meter. The latter methodology was unsuccessful due to uncontrolled variability in the readings. Continuing the characterization of our mutant population, we used thermal asymmetric interlaced PCR (TAIL-PCR) to obtain the nucleotide sequence of the genomic DNA regions that flank the T-DNA insertion sites in independent transgenic strawberry lines. Primers were designed that would amplify the derived strawberry flanking sequences in the two parents of an interspecific mapping population between the two diploid species, *F. vesca* x *F. bucharica.* The amplified products were sequenced and examined for the occurrence of SNPs (single nucleotide polymorphisms). The same primers were then used on the F_2 mapping population. Segregation of SNP markers with previously mapped genetic markers allowed us to position 74 SNP markers, and hence their corresponding insertional mutants, on a well-populated genetic linkage map for the diploid strawberry. Finally, we analyzed the insertion site from more

than 190 mutants looking at both the right and left borders of the T-DNA where microsimilarities of a few base pairs between ends of T-DNA and genomic DNA were observed, indicating that T-DNA integration had not occurred randomly in strawberry. We have also characterized the insertion sites through gene annotation found in the strawberry genome database.

Acknowledgements

My heartfelt thanks to my supervisor, Dr Richard Veilleux, whose encouragement, guidance, confidence and support from the beginning enabled me to develop and better understand my research every day. I would like to thank my co-advaisor Dr Vladimir Shulaev as well as the rest of my committee members Dr. Bingyu Zhao, Dr. Boris Vinatzer, and Dr. M. A. Saghai Maroof whose advice and guidance has made this project possible.

A very special thanks to Dr. Daniel J. Sargent for his invaluable support in his laboratory at East Malling, Kent UK, and Dr Jeremy Pattison for his expert technical instruction during my 2007 summer work at the Piedmont Research Station located in Blackstone, Virginia.

I want to thank the entire faculty and staff from the Horticulture Department at Virginia Tech, whose support has been priceless, in particular to Dr. Jerzy Nowak.

To my family, especially my mother, Bellanith, who has always given me the encourage to keep going forward in my life, to my endlessly loved brothers and sisters, who are my motivation everyday to succeed, and to my nephews who are my bright light and who remind me that there is always tomorrow. To all my friends that became my family during the completion of my program, Mihaela Babiceanu, Norma Manrique, Yeun Kyung, Nan Lu, Linda Taylor, Rachel Layman, Sarah Holt, Sukhwinder Aulak, Hope Gruszewski, and lastly, I offer my gratitude to all of those who supported me, all of those who walked in the same direction with me, all of those who were important in my life, all of those who for one or another reason are not around anymore.

Life changes, people change, we change, and I have changed, so thanks to all of you because you make me who I am now, for the better or the worst.

Thank you,

Table of Contents

Chapter 4: **Implementing reverse genetics in Rosaceae: analysis of T-DNA flanking sequences of insertional mutant lines in the diploid strawberry, *Fragaria vesca***

List of Tables

Chapter 4: **Implementing reverse genetics in Rosaceae: analysis of T-DNA flanking sequences of insertional mutant lines in the diploid strawberry, *Fragaria vesca***

List of Figures

Chapter 2: **Methods for identifying homozygous lines in a population of T-DNA insertional mutants of the diploid strawberry, *Fragaria vesca***

Chapter 4: **Implementing reverse genetics in Rosaceae: analysis of T-DNA flanking sequences of insertional mutant lines in the diploid strawberry, *Fragaria vesca***

Chapter 1:

Introduction:

Some of the most economically important cultivated crops in temperate regions are within the Rosaceae family. Crops like apple, peach, strawberry, apricot, pear, rose among others, are economically important not only in the United States but also internationally. Currently, the best-developed model species for Rosaceae include apple, peach (*Prunus persica*), and diploid strawberry (*Fragaria vesca*). In addition, many members of this family provide high value nutritional foods, ornamental plants, wood resources, etc. (Shulaev et al., 2008). The genus *Fragaria* Duch. consist of at least 15 different species with ploidy levels ranging from diploid in most wild type species to octoploids in the cultivated strawberry, however the most widely distributed natural species is the diploid European *Fragaria vesca* (Heide and Sonsteby, 2007; Shulaev et al., 2008). The application of molecular techniques for genetic improvement of most rosaceous crops is limited by genome size and complexity. Particularly in strawberry and related species the effort has focused on developing linkage maps to generate molecular marker information for use in breeding programs. Cultivated strawberry (*Fragaria ×ananassa*) is an octoploid (2*n*=8*x*=56) species and the high ploidy level makes genetic and molecular studies difficult. However, its commercial success because of its unique flavor and nutritious qualities has increased interest in the development of genomic studies, especially linkage maps (Sargent et al., 2004a; Sargent et al., 2009b). The alpine strawberry, *F. vesca* L., however, has several advantages to be considered as an attractive model within Rosaceae for genomic investigation. *F. vesca*, the most closely related diploid (2*n*=2*x*=14) to the cultivated octoploid strawberry, *Fragaria ×ananassa* (Ontivero et al., 2000; Potter et al., 2000), has a small genome size compared to the cultivated (206 vs. 600 Mbp per 1C nucleus) (Sargent et al., 2009a), short generation time (about 4 months), facile vegetative and seed propagation, and it is easy to propagate and manipulate both *in vitro* and in greenhouse conditions. Several genetic linkage maps have been constructed for *F. vesca* and for interspecific crosses of *Fragaria* spp. based on different types of molecular markers, including Random Amplified Polymorphic DNA (RAPDs) (Davis and Yu, 1997), microsatellites (Sargent et al., 2004b; Sargent et al., 2006) and more recently Sequence Tagged Sites (STS) (Nier et al., 2006), and Single Nucleotide Polymorphism

(SNPs) (Ruiz-Rojas et al., 2010). These characteristics and the phylogenetic relationships of *F. vesca* with other members of the Rosaceae family also support using the alpine strawberry as an attractive model for functional genomics (Sargent et al., 2009a; Sargent et al., 2004a).

A greater knowledge of the phylogenetic relationships between the diploid strawberry species is important for understanding the origins of the progenitors of the octoploid cultivated strawberry. For example, the physiology of flowering of the cultivated strawberry has been extensively researched because of the commercial importance of this plant, while flowering of wild species has been studied less (Battey NH et al., 1998; Heide and Sonsteby, 2007) Furthermore, developing genomic tools to understand the genome organization of wild species like *F. vesca* can facilitate the reconstruction of the genome structure in the cultivated strawberry and transferring or improvement of agronomic traits of interest (Folta and Davis, 2006; Sargent et al., 2004a; Sargent et al., 2009b; Sargent et al., 2009c).

Insertional mutagenesis

In addition to the construction of genetic maps, another tool used to understand the genome structure and gene function is insertional mutagenesis. This functional analysis approach of disrupting gene function is based on the insertion of foreign DNA (T-DNA) into the gene of interest. The effect of T-DNA insertion is most commonly evaluated by monitoring expression of mutated genes in homozygous insertion mutants. The foreign DNA not only disrupts the expression of the gene into which it is inserted but also acts as a marker for subsequent identification of the mutation. An advantage of using T-DNA as the insertional mutagen, as opposed to transposons, is that T-DNA insertions will not transpose subsequent to integration within the genome and are therefore chemically and physically stable through multiple generations (Krysan et al., 1999; Wang, 2008; Yu et al., 2010)

The use of insertional mutagenesis in functional genomics is well-documented in several plant species like Arabidopsis(Alonso et al., 2003; Tadege et al., 2008), tomato (Pineda et al., 2010), rice (Li et al., 2006; Sripriya et al., 2008), maize (Oltmanns et al., 2010), etc. In *Arabidopsis*; more than 21,700 T-DNA insertional mutants were identified from 225,000 independent events that were created. T-DNA insertional mutations were identified for almost every single gene known and predicted in the *Arabidopsis* genome, making this mutant

collection important for understanding the function of each single gene, not only in *Arabidopsis* but also in other species (Alonso et al., 2003). In tomato, more than 2000 T-DNA insertion mutant lines were generated by *Agrobacterium*-based transformation leading to the finding of a mutation that affected the reproductive development and conversion of sepals into fleshy fruit-like organs which revealed a key for reproductive development in tomato (Pineda et al., 2010). In rice, a collection of mutants was generated by T-DNA insertion; 4416 rice T_1 tagged lines were screened and genetically analyzed. Several lines with morphological trait mutations were found as a result of loss of function by the T-DNA (Li et al., 2006).

Despite the proven use of insertional mutagenesis in different plant systems (Alonso et al., 2003; Li et al., 2006; Oltmanns et al., 2010; Pineda et al., 2010; Sripriya et al., 2008; Tadege et al., 2008), this strategy has not been utilized in any rosaceous species. With the growing use of *F. vesca* as a model plant in genetic studies, an initial effort has been developed to establish forward and reverse genetic platforms in strawberry as a model rosaceous crop, by implementing T-DNA mutagenesis (Oosumi et al., 2010; Ruiz-Rojas et al., 2010; Sargent et al., 2009a). According to Oosumi et al. (2006), approximately 255,000 independent T-DNA-transformed lines are required to have a high probability that every single gene in the *F. vesca* genome has been mutated. Before beginning an enterprise such as developing a large collection of mutants, we need to document the ability to characterize genes in *F. vesca* by insertional mutagenesis on a small scale. Therefore, it is necessary to evaluate a pilot collection of mutants for the gene function that has been interrupted and furthermore locate those genes in the most complete genetic map available for *F. vesca* × *F. bucharica* (Sargent et al., 2004a)

A research project to implement a reverse genetics approach of functional genomics in *F. vesca* has been recently established. The availability of an efficient transformation methodology has facilitated the development of a T-DNA mutagenesis system. As a preliminary effort to develop a collection of insertional T-DNA mutants in *F. vesca*, our laboratory at Virginia Tech has generated more than 500 T-DNA insertion lines that can be used for functional genomics. Initial morphological screening of T_1 progenies of 80 T-DNA insertion lines showed four leaf-altered phenotypes (Sargent et al., 2009a). More recently, this collection of mutants was expanded to 700 T-DNA insertions and more than 300 have been characterized by hi-TAIL PCR, SNPs discovery and genetic mapping (Oosumi et al., 2010; Ruiz-Rojas et al., 2010). Further

segregation and genomic analysis, as well as phenotypic screening is needed to identify the genomic location of each T-DNA insert and the possible gene disrupted by each insertion. Finally, development of the existing gene-tagging system in *F. vesca* would complement the current genomics tools in strawberry and greatly facilitate functional gene annotation and analysis.

Thermal asymmetric interlaced - PCR (TAIL-PCR)

An alternative approach for the initial screening of transgenic events is to characterize the genomic DNA regions flanking the T-DNA insertion sites by TAIL-PCR. TAIL-PCR was originally described by Liu and Whittier in 1995. It is an efficient method to amplify unknown sequences adjacent to known insertion sites. This technique basically involves PCR using nested specific primers and arbitrary degenerate primers for the unknown genomic DNA region flanking the insertion site (Liu and Whittier, 1995). As a result, priming creates specific and nonspecific products that are thermally controlled by PCR cycles carried out with alternating high annealing temperatures and lower temperatures, allowing both primers to function. More recently, an improved protocol called high-efficiency thermal asymmetric interlaced PCR (hiTAIL-PCR) was developed by Liu and Cheng to optimize TAIL-PCR in order to increase the success rate and generate larger product sizes from T-DNA insertions flanking unknown sequences from transgenic rice lines. In this study, four relatively longer arbitrary degenerate primers were designed of 33 or 34 nucleotides for hiTAIL-PCR and taking advantage of the pCAMBIA binary vector that was used to generate transgenic rice lines, a set of three nested specific primer was designed according to the sequences adjacent to the T-DNA right border. Optimization of hiTAIL-PCR was achieved to generate product sizes between 1 and 3 kb with a 90% success rate (Liu and Chen, 2007). One of the greatest advantages of these primers is that they are universally applicable for various organisms.

TAIL-PCR has been widely used to recover flanking sequences of T-DNA and/or transposon insertion for functional genomics, cloning such insertion-tagged genes, recovering upstream regulatory sequences of tagged genes, cloning genes from large molecules such as P1 phage, yeast and bacterial artificial chromosomes, etc. (Liu and Whittier, 1995; Liu and Chen, 2007). Like TAIL-PCR, many other PCR-based techniques such as inverse PCR, gene walking

PCR and single primer PCR can be used for the same purpose, but they employ special steps before PCR such as Southern analysis, restriction digestion and ligation, making the task laborious, and they also tend to generate non-specific products (Ochman et al., 1990).

Everyday there are more and more crops for which TAIL-PCR is utilized for isolation of unknown sequences flanking T-DNA insertions. In rice, more than 361 flanking sequences were amplified by TAIL-PCR from 5,200 independent T-DNA tagged lines (Sha et al., 2004). Miniature inverted repeat transposable elements (MITEs) were recovered from peanut by TAIL-PCR modifying peanut MITE specific nested primers to recover the 5' genomic flanking sequence tag (Bhat et al., 2008). Isolation of genomic sequences flanking T-DNA insertion from transgenic lines of *Arabidopsis thaliana* was successfully achieved by TAIL-PCR (Liu and Whittier, 1995). In strawberry, verification of T-DNA integration was successfully done by TAIL-PCR (Hanhineva and Karenlampi, 2007). As part of the current research project a population of 321 T-DNA insertional mutants of diploid strawberry has been characterized (Oosumi et al., 2010; Ruiz-Rojas et al., 2010)

Single Nucleotide Polymorphism (SNP)

Before the advent of molecular markers, plant breeders based their selection on phenotypic agronomic traits such as plant height and grain yield. Even today, such an approach has considerable merit. However, with the development of marker technology, plant breeders are provided with tools capable of monitoring the whole genome in the absence of a phenotype (Edwards and Mogg, 2001). Molecular markers are polymorphic when there is DNA sequence variation between the individuals under study. Molecular markers are, therefore, simply an indicator of sequence polymorphism. Sequence polymorphism between individuals can take many forms, for instance, it can be due to the insertion or deletion of multiple bases, or it can be due to single nucleotide polymorphism or SNPs (Brookes, 1999; Edwards and Mogg, 2001). Insertions, deletions and SNPs are important in determining sequence variation between individuals.

SNPs are DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered, making them the most frequent form of DNA variation in the genome (Thiel et al., 2004). For example a SNP might change the DNA sequence A**A**GGCTAA

to A**T**GGCTAA. For a variation to be considered a SNP, it must occur in at least 1% of the population (Kruglyak, 1997; Ogren, 2003). SNPs, which make up about 90% of all human genetic variation, occur every 100 to 300 bases along the 3-billion-base human genome. Two of every three SNPs involve the replacement of cytosine (C) with thymine (T). Many SNPs have no effect on cell function, but scientists believe others could predispose people to disease or influence their response to a drug. SNPs may fall within coding sequences of genes, non-coding regions of genes, or in the intergenic regions between genes. SNPs within a coding sequence will not necessarily change the amino acid sequence of the protein that is produced, due to redundancy in the genetic code (Brookes, 1999; Collins et al., 1998; Gut, 2005.)

In barley, sixteen genotypes were targeted to amplify and sequence the *Isa* gene region to detect sequence polymorphisms. The *Isa* gene from barley, an intron-less gene expressed in maternal tissues of the seed has a likely role in defense against pathogens. The protein product—bifunctional α-amylase/subtilisin inhibitor—inhibits the seed's own amylase in addition to the bacterial protease subtilisin and fungal xylanase. Little is known about genetic diversity at this locus. A total of 80 SNPs and 23 indels was detected in 2,164 bp of sequence containing the *Isa* transcript, promoter and 3′ non-transcribed region (overall one SNP per 27 bp and one indel per 94 bp), with eight sequence-based haplotypes distinguishable amongst the 16 varieties (Bundock and Henry, 2004).

Supernodulation in soybean (*Glycine max* L. Merr.) is an important source of nitrogen supply to subterranean ecological systems. The gene *GmNARK* (*Glycine max* nodule autoregulation receptor kinase), controlling autoregulation of nodulation, was found to have a SNP between the wild-type cultivar Sinpaldalkong 2 and its supernodulating mutant, SS2-2 (Kim et al., 2005). Transversion of A to T at the 959-bp position of the *GmNARK* sequence resulted in a change of lysine (AAG) to a stop codon (TAG), thus terminating its translation in SS2-2. Based on the identified SNP in *GmNARK*, among 28 individuals with the normal phenotype, eight individuals having only the A-allele-specific band were homozygous and normal, while 20 individuals were found to be heterozygous at the SNP having both A and T bands. Twelve supernodulating individuals showed only the band specific to the T allele (Kim et al., 2005).

In the last decades, the use of DNA markers for the study of crop genetic diversity has become routine, and has revolutionized biology. Increasingly, techniques are being developed to more precisely, quickly and cheaply assess genetic variation. These techniques have changed the standard equipment of many labs, and most scientists who study plant germplasm are expected to be trained in DNA data generation and interpretation (Karp et al., 1997). This is how a procedure called Cleaved Amplified Polymorphic Sequence (CAPS) was developed to facilitate the mapping of gene mutations in *Arabidopsis* by a simple PCR-based strategy. This technique detects a single-base change that creates a unique restriction enzyme site in one of a pair of alleles (Konieczny and Ausubel, 1993; Michaels and Amasino, 1998). Unfortunately, the majority of single nucleotide changes does not create such sites and cannot be used to create CAPS markers. However, a modification of the CAPS technique allowing the introduction of one or two mismatches by the introduction of those mismatches on a new primer that changes one base next to the original single-base change allows the creation of a unique restriction site in one of the alleles, this modification is called derived Cleaved Amplified Polymorphic Sequence (dCAPS) (Michaels and Amasino, 1998). In rice and strawberry, the CAPS/dCAPS methods were utilized to convert SNP markers generating a restriction site difference between lines into PCR-based markers that were rapidly and reliably analyzed (Komori and Nitta, 2005; Ruiz-Rojas et al., 2010).

Genetic linkage map as a tool to position a collection of T-DNA strawberry mutants

The use of DNA polymorphisms as genetic markers for organisms with complex genomes has made the mapping, manipulation, and cloning of genes associated with interesting biological traits possible. DNA polymorphisms are particularly useful for genetic analysis because the DNA of any organism exhibits an abundance of individual variation (Green, 1997). DNA marker analysis of genomes has revolutionized genetic studies of organisms in the past decade; genetic maps have been developed for all major crops like Arabidopsis (Zhang et al., 2006), tomato (Liu, 2005), maize (Quint et al., 2002), apple (Han et al., 2009), strawberry (Rousseau-Gueutin et al., 2008; Sargent et al., 2004a), etc. The development of technologies for large DNA fragment cloning such as bacterial artificial chromosome (BACs), for fingerprinting and for contig assembly, as well as the increasing understanding of how DNA mobile elements work, provided powerful tools to rapidly generate genome wide physical maps and characterize

the function of genes. A physical map integrated with a DNA marker genetic map and the exact location of T-DNA knockout insertions, provides improved efficiency in isolating genes known only by their phenotypes and the study of genome structure and organization (Wing et al., 1994; Zhang et al., 2006)

In genetics and breeding, mapping populations are the tools used to identify the genetic loci or genes controlling measurable phenotypic traits. Mapping populations consist of individuals of one species, or in some cases they derive from a cross between related species where the parents differ in the traits to be studied. These genetic tools are used to identify factors or loci that influence phenotypic traits and to determine the recombination distance between loci. In different organisms of the same species, the genes, represented by alternate allelic forms are arranged in a fixed linear order on the chromosomes. Linkage values among genetic factors are estimated based on recombination events between alleles of different loci, and linkage relationships along all chromosomes provide a genetic map of an organism (Meksem and Kahl, 2005). A genetic linkage map has been constructed based on an F_2 population derived from an interspecific cross between two diploid *Fragaria* species, *F. vesca* f. *semperflorens,* and *F. bucharica*, using informative sequence tagged-site (STS) markers developed for *Fragaria*, including five microsatellite (SSR) loci, and three morphological markers (Sargent et al., 2004a). Additionally, a linkage map of a diploid *Fragaria* backcross progeny (*F. vesca* × [*F. vesca* × *F. viridis*]) from 31 SSR and two gene-specific markers selected from the diploid *Fragaria* reference map *F. vesca* × *F bucharica*, has been produced. Twenty three selected STS markers segregated in seven linkage groups and marker order was conserved between *F. vesca* and *F. bucharica*. This indicates that no major genomic reorganization has occurred during speciation in the genus, and suggests that STS linkage maps of other *Fragaria* populations, including those of the cultivated strawberry *F.* ×*ananassa*, could be constructed using the diploid *Fragaria* reference map, *F vesca* × *F bucharica*, as a standard reference point (Nier et al., 2006). In addition to that, 22 microsatellite markers have been characterized from *F. viridis*, and their amplification in the parents of a *F. vesca* mapping population, were conducted by Sargent and collaborators (Sargent et al., 2003). More recently 74 new SNPs markers were mapped onto the strawberry reference map that correspond to the genetic location of T-DNA insertions in the respective transgenic strawberry lines (Ruiz-Rojas et al., 2010).

Establishing the function of unknown genes is one of the main goals in biological sciences. One approach known to determine the function of those genes is by discovering the genes associated with a particular phenotype. This is known as forward genetics. The opposite would be reverse genetics, when the obvious phenotype is unknown, but a gene is known. As genome sequencing of different plant species has become possible, an essential tool for the functional analysis of these completely sequenced genomes has focused on the ability to create loss-of-function mutations for all genes. Gene tagging is one technique used for loss-of-function in genes using mobile or introduced DNA with a known sequence (T-DNA). The insertion of foreign T-DNA into a gene both mutates the gene and serves as a vehicle for the isolation of the unknown flanking DNA. *Fragaria vesca,* a wild type strawberry has become a diploid model system for genetic and genomic studies in order to bypass the complexities inherent in the octoploid genome structure of the cultivated strawberry (Sargent et al., 2009a; Sargent et al., 2004a; Shulaev et al., 2008).

In this study, the main goal was to characterize an insertional mutant collection generated by *Agrobacterium* transformation of *Fragaria vesca* L, as a part of the development of this genomic tool in the model system. During the transformation and regeneration process, we selected transgenic strawberry lines by GFP expression, and used segregation of GFP expression in T_1 and T_2 progenies to identify homozygous mutants that were entered into a strawberry FileMaker Pro database (www.hort.vt.strawberry.edu). We isolated the T-DNA flanking regions of the collection of strawberry mutants through hiTAIL-PCR and performed gene annotation using the strawberry genome browser (www.strawberrygenome.org). In addition, SNPs discovered in the flanking regions were converted into PCR-based CAPS/dCAPS markers allowing us to map each T-DNA insertion to a *F. vesca* × *F. bucharica* reference mapping population. These SNPs in the flanking regions enriched the genetic map and were used extensively in anchoring scaffolds obtained from 454 sequencing to linkage groups in the strawberry genome assembly (Shulaev, et al. unpublished).

References:

Alonso J.M., Stepanova A.N., Leisse T.J., Kim C.J., Chen H., Shinn P., Stevenson D.K., Zimmerman J., Barajas P., Cheuk R., Gadrinab C., Heller C., Jeske A., Koesema E., Meyers C.C., Parker H., Prednis L., Ansari Y., Choy N., Deen H., Geralt M., Hazari N.,

Hom E., Karnes M., Mulholland C., Ndubaku R., Schmidt I., Guzman P., Aguilar-Henonin L., Schmid M., Weigel D., Carter D.E., Marchand T., Risseeuw E., Brogden D., Zeko A., Crosby W.L., Berry C.C., Ecker J.R. (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. Science 301:653-657.

Battey NH, LeMière P, Teranifar A, Cekic C, Taylor S, Shrives KJ, Greenland AJ, Darby J, . W.M. (1998) Genetic and environmental control of flowering in strawberry. In: CockshullKE, GrayD, ThomasB (eds) Genetic and environmental manipulation of horticultural crops. CAB International Publications, Oxford, UK, pp 111-131.

Bhat R.S., Patil V.U., Chandrashekar T.M., Sujay V., Gowda M.V.C., Kuruvinashetti M.S. (2008) Recovering flanking sequence tags of a miniature inverted-repeat transposable element by thermal asymmetric interlaced-PCR in peanut. Curr. Sci. 95:452-453.

Brookes A.J. (1999) The essence of SNPs. Gene 234:177-186.

Bundock P.C., Henry R.J. (2004) Single nucleotide polymorphism, haplotype diversity and recombination in the Isa gene of barley. Theor. Appl. Genet. 109:543-51.

Collins F.S., Brooks L.D., Chakravarti A. (1998) A DNA polymorphism discovery resource for research on human genetic variation. Genome Res. 8:1229-31.

Davis T.M., Yu H. (1997) A linkage map of the diploid strawberry, *Fragaria vesca*. J Hered 88:215-221.

Edwards K., Mogg R. (2001) Plant genotyping by analysis of single nucleotide polymorphisms. Henry, R. J ed. CABI Publishing, Lismore.

Folta K.M., Davis T.M. (2006) Strawberry genes and genomics. Crit. Rev. Plant Sci. 25:399-415.

Green E.D. (1997) Genome analysis : a laboratory manual Cold Spring Harbor Laboratory Press. pp 493, Plainview, N.Y.

Gut I.G. (2005.) An overview of genotyping and single nucleotide polymorphism (SNP), in: R. Rapley and S. Harbron (Eds.), Molecular analysis and genome discovery, John Wiley & Sons, Ltd, Chichester, UK.

Han Y., Chagné D., Gasic K., Rikkerink E.H.A., Beever J.E., Gardiner S.E., Korban S.S. (2009) BAC-end sequence-based SNPs and Bin mapping for rapid integration of physical and genetic maps in apple. Genomics 93:282-288.

Hanhineva K.J., Karenlampi S.O. (2007) Production of transgenic strawberries by temporary immersion bioreactor system and verification by TAIL-PCR. BMC Biotechnol. 7:11.

Heide O.M., Sonsteby A. (2007) Interactions of temperature and photoperiod in the control of flowering of latitudinal and altitudinal populations of wild strawberry (*Fragaria vesca*). Physiol. Plant. 130:280-289.

Karp A., Edwards K.J., Bruford M., Funk S., Vosman B., Morgante M., Seberg O., Kremer A., Boursot P., Arctander P., Tautz D., Hewitt G.M. (1997) Molecular technologies for biodiversity evaluation: opportunities and challenges. Nat. Biotechnol. 15:625-8.

Kim M.Y., Van K., Lestari P., Moon J.K., Lee S.H. (2005) SNP identification and SNAP marker development for a GmNARK gene controlling supernodulation in soybean. Theor. Appl. Genet. 110:1003-10.

Komori T., Nitta N. (2005) Utilization of the CAPS/dCAPS method to convert rice SNPs into PCR-based markers. Breeding Sci 55:93-98.

Konieczny A., Ausubel F.M. (1993) A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. Plant J. 4:403-10.

Kruglyak L. (1997) The use of a genetic map of biallelic markers in linkage studies. Nat. Genet. 17:21-4.

Krysan P.J., Young J.C., Sussman M.R. (1999) T-DNA as an insertional mutagen in *Arabidopsis*. Plant Cell 11:2283-90.

Li A.H., Zhang Y.F., Wu C.Y., Tang W., Wu R., Dai Z.Y., Liu G.Q., Zhang H.X., Pan X.B. (2006) Screening for and genetic analysis on T-DNA-inserted mutant pool in rice. Acta Gen. Sin. 33:319-29.

Liu Y.-G., Whittier R.F. (1995) Thermal asymmetric interlaced PCR: automatable amplification and sequencing of insert end fragments from P1 and YAC clones for chromosome walking. Genomics 25:674-681.

Liu Y. (2005) Construction of a genetic map and location of quantitative trait loci for number of flowers per truss traits in tomato by SSR markers. J. Nanjing Agric. Univ. 28:30-34.

Liu Y.G., Chen Y. (2007) High-efficiency thermal asymmetric interlaced PCR for amplification of unknown flanking sequences. Biotechniques 43:649-50.

Meksem K., Kahl G. (2005) The handbook of plant genome mapping : genetic and physical mapping Wiley-VCH. pp 380, Weinheim; [Great Britain].

Michaels S.D., Amasino R.M. (1998) A robust method for detecting single-nucleotide changes as polymorphic markers by PCR. Plant J. 14:381-5.

Nier S., Simpson D.W., Tobutt K.R., Sargent D.J. (2006) A genetic linkage map of an inter-specific diploid *Fragaria* BC1 mapping population and its comparison with the *Fragaria* reference map (FV × FN) J Hortic Sci Biotech 81:645-650.

Ochman H., Ajioka J.W., Garza D., Hartl D.L. (1990) Inverse polymerase chain reaction. Biotechnology (N Y) 8:759-60.

Ogren M. (2003) Whole-genome SNP genotyping. The Scientist. 17:11-42.

Oltmanns H., Frame B., Lee L.Y., Johnson S., Li B., Wang K., Gelvin S.B. (2010) Generation of backbone-free, low transgene copy plants by launching T- DNA from the *Agrobacterium* chromosome. Plant Physiol. 152:1158-1166.

Ontivero M., Arias M., Diaz J., babot J., Albornoz P., Castagnaro A. (2000) Analysis of genetic similarities among species of *Fragaria*, *Potentilla*, and *Duchesnea* found in northwest Argentina by using morphological, anatomical, and molecular characters. Can. J. Bot. 78:547-556.

Oosumi T., Ruiz-Rojas J.J., Veilleux R.E., Dickerman A., Shulaev V. (2010) Implementing reverse genetics in Rosaceae: analysis of T-DNA flanking sequences of insertional mutant lines in the diploid strawberry, *Fragaria vesca*. Physiol. Plant. 140:1-9.

