2011-07-01

Genome Snapshot and Molecular Marker Development in *Penstemon* (Plantaginaceae)

Rhyan B. Dockter
Brigham Young University - Provo

Genome Snapshot and Molecular Marker Development in *Penstemon* (Plantaginaceae)

Rhyan B. Dockter

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Mikel R. Stevens
Brad Geary
P. Jeff Maughan
Leigh A. Johnson

Department of Plant and Wildlife Sciences

Brigham Young University

August, 2011

ABSTRACT

Genome Snapshot and Molecular Marker Development in
Penstemon (Plantaginaceae)

Rhyan B. Dockter
Department of Plant and Wildlife Sciences
Master of Science

Penstemon Mitchell (Plantaginaceae) is one of the largest, most diverse plant genera in North America. Their unique diversity, paired with their drought-tolerance and overall hardiness, give *Penstemon* a vast amount of potential in the landscaping industry—especially in the more arid western United States where they naturally thrive. In order to develop *Penstemon* lines for more widespread commercial and private landscaping use, we must improve our understanding of the vast genetic diversity of the genus on a molecular level. In this study we utilize genome reduction and barcoding to optimize 454-pyrosequencing in four target species of *Penstemon* (*P. cyananthus*, *P. davidsonii*, *P. dissectus* and *P. fruticosus*). Sequencing and assembly produced contigs representing an average of 0.5% of the *Penstemon* species. From the sequence, SNP information and microsatellite markers were extracted. One hundred and thirty-three interspecific microsatellite markers were discovered, of which 50 met desired primer parameters, and were of high quality with readable bands on 3% Metaphor gels. Of the microsatellite markers, 82% were polymorphic with an average heterozygosity value of 0.51. An average of one SNP in 2,890 bp per species was found within the individual species assemblies and one SNP in 97 bp were found between any two supposed homologous sequences of the four species. An average of 21.5% of the assembled contigs were associated with putative genes involved in cellular components, biological processes, and molecular functions. On average 19.7% of the assembled contigs were identified as repetitive elements of which LTRs, DNA transposons and other unclassified repeats, were discovered. Our study demonstrates the effectiveness of using the GR-RSC technique to selectively reduce the genome size to putative homologous sequence in different species of *Penstemon*. It has also enabled us the ability to gain greater insights into microsatellite, SNP, putative gene and repetitive element content in the *Penstemon* genome which provide essential tools for further genetic work including plant breeding and phylogenetics.

Key Words: *Penstemon*, genome reduction, pyrosequencing, repetitive elements, gene ontology

ACKNOWLEDGMENTS

I would like to thank Dr. Mikel Stevens for being a great teacher, mentor and friend. Without his guidance, this would not have been possible. I would also like to thank Dr. Brad Geary for financing undergraduate student help and providing lab space and supplies from time to time; Dr. Jeff Maughan for his expertise in marker development and guidance as I progressed through the genome reduction steps of my project; and Dr. Leigh Johnson for his guidance and expertise in plant systematics.

I would also like to acknowledge the assistance of several coworkers who patiently tutored me in the use of the computer programs mentioned in this thesis: David Elzinga, Robert Byers, Scott Yourstone and Prabin Bajgain.

Also I would like to thank all of the undergraduate students who assisted in this project over the years, especially Janalynn Franke and Danika Tumbleson who often provided more support than was asked of them.

Most of all, I would like to thank my wife Keri Dockter, who not only provided moral support but as often as occasion would permit, assisted me in the lab and my daughter Adelle for providing sufficient distraction.

TABLE OF CONTENTS

CHAPTER 1:

A GENOME SNAPSHOT AND MOLECULAR MARKER DEVELOPMENT IN FOUR SPECIES OF *PENSTEMON*

INTRODUCTION

Penstemon Mitchell (Plantaginaceae) is one of the largest, most diverse plant genera in North America with roughly 270 identified species, most of which are native to the United States Intermountain West (Lodewick and Lodewick 1999; Wolfe et al. 2006). *Penstemon* possess a diverse array of flower and leaf structure, color, and texture (Way and James 1998; Nold 1999; Lindgren 2000; Lindgren and Wilde 2003). Their unique diversity, paired with their drought-tolerance and overall hardiness, give *Penstemon* a vast amount of nearly "untapped" potential in the landscaping industry; especially for the more arid western United States where they naturally thrive and where much of the fresh water reserves are utilized for landscaping (Way and James 1998; Nold 1999; Lindgren 2000; Dougher 2003; Zollinger et al. 2006; Zollinger et al. 2007). A way to increase water use efficiency is to utilize landscape plants adapted to arid conditions (Helmle 2005). Additionally, a few taxa are being monitored, in consideration of Endangered Species Act protection (Kramer and Fant 2007). Thus, multiple users will benefit from an improved understanding of its genetics.

Aside from a few phylogenetic studies using genomic size, isozymes, eight ISSRs (intersimple sequence repeat), eight microsatellites or simple-sequence repeats (SSRs), ITS and chloroplast DNA, little molecular or classical genetic work has taken place in the *Penstemon* genus (Wolfe and Elisens 1993; Wolfe and Elisens 1994; Wolfe and Elisens 1995; Wolfe et al. 1998a; Wolfe et al. 1998b; Wolfe et al. 2002; Wolfe et al. 2006; Kramer and Fant 2007; Broderick et al. 2011). Wolfe et al. (2006) found that the phylogenetic relationships were difficult to determine between many species due its apparent recent speciation and dissemination across North America and concluded that additional molecular studies were needed to more clearly understand relationships between the species.

In addition to the use of molecular markers in phylogenetic studies, they are essential to modern plant breeding, genetic diversity studies, genetic mapping and germplasm conservation programs (Tanksley and McCouch 1997; Bernardo 2008; Cramer et al. 2008; Maughan et al. 2011). In order to create two of the most informative molecular markers, microsatellites and SNPs (single nucleotide polymorphisms), a vast amount of sequence must be obtained from the organism of interest in order to recover enough markers to adequately accomplish a number of genomic studies (Maughan et al. 2009; Santana et al. 2009). Thus the development of molecular markers from sequence in organisms with little to no prior sequence data, like *Penstemon*, has historically been time consuming and costly. However, the development of the next generation sequencing instrumentation (e.g., Roche 454-pyrosequencing) has greatly facilitated a high-throughput method of marker discovery. For instance Santana et al. (2009) found that microsatellite marker development increased tenfold using the 454-pyrosequencer over the traditional cloning techniques, at a greatly reduced cost per marker. Maughan et al. (2009) described a method using 454-pyrosequencing to sequence homologous loci known as genome reduction using restriction enzyme site conservation (GR-RSC). This technique leads to a > 90% reduction of the genome and allows for the focused sequencing of homologous regions between two different restriction enzymes recognition sites. The use of genomic reduction techniques such as AFLPs (amplified fragment length polymorphism) and GR-RSC has confidently identified homologous sequences between closely related species (Althoff et al. 2007; Maughan et al. 2011) allowing for the identification and development of single nucleotide polymorphisms (Maughan et al. 2011).

The objectives of this study were two-fold. First, we identified putative homologous sequences from four *Penstemon* species (three of the four were from distantly related subgenera)

in order to identify potential interspecific molecular markers for phylogenetic studies. Second, we used the GR-RSC high-throughput 454 sequence data to gain an improved understanding of the *Penstemon* genome's microsatellites, SNPs, putative genes, and repetitive elements.

MATERIALS AND METHODS

Plant material and DNA extraction

DNA from *P. cyananthus* var. *cyananthus* Hooker, *P. davidsonii* Greene, *P. dissectus* Elliot, and *P. fruticosus* var. *fruticosus* Greene were extracted using the CTAB purification method (Sambrook et al. 1989) with modifications used by Todd and Vodkin (1996). These plants were from the same sources and identified as described by Broderick et al. (2011) (Table 1). A single sample from each species with the highest quality, and DNA concentration, as determined using a ND-1000 spectrophotometer, (NanoDrop Technologies Inc., Montchanin, DE) was selected to produce a total of 500 ng of DNA necessary for genome reduction steps.

Genome reduction, barcode addition and 454 pyrosequencing

Genome reduction was performed according to the methods described by Maughan et al. (2009). Briefly, *Eco*RI and *Bfa*I were selected for the initial restriction digest step, after which a biotin-labeled adapter was ligated to the *Eco*RI restriction site and a non-labeled adapter was ligated to the *Bfa*I restriction site of each species independently. Following adaptor ligation, a non-labeled size exclusion step utilizing Chroma Spin +TE-400 columns (Clontech Laboratories, Inc., Mountain View, CA) and magnetic biotin-streptavidin separation (Dynabeads M-280 Streptavidin, Invitrogen Life Science Corporation, Carlsbad, CA) was performed. Unique multiplex identifiers (MID) barcodes were added independently to each of the four species using

primers complementary to the adapter and cut sites (Table 1). Preliminary amplification was performed according to the following thermocycling profile: 95º C for 1 minute, 22 cycles of 95º C for 15 seconds, 65º C for 30 seconds, and 68º C for 2 minutes (Maughan et al. 2009). PCR products were loaded into a 1.2% agarose Flashgel DNA Cassette (Lonza Corporation; Rockland, ME) to verify smearing and adequate amplification in preparation for pyrosequencing.

After the PCR step the concentrations of each of the four species samples were determined fluorometrically using PicoGreen® dye (Invitrogen, Carlsbad, CA). Samples were then combined with approximately equal molar concentrations of each species except for *P. cyananthus* (genome size = 1C = 893 Mbp) where the molar concentration was doubled to maintain a similar genomic representation compared to the other three species with smaller genome sizes (*P. davidsonii*, 1C = 483 Mbp; *P. dissectus*, 1C = 462 Mbp; and *P. fruticosus* var. *fruticosus*, 1C = 476 Mbp; [Broderick et al. 2011]). DNA fragments between 500-600 bp were selected using the method described by Maughan et al. (2009). Sequencing was performed by the Brigham Young University DNA Sequencing Center (Provo, UT) using a half 454-pyrosequencing plate, Roche-454 GS GLX instrument and Titanium reagents (Brandord, CT). Raw sequences will be added to GenBank.

Sequence assembly

Sequence data from each species was separated by the software CLC Bio Workbench (v. 2.6.1; Katrinebjerg, Aarhus N, Denmark) using the four MID barcodes (Table 1) into four separate FASTA files and assemblies were performed using Roche Newbler de novo assembler (v. 2.0.01). Consensus sequences of the individual sequence assemblies (contigs) were created and used for further analyses. The parameters of the full *Penstemon* assembly (sequence from all

four species pooled into one file) were left as default except that the large/complex genome option was selected and a trimming file was uploaded to trim the barcodes from the sequencing reads. For all subsequent species assemblies default parameters were used except for the following exceptions: the large/complex genome option selected, an expected depth of '10' (0 default), trimming file included, a minimum overlap length of '50' (40 default), and a minimum overlap identity of 95% (90% default).

Repeat element identification

Assembled sequences from all four species were masked using a combination of RepeatModeler (v. 1.0.4; Smit and Hubley [*RepeatModeler Open-1.0*. 2008-2010, http://www.repeatmasker.org]) and RepeatMasker (v. 3.2.9; Smit et al. 2010 [*RepeatMasker Open-3.0*. 1996-2010, (RMLib: 20110419) http://www.repeatmasker.org]) was implemented. RepeatModeler is a de-novo repeat element family identification algorithm that implements RECON (Bao and Eddy, 2002) and RepeatScout (Price et al. 2005). RepeatModeler scanned the contigs from all four *Penstemon* species and produced the predicted repeat element models. Using these models as a reference database, RepeatMasker then scanned the four *Penstemon* species to filter out the repetitive elements. Singletons were omitted from the analysis. We also included two genome reduced sequences from the *Arabidopsis* RIL lines Ler-O and Col-4 (data from Maughan et al. 2010; RIL lines originally described by Lister and Dean 1993) and a non-genome reduced *Arabidopsis* (downloaded as whole chromosomes from TAIR [The *Arabidopsis* Information Resource]) as references.

Microsatellite and SNP detection

The computer programs MISA and SNP_Finder_Plus (custom Perl-script) were used to identify microsatellites and SNPs, respectively (Stajich et al. 2002; Thiel et al. 2003; Maughan et al. 2009). RepeatMasker was used to identify/mask transposable elements (to eliminate possible multi-loci microsatellites for marker creation) and calculate GC content for the fragments. MISA parameters were set as follows: Di-nucleotide motifs had a minimum of eight repeats, tri-nucleotide motifs had a minimum of six repeats, tetra-nucleotide motifs had a minimum of five repeats, and 100 bp was set as the interruption (max difference between two microsatellite alleles). For the comparison of microsatellite frequency and repeat motifs across species, "unmasked" assembly files were used in order to remove bias caused by masking low complexity reads. For SNP_Finder_Plus the following parameters were used: 8X minimum read depth for the SNP, 30% proportion of the reads representing the minor allele and 90% identity (an indication of homozygosity within a single species used in a dual-species assembly) required for each SNP locus. These parameters also compensate for sequencing and assembly errors, which allow us to have greater confidence in calling base pair discrepancies as actual SNPs in the dual-species assemblies and heterozygosity in the individual assemblies. For both individual assemblies and dual species assemblies the only SNPs reported are those abiding by aforementioned parameters.

