Characterization of duplicate gene evolution in the recent natural allopolyploid *Tragopogon miscellus* by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping

Richard J. A. Buggs[1,2,*], Srikar Chamala[1,3,*], Wei Wu[4], Lu Gao[4], Gregory D. May[5], Patrick S. Schnable[4], Douglas E. Soltis[1,3], Pamela S. Soltis[2,3], W. Brad Barbazuk[1,3]

[1]Department of Biology, University of Florida, Gainesville, Florida 32611.
[2]Florida Museum of Natural History, University of Florida, Gainesville, Florida 32611.
[3]Genetics Institute, University of Florida, Gainesville, Florida 32610.
[4]Center for Plant Genomics, Iowa State University, Ames, Iowa 50011.
[5]National Center for Genome Resources, Santa Fe, New Mexico 87505.
*These authors contributed equally to this manuscript.

Corresponding author: Dr W. Brad Barbazuk, Department of Biology, University of Florida, Gainesville, Florida 32611, USA.
Email: bbarbazuk@ufl.edu
Telephone: 352-273-8624

Key words: polyploidy, whole genome duplication, *Tragopogon miscellus*, homoeolog evolution, pyrosequencing
Running title: Homoeolog loss in a natural allopolyploid
Word count: 5813

Abstract

Tragopogon miscellus (Asteraceae) is an evolutionary model for the study of natural allopolyploidy, but until now has been under-resourced as a genetic model. Using 454 and Illumina expressed sequence tag sequencing of the parental diploid species of *T. miscellus*, we identified 7782 single nucleotide polymorphisms that differ between the two progenitor genomes present in this allotetraploid. Validation of a sample of 98 of these SNPs in genomic DNA using Sequenom MassARRAY iPlex genotyping confirmed 92 SNP markers at the genomic level that were diagnostic for the two parental genomes. In a transcriptome profile of 2989 SNPs in a single *T. miscellus* leaf, using Illumina sequencing, 69% of SNPs showed approximately equal expression of both homeologs (duplicate homologous genes derived from different parents), 22% showed apparent differential expression, and 8.5% showed apparent silencing of one homeolog in *T. miscellus*. The majority of cases of homeolog silencing involved the *T. dubius* SNP homeolog (164/254; 65%) rather than the *T. pratensis* homeolog (90/254). Sequenom analysis of genomic DNA showed that in a sample of 27 of the homeologs showing apparent silencing, 23 (85%) were due to genomic homeolog loss. These methods could be applied to any organism, allowing efficient and cost-effective generation of genetic markers.

Introduction

Many natural and domesticated plant species are hybrids which have undergone whole-genome duplication. This condition, known as allopolyploidy (Kihara & Ono 1927), may have large effects on both the ecology (e.g. Stebbins 1942; Buggs & Pannell 2007) and evolution (Soltis & Soltis 1999; Adams & Wendel 2005) of a lineage. Genome evolution of allopolyploids has been extensively studied in crop species such as cotton (Adams & Wendel 2004; Udall & Wendel 2006), wheat (Feldman *et al.* 1997; Levy & Feldman 2004; Dong *et al.* 2005; Bottley *et al.* 2006), soybean (Joly *et al.* 2004), and tobacco (Lim *et al.* 2004; Petit *et al.* 2007), as well as genetic models such as *Arabidopsis* (Chen *et al.* 2004; Chen *et al.* 2008). These studies demonstrate dynamic patterns of evolution, but have limitations due to uncertainties about the precise history and ecological context of the lineages. Furthermore, they cannot provide insights into the early stages of polyploid evolution in nature. It is therefore difficult to know whether certain evolutionary changes took place in the progenitor diploids, upon allopolyploidization, or in the subsequent generations.

A need therefore exists for natural allopolyploid model organisms with a known history and ecological context (Soltis *et al.* 2004b; Buggs 2008). A handful of species have been identified for this purpose, such as *Senecio cambrensis* (Hegarty *et al.* 2005), *Spartina anglica* (Ainouche *et al.* 2004), *Tragopogon mirus* and *T. miscellus* (Soltis *et al.* 2004a). *Tragopogon miscellus* is a particularly tractable evolutionary model for the study of the early generations of allopolyploidy. Its origin can be accurately dated to about 80 years ago (Ownbey 1950; Soltis *et al.* 2004a). The parental diploid species are known and still coexist with their allopolyploid derivative; both reciprocal crosses of the parents exist in natural populations, and at least one of them appears to have originated multiple times (Novak *et al.* 1991; Soltis *et al.* 1995; Symonds *et al.* 2009). *Tragopogon miscellus* is a textbook example of allopolyploid speciation (e.g. Judd *et al.* 2007; Sadava *et al.* 2008).