Pineda B., Gimenez-Caminero E., Garcia-Sogo B., Anton M.T., Atares A., Capel J., Lozano R., Angosto T., Moreno V. (2010) Genetic and physiological characterization of the *Arlequin* insertional mutant reveals a key regulator of reproductive development in tomato. Plant Cell Physiol. 51:435-447.

Potter D., Luby J.J., Harrison R.E. (2000) Phylogenetic relationships among species of *Fragaria* (Rosaceae) inferred from non-coding nuclear and chloroplast DNA sequences. Syst. Bot. 25:337-348.

Quint M., Mihaljevic R., Dussle C.M., Xu M.L., Melchinger A.E., Lübberstedt T. (2002) Development of RGA-CAPS markers and genetic mapping of candidate genes for sugarcane mosaic virus resistance in maize. Theor. Appl. Genet. 105:355-363.

Rousseau-Gueutin M., Lerceteau-Kohler E., Barrot L., Sargent D.J., Monfort A., Simpson D., Arus P., Guerin G., Denoyes-Rothan B. (2008) Comparative genetic mapping between octoploid and diploid *Fragaria* species reveals a high level of colinearity between their genomes and the essentially disomic behavior of the cultivated octoploid strawberry. Genetics 179:2045-2060.

Ruiz-Rojas J., Sargent D., Shulaev V., Dickerman A., Pattison J., Holt S., Ciordia A., Veilleux R. (2010) SNP discovery and genetic mapping of T-DNA insertional mutants in *Fragaria vesca*; L. Theor. Appl. Genet. 121:449-463.

Sargent D.J., Hadonou A.M., Simpson D.W. (2003) Development and characterization of polymorphic microsatellite markers from *Fragaria viridis*, a wild diploid strawberry. Mol. Ecol. Notes 3:550-552.

Sargent D.J., Davis T.M., Simpson D.W. (2009a) Strawberry (*Fragaria* spp.) structural genomics, in: K. M. Folta and S. E. Gardiner (Eds.), Genetics and genomics of Rosaceae, Springer New York. pp. 437-456.

Sargent D.J., Geibel M., Hawkins J.A., Wilkinson M.J., Battey N.H., Simpson D.W. (2004a) Quantitative and qualitative differences in morphological traits revealed between diploid *Fragaria* species. Ann. Bot. 94:787-96.

Sargent D.J., Davis T.M., Tobutt K.R., Wilkinson M.J., Battey N.H., Simpson D.W. (2004b) A genetic linkage map of microsatellite, gene-specific and morphological markers in diploid *Fragaria*. Theor. Appl. Genet. 109:1385-91.

Sargent D.J., Fernandéz-Fernandéz F., Ruiz-Rojas J., Sutherland B., Passey A., Whitehouse A., Simpson D. (2009b) A genetic linkage map of the cultivated strawberry (*Fragaria* × *ananassa*) and its comparison to the diploid *Fragaria* reference map. Mol. Breed. 24:293-303.

Sargent D.J., Marchese A., Simpson D.W., Howad W., Fernández-Fernández F., Monfort A., Arús P., Evans K.M., Tobutt K.R. (2009c) Development of "universal" gene-specific markers from *Malus spp*. cDNA sequences, their mapping and use in synteny studies within Rosaceae. Tree Genet. Genom. 5:133-145.

Sargent D.J., Clarke J., Simpson D.W., Tobutt K.R., Arus P., Monfort A., Vilanova S., Denoyes-Rothan B., Rousseau M., Folta K.M., Bassil N.V., Battey N.H. (2006) An enhanced microsatellite map of diploid *Fragaria*. Theor. Appl. Genet. 112:1349-59.

Sha Y., Li S., Pei Z., Luo L., Tian Y., He C. (2004) Generation and flanking sequence analysis of a rice T-DNA tagged population. Theor. Appl. Genet. 108:306-314.

Shulaev V., Korban S.S., Sosinski B., Abbott A.G., Aldwinckle H.S., Folta K.M., Iezzoni A., Main D., Arus P., Dandekar A.M., Lewers K., Brown S.K., Davis T.M., Gardiner S.E., Potter D., Veilleux R.E. (2008) Multiple models for Rosaceae genomics. Plant Physiol. 147:985-1003.

Sripriya R., Raghupathy V., Veluthambi K. (2008) Generation of selectable marker-free sheath blight resistant transgenic rice plants by efficient co-transformation of a cointegrate vector T-DNA and a binary vector T-DNA in one *Agrobacterium tumefaciens* strain. Plant Cell Rep. 27:1635-1644.

Tadege M., Wen J.Q., He J., Tu H.D., Kwak Y., Eschstruth A., Cayrel A., Endre G., Zhao P.X., Chabaud M., Ratet P., Mysore K.S. (2008) Large-scale insertional mutagenesis using the *Tnt1* retrotransposon in the model legume *Medicago truncatula*. Plant J. 54:335-347.

Thiel T., Kota R., Grosse I., Stein N., Graner A. (2004) SNP2CAPS: a SNP and INDEL analysis tool for CAPS marker development. Nucleic Acids Res. 32:e5.

Wang Y.H. (2008) How effective is T-DNA insertional mutagenesis in *Arabidopsis*? J. Biochem. Technol 1:11-20.

Wing R.A., Zhang H.B., Tanksley S.D. (1994) Map-based cloning in crop plants. Tomato as a model system: I. Genetic and physical mapping of jointless. Mol. Gen. Genet. 242:681-8.

Yu J.G., Lee G.H., Kim J.S., Shim E.J., Park Y.D. (2010) An insertional mutagenesis system for analyzing the Chinese cabbage genome using *Agrobacterium* T-DNA. Mol. Cells 29:267-275.

Zhang S., Zhou Q., Chen H., Yi J., Gao J.F., Yang Z.N. (2006) Identification and genetic mapping of a novel gene involved in pollen development in *Arabidopsis*. J. Mol. Cell Biol. 39:163-8.

Chapter 2:

Methods for identifying homozygous lines in a population of T-DNA insertional mutants of the diploid strawberry, *Fragaria vesca*.

Ruiz-Rojas JJ[1], Baxter AJ[1], Holt SH[1], Mills K[1], Davis C[1], Morris, B[1], Upham K[1], Oosumi T[2], Shulaev V[1,2], and Veilleux RE[1](✉)

Author contribution:

Aaron J. Baxter was a technician working at Horticulture department at Virginia Tech on the strawberry project during 2007; he contributed to the quantification of GFP by troubleshooting the GFP meter and working with me to collect the data. He also conducted strawberry transformation experiments that engendered some of the plant material used in the study. Sarah Holt is a Ph.D. candidate in Dr Veilleux's laboratory; she was also involved in strawberry transformation experiments and worked on the team organizing the experiment and collecting samples from the transgenic plants in the greenhouse. Kerri Mills is a technician working in our laboratory; she did almost all the PCR zygosity testing and primer design. Cherish Davis is a technician responsible for the strawberry database; she developed and maintains the database, and will make it available to the public. Benjamin Morris and Kendal Upham were undergrad students who worked on surface sterilization of seeds, GFP screening under UV light and greenhouse work. Teruko Oosumi was a postdoctoral associate working at the Virginia Bioinformatics Institute until 2005. She worked under Dr. Vladimir Shulaev's supervision and she started strawberry transformations and characterization of the insertional mutant populations used in this investigation. Dr Richard Veilleux and Dr Vladimir Shulaev are the principal investigators on the strawberry project; they directed the investigation and edited the article. Finally, I contributed in all topics related with this publication; I collected and processed all the data as well as statistical analysis guided by Dr. Guillaume Pilot from the Department of Plant Pathology, Physiology and Weed Science at Virginia Tech. I also wrote the first draft of this paper.

[1]*Department of Horticulture, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA*
[2]*Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA*

Methods for identifying homozygous lines in a population of T-DNA insertional mutants of the diploid strawberry, *Fragaria vesca*

Ruiz-Rojas JJ[1], Baxter AJ[1], Holt SH[1], Mills K[1], Davis C[1], Morris, B[1], Upham K[1], Oosumi T[2], Shulaev V[1,2], and Veilleux RE[1](✉)

(✉) Author for correspondence
Tel: 540 231-5584
Fax: 540 231-3083
E-mail: potato@vt.edu

[1]Department of Horticulture, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA
[2]Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

Abstract

Insertional mutant population has become an important tool for understanding the complexity of plant genomes, not only because genes can be disrupted and it phenotypic changes might be allow to see, but also because it can lead the way to determine where those genes are located within the genome. Reporter genes as insertional elements have been utilized to aid in the identification of insertions within functional genes. On this study we have attempted 3 different methods for identifying homozygous knockout mutants from a T-DNA insertional population using GFP-based marker as a reporter gene. Three different approaches were used in this investigation, we have been able to identify homozygous lines both transgene inheritance and segregation patterns of GFP from the T_2 segregating population and by PCR on T_1 lines. Our attempt to identify homozygous lines from T_1 plants using a commercially GFP meter was not possible for the diploid strawberry, *Fragaria vesca* L. in this investigation.

Keywords: Rosaceae, Green Fluorescent Protein, PCR, homozygous, hemizygous, insertional mutant population

Introduction

Classical selection of transformed organisms using antibiotic resistance, proteins able to metabolize chromogenic chemical agents or cell surface proteins detected by immunological techniques often involve the use of several steps in the process of eliminating escapes. Transgenesis, gene therapy and numerous basic research projects have relied on the design of efficient gene transfer strategies and gene reporters which allow the visualization and localization of the transgene (Hicks, 2009). Since the discovery of the Green Florescent Protein (GFP) from the jellyfish *Aequorea aequirea* by Shimomura et al.(1962) who was awarded the Nobel Prize in Chemistry in 2008, countless studies have been conducted using GFP as a reporter gene because it is the most useful visual marker for screening genetic transformation in both animal and plants. The protein has the unique ability to transduce UV or blue light to green light, which makes it possible to visualize fluorescence in transgenic organisms by shining a bright ultraviolet light in darkened locations (Stewart, 2001). GFP not only allows for the convenient and immediate visualization and subcellular localization of transformed cells, but also permit the real time imaging of subcellular movement of tagged proteins in living organism without destruction of tissue (Leclercq et al., 2010).

Numerous studies have been conducted using GFP as a selectable marker in several transgenic plant species including *Citrus* (Tan et al., 2009), *Arabidopsis thaliana* (Shimada et al., 2010), *Coffea* (Mishra et al., 2010), *Oryza sativa* (Afolabi et al., 2005), and *Fragaria* (Oosumi et al., 2006; Slovin et al., 2009), among others. An efficient system for *Agrobacterium*-mediated transformation and plant regeneration has already been established for the diploid strawberry *Fragaria vesca* L. using a binary vector pCAMBIA 1304 that contains hygromycin and GFP as selectable markers (Oosumi et al., 2006).

Based on the principle that monitoring the quantification of GFP fluorescence can accurately assess the concentration of genetically linked proteins of interest, GFP expression was used to monitor leaf canopy development in transgenic oilseed rape (*Brassica napus)* using the pSAM 12 plasmid containing genetically linked GFP and *Bacillus thuringiensis* (Bt) *cry1Ac* cassettes regulated by independent CaMV 35S

promoters to regulate transgene expression (Halfhill et al., 2003). Another example of the use of GFP fluorescence in plants has been in horseweed (*Conzya Canadensis*) to determine glyphosate resistant plants. Explants were infected with *Agrobacterium tumefaciens* strain GV3850 harboring the pBIN–mGFP5–ER plasmid that featured a cauliflower mosaic virus 35S promoter 5' to a GFP (mGFP5–ER) reporter gene and nos terminator including a kanamycin selection cassette within the T-DNA borders (Halfhill et al., 2007). In sugarcane, the use of GFP as a selectable marker facilitated the initial visual selection of transgenic sugarcane callus without antibiotics, herbicides or an assay. The binary vector used in transformation was pTO134, which contains a synthetic GFP gene (s*GFP*S65T) and the bialaphos resistance gene (*bar*) both controlled by CaMV 35S promoters (Elliott et al., 1998).

One of the practical considerations in developing an insertional mutant population is to isolate homozygous knockout mutants once unique hemizygous T_0 lines have been regenerated after transformation. There are several possibilities, including gene dosage related expression of GFP fluorescence in the T_1 generation, segregation of GFP expression in T_2 seedlings obtained from GFP positive T_1, or PCR-based screening of T_2 seedlings for the presence of both wild type and mutant alleles. Dosage related expression would certainly be the easiest method as it would identify T_1 homozygotes directly so that seed could be collected from only the high expressing T_1 plants with the expectation that their T_2 progeny would be homozygous. The other two methods require growing the T_2 generation, identification of the homozygous T_2, and collection of the seed from them to serve as the homozygous mutant stock. In the present study we have attempted to use all of these methods to determine their relative efficiency in strawberry.

Materials and methods

Plant material

The diploid *Fragaria vesca* PI 551572 (FV10) was transformed with either pCAMBIA 1302 (GFP) or 1304 (GFP/gus) using an *Agrobacterium*-mediated transformation protocol (Oosumi et al., 2006). During the transformation and regeneration process GFP expression was used to identify transformed callus colonies as

well as transgenic roots on plantlets prior to their acclimation to the greenhouse. A total of 700 T_0 mutant lines was grown at the Virginia Tech greenhouse and T_1 seeds were harvested. We have attempted to analyze a subset 240 T_0 lines to identify homozygous lines. A flow diagram has been outlined on Figure 1 to show the plant material that was used on this investigation.

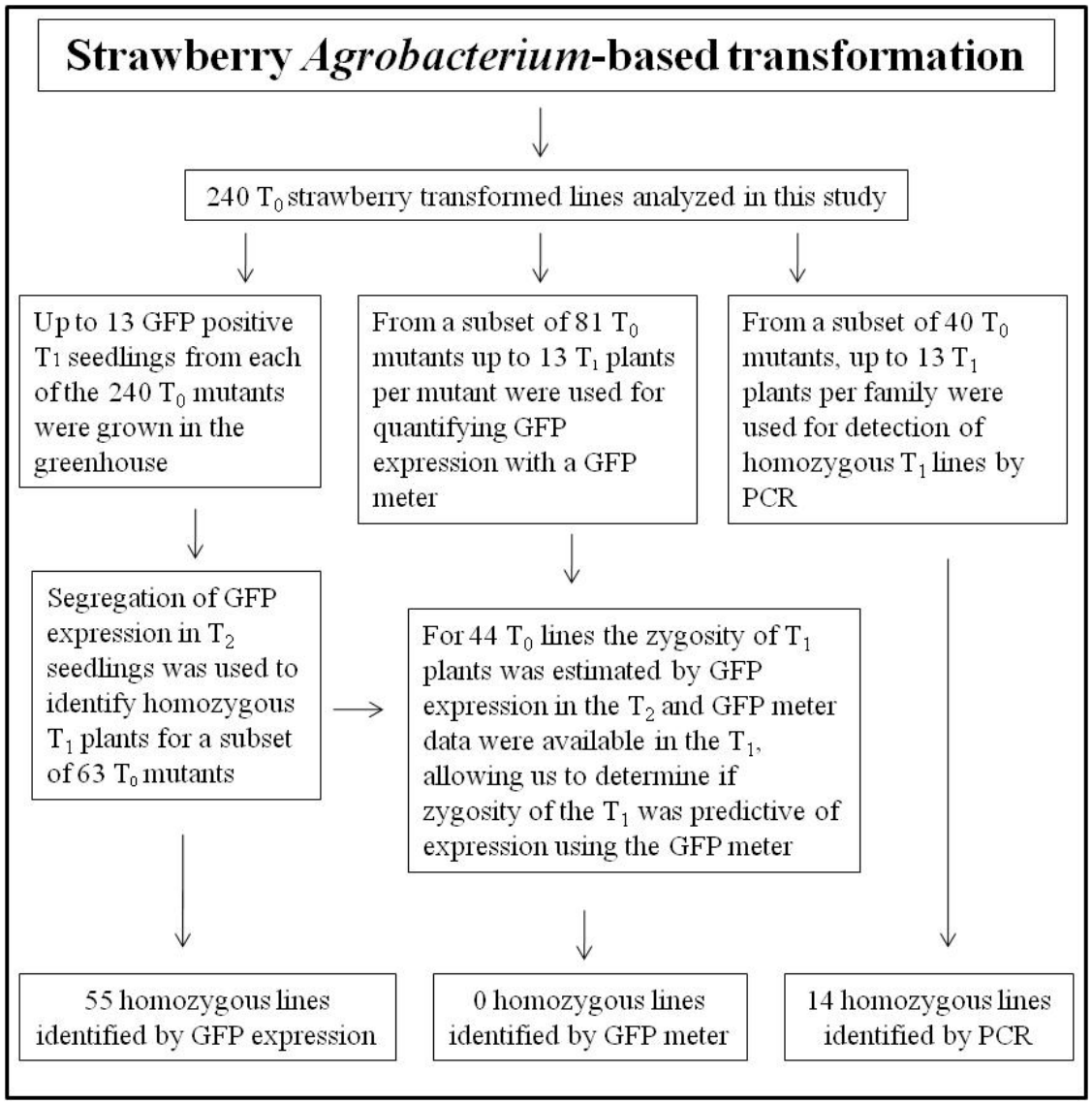


Figure 1: Flow diagram of T-DNA mutant lines from the diploid strawberry *Fragaria vesca* used in this investigation.

pCAMBIA 1304 and 1302

Around 500 strawberry T-DNA insertional lines have been generated using

Agrobacterium-mediated transformation protocol (Oosumi et al., 2006). A binary vector pCAMBIA-1304 and [GenBank accession: AF234300, Center for the Application of Molecular Biology to International Agriculture (CAMBIA), Black Mountain, Australia], containing mGFP5-gusA-His6 fusion, was introduced into two *A. tumefaciens* strains, GV3101 (pMP90) (Koncz, 1986) and LBA4404 carrying the ternary plasmid pBBR1MCS-5.virGN54D containing constitutive *virG* mutant gene (van der Fits et al., 2000). Additionally, more than 200 strawberry T-DNA insertional lines have been generated using *Agrobacterium* mediated transformation protocol (Oosumi et al., 2006) carrying a binary vector pCAMBIA-1302 (GenBank accession: AF234298).

Strawberry seed disinfection, germination, transplanting and growth conditions

Up to 50 seeds were counted out for each single T_1 and T_2 mutant lines. Surface sterilization was performed. Briefly, seeds were placed in a 1.5 ml eppendorf tube, soaked in 1 ml of 70% ethanol for 5 min. Seeds were drained and rinsed once with 1 ml autoclaved DI water, drained again and kept overnight in 1 ml autoclaved DI water. On the next day, seeds were drained and 1 ml 1% sodium hypochlorite was added for 15 min, shaking every 5 min, followed by three rinses with autoclaved DI water. The seeds were then soaked in 1 ml 50% filter sterilized preservative for plant tissue culture media (PPM), Plant Cell Technology for 15 min, shaking every 5 min and drained again. PPM may be reused after filter-sterilized up to 4 times. Finally using sterile forceps, we placed surface sterilized seeds on a petri dish containing blue blotter paper (227968-S0 19x24, S102070) Anchor Paper Company at Hudson, WI, for germination. Blue blotter paper was saturated with regular autoclaved water, and then plates were sealed with micropore tape and placed in the dark at room temperature for germination. Germination was scored every week up to 8 weeks after sowing. Seedlings that emerged were screened for GFP, transplanted to a growth chamber and finally to greenhouses according to Oosumi et al, (, 2006)

Detection of homozygous lines by transgene inheritance and segregation patterns of GFP

Developing T_2 seedlings were tested for GFP expression, using an Olympus

fluorescence microscope SZXZRFL3 (Olympus America, Melville, NY, USA) with a 100 W mercury burner (Olympus BH2-RFL-T3) and filter sets for GFP long pass, narrow-band GFP, and FITC/TRITC (filter sets no. 41018, 41020, and 52004V2, respectively, Chroma Technology Corp., Rockingham, VT, USA), and positive plants were transferred to autoclaved soil. Since strawberry germination is not synchronized, we conducted the screening for GFP positive plants for up to 8 weeks after the beginning of the experiment. Selected GFP positive seedlings were grown for 3 to 4 weeks in a growth chamber before they were transferred to a greenhouse.

Detection of homozygous lines by GFP meter

In addition to qualitative evaluation of GFP fluorescence in germinating seedlings; an effort to quantify GFP on well-developed plants was made on up to 13 T_1 GFP positive plants from each line. A commercially available GFP meter (Opti-Sciences, Tyngsboro, MA) was used, this instrument uses a fiber optic cable to emit and detect specific wavelengths of light (emit: 395nm/495nm, detect: 530nm) and records GFP expression in relative units called tics (Millwood et at. 2003). In an attempt to identify homozygous T_1 lines from hemizygous T_0 lines by relative fluorescence, we screened two different strawberry tissues: corolla and emerging roots from runners. Screening was conducted during hours of daylight at the Virginia Tech greenhouses.

Detection of homozygous lines by the polymerase chain reaction

We employed a strategy using three PCR primers in a single reaction on up to 13 GFP positive T_1 plants per line. Previously conducted hi-TAIL-PCR had provided us with T-DNA flanking sites that gave unique hits for each mutant within the scaffold assembly of the strawberry genome (www.strawberrygenome.org) (Shulaev, et al. unpublished), allowing us to predict the nucleotide sequence expected on the opposite side of the T-DNA from which the hi-TAIL product had been determined. Our three-primer reaction then included two from the scaffold expected to straddle the T-DNA insertion and another from within the T-DNA. The scaffold primers were designed to yield a unique size wild type product that could be easily differentiated from the mutant product expected from amplification between a T-DNA-based primer 1304RBF (5’

TATGATTAGAGTCCCGCAAT 3') common to all of the mutants and one of the scaffold primers listed on supplemental Table 1. A long product between flanking primers would not be obtained if T-DNA intervened under the PCR conditions used. Therefore, DNA extracted from homozygous mutant or homozygous wild type plants would yield single products of obviously different size and that of hemizygous (wild type and mutant allele) lines would yield both products. Flow diagram has been outlined in Figure 2. This idea was adapted and modified to our conditions from Ahern et al. 2009 (Ahern et al., 2009). PCR primers were designed according to Ruiz-Rojas et al (, 2010). PCR was performed on genomic DNA of 29 T-DNA mutants in a final volume of 25 µl following the touchdown protocol of Sargent et al. (2003) with an annealing temperature from 55 to 50ºC (Sargent et al., 2003), in a iCycler^{TM} Thermal Cycler – BioRad. Bands were separated through a 1.2% agarose gel (Promega) containing 0.1% ethidium bromide and run at 150 V, 400 amps, for 60 min. PCR Products were visualized over UV light.

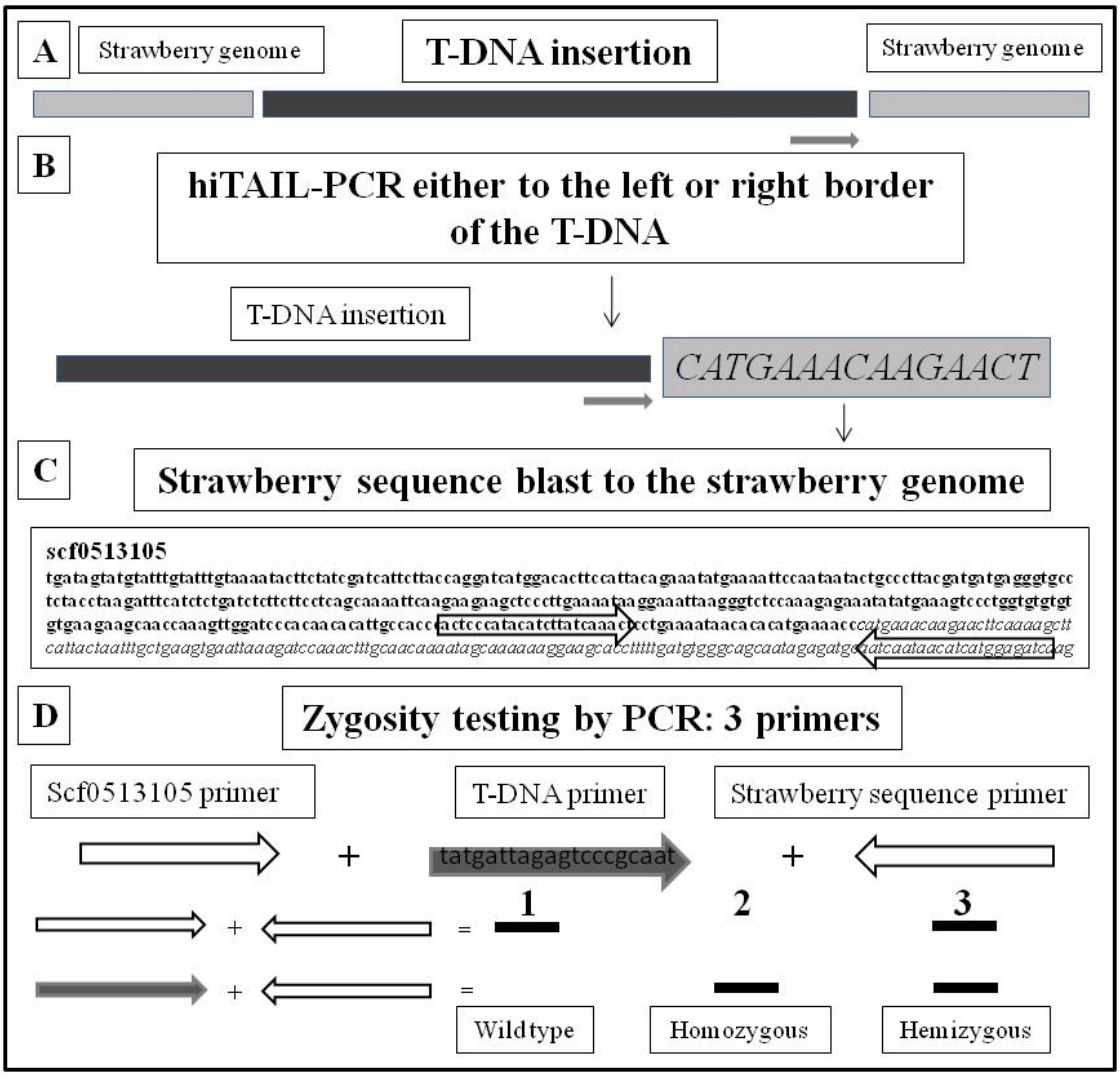


Figure 2: Flow diagram of zygosity testing developed on T-DNA mutant lines from the diploid strawberry *Fragaria vesca*; **A)** T-DNA integration into the strawberry genome by *Agrabacterium*-based transformation; **B)** Strawberry flanking sequences were isolated by hiTAIL-PCR either from the right or left border of the T-DNA; **C)** Flanking sequences were blasted against the strawberry genome and primers were designed from both the scaffold (bold lower case letters) and the flanking sequence (italic lower case letters) as shown by the arrows; **D)** Wild type and homozygous plants were differentiated by presence or absence of a product amplified from both flanking primers (wild type) or from the T-DNA primer and one flanking primer (mutant). The 128 bp sequence between the T-DNA right border and the T-DNA primer was taken into consideration when selecting the left primer from strawberry flanking sequence to ensure that the difference between wild types and homozygous PCR bands was sufficient for easy visual identification. Hemizygous plants would be expected to exhibit both bands.

Results

Detection of homozygous lines by transgene inheritance and segregation patterns of GFP

So far we have surface sterilized and processed 8,075 T_1 seeds from 240 independent T_0 lines. Germination was scored weekly and as seedlings were large enough (approx. 1 cm) GFP expression was observed under UV light and sprouted seedlings were transferred to growth chambers. Out of 8,075 seeds we observed 68.7% germination (5,551 seeds germinated within 8 weeks). Figure 3 shows percentage germination and GFP expression per week.

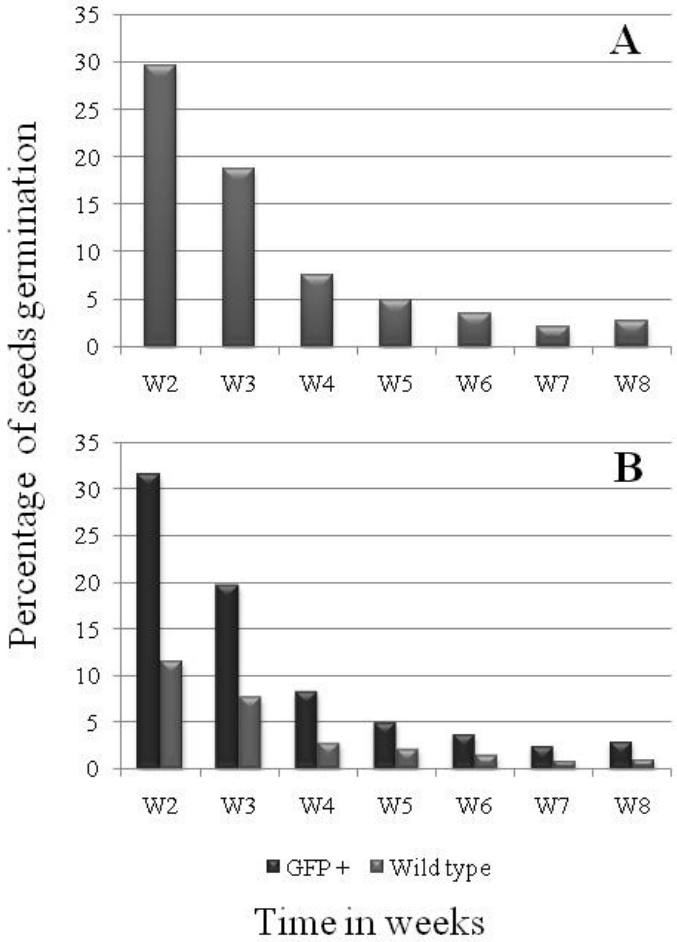


Figure 3: Seed germination and GFP segregation in the diploid strawberry *Fragaria vesca*: **A)** percentage of seeds germinated each week between 2 and 8 weeks after sowing, and **B)** percentage as in A of both GFP positive and wild type plants each week. The bars represent combined data for 240 mutant lines.