Microsatellite marker development and validation

The computer program Primer3 v2.0 (Rozen and Skaletsky 2000) was used to develop primers for the microsatellite markers, with the parameters set as follows: An optimal primer size

of 20 with a minimum of 18 and a maximum of 27, product sizes ranging from 100-500 base pairs, a Tm range of 50-60 with an optimum Tm of 55 ^{o}C and a maximum polynucleotide of 3.

The microsatellite markers were validated using the original four species as template DNA. PCR reactions were conducted as follows: for each cocktail of four reactions, 27µl of double distilled water, 2.25µl of 2 mM dNTPs, 4.5µl of cresol red, 4.5µl of 10X PCR buffer (Sigma-Aldrich, St. Louis, MO), 0.9µl of JumpStart^{TM} Taq DNA Polymerase (Sigma-Aldrich, St. Louis, MO) and 0.68µl (each) of 10mM forward and reverse primers. The thermo cycler (Mastercycler® Pro; Eppendorf International; Hamburg, Germany) was set as follows: 94º C for 30 seconds, 45 cycles of 92º C for 20 seconds, (specific annealing temperature)º C for 1 minute 30 seconds, 72º C for 2 minutes, and 72º C for 7 minutes (final extension).

Following PCR reactions, DNA was loaded into 3% Metaphor® Agarose (Lonza Corporation; Rockland, ME) gels and run using a gel electrophoresis box at 100V for 2 hours. Optimal annealing temperatures for each marker were selected based on clarity of bands produced over varying annealing temperatures. Quality scores were assigned to the markers with a scale ranging from 1-4 as follows: 1 = clear bands with no apparent stutters, 2 = some clear bands and no more than 2 faint bands with no apparent stutters, 3 = many faint bands but no more than 2 per species, few stutters, 4 = many faint bands more than 2 per species (multiple loci likely) and few to many stutters.

Data from the validation gels were then utilized to create a binary matrix scored as: 1 = allele present; 0 = allele absent. Following matrix creation, a simple neighbor-joining analysis was performed using PAUP (v. 4.0b10 [Swofford, DL (2003) PAUP: Phylogenetic analysis using parsimony and other methods]) (Sinauer Associates; Sunderland, MA).

Gene ontology

From assembled sequences of all four species, BLASTX (Altschul et al. 1990), which, for our study, compared sequence to the GenBank refseq-protein database (Benson et al. 2011), was used with a e-value threshold of $<1.0e^{-15}$. Blast2GO (v2.4.2) (Conesa et al., 2005) was used to map the blast hits and annotated them to putative cellular components, biological processes, and molecular functions found in the blast database. For species comparisons the GO level of 3 was used for cellular components and level 2 was used for both biological processes and molecular functions.

For further comparisons to the related genera *Antirrhinum* and *Mimulus*, assembled sequences of all four species were compared to all available *Antirrhinum* and *Mimulus* genes on GenBank (downloaded June 23rd, 2011). Comparisons were done using BLASTN (Altschul et al. 1990) with an e-value threshold of $<1.0e^{-6}$.

RESULTS

Genome reduction, pyrosequencing and species assemblies

Based on an 35% GC content for dicots (Kawabe and Miyashita 2003), a theoretical frequency of *Bfa*I and *Eco*RI recognition sites, and the genome sizes of our four species, we estimated a 104 fold reduction of sequence for each species, allowing for an estimated 11X coverage of remaining sequence per species. Our pyrosequencing run produced 287 Mbp from 733,413 reads, with an average of 392 bp per read (Table 2). In total, 93,828,389, 46,381,818, 48,808,929, and 53,287,470 bp were sequenced from *P. cyananthus*, *P. dissectus*, *P. davidsonii* and *P. fruticosus*, respectively, which closely represents the 2:1:1:1 ratio of the combined DNA

samples prior to sequencing. The GC content of these samples was 36.4%, 34.5%, 35.3%, and 35.15% from each of the above listed respective species (Table 2).

The full assembly of all four *Penstemon*, using the Newbler de novo assembler, produced a total of 44,966 contigs, representing 16.4 Mbp, or 5.7% of our total sequence. In the individual species assemblies of *P. cyananthus*, *P. dissectus*, *P. davidsonii*, and *P. fruticosus*, a total of 9,714, 5,364, 4,882, and 4,777 contigs were created representing 4.6, 2.6, 2.4 and 2.3 Mbp of assembled bases respectively, averaging 0.5% of the total genomes being sequenced with an average coverage of 8.5X for a given base (Table 2).

Microsatellite analysis

From the "masked" assembly containing contigs from all four species, MISA identified 133 microsatellite motifs. Of the 133 microsatellites, Primer 3 was able to design 77 primer combinations meeting the desired parameters for marker creation (Table 3). There were 97, 113, 49 and 58 identified microsatellites in the "unmasked" individual assemblies of *P. cyananthus*, *P. dissectus*, *P. davidsonii*, and *P. fruticosus* respectively (Table 4).

Of the 77 microsatellite markers developed, 50 possessed a quality score of 1 or 2 (clear bands with no or few faint bands and no stutters). Of the 50 high-quality markers, there was an overall success rate (samples with at least one visible amplicon) of 76% and 41 markers (82%) appeared polymorphic between the four species (Table 3). An average of 2.5 alleles per marker was found across the four species and 1.0 allele per marker per species average. Among all 50 markers, the average heterozygosity value was 0.51 (Table 3). Putative alleles were scored as either '1' (present) or '0' (absent) and analyzed in PAUP to create a simple unrooted tree (Fig. 1).

For the di-nucleotide repeat motifs, AT/TA represented 51.7%, AG/TC represented 25.0%, and AC/TG represented 23.3% of the repeats on average among the four species. Similarly, motifs rich in adenine and thymine were the largest categories in the tri- and tetra-nucleotide motifs. AAT/TTA represented 40.1% of the tri-nucleotide repeat motifs and AAAT/TTTA represented 20.0% of the tetra-nucleotide motifs (Fig. 2).

SNP analysis

Using our SNP discovery parameters of a minimum of 8X coverage, and 30% representation of the minor allele we identified an average of one SNP per 2,890 bp across the four species ranging from *P. cyananthus* which had one SNP per 1,855 bp to *P. fruticosus* (1 SNP per 3,777 bp). The three species with similar genome sizes all had similar frequencies of SNPs (Table 5). The SNP transition mutations (A↔G or C↔T) were more common compared to transversions (A↔T, A↔C, G↔C, G↔T) by an average factor of (1.5; Table 5).

In the dual species assembly, utilizing the same conservative parameters and a 90% SNP identity, we found an average of 1 SNP in 97 bp between homologous sequence assemblies of any two of the four species. The density of SNPs between homologous sequences of *P. dissectus* and *P. davidsonii* was the highest at one per 64 bp with the lowest being between *P. cyananthus* and *P. davidsonii* at one SNP per 124 bp. In the dual species assemblies, the average margin of transition to transversions mutation was lower (1.2; Table 5).

Gene ontology

BLASTX identified an average of 21.5% of the contigs across the four species as putative genes and an average of 13.9% were annotated by Blast2Go (Table 4). In consideration of the

cellular components, most sequences fell into the categories "cell part" (43-45%) or "membrane bound organelle" (32-36%) with "organelle", "vesicle", and "non-membrane-bound organelle" putative genes also being present (Fig. 3). A majority of the biological processes found in the putative genes were divided into "metabolic processes" (29-34%) and "cellular processes" (28-33%). Other categories include: "biological regulation", "response to stimulus", and "localization" (Fig. 3). Most of the putative genes associated with molecular function fell into two categories; "binding" (46-51%) and "catalytic activity" (38-42%). Other categories included: "transporter activity", "transcription regulator activity", and "molecular transducer activity" (Fig. 3). Overall there was little difference between the four species across all putative gene categories (std. dev ≤ 2.8%).

By comparing our 454 sequence data to known genes from the related genera *Antirrhinum* and *Mimulus*, we were able to discover ten putative *Penstemon* genes from *Antirrhinum* and 14 putative *Penstemon* genes from *Mimulus* with an e-value below $1.0e^{-6}$ (Table 6). Three genes (NADH dehydrogenase from *M. aurantiacus*, ribosomal protein L10 from *M. guttatus*, and ribosomal protein subunit 2 from *M. aurantiacus*, *M. szechuanensis*, and *M. tenellus var. tenellus*) were perfect hits (e-value = 0.0; Table 6).

Repetitive elements

We identified 28.5%, 16.8%, 17.4% and 16.1% of the respective sequence from *P. cyananthus*, *P. dissectus*, *P. davidsonii*, and *P. fruticosus* as repeat elements using RepeatModeler and RepeatMasker. Of these elements, 3.0-7.8% were identified as LTR (long terminal repeat) retroelements, 0.3-1.0% were DNA transposons and the remainder were unclassified (Table 4). Since RepeatModeler utilizes RECON and RepeatScout to create a de

novo model to be used in RepeatMasker in place of the *Arabidopsis* model, details about the subcategories of LTRs and DNA transposons which are included in the model could not be addressed. Between the two genome reduced *Arabidopsis* plants (Ler-0 and Col-4), an average of 6.2% of the sequence was filtered as repeat elements, of which an average of 4.4% were identified as LTR retroelements and 0.4% were identified as DNA transposons using RepeatModeler and then RepeatMasker. From the non-genome reduced *Arabidopsis* (TAIR10) 7.4% of the bases were identified as repeat elements of which 3.0% were LTR retroelements and 0.2% were DNA transposons (Table 4 and Fig. 4 and 5).

DISCUSSION

In this study, we used the GR-RSC technique on four *Penstemon* species to gain an understanding of the breadth of genomic differences within the genus. These four species were selected according to current taxonomic classifications utilized by Broderick et al. (2011) and Wolfe et al. (2006). Two of the species were closely related and from the subgenus *Dasanthera* (*P. davidsonii* and *P. fruticosus*), the third was from the *Habroanthus* subgenus (*P. cyananthus*), and the fourth was a monotypic species *P. dissectus* (subgenus *Dissecti*) that is phenotypically distinct from all other species in the genus. Under the assumption that more closely related species (within the same subgenus) would have similar genomic sequences, this experimental design allows us to make broader inter- and intra-subgenera comparisons in the *Penstemon* genus.

Genome reduction, pyrosequencing, and species assemblies

Given that a full 454 pyrosequencing plate using Titanium reagents is capable of producing 1.3 million reads averaging ~400 bp each (Maughan et al. 2009), we expected a half plate to produce approximately 250 Mbp from 650,000 reads. The reaction produced 287 Mbp from 733,413 reads which was higher than expected with an average read length of 392 which was nearly identical to the expected 400 bp. The average GC content across the four species was 35.3% which is in harmony with both our assumption of GC content in deciding which restriction enzymes to use in the GR-RSC technique as well as the theoretical GC content for dicots (Kawabe and Miyashita 2003).

From our de novo assemblies we identified nearly twice as many contigs (9,714) in *P. cyananthus* as the 4,777 found in *P. fruticosus* (Table 2) which is expected since we added approximately twice as much DNA from *P. cyananthus* as we did from the other three species. Those 9,714 contigs represented 0.6% of the total *P. cyananthus* genome which was similar to the 0.5% average of the other three genomes which demonstrate that we were able to produce nearly equal genome representation from each of the four species with differences in genome size by mixing molar concentrations relative to the genome size in the sequencing reaction. The portions of sequence that formed contigs in the individual species assemblies averaged 48.2%, suggesting that about half of the sequence lacked sufficient coverage or were too short to be assembled into contigs (Table 2).

Microsatellite analysis

Using MISA we identified microsatellites in the "unmasked" contigs of all four species. The "masked" assembly was not used because masking the sequence file removes some of the

adenine and thymine rich microsatellites and categorizes them as "low complexity reads" which may bias the data for general species comparisons. There were more microsatellites in *P. dissectus* (113) than *P. cyananthus* (97) which has a 1.9 times larger genome (Broderick et al. 2011) and a higher representation of sequence than *P. dissectus* (Table 4). Finding this inverse relationship between genome size and the number of microsatellites agrees with what has been identified in other plant genomes (Morgante et al. 2002).

We also examined trends in repeat motifs between the four species of *Penstemon* (Fig. 2). We found that, on average, adenine and thymine rich repeat motifs were the most common repeat type in the di-, tri-, and tetra-nucleotide repeat motifs. In general, AT motifs are the most common in noncoding regions of most plant genomes (Morgante et al. 2002). This is consistent with our findings that adenine and thymine rich repeat motifs were the most common category among the di-, tri-, and tetra- nucleotide repeat motifs. There seemed to be much more variation in the repeat motifs in the tetra-nucleotide repeats across the four species. Even the closely related *P. fruticosus* and *P. davidsonii* had completely distinct tetra-nucleotide repeat motifs (Fig. 2). This is likely due in part to the rarity of the motifs and high number of possible nucleotide combinations. Several studies have found that the hypothetical origins of some microsatellites are retrotransposition events (Nadir et al. 1996; Temnykh et al. 2001; Parida et al. 2009) and as such may be useful, once studied, as part of a unique "fingerprint" for a given species.