By week 3, 48% of the seeds had germinated with decreasing germination over the next few weeks (Figure 3A). The frequency of GFP positive and wild type plants per week showed similar germination trends (Figure 3B), with 51% of GFP positive plants germinated and scored 3 weeks after surface sterilization.

Of 240 T_1 lines analyzed 83% exhibited a 3:1 segregation (GFP positive:wild type) according to χ^2 (Table 1). Nineteen additional lines exhibited segregation not significantly different from 15:1. Other unexpected segregation patterns (1:1, 0:1) were observed as well. A minimum of 13 GFP positive T_1 plants was required to have 99% confidence of finding at least one homozygous line, given the expectation of 3:1 segregation GFP:wild type. A subset of 104 T_1 lines, where sufficient germination (>13 seedlings per line) was obtained to allow confident χ^2 analysis, was used to correlate GFP segregation by transgene inheritance and GFP expression measured on a GFP meter.

Table 1: GFP segregation ratios of 240 insertional mutant lines of the diploid strawberry *Fragaria vesca*, for expected Mendelian segregation with one (3:1) and two (15:1) T-DNA insertions. Other unexpected ratios are also shown.

Expected segregation	Number of T_1 lines	Percentage	Subset of T_1 plants evaluated	Percentage
Likely 3:1	198*****	83	92*	88
Likely 15:1	19	8	3	3
Likely 1:1	10	4	4	4
Likely 0:1	8	3	3	3
Others	5	2	2	2
n=	**240**		**104**	

*$P \geq 0.05$

We collected T_2 seed from 240 T_1 lines (up to 13 T_1 lines per T_0 insertional mutant) and germinated up to 50 T_2 seeds per T_1 line to identify homozygous T_1 plants that did not segregate for GFP expression in the T_2. We germinated 14,446 T_2 seeds representing 63 T_1 lines (up to 13 T_2 lines per T_1 line) to observe GFP segregation in the T_2. A total of 154 T_2 homozygous plants have been identified among the progeny of 63 T_1 lines; however a range of 0 to 9 homozygous plants have been observed within lines, as a result 55 independent homozygous mutants have been identified by transgene inheritance and segregation patterns of GFP (Figure 4), there were 8 T_1 lines where we were unable to identify a homozygous line by GFP segregation.

Figure 4: GFP expression under UV (left) and tungsten (right) light of T_2 seedlings from T-DNA insertional mutants of the diploid strawberry, *Fragaria vesca.* The P-CAMBIA T-DNA construct carried mGFP5 driven by a p35S promoter. Segregation of GFP positive (white arrow) and negative (red arrow) seedlings can be observed.

Detection of homozygous lines by GFP meter

Before taking GFP meter readings at the greenhouse, GFP expression was observed under UV light on different plant tissues. Tissue like emerging roots from runners, corolla, receptacle, and fruits expressed GFP fluorescence (Figure 5). To attempt to identify T_1 homozygous lines in the greenhouse after we confirmed that GFP expression was present, we used the GFP meter to take 3 readings both from corolla and emerging roots from runners from each plant. However, there were not enough data to analyze the corolla readings. We selected a subset of 736 T_1 lines representing 81 T_0 mutants (Figure 1). All 736 T_1 plants were screened with the GFP meter, with the expectation that higher GFP (tic) values would found for homozygous lines, and medium-low values for hemizygous lines as observed in tobacco seedlings (Molinier et al., 2000). Supplemental Figure 1 shows error bars representing the standard deviation of the mean of 3 emerging roots from runners on the same plant. Tic values have been arranged from high to low, but the high standard errors indicate high variation for GFP measurements within plants. In addition, we tried to correlate GFP meter readings with GFP segregation by transgene inheritance and segregation patterns. A subset of 44 mutant lines with complete GFP meter readings was analyzed with its corresponding T_2 segregation data. This gave us a population of 215 T_1 plants with both GFP meter data and zygosity estimation by GFP segregation in T_2 progeny. Figure 6 shows the frequency

of homozygous and hemizygous plants with the corresponding GFP meter reading (tic) from the T_1 subset. Most of the plants that were identified as homozygous by GFP segregation in the T_2 were found among the population with the lowest GFP meter readings (Figure 6). Therefore, the GFP meter readings were not predictive of the T_1 zygosity.

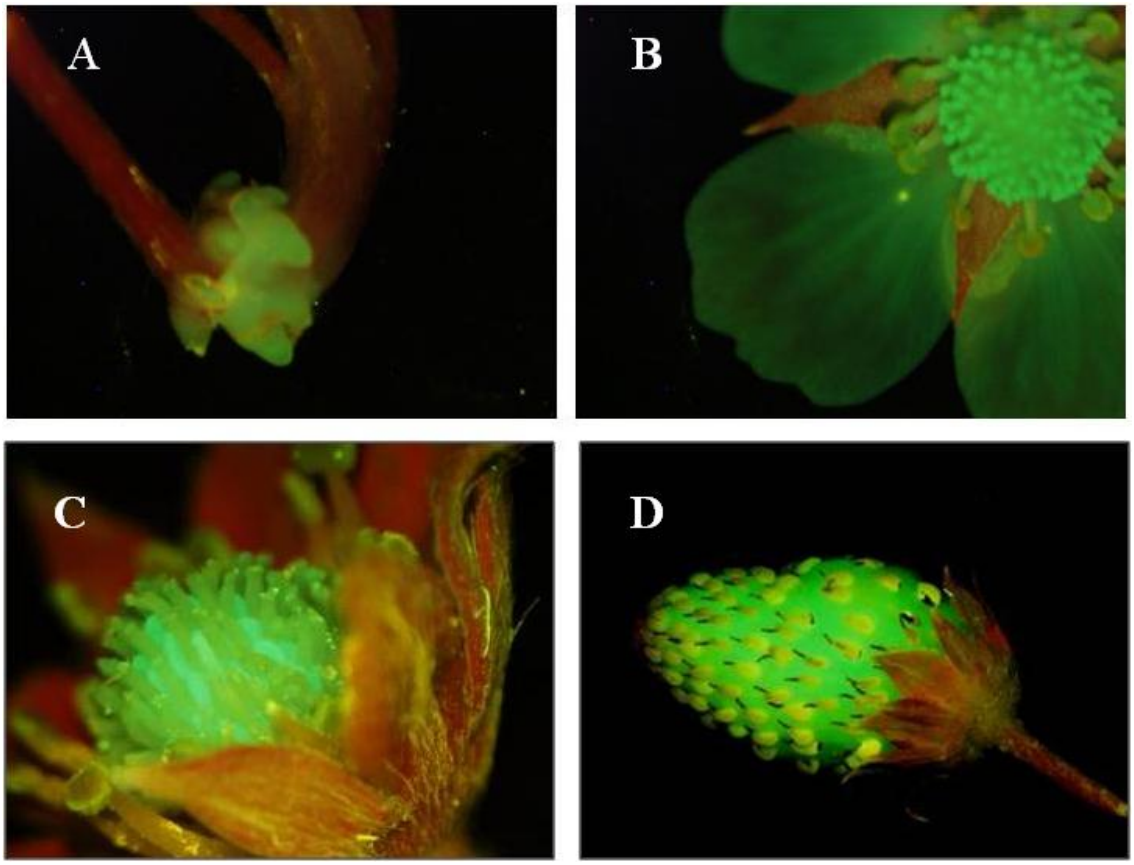


Figure 5: Images of T_1 plants of T-DNA insertional mutants of the diploid strawberry *Fragaria vesca* expressing *mGFP5* driven by a p35S promoter: **A)** emerging roots from a runner; **B)** corolla with stigmatic tissue; **C)** receptacle in an immature fruit; and **D)** mature fruit.

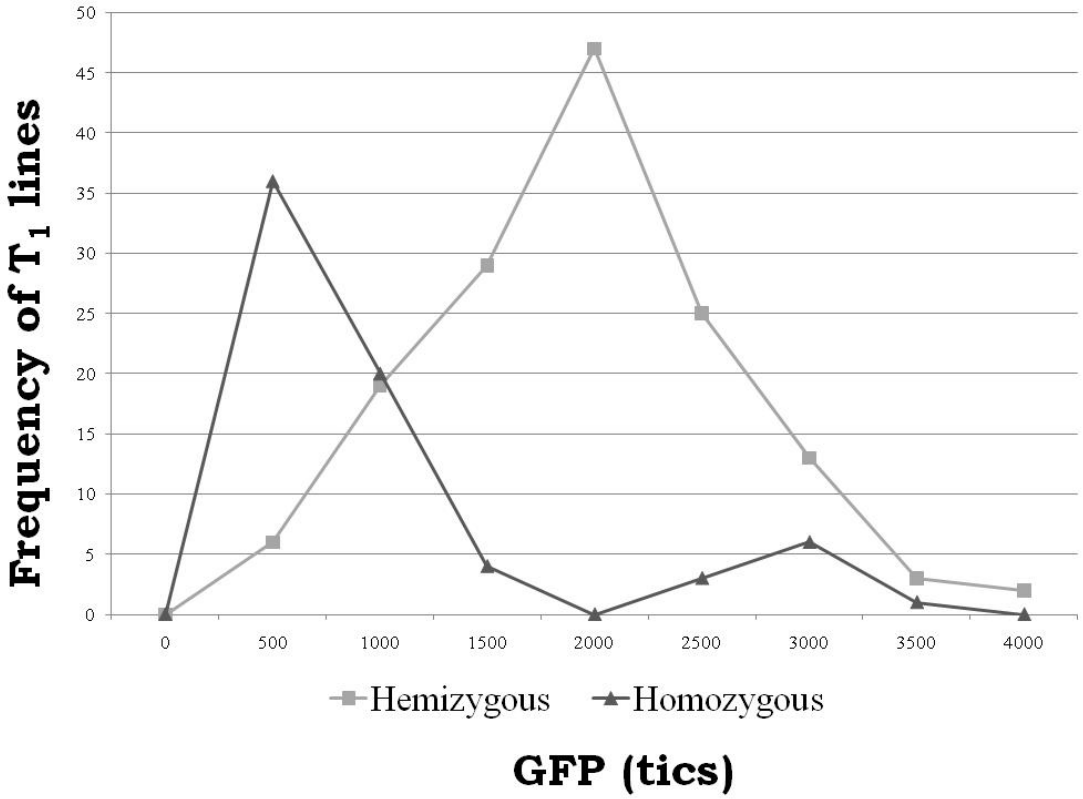


Figure 6: Range of GFP expression measured in arbitrary units (tics) using a GFP meter on developing roots from runners among 215 T_1 plants derived from 44 independent T_0 insertional mutants of *Fragaria vesca.* Each of the T_1 plants was determined to be homozygous or hemizygous for the T-DNA insertion by observing segregation of GFP expression in up to 13 T_2 seedlings per T_1 plant.

Detection of homozygous lines by the polymerase chain reaction

DNA was extracted on up to 13 T_1 GFP positive plants for each of 29 different T_0 insertional mutants. Zygosite primers as shown in Figure 2 were used to attempt to identify homozygous lines. The advantage of this technique is that homozygous lines can be identified directly in the T_1 stage, and homozygous T_2 seed collected only from the plants identified with all of the hemizygous T_1 plants discarded. A list of all the expected and observed band sizes is shown in supplemental Table 1 as well as mutant genbank ID. Figure 7 shows mutant VT_730, genbank ID HM012636 with 2 homozygous lines easily scored in lanes 8 and 13. We struggled trying to design good primers for some mutants from the flanking sequences due to the fact that the T-DNA insertion was in a highly repetitive region. In several cases the primers did not amplify the expected bands. From

29 different lines we were only able to identify 14 homozygous lines by PCR using this approach.

Figure 7: Agarose gel electrophoresis of PCR products of wild type strawberry, a T_0 insertional mutant line and 13 T_1 plants after self-pollination using zygosity primers as described in Figure 2. From left to right, the lanes are: **L)** 1000 bp marker, **1)** FV10 wild type strawberry, **2)** T_0 Mutant VT_730, **3 – 15)** T_1 GFP positive plants. The upper band derives from the flanking sequence primers and indicates one or two wild type alleles (790 bp), the lower band derives from a T-DNA primer and a primer from the right flanking sequence and indicates one or two mutant alleles (640 bp). The plants represented in lanes 8 and 13 would be homozygous mutant lines selected for perpetuation of the mutant line.

Discussion:

In this study we were able to process 240 T_0 lines (5551 seeds) and 63 T_2 lines (14446 seeds) from the diploid strawberry *Fragaria vesca* (flow diagram 1). We observed germination capacity of 68.7% from the 240 T_1 lines, having 48% germination in the first 3 weeks as shown on Figure 3A. There is not much documentation about seed germination on the diploid strawberry, and more importantly on a population of insertional mutants, however there is evidence that atypical ecotypes of *Fragaria vesca* from West Siberia have shown germination after the 5th day of planting and at 21 days germinations capacity was from 99% to 59%, depending on the ecotype (Baturin, 2009). Our results from the diploid *Fragaria vesca* PI 551572 (FV10) are within their range. Additionally, a study conducted at the greenhouse level on an inbred line, YW5AF7, of a diploid strawberry *Fragaria vesca* f. *semperflorens* line called "Yellow Wonder", shows that more uniform germination can be achieved by cold treatment of moist seed. By

doing that for 3 weeks, germination capacity from 87% in the first 3 weeks was increased to 91% in the first 2 weeks (Slovin et al., 2009). Looking at Figure 3B, germination from T-DNA insertional population seems to not be affected by transgenesis compared to the distribution of wild type, and germination from the all T_2 segregation which have same tendency as the T_1 population (results not shown). However a more detailed study will be necessary to support this consideration.

Detection of homozygous lines

Our GFP marker segregation analysis from the 240 T_1 lines showed that the majority of our insertional population (83%) has a 3:1 Mendelian segregation ($p>0.05$) as expected for a single insertion, and 8% of them seem to be 15:1 most likely because of the double integration of T-DNA (Table 1). In Arabidopsis, out of 1000 putative mutants that were identified, 80% showed a 3:1 or a 15:1 segregation of the T-DNA-encoded kanamycin resistance marker (Koncz et al., 1992). A low number of transgenic lines (9%) showed deviation from the expected segregation ratio difficult to explain. In rice, more than 100 independent lines were generated using the Cry1AB gene, expected Mendelian segregation was observed on 2 of the 3 lines analyzed, and deviation from the expected segregation was also observed in a small number of lines (Kim et al., 2008). Homozygous lines were really easy to identify by GFP over UV light, up to 50 seeds were screened and as shown on Figure 3, if 100% expression was observed, homozygous lines were found. So far, we have found 55 homozygous lines by GFP expression over UV light. GFP as a visual selectable marker has been successfully and widely used in different crops (Leclercq et al., 2010; Mishra et al., 2010; Palumbo and Veilleux, 2007; Shimada et al., 2010; Stewart, 2005; Tan et al., 2009), however harvesting seeds from the T_1 lines, and going through all surface sterilization process and screening for GFP takes a lot of time and high labor. For those 8 lines were homozygous lines have not been found yet, T_2 GFP positive plants were grown again and their progeny are undergoing. All information and a genetic characterization of our insertional mutant population as well as availability of homozygous line seeds will be publicly available at our strawberry database constructed on FileMaker pro 10 (http://www.strawberry/hort.vt.edu).

Identification of homozygous mutant lines directly from T_1 plants at the greenhouse level using a GFP meter was not possible in this investigation. GFP meter has been successfully used before in crops like tobacco, canola and potato were low, medium and high GFP meter were correlated with wild type, hemizygous and homozygous lines respectively (Millwood et al., 2003; Palumbo and Veilleux, 2007). Our data shows no correlation at all between GFP readings taken over UV light from the T_2 segregating population and GFP meter taken from its corresponding T_1 parental lines. We have also grouped all our GFP meter readings in a single Figure (supplemental Figure 1) to try to see the range of GFP quantification, however high standard deviation from the mean was observed among almost all population. Our correlation between the T_2 GFP data and T_1 GFP meter data shows that low GFP meter data are highly grouped between 500 tics were almost all the homozygous T_2 were located (Figure 6) In addition, a wide distribution of the frequencies of hemizygous was observed from 500 to 3000 tics, making impossible to detect on early stages homozygous lines for the diploid strawberry *Fragaria vesca.*

Our final approach of finding homozygous lines was done by PCR reaction; we successfully identified homozygotes for 14 T_1 lines. As shown in our flow diagram (Figure 2), and Figure 7, homozygous lines will be detected only if a single T-DNA insert is present, other ways, primers designated from strawberry genome will result on wild type plants among the population under investigation. Our 3 different attempts of homozygous detection lead us to conclude that GFP meter was not successful in this investigation, PCR works; however our low efficiency of 48.2% makes it difficult and costly to keep doing it. Right now we are screening for homozygous lines by GFP under the UV light.

Acknowledgments: This work was supported by the USDA/NRI 2008-02284, Hatch VA-135816, the East Malling Trust for Horticulture Research, the BBSRC, and a Virginia Tech ASPIRES Grant. We also thank Dr. Guillaume Pilot from the Department of Plant Pathology, Physiology and Weed Science at Virginia Tech, for his kind collaboration on the statistical analysis.

References

Afolabi A.S., Worland B., Snape J., Vain P. (2005) Novel pGreen/pSoup dual-binary vector system in multiple T-DNA co-cultivation as a method of producing marker-free (clean gene) transgenic rice (*Oriza sativa* L) plant. Afr. J. Biotechnol. 4:531-540.

Ahern K.R., Deewatthanawong P., Schares J., Muszynski M., Weeks R., Vollbrecht E., Duvick J., Brendel V.P., Brutnell T.P. (2009) Regional mutagenesis using Dissociation in maize. Methods 49:248-254.

Baturin S.O. (2009) Seed germination of *Fragaria vesca* L. from atypical ecotopes of West Siberia. Contemp. Probl. Ecol. 2:556-559.

Elliott A.R., Campbell J.A., Brettell R.I.S., Grof C.P.L. (1998) *Agrobacterium*-mediated transformation of sugarcane using GFP as a screenable marker. Aust. J. Plant Physiol. 25:739-743.

Halfhill M.D., Millwood R.J., Rufty T.W., Weissinger A.K., Stewart C.N. (2003) Spatial and temporal patterns of green fluorescent protein (GFP) fluorescence during leaf canopy development in transgenic oilseed rape, *Brassica napus* L. Plant Cell Rep. 22:338-343.

Halfhill M.D., Good L., Basu C., Burris J., Main C., Mueller T., Stewart C. (2007) Transformation and segregation of GFP fluorescence and glyphosate resistance in horseweed (*Conyza canadensis*) hybrids. Plant Cell Rep. 26:303-311.

Hicks B. (2009) Viral applications of green fluorescent protein: methods and protocols New York: Humana.

Kim S., Kim C., Li W., Kim T., Li Y., Zaidi M.A., Altosaar I. (2008) Inheritance and field performance of transgenic Korean *Bt* rice lines resistant to rice yellow stem borer. Euphytica 164:829-839.

Koncz C., Nemeth K., Redei G.P., Schell J. (1992) T-DNA insertional mutagenesis in *Arabidopsis*. Plant Mol. Biol. 20:963-976.

Koncz C.a.S., J. (1986) The promoter of TL-DNA gene 5 controls the tissue-specific expression of chimaeric genes carried by a novel type of *Agrobacterium* binary vector. Mol. Gen. Genet. 204:383-396.

Leclercq J., Lardet L., Martin F., Chapuset T., Oliver G., Montoro P. (2010) The green fluorescent protein as an efficient selection marker for *Agrobacterium tumefaciens*-mediated transformation in *Hevea brasiliensis* (Müll. Arg). Plant Cell Rep. 29:513-522.

Millwood R.J., Halfhill M.D., Harkins D., Russotti R., Stewart C.N. (2003) Instrumentation and methodology for quantifying GFP fluorescence in intact plant organs. Biotechniques 34:638-643.

Mishra M., Devi S., McCormac A., Scott N., Chen D., Elliott M., Slater A. (2010) Green fluorescent protein as a visual selection marker for coffee transformation. Biologia 65:639-646.

Molinier J., Himber C., Hahne G. (2000) Use of green fluorescent protein for detection of transformed shoots and homozygous offspring. Plant Cell Rep. 19:219-223.

Oosumi T., Gruszewski H.A., Blischak L.A., Baxter A.J., Wadl P.A., Shuman J.L., Veilleux R.E., Shulaev V. (2006) High-efficiency transformation of the diploid strawberry (*Fragaria vesca*) for functional genomics. Planta 223:1219-30.

Palumbo R., Veilleux R. (2007) Variation of GFP expression in diploid and tetraploid transgenic potato. Am. J. Potato Res. 84:393-401.

Ruiz-Rojas J., Sargent D., Shulaev V., Dickerman A., Pattison J., Holt S., Ciordia A., Veilleux R. (2010) SNP discovery and genetic mapping of T-DNA insertional mutants in *Fragaria vesca*; L. Theor. Appl. Genet. 121:449-463.

Sargent D.J., Hadonou A.M., Simpson D.W. (2003) Development and characterization of polymorphic microsatellite markers from *Fragaria viridis*, a wild diploid strawberry. Mol. Ecol. Notes 3:550-552.

Shimada T.L., Shimada T., Hara-Nishimura I. (2010) A rapid and non-destructive screenable marker, FAST, for identifying transformed seeds of *Arabidopsis thaliana*. Plant J. 61:519-528.

Shimomura O., Johnson F.H., Saiga Y. (1962) Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, *Aequorea*. J. Cell. Comp. Physiol. 59:223-239.

Slovin J., Schmitt K., Folta K. (2009) An inbred line of the diploid strawberry *Fragaria vesca* f. semperflorens for genomic and molecular genetic studies in the Rosaceae. Plant Methods 5:1-10.

Stewart C.N. (2001) The utility of green fluorescent protein in transgenic plants. Plant Cell Rep. 20:376-382.

Stewart C.N. (2005) Monitoring the presence and expression of transgenes in living plants. Trends Plant Sci. 10:390-396.

Tan B., Li D.-L., Xu S.-X., Fan G.-E., Fan J., Guo W.-W. (2009) Highly efficient transformation of the *GFP* and *MAC12.2* genes into precocious trifoliate orange (*Poncirus trifoliata* [L.] Raf), a potential model genotype for functional genomics studies in *Citrus*. Tree Genet. Genom. 5:529-537.

van der Fits L., Deakin E.A., Hoge J.H., Memelink J. (2000) The ternary transformation system: constitutive *virG* on a compatible plasmid dramatically increases *Agrobacterium*-mediated plant transformation. Plant Mol. Biol. 43:495-502.

Supplemental material

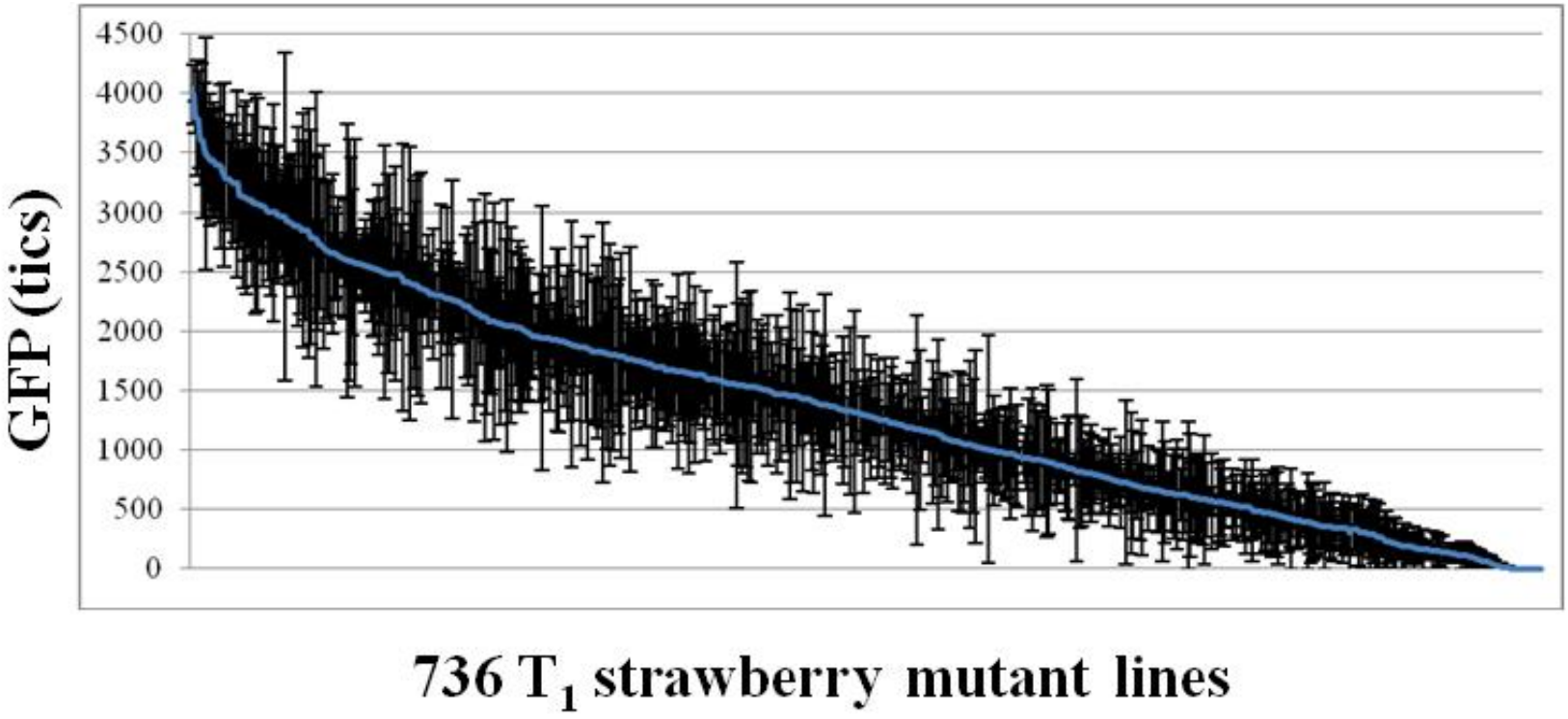
Supplemental Table 1: Insertional mutants, their GenBank identity and primers used for zygosity testing in the diploid strawberry *Fragaria vesca*. A total of 29 T_1 lines was analyzed. The expected and observed band size are provided.