We utilized the assembly containing contigs from all four species with "masked" multi-loci repeats (such as transposons) in order to design microsatellite marker primers (Table 3). Those markers have potential for use in intra and interspecific molecular marker studies. Since the microsatellite markers were validated across species boundaries, it was expected that

polymorphism would potentially be high. The average heterozygosity value found in this study (0.51) is lower than the typical 0.60 heterozygosity of microsatellite markers in general (Powell et al. 1996). However, the small number of samples used to test the microsatellite markers may have biased the allele frequencies resulting in the putatively reduced heterozygosity value identified. The 76% success amplification rate was unexpectedly high for inter-specific microsatellites. It is possible that many *Penstemon* species have highly conserved microsatellite regions, although the microsatellites themselves vary greatly in length. Indeed Morgante et al. (2002) suggested that microsatellites in non-coding regions of plant genomes predate the most recent genome expansions of many plants. However, some of the bands differed in length as much as 570 bp (Table 3) which is probably not a microsatellite motif variation. It is more probable that these large differences in bp length between species are reflective of insertions/deletions events independent of SSR length variation. At present, more work needs to be done to elucidate the cause of such dramatic size differences in microsatellite alleles.

Our simple neighbor-joining analysis of the microsatellites revealed what was expected from the known phylogeny of the *Penstemon* genus. The two representatives from the Dasanthera subgenus (*P. davidsonii* and P. *fruticosus*) clustered together and the other two species (*P. cyananthus* and *P. dissectus* from the *Habroanthus* and *Dissecti* subgenera respectively) remained on separate branches (Fig. 1). These data are in agreement with the research done by Wolfe et al. (2006) and demonstrate the possibility that these microsatellite markers have potential for use in inter-specific phylogenetic studies.

SNP analysis

As reported in other plant species (Zhang and Zhao 2004; Morton et al. 2006), we found that the frequencies of transitions mutations were more common than transversions in *Penstemon* by an average factor of 1.5 (Table 5) which is close to the factor of 1.4 found in *Arabidopsis* (Maughan et al. 2010). Since it appears that many plant species show a similar ratio of transitions to transversions, it is possible that this trend is predictable across plant genomes. Zhao et al. (2006) found similar ratios of transitions to transversions in the genomes of *Arabidopsis* and rice (*Oryza sativa* L.) but they also found that the bases surrounding the SNP had an effect on the ratio of transitions to transversions overall. When a SNP occurs within a few base pairs of an adenine or thymine, the likelihood of transversions increases and transitions decrease (Zhao et al. 2006). Since the GC content in *Penstemon* was similar between all four species and to other dicots, there was a similar probability of a SNP being adjacent to either an adenine or a thymine; thus, our ratio of transition SNP to transversion SNP was expected.

As could be expected from molecular based phylogenetic studies (Wolfe et al. 2006; Broderick et al. 2011), one of the lowest interspecific SNP densities was found to be between homologous sequences of *P. davidsonii* and *P. fruticosus* (one SNP per 119 bp) which are both members of the *Dasanthera* subgenus. Comparisons of *P. cyananthus* and either *P. davidsonii* or *P. fruticosus* homologous sequences yielded equally low results (1/124 and 1/119, respectively). However any homologous sequence comparison involving *P. dissectus* consistently had the highest density of SNPs. This suggests that *P. dissectus* is the most evolutionary distant of the four species used in this study.

Due to the functionality of the program SNP_Finder_Plus, two sequence files are required to satisfy the "SNP identity" parameter. For the individual assembly files, this

requirement was not met which introduces a potential weakness to our study. However, due to our coverage requirement (8X) and the minimum minor allele frequency (30%), the same putative SNP must be represented in at least three of eight reads. This provides some protection from mislabeling a sequencing/assembly error as a SNP or heterozygosity. Furthermore the average SNP coverage was actually 14.4 across the four species (Table 5). This suggests that, on average, five identical putative SNPs represented the minor allele, further supporting the validity of our claim.

Gene ontology

The highest represented categories from cellular components, biological processes, and molecular function were "cell part", "metabolic processes", and "binding" respectively. Hao et al. (2011) reported finding the same categories as the highest represented from the *Taxus* genome despite the fact that a completely different sequence subset via fosmid library was used. In addition many of the other putative gene categories found in their study was also found in ours. Although it is probable that this similarity is due to highly conserved, abundantly represented genes in plant genomes, it may, in part, be due to the functionality of the Blast2GO program (categorization of putative genes) or bias imposed by over-representation of certain genes in the GenBank refseq-protein database.

Since many of those comparisons were with more distantly related plant species, we wanted to explore potential sequence similarities with genes available on GenBank from the closely related genera *Mimulus* and *Antirrhinum* which, according to some taxonomic keys, belong to the same plant family as *Penstemon* (Welsh et al. 2008). These genera have also had more molecular work done. As expected many of the "hits" found were highly conserved plant

genes involved in processes ranging from metabolism to protein synthesis such as: NADH dehydrogenase, MatK, ATPases, and ribosomal protein subunits. Some of these genes were perfect matches (Table 6). These data demonstrate that our 454 sequence adds to the collective molecular work of both related and distant plant species and provides a data source for validation work of several putative *Penstemon* genes.

Repetitive elements

Broderick et al. (2011) hypothesized that the diversity in genome sizes between *Penstemon* species of the same ploidy may be explained by retrotransposons. Lynch (2007) detailed a relationship between genome size numbers of repeat elements suggesting that after threshold of 100 Mbp, all species possess all three main classes of repeat elements: LTRs, non-LTRs, and DNA transposons with a linear relationship between number of elements and genomes size (Kidwell 2002; Lynch and Conery 2003; Lynch 2007). The four *Penstemon* species used in this study provide insufficient evidence to establish whether a linear relationship between genome size and repeat elements in *Penstemon*. However, the three smaller, similar sized *Penstemon* genomes possess an equally similar amount of repetitive elements whereas *P. cyananthus* (the largest genome) showed a much higher number of repeat elements (nearly double) compared to the other three species (Fig. 5).

Not only do repetitive elements largely influence the size of a genome, but they also are more likely to evolve more rapidly than do low-copy sequence (Kidwell 2002; Raskina et al. 2008). Thus, repetitive elements of a species take on unique "fingerprints" which become valuable in phylogenetic relationship studies (Kubis et al. 1998; Kolano et al. 2011). It is clear in our study that *P. dissectus* has a lower percentage of LTR repeat elements than the other species.

Thus, our limited genomic data set of the four *Penstemon* species suggest agreement with both hypotheses that repetitive elements are probably a major component of the genome size variation identified by Broderick et al. (2011) and these elements are variable between the species; thus, suggesting the possibility of identifying species specific repetitive elements. However, without further studies, we were unable to elucidate the specific repetitive elements associated with our four *Penstemon* species.

CONCLUSIONS

We have demonstrated that GR-RSC is a highly effective way of mass sequencing putative homologous alleles in the *Penstemon* genus. Using this technique we were able to sequence approximately 0.5% of the genomes of *P. cyananthus*, *P. dissectus*, *P. davidsonii*, and *P. fruticosus* involving three subgenera of the genus. In our study 50 high-quality interspecific microsatellite markers and several thousand inter- and intraspecific SNPs were identified. Information on ~3,500 putative *Penstemon* genes was obtained as well as linking 24 of them to the closely related genera *Mimulus* and *Antirrhinum*. Furthermore, valuable information on repetitive element content found within the genomes of our four species may lead to further studies on the influence of repetitive elements in the speciation of this large plant genus. Additional studies using the same parameters of this GR-RSC study on other species within the genus would allow broader comparisons of putative genes and repeat elements, SNPs and microsatellites for phylogenetics and related studies. Finally, using this technique on carefully selected parent plants would allow for the creation of populations for genetic mapping studies of *Penstemon*. Genetic markers and mapping studies would dramatically facilitate phylogenetic, population genetics and greatly enhance the ability to do breeding studies within this genus.

REFERENCES CITED

Althoff DM, Gitzendanner MA and Segraves KA (2007) The utility of amplified fragment length polymorphisms in phylogenetics: A comparison of homology within and between genomes. Syst Biol 56:477-484

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

Bao Z and Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12:1269-1276

Bernardo R (2008) Molecular markers and selection for complex traits in plants: Learning from the last 20 years. Crop Sci 48:1649-1664

Broderick SR, Stevens MR, Geary B, Love SL, Jellen EN, Dockter RB, Daley SL and Lindgren DT (2011) A survey of *Penstemon's* genome size. Genome 54:160-173

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M and Robles M (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676

Cramer ERA, Stenzler L, Talaba AL, Makarewich CA, Vehrencamp SL and Lovette IJ (2008) Isolation and characterization of SNP variation at 90 anonymous loci in the banded wren (*Thryothorus pleurostictus*). Conserv Genet 9:1657-1660

Dougher T (2003) Commercializing production of native Intermountain species. HortScience 38:752-753

Hao D, Yang L and Xiao P (2011) The first insight into the *Taxus* genome via fosmid library construction and end sequencing. Mol Genet Genomics 285:197-205

Helmle SF (2005) Water Conservation Planning: Developing a Strategic Plan for Socially Acceptable Demand Control Programs. An Applied Research Project, Texas State University-San Marcos

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, and Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogent Genome Res 110:462-467

Kawabe A and Miyashita NT (2003) Patterns of codon usage bias in three dicot and four monocot plant species. Genes Genet Syst 78:343-352

Kidwell MG (2002) Transposable elements and the evolution of genome size in eukaryotes. Genetica 115:49-63

Kolano B, Gardunia BW, Michalska M, Bonifacio A, Fairbanks D, Maughan PJ, Coleman CE, Stevens MR, Jellen EN and Maluszynska J (2011) Chromosomal localization of two novel repetitive sequences isolated from the *Chenopodium quinoa* Willd. genome. Genome (Accepted)

Kramer AT and Fant JB (2007) Isolation and characterization of microsatellite loci in *Penstemon rostriflorus* (Plantaginaceae) and cross-species amplification. Mol Ecol Notes 7:998-1001

Kubis S, Schmidt T and Heslop-Harrison JS (1998) Repetitive DNA elements as a major component of plant genomes. Ann Bot 82(Supl A):45-55

Lindgren D and Wilde E (2003) Growing *Penstemons*: Species,cultivars and hybrids. Haverford

Lindgren DT (2000) Breeding *Penstemon*, In: Callaway DJ and Callaway MB (eds) Breeding ornamental plants. Timber Press, Portland, pp. 196-212

Lodewick K and Lodewick R (1999) Key to the genus *Penstemon* and its related genera in the tribe Cheloneae (Scrophulariaceae). K. Lodewick, Eugene

Lister C and Dean C (1993) Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. Plant J 4:745-750

Lynch M (2007) The Origins of Genome Architecture. Sinauer Associates, Inc. , Sunderland

Lynch M and Conery JS (2003) The origins of genome complexity. Science 302:1401-1404

Maughan PJ, Smith SM, Fairbanks DJ and Jellen EN (2011) Development, characterization, and linkage mapping of single nucleotide polymorphisms in the grain amaranths (*Amaranthus* sp.). Plant Gen 4:1-10

Maughan PJ, Yourstone SM, Byers RL, Smith SM and Udall JA (2010) Single-nucleotide polymorphism genotyping in mapping populations via genomic reduction and next-generation sequencing: Proof-of-concept. Plant Gen 3:1-13

Maughan PJ, Yourstone SM, Jellen EN and Udall JA (2009) SNP discovery via genomic reduction, barcoding and 454-pyrosequencing in amaranth. Plant Gen 2:260-270

Morgante M, Hanafey M and Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Biotechnol 30:194-200

Morton BR, Bi IV, McMullen MD and Gaut BS (2006) Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition. Genetics 172:569-577

Nadir E, Margalit H, Gallily T and Ben-Sasson SA (1996) Microsatellite spreading in the human genome: Evolutionary mechanisms and structural implications. Proc Natl Acad Sci USA 93:6470-6475

Nold R (1999) *Penstemons.* Timber Press, Portland

Parida SK, Kalia SK, Kaul S, Dalal V, Hemaprabha G, Selvi A, Pandit A, Singh A, Gaikwad K, Sharma TR, Srivastava PS, Singh NK and Mohapatra T (2009) Informative genomic microsatellite markers for efficient genotyping applications in sugarcane. Theor Appl Genet 118:327-338

Pettersson A, Winer ES, Weksler-Zangen S, Lernmark A and Jacob HJ (1995) Predictability of heterozygosity scores and polymorphism information content values for rat genetic markers. Mamm Genome 6:512-520

Powell W, Morgante M, Andre C, Hanafey M, Vogel J, Tingey S and Rafalski A (1996) The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Mol Breed 2:225-238

Price AL, Jones NC and Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21(Supl1):i351-i358

Raskina O, Barber JC, Nevo E and Belyayev A (2008) Repetitive DNA and chromosomal rearrangements: Speciation-related events in plant genomes. Cytogenet Genome Res 120:351-357

Rozen S and Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S and Misener S (eds) Bioinformatics methods and protocols: methods in molecular biology. Humana Press, Totowa, pp. 365-386

Sambrook J, Fritcsh EF and Maniatis T (1989) Molecular Cloning: A Laboratory Manual (Cold Spring Harbor Lab.). Cold Spring Harbor, New York

Santana QC, Coetzee MPA, Steenkamp ET, Mlonyeni OX, Hammond GNA, Wingfield MJ and Wingfield BD (2009) Microsatellite discovery by deep sequencing of enriched genomic libraries. BioTechniques 46:217-223

Schmuths H, Meister A, Horres R and Bachmann K (2004) Genome size variation among accessions of *Arabidopsis thaliana*. Ann Bot 93:317-321

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD and Birney E (2002) The bioperl toolkit: Perl modules for the life sciences. Genome Res 12:1611-1618

Tanksley SD and McCouch SR (1997) Seed banks and molecular maps: Unlocking genetic potential from the wild. Science 277:1063-1066

Thiel T, Michalek W, Varshney RK and Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet 106:411-422

Todd JJ and Vodkin LO (1996) Duplications that suppress and deletions that restore expression from a chalcone synthase multigene family. Plant Cell 8:687-699

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M and Zabeau M (1995) AFLP: A new technique for DNA fingerprinting. Nucl Acids Res 23:4407-4414

Way D and James P (1998) The Gardener's Guide to Growing *Penstemon.* Timber Press, Portland

Welsh SL, Atwood ND, Goodrich S, and Higgins LC (2008) *A Utah Flora*. Brigham Young University, Provo

Wolfe AD, Datwyler SL and Randle CP (2002) A phylogenetic and biogeographic analysis of the Cheloneae (Scrophulariaceae) based on ITS and *matK* sequence data. Syst Bot 27:138-148

Wolfe AD and Elisens WJ (1993) Diploid hybrid speciation in *Penstemon* (Scrophulariaceae) revisited. Amer J Bot 80:1082-1094

Wolfe AD and Elisens WJ (1994) Nuclear ribosomal DNA restriction site variation in *Penstemon* section *Peltanthera* (Scrophulariaceae) - an evaluation of diploid hybrid speciation and evidence for introgression. Amer J Bot 81:1627-1635

Wolfe AD and Elisens WJ (1995) Evidence of chloroplast capture and pollen-mediated gene flow in *Penstemon* sect. *Peltanthera* (Scrophulariaceae). Syst Bot 20:395-412

Wolfe AD, Randle CP, Datwyler SL, Morawetz JJ, Arguedas N and Diaz J (2006) Phylogeny, taxonomic affinities, and biogeography of *Penstemon* (Plantaginaceae) based on ITS and cpDNA sequence data. Amer J Bot 93:1699-1713

Wolfe AD, Xiang QY and Kephart SR (1998a) Assessing hybridization in natural populations of *Penstemon* (Scrophulariaceae) using hypervariable intersimple sequence repeat (ISSR) bands. Mol Ecol 7:1107-1125

Wolfe AD, Xiang QY and Kephart SR (1998b) Diploid hybrid speciation in *Penstemon* (Scrophulariaceae). Proc Natl Acad Sci USA 95:5112-5115

Zhang FK and Zhao ZM (2004) The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs. Genomics 84:785-795

Zhao H, Li QZ, Li J, Zeng CQ, Hu SN and Yu J (2006) The study of neighboring nucleotide composition and transition/transversion bias. Sci China Ser C 49:395-402

Zollinger N, Kjelgren R, Cerny-Koenig T, Kopp K and Koenig R (2006) Drought responses of six ornamental herbaceous perennials. Sci Hort 109:267-274

Zollinger N, Koenig R, Cerny-Koenig T and Kjelgren R (2007) Relative salinity tolerance of intermountain western United States native herbaceous perennials. HortScience 42:529-534

CHAPTER 2:

LITERATURE REVIEW

INTRODUCTION

We live during a time of growing concern for the availability of natural resources for future generations. Much of this concern is centered around increased levels of atmospheric carbon dioxide and its impacts on the environment; however the vast plant kingdom possesses the ability to re-sequester CO_2 into biomaterial provided that below ground resources (water, mineral nutrients, etc.) are abundant (Hungate et al. 1997). Water is often the most limiting resource for plant productivity (Boyer, 1982). Furthermore, a compelling study by Zhao and Running (2010) shows that terrestrial net primary production (atmospheric carbon fixation by plants) has decreased by 0.55 pentagrams of carbon in the last 9 years. Their data suggest that this decrease is primarily due to drought conditions in the southern hemisphere. The preservation of clean fresh water sources for plants should be one of the most important considerations for the continued health and sustainability of our population.

The United States Geological Survey agency (USGS) is responsible for gathering water use information and providing estimates of future use (USGS 2009) such as, where and how much water is withdrawn yearly and how that water is used. In 2005, the United States withdrew about 410 billion gallons of water per day; of that 31% (128 billion gallons) was used for irrigation, 44.2 billion gallons was withdrawn for the public supply, of that 58% was for domestic use. Domestic use includes drinking water, sanitation as well as outdoor residential use (USGS 2009). Estimates have shown that the typical suburban lawn uses about 10,000 gallons of water annually above normal rainfall (EPA 2002). It is apparent that large quantities of water every year are used for agriculture and residential landscaping practices. Despite the fact that the amount of total water withdrawn yearly has stayed relatively constant since 1980 (Roy et al. 2003), domestic water use has generally increased since 1950 (USGS 2009). Moreover, some

feel that estimates of long-term sustainability do not always accurately reflect changes in demand and scarcities on the local level (Roy et al. 2003). Having enough water to meet our needs may become a logistical problem for some areas of the country including the dryer climates of the western United States.

A national or global water supply problem cannot be overcome by any one solution. Rather it must be an integrated solution involving the promotion of more sustainable water use practices on a local level. This will help individual homeowners to decrease their annual water use. Water conscious individuals often purchase more water efficient shower heads or toilets (EPA 2002) but more could be done to promote water efficient landscaping practices. Xeriscape landscaping is one such approach as it employs the use of native and drought-tolerant plants in order to lower the necessary amount of water used in landscape maintenance (EPA 2002). Many of the dry climates of the United States are located in the Intermountain West. Since many of the native plants of the Intermountain West are adapted to the dry environment, it is an excellent resource for drought tolerant species. One such genus of drought-tolerant plants native to the Intermountain West, which is both hardy in dry environments and genetically diverse, is *Penstemon* (Broderick 2010; Lindgren 2000; Nold 1999; Way and James 1998).

PENSTEMON

Penstemon Mitchell (Plantaginaceae) is one of the largest, most diverse plant genera in North America with roughly 270 identified species, most of which are native to the United States Intermountain West (Wolfe 2006; Nold 1999). One particular state, Utah, is home to more *Penstemon* species than any other area with approximately 70 species found within its borders (Nold 1999; Way and James 1998). Generally, *Penstemon*s are characterized by: 1) a tube-

shaped (often bilabiate) shaped flower that separates into five petals (two petals forming an upper lip and a three petals forming a lower lip), 2) four fertile stamens, 3) an infertile, often pubescent, staminode, 4) opposite leaf pairs forming right angles with the leaf pairs both above and below (Way and James 1998), 5) stems which can either be woody, suffrutescent, or herbaceous, arising from a woody caudex (Wolfe 2006). In addition, they possess a diverse array of flower and leaf structure, color, and texture (Nold 1999; Way and James 1998). For most of the species of this genus, an arid sloping soil is preferred with little organic matter, which allows for plenty of drainage (Nold 1999). Their unique diversity, paired with their drought-tolerance and overall hardiness (Dougher 2003; Nold 1999; Zollinger et al. 2006), give *Penstemon* a vast amount a nearly "untapped" potential in the landscaping industry—especially in the more arid western United States where they naturally thrive (Lindgren 2000; Way and James 1998).

Penstemon, however, are not ready for a full-blown overhaul of arid urban and suburban landscapes. Despite the benefits of both drought-tolerance and diversity of form within the xeric Intermountain *Penstemon* they lack some of the features that popular ornamentals possess: such as a diverse set of phenotypic characteristics such as diverse showy colors with compact habits, as well as lacking long blooming time habits. These and other desirable characteristics have often been achieved in broadly accepted commercial cultivars. Generally, these successful cultivars have been created through years of plant breeding using inter-specific hybridization and other specialized breeding techniques. While the horticulture industry has been developing cultivars and unique hybrids of other species for centuries, including moisture loving *Penstemon*, the arid species of this genus have been relatively untouched (Lindgren 2000; Way and James 1998). Additionally, the genus is so large, getting a clear picture of the taxonomic classification and phylogenetic relationships has been problematic further complicating the efforts to breed

"xeric friendly" cultivars (Broderick 2010). However, the benefits of using *Penstemon* outweigh the costs that may be required to make them more appealing to the "mainstream" xeric landscape industry. *Penstemon* breeding (to both select for desirable traits and to create new hybrids) needs to be done to ensure the broad success of horticulturally acceptable drought tolerant cultivars. In order to facilitate the breeding of the xeric species of *Penstemon* there is a great need to increase our understanding of the genetics of genus; especially, to improve our understanding of its intrageneric phylogenetic relationships.

Through the creation of high-informative polymorphic markers that act inter-specifically in the *Penstemon* genus, phylogenetic inconsistencies and taxonomic perturbances can be alleviated leading to greater strides in the breeding and domestication of this highly diverse and beautiful plant genus. Furthermore, the use of 454 pyrosequencing and genome reduction can aid in marker development by reducing the cost per data point; providing the means by which markers can be created for what some consider are less economically important ornamental plants.

TAXONOMY AND PHYLOGENY OF *PENSTEMON*

Various attempts and methodologies have been utilized to elucidate the complex taxonomic classifications of the genus *Penstemon* (for a review see Wolfe et al. 2006). One of the most widely used and accepted methodologies in *Penstemon* classification is based primarily on the pattern of dehiscence of the anther-cells; secondarily, by using the morphology of the inflorescence, leaves, and stems, including the appearance of pubescence on those structures (Cronquist et al. 1984; Lodewick and Lodewick 1999; Welsh et al. 2008). Even though the taxonomy of the *Penstemon* genus has had multiple name revisions, it is generally accepted to be

separated into six subgenera (two of which are monotypic), 13 sections, and 22 subsections (Broderick 2010; Lodewick and Lodewick 1999; Nold 1999; Wolfe et al. 2006). It is clear that the morphological diversity, which in many lights is beneficial, has made taxonomical classification based on phenotypic characters much more difficult, and made the evolutionary relationships between species convoluted (Broderick 2010; Wolfe et al. 2006).

A few studies have employed the use of genetic information in an attempt to produce more realistic and simple taxonomic classifications and phylogenetic trees (Datwyler and Wolfe 2004; Wilson et al. 2007; Wolfe et al. 2002; Wolfe and Elisens 1993; Wolfe and Elisens 1994; Wolfe and Elisens 1995; Wolfe et al. 2006; Wolfe et al. 1998a; Wolfe et al. 1998b). Wolfe et al. (2006) reported on the most comprehensive phylogenetic study to date on *Penstemon* where they compared nucleotide information from both ITS (internal transcribed spacer) and cpDNA (chloroplast DNA) regions in 163 species. However, they found that these two genes were insufficient to clearly resolve the taxonomic and phylogenetic complications. More recently, Broderick (2010) employed the use of flow cytometry to measure the nuclear DNA content of over 100 *Penstemon* species. His hypothesis was that the evolutionary relationships of species of *Penstemon* are mirrored by their genome size, several putative rearrangements were suggested of which some confirmed the findings of Wolfe et al. (2006). There is no doubt that progress is being made toward an improved understanding of the phylogenetic relationships of the species within *Penstemon*; however, more needs be done to further unravel the genetic relationships within the genus.

MARKER DEVELOPMENT IN *PENSTEMON*

Molecular markers have been developed and utilized in many genetic studies in order to more efficiently evaluate genetic variability, select for desirable traits in cultivars and establish phylogenetic relationships in both plants and animals. Rapid progress in the development of these markers has greatly facilitated the creation of improved cultivars (Eathington et al. 2007). Although some markers have been developed for *Penstemon*, compared to more "mainstream" agronomic and ornamental plant species, the number of markers available are scant at best. One of the first studies utilizing molecular markers in *Penstemon* was conducted by Wolfe and Elisens in 1994 using rDNA restriction-site variation markers. They tested for the possibility of diploid hybrid speciation and introgression in natural populations of *P. centranthifolius*, *P. grinnellii*, *P. spectabilis*, and *P. cleavlandii*. The same research was later revisited in 1998 utilizing eight inter-simple sequence repeat (ISSR) markers (Wolfe et al. 1998). Also, Datwyler and Wolfe (2004) utilized five ISSR markers to amplify 131 loci from nine species of the subgenus *Dasanthera* in conjunction with their ITS and chloroplast matK sequence to study both the phylogeny and biogeography of those species. From their data they were able to produce a phylogenetic tree of the subgenus.