Mutant name	***GenBank ID***	***Primer name***	***Primer sequence 5′ - 3′***	***Expected band size (bp)***		***Observed band size (bp)***	
				Mutant	**WT**	**Mutant**	**WT**
pCAMBIA T-DNA	AF234300	**1304RBF***	**TATGATTAGAGTCCCGCAAT**				
FV10-C02-678 (J1)	HQ185080	678R	tcataaacgccgacgtacgt	382	537	382	537
		Zyg678F	caagtaagcaacgtaccttccg				
FV10-C02-679 (J2)	HM012652	Zyg679R	acatctgggttcaagccgat	580	519	580	519
		Zyg679F	tagatcaggatggcattcaggg				
FV10-C02-682 (J4)	HM012651	zVT682R	aggaagcttcgtacttggttcat	429	564	429	564
		zVT682F	gaagctcttagctggaagttgtgg				
FV10-C02-688 (J9)	HQ185081	688R	ttctccgaactcgccgaaatca	305	787	305	787
		Zyg688	ccacccacaaacagaacaaactcc				
FV10-C02-691 (J11)	HM012649	zVT691R	ccaaccccaagctctcagatat	828	988	828	988
		zVT691F	atccatataagcctgtcccagc				
FV10-C02-696 (J16)	HQ185083	zyg 696R	ttagccttgccagtgttagtcg	592	746	700	746
		zyg 696F	accaaccaccaagaacacaagg				
FV10-C02-710 (K4)	HM012646	zVT710R	tgggtttttccggtctcctt	291	199	291	199
		zVT710F	tgccatccctagtctagaagaacc				
FV10-C02-712 (K6)	HM012644	zVT712R	gctttgcatgagaggtctccat	509	589	400	589
		zVT712F	ctccaacaatgtgcagaagagc				
FV10-C02-714 (K8)	HM012643	714_R	ggagaggaaaccgatcaagcaatg	543	831	700	831
		Zyg714	tgtccagtcatcccttcaaacc				
FV10-C02-721 (K14)	HM012640	zVT721R	ccatggcttgtgcttctgaa	672	851	672	851
		zVT721F	ccagttgcaatgacttcttggc				
FV10-C02-723 (K16)	HM012639	723_AC1_R	gatccagtcagaaaggcctaactac	600	713	650	713

		Zyg723	ataaatacagatatccacaaagctctgagg				
FV10-C02-725 (K18)	HM012638	725_AC1_R	ccatctttcttcccttcagcaaa	800	1088	800	1088
		Zyg725	ccagactggtaaggttgtatgtgagg				
FV10-C02-729 (K21)	HQ185100	zVT729R	gcagtgaagatagagctggagcaa	558	648	558	648
		zVT729F	aagctccaagatcatgtgctcc				
FV10-C02-730 (K22)	HM012636	zVT730R	cgaggaagaggacgttgacatt	682	783	682	783
		zVT730F	attaacctcggactccatcagc				
FV10-C02-650 (I2)	HM012665	VT650ZYG	aatgtttggacaccaaatagaataggg	516	738	Failed	Failed
		650R	gcttttgtacgagaagatgctca				
FV10-C02-655 (I5)	HM012664	VT655ZYG	tttggtgagtgttaaggccctc	826	515	Failed	Failed
		655R	gagacccttcttcttgatgatgaa				
FV10-C02-656 (I6)	HM012663	VT656ZYG	gcgattacaagtgaagatgtgg	798	432	Failed	Failed
		656R	ccactttctcaatttctagt				
FV10-C02-669 (I17)	HM012658	VT669ZYG	ttctgattgcagctggactctacg	779	388	Failed	Failed
		669R	ggacatggtctgagaaagtggct				
FV10-C02-684 (J5)	HM012650	zVT684F	aattgcccatctaacccactcc	446	590	Failed	Failed
		zVT684R	tgctaatcgcttacctcctgttt				
FV10-C02-695 (J15)	HQ185082	Zyg695	cttgctttgaagacgtggttgg	300	495	Failed	Failed
		695_AC1_R	gccatgacgttatttatgagatggg				
FV10-C02-698 (J18)	HQ185084	zyg 698F	cttcatggacttccgtcctaacag	422	851	Failed	Failed
		zyg698R	actccaaccacgggttcaagta				
FV10-C02-701 (J21)	HM012648	zVT701F	tgagaaggttggactaggagatgg	687	945	Failed	Failed
		zVT701R	ccctttcacaccgatcagtgat				
FV10-C02-706 (J25)	HQ185085	zyg706F	aaaccgtgccatgaaactcg	766	1196	Failed	Failed
		zyg 706R	cgggtgagaactgagtatggagtt				
FV10-C02-708 (K2)	HQ185086	Zyg708	aaatatgttggggtctgcgagc	550	705	Failed	Failed
		708_AC1_R	actctggaacatttccgttacttg				
FV10-C02-709 (K3)	HQ185087	Zyg709	atgtcaaaacccactgaccagc	550	809	Failed	Failed
		709_AC1_R	ttatcctgcgctggtgtccc				

FV10-C02-719 (K12)	HQ185129	Zyg719	tgagtatacagttggtggcagagg	750	858	Failed	Failed
		VT719R	ccgagagggacaatgtgcttat				
FV10-C02-732 (K23)	HQ185127	Zyg732	gccatgtgtgtgactgatcttagc	650	940	Failed	Failed
		732_AC1_R	agaagttgcccctctagaat				
FV10-C02-733 (K24)	HQ185114	Zyg733	ataacgtggtcctgcaattgcc	550	821	Failed	Failed
		733_LBAllR	acgacccgccttataaatgcttc				
FV10-C02-734 (K25)	HM012635	zVT734F	tataatggatcacaaacctgcagg	646	713	Failed	Failed
		zVT734R	cacatctccggctgtcttcatt				

*** T-DNA primer 1304RBF was common to all 29 PCR reactions**



Supplemental Figure 1: Range of GFP expression among 736 T_1 strawberry plants derived from 81 T_0 insertional mutants of the diploid strawberry, *Fragaria vesca*. Each seedling was selected for GFP expressing under UV light prior to greenhouse establishment. GFP expression was measured in arbitrary units (tics) using a GFP meter. Seedlings are arranged from highest to lowest GFP expression. Error bars represent the standard deviation of the mean of three emerging roots from runners on the same plant.

Chapter 3:

SNP discovery and genetic mapping of T-DNA insertional mutants in *Fragaria vesca* L.

J.J. Ruiz-Rojas[1], D.J. Sargent[2], V. Shulaev [1,4], A. W. Dickerman[4], J. Pattison[3], S. H. Holt[1], A. Ciordia[2] and R.E. Veilleux[1](✉)

Author contribution:

Dr Daniel J. Sargent at the East Malling Research Institute, UK has developed the most complete genetic mapping of the diploid *Fragaria vesca* L. As a part of my program, I spent 3 months at East Malling Research, working under Dr. Sargent's supervision. We conducted the scoring and genetic mapping of the SNP markers developed at Virginia Tech. Dr. Jeremy Pattison and I initiated searching for SNP markers between the parents of the mapping population (*Fragaria vesca* × *Fragaria bucharica*), and converted them to CAPS/dCAPS-based markers at the Virginia Tech Southern Piedmont Agricultural Research and Extension Center, located in Blackstone Virginia. Sarah Holt is a Ph.D. candidate in Dr Veilleux's laboratory; she adjusted the hiTAIL-PCR protocol and accommodated it to the specific equipment in our laboratory. She also contributed to the maintenance and cataloguing of the mutant collection. Anna Ciordia was a technician working at the East Malling Research UK; she and Dr. Daniel Sargent scored and mapped the first 11 markers developed at Virginia Tech. Dr Allan W. Dickerman contributed to the bioinformatic analysis and gene annotation of the flanking sequences generated in this investigation. Dr Richard Veilleux and Dr Vladimir Shulaev are the principal investigators on the strawberry project; they directed the investigation and edited the manuscript. Finally, I contributed to all topics related to this publication; it was the main subject of my research and I wrote the first draft of this paper.

[1]Department of Horticulture, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA
[2]East Malling Research (EMR), New Road, East Malling, Kent ME19 6BJ, UK
[3]Department of Horticultural Science, North Carolina State University, 600 Laureate Way, Kannapolis, NC 28081, USA
[4]Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

Theor Appl Genet (2010) 121:449–463
DOI 10.1007/s00122-010-1322-9

ORIGINAL PAPER

SNP discovery and genetic mapping of T-DNA insertional mutants in *Fragaria vesca* L.

J. J. Ruiz-Rojas · D. J. Sargent · V. Shulaev · A. W. Dickerman · J. Pattison · S. H. Holt · A. Ciordia · Richard E. Veilleux

Received: 20 November 2009 / Accepted: 5 March 2010 / Published online: 27 March 2010

Abstract As part of a program to develop forward and reverse genetics platforms in the diploid strawberry [*Fragaria vesca* L.; ($2n = 2x = 14$)] we have generated insertional mutant lines by T-DNA mutagenesis using pCAMBIA vectors. To characterize the T-DNA insertion sites of a population of 108 unique single copy mutants, we utilized thermal asymmetric interlaced PCR (hiTAIL-PCR) to amplify the flanking region surrounding either the left or right border of the T-DNA. Bioinformatics analysis of flanking sequences revealed little preference for insertion site with regard to G/C content; left borders tended to retain more of the plasmid backbone than right borders. Primers were developed from *F. vesca* flanking sequences to attempt to amplify products from both parents of the reference *F. vesca* 815 × *F. bucharica* 601 mapping population. Polymorphism occurred as: presence/absence of an amplification product for 16 primer pairs and different size products for 12 primer pairs, For 46 mutants, where polymorphism was not found by PCR, the amplification products were sequenced to reveal SNP polymorphism. A cleaved amplified polymorphic sequence/derived cleaved amplified polymorphism sequence (CAPS/dCAPS) strategy was then applied to find restriction endonuclease recognition sites in one of the parental lines to map the SNP position of 74 of the T-DNA insertion lines. BLAST search of flanking regions against GenBank revealed that 46 of 108 flanking sequences were close to presumed strawberry genes related to annotated genes from other plants.

Communicated by A. Schulman.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-010-1322-9) contains supplementary material, which is available to authorized users.

J. J. Ruiz-Rojas · V. Shulaev · S. H. Holt · R. E. Veilleux (✉)
Department of Horticulture, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA
e-mail: potato@vt.edu

D. J. Sargent · A. Ciordia
East Malling Research (EMR), New Road, East Malling, Kent ME19 6BJ, UK

J. Pattison
Department of Horticultural Science, North Carolina State University, 600 Laureate Way, Kannapolis, NC 28081, USA

V. Shulaev · A. W. Dickerman
Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

Introduction

The application of molecular techniques for genetic improvement of most rosaceous crops is limited by large genome size and complexity (Shulaev et al. 2008). The alpine strawberry, *Fragaria vesca* L., however, has several advantages for genomic investigation. *F. vesca* is the most closely related diploid strawberry species ($2n = 2x = 14$) to the cultivated octoploid strawberry, *Fragaria* × *ananassa* Duch. ($2n = 8x = 56$) (Potter et al. 2000; Rousseau-Gueutin et al. 2008), has a small genome estimated at 206 Mbp (Folta and Davis 2006), a short generation time (about 4 months), is easy to propagate from seed or clonally, and selected accessions can be easily transformed by *Agrobacterium tumefaciens* (Smith and Townsend 1907; Conn 1942) in tissue culture. With the growing use of *F. vesca* as a model rosaceous plant, we are working to develop forward and reverse genetic platforms in this diploid strawberry by implementing T-DNA mutagenesis (Oosumi et al. 2006).

T-DNA mutagenesis or 'gene tagging' is a technique used for generating loss-of-function mutations in genes by mobile or introduced DNA with a known sequence (T-DNA). The insertion of foreign T-DNA into a gene both mutates the gene and serves as a vehicle for the isolation of the unknown genomic DNA flanking the insertion site. The foreign DNA not only disrupts the expression of the gene into which it has been inserted, but also acts as a marker for subsequent identification of the mutation. An advantage of using T-DNA as the insertional mutagen, as opposed to transposon tagging, is that T-DNA insertions will not transpose subsequent to integration within the genome and are therefore chemically and physically stable through multiple generations (Krysan et al. 1999). In insertional mutant populations of Arabidopsis and rice, T-DNA has shown a preference for integration into gene-rich chromosomal regions rather than centromeric regions that contain fewer genes (Li et al. 2006; Shuangyan et al. 2003).

Characterization of the genomic DNA flanking the T-DNA insertion sites (IS) can be done by TAIL-PCR, originally described by Liu and Whittier (1995). This technique employs nested, specific primers designed outward from the T-DNA used sequentially in combination with arbitrary degenerate primers for the unknown genomic DNA region flanking the insertion site (Liu and Whittier 1995). As a result, priming creates specific and nonspecific products that are thermally controlled by PCR cycles carried out with alternating high and low annealing temperatures, allowing both types of primers to function at optimum efficiency. TAIL-PCR has been widely used to recover flanking sequences of T-DNA and/or transposon insertions for functional genomics, cloning such insertion-tagged genes, recovering upstream regulatory sequences of tagged genes, and cloning genes from large vectors such as P1 phage, yeast and bacterial artificial chromosomes (Liu and Whittier 1995; Liu and Chen 2007).

DNA marker analysis of genomes including those of rosaceous plants along with all other families of economically important plant species has revolutionized genetic studies in the past decade. Cleaved amplified polymorphic sequences (CAPS) and derived cleaved amplified polymorphism sequences (dCAPS) are PCR-based markers in which a restriction site is present or introduced in only one of two amplified sequences. This restriction site allows the use of restriction enzymes to cut one of two alleles and screen a mapping population easily and more economically than several other methods (Konieczny and Ausubel 1993; Neff et al. 1998; Neff et al. 2002). The CAPS method has been used to convert SNPs into PCR-based markers in rice (Komori and Nitta 2005). Genetic maps have been developed for all major crops including tomato (Liu 2005), sweet cherry (Olmstead 2008), rice (Murai et al. 2003) and apple (Sargent et al. 2009). In the diploid strawberry, the most complete genetic linkage map has been constructed from an F_2 population derived from an interspecific cross between two diploid *Fragaria* species, *F. vesca* 815 × *F. bucharica* 601 Losinsk. [formerly *F. nubicola* (Hook.f.) Lindl. ex Lacaita] (Sargent et al. 2004). The linkage map consists of a total of 348 molecular markers, which includes 272 microsatellite, or simple sequence repeat (SSR) loci, 35 gene-specific and STS markers, 40 RFLPs and 1 SCAR spanning 568.8 cM over seven linkage groups, as well as three morphological traits. Due to the abundance of transferable markers employed in the construction of this map, it has been adopted as the standard reference map for the genus, and has been used to build maps of the cultivated octoploid strawberry *F.* × *ananassa* (Rousseau-Gueutin et al. 2008; Sargent et al. 2009).

Insertional mutagenesis using *Agrobacterium* is an imprecise process. In addition to the desirable perfect single insertions of T-DNA into a gene to knock out its function, Krysan et al. (1999) describe a range of undesirable alternative integration patterns that may occur. The randomness of insertion has been questioned as T-DNA may show preference for integration into particular nucleotide regions. Our objective was to examine the range of insertion sites of a collection of strawberry transformants and position them on the strawberry genetic map. In this endeavor we recovered the genomic insertion sites of T-DNA strawberry insertional mutants using hiTAIL-PCR and examined the flanking sequences to develop allele-specific primers adjacent to T-DNA insertions. Single nucleotide polymorphisms (SNP) in these flanking regions between the parents of the reference mapping population (*F. vesca 815* and *F. bucharica 601*) were converted into CAPS/dCAPS markers that could be placed on the strawberry genetic map. In addition, we initiated functional annotation of our T-DNA insertional mutant collection.

Materials and methods

Isolation of unknown flanking sequences by hiTAIL-PCR

Independent T-DNA insertional mutant lines of *F. vesca* PI 551572 (FV10 or Hawaii 4) were generated using the *Agrobacterium*-mediated transformation protocol described by Oosumi et al. (2006). Transgenic lines carrying either of two binary vectors, pCAMBIA-1304 (GUS/GFP as selectable marker) or pCAMBIA 1302 (GFP selectable marker), were used. Genomic DNA was extracted from T_0 plants that were grown in soilless medium in a growth chamber following the previously described methods (Oosumi et al. 2006). The number of T-DNA insertions in each plant was estimated by Southern blot analysis (Lofstrand Labs

Limited, Gaithersburg, MD, USA). High-efficiency thermal asymmetric interlaced PCR (hiTAIL-PCR) according to Liu and Chen (2007) was performed on a subset of plants expected to carry a single copy of the T-DNA according to Southern blot results (data not shown). Initially, we used the right border nested primers of Liu and Chen (2007) on all the mutants. When right border reactions failed, we used left border nested primers (kindly provided by YG Liu, personal communication, Supplemental Table 1).

Primary and secondary reaction products were separated through a 1.2% agarose gel containing 0.1% ethidium bromide at 150 V for 45 min. Products were visualized over UV light and the largest band (not less than 600 bp) for each amplification was excised and purified using a QiaQuick gel extraction kit (Qiagen, Valencia, CA, USA) according to the manufacturer's recommendations and resuspended in 30 µl autoclaved distilled water. Products were sequenced at the Tufts University Core Facility (Boston, MA, USA).

Allele specific primer design from flanking sequences

Sequences generated from the hiTAIL-PCR fragments were analyzed using Vector NTI-10 software (http://www.invitrogen.com). Each sequence was aligned against the pCAMBIA vector to identify and remove any T-DNA or plasmid backbone sequence. Primers were designed from the remaining *Fragaria* DNA sequence to have a melting temperature between 55 and 65°C, a length between 20 and 30 nucleotides, and a GC content above 50%, using the primer design function of Vector NTI-10 software. Sequences were denoted VT (Virginia Tech) followed by the mutant number for each transgenic line.

Discovery of SNPs using PCR CAPS/dCAPS based markers

PCR was performed on genomic DNA of the parents of the diploid *Fragaria* reference linkage map (*F. vesca* 815 and *F. bucharica 601*) (Sargent et al. 2004) using primers designed from each T-DNA insertion flanking region in a final volume of 25 µl following the touchdown protocol of Sargent et al. (2003) with an annealing temperature from 55 to 50°C. Amplicons generated from each parent were sequenced twice to ensure fidelity at the Tufts University Core Facility and confirm the sequence polymorphism between the parental genotypes. Sequences were aligned using Vector NTI-10 to detect SNPs. CAPS/dCAPS markers were developed following the procedures of Konieczny and Ausubel (1993) and Michaels and Amasino (1998). Briefly, each potential SNP was isolated with a maximum of 25 flanking nucleotides on each side of a parental DNA sequence. Sequences were analyzed with the software dCAPS finder 2.0 (Neff et al. 2002), (http://helix.wustl.edu/dcaps/dcaps.html) to identify restriction enzymes that cleaved at the SNP position of one of the two parental sequences. If no restriction endonuclease recognition site was found, the program permits design of a PCR primer that introduces a mismatch into one of the two sequences flanking the SNP to create a recognition site and therefore allow allelic discrimination.

Genetic linkage mapping of SNPs in the diploid *Fragaria* reference map (*F. vesca* 815 × *F. bucharica* 601)

Primer pairs were used to generate polymorphic amplicons in the diploid *Fragaria* reference mapping population *F. vesca* 815 × *F. bucharica* 601. Polymorphisms were scored on 74 individuals as either presence or absence of an amplicon in the parents, or a length polymorphism created by a SNP site. PCR was performed in a final volume of 12.5 µl following the touchdown protocol of Sargent et al. (2003) using an annealing temperature from 55 to 50°C. Where required, products were digested for 2 h with the appropriate restriction endonuclease listed in Table 1 in a final volume of 20 µl containing 1× reaction buffer, 1 U restriction endonuclease and 1× BSA where required. Products were separated in a 1.2% agarose gel at 120 V for up to 3.5 h and visualized over UV light. Where polymorphisms between PCR products were too small to score reliably through agarose gel electrophoresis, the forward primer was fluorescently labeled with either 6-FAM or HEX fluorescent dyes (VHBio, Newcastle, UK) and products were scored after capillary electrophoresis as described by Sargent et al. (2003).

Chi-squared tests of goodness-of-fit to an expected segregation ratio of 1:2:1 or 3:1 were performed for all segregating T-DNA insertional mutant line markers using JOINMAP 4.0 software (Van Ooijen and Voorrips 2001). Marker placement was determined on the *F. vesca* 815 × *F. bucharica* 601 map using the mapping data of Sargent et al. (2008) following the procedures described therein. The map presented was constructed using MAPCHART 2.2 software for Windows (Voorrips 2002).

Functional annotation of the sequences derived from hiTAIL-PCR of T-DNA insertional mutant DNA sequences

Each *Fragaria* sequence flanking a T-DNA insertion site was blasted against the GenBank database (http://www.ncbi.nlm.nih.gov/) to identify proteins that may be homologous to the strawberry DNA sequences flanking the T-DNA insertions. Blastx [version 2.2.18 (Altschul et al. 1997)] was used to search flanking DNA sequences against the protein "nr" database downloaded on 15 May 2009

Table 1 Locus names, primer sequences, restriction enzymes, and product sizes of the 74 flanking sequence amplification products that were mapped by PCR CAPS/dCAPS-based markers from the *F. vesca* insertional mutants

Locus name	Forward primer (5′ to 3′)	Reverse primer (5′ to 3′)	Restriction enzyme	Product sizes (bp)	
				F. vesca 815	*F. bucharica* 601
VT-008a	aacttgacagctccgggattatgc	ctgcagttgtgcaacttaatcaacagt		650	
VT-008	tggtttgacctactgatcctctctgg	cactggattccaagtacttcctgat	*Mbo*I	550	680
VT-010	tgcccagtgctccctgtttgaaat	ctactgcatcaaactgacatgcgt	*Hpa*II	300/400	300/375
VT-021	agtggtgtaggctgtagctgaagt	cggcaataatcttgtccgtgtatgg	*Hae*III	200/250	500
VT-022	tcccaatccctcttattcttctgaaggc	ttgccagtcgggaattcctgttac	*Mbo*I	150	
VT-025	ccctttctgaacatgcatcccact	cagttgagtcactctaatttccttggg	*Mbo*I	550	680
VT-026	gcattctaagtcatggtcgacggg	gcaaccctagacataggaatatcgc		250	
VT-029	gccactttagaaatgcggct	tgttaatgtatgtgaatgatgtgacgc		600	750
VT-032	ttgtttcgcacgaggaagagggaa	aaaggatgggcgcagttgttcgat	*Mae*II	700	300
VT-034	gcaactgtccatgggcatctcaat	tgagcaccagcttgttgtatcagc	*Alu*I	740	300
VT-036	ccacaatgggtccacaattacaac	accatgtcggtaagcatcttct		200	230
VT-076	tgttgctgctcttagtgtgtgac	ggtccaaagagtctaaggtggtga		800	350
VT-081	tggaggcctttccataacccatga	accttctggcctctctcaatgaac	*Tsp*4CI	320	250
VT-087	cccaaagctctctctttccacaac	gatagtgcgtctgaggtgaatt	*Xap*I	150	190
VT-136	ttgccttgacctattccaggta	ccagagtaacaccaagttgaggca	*Rsa*I	150	200
VT-353	ttaacgtactaaaagatctccggtg	gggggatcggtatggaaatt	*ACi*I	300	450
VT-378	gcacgagtaaactttcccagttctg	acagaaggcttcctgaaactcagc		450	490
VT-382	ttaacgtactaaaagatctccggtg	tgttagctggaatttaatgtcaacc	*ACi*I	250	400
VT-393	ccacctcacccctctttttctctctctctt	ccagttcctgctcctcaacatccttaag		350	400
VT-398	acgttccaatcaaatttaaacatgt	gaagagaatggattgaccaaaaca		400	
VT-402	cccttggatcttagtttatcacatg	ttctaatacccatacgaaagggttt	Alu*I*	450	250
VT-408	atatgcatctggtcttaatcgatca	agtaatgaagctagcgtggtgtaga	*Tsp4CI*	450	380
VT-413	gtctaagcgtcatgatatccaacccc	aaggacaagatacgcgagacagtgag	*ApoI*	380	280
VT-423	agaacaaaaataaccgataccttgc	caaaatcatcaccacacccatcta		300	
VT-637	acaatattgcaaaagctcacccaag	ggtgccatggtacccctataattcat		1,100	
VT-643	aaatcgtaaaatttgggacgtgacat	attcttcattttctaatgct	*Tsp*4*CI*	350	280
VT-650	agggacacgtgtttgctaccagtt	gataaagcagtgaaaccgcatttta	*TspEI*	330	180
VT-656	aattactgtattagctccatcc	ccactttctcaatttctagt		300	
VT-659	aaccagggttttgtttcttgttg	gcaaacacaaaagaatgagactgaa	*Mbo*II	150	280
VT-661	accctcactacacatgcacttca	tgtcgtaagagggttcgtgtttaat	*Hin*dIII	150	320
VT-666	ttcacttctggtgtttctctttgtc	tggagagatggaattatagcagcta		300	
VT-673	tcatggtgtttgcactccacttt	aggcataagcagtgtgcgctatac	*Mae*II	180	300
VT-674	agtcaccaaaaccaccacgg	tgggcttagcgtatgtagtaggat	*Cvi*RI	250	280
VT-675	atatccaatacgacacagtcctgaa	gctttcatcctgttttggttcaaa	*Bsm*I	550	400
VT-679	aaagggcttttatatgggcttg	gcttaacctcaagcctcctaaatta		250	
VT-701	ttttggattggataagataaggcg	tcttctttgttagggtctgtgtgat	*Taq*I	300	420
VT-704	cgatagcatacacgcacccc	gcaccaaccaaagtgataatggatt	*Tsp*EI	180	250
VT-710A	aaaagagagtcctagaaaaacccagac	atttgctctgagtttgggctgctc		280	250
VT-710B	aaaagagagtcctagaaaaacccagac	atttgctctgagtttgggctgctc		280	250
VT-712	tagagcttagctgtcggtcattatg	ggttctggtgagaatgatttcatta	*Cvi*R1	230	280
VT-714	accaattgggtgttgccatg	ggagaggaaaccgatcaagcaatg		420	450
VT-720	accagtccctgacttcaaatatttt	agcaaaagtatgacacgaactgc	*Mbo*I	350	200
VT-723	cagaatcggacccccttgactgtt	ggaaccttgcacttgctgctttta	*Mn1*I	280	180
VT-725	catcttggatgacatagatgctctatccca	cagaatgtgtgggtaatagttggtctttga	*Eco*RI	680	450
VT-734	ataagggaattggtcgaaggtg	tgaacacaacacaatccaagctt	*Tsp*4CI	300	380
VT-735	ctgatcgagtttcagagagaatggc	gaatgagatagagagaaaagagaagaaaca		250	320

Table 1 continued

Locus name	Forward primer (5′ to 3′)	Reverse primer (5′ to 3′)	Restriction enzyme	Product sizes (bp)	
				F. vesca 815	*F. bucharica* 601
VT-737	ggctatcttggctagtgtgatgca	ttccaattcaggcctagtaggagc			270
VT-739	atcattctccttagcaatcgtgttcatctg	cgctaggcttgttaattactgcagttctat		1,000	
VT-743	ccaataacacaagaaggtatggca	caagagaggggggattagaccaaa	*Mbo*II	150	200
VT-744	ctggttgttgtttgactagttgca	aactagtatttgagtccaaagccg		750	350
VT-748	ctggatttgggatcaacctcagac	ggaccaaggctaccagaatattcaa		550	
VT-752	ctctcaagggtacgtacaagtactatttgc	cgccatagtgtccctagattttaagtcttt	*Mbo*II	250	450
VT-753	attcaatccagttcagcattcaaaggg	gcaattaagtttggtcaatcatggagactc	*Cvi*RI	180	350
VT-756	tcagagccccaagatcgattt	cacacacagtagactccagtacaaccct	*Hpy*188I	250	210
VT-763	gggtgataatgatcagtagccgttgc	gcccttctcttacagacgtgtcttga			1,400
VT-773	gctgctccttgggaatgttacttg	agaagtcctaaataaccttaatgtcgaca	*Eco*RV	420	280
VT-792	gtttctaggccttaggctctttgg	cttcaggcatggcactctctc		400	550
VT-794	ataaacgcggggcaaaatcg	cagagatggcggaggtcatc	*Aci*I	240	280
VT-803	tatgctgaagtaacatatattctgctttcc	tcttgctacatccaaatgataactatcaa		500	400
VT-811	gtcaaattgatcctcggcagt	cggttgtttgatgcaatgcc		420	
VT-816	gataagtgggaagtcaagagtcagg	tggtcccttattaagtctatcggtc	*Mn1*I	350	250
VT-824	ggtatagacccaaattactgctaatctaca	aactatcctcccgagtcccaagta	*Hpy*188I	290	220
VT-825	tggtttttggatgatctatcaactagc	tgtgctgcactatacaaatatagttgtcct		450	550
VT-837	agtggtcactgtgctcttgtatcttcc	ggggtaggaagaattccaaaagttttg	*Apa*I	200	220
VT-838	ttattatgatgttgattgggctgctcg	ggatgatggatctttatataattggtgtt	*Hpy*188I	120	290
VT-846	atttcaatctagactgatacgaagtgagcg	ccaaatcccagacccaatcaaaa		1,200	1,400
VT-850	atatgtagcgtatgctcttacttaattggt	ggcagggaatgaagtcacctatgt	*Hpy*188I	300	400
VT-851	tactgagcacagggtgtataaaggga	ttcctaggagtgcagctttgttcc	*Bbv*I	300	200
VT-855	tttgaccatataataaatccttcactgtca	gaacaccaaatatgcataaaagttgtctat		400	
VT-861	cacaaattgctttggtcaaggttag	catagtacagctgccttccacaga	*Mse*I	400	280
VT-869	atatgtagcgtatgctcttacttaattggt	acaccatcatctccaccaccat	*Hpy*188I	300	350
VT-872	cacaaattgctttggtcaaggttag	aacagcaagacccatgagagca	*Mse*I	320	250
VT-873	cgggtcatttaacacgtcaagc	catgacatgtgacacgcgttgt	*Mbo*I	650	450
VT-877	aattggtcatcgtattggtaatttcc	ccaatctcaattcagaggtcaaaga	*Aci*I	500	600

with the expectation threshold set to $1e^{-20}$ and other parameters at default values. The top-scoring hit to each flanking sequence was used as evidence of proximity of a coding region.

Results

Isolation of unknown flanking sequences by hiTAIL-PCR

Flanking regions were recovered from either the left or right borders of the T-DNA insertion sites (Fig. 1). An average fragment size of 800 bp (ranging from 500 to 2,400 bp) was obtained. Only fragments greater than 600 bp were excised from the gel for sequencing. Specific products were not always observed in the primary reactions; however, once the primary reactions were diluted

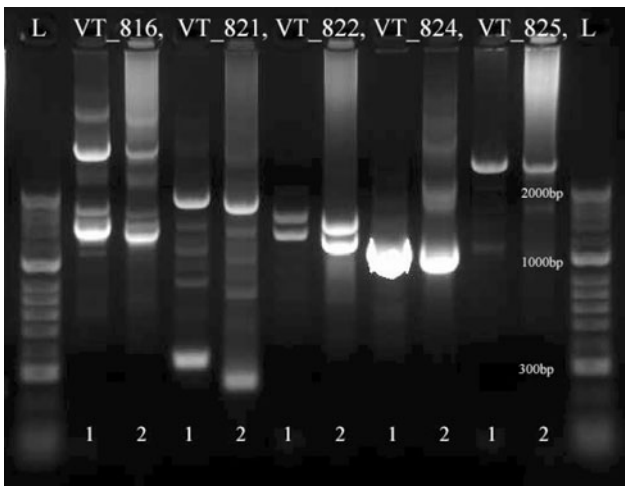


Fig. 1 Agarose gel electrophoresis of hiTAIL-PCR products of T-DNA insertional mutants of *F. vesca*. *L* 2,000 kb marker; under each of the five VT Mutant IDs, *lanes 1* and *2* represent the primary and secondary hiTAIL-PCR products, respectively

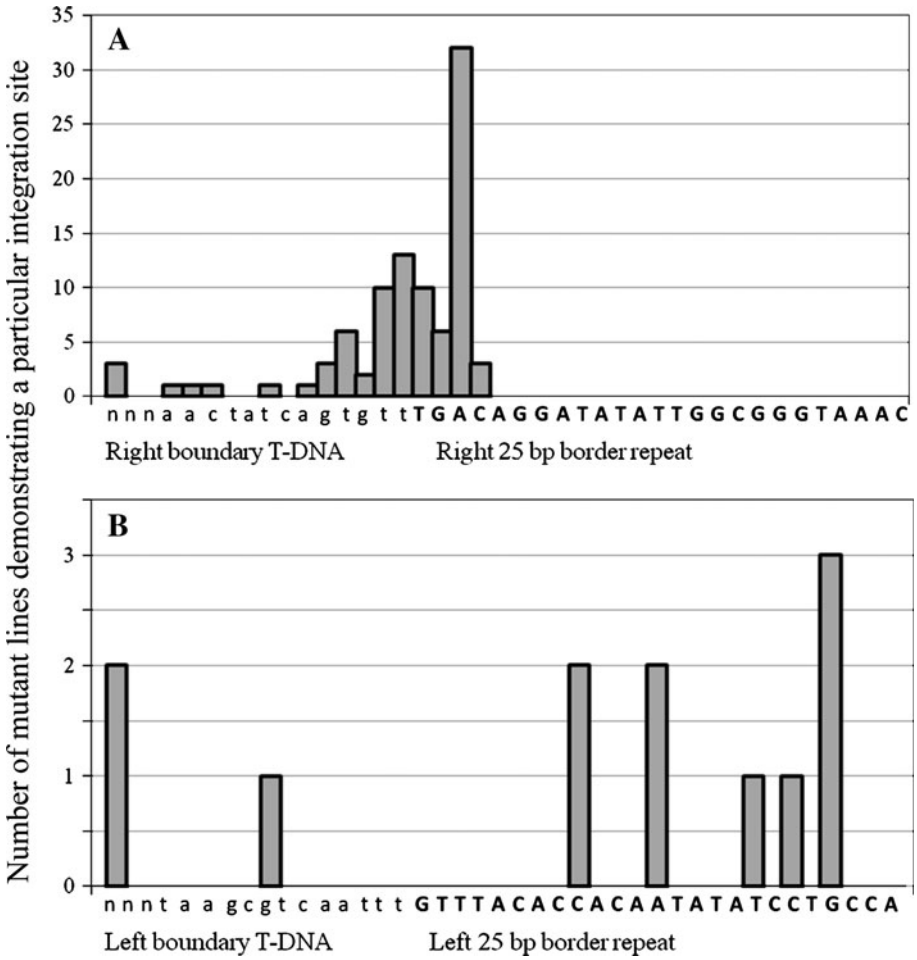


Fig. 2 Sequence analysis of pCAMBIA T-DNA insertion sites (IS) adjacent to the **a** right and **b** left border repeats. On the *x*-axis, nucleotides in lower case represent the T-DNA, those in bold and capital letters represent the 25 bp right and left imperfect repeat sequence adjacent to the T-DNA in the plasmid backbone. **a** Shows 96 mutants and **b** shows 12 mutants exhibiting each insertion. The *Y*-axis is the number of all mutants with each integration pattern

1:10 and used as a template for the secondary reactions, the specific products amplified during the subsequent secondary reactions were visible on a 1.2% agarose gel (Fig. 1). All sequences generated by hiTAIL-PCR contained approximately 128 bp of T-DNA right or left border sequence in addition to the flanking genomic sequence of the strawberry genome. In addition, the alignment of sequences revealed three duplicate sets of mutant lines (VT-861 and VT-872; VT-850 and VT-869; and VT-353 and VT-382), suggesting that more than one transgenic shoot that emerged from a callus representing the same transgenic event had been selected and transferred to the glasshouse.

Analysis of the boundaries of the T-DNA

In total, 108 transformation lines appeared to contain a single T-DNA insertion site that was stably integrated into the strawberry genome. Of these, 51 plants (53%) had T-DNA insertion sites that contained at least the first 4 bp of the 25 bp right border repeat sequence, while the remaining 45 (47%) lacked any of the right border repeat sequence. Nine (75%) of the sequences obtained with the left T-DNA border primers showed between 9 and 22 bp of the 25 bp left border repeat sequence, with the remaining three (25%) lacking any of the left border repeat sequence (Fig. 2).

Product amplification and scoring in the *F. vesca* 815 × *F. bucharica* 601 mapping progeny

All 108 primer pairs designed from transformation lines containing a single T-DNA insertion generated amplicons in one or both of the parents of the *F. vesca* 815 × *F. bucharica* 601 mapping population. Amplicons were obtained from just a single parent with 24 (22%) of the primer pairs: of these 20 (19%) amplified a product only in *F. vesca* 815 whereas four (2%) amplified a product only in *F. bucharica* 601. In subsequent analyses, 12 of these with only a *F. vesca* amplicon segregated 3:1 in the F_2 population as expected, two did not segregate as expected (assuming homozygosity in the parents), four displayed an amplification product for the other parent when the reaction was performed on the F_2, and two were monomorphic (*F. vesca* allele only) in the F_2. A further 15 (14%) primer pairs generated amplicons of different sizes between *F. vesca* 815 and *F. bucharica* 601 that could be resolved after agarose gel electrophoresis (Fig. 3); however, one of these could not be scored unambiguously in the

Fig. 3 Agarose gel demonstrating amplification patterns using primers developed from T-DNA flanking regions recovered from four different insertional mutants (VT) of *F. vesca.* Under each VT mutant number, *lane 1* is *F. vesca* 815, *lane 2* is *F. bucharica* 601 and *lane 3* is wild type *F. vesca* FV10, *L* 2,000 bp marker

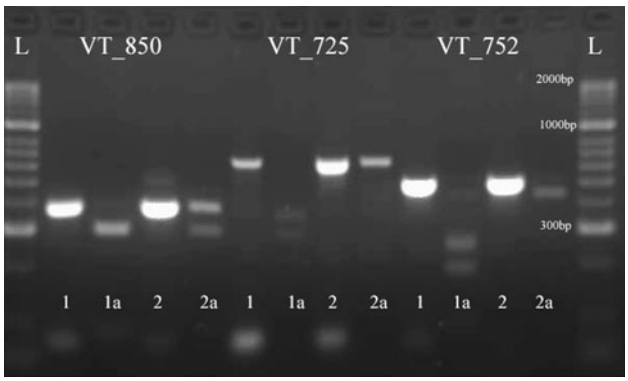


Fig. 4 Restriction enzyme digestion of PCR products generated from amplifying in T-DNA flanking sequences of three insertional mutants (VT) of *F. vesca* where the initial amplification products had been indistinct. Under each VT mutant number, *lanes 1* and *1a* are undigested and digested *F. vesca* 815, respectively, *lanes 2* and *2a* are undigested and digested *F. bucharica* 601, respectively; *L* 2,000 bp marker

F. vesca 815 × *F. bucharica* 601 progeny. A total of 79 (73.2%) primer pairs generated PCR products of the same size in the parents of the *F. vesca* 815 × *F. bucharica* 601 mapping population. Of these, 14 contained no SNP and thus were not polymorphic between *F. vesca* 815 and *F. bucharica* 601. Of the 65 (60%) sequences containing one or more SNPs, nine were heterozygous in the *F. bucharica* 601 parent but contained an allele common to the *F. vesca* 815 parent (*ab*×*aa*) and were subsequently homozygous in the F_1 plant (*aa*) from which the F_2 population derived, and thus did not segregate in the progeny. Ten primer pairs produced weak PCR amplicons and could not be scored following restriction enzyme digestion. The remaining 46 primer pairs yielded amplicons that could be cut with a restriction enzyme at a SNP position to reveal polymorphism between *F. vesca* 815 and *F. bucharica* 601 that segregated in the mapping population (Fig. 4). Table 1 lists locus name, forward and reverse primers, and the restriction endonucleases used to digest products containing SNPs, along with the amplicon sizes observed after digestion in the parents of the *F. vesca* 815 × *F. bucharica* 601 mapping population for the 74 mapped loci. Supplemental Table 2 lists similar information for the remaining 34 loci that could not be mapped using this strategy.

Thus, a total of 74 polymorphic markers segregated in the *F. vesca* 815 × *F. bucharica* 601 mapping population. Of these, 48 were co-dominant and segregated in an expected 1:2:1 Mendelian ratio in the progeny, 21 were dominant in *F. vesca* 815 and segregated in a 3:1 ratio in the F_2, and a further five were dominant in the *F. bucharica* 601 parent and approximated a 1:3 segregation in the progeny. Table 2 details the segregation data and chi-square values for goodness-of-fit to the expected and observed Mendelian segregation ratios for the 74 polymorphic markers segregating in the *F. vesca* 815 × *F. bucharica* 601 progeny. In total, 31 of the 74 markers (42%) exhibited significant deviation from the expected Mendelian segregation ratios (Table 2). In addition, the two markers in each of the three sets of duplicate (VT-861 and VT-872; VT-850 and VT-869; VT-353 and VT-382) T-DNA transgenic lines that were identified by hi-TAIL PCR produced identical segregation scored as expected.

Genetic linkage mapping of SNPs in the diploid *Fragaria* reference map (*F. vesca* 815 × *F. bucharica* 601)

After cosegregation analysis, all 74 polymorphic, segregating markers mapped to discrete loci on the *F. vesca* 815 × *F. bucharica* 601 reference map (Fig. 5). An average of 10.5 VT markers mapped to each of the seven *F. vesca* 815 × *F. bucharica* 601 linkage groups. The lowest number of markers (5) mapped to linkage group FG1, the greatest (16) mapped to linkage group FG5. Eight markers mapped to FG2, 9 to FG3, 13 to FG4, 14 to FG6, and 9 to FG7. Eight markers were located at the ends of six of the *F. vesca* 815 × *F. bucharica* 601 linkage groups, extending these groups by a total of 66.7 cM. The VT markers distributed across the seven linkage groups of the *F. vesca* 815 × *F. bucharica* 601 map and markers VT-861 and 872 (FG4) were located within a gap of more than 20 cM between markers BG-1 and CEL-1 on the previously published map of Sargent et al. (2008).

Functional annotation of the flanking regions for insertional mutant lines

All 108 T-DNA strawberry flanking sequences were blasted against the GenBank nr database (http://www.ncbi.nlm.

Table 2 Segregation analysis (chi-square values for goodness-of-fit to the expected Mendelian segregation ratios), linkage group (FG), map position, sequence characterization of T-DNA right (R) or left (L) border, and percentage of G/C content in each flanking sequence of 74 PCR markers derived from the flanking DNA sequence of T-DNA mutant strawberry lines in the diploid strawberry mapping population

Locus name	Classification	Expected	Observed	χ^2	*df*	Significance[a]	Linkage group	Map position	T-DNA border	% G/C content
VT-008	aa:ab:bb	1:2:1	19:42:13	2.32	2	–	FG6	42.7	RB	36.8
VT-008a	a_:bb	3:1	37:24	6.69	1	**	FG2	0.0	RB	38.0
VT-010	a_:bb	3:1	48:25	3.33	1	–	FG5	50.0	LB	33.1
VT-021	aa:ab:bb	1:2:1	10:30:32	15.44	2	***	FG4	32.1	LB	38.0
VT-022	aa:ab:bb	1:2:1	13:42:17	2.44	2	–	FG3	43.7	RB	37.6
VT-025	aa:ab:bb	1:2:1	20:42:13	2.39	2	–	FG6	42.4	RB	44.0
VT-026	a_:bb	3:1	57:17	0.16	1	–	FG1	31.9	RB	33.9
VT-029	aa:ab:bb	1:2:1	19:30:20	1.2	2	–	FG4	51.0	RB	32.9
VT-032	aa:ab:bb	1:2:1	10:33:29	10.53	2	**	FG4	31.0	LB	44.2
VT-034	aa:ab:bb	1:2:1	12:26:35	20.53	2	***	FG7	18.7	RB	33.8
VT-036	aa:ab:bb	1:2:1	3:30:39	38.0	2	***	FG2	18.0	RB	31.6
VT-076	a_:bb	3:1	52:23	1.28	1	–	FG6	72.8	RB	34.7
VT-081	aa:ab:bb	1:2:1	14:30:11	0.78	2	–	FG1	64.3	LB	38.6
VT-087	aa:ab:bb	1:2:1	10:37:29	9.55	2	**	FG5	51.8	RB	42.6
VT-136	aa:ab:bb	1:2:1	13:33:28	6.95	2	*	FG7	15.3	RB	38.5
VT-353	aa:ab:bb	1:2:1	12:17:25	13.67	2	***	FG7	5.2	RB	35.5
VT-378	aa:ab:bb	1:2:1	32:40:3	22.76	2	***	FG6	12.2	RB	36.5
VT-382	aa:ab:bb	1:2:1	12:17:25	13.67	2	***	FG7	5.2	RB	36.4
VT-393	aa:ab:bb	1:2:1	18:16:24	12.9	2	***	FG5	21.3	RB	40.4
VT-398	a_:bb	3:1	46:30	8.49	1	***	FG3	14.7	RB	32.7
VT-402	aa:ab:bb	1:2:1	13:48:6	14.01	2	***	FG4	0.0	RB	29.0
VT-408	aa:ab:bb	1:2:1	12:39:22	3.08	2	–	FG3	28.3	RB	36.0
VT-413	aa:ab:bb	1:2:1	9:35:26	8.26	2	–	FG5	50.6	RB	44.3
VT-423	a_:bb	3:1	44:32	11.86	1	**	FG7	22.3	RB	39.0
VT-637	a_:bb	3:1	38:38	25.33	1	**	FG2	22.3	RB	38.7
VT-643	a_:bb	3:1	45:30	9.0	1	**	FG5	55.0	RB	33.2
VT-650	aa:ab:bb	1:2:1	14:27:22	3.32	2	–	FG1	15.3	RB	42.6
VT-656	a_:bb	3:1	47:27	5.21	1	–	FG5	54.5	RB	29.1
VT-659	aa:ab:bb	1:2:1	15:39:22	1.34	2	–	FG7	63.7	RB	39.1
VT-661	aa:ab:bb	1:2:1	12:28:32	14.67	2	**	FG7	18.0	RB	36.6
VT-666	a_:bb	3:1	42:34	15.79	1	**	FG5	59.8	RB	41.6
VT-673	aa:ab:bb	1:2:1	11:36:25	5.44	2	–	FG3	20.8	RB	34.0
VT-674	aa:b_	1:3	17:56	0.11	1	–	FG6	50.0	RB	44.6
VT-675	aa:ab:bb	1:2:1	10:37:22	4.54	2	–	FG5	56.2	RB	40.5
VT-679	a_:bb	3:1	50:26	3.44	1	–	FG6	78.9	RB	42.3
VT-701	aa:ab:bb	1:2:1	15:40:19	0.92	2	–	FG4	76.3	RB	42.2
VT-704	aa:ab:bb	1:2:1	23:22:11	7.71	2	*	FG6	0.0	RB	41.7
VT-710A	aa:ab:bb	1:2:1	9:33:22	5.34	2	–	FG5	64.8	RB	46.2
VT-710B	aa:ab:bb	1:2:1	11:38:16	2.63	2	–	FG6	76.2	RB	46.2
VT-712	a_:bb	3:1	49:25	3.05	1	–	FG3	25.1	RB	39.7
VT-714	aa:ab:bb	1:2:1	5:30:34	25.55	2	***	FG2	26.0	RB	41.7
VT-720	aa:ab:bb	1:2:1	14:35:12	1.46	2	–	FG6	57.1	RB	34.2
VT-723	a_:bb	3:1	25:48	3.33	1	–	FG2	40.3	RB	41.8
VT-725	aa:ab:bb	1:2:1	12:32:16	0.8	2	–	FG3	52.5	RB	40.6
VT-734	aa:ab:bb	1:2:1	16:30:27	5.63	2	–	FG5	18.5	RB	46.9
VT-735	aa:ab:bb	1:2:1	24:18:12	11.33	2	***	FG5	81.3	RB	33.6

Table 2 continued

Locus name	Classification	Expected	Observed	χ^2	*df*	Significance[a]	Linkage group	Map position	T-DNA border	% G/C content
VT-737	a_:bb	3:1	40:35	32.11	1	***	FG2	16.7	RB	46.8
VT-739	a_:bb	3:1	65:6	10.37	1	***	FG6	35.9	RB	37.6
VT-743	aa:ab:bb	1:2:1	17:37:19	0.12	2	–	FG4	81.1	RB	39.9
VT-744	aa:ab:bb	1:2:1	16:36:26	5.35	2	–	FG3	24.3	RB	37.0
VT-748	a_:bb	3:1	56:20	0.07	1	–	FG6	56.1	RB	32.8
VT-752	aa:ab:bb	1:2:1	12:32:30	10.11	2	*	FG5	36.0	LB	36.0
VT-753	aa:ab:bb	1:2:1	20:29:19	1.5	2	–	FG4	85.9	LB	36.8
VT-756	aa:ab:bb	1:2:1	11:31:32	13.86	2	**	FG4	34.6	RB	37.9
VT-763	aa:ab:bb	1:2:1	11:41:20	3.64	2	–	FG3	27.5	RB	45.0
VT-773	aa:ab:bb	1:2:1	8:35:27	10.31	2	**	FG5	49.6	RB	38.1
VT-792	a_:bb	3:1	52:24	1.75	1	–	FG6	64.3	RB	37.0
VT-794	aa:b_	1:3	10:63	4.97	1	**	FG4	31.5	RB	45.5
VT-803	aa:ab:bb	1:2:1	10:32:11	2.32	2	–	FG7	77.8	RB	32.2
VT-811	a_:bb	3:1	45:31	10.11	1	***	FG5	53.7	RB	43.0
VT-816	aa:ab:bb	1:2:1	8:22:29	18.76	2	***	FG4	40.2	RB	41.4
VT-824	aa:ab:bb	1:2:1	17:42:11	3.83	2	–	FG6	43.5	RB	43.0
VT-825	aa:ab:bb	1:2:1	6:35:32	18.64	2	***	FG2	24.9	RB	52.2
VT-837	a_:bb	3:1	30:29	18.36	1	***	FG2	8.3	RB	36.1
VT-838	aa:b_	1:3	20:55	0.11	1	–	FG7	45.4	RB	34.9
VT-846	aa:b_	1:3	18:58	0.07	1	–	FG1	27.8	RB	37.2
VT-850	aa:ab:bb	1:2:1	13:30:30	10.23	2	**	FG5	34.7	RB	48.1
VT-851	a_:bb	3:1	39:26	7.8	1	**	FG4	20.2	RB	39.5
VT-855	a_:bb	3:1	52:24	1.75	1	–	FG3	34.1	RB	32.8
VT-861	aa:ab:bb	1:2:1	10:38:23	5.11	2	–	FG4	60.4	RB	35.8
VT-869	aa:b_	1:3	13:55	1.25	1	–	FG5	34.7	RB	45.9
VT-872	aa:ab:bb	1:2:1	11:40:23	4.38	2	–	FG4	60.4	RB	35.3
VT-873	aa:ab:bb	1:2:1	23:37:11	4.18	2	–	FG1	51.4	RB	33.7
VT-877	aa:ab:bb	1:2:1	5:41:11	12.23	2	***	FG6	57.1	RB	33.9

[a] *,**,***, significantly different from expected segregation at $p < 0.05$, 0.01, and 0.001 respectively; – not significantly different from expectation

nih.gov/). Of these, 49 (45%) showed high homology to proteins with a predicted or specific functionality. Table 3 lists the 49 VT loci and their putative protein function, GenBank accession number of the protein to which they were most closely matched and the species from which the protein was identified.

Discussion

Isolation of unknown flanking sequences by hiTAIL-PCR and analysis of the boundaries of the T-DNA

Agrobacterium-mediated transformation has many advantages over other transformation methods, including high efficiency, transfer of defined pieces of DNA (T-DNA), and the ability to transfer relatively large segments of DNA. This transformation process is controlled chromosomally and by plasmid-virulence encoded proteins. These virulence genes are located on the *Ti* plasmid and are responsible for the success of transfer of the T-DNA into the plant genome by illegitimate recombination, a process that joins two DNA molecules that do not share extensive homology. However the mechanism by which T-DNA integrates into the host genomes and the plant factors involved in the integration process are poorly understood (Brunaud et al. 2002; Bundock and Henry 2004; Dafny-Yelin et al. 2009; Windels et al. 2008). Finding the location of inserted T-DNA elements and characterizing the genomic DNA sequence into which they have been inserted using hiTAIL-PCR encompass a strategy for understanding gene function by forward genetics. We used hiTAIL-PCR to characterize strawberry T-DNA flanking sequences using nested specific primers from both right and left

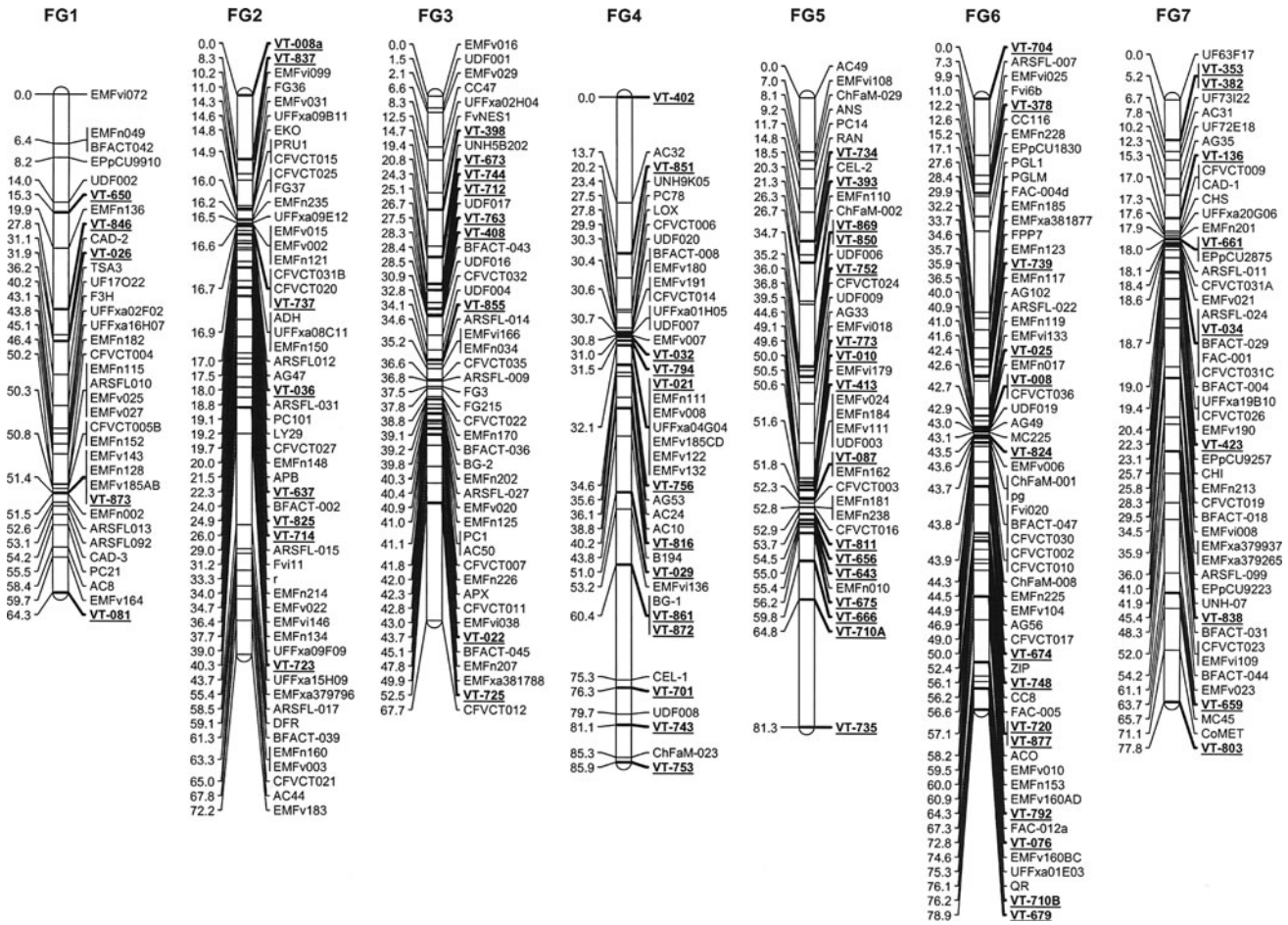


Fig. 5 Genetic linkage map of diploid strawberry demonstrating the positions of 74 insertional mutants of *F. vesca*. SNPs in the flanking sequences were used to develop CAPS/dCAPS markers. The new SNP-based markers as well as indel-based markers are identified by their VT mutant number and highlighted in bold. *Map distances* on the *left* are in cM

borders of the T-DNA, and LAD primer pools (Fig. 1). Although the LAD primers used in the pre-amplification reaction are nonspecific (Liu and Chen 2007), all four were able to amplify strawberry DNA to create one or more annealing site(s) for the primary and secondary hiTAIL-PCR reactions. Flanking sequences generated from more than 100 transgenic strawberry plants exhibited approximately 128 bp of the inserted T-DNA and then a flanking strawberry sequence. These results concur with Liu and Chen (2007), where all transgenic rice lines that were sequenced contained the inserted T-DNA as well as its flanking sequence.

Sequence analysis of both left and right borders of the T-DNA revealed variation of integration patterns of the T-DNA: the first 4 bp from the T-DNA 25-bp right border and up to 22 bp from the T-DNA 25-bp left border had integrated in different mutants. In rice, sequence analysis between the genomic DNA and T-DNA showed integration ranging from 1 to 3 bp of the 25-bp right border, and 6 up to 20-bp of the 25-bp left border (Hiei et al. 1994). Similar results have also been found in tobacco (Zambryski et al. 1982) and Arabidopsis (Brunaud et al. 2002).

We found that the flanking sequences analyzed showed an average GC content of 39.74% (SD = 0.053%). This deviates strongly from the GC content of 573 non-chloroplast protein coding regions found in GenBank (24 Feb 2010) with an average GC content of 44.6 ± 5.1% (*P* value from two-tailed *t* test = $1.8e^{-93}$). The background GC content of a typical rosaceous genome was estimated from a sample of 2,000 genome survey sequences from peach (*Prunus persica*) available in GenBank: mean = 38.14%, SD = 5.37%. To the extent that the *Prunus* genome approximates the *Fragaria* background GC content, tDNA insertion sites are not strongly biased by GC level (Fig. 6). Pan et al. (2005) and Brunaud et al. (2002) in analyses of T-DNA integration into the Arabidopsis genome found a preference for integration into regions ranging from 20 to 30% GC; however, the genomic GC content of Arabidopsis is about 36% and is low relative to other plant genomes.