The first microsatellite or simple-sequence repeat (SSR) markers for *Penstemon* were developed by Kramer and Fant in 2007. For their study, genomic DNA was extracted from *P. rostriflorus* and a microsatellite library was created that was enriched for TG, GA, AAG, AAC, and GATA. Their library contained 30 microsatellites of which they selected eight of the most polymorphic repeatable markers. They validated those markers in 20 individuals of *P. rostriflorus* and calculated their heterozygosities which ranged from 0.40 to 0.95. In addition, they tested their eight markers on 40 *Penstemon* species and they found a high percentage of

amplification in other species. This suggests that the markers they created from the microsatellite enriched genomic library are highly polymorphic markers that can act interspecifically. A subsequent study by Kramer et al. demonstrated the usefulness of those markers in population genetics. In their study they used those same eight microsatellite markers to identify differences in neutral population genetic structure between mountainous populations of *P. deustus*, *P. pachyphyllus* and *P. rostriflorus* (Kramer et al. 2011). Generally, microsatellite markers have traditionally been expensive to produce but the high initial cost of development is predominately determined by the expense of obtaining sequence information. Once the microsatellite markers are produced, they are easy to use and share among different laboratories, highly reproduceable, and are abundant in most plant genomes making them ideal for plant genetic studies (Mason et al. 2005; Maughan et al. 1995). Recent developments in sequencing technology, like the 454 pyrosequencer provide the means by which more microsatellites can be developed in *Penstemon*.

454 PYROSEQUENCING

The development of molecular markers from sequence in organisms with little to no prior sequence data has been extremely problematic. This is certainly the case with the genus *Penstemon*. Traditional sequencing methods are expensive and labor intensive, making them less efficient for use in marker discovery. However, the development of the Roche GS FLX 454 pyrosequencer has greatly facilitated a high-throughput method of marker discovery. In one study conducted by Santana et al. (2009), they discovered that microsatellite detection was increased by one order of magnitude (10 fold increase) by using the 454 pyrosequencer over the traditional cloning techniques. This relatively new technique allows for more bases to be

sequenced at a lower price per base, thus making it more feasible to sequence the genomes of many species.

Pyrosequencing technology is based on the release of pyrophosphate, as DNA is being synthesized complementary to the template strand within a well of the sequencer (Ronaghi 2001). While integrating the dNTPs during synthesis, the DNA polymerase releases inorganic pyrophosphate, which is then converted to ATP by sulfurylase. The ATP provides chemical energy to luciferase, which oxidizes luciferin, generating light. With each additional wash of nucleotides, the light is emitted equivalent to the number of bases incorporated during the synthesis of DNA. Since each "stepwise" wash contains only one type of dNTP, the number and identity of the bases can be inferred for a given well through the use of a photodiode, a photomultiplier tube, or a charge-coupled device camera. This produces a "pyrogram," representing a sequence of bases for each well (Ronaghi 2001). This method of sequencing is fast, relatively reliable, and can produce massive amounts of sequence data for a variety of genetic analyses including the creation of molecular markers.

MARKER DEVELOPMENT FROM 454 PYROSEQUENCING

Although the 454 pyrosequencing technology has not been used extensively with ornamentals, a few studies have involved the creation of molecular markers in other plant species. In a recent study by Wicker et al. (2009) part of the haploid genome of barley (*Hordeum vulgare* cv. Morex), which contains a particularly large and complex genome, was sequenced in two runs of a 454 sequencer yielding approximately 1% of the genome (58.91 Mb). Subsequently, they searched their sequence data for microsatellites, transposable elements, genes, conserved non-coding regions, and other repeatable elements. They were able to discover

11,876 reads containing microsatellites along with other useful data important to their genomic analysis. This research shows the value of 454 pyrosequencing. It can provide a massive amount of data and the potential to extract large numbers of microsatellite markers in a single sequencing run (plate).

In another study, Cañahua (*Chenopodium pallidicaule* Aellen) (an important subsistence crop of the South American Andes) was sequenced using the 454 pyrosequencer producing 192 microsatellite markers from random genomic DNA (Vargas 2010). Of the microsatellite markers derived from the Cañahua sequence information, 48 were polymorphic with an average heterozygosity value of 0.47. Cañahua is a narrowly studied plant with which sequence information was unavailable previous to this study. This study demonstrates the effectiveness of 454 pyrosequencing technology in producing microsatellite markers in a plant with little to no genetic information (like *Penstemon*).

In addition to microsatellite marker development, single nucleotide polymorphisms (SNPs) can be discovered from 454 sequence information. Maughan et al. (2009) extracted 27,658 SNPs from sequence obtained from a single pyrosequencing run containing genomic DNA of four populations of amaranth (genus: *Amaranthaceae*) (a new world grain). SNPs, characterized by single base pair changes, can be used in a variety of genetic studies and as molecular markers as well. Such variations in genetic sequence are useful in the analysis of inter-specific diversity, gene function/structure and in evolutionary relationships (Zhu et al. 2003). Such associations make SNPs an ideal candidate for simple molecular markers that can also be developed from 454 pyrosequencing data.

454 PYROSEQUENCING USING GENOME REDUCTION

With about 500 mega bases as a standard capacity of a FLX 454 pyrosequencing run (plate), it becomes necessary to determine what type of data is needed. One can then electively amplify the needed sequence in order to get enough coverage in a sequencing run. A method of genome reduction, in species with limited to no sequence information, was developed to reduce the complexity of target genomes using restriction enzyme site conservation (GR-RSC) (Maughan et al. 2009). In this method DNA is extracted and double-digested using a six-base rare-cutting restriction enzyme and a four-base frequent-cutting restriction enzyme. This produces three types of fragments: rare by rare, rare by frequent, and frequent by frequent (Maughan et al. 2009). Adapters (double stranded oligonucleotides) are ligated to the restriction sites of the fragments. Primers, complementary to the adapter and the restriction site, are used to amplify the fragments. The adapter ligated to the rare-cutting restriction site is labeled with a 5' biotin molecule, which is used to separate the less-common rare by rare and rare by frequent-cutting fragments.

This separation leads to greater than 90% reduction of the genome data (Maughan et al. 2009). Since this technique uses genomic DNA rather than transcripts, it gives a more random sampling of the genome of interest but with the added benefit of homology based on conserved restriction sites. Since the restriction enzyme cut sites would be conserved within related species or individuals, the fragments produced by this technique will also have a high probability of homology between samples. This would increase the likelihood of being able to adequately compare the genomes and create markers.

454 PYROSEQUENCING USING BARCODES

In addition to genome reductions, 454 pyrosequencing can be greatly enhanced by giving it the ability to sequence several samples at once. This can help alleviate the cost of using a next-generation sequencer, especially since the creation of intra- or inter-specific markers is dependent upon the comparison of two different individuals or species. Manifolds (which physically separate one sample from another), allow for multiple samples on the same plate and are often included with 454 pyrosequencing plates. However, the manifold covers some wells, which decreases the capacity of the sequencer (Parameswaran et al. 2007). To overcome this problem, a novel multiplexing techniques has been developed that involves the use of barcoding technology. In DNA barcoding, ten base pair sequences are incorporated into the fragments that are to be sequenced during PCR by adding the barcode sequence to the forward and reverse primers (Maughan et al. 2009; Parameswaran et al. 2007). This allows all samples to be added into the same sequencing sample mixture (without manifolds) and is bioinformatically separated afterwards. Using a ten base pair barcode with a three base pair linker, up to 48 different samples can be separated with an inaccurate assignment rate less than 0.005% (Parameswaran et al. 2007). Even more samples can be multiplexed if longer barcodes are used.

The 454 pyrosequencer is especially adept to the use of this technique because it produces longer average read lengths than Solexa or SOLiD next-generation sequencers. This allows for more accurate reads of longer, more complex barcodes (Parameswaran et al. 2007). Also, longer read length aids in the process of finding primer combinations that have similar annealing temperatures. This technology coupled with a genome reduction can aid in SNP and microsatellite discovery because multiple samples can be sequenced simultaneously in the same 454-pyrosequencing run; thus decreasing the cost for valuable marker development.

CONCLUSION

In order to breed xeric adapted *Penstemon* cultivars for more widespread landscaping uses, we must develop a better understanding of the vast genetic diversity within the genus on a molecular level. However, that knowledge is limited by the meager number of molecular markers available. Microsatellites are an ideal marker for *Penstemon* because of their co-dominance, reproducibility and frequency in plant genomes. Although microsatellite markers are historically expensive to produce, recent advances in sequencing technology (such as the 454 pyrosequencer) make gathering sequence information cheaper and thus the cost of the production of markers is steadily decreasing. In addition, the use of genome reduction and DNA barcoding can optimize sequencing to extract the most information from fewer runs. The sequence and subsequent molecular markers developed from these "second generation" sequencing technologies will allow the improvement of *Penstemon* through breeding with modern tools. Our specific objective is to develop cultivars which have a wide public acceptance and are highly adapted to the xeric landscape which will only become possible as we learn more about the genome of the *Penstemon* genus, create molecular markers, implement them in revising and updating the genus' phylogeny and making them available for future molecular work and plant breeding in *Penstemon.*

REFERENCES CITED

Boyer JS (1982) Plant productivity and environment. Science 218:443-448

Broderick SR (2010) An examination of the DNA content, taxonomy and phylogeny of *Penstemon* (Plantaginaceae). Brigham Young University, Provo, Masters

Cronquist A, Holmgren AH, Holmgren NH, Reveal JL and Holmgren PK (1984) Intermountain Flora: Vascular Plants of the Intermountain West, U.S.A.. New York Botanical Garden Press, New York

Datwyler SL and Wolfe AD (2004) Phylogenetic relationships and morphological evolution in *Penstemon* subg. *Dasanthera* (Veronicaceae). Syst Bot 29:165-176

Dougher T (2003) Commercializing production of native Intermountain species. Sci Hort 38:752-753

Eathington SR, Crosbie TM, Edwards MD, Reiter RS and Bull JK (2007) Molecular markers in a commercial breeding program. Crop Sci 47(S3): S154-S163

Hungate BA, Holland EA, Jackson RB, Chapin FS, Mooney HA and Field CB (1997) The fate of carbon in grasslands under carbon dioxide enrichment. Nature 388:576-579

Kramer AT and Fant JB (2007) Isolation and characterization of microsatellite loci in *Penstemon rostriflorus* (Plantaginaceae) and cross-species amplification. Mol Ecol Notes 7:998-1001

Kramer AT, Fant JB and Ashley MV (2011) Influences of landscape and pollinators on population genetic structure: examples from three *Penstemon* (Plantaginaceae) species in the great basin. Am J Bot 98:109-121

Lindgren DT (2000) Breeding *Penstemon*: Breeding Ornamental Plants. Timber Press, Portland, pp. 196-212

Lodewick K and Lodewick R (1999) Key to the genus *Penstemon* and its related genera in the tribe Cheloneae (Scrophulariaceae). K. Lodewick, Eugene

Mason SL, Stevens MR, Jellen EN, Bonifacio A, Fairbanks DJ, Coleman CE, McCarty RR, Rasmussen AG and Maughan PJ (2005) Development and use of microsatellite markers for germplasm characterization in quinoa (*Chenopodium quinoa* Willd.) Crop Sci 45:1618-1630

Maughan PJ, Saghai Maroof MA and Buss GR (1995) Microsatellite and amplified sequence length polymorphisms in cultivated and wild soybean. Genome 38:715-723

Maughan PJ, Yourstone SM, Jellen EN and Udall JA (2009) SNP discovery via genomic reduction, barcoding and 454-pyrosequencing in amaranth. Plant Gen 2:260-270

Nold R (1999) *Penstemons*. Timber Press, Portland

Parameswaran P, Jalili R, Tao L, Shokralla S, Gharizadeh B, Ronaghi M and Fire AZ (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. Nucleic Acids Res 35:19

Ronaghi M (2001) Pyrosequencing sheds light on DNA sequencing. Genome Res 11:3-11

Roy SB, Summers KV and Goldstein RA (2003) Water sustainability in the United States and cooling water requirements for power generation. Water Resources Update (UCOWR) 126:94-99

Santana QC, Coetzee MPA, Steenkamp ET, Mlonyeni OX, Hammond GNA, Wingfield MJ and Wingfield BD (2009) Microsatellite discovery by deep sequencing of enriched genomic libraries. BioTechniques 46:217-223

United States Environmental Protection Agency (2002) Water efficient landscaping: preventing pollution & using resources wisely. [Online]. Available at http://www.epa.gov/owm/water-efficiency/docs/water-efficient_landscaping_508.pdf

United States Geological Survey (2009) Summary of estimated water use in the United States in 2005. [Online]. Available at http://pubs.usgs.gov/fs/2009/3098/

Vargas A (2010) Development and use of microsatellite markers for genetic diversity analysis of Cañahua (*Chenopodium pallidicaule* Aellen). Brigham Young University, Provo, Masters