Table 3 BLASTx analyses and predicted gene function of 49 T-DNA flanking strawberry sequences generated by hiTAIL-PCR from GenBank (http://www.ncbi.nlm.nih.gov/), as well as locus name, GenBank accession number, score, *E* value, and species where it was identified first

Locus name	GenBank accession No	Predicted gene function	GenBank NR protein data set result		Species
			Score	*E* value	
VT-008a	EEF41624.1	Peroxisome assembly factor-2, putative	113	$1e^{-22}$	*Ricinus communis*
VT-021	AAD49734.1	AF169795-1 glutamine synthetase precursor	130	$9e^{-28}$	*Juglans nigra*
VT-022	XP-002313228.1	Predicted protein	107	$8e^{-21}$	*Populus trichocarpa*
VT-025	EEF46369.1	Stem 28 kDa glycoprotein precursor, putative	186	$1e^{-44}$	*Ricinus communis*
VT-032	EEF41851.1	RNA binding motif protein, putative	273	$7e^{-71}$	*Ricinus communis*
VT-033	EEF36498.1	Protein pof4, putative	80	$8e^{-29}$	*Ricinus communis*
VT-054	XP-002313425.1	Cytochrome P450	245	$1e^{-62}$	*Populus trichocarpa*
VT-076	EEF49342.1	Lysosomal alpha-mannosidase, putative	102	$6e^{-57}$	*Ricinus communis*
VT-087	lNP-177103.1	AGO7 (ARGONAUTE7); nucleic acid binding	112	$5e^{-40}$	*Arabidopsis thaliana*
VT-136	EEF37316.1	Erythroblast macrophage protein emp	132	$5e^{-42}$	*Ricinus communis*
VT-393	EEF51349.1	Protein binding protein, putative	116	$1e^{-23}$	*Ricinus communis*
VT-413	EEF48596.1	Transcription factor, putative	310	$7e^{-82}$	*Ricinus communis*
VT-637	EEF41389.1	Interferon-induced guanylate-binding protein, putative	109	$9e^{-23}$	*Ricinus communis*
VT-640	EEF41389.1	Interferon-induced guanylate-binding protein, putative	109	$9e^{-23}$	*Ricinus communis*
VT-660**	CAA71490.1	Peroxidase	117	$4e^{-24}$	*Spinacia oleracea*
VT-666	EEF48571.1	Clathrin assembly protein, putative	157	$3e^{-36}$	*Ricinus communis*
VT-669**	EEF33958.1	Amino acid transporter, putative	169	$1e^{-39}$	*Ricinus communis*
VT-674	EEF51150.1	Leucine-rich repeat receptor protein kinase EXS precursor	213	$7e^{-53}$	*Ricinus communis*
VT-675	XP-002314629.1	SIN3 component, histone deacetylase complex	203	$9e^{-50}$	*Vitis vinifera*
VT-676**	EEF51150.1	Leucine-rich repeat receptor protein kinase EXS precursor, putative	714	0.0	*Ricinus communis*
VT-679	NP-198784.1	EMB2744 (EMBRYO DEFECTIVE 2744)	483	e^{-134}	*Arabidopsis thaliana*
VT-682**	lXP-002316906.1	Predicted protein	223	$5e^{-56}$	*Populus trichocarpa*
VT-691**	EEF39690.1	Leucoanthocyanidin dioxygenase, putative	226	$4e^{-72}$	*Ricinus communis*
VT-701	EEF36999.1	Polygalacturonase, putative	219	$1e^{-54}$	*Ricinus communis*
VT-714	XP-002264804.1	PREDICTED: hypothetical protein	114	$6e^{-23}$	*Vitis vinifera*
VT-715	EEF49033.1	UDP-glucosyltransferase, putative	321	$9e^{-86}$	*Ricinus communis*
VT-721	EEF46451.1	Cytoplasmic dynein light chain, putative	94	$1e^{-23}$	*Ricinus communis*
VT-723	EEF36880.1	Nucleic acid binding protein, putative	334	$2e^{-89}$	*Ricinus communis*
VT-725	EEF35141.1	DNA binding protein, putative	465	e^{-143}	*Ricinus communis*
VT-729	NP-190817.1	Heat shock protein-related	483	e^{-134}	*Arabidopsis thaliana*
VT-730	EEF29797.1	Cop9 signalosome complex subunit, putative	91	$2e^{-28}$	*Ricinus communis*
VT-734	NP-849842.1	AAA-type ATPase family protein	103	$3e^{-32}$	*Arabidopsis thaliana*
VT-735	NP-173839.1	AAA-type ATPase family protein	301	$3e^{-79}$	*Arabidopsis thaliana*
VT-739	XP-002271602.1	PREDICTED: hypothetical protein	161	$4e^{-37}$	*Vitis vinifera*
VT-750**	XP-002311853.1	Glycosyltransferase, CAZy family GT8	563	e^{-158}	*Populus trichocarpa*
VT-753	XP-002272777.1	PREDICTED: hypothetical protein	125	$2e^{-26}$	*Vitis vinifera*
VT-756	EEF50019.1	ATP-citrate synthase, putative	247	$5e^{-63}$	*Ricinus communis*
VT-763	XP-002300564.1	Predicted protein	131	$3e^{-28}$	*Populus trichocarpa*
VT-794	NP-194760.2	Ribitol kinase, putative proteína	119	$1e^{-24}$	*Arabidopsis thaliana*
VT-798	lNP-849972.1	AAA-type ATPase family protein	503	e^{-140}	*Arabidopsis thaliana*
VT-811	EEF48468.1	Pollen-specific protein C13 precursor, putative	150	$6e^{-34}$	*Ricinus communis*
VT-816	EEF47537.1	Anthranilate N-benzoyltransferase protein, putative	246	$7e^{-63}$	*Ricinus communis*

Table 3 continued

Locus name	GenBank accession No	Predicted gene function	GenBank NR protein data set result		Species
			Score	*E* value	
VT-824	ABQ53132.1	Putative CBF/DREB transcription factor	113	$5e^{-23}$	*Rosa chinensis*
VT-850	ACN38268.1	Flavonoid 3′ hydroxylase	201	$4e^{-49}$	*Vitis amurensis*
VT-851	CAN83125.1	PREDICTED: hypothetical protein	176	$2e^{-42}$	*Vitis vinifera*
VT-861	XP-002274118.1	PREDICTED: hypothetical protein	367	$1e^{-99}$	*Vitis vinifera*
VT-865	NP-001151166.1	Acetolactate synthase/amino acid binding protein	128	$3e^{-27}$	*Zea mays*
VT-869	ACN38268.1	Flavonoid 3′ hydroxylase	201	$4e^{-49}$	*Vitis amurensis*
VT-872	XP-002274118.1	PREDICTED: hypothetical protein	367	$1e^{-99}$	*Vitis vinifera*

** Monomorphic markers

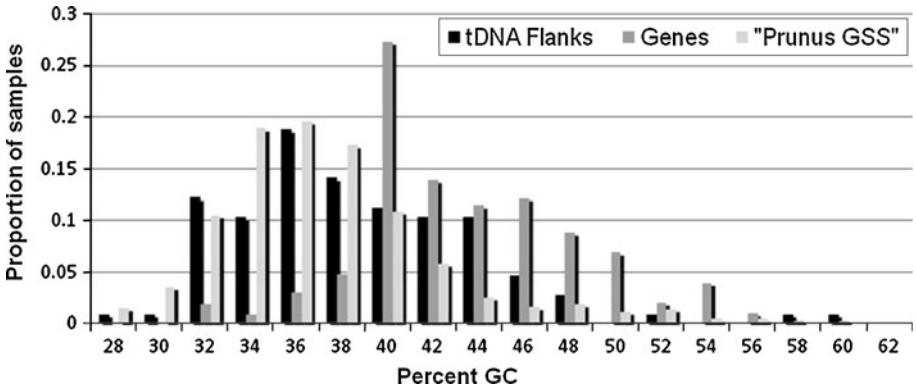


Fig. 6 Comparison of GC content among 108 tDNA flanking sequences (from hiTAIL-PCR), 573 *Fragaria* protein coding genes, and 2,000 *Prunus persica* genome survey sequences (GSS). The GSS from a closely related rosaceous genome estimate the background GC content of *Fragaria*

Amplification of products from strawberry flanking sequences and SNP discovery between *F. vesca* 815 and *F. bucharica* 601

After sequencing the flanking region of more than a 100 T-DNA strawberry mutants, all regenerated from *F. vesca* PI 551572, we designed and tested sequence-based primers on genomic DNA of the parental lines of the *F. vesca* 815 × *F. bucharica* 601 mapping population. For 39 (36%) of the primer pairs that revealed polymorphic products between the parental lines, we did not observe segregation in the F_2 mapping population due to the fact that the heterozygous *F. bucharica* 601 shared an allele in common with *F. vesca* 815 and this allele was transmitted to the F_1 plant used to raise the *F. vesca* 815 × *F. bucharica* 601 mapping population. However, 28 (26%) of the primer pairs revealed polymorphic products between the parental lines and segregated in the F_2 progeny allowing the markers to be placed on the *F. vesca* 815 × *F. bucharica* 601 map. The remaining 46 primer pairs amplified similar product size between the mapping parents. After sequencing those fragments we could design assays that utilized inherent differences between the parents in restriction sites (CAPS) or SNPs. Genotyping SNPs in a mapping population or among a panel of genotypes representative of a crop species can be accomplished using DNA sequencing platforms, microarrays, Taqman assays, high resolution melting analysis among others. Our approach was modeled after the efforts of Neff et al. (1998) who utilized incomplete restriction recognition sites in combination with the polymorphic SNP for allelic detection. The use of restriction enzymes with no more than 4 or 5 bp recognition sites decreased the cost of the procedure. For dCAPS, we designed primers around the incomplete recognition site with a mismatched nucleotide; when used to amplify the loci, a complete recognition site was created in one of the mapping parents thus allowing allelic discrimination. This approach has been used for SNP assaying to identify a functional polymorphism for the waxy character (*Wx*) of wheat (Yanagisawa et al. 2003), herbicide resistance via mutations in acetolactate-synthase (ALS) in ryegrass (Délye et al. 2009), and alternative marker developed for saturating the genetic linkage map of chickpea (Muehlbauer and Rajesh 2008). Our approach utilized the CAPS/dCAPS methodology to reveal the genetic location of the T-DNA insertion to characterize individual mutant lines.

SNPs represent the most common type of sequence polymorphism in plants, animals and humans (Brookes 1999; Henry 2008). In the present study, an average of one

SNP was found between *F. vesca* 815 and *F. bucharica* 601 every 60 bp, higher than the frequency previously observed between these two species accessions by Sargent et al. (2009). The abundance of nucleotide variation (SNPs, insertions and deletions) between *F. vesca* 815 and *F. bucharica* 601 facilitated the development of robust PCR markers that allowed us to place 74 new markers on the diploid strawberry genetic linkage map that correspond to the genetic location of T-DNA insertions in the respective transgenic strawberry lines. Flanking sequences that exhibited no SNP polymorphism were observed in nine cases, suggesting that these represented highly conserved regions. Some of these monomorphic sequences are shown in Table 3 with the predicted function of the interrupted gene as previously described in other crops such as coffee (Aggarwal et al. 2007).

Genetic linkage mapping of SNPs in the diploid *Fragaria* reference map (*F. vesca* 815 × *F. bucharica* 601)

The PCR CAPS/dCAPS-based markers showing polymorphism between the parents of the *F. vesca* 815 × *F. bucharica* 601 reference map were used to genotype the F_2 progeny analyzed in this investigation. We added and determined the location of 74 unique, independent single copy T-DNA insertion sites and increased the density of DNA markers saturating the reference *Fragaria* map. By comparing the position of our PCR CAPS/dCAPS-based markers within and between linkage groups, we deduced that eight VT markers were located at the distal regions of six linkage groups, thus elongating the linkage group beyond previously placed SSR markers (Fig. 5) (Hadonou et al. 2004; Sargent et al. 2004, 2006, 2007). We also found in linkage group FG3, FG4, and FG5 that four groups of three markers were closely linked to each other. One possible explanation is that these insertions occurred in genetic regions closely linked to the centromeres where recombination is generally repressed, thus giving the illusion of proximity. Alternatively, these genomic regions may harbor nucleotide sequences that are particularly susceptible to T-DNA integration.

Distorted segregation has been detected with almost all molecular and genetic markers due to factors like structure of the mapping population, genetic transmission, gametic and zygotic selection, non-homologous recombination, gene transfer, transposable elements and environmental agents (Jing et al. 2007; Knox and Ellis 2002; Zhao et al. 2006a). In our strawberry segregation data, we observed 42% of cases with significant deviation ($P \leq 0.05$) from the expected Mendelian segregation ratios (Table 2). Similar results were reported in the reference *Fragaria* map by Sargent et al. (2004) with 54% of SSR markers displaying distorted segregation. Molecular markers scored in an F_2 population of rice showed 33% genetic distortion (Zhao et al. 2006a), 54% in *Avena sativa* (Tanhuanpää et al. 2008), and 11% in sunflower (Lai et al. 2005). There were five cases (including 11 VT markers) where skewed markers grouped together, whereas the 21 remaining distorted VT markers mapped independently (Table 2; Fig. 5).

In two cases we observed an amplicon for *F. vesca* 815 but not *F. bucharica* 601 and expected a 3:1 segregation (presence:absence) in the F_2; however, we observed monomorphic F_2 with all genotypes exhibiting the *F. vesca* 815 amplicon. It is possible that this represents extreme distortion where the null allele of the *F. bucharica* 601 parent that was expected to be heterozygous in the F_1 could not be tolerated as homozygous in the F_2 with *F. vesca* 815 cytoplasm. The most likely map position for these markers would be at the proximal end of FG2 where the most extreme segregation distortion occurred on the *F. vesca* 815 × *F. bucharica* 601 map (Sargent et al. 2004).

Functional annotation of the hiTAIL PCR of T-DNA insertional mutant DNA sequences

To begin the functional characterization of T-DNA flanking strawberry sequences we used the GenBank NR protein data set to translate nucleotide sequences into encoded proteins. Almost 50% of the 108 T-DNA strawberry flanking sequences analyzed in this investigation were closely related to proteins with a predicted or specific functionality that have been previously characterized in other plant species. We found T-DNA mutants that likely alter transcription of genes affecting proteins related to: RNA binding motif, cytochrome P450, amino acid transport, cytoplasmic dynein light chain, ATPase, pollen-specific C13 precursor, flavonoid 3′ hydroxylase, among others (Table 3). Mutant VT_850 with a strong hit to a gene involved in flavonoid synthesis in *Vitis vinifera* is worthy of future investigation considering the important contribution of flavonoids to the health promoting properties of strawberry and other fruit crops (Shulaev et al. 2008). The remaining sequences (over half) had no predicted functionality. Mapped insertional mutants related to gene functions as described in this study provide an invaluable resource for forward genetics in fruit crops. Despite high transformation efficiency of *F. vesca* using the tissue culture methodology described by Oosumi et al. (2006), generating sufficient insertional mutants to saturate the strawberry genome would be prohibitively laborious to adopt a strategy of using a single transformation event to generate each insertional mutant. The *in planta* transformation protocol used to develop large knockout collections in Arabidopsis (Krysan et al. 1999; Robinson et al. 2009;

Woody et al. 2007) represents a much more efficient system but is unavailable for strawberry. Therefore, we are in the process of developing a transposon tagging strategy where unlimited transposants can be generated from a single launch pad transgenic plant carrying an *AcDs* construct for generating a more extensive population, as in rice and barley (Kolesnik et al. 2004; Upadhyaya et al. 2006; Zhao et al. 2006b).

Acknowledgments This work was supported by the USDA/NRI 2008-02284, Hatch VA-135816, the East Malling Trust for Horticulture Research, the BBSRC, and a Virginia Tech ASPIRES Grant.

References

Aggarwal RK, Hendre PS, Varshney RK, Bhat PR, Krishnakumar V, Lalji S (2007) Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. Theor Appl Genet 114:359–372

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

Brookes AJ (1999) The essence of SNPs. Gene 234:177–186

Brunaud V, Balzergue S, Dubreucq B, Aubourg S, Samson F, Chauvin S, Bechtold N, Cruaud C, DeRose R, Pelletier G, Lepiniec L, Caboche M, Lecharny A (2002) T-DNA integration into the Arabidopsis genome depends on sequences of pre-insertion sites. EMBO Rep 3:1152–1157

Bundock PC, Henry RJ (2004) Single nucleotide polymorphism, haplotype diversity and recombination in the *Isa* gene of barley. Theor Appl Genet 109:543–551

Conn HJ (1942) Validity of the genus *Alcaligenes*. J Bacteriol 44:353–360

Dafny-Yelin M, Tovkach A, Tzfira T (2009) Integration of *Agrobacterium* T-DNA in plant cells. In: Meier I (ed) Functional organization of the plant nucleus. Springer, Berlin, pp 157–186

Délye C, Boucansaud K, Pernin F, Le Corre V (2009) Variation in the gene encoding acetolactate-synthase in *Lolium* species and proactive detection of mutant, herbicide-resistant alleles. Weed Res 49:326–336

Folta KM, Davis TM (2006) Strawberry genes and genomics. Crit Rev Plant Sci 25:399–415

Hadonou AM, Sargent DJ, Wilson F, James CM, Simpson DW (2004) Development of microsatellite markers in *Fragaria*, their use in genetic diversity analysis, and their potential for genetic linkage mapping. Genome 47:429–438

Henry RJ (2008) Future prospects for plant genotyping. In: Henry RJ (ed) Plant genotyping II: SNP technology. CABI, Wallingford, pp 272–280

Hiei Y, Ohta S, Komari T, Kumashiro T (1994) Efficient transformation of rice (*Oryza sativa* L.) mediated by *Agrobacterium* and sequence analysis of the boundaries of the T-DNA. Plant J 6:271–282

Jing R, Johnson R, Seres A, Kiss G, Ambrose MJ, Knox MR, Ellis THN, Flavell AJ (2007) Gene-based sequence diversity analysis of field pea (*Pisum*). Genetics 177:2263–2275

Knox MR, Ellis THN (2002) Excess heterozygosity contributes to genetic map expansion in pea recombinant inbred populations. Genetics 162:861–873

Kolesnik T, Szeverenyi I, Bachmann D, Kumar CS, Jiang S, Ramamoorthy R, Cai M, Ma ZG, Sundaresan V, Ramachandran S (2004) Establishing an efficient *Ac/Ds* tagging system in rice: large-scale analysis of *Ds* flanking sequences. Plant J 37:301–314

Komori T, Nitta N (2005) Utilization of the CAPS/dCAPS method to convert rice SNPs into PCR-based markers. Breed Sci 55:93–98

Konieczny A, Ausubel FM (1993) A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. Plant J 4:403–410

Krysan PJ, Young JC, Sussman MR (1999) T-DNA as an insertional mutagen in *Arabidopsis*. Plant Cell 11:2283–2290

Lai Z, Livingstone K, Zou Y, Church SA, Knapp SJ, Andrews J, Rieseberg LH (2005) Identification and mapping of SNPs from ESTs in sunflower. Theor Appl Genet 111:1532–1544

Li Y, Rosso MG, Ülker B, Weisshaar B (2006) Analysis of T-DNA insertion site distribution patterns in *Arabidopsis thaliana* reveals special features of genes without insertions. Genomics 87:645–652

Liu Y (2005) Construction of a genetic map and location of quantitative trait loci for number of flowers per truss traits in tomato by SSR markers. J Nanjing Agric Univ 28:30–34

Liu YG, Chen Y (2007) High-efficiency thermal asymmetric interlaced PCR for amplification of unknown flanking sequences. BioTechniques 43:649–656

Liu Y-G, Whittier RF (1995) Thermal asymmetric interlaced PCR: automatable amplification and sequencing of insert end fragments from P1 and YAC clones for chromosome walking. Genomics 25:674–681

Michaels SD, Amasino RM (1998) A robust method for detecting single-nucleotide changes as polymorphic markers by PCR. Plant J 14:381–385

Muehlbauer FJ, Rajesh PN (2008) Chickpea, a common source of protein and starch in the semi-arid tropics. In: Jorgensen RA (ed) Plant genetics and genomics: crops and models. Springer, Heidelberg, pp 171–186

Murai H, Sharma PN, Murata K, Hashimoto Z, Ketipearachi Y, Shimizu T, Takumi S, Mori N, Kawasaki S, Nakamura C (2003) Constructing linkage maps of brown planthopper resistance genes *Bph1*, *bph2*, and *Bph9* on rice chromosome 12. In: Khush GS, Brar DS, Hardy B (eds) Rice genetics collection. World Scientific Publishing, Singapore, pp 263–265

Neff MM, Neff JD, Chory J, Pepper AE (1998) dCAPS, a simple technique for the genetic analysis of single nucleotide polymorphisms: experimental applications in *Arabidopsis thaliana* genetics. Plant J 14:387–392

Neff MM, Turk E, Kalishman M (2002) Web-based primer design for single nucleotide polymorphism analysis. Trends Genet 18:613–615

Olmstead JW (2008) Construction of an intra-specific sweet cherry (*Prunus avium* L.) genetic linkage map and synteny analysis with the *Prunus* reference map. Tree Genet Genom 4:897–910

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 223:1219–1230

Pan XK, Li Y, Stein L (2005) Site preferences of insertional mutagenesis agents in Arabidopsis. Plant Physiol 137:168–175

Potter D, Luby JJ, Harrison RE (2000) Phylogenetic relationships among species of *Fragaria* (Rosaceae) inferred from non-coding nuclear and chloroplast DNA sequences. Syst Bot 25:337–348

Robinson SJ, Tang LH, Mooney BAG, McKay SJ, Clarke WE, Links MG, Karcz S, Regan S, Wu YY, Gruber MY, Cui DJ, Yu M, Parkin IAP (2009) An archived activation tagged population of *Arabidopsis thaliana* to facilitate forward genetics approaches. BMC Plant Biol 9:101

Rousseau-Gueutin M, Lerceteau-Köhler E, Barrot L, Sargent DJ, Monfort A, Simpson D, Arús P, Guérin G, Denoyes-Rothan B (2008) Comparative genetic mapping between octoploid and diploid *Fragaria* species reveals a high level of colinearity between their genomes and the essentially disomic behavior of the cultivated octoploid strawberry. Genetics 179:2045–2060

Sargent DJ, Hadonou AM, Simpson DW (2003) Development and characterization of polymorphic microsatellite markers from *Fragaria viridis*, a wild diploid strawberry. Mol Ecol Notes 3:550–552

Sargent DJ, Davis TM, Tobutt KR, Wilkinson MJ, Battey NH, Simpson DW (2004) A genetic linkage map of microsatellite, gene-specific and morphological markers in diploid *Fragaria*. Theor Appl Genet 109:1385–1391

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 112:1349–1359

Sargent DJ, Rys A, Nier S, Simpson DW, 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

Sargent DJ, Cipriani G, Vilanova S, Gil-Ariza D, Arús P, Simpson DW, Tobutt KR, Monfort A (2008) The development of a bin mapping population and the selective mapping of 103 markers in the diploid *Fragaria* reference map. Genome 51:120–127

Sargent DJ, Marchese A, Simpson DW, Howad W, Fernández-Fernández F, Monfort A, Arús P, Evans KM, Tobutt KR (2009) Development of "universal" gene-specific markers from *Malus* spp. cDNA sequences, their mapping and use in synteny studies within Rosaceae. Tree Genet Genom 5:133–145

Shuangyan C, Weizheng J, Mingyi W, Fan Z, Jie Z, Qiaojun J, Yunrong W, Feiyan L, Ping W (2003) Distribution and characterization of over 1000 T-DNA tags in rice genome. Plant J 36:105–113

Shulaev V, Korban SS, Sosinski B, Abbott AG, Aldwinckle HS, Folta KM, Iezzoni A, Main D, Arús P, Dandekar AM, Lewers K, Brown SK, Davis TM, Gardiner SE, Potter D, Veilleux RE (2008) Multiple models for Rosaceae genomics. Plant Physiol 147:985–1003

Smith EF, Townsend CO (1907) A plant tumor of bacterial origin. Science 25:671–673

Tanhuanpää P, Kalendar R, Schulman AH, Kiviharju E (2008) The first doubled haploid linkage map for cultivated oat. Genome 51:560–569

Upadhyaya NM, Zhu QH, Zhou XR, Eamens AL, Hoque MS, Ramm K, Shivakkumar R, Smith KF, Pan ST, Li SZ, Peng KF, Kim SJ, Dennis ES (2006) *Dissociation* (*Ds*) constructs, mapped *Ds* launch pads and a transiently-expressed transposase system suitable for localized insertional mutagenesis in rice. Theor Appl Genet 112:1326–1341

Van Ooijen JW, Voorrips R (2001) Joinmap 3.0: software for the calculation of genetic linkage maps. Plant Research International, Wageningen

Voorrips RE (2002) MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered 93:77–78

Windels P, De Buck S, Depicker A (2008) *Agrobacterium tumefaciens*-mediated transformation: Patterns of T-DNA integration into the host genome. In: Tzfira T, Citovsky V (eds) Agrobacterium: from biology to biotechnology. Springer, New York, pp 441–481

Woody ST, Austin-Phillips S, Amasino RM, Krysan PJ (2007) The WiscDsLox T-DNA collection: an Arabidopsis community resource generated by using an improved high-throughput T-DNA sequencing pipeline. J Plant Res 120:157–165

Yanagisawa T, Kiribuchi-Otobe C, Hirano H, Suzuki Y, Fujita M (2003) Detection of single nucleotide polymorphism (SNP) controlling the waxy character in wheat by using a derived cleaved amplified polymorphic sequence (dCAPS) marker. Theor Appl Genet 107:84–88

Zambryski P, Depicker A, Kruger K, Goodman HM (1982) Tumour induction by *Agrobacterium tumefaciens*: analysis of the boundaries of T-DNA. J Mol Appl Genet 1:361–370

Zhao B, Deng Q-M, Zhang Q-J, Li J-Q, Ye S-P, Liang Y-S, Peng Y, Li P (2006a) Analysis of segregation distortion of molecular markers in F_2 population of rice. Acta Genet Sin 33:449–457

Zhao T, Palotta M, Langridge P, Prasad M, Graner A, Schulze-Lefert P, Koprek T (2006b) Mapped *Ds*/T-DNA launch pads for functional genomics in barley. Plant J 47:811–826

Supplemental data
Ruiz-Rojas et al.
SNP Discovery and Genetic Mapping of T-DNA Insertional Mutants in *Fragaria vesca* L.

Supplemental Table 1 Primer sequences utilized in hiTAIL-PCR. Leaf border primer sequences were kindly provided by YG Liu

Primer name:	**Primer sequence**	**Reaction in which each primer is utilized**
LAD1-1	5'–ACGATGGACTCCAGAGCGGCCGC(G/C/A)N (G/C/A) NNNGGAA-3'	Preamplification
LAD1-2	5'–ACGATGGACTCCAGAGCGGCCGC(G/C/T)N (G/C/T) NNNGGTT-3'	Preamplification
LAD1-3	5'–ACGATGGACTCCAGAGCGGCCGC(G/C/A)N (G/C/A) N(G/C/A)NNNCCAA-3'	Preamplification
LAD1-4	5'–ACGATGGACTCCAGAGCGGCCGC(G/C/T)N (G/A/T) N(G/C/T)NNNCGGT-3'	Preamplification
AC1	5'-ACGATGGACTCCAGAG-3'	Primary and secondary TAIL-PCR
RB-0a	5'GGCAATAAAGTTTCTTAAGATTGAACCTGT-3'	Preamplification
RB-1a	5'-ACGATGGACTCCACTCCGGCCTGTTGCCGG TCTTGCGATGATTATCA-3'	Primary TAIL-PCR
RB-2a	5'-GTAATGCATGACGTTATTTATGAGATGGGT T-3'	Secondary TAIL-PCR
LB-0a	5'-GGACCCTAATTCCCTTATCTGGGAACTAC-3'	Preamplification
LB-1a	5'-TCAACACATGAGCGAAACCCTATAACGATGG ACTCCAGTCCGGCC-3'	Primary TAIL-PCR
LB-2a	5'-TGGATTTTAGTACTTGGATTTTGGTTTTAGG-3'	Secondary TAIL-PCR

Supplemental Table 2 Locus names, primer sequences, restriction enzymes, and product sizes of the 34 flanking sequences analyzed in this investigation but not mapped.

Locus name	Forward primer (5'to 3')	Reverse primer (5' to 3')	Restriction enzyme	Product sizes (bp)	
				F. vesca 815	*F. bucharica* 601
VT-031	tgaggaatgcatcctttgctcgga	aggtccttggggtcgagaatgaatg	*BsmI*	420	400
VT-033	gctctttgtattagaaacattatagttgactg	tcctcatcagccgctcaaacacta	*Tsp4CI*	180	
VT-054	cacacaggaatgacaaggacatga	tggaagacgtggagtttgagggat		150	
VT-358	gtgtgacaccaaaggaggaaaatt	gaagtttggtaattttggcggat		480	
VT-363	aacataattctgagccaccaactt	tggtagtaccttcatggaatccag	*MboII*	320	380
VT-397	tactaacaatgagaatccatcaggg	agactcacagagacttcaagctcca		500	800

VT-427	cgttggcctcacgatatgtcagctagtt	ccctcgtagtctcatactccctctgttgtt		500	500
VT-640	acaatattgcaaaagctcacccaag	ttcctgaacaatttcattcctccct		250	250
VT-648	ttggtaagtatcctatttttccttttgcaa	gctaccagttcatcacttttgggttttctt		950	
VT-655	ctttcttttctttcatttcctccac	gagacccttcttcttgatgatgaa	*BccI*	200	400
VT-660	cagtctatcacgtagtgagcagtca	ccacctacctatctttcgagcatacc	*TspEI*	280	300
VT-669	ctatatttgtatgcagtgcgtctgg	ggacatggtctgagaaagtggct	*HindIII*	150	280
VT-676	ccattgaattggtctgcttcc	cattgaagcctgcacggaaa	*BccI*	220	180
VT-677	tcacttctggtgtttctctttgtcg	tttgtcatcatgttaaggcactgta		400	
VT-682	atctctaggctactacttgggctcc	gaggtaggaaagtttacgtcttggt	*Hpy188I*	280	230
VT-684	ccgaattttcttgcagccacg	gatctgctggagttgggaaggaa	*TaqI*	950	600/400
VT-691	aaaccaaagtagcaagcatcaaga	tgattggaactccttcttctcctc	*HaeIII*	320	450/320
VT-715	agggaaccctagggaagacatt	ttaatggactcagagtgaagcaaag			320
VT-721	ctatcgatcattcttaccaggatca	gaagttcttgtttcatgggttttca			320
VT-729	tcttcaccaaattctctagaacactcgtca	aaatctccccataagacccttttgtc		400	400
VT-730	ataaagactaattgccttataggcccaagc	gaagaggacgttgacattgagaat	*Mn1I*	220	200
VT-746	aggtatggagctggtacgagagacc	ccagagcggccacatggact		500	
VT-750	ttgccccgaaaacgtgttct	ttggcgaggcaatactccgg	*SmaI*	300	380
VT-757	ctctgcctaatttcattttgatcg	aactaccaagggttttagataccactaact		350	
VT-765	ttattggatagctacatatatc	agccatgtgaattaatgtgatccc		300	
VT-780	aggcatatctccaaagtgggactct	tgagccgcgtcaatatctcctt	*HphI*	800	680
VT-782	tagtcgcggaggctatggat	ggtcaagaccaatgcggagcatat	*HphI*	680	400
VT-798	tgcttgatcagctcatcccc	ttcttcaccgtcagctcaaact	*SacI*	300	370
VT-805	acaggcgcttgttgtggtgtaag	ggggatggaactttgagtgatgc	*AccI*	1200/700	900
VT-813	aaagattagtgggatcgtga	cacgtacaacttactgctatatatg	*AciI*	350	400
VT-844	ccgtcatacaagtatgatgaataacgc	agaatagcgtacgctctctctctttc		850	
VT-856	ttagacttttagagctttgatttggcaggg	catggttcacacacacgcacgtat	*AluI*	220	280
VT-865	tggccgtgatagttgggttagt	caacatctagagagtagagagtacgtgtca	*BseMII*	180	250
VT-875	gtatatgagcgttctgaaagaaaggtaa	tcctttcaaccaatgccattca		280	280

Chapter 4:

Implementing reverse genetics in Rosaceae: analysis of T-DNA flanking sequences of insertional mutant lines in the diploid strawberry, *Fragaria vesca*

Teruko Oosumi[a], Juan Jairo Ruiz-Rojas[b] , Richard E. Veilleux[b], Allan W. Dickerman[a], and Vladimir Shulaev [a,b]

Author contribution:

Dr Teruko Oosumi was a postdoctoral associate working at the Virginia Bioinformatics Institute at Virginia Tech until 2005 under Dr. Vladimir Shulaev's supervision. Dr Oosumi developed the strawberry transformation protocol and characterized the first set of insertional mutants used in this investigation. However, over the past 5 years several hundred additional mutants were generated in Dr Veilleux's laboratory. I analyzed many of the latter mutants and combined my data with that of Dr. Oosumi to make a more robust data set. In doing so, I updated the results, discussion, figures and tables of Dr. Oosmui's draft manuscript to increase its likelihood of publication; therefore, the both she and I, the first 2 authors on the published manuscript, contributed equally to the publication. Dr Allan W. Dickerman contributed the bioinformatics analysis and gene annotation of the flanking sequences. Dr Richard Veilleux and Dr Vladimir Shulaev are the principal investigators on the strawberry project; they directed the investigation and edited the manuscript.

[a]Department of Horticulture, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA
[b]Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

Physiologia Plantarum
An International Journal for Plant Biology

Physiologia Plantarum 140: 1–9. 2010

Implementing reverse genetics in Rosaceae: analysis of T-DNA flanking sequences of insertional mutant lines in the diploid strawberry, *Fragaria vesca*

Teruko Oosumi[a,†], Juan Jairo Ruiz-Rojas[b], Richard E. Veilleux[b], Allan Dickerman[a] and Vladimir Shulaev[a,b,*]

[a]Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA
[b]Department of Horticulture, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

Correspondence
*Corresponding author,
e-mail: vshulaev@vbi.vt.edu

Received 1 November 2009;
revised 21 April 2010

doi:10.1111/j.1399-3054.2010.01378.x

Reverse genetics is used for functional genomics research in model plants. To establish a model system for the systematic reverse genetics research in the Rosaceae family, we analyzed genomic DNA flanking the T-DNA insertions in 191 transgenic plants of the diploid strawberry, *Fragaria vesca*. One hundred and seventy-six T-DNA flanking sequences were amplified from the right border (RB) and 37 from the left border (LB) by thermal asymmetric interlaced PCR. Analysis of the T-DNA nick positions revealed that T-DNA was most frequently nicked at the cleavage sites. Analysis of 11 T-DNA integration sites indicated that T-DNA was integrated into the *F. vesca* genome by illegitimate recombination, as reported in other model plants: *Arabidopsis*, rice and tobacco. First, deletion of DNA was found at T-DNA integration target sites in all transgenic plants tested. Second, microsimilarities of a few base pairs between the left and/or right ends of the T-DNA and genomic sites were found in all transgenic plants tested. Finally, filler DNA was identified in four break-points. Out of 191 transgenic plants, T-DNA flanking sequences of 79 plants (41%) showed significant similarity to genes, elements or proteins of other plant species and 67 (35%) of the sequences are still unknown strawberry gene fragments. T-DNA flanking sequences of 126 plants (66%) showed homology to plant ESTs. This is the first report of T-DNA integration in a sizeable population of a rosaceous species. We have shown in this paper that T-DNA integration in strawberry is not random but directed by sequence microsimilarities in the host genome.

Introduction

The Rosaceae family comprises approximately 3000 species, including economically important fruit trees (apples, pears, peaches, apricots, plums and cherries), berries (strawberries), nuts (almonds) and ornamentals (roses). Several model species are now being used for genetics and genomics research in Rosaceae (Shulaev et al. 2008). Each model represents an individual rosaceous subtaxon and has unique features making it particularly useful for different research areas. The apple (*Malus* × *domestica*), representing the tribe Pyreae in

Abbreviations – EST, expressed sequence tag; LB, left border; PCR, polymerase chain reaction; RB, right border; TAIL-PCR, thermal asymmetric interlaced PCR; T-DNA, transferred DNA.

[†]Present address: Simplot Company, Simplot Plant Sciences, Boise, ID 83706, USA.

subfamily Spiraeoideae, is one of the most economically important fruit crops worldwide. It is currently the best developed model for functional genomics studies in the family. The peach (*Prunus persica*) representing the tribe Amygdaleae of subfamily Spiraeoideae is considered a genetic model for the family because it is the best genetically characterized species. It has a small genome size [290 Mb (Baird et al. 1994)] and well-developed genetic and physical maps. The physical map of peach is estimated to cover 303 Mb of the peach genome and it is comprised of 2138 contigs, of which 252 contigs are anchored to eight linkage groups of the *Prunus* reference map (Jung et al. 2008). However, the long generation time of apple and peach has a disadvantage for high throughput functional genomics studies.

The diploid strawberry, *Fragaria vesca* L. ($2n = 2x = 14$) is a member of the Rosoideae subfamily. It has emerged recently as a rosaceous genetic model because of its small genome size, short generation time (about 4 months) and small plant size (Battey et al. 1998, Folta and Davis 2006). Its small genome size [estimated at 206 Mb is only 1.7 times larger than that of *Arabidopsis thaliana* (125 Mb)]. Genetic linkage maps have been constructed for *F. vesca* based on both randomly amplified polymorphic DNA (RAPD) (Davis and Yu 1997) and microsatellite markers (Sargent et al. 2004, 2006, 2009).

As a result of concerted effort in all species, considerable sequence information has been accumulated for the family. Over 359 000 ESTs are currently available for Rosaceae species in NCBI dbEST and the Genome Database for Roscaeae (GDR, www.rosaceae.org). About 70% of all Rosaceae ESTs are from apple because of several high throughput sequencing projects. The total number of apple ESTs are over 250 000 sequences aligned in about 82 000 unigenes (Jung et al. 2008). A putative peach fruit unigene set of about 23 000 ESTs was defined and more than 1200 of the ESTs were mapped to develop the first transcript map for peach (Horn et al. 2005). ESTs are continuously deposited in the GenBank for both *F. vesca* and the cultivated octoploid strawberry *F.* × *ananassa* (currently 47 606 and 6176, respectively). Comparative genetic maps have been developed between different Rosaceae species (Dirlewanger et al. 2004, Sargent et al. 2009). Genomes of apple, peach and woodland strawberry are currently being sequenced (Sosinski et al. 2009). Availability of this vast amount of sequence information makes it extremely urgent that functional genomics studies aim at elucidating gene functions using an array of modern genomics tools.

Two complementary genetic approaches, namely forward and reverse genetics, are currently used for functional genomics research. Forward genetics starts with screening for mutant phenotypes and leads to identification of the gene responsible for the phenotype, whereas reverse genetics starts with a known mutant sequence and ends with the corresponding mutant phenotype. For both forward and reverse genetics, it is necessary to produce a population of mutants with a high probability of disrupting every gene in the genome. In addition, these mutants must be tagged by a known sequence. A collection of sequence-tagged mutants can be generated by the insertion of transposons or by T-DNA mutagenesis via *Agrobacterium*-mediated transformation.

Identification of mutants for reverse genetics can be done in several ways. Mutations in known genes of interest can be found by PCR screening using pools of genomic DNA from transgenic plants as templates (Galbiati et al. 2000, Krysan et al. 1996, 1999, McKinney et al. 1995, Rios et al. 2002, Tissier et al. 1999, Winkler et al. 1998). An alternative approach is to establish a database of tagged genes with functional annotations using systematic high throughput sequencing of genomic DNA flanking T-DNA or transposon insertion. Genomic sequences flanking the insertion sites can be amplified by inverse PCR (Ochman et al. 1988), thermal asymmetric interlaced (TAIL) PCR (Liu et al. 1995), suppression PCR (Schupp et al. 1999) or DNA fingerprinting (Theuns et al. 2002). Systematic study of the T-DNA insertional mutant collections in *Arabidopsis* and rice revealed that 48–63% of T-DNAs were integrated into genetic regions including 300 bp outside the start and stop codons (An et al. 2003, Ryu et al. 2004, Szabados et al. 2002).

To effectively use reverse genetics approach in Rosaceae, we need a high efficiency stable transformation system in a model species to generate a mutant collection large enough to induce mutation in every gene. We previously reported the development of an efficient transformation method in *F. vesca* using combination screening for hygromycin resistance and the expression of green fluorescence protein (Oosumi et al. 2006). We achieved 100% transformation efficiency for 6 of 14 *F. vesca* accessions tested. Based on this transformation protocol, we have developed a high throughput system for the systematic generation of a collection of T-DNA-tagged mutant lines and reported production of a set of T-DNA-tagged lines (Oosumi et al. 2006). Here, we present data on the analysis of genomic DNA flanking the T-DNA insertions in transgenic *F. vesca* plants, and show that T-DNA tagging in *F. vesca* can be used as a model system for systematic reverse genetics research in the Rosaceae family.

Materials and methods

Plant transformation and growth conditions

Agrobacterium-mediated transformation and growth of plants of *F. vesca* accessions, PIs 551572, 602578, and 'Alpine' were performed as described previously (Oosumi et al. 2006).

Genomic DNA preparation

Genomic DNA was isolated from folded trifoliate leaves of T_0 transgenic plants using a modified rapid CTAB method (Doyle and Doyle 1990, Lodhi et al. 1994, Porebski et al. 1997). One single folded trifoliate leaf harvested from each plant was ground in liquid nitrogen. Subsequently 180–200 µl of the extraction buffer [100 m*M* Tris–HCl (pH 8.0), 1.4 *M* NaCl, 20 m*M* EDTA (pH 8.0), 2% (w/v) cetyltrimethylammonium bromide, 1% (w/v) polyvinylpyrrolidone-40, 0.2% (v/v) β-mercaptoethanol] was added and vortexed. The extraction mix was heated at 65°C for 15 min. DNA solution was extracted twice with equal volumes of chloroform:isoamyl alcohol (24:1). DNA was precipitated with two volumes of ethanol and 0.5 volumes of 5 *M* NaCl and dissolved in 50 µl TE buffer containing 20 µg ml^{-1} RNase.

PCR amplification of T-DNA flanking sequences

TAIL-PCR (Liu et al. 1995) or hi-TAIL-PCR (Liu and Chen 2007) were used to amplify flanking regions adjacent to the pCAMBIA T-DNA. Primary or secondary TAIL-PCR products were purified using PCR purification kits (Roche, Basel, Switzerland or Qiagen, Valencia, CA), eluted with 50 µl of the elution buffer, and sequenced directly either the Core Laboratory Facility at the Virginia Bioinformatics Institute using a 3730 DNA Analyzer (Applied Biosystems, Foster City, CA) or Tufts University Core Facility using a 3130XL DNA sequencer.

PCR amplification of T-DNA integration sites

For analysis on T-DNA integration sites, in 11 cases where both right and left border flanking sequences had been determined by TAIL-PCR, we designed primers to amplify from the left border sequence to the right border sequence in wild-type strawberry. For each 50 µl PCR reaction, 0.5 U of Ex Taq polymerase (Takara Bio Inc., Tokyo, Japan), 1× Ex Taq polymerase buffer, 200 µ*M* of each dNTP, 2 µ*M* of each primer and about 200 ng DNA of wild-type PI 551572 plants were used. A GeneAmp PCR system 9700 was used for PCR amplifications using the following step cycle program: 94°C, 30 s; 54°C, 30 s and 72°C, 60 s for 30 cycles. The products were purified and sequenced as described above and the sequence obtained compared with the T-DNA mutant left and right border flanking sequences.

Sequence analysis

A homology search was carried out using a BLAST database search (http://www.ncbi.nlm.nih.gov/blast) (Altschul et al. 1990). Three methods (blastn, blastx and tblastx) and two databases (nr and EST) were used. T-DNA integration sites were analyzed by aligning genomic sequences of a wild-type plant and the T-DNA locus in a transgenic plant using the BLAST2 sequence tool (Tatusova and Madden 1999) at NCBI.

Results

Amplification of T-DNA flanking genomic sequences

From 431 T_0 transgenic plants, we obtained 283 RB and 51 LB-TAIL-PCR products ranging in size from 400 bp to 5.0 kb in length with an average size of 1.4 kb (Fig. 1). Genomic sequences flanking RB or LB were determined by direct sequencing TAIL-PCR products. For RB and LB-flanking regions 176 and 37 unique PCR products were sequenced, respectively, ranging in size from 51 to 886 bp of unambiguous sequence.

T-DNA nick position

To determine the T-DNA nick positions in the tagged *F. vesca* genome, we analyzed the sequence of the junction between the T-DNA and genomic DNA. A total of 176 RB and 37 LB junction sequences were analyzed. In the case of RB, T-DNA of 52 of 176 TAIL-PCR products (30%) was nicked after the third base of

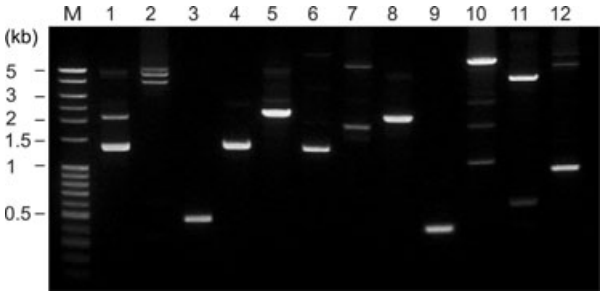


Fig. 1. Secondary TAIL-PCR amplification of *F. vesca* genomic sequences flanking the right border of pCAMBIA-1304 T-DNA using arbitrary degenerate primer 2. M:DNA size marker, lanes 1–12: transgenic plants of PI551572.

the RB, the most frequent nick position (Table 1). This result was similar to previous reports in other plants, such as *Arabidopsis*, rice and tobacco (Gambino et al. 2009, Kim et al. 2003, Sha et al. 2004, Tinland 1996). This site was reported as a cleavage site for generating single-stranded T-DNA fragments (Stachel et al. 1987, Yanofsky et al. 1986). The second most frequent nick position (17%) was after the first base of RB. The deletion of RB was observed in 62 (37%) of 176 transgenic plants. The left border junctions showed much more variation. The most frequent nick position (16%) was after the 23rd base of the LB (Table 2) in 6 of 37 TAIL-PCR products. This site was also a cleavage site reported previously (Stachel et al. 1987, Yanofsky et al. 1986). LB-flanking sequences of the transgenic plant 10–44 had 109 bp of vector backbone sequence caused by reading-through the vector backbone at LB. The deletion of LB was observed in nine (24%) transgenic plants.

Table 1. Analysis of junction sequences of T-DNA RB and genomic DNA in 176 T_0 transgenic plants. T-DNA sequences are shown in lowercase letters and RB sequences are bold and italic. Variable genomic DNA sequences are shown in bold italic '***n***'.

Right boundary T-DNA|right 25-bp border repeat	No. of flanking sequences analyzed
ttaaactatcagtgtt***TGACAGGATATATTGGCGGGTAAAC***	pCAMBIA
ttaaactatcagtgtt***TGACA*nnnnnnnnnnnnnnnnnnnnnnnnn**	2
ttaaactatcagtgtt***TGAC*nnnnnnnnnnnnnnnnnnnnnnnnn**	7
ttaaactatcagtgtt***TGA*nnnnnnnnnnnnnnnnnnnnnnnnn**	52
ttaaactatcagtgtt***TG*nnnnnnnnnnnnnnnnnnnnnnnnn**	19
ttaaactatcagtgtt***T*nnnnnnnnnnnnnnnnnnnnnnnnn**	30
ttaaactatcagtgtt(none)***n*nnnnnnnnnnnnnnnnnnnnnnnnn**	66

Table 2. Analysis of junction sequences of T-DNA LB and genomic DNA in 37 T_0 transgenic plants. T-DNA sequences are shown in lowercase letters and LB sequences are bold upper case italic. Genomic DNA sequences are shown in bold italic '***n***'.

Left boundary T-DNA|left 25-bp border repeat	No. of flanking sequences analyzed
agcgtcaatt***TGTTTACACCACAATATATCCTGCC***	pCAMBIA
agcgtcaatt***TGTTTACACCACAATATATCCTGCC***	3
agcgtcaatt***TGTTTACACCACAATATATCCTGC*n**	1
agcgtcaatt***TGTTTACACCACAATATATCCTG*nn**	6
agcgtcaatt***TGTTTACACCACAATATATCCT*nnn**	1
agcgtcaatt***TGTTTACACCACAATATATCC*nnnn**	2
agcgtcaatt***TGTTTACACCACAATATAT*nnnnnnn**	2
agcgtcaatt***TGTTTACACCACAATAT*nnnnnnnnn**	1
agcgtcaatt***TGTTTACACCACAATA*nnnnnnnnnn**	1
agcgtcaatt***TGTTTACACCACAAT*nnnnnnnnnnnn**	1
agcgtcaatt***TGTTTACACCACAA*nnnnnnnnnnnnn**	4
agcgtcaatt***TGTTTACACCACA*nnnnnnnnnnnnnn**	1
agcgtcaatt***TGTTTACACCA*nnnnnnnnnnnnnnnn**	1
agcgtcaatt***TGTTTACACC*nnnnnnnnnnnnnnnnn**	1
agcgtcaatt***TGTT*nnnnnnnnnnnnnnnnnnnnnnnn**	1
agcgtcaatt***TG*nnnnnnnnnnnnnnnnnnnnnnnnnn**	2
agcgtcaatt(none)***n*nnnnnnnnnnnnnnnnnnnnnnnn**	9

T-DNA integration sites

T-DNA integration sites for 11 transgenic plants were obtained and compared with the native DNA sequence in wild-type PI551572 plants. Deletion of 2–26 bp of native DNA was observed for all transgenic plants analyzed (Fig. 2). These deleted sequences are referred to as target sites (Mayerhofer et al. 1991). We did not find any consensus sequence for T-DNA integration. Alignment of the integration target sites to the T-DNA insertion loci revealed microsimilarities of a few base pairs between the left and/or right ends of the T-DNA and genomic sites in all transgenic plants we tested. Microsimilarities of the left end were observed in 9 of 11 break-points of target deletions. However, microsimilarities of the right end were not necessarily at break-points. In fact, 4 of 11 break-points and their corresponding T-DNA ends were not identical. Moreover, similar sequences downstream of the T-DNA cleavage site were not used as break-points in transgenic plants VT-16, VT-29, VT-41 and VT-87. In these cases, T-DNA was nicked at the cleavage site (VT-29) or after the second base upstream from the RB (the remaining plants).

T-DNA insertion events resulting in a precise replacement of a target site deletion by an insertion with break-points identical to T-DNA ends are referred to as precise target replacements (Mayerhofer et al. 1991). Such events were observed in 7 of 11 transgenic plants. In transgenic plants VT-19, VT-26 and VT-136, microsimilarities of 1–5 bp were observed between both T-DNA ends and the genomic break-points. Significant microsimilarities were also observed between genomic DNA or deleted sequences and T-DNA ends in those plants.

Imprecise junctions were characterized by the presence of additional DNA sequences filling the gap between the break-points of target deletions and T-DNA insertions. These additional DNA sequences are referred to as filler DNA (Mayerhofer et al. 1991). In transgenic plant VT-32, a filler DNA of 22 bp was identified between the left break-point of T-DNA insertion and target deletion. One independent and two overlapping sequence motifs (arrows A, B and C in Fig. 2) were detected in plant genomic DNA just outside the left break-point of target deletion. In transgenic plants VT-15, VT-21 and VT-66, 2–4 bp of filler DNA were identified between the right or left break-point of T-DNA insertion and target deletion.

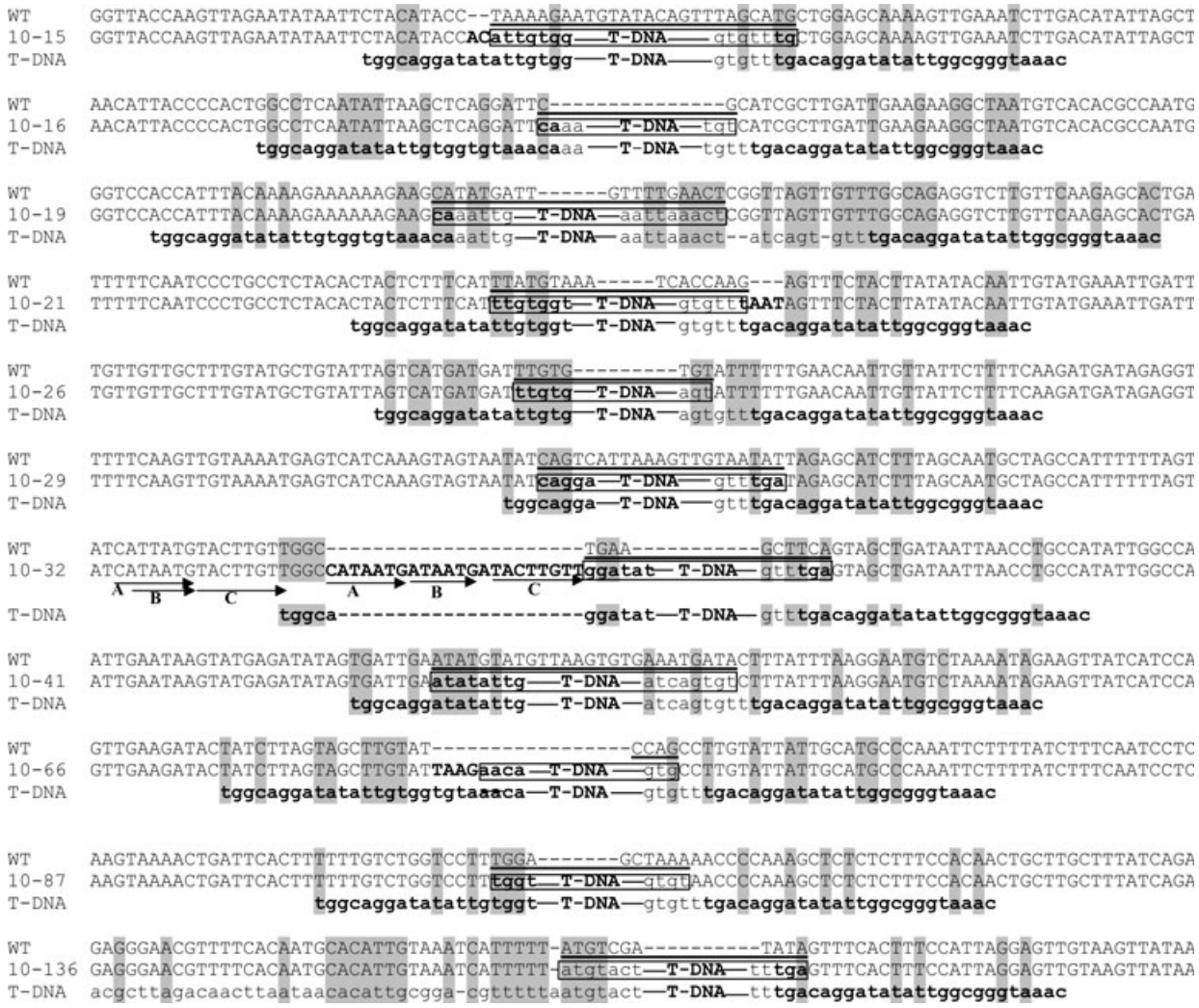


Fig. 2. Comparison of the wild-type (top) and the T-DNA-tagged (middle) loci and T-DNA (bottom). *F. vesca* genomic sequence is shown in uppercase and T-DNA sequence in lowercase letters. The LB and RB sequences are shown in bold lower case. The filler DNAs are shown in bold uppercase and their possible origins are shown by arrows. The T-DNA insertion in transgenic plants is boxed. Identical nucleotides are shaded.

Homology search of T-DNA flanking sequences

To construct a database of *F. vesca* T-DNA tags, we performed a homology search using 213 T-DNA flanking sequences as queries. T-DNA-tagged genomic sequences of 191 transgenic plants were determined for RB, LB or both RB and LB. GenBank accessions of the BLAST hits with alignment scores greater than 50 were detected from nr and EST databases are shown in Table S1 (supporting information). T-DNA-tagged genomic sequences of 79 transgenic plants (41%) had significant similarity to genes, elements, proteins or BAC clones of plant species, including one known strawberry gene encoding alcohol dehydrogenase (GenBank accession number X15588.1). Tagged sequences for 126 of 191 transgenic plants (66%) had homology to plant ESTs including 85 *F. vesca* ESTs, 32 ESTs of other rosaceous species and 2 ESTs of fruit species. Most of the 85 sequences showing homology to *F. vesca* ESTs also had similarity to ESTs of other rosaceous species. Forty-eight had no significant similarity to any gene or protein in the database, while they had similarity to ESTs, 67 had similarity neither to any gene or protein nor to ESTs in the database with alignment scores greater than 50.

Discussion

In the present study, we report the development of efficient T-DNA tagging in *F. vesca* as a model for insertional mutagenesis in Rosaceae. Our data show

that the TAIL-PCR method (Liu and Chen 2007, Liu et al. 1995) can be used efficiently to amplify *F. vesca* genomic sequence flanking T-DNA insertion. Our amplification frequencies were similar to those reported for *Arabidopsis* (Szabados et al. 2002). The average size of TAIL-PCR products was greater for RB than for LB. Although the majority of TAIL-PCR products from LB ranged from 0.5 to 3 kb, RB-TAIL-PCR produced products ranging from 0.5 to 6.5 kb. The distribution of LB-TAIL-PCR products in *F. vesca* was similar to that of long-range iPCR (Li-PCR) products in *Arabidopsis* (Szabados et al. 2002), in which 90% of PCR products were 0–3 kb in size.

Analysis of the junctions between the T-DNA borders and *F. vesca* genomic DNA revealed that T-DNA nick positions were similar to those reported for *Arabidopsis*, tobacco, grape and rice (Gambino et al. 2009, Kim et al. 2003, Sha et al. 2004, Tinland 1996). Alignment of the integration target sites with the T-DNA insertion loci revealed microsimilarities of a few base pairs between the left and/or right ends of the T-DNA and genomic sites in all transgenic plants tested. Analogous results have been observed in grapevine obtained by *Agrobacterium*-mediated transformation where perfect microsimilarities were found not only near the LB but also near the RB (Gambino et al. 2009). Precise and imprecise junctions between T-DNA ends and target sequences were observed. These results indicated that T-DNA was integrated into the *F. vesca* genome by illegitimate recombination as reported in tobacco, *Arabidopsis*, aspen and rice (Brunaud et al. 2002, Gheysen et al. 1991, Kim et al. 2003, Kumar and Fladung 2002, Matsumoto et al. 1990, Mayerhofer et al. 1991, Wang et al. 2005).

We have reported that 154 of 213 (72%) of the T-DNA-tagged genomic sequences showed homology to plant genes, proteins and ESTs indicating that about 60% of T-DNAs were integrated into genetic regions. Although we do not yet know if these genes are active, this frequency is close to that reported in *Arabidopsis* and rice (An et al. 2003, Szabados et al. 2002). Folta et al. (2005) reported that 64% of unigenes of octoploid strawberry *F* × *ananassa* had significant homology to other Rosaceae ESTs. Our data in *F. vesca* showed that 85 of 126 (67%) T-DNA-tagged sequences of putative genes had homology to other Rosaceae ESTs. These results suggest that about half of T-DNA-tagged genomic sequences can transfer to other rosaceous species via ESTs as sequence data correlated with phenotype information.

EST projects in fruit crops are in progress and EST databases are still developing. It is possible that an increase of EST data accompanies an increase of the percentage of tagged sequences with homology to other sequences in the GenBank database. Horn et al. (2005) generated about 10 000 ESTs from a peach fruit and defined a putative peach unigene set of 3842 ESTs. Seventy-six percentage of the ESTs had homology in the GenBank. Newcomb et al. (2006) collected about 152 000 ESTs from different tissues of apple and created a set of 42 938 non-redundant sequences (NRs); 94% of apple NRs had matches in predicted proteins of *Arabidopsis*. For tagged sequences of 126 *F. vesca* transgenic plants that had matches in plant ESTs, 48 (38%) did not have any significant similarity to any gene or protein in the GenBank database. One possibility is that predicted proteins of these putative genes have limited conserved domains. In this case, if an EST database of the Rosaceae family saturates, it would be possible that the percentage of tagged sequences with homology to genes and proteins increases to the level that of apple NRs.

This is the first report of T-DNA integration in a sizeable population of a rosaceous species. There is no guarantee that integration is similar across species and a population of several hundred independently transformed plants with the same vector has not been available in many species. We have shown in this paper that T-DNA integration in strawberry is not random but directed by sequence microsimilarities in the host genome. This is important in the selection of a vector for reverse genetics as the sequence of the T-DNA borders can be expected to influence integration sites. Our strategy has changed from generating single knockout mutants by independent *Agrobacterium*-mediated transformation using pCAMBIA vectors to transposon tagging. The set of genetic tools in rosaceous crops is under development and increasing rapidly with the availability of genome sequences, genome browsers, inbred lines and insertional mutants as we describe here. It is important for the community to understand not only the importance of these tools but also their limitations.

The pCAMBIA mutants were especially adept at integration into genomic regions and will provide a valuable resource. Obtaining a sufficient number of these mutants to interrupt every gene in the strawberry genome, however, would be prohibitively expensive in time and resources, making transposon tagging far more attractive. If similar principles hold for integration of dissociator elements into the strawberry genome then transposon insertion sites may differ significantly from those of the pCAMBIA mutants.