Way D and James P (1998) The Gardener's Guide to Growing *Penstemon.* Timber Press, Portland

Welsh SL, Atwood ND, Goodrich S and Higgins LC (2008) A Utah Flora. Brigham Young University, Provo

Wicker T, Taudien S, Houben A, Keller B, Graner A, Platzer M and Stein N (2009) A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. Plant J 59:712–722

Wilson P, Wolfe AD, Armbruster WS and Thomson JD (2007) Constrained lability in floral evolution: Counting convergent origins of hummingbird pollination in *Penstemon* and *Keckiella*. New Phytol 176:883-890

Wolfe AD, Datwyler SL and Randle CP (2002) A phylogenetic and biogeographic analysis of the Cheloneae (Scrophulariaceae) based on ITS and *matK* sequence data. Syst Bot 27:138-148

Wolfe AD and Elisens WJ (1993) Diploid hybrid speciation in *Penstemon* (Scrophulariaceae) revisited. Amer J Bot 80:1082-1094

Wolfe AD and Elisens WJ (1994) Nuclear ribosomal DNA restriction site variation in *Penstemon* section *Peltanthera* (Scrophulariaceae) - an evaluation of diploid hybrid speciation and evidence for introgression. Amer J Bot 81:1627-1635

Wolfe AD and Elisens WJ (1995) Evidence of chloroplast capture and pollen-mediated gene flow in *Penstemon* sect. *Peltanthera* (Scrophulariaceae). Syst Bot 20:395-412

Wolfe AD, Randle CP, Datwyler SL, Morawetz JJ, Arguedas N and Diaz J (2006) Phylogeny, taxonomic affinities, and biogeography of *Penstemon* (Plantaginaceae) based on ITS and cpDNA sequence data. Amer J Bot 93:1699-1713

Wolfe AD, Xiang QY and Kephart SR (1998a) Assessing hybridization in natural populations of *Penstemon* (Scrophulariaceae) using hypervariable intersimple sequence repeat (ISSR) bands. Mol Ecol 7:1107-1125

Wolfe AD, Xiang QY and Kephart SR (1998b) Diploid hybrid speciation in *Penstemon* (Scrophulariaceae). Proc Natl Acad Sci USA 95:5112-5115

Zhao M and Running SW (2010) Drought-induced reduction in global terrestrial net primary production from 2000 through 2009. Science 329:940-943

Zhu YL, Song JQ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND and Cregan PB (2003) Single-nucleotide polymorphisms in soybean. Genetics 163:1123–1134

Zollinger N, Kjelgren R, Cerny-Koenig T, Kopp K and Koenig R (2006) Drought responses of six ornamental herbaceous perennials. Sci Hort 109:267-274

TABLES AND FIGURES

Table 1. Plant tissue sources and summary data from the four multiplex identifiers (MID) barcodes (adapter) primers used for the genomes of *P. cyananthus*, *P. dissectus*, *P. davidsonii*, and *P. fruticosus*. Blue sequences show where adapters complement enzyme cut site and red bases show where the base was changed to avoid further enzymatic cleavage of fragment.

Species	Source / Collection Site(s)	MID ID #	EcoR1 MID Primer	Bfa1 MID Primer
P. cyananthus	Utah Co., UT; Wasatch Co., UT	MID 1	5'- ACGAGTGCGTGACTGCGTACCAATTC	5'- ACGAGTGCGTGATGAGTCCTGAGTA
P. dissectus	Purchased[a] FL	MID 2	5'- ACGCTCGACAGACTGCGTACCAATTC	5'- ACGCTCGACAGATGAGTCCTGAGTA
P. davidsonii	Purchased[b] Jackson Co, CO	MID 3	5'- AGACGCACTCGACTGCGTACCAATTC	5'- AGACGCACTCGATGAGTCCTGAGTA
P. fruticosus	Purchased[a]	MID 4	5'- AGCACTGTAGGACTGCGTACCAATTC	5'- AGCACTGTAGGATGAGTCCTGAGTA

[a] Purchased as a mature plant from a nursery

[b] Purchased as seed from alplains.com

Table 2. Summary data from 454-pyrosequencing and Newbler de-novo assembly (v.2.0.01) of *P. cyananthus*, *P. dissectus*, *P. davidsonii*, and *P. fruticosus*.

Assembly	Genome Size (Mbp)[a]	GC Content	Reads	Bases[b]	% Reads Assembled	% Bases Assembled	Contigs Created	Bases in Assembly	% Genome Represented	Average Coverage	Bases Shared Between Assemblies
P. cyananthus	893	36.4%	199,329	87,753,792	53.1%	50.0%	9,714	4,623,755	0.5%	7.7X	
P. dissectus	462	34.5%	98,868	43,304,550	52.8%	50.9%	5,364	2,629,819	0.6%	8.2X	
P. davidsonii	483	35.3%	103,963	45,599,742	45.8%	43.5%	4,882	2,376,141	0.5%	9.1X	
P. fruticosus	476	35.2%	113,146	49,786,980	41.0%	38.8%	4,777	2,322,606	0.5%	8.9X	
P. cyananthus x P. dissectus			298,197	131,058,342	53.0%	50.1%	14,523	6,915,079			338,495
P. cyananthus x P. davidsonii			303,292	133,353,534	49.9%	46.9%	14,254	6,757,023			242,873
P. cyananthus x P. fruticosus			312,475	137,540,772	47.8%	44.9%	14,134	6,705,536			240,825
P. dissectus x P. davidsonii			202,831	88,904,292	48.3%	46.1%	10,053	4,855,491			150,469
P. dissectus x P. fruticosus			212,014	93,091,530	45.7%	43.5%	9,873	4,774,539			177,886
P. davidsonii x P. fruticosus			217,109	95,386,722	44.0%	41.7%	9,184	4,442,194			256,553
Full Penstemon Assembly			730,215	265,987,500	47.9%	46.4%	44,966	16,363,589			

[a] Genome size as reported by (Broderick et al. 2011)

[b] Bases denotes the total number of bases used to create the assembly and not the total number of bases sequenced

Table 3. Summary of microsatellite marker characteristics including the primary microsatellite motif, primer sequences, expected fragment length (EFL), total alleles, fragment sizes (estimated from 3% Metaphor gel ran for 2 hours), and heterozygosity (H).

Marker Name[a]	Primary Motif	Forward Primer (5'-3')	Reverse Primer (5'-3')	EFL	Total Alleles	Fragment Size				H[b]
						P. cyananthus	*P. davidsonii*	*P. dissectus*	*P. fruticosus*	
PenSSR_003 (di,f)	(AT)8	TGCCTCTGTCTTTACATTCCAA	CATGAAGCACTGCAAATCCA	217	3	360	260	250	260	0.625
PenSSR_004 (da,f)	(ATT)6	TGTTTCAATTGCTGTCCACAT	TTGTCTGTCCAAACGGTAGGT	476	3	420	460	440	460	0.625
PenSSR_005 (c,di,f)	(GAA)6	GCCCAACTTCCGTAATTGAA	AACTGCTTGCCACTCGACTC	303	3	260,300	260	280	280	0.438
PenSSR_009 (c,da,f)	(TGA)6	ACCTCGAACTTGACGGTCC	TTCTGAGGAGAAACCAAGGG	466	4	370, 650	540	650	600	0.563
PenSSR_011 (da,f)	(GA)8	AAGTGCGACACTGGATGTCTT	GCAGCTTCAGCTCCAGAAAT	435	2	860	500	860	500	0.500
PenSSR_012 (c)	(TA)8	TCCATATTGTAACCAACAATGACTG	TGAATGGCAAACCGTAATCA	402	3	400	360	370	360	0.625
PenSSR_013 (f)	(TA)8	GAAGAATTGATTTAAACAAGATGCAA	TCAGTACGTGAGAAACTTGATCAATAA	399	2	400	650	650	400	0.500
PenSSR_014 (c)	(TGA)6	CGATTTGGTATAGTTGGATTACGA	CCTTCATCACCCGGTACTTG	409	3	410	370	380	410	0.625
PenSSR_015 (di)	(TCG)6	GCCGAGTTTCAAGAAAGCAA	AATTACGACCTGCCACGC	409	2	490	500	490	490	0.375
PenSSR_016 (c,di)	(CT)8	CATGGCCCTTTCTTCACACT	GACGCGGTTGGCTATACAGT	447	3	NA[c]	1100	1060	1030	0.813
PenSSR_017 (da,di)	(AG)9	GAAGGCTTAGCATAAATCCTCAAA	ATTAGGCTCCCACGAACAAA	455	2	750	700	750	700	0.500
PenSSR_019 (c,di)	(AG)8	AATCCCACAGCCCATACAAA	TGAATTGAGTCCTATACCCTATTTCAA	473	1	380	380	380	380	0.000
PenSSR_021 (f)	(CT)8	CTTTAGCTTAGCTGGAATACACGTT	AGATTCTTGCATCACAGTTCAATTA	386	3	350	450	450	420	0.625
PenSSR_023 (da)	(AG)8	GCTGGAGAATAACATGGCG	CCATCTTGCAAGTCCATACG	469	4	310	480	120, 740	480	0.563
PenSSR_024 (da,f)	(CTG)6	CTTCTTGCCCTGTGCCTCT	CCACCACCAACAACAACAAC	403	2	430	430	400	430	0.375
PenSSR_025 (c,di)	(TC)9	GCACATGAATGAAGGAATGC	ACGATCTGTGAAGGAACCCA	440	3	440	410	440	400	0.625
PenSSR_026 (c,da,di,f)	(CTT)6	ACTTAATAATGCCTCCTTGTGTCA	TTCCGCAACGTTGTATTTGA	465	1	460	NA[c]	460	460	0.438
PenSSR_028 (di)	(AC)9	GGGAGGCAGGTAACAACAAA	TACCTCTGCCGAACTGGATT	316	4	950	400, 460	320	400, 460	0.375
PenSSR_029 (di)	(TA)8	ACCAAGTTGTTGGATGTTTGG	GGTTTGGAATGAGACTTAGAAGGA	440	3	840	500	500	420	0.625
PenSSR_032 (c,di)	(GT)9	ACAAAGTCTCCTCAATCGCC	GCATGTACCGTGCACACACT	328	2	370	370	370	340	0.375
PenSSR_035 (da,f)	(TC)9	TTGCACAGCTACTTTGGCAT	ATCTGTCCAAGGCATGGAAT	486	3	630	520	920	520	0.625
PenSSR_036 (c,di)	(TA)8	TTCCTAATTTGGTAGCTGCAATC	TCCGAGGAACTATTGCCATT	405*	3	770	770, 820	590	770	0.313
PenSSR_038 (c,da)	(TA)8	GTAATTACTTCGGCAGTTTGTTAATTT	GGTGCGACCTAATTACGTTTCTAT	100	1	NA[c]	100	100	NA[c]	0.750
PenSSR_040 (da)	(CA)9	TAAAGAGGCTTAAGCGCGG	ACCTGAAGAGCTGCGGAGTA	399	3	380	390	410	390	0.625
PenSSR_041 (c,da,di,f)	(AT)8	TTTCCGCAAGAGAAGAGCAT	CTTGTGCACGATTCCATTGT	249	3	270	670	270	240	0.625
PenSSR_045 (c,da)	(CT)8	GCCACATACATGAAACGTGAA	CGAACTCTCTTGTGTTTCTCCC	366	4	460	NA[c]	440	120, 400	0.750
PenSSR_047 (c,di,f)	(AC)8	ACACGACATCGTTTCAGCAA	GCGTATGGAGAGATTTGGGA	428	3	470, 510	440	470	470	0.313
PenSSR_048 (c,di)	(CA)9	GCATTAGATGCCGAAATATCTACAA	TGCCTGTAGGTTGATTTCCTTT	436	3	420	440	380	420	0.625
PenSSR_049 (c,da,di,f)	(AG)8	CCCATCAATAAAGAAAGAAAGAAAGA	GGTGAAACCCTGTCCTAAACC	436	2	460	460	1000	460	0.375
PenSSR_050 (c,di)	(AT)9	GTGTAACCTCTGAACAAGTTTACTGAA	TGCAGTGAGCCATGCTATTC	434	2	480	460	480	460	0.500
PenSSR_051 (c,di,f)	(TG)8	TGTAACACGACAATTTAACTCTTTCA	CGAGAACTCTTTCCGAGAACC	352	1	280	NA[c]	280	280	0.438
PenSSR_052 (c,da,di,f)	(AC)9	CGCGGTCAATCTTGAAATCT	TGACTTCCTCTCTCTCTCACAC	206	1	220	220	220	220	0.000
PenSSR_053 (di)	(AC)8	AATCATAGTCTCGAGCGCGT	GAGATAAATTAGATCAGCGCATCA	410	3	160, 320	320	320, 450	320	0.875
PenSSR_054 (c,da,f)	(GA)8	TCGTTAAGCAATCTCGGAGC	TCGACTGGAGAGCAAAGCA	192	3	200	200	180	190	0.625

PenSSR_055 (f)	(AG)8	TGTGGTCCGGTTCCATAAAC	TTTGTCTCCCTAATATGTGTGATGAT	412	4	960	500	1040	470	0.750
PenSSR_056 (da,di)	(TG)8	CATGTTTCAGGATTGGGCTT	CGGTTACACACAGGTTGTTGA	319	4	690	450	230	340	0.750
PenSSR_057 (da,f)	(AT)8	TGCCTAATGGACCTGATCCT	CCCAATTGTTTGAAGAAAGAACA	402	2	570	440	570	440	0.500
PenSSR_058 (da)	(AT)9	GTGCAACCAATGCAACTAATTC	TCTCTCATTTCCAATGATTTCTCA	469*	1	NA[c]	720	NA[c]	720	0.750
PenSSR_059 (di)	(CT)8	CATCAATTGACACACAAGCAGA	TCGAATCTTAAAGAAACACATCCA	312	2	930	340	340	340	0.375
PenSSR_060 (c,di)	(AC)9	CCATGAGAAGTAGATGACTGGGA	TTGTAATTATGATTAACTTCCCTCGTT	484	2	560	560	560	540	0.375
PenSSR_061 (da,f)	(TA)8	CGACCAATCATCAACCAACA	GACGGGCAGAATAATTGGAA	453	3	480	480, 530	450, 480	NA[c]	0.313
PenSSR_062 (c,di)	(TA)9	TGGAGAGGGTACGAAAGTGC	CAACGATCGATTATTAGCACCA	320	2	350	290	350	290	0.500
PenSSR_064 (c,da,f)	(AG)8	ATGGATGCCCTATGGGTACA	TGAAATGGAGGGAGTAATATAAACAA	437	4	490	500, 680	470	470	0.563
PenSSR_066 (di)	(GA)9	CAAGGATGCAGGCTCTCATT	CTCTGCTCGTCGTAGTGCAA	434	2	250	480	250, 480	480	0.188
PenSSR_068 (c,da,di,f)	(GA)8	TTTGGGATGCATTTCTCCAC	TCAAAGTGACATCTTCCAACAAA	463	2	500	500	480	480	0.500
PenSSR_069 (di)	(GT)8	CATTGGGTCAGATTTGGCTT	GCTTTCAGTTTGTATATTTGTGCC	309	4	220	210	390	350	0.750
PenSSR_071 (c,di)	(AT)8	AAGATGGCCCTGATCTGTTG	TTCGTGGGAGTTGCAAATTA	446	1	NA[c]	NA[c]	490	NA[c]	0.938
PenSSR_074 (c,da,di,f)	(AAG)6	AGAAATCTCGCTCTCCACGA	CGACAACCTTAGTGATCGCTTT	168	1	170	170	170	170	0.000
PenSSR_075 (c,da,di,f)	(TA)8	CACCACTTTCGCAGCATTTA	CAAATTACATTATTGTATGGAAACACG	120	2	160	140	140	140	0.375
PenSSR_076 (c,di)	(GTG)6	CTGACAGCAACATGAACATGAA	CAATCTTTGCCAATTTCCCA	161	1	170	170	170	170	0.000

[a] Parentheses indicates the species possessing sequence from which primers were designed (c = *P. cyananthus*, da = *P. davidsonii*, di = *P. dissectus*, f = *P. fruticosus*)

[b] Heterozygosity calculated according to the formula $H = 1-\Sigma p_i^2$ where p_i represents the frequency of allele "i" (Pettersson et al. 1995)

[c] NA = no allele

Table 4. Data obtained from MISA (SSR), Blast2Go (GO) and RepeatMasker (RM).

			Penstemon Species			
			P. cyananthus	*P. dissectus*	*P. davidsonii*	*P. fruticosus*
SSR	Total SSRs[a]		97	113	49	58
	SSRs / Assembly Length		2.1E-05 (~1/48000)	4.3E-05 (~1/23000)	2.1E-05 (~1/48000)	2.5E-05 (~1/40000)
	Repeat Type	di-	44.3%	40.7%	46.9%	48.3%
		tri-	45.4%	43.4%	44.9%	41.4%
		tetra-	10.3%	15.9%	8.2%	10.3%
GO	Contigs Analyzed		9,714	5,364	4,882	4,777
	Blast Hits Found[b]		1,899	1,125	1,121	1,091
	Annotated Hits		1,430	844	388	826
	% Blast Hits		19.5%	21.0%	23.0%	22.8%
	% Annotated		14.7%	15.7%	7.9%	17.3%
RM	Masked Repeat Elements		28.5%	16.8%	17.4%	16.1%
	Retroelements (LTR)		7.8%	3.0%	4.9%	4.6%
	DNA Transposons		0.3%	0.9%	1.0%	1.0%
	Other Repeats[c]		20.4%	12.9%	11.6%	10.5%

[a] For MISA, "unmasked" individual species assemblies were used

[b] sequence compared to the GenBank refseq-protein database e-value threshold of $<1.0e^{-15}$

[c] Other Repeats includes: lines, unclassified repeats, satellites, simple repeats, and low complexity sequence

Table 5. Comparisons of SNPs found within species (heterozygosity) and homologous sequences between species using individual and dual species assemblies using SNP_Finder_Plus (8X min. coverage, 30% min. minor allele, 90% min. identity). In addition the average coverage per SNP per Assembly and the SNP type distribution are shown.

Species Assembly	SNP	Average Coverage	SNPs / Assembly Length[a]	SNP distribution					
				A↔C	A↔G	A↔T	C↔G	C↔T	G↔T
P. cyananthus	2,493	16.4	0.000539 (~1/1855 bp)	10.7%	29.5%	13.9%	4.3%	30.2%	9.5%
P. dissectus	737	14.3	0.000280 (1/3568 bp)	9.8%	30.7%	15.6%	4.6%	27.4%	9.8%
P. davidsonii	713	14.4	0.000300 (~1/3333 bp)	11.9%	26.4%	15.2%	3.9%	28.3%	11.8%
P. fruticosus	615	12.4	0.000265 (~1/3777 bp)	11.7%	27.2%	17.9%	4.2%	25.4%	12.0%
Species Assembled	SNP	Average Coverage	SNPs / Assembly Length[a]	SNP distribution					
				A↔C	A↔G	A↔T	C↔G	C↔T	G↔T
P. cyananthus x P. dissectus	3,253	10.6	0.009610 (~1/104 bp)	11.7%	27.5%	16.0%	7.1%	27.1%	10.6%
P. cyananthus x P. davidsonii	1,958	10.7	0.008062 (~1/124 bp)	11.1%	27.6%	15.8%	7.1%	28.5%	9.9%
P. cyananthus x P. fruticosus	2,015	10.6	0.008367 (~1/119 bp)	10.6%	27.2%	16.7%	6.8%	28.7%	10.1%
P. dissectus x P. davidsonii	2,348	10.8	0.015605 (~1/64 bp)	12.6%	26.7%	15.5%	7.5%	27.3%	10.4%
P. dissectus x P. fruticosus	2,133	10.0	0.011991 (~1/83 bp)	12.0%	26.4%	16.5%	7.6%	27.2%	10.4%
P. davidsonii x P. fruticosus	2,156	10.1	0.008404 (~1/119 bp)	12.8%	28.2%	14.5%	7.2%	27.2%	10.1%

[a] Assembly Length is bases shared between assemblies (see Table 2)

Table 6. Sequence comparison to known genes in the genera *Antirrhinum* and *Mimulus* using BLASTN and an e-value of <1.0e^{-6}.

Gene/Sequence	Gen Bank Accession	Species	*P. cyanathus*		*P. dissectus*		*P. davidsonii*		*P. fruticosus*	
			Contig	e-value	Contig	e-value	Contig	e-value	Contig	e-value
PLENA protein	AB516404.1	*A. majus*	C04500	2.0e-56						
cyclin D1 (cycD1)	AJ250396.1	*A. majus*			C00374	1.0e-27				
putative transposase (rsi-AT1)	AJ849555.1	*A. majus*	C02682	5.0e-08						
YA3 (nf-YA)	AM422772.1	*A. majus*			C03142	3.0e-14	C05173	5.0e-16		
RNA polymerase II second largest subunit (RPB2)	DQ020637.1	*A. majus*			C02648	2.0e-104				
RNA polymerase II largest subunit (RPB1)	DQ020643.1	*A. majus*	C18096	2.0e-103	C06219	6.0e-102	C06342	6.0e-102	C06825	6.0e-102
ROSEA2 (Rosea2)	DQ275530.1	*A. majus*			C03200	3.0e-36				
delila (DEL)	M84913.1	*A. majus*					C01605	1.0e-28		
cyclin-dependent kinase (cdc2b)	X97638.1	*Antirrhinum* sp.	C05259	2.0e-42	C00902	9.0e-43				
bZIP DNA-binding protein	Y13675.1	*A. majus*							C01674	8.0e-28
ribulose 1,5-bisphosphate carboxylase large subunit (rbcL)	AF026835.1	*M. aurantiacus* (3)[a]			C01599	2.0e-10	C00408	1.0e-155		
NADH dehydrogenase (ndhF)	AF188186.1	*M. aurantiacus*							C05403	0.0
tRNA-Leu (trnL) and trnL-trnF intergenic spacer	AF479000.1	*M. ringens* (117)[a]			C05569	1.0e-105	C02286	3.0e-106		
ribosomal protein S16 (rps16)	AJ609163.1	*M. aurantiacus* (3)[ab]	C01797	7.0e-70					C05122	5.0e-44
maturase K (matK)	AY849605.1	*M. aurantiacus*			C03520	6.0e-177				
large ribosomal protein 16 (rpl16)	DQ090908.1	*M. guttatus* (15)[ac]	C06023	3.0e-59					C04498	1.0e-173
ATPase subunit 1 (atp1)	EU551652.1	*M. guttatus*					C02505	1.0e-124		
ATPase subunit 8 (atp8)	EU551655.1	*M. guttatus*	C08546	1.0e-168	C03976	8.0e-169	C03430	7.0e-169	C03231	7.0e-169
ATPase subunit 9 (atp9)	EU551656.1	*M. guttatus*							C01152	9.0e-35
NADH dehydrogenase subunit A (nad3)	EU551661.1	*M. guttatus*	C04937	1.0e-147	C01200	8.0e-148	C02107	7.0e-148	C03997	7.0e-148
ribosomal protein subunit 2 (rps2)	FJ172718-9.1, AF055154.1	*Mimulus* (3)[d]	C09121	0.0	C03314	0.0	C03887	0.0	C02758	0.0
ribosomal protein L10 (rpl10)	GQ402499.1	*M. guttatus*	C05893	0.0	C03195	0.0	C03361	0.0	C03196	0.0
maturase (matR)	HQ593753.1	*M. guttatus*	C00372	1.0e-35					C04420	4.0e-68
ATP synthase subunit 6 (atp6)	HQ593782.1	*M. guttatus*	C07054	7.0e-49						

[a] multiple species (number shown in parenthesis) with hits; lowest species e-value is shown
[b] In *P. fruticosus*, *Mimulus szechuanensis* (FJ172706.1) was the closest match
[c] In *P. fruticosus*, *Mimulus latidens* (DQ090903.1) was the closest match and there were 19 total species hits
[d] multiple species (*M. aurantiacus*, *M. szechuanensis*, and *M. tenellus var. tenellus*) with hits; e-value equal

Figure 1. Un-rooted neighbor-joining tree created on PAUP v. 4.0b10 using a binary matrix (1 = allele present; 0 = allele absent) derived from 50 validated microsatellite markers.

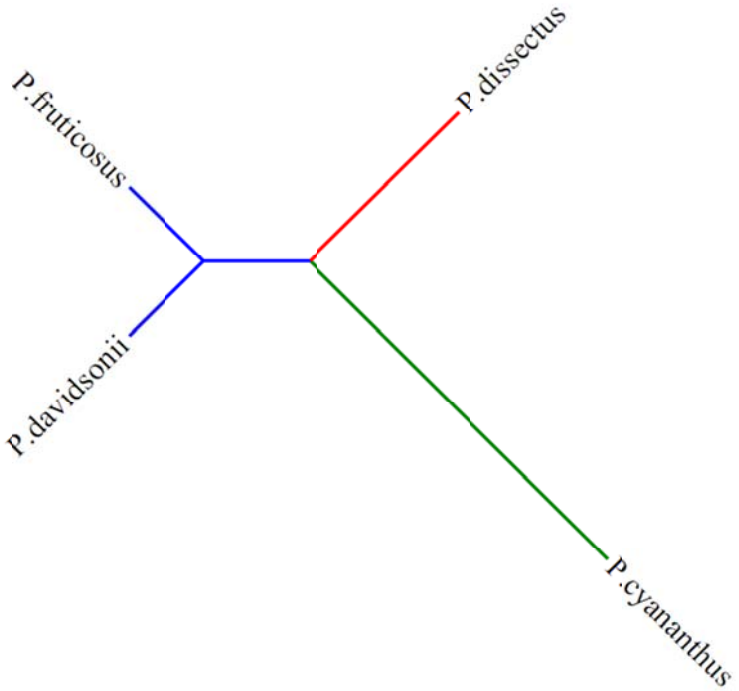


Figure 2. Microsatellite Repeat Motif distributions identified in each of the four *Penstemon* sequences using the program MISA.