Acknowledgements – This project was funded in part by a Virginia Tech ASPIRES grant, USDA/NRI award no. 2008-35300-04458 and USDA Proj. No. VA-135816. We thank Dr Joel Shuman, Hope Gruszewski, Kerri Mills, Sarah H. Holt and Benjamin Morris for help in plant growth

and GFP screening. We thank Selene Willis for help in tissue collection for DNA preparation. We also thank Jim Walke for critical reading of this manuscript.

References

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

An SY, Park S, Jeong DH, Lee DY, Kang HG, Yu JH, Hur J, Kim SR, Kim YH, Lee M, Han SK, Kim SJ, Yang JW, Kim E, Wi SJ, Chung HS, Hong JP, Choe V, Lee HK, Choi JH, Nam JM, Park PB, Park KY, Kim WT, Choe S, Lee CB, An GH (2003) Generation and analysis of end sequence database for T-DNA tagging lines in rice. Plant Physiol 133: 2040–2047

Baird WV, Estager AS, Wells JK (1994) Estimating nuclear DNA content in peach and related diploid species using laser flow cytometry and DNA hybridization. J Am Soc Hortic Sci 119: 1312–1316

Battey NH, Mière PI, Tehranifar A, Cekic C, Taylor S, Shrives KJ, Hadley P, Greenland AJ Darby J, Wilkinson MJ (1998) Genetic and environmental control of flowering in strawberry. In: Cockshull KE, Gray D, Seymour GB, Thomas B (eds) Genetic and Environmental Manipulation of Horticultural Crops. CAB International, Wallingford, UK, pp 111–131

Brunaud V, Balzergue S, Dubreucq B, Aubourg S, Samson F, Chauvin S, Bechtold N, Cruaud C, DeRose R, Pelletier G, Lepiniec L, Caboche M, Lecharny A (2002) T-DNA integration into the Arabidopsis genome depends on sequences of pre-insertion sites. EMBO Rep 3: 1152–1157

Davis TM, Yu H (1997) A linkage map of the diploid strawberry, *Fragaria vesca*. J Hered 88: 215–221

Dirlewanger E, Graziano E, Joobeur T, Garriga-Caldere F, Cosson P, Howad W, Arus P (2004) Comparative mapping and marker-assisted selection in Rosaceae fruit crops. Proc Natl Acad Sci USA 101: 9891–9896

Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue. Focus 12: 13–15

Folta KM, Davis TM (2006) Strawberry genes and genomics. Crit Rev Plant Sci 25: 399–415

Folta KM, Staton M, Stewart PJ, Jung S, Bies DH, Jesdurai C, Main D (2005) Expressed sequence tags (ESTs) and simple sequence repeat (SSR) markers from octoploid strawberry (*Fragaria* × *ananassa*). BMC Plant Biol 5: 12.

Galbiati M, Moreno MA, Nadzan G, Zourelidou M, Dellaporta SL (2000) Large-scale T-DNA mutagenesis in Arabidopsis for functional genomic analysis. Funct Integr Genomics 1: 25–34

Gambino G, Chitarra W, Maghuly F, Laimer M, Boccacci P, Torello Marinoni D, Gribaudo I (2009) Characterization of T-DNA insertions in transgenic grapevines obtained by *Agrobacterium*-mediated transformation. Mol Breed 24: 305–320

Gheysen G, Villarroel R, Vanmontagu M (1991) Illegitimate recombination in plants: a model for T-DNA integration. Genes Dev 5: 287–297

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 YS, Kim HR, Jesudurai C, Sosinski B, Arús P, Baird V, Parfitt D, Reighard G, Scorza R, Tomkins J, Wing R, Abbott AG (2005) Candidate gene database and transcript map for peach, a model species for fruit trees. Theor Appl Genet 110: 1419–1428

Jung S, Staton M, Lee T, Blenda A, Svancara R, Abbott A, Main D (2008) GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. Nucleic Acids Res 36: D1034–D1040

Kim SR, Lee J, Jun SH, Park S, Kang HG, Kwon S, An G (2003) Transgene structures in T-DNA-inserted rice plants. Plant Mol Biol 52: 761–773

Krysan PJ, Young JC, Sussman MR (1999) T-DNA as an insertional mutagen in *Arabidopsis*. Plant Cell 11: 2283–2290

Krysan PJ, Young JC, Tax F, Sussman MR (1996) Identification of transferred DNA insertions within Arabidopsis genes involved in signal transduction and ion transport. Proc Natl Acad Sci USA 93: 8145–8150

Kumar S, Fladung M (2002) Transgene integration in aspen: structures of integration sites and mechanism of T-DNA integration. Plant J 31: 543–551

Liu YG, Chen Y (2007) High-efficiency thermal asymmetric interlaced PCR for amplification of unknown flanking sequences. BioTechniques 43: 649–656

Liu YG, Mitsukawa N, Oosumi T, Whittier RF (1995) Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. Plant J 8: 457–463

Lodhi MA, Ye GN, Weeden NF, Reisch BI (1994) A simple and efficient method for DNA extraction from grapevine cultivars and *Vitis* species. Plant Mol Biol Rep 12: 6–13

Matsumoto S, Ito Y, Hosoi T, Takahashi Y, Machida Y (1990) Integration of *Agrobacterium* T-DNA into a tobacco chromosome: possible involvement of DNA homology between T-DNA and plant DNA. Mol Gen Genet 224: 309–316

Mayerhofer R, Konczkalman Z, Nawrath C, Bakkeren G, Crameri A, Angelis K, Redei GP, Schell J, Hohn B, Koncz C (1991) T-DNA integration – a mode of illegitimate recombination in plants. EMBO J 10: 697–704

McKinney EC, Ali N, Traut A, Feldmann KA, Belostotsky DA, McDowell JM, Meagher RB (1995) Sequence-based identification of T-DNA insertion mutations in Arabdopsis: actin mutants *act2-1* and *act4-1*. Plant J 8: 613–622

Newcomb RD, Crowhurst RN, Gleave AP, Rikkerink 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, Yauk YK (2006) Analyses of expressed sequence tags from apple. Plant Physiol 141: 147–166

Ochman H, Gerber AS, Hartl DL (1988) Genetic applications of an inverse polymerase chain reaction. Genetics 120: 621–623

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: 1219–1230

Porebski S, Bailey LG, Baum BR (1997) Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. Plant Mol Biol Rep 15: 8–15

Rios G, Lossow A, Hertel B, Breuer F, Schaefer S, Broich M, Kleinow T, Jasik J, Winter J, Ferrando A, Farras R, Panicot M, Henriques R, Mariaux JB, Oberschall A, Molnar G, Berendzen K, Shukla V, Lafos M, Koncz Z, Redei GP, Schell J, Koncz C (2002) Rapid identification of Arabidopsis insertion mutants by non-radioactive detection of T-DNA tagged genes. Plant J 32: 243–253

Ryu CH, You JH, Kang HG, Hur JH, Kim YH, Han MJ, An KS, Chung BC, Lee CH, An GH (2004) Generation of T-DNA tagging lines with a bidirectional gene trap vector and the establishment of an insertion-site database. Plant Mol Biol 54: 489–502

Sargent DJ, Clarke J, Simpson DW, Tobutt KR, Arús 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: 1349–1359

Sargent DJ, Davis TM, Tobutt KR, Wilkinson MJ, Battey NH, Simpson DW (2004) A genetic linkage map of microsatellite, gene-specific and morphological markers in diploid *Fragaria*. Theor Appl Genet 109: 1385–1391

Sargent DJ, Marchese A, Simpson DW, Howad W, Fernández-Fernández F, Monfort A, Arús P, Evans KM, Tobutt KR (2009) Development of 'universal' gene-specific markers from *Malus* spp. cDNA sequences, their mapping and use in synteny studies within Rosaceae. Tree Genet Genom 5: 133–145

Schupp JM, Price LB, Klevytska A, Keim P (1999) Internal and flanking sequence from AFLP fragments using ligation-mediated suppression PCR. BioTechniques 26: 905

Sha Y, Li S, Pei Z, Luo L, Tian Y, He C (2004) Generation and flanking sequence analysis of a rice T-DNA tagged population. Theor Appl Genet 108: 306–314

Shulaev V, Korban SS, Sosinski B, Abbott AG, Aldwinckle HS, Folta KM, Iezzoni A, Main D, Arús P, Dandekar AM, Lewers K, Brown SK, Davis TM, Gardiner SE, Potter D, Veilleux RE (2008) Multiple models for Rosaceae genomics. Plant Physiol 147: 985–1003

Sosinski B, Shulaev V, Korban SS, Dhingra A, Kalyanaraman A, Bumgarner R, Rokhsar D, Verde I, Velasco R, Abbott AG (2009) Rosaceous genome sequencing: Perspectives and progress. In: Folta KM, Gardiner SE (eds) Genetics and Genomics of Rosaceae. Springer, New York, pp 601–616

Stachel SE, Timmerman B, Zambryski P (1987) Activation of *Agrobacterium tumefaciens vir* gene expression generates multiple single-stranded T-strand molecules from the pTiA6 region: requirements for 5′ *virD* gene products. EMBO J 6: 857–863

Szabados L, Kovacs I, Oberschall A, Abraham E, Kerekes I, Zsigmond L, Nagy R, Alvarado M, Krasovskaja I, Gal M, Berente A, Redei GP, Ben Haim A, Koncz C (2002) Distribution of 1000 sequenced T-DNA tags in the Arabidopsis genome. Plant J 32: 233–242

Tatusova TA, Madden TL (1999) BLAST 2 SEQUENCES, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 174: 247–250

Theuns I, Windels P, De Buck S, Depicker A, Van Bockstaele E, De Loose M (2002) Identification and characterization of T-DNA inserts by T-DNA fingerprinting. Euphytica 123: 75–84

Tinland B (1996) The integration of T-DNA into plant genomes. Trends Plant Sci 1: 178–184

Tissier AF, Marillonnet S, Klimyuk V, Patel K, Torres MA, Murphy G, Jones JDG (1999) Multiple independent defective suppressor-mutator transposon insertions in Arabidopsis: a tool for functional genomics. Plant Cell 11: 1841–1852

Wang J, Li L, Shi ZY, Wan XS, An LS, Zhang JL (2005) T-DNA integration category and mechanism in rice genome. J Integr Plant Biol 47: 350–361

Winkler RG, Frank MR, Galbraith DW, Feyereisen R, Feldmann KA (1998) Systematic reverse genetics of transfer-DNA-tagged lines of Arabidopsis – isolation of mutations in the cytochrome P450 gene superfamily. Plant Physiol 118: 743–749

Yanofsky MF, Porter SG, Young C, Albright LM, Gordon MP, Nester EW (1986) The *virD* operon of *Agrobacterium tumefaciens* encodes a site-specific endonuclease. Cell 47: 471–477

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. BLAST analyses and predicted gene function of 213 T-DNA flanking strawberry sequences generated by TAIL-PCR from GenBank, (http://www.ncbi.nlm.nih.gov/), as well as sequence ID, GenBank accession number, and species where it was identified.

Edited by R. Mittler

Table S1. BLAST analyses and predicted gene function of 213 T-DNA flanking strawberry sequences generated by TAIL-PCR from Genbank, (http://www.ncbi.nlm.nih.gov/), as well as sequence ID, GenBank accession number, and species where it was identified.

Sequence ID	Blast hits (Genebank accession, alignment score)	
	Gene, element, protein or BAC clon	EST
VT-002R	None	*Fragaria vesca* (EX659972.1)
VT-A02???	None	*None*
VT-003R	(AAA33948.1) Calmodulin (*Glycine max*)	*Populus tremula* (BU862668.1)
VT-A03R	None	*Fragaria vesca* (DY668798.1)
VT-A04R	None	*Fragaria vesca* (EX659431.1)
VT-A05R	(XP_002527022.1) Conserved hypothetical protein (*Ricinus communis*)	*Fragaria vesca* (DY673897.1)
VT-008R	(EEF41624.1) Peroxisome assembly factor-2, putative (*Ricinus communis*)	*Prunus persica* (DW343755.1)
VT-A08R	(XP_002521955.1) Associate of C-myc, putative (*Ricinus communis*)	*Fragaria vesca* (DY673987.1)
VT-009R	(XM_002298609.1) Hydroxycinnamoyl CoA shikimate/quinate hydroxycinnamoyltransferase-like protein (HCQL6), mRNA (*Populus trichocarpa*)	*Fragaria vesca* (DY669783.1)
VT-A09R	(X71654.1) CYP71A2 mRNA for hydroxylase (*Solanum melongena*)	*Malus ×domestica* (CV186493.1)
VT-010L	(XP_002529754.1) Isoleucyl tRNA synthetase, putative(*Ricinus communis*)	*Fragaria vesca* (DY673231.1)
VT-010R	(XP_002318916.1) Membrane bound o-acyl transferase (*Populus trichocarpa*)	*Fragaria vesca* (DV438852.1)
VT-A11R	None	None
VT-012R	None	*Oryza sativa* (CI304327.1)
VT-A12R	(ABA94690.1) Lipin, N-terminal conserved region family protein, expressed (*Oryza sativa*)	*Malus ×domestica* (CN876923.1)
VT-014R	None	None
VT-015L	None	*Fragaria vesca* (EX684192.1)
VT-015R	None	*Fragaria vesca* (EX684192.1)
VT-016L	None	None
VT-016R	None	None
VT-017R	None	None
VT-018L	None	*None*
VT-018R	None	*Panicum virgatum* (FL798271.1)
VT-A18R	None	*Fragaria vesca* (DY668080.1)
VT-019L	None	*Fragaria vesca* (DY667977.1)
VT-019R	None	*Fragaria vesca* (DY667977.1)
VT-020L	None	None
VT-020R	None	None
VT-021L	(AAD49734.1) Glutamine synthetase precursor (*Juglans nigra*)	*Fragaria vesca* (DY674547.1)
VT-021R	(AAD49734.1) Glutamine synthetase precursor (*Juglans nigra*)	*Fragaria vesca* (DY674547.1)

VT-022R	(XP_002313228.1) Predicted protein (*Populus trichocarpa*)	*Fragaria vesca* (DY673563.1)
VT-A23R	(XP_002300628.1) Predicted protein (*Populus trichocarpa*)	*Fragaria vesca* (EX671983.1)
VT-A24R	None	None
VT-025R	(EEF46369.1) Stem 28 kDa glycoprotein precursor, putative (*Ricinus communis*)	*Malus ×domestica* (GO497633.1)
VT-A25R	None	*Fragaria vesca* (EX685821.1)
VT-026L	None	*Prunus persica* (DW344278.1)
VT-026R	None	*Fragaria vesca* (DY670903.1)
VT-A26R	(YP_762482.1) Rribosomal protein S3 (*Tripsacum dactyloides*)	*Fragaria vesca* (EX670071.1)
VT-028L	None	None
VT-028R	None	None
VT-029L	(NP_193580.1) Ethylene-responsive factor, putative (*Arabidopsis thaliana*)	*Vitis vinifera* (FC066471.1)
VT-029R	(NP_193580.1) Ethylene-responsive factor, putative (*Arabidopsis thaliana*)	*Vitis vinifera* (FC066471.1)
VT-031R	None	*Fragaria vesca* (EX685695.1)
VT-032L	(EEF41851.1) RNA binding motif protein, putative (*Ricinus communis*)	*Fragaria vesca* (EX683367.1)
VT-032R	(EEF41851.1) RNA binding motif protein, putative (*Ricinus communis*)	*Fragaria vesca* (EX683367.1)
VT-A32R	None	None
VT-033L	None	*Fragaria vesca* (GH202415.1)
VT-033R	(XP_002525893.1) Glutamyl-tRNA (gln) amidotransferase subunit A, putative (*Ricinus communis*)	*Fragaria vesca* (GH202415.1)
VT-034R	None	*Fragaria vesca* (EX685750.1)
VT-035R	(XP_002532062.1) Glucan endo-1,3-beta-glucosidase precursor, putative (*Ricinus communis*)	*Fragaria vesca* (GH201886.1)
VT-036R	(X15588.1) Strawberry Adh gene for alcohol dehydrogenase (*Fragaria ×ananassa*)	*Malus ×domestica* (GO518826.1)
VT-038R	None	*Fragaria vesca* (EX679240.1)
VT-039R	(XP_002278257.1) Predicted: hypothetical protein (*Vitis vinifera*)	*Malus ×domestica* (GO500904.1)
VT-041L	None	*Fragaria vesca* (DY670791.1)
VT-041R	None	*Fragaria vesca* (DY670791.1)
VT-042R	(XP_002528849.1) Serine-threonine protein kinase, plant-type, putative (*Ricinus communis*)	None
VT-043L	None	*Prunus persica* (DY646081.1)
VT-043R	None	None
VT-044L	None	None
VT-044R	None	None
VT-051L	None	None
VT-052R	None	None
VT-054L	(XP-002313425.1) Cytochrome P450 (*Populus trichocarpa*)	*Malus ×domestica* (CO418222.1)
VT-054R	None	None
VT-057R	(XP_002521694.1) Adipocyte plasma membrane-associated protein, putative (*Ricinus communis*)	*Fragaria vesca* (EX668994.1)

VT-058L	None	*Fragaria vesca* (EX681135.1)
VT-059R	(XP_002522973.1) Vacuolar sorting protein SNF8, putative (*Ricinus communis*)	*Fragaria vesca* (EX660290.1)
VT-063L	None	None
VT-063R	None	*Malus ×domestica* (GO538836.1)
VT-066L	None	None
VT-066R	None	None
VT-067L	(XP_002530505.1) Protein trm112, putative (*Ricinus communis*)	*Fragaria ×ananassa* (CO816790.1)
VT-067R	(XP_002314168.1) GRAS family transcription factor (*Populus trichocarpa*)	*Prunus persica* (FC867070.1)
VT-071L	(XP_002510286.1) Amino acid transporter, putative (*Ricinus communis*)	*Malus ×domestica* (CV150930.1)
VT-072R	(XP_002510457.1) ATP-dependent peptidase, putative (*Ricinus communis*)	*Fragaria vesca* (EX679614.1)
VT-074L	None	*Fragaria vesca* (EX669187.1)
VT-074R	None	None
VT-075R	None	*None*
VT-076R	(EEF49342.1) Lysosomal alpha-mannosidase, putative (*Ricinus communis*)	*Fragaria vesca* (DV438694.1)
VT-077R	None	None
VT-081L	(BAF44192.1) SNF1-related kinase (*Solanum lycopersicum*)	*Fragaria vesca* (DY667382.1)
VT-082R	None	None
VT-086L	None	*None*
VT-087L	(Q75HC2.2) RecName: Full=Protein argonaute 7; Short=OsAGO7 (*Oryza sativa*)	*Aristolochia fimbriata* (FD758003.1)
VT-087R	(Q75HC2.2) RecName: Full=Protein argonaute 7; Short=OsAGO7 (*Oryza sativa*)	*Aristolochia fimbriata* (FD758003.1)
VT-103R	None	*Fragaria vesca* (GH203245.1)
VT-136L	None	*None*
VT-136R	(EEF37316.1) Erythroblast macrophage protein emp (*Ricinus communis*)	*Fragaria vesca* (EX676800.1)
VT-351R	None	None
VT-352R	None	None
VT-357R	(XP_002531299.1) Calcium ion binding protein, putative (*Ricinus communis*)	*Fragaria vesca* (DY667959.1)
VT-360R	None	*Fragaria vesca* (GO578585.1)
VT-361R	(NP_172146.2) Unknown protein (*Arabidopsis thaliana*)	*Fragaria vesca* (DY668345.1)
VT-362R	(NP_850319.1) Radical SAM domain-containing protein (*Arabidopsis thaliana*)	*Fragaria vesca* (EX656886.1)
VT-365R	None	*Fragaria ×ananassa* (GO578585.1)
VT-367R	(XP_002311340.1) Predicted protein (*Populus trichocarpa*)	***Malus ×domestica*** (EB131731.1)
VT-368R	(XP_002519354.1) Anthranilate N-benzoyltransferase protein, putative (*Ricinus communis*)	*Prunus persica* (AM291707.1)
VT-370R	(XP_002527245.1) Lyase, putative (*Ricinus communis*)	None
VT-371R	None	None
VT-372R	(BAE95414.1) Mitogen-activated protein kinase kinase (*Nicotiana benthamiana*)	*Fragaria vesca* (EX670048.1)

VT-374R	None	None
VT-379R	(XP_002515949.1) Stem 28 kDa glycoprotein precursor, putative (*Ricinus communis*)	*Malus ×domestica* (GO562364.1)
VT-380R	(XP_002305325.1) Predicted protein (*Populus trichocarpa*)	*Fragaria ×ananassa* (CO817978.1)
VT-386R	(XP_002440915.1) hypothetical protein SORBIDRAFT_09g016440 (*Sorghum bicolor*)	*Fragaria vesca* (EX663441.1)
VT-387R	None	None
VT-388R	None	*Nicotiana benthamiana* (CK298208.1)
VT-390R	None	None
VT-394R	None	None
VT-395R	None	*Fragaria vesca* (EX661436.1)
VT-396R	None	None
VT-400R	None	*Arachis hypogaea* (GO329515.1)
VT-403R	None	*Fragaria vesca* (EX687219.1)
VT-406R	(XP_002313425.1) Cytochrome P450 (*Populus trichocarpa*)	*Malus ×domestica* (CO418222.1)
VT-409R	None	None
VT-411R	(XP_002281043.1) Predicted: hypothetical protein (*Vitis vinifera*)	None
VT-416R	(XP_002283946.1) Hypothetical protein (*Vitis vinifera*)	None
VT-424R	None	None
VT-435R	None	None
VT-437R	None	*Fragaria vesca* (EX673103.1)
VT-448R	None	*Fragaria vesca* (GO578585.1)
VT-449R	(XP_002533036.1) Magnesium and cobalt efflux protein corC, putative (*Ricinus communis*)	*Malus ×domestica* (GO562508.1)
VT-451R	None	*Fragaria vesca* (EX668351.1)
VT-453R	None	*Fragaria vesca* (DY675759.1)
VT-461R	None	None
VT-465R	(XP_002324759.1) Predicted protein (*Populus trichocarpa*)	*Fragaria vesca* (EX683670.1)
VT-467R	None	*Malus ×domestica* (GO537620.1)
VT-468R	None	None
VT-473R	(XP_002517122.1) Conserved hypothetical protein (*Ricinus communis*)	*Fragaria vesca* (EX677160.1)
VT-476R	None	*Fragaria vesca* (DY671427.1)
VT-480R	(XP_002509487.1) Phenylalanyl-tRNA synthetase beta chain, putative (*Glycine max*)	*Fragaria vesca* (EX667479.1)
VT-482R	(XP_002529455.1) Proline synthetase associated protein, putative (*Ricinus communis*)	*Fragaria vesca* (EX683911.1)
VT-485R	None	*Rosa montezumae* (BQ106418.1)
VT-487R	(CAN63667.1)Hypothetical protein (*Vitis vinifera*)	*Fragaria vesca* (EX671127.1)
VT-489R	(ABA62612.1) Expansin (*Fragaria ×ananassa*)	*Fragaria vesca* (EX666094.1)
VT-491R	(XP_002534105.1) Spotted leaf protein, putative (*Ricinus communis*)	*Fragaria vesca* (CX661914.1)

VT-496R	None	*Fragaria vesca* (DY674246.1)
VT-497R	(XP_002308949.1) Predicted protein (*Populus trichocarpa*)	*Fragaria vesca* (DY672522.1)
VT-499R	(XP_002518809.1) Phytosulfokine receptor precursor, putative (*Ricinus communis*)	*Prunus persica* (DY636269.1)
VT-500R	(XP_002456163.1) Hypothetical protein SORBIDRAFT_03g031480 (*Sorghum bicolor*)	*Fragaria vesca* (EX687133.1)
VT-507R	None	*Brassica napus* (EV223117.1)
VT-511R	None	None
VT-514R	(NP_001119319.1) Disease resistance protein (TIR-NBS-LRR class), (*Arabidopsis thaliana*)	None
VT-515R	None	*Fragaria vesca* (EX661133.1)
VT-518R	None	*Fragaria vesca* (DY672027.1)
VT-520R	None	*Fragaria vesca* (DY675759.1)
VT-523R	None	*Fragaria vesca* (DV438279.1)
VT-534R	None	None
VT-639R	None	None
VT-642R	(XP_002284423.1)| Predicted: hypothetical protein (*Vitis vinifera*)	*Malus ×domestica* (GO522501.1)
VT-645R	None	*Rosa montezumae* (BQ104039.1)
VT-647R	None	*Fragaria vesca* (CO382081.1)
VT-653R	None	*Malus ×domestica* (CN911892.1)
VT-662R	None	None
VT-665R	(NP_198669.1) Tryptophan synthase-related (*Arabidopsis thaliana*)	*Malus ×domestica* (DT003514.1|)
VT-670R	None	*Fragaria vesca* (EX683617.1|)
VT-671R	None	None
VT-678R	(XP_002527483.1) Protein phosphatase 2c, putative(*Ricinus communis*)	*Fragaria vesca* (DY668463.1)
VT-688R	None	None
VT-695R	(Q0E0Y3.2) Full=RNA pseudourine synthase 7 (*Oryza sativa*)	*Prunus persica* (FC864610.1)
VT-696R	(AAD17395.1) Putative non-LTR retroelement reverse transcriptase (*Arabidopsis thaliana*)	None
VT-697L	(NP_181016.1) ACD11 (Accelerated cell death 11); sphingosine transmembrane transporter (*Arabidopsis thaliana*)	*Fragaria vesca* (DY672522.1)
VT-698R	None	None
VT-705L	None	None
VT-706R	None	None
VT-708R	(XP_002517487.1) Aldose-1-epimerase, putative (*Ricinus communis*)	*Fragaria vesca* (EX666191.1)
VT-709R	None	None
VT-713L	None	None
VT-719R	None	*Fragaria vesca* (EX678080.1)
VT-726R	None	*Brassica napus* (EV223117.1)
VT-728R	(NP_190817.1) Heat shock protein-related. (*Arabidopsis thaliana*)	None

VT-729R	(NP_190817.1) Heat shock protein-related (*Arabidopsis thaliana*)	*Malus ×domestica* (CN911130.1)
VT-732R	None	*Fragaria vesca* (EX665050.1)
VT-733L	(AAM65752.1) Putative glucosyltransferase (*Arabidopsis thaliana*)	*Fragaria vesca* (DY667402.1)
VT-745R	(XP_002512536.1) Beta-fructofuranosidase, putative (*Ricinus communis*)	*Rosa luciae* (EC588757.1)
VT-749R	None	None
VT-754R	Galactoside 2-alpha-L-fucosyltransferase, putative (*Ricinus communis*)	*Malus ×domestica* (CO904049.1)
VT-755R	None	*Malus ×domestica* (CO899840.1)
VT-758R	None	None
VT-759R	None	None
VT-760L	None	None
VT-762R	None	*Fragaria vesca* (EX674682.1)
VT-764R	(XP_002511705.1) Conserved hypothetical protein (*Ricinus communis*)	*Fragaria vesca* (DV438287.1)
VT-766L	(XP_002532229.1) Transparent Testa 12 protein, putative (*Ricinus communis*)	*Prunus persica* (DY646522.1)
VT-767R	(XP_002526996.1) Conserved hypothetical protein (*Ricinus communis*)	*Fragaria vesca* (EX673693.1)
VT-768R	None	None
VT-770R	(XP_002520883.1) Conserved hypothetical protein (*Ricinus communis*.)	*Fragaria vesca* (EX672516.1)
VT-771L	None	*Fragaria vesca* (EX663019.1)
VT-777R	None	*Fragaria vesca* (EX658746.1)
VT-778R	(XP_002510601.1) Tetracycline transporter, putative (*Ricinus communis*.)	*Fragaria vesca* (EX681481.1)
VT-779R	None	*Fragaria vesca* (EX670071.1)
VT-781R	None	None
VT-785R	(XP_002297736.1) cc-nbs-lrr resistance protein (*Populus trichocarpa*)	None
VT-789R	None	None
VT-791L	None	None
VT-793R	None	*Fragaria vesca* (EX688373.1)
VT-795R	None	*Fragaria vesca* (DV440340.1)
VT-800R	(XP_002526021.1) Vitellogenic carboxypeptidase, putative (*Ricinus communis*)	*Fragaria vesca* (DY669426.1)
VT-808R	None	None
VT-810R	None	None
VT-814R	(BAE97370.1) Endo-beta-mannosidase (*Brassica oleracea*)	*Fragaria vesca* (DY667724.1)
VT-815L	None	None
VT-821R	None	*Brassica napus* (EV223117.1)
VT-822R	None	None
VT-827L	(ABK60177.1) Putative reverse transcriptase (*Zingiber officinale*)	*Zea mays* (FL432930.1)
VT-831R	(ACJ37406.1) Stress-induced receptor-like kinase (*Glycine max*)	*Fragaria vesca* (DY671345.1)

VT-832R	None	None
VT-839R	None	None
VT-842R	None	None
VT-857R	None	*Fragaria vesca* (EX688115.1)
VT-858R	(XP_002329162.1) Tir-nbs-lrr resistance protein (*Populus trichocarpa*)	*Fragaria vesca* (EX666722.1)
VT-863R	(ACH61565.1) Truncated MADS-box transcription factor (*Pseudotsuga menziesii*)	*Arachis hypogaea* (GO329563.1)
VT-866R	None	*Fragaria vesca* (DY675759.1|)
VT-867R	None	None