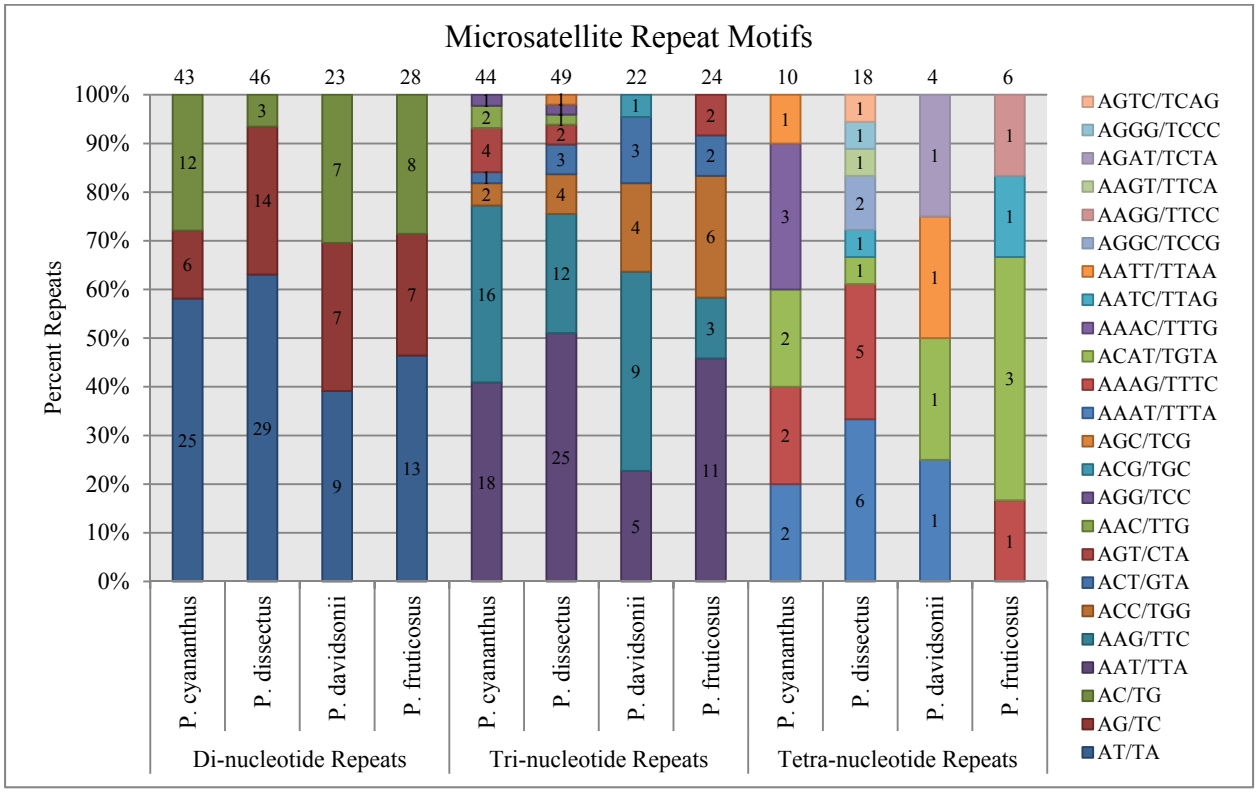


Figure 3. Cellular, biological, and molecular gene ontology comparisons between *P. cyananthus*, *P. dissectus*, *P. davidsonii*, and *P. fruticosus*.

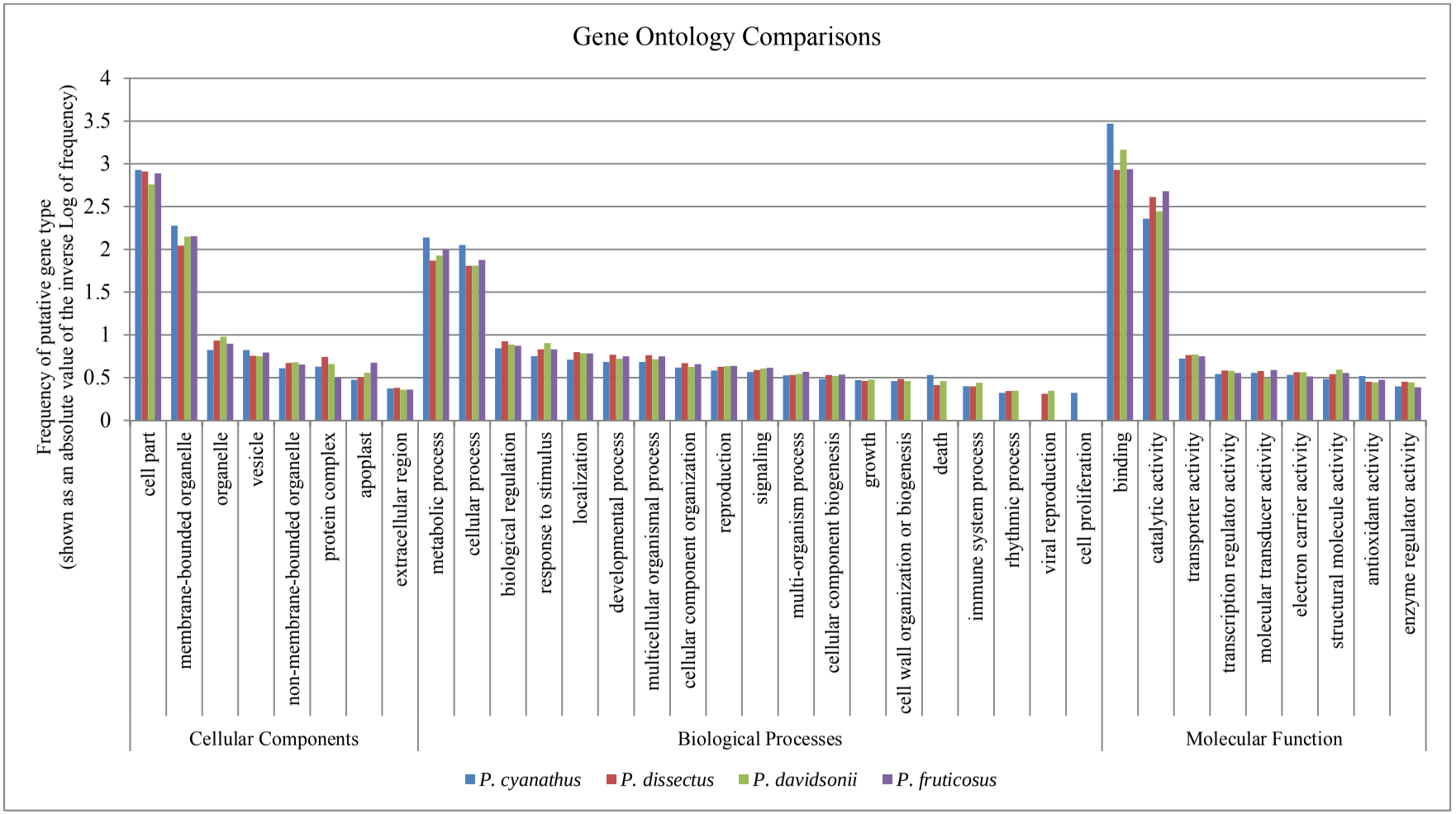
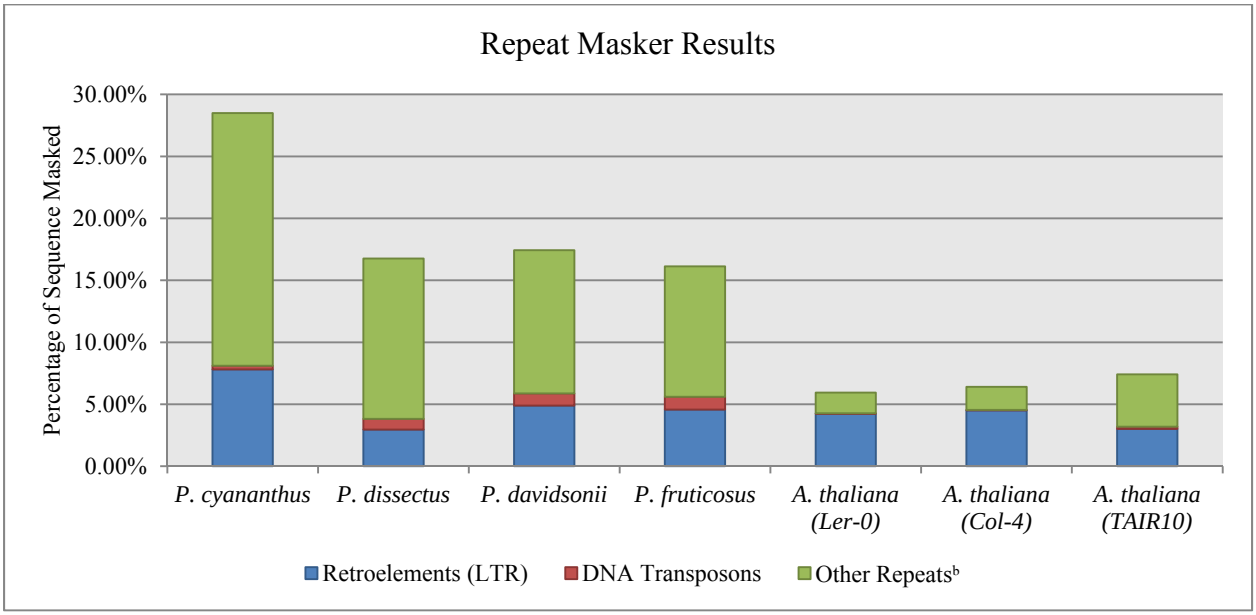


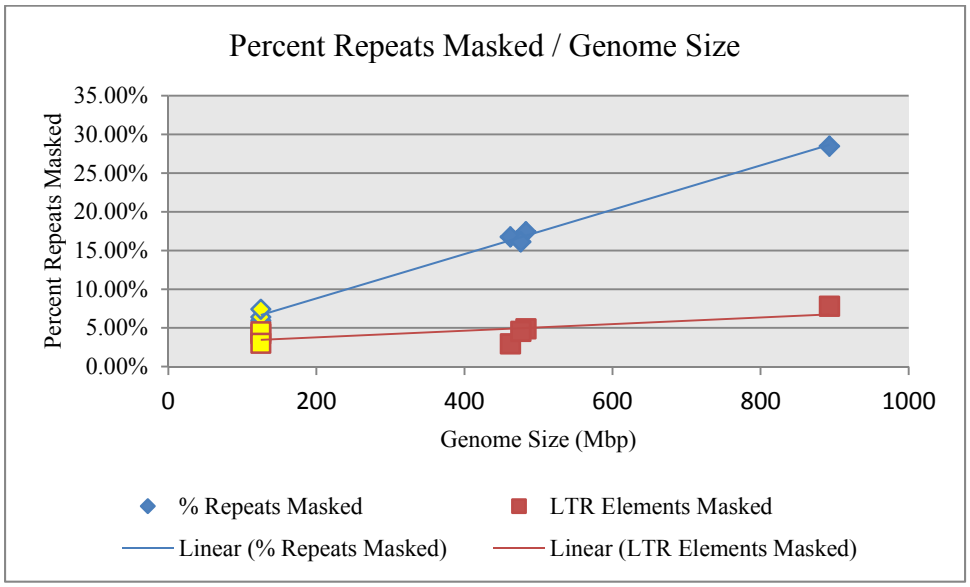
Figure 4. Percentage of retroelements, DNA transposons and other unclassified repeats in *P. cyananthus*, *P. dissectus*, *P. davidsonii*, *P. fruiticosus, and* both genome reduced and non-genome reduced *Arabidopsis*[a].



[a] Genome reduced *A. thaliana* sequence from Maughan et al. 2010; *A. thaliana* RIL lines Ler-0 and Col-4 originally described by Lister and Dean 1993; Non-genome reduced *A. thaliana* (TAIR10) sequence downloaded from TAIR (The *Arabidopsis* Information Resource) as whole chromosomes

[b] Other Repeats includes: lines, unclassified repeats, satellites, simple repeats, and low complexity sequence

Figure 5. Relationship between genome size and repeat elements in *Penstemon* including the relationship of both LTRs and total repeat elements to genome size for both genome reduced *Penstemon* and genome reduced/non-genome reduced *Arabidopsis*[a] (yellow).



[a] Genome reduced *A. thaliana* sequence from Maughan et al. 2010; *A. thaliana* RIL lines Ler-0 and Col-4 originally described by Lister and Dean 1993; Non-genome reduced *A. thaliana* sequence downloaded from TAIR (The *Arabidopsis* Information Resource) as whole chromosomes; Genome size as reported by Schmuths et al. 2004 and Broderick et al. 2011