Unlike the crop species that have been used to study allopolyploid evolution, the natural allopolyploid evolutionary model systems are under-resourced as genetic models. To date, the best resourced is *S. cambrensis* for which cDNA microarrays have been made to study gene expression (Hegarty *et al.* 2005; Hegarty *et al.* 2006). Until now resources for *T. miscellus* have consisted of DNA sequence tags for only 23 duplicate gene pairs (Tate *et al.* 2006; Buggs *et al.* 2009; Tate *et al.* 2009a), a handful of phylogenetic markers (Mavrodiev *et al.* 2005), and 2000 uncharacterized Sanger ESTs (J. Koh, J. Tate, D. Soltis and P. Soltis, unpubl. data). This paucity of sequence data contrasts with the usefulness of *T. miscellus* as an evolutionary model.

One key issue in the evolution of allopolyploids is the fate of duplicated genes. Duplicate gene evolution is important for understanding the evolution of the allopolyploids themselves, and may allow for more general statements about the evolution of duplicated genes in non-polyploid organisms. Natural allopolyploid models present systems containing a whole genome’s worth of duplicated genes of identical and known age. Duplicated genes may have a variety of evolutionary fates: non-functionalization, subfunctionalization and neo-functionalization (Lynch & Conery 2000). Several studies have examined the evolution of homeologs (genes duplicated by whole-genome duplication) in allopolyploids. Studies in crop species have shown homeolog loss (e.g. Song *et al.* 1995; Kashkush *et al.* 2002) and patterns of homeolog expression suggestive of subfunctionalization (e.g. Adams *et al.* 2003; Flagel *et al.* 2008).

In natural models, our knowledge of homeolog evolution is limited. In the *S. cambrensis* cDNA microarray, the oligo-nucleotides used did not distinguish between homeologs: measures of gene expression were the total expression of both homeologs. In *T. miscellus*, loss and silencing of homeologs occurred in the early generations of allopolyploidy (Tate *et al.* 2006; Buggs *et al.* 2009; Tate *et al.* 2009a) based on analysis of only 20 homeolog pairs using PCR-based methods. New surveys are needed that will move us from a gene-by-gene approach to a genomic level. This requires a dramatic increase in the genomic resources available for plants that are good evolutionary models but not genetic models. We wished to develop a protocol that would produce a large number of homeolog-specific markers in *T. miscellus* at minimal time and expense, allowing us to assess homeologous gene loss and silencing.

Sequencing of cDNA or expressed sequence tags (EST) provides a rapid method for gene discovery and can be used to identify transcripts associated with specific biological processes. As such, it is often a first step in the genomic characterization of an organism. Variation in ESTs can be characterized by single nucleotide polymorphisms (SNPs), which are single-base differences between haplotypes. Transcript-associated SNPs can be used to develop allele-specific assays for the examination of *cis*-regulatory variation within a species (Guo *et al.* 2004; Stupar & Springer 2006) and may provide a rapid means to investigate differential expression and gene gain/loss within polyploids. EST collections and SNP discovery rely on DNA sequencing, which until recently was prohibitively costly for most evolutionary studies.

Recent advances in high-throughput sequencing technology provide rapid and cost-effective means to generate sequence data (Stupar & Springer 2006; Ellegren 2008; Hudson 2008). This new paradigm, termed flow-cell sequencing (reviewed in Holt & Jones 2008), consists of stepwise determination of DNA sequence by iterative cycles of nucleotide extensions done in parallel on huge numbers of clonally amplified template molecules. This massively parallel approach enables DNA sequence to be acquired at extremely high depths of coverage in less time and for less cost than traditional sequencing. The 454-FLX produces 200,000 sequences per run with ~200-300 bp lengths (100 Mb). With new Titanium reagents, this can be increased to over 1 million sequences with ~350-400 bp read lengths (400-600 Mb per run). In contrast, the Illumina Genome Analyzer (GA) II DNA sequencing instrument can produce >80 million sequences, each of which is 36bp in length (> 2 Gb). Short read lengths can confound assembly and alignment programs, but the reduction in read length vs. increased depth of coverage is an acceptable trade-off for many re-sequencing applications such as transcript expression profiling (Eveland *et al.* 2008), *in vivo* DNA binding site detection (Johnson *et al.* 2007) and polymorphism detection (Barbazuk *et al.* 2007; Novaes *et al.* 2008; Van Tassell *et al.* 2008). In the latter application, a high volume of short reads is very powerful in discriminating sequence variants, enabling reliable SNP discovery, so long as each read is long enough and accurate enough to align uniquely to the reference sequences.

To permit gene discovery and genomic tool development in species with few genomic resources, we designed a hybrid sequencing approach. In this approach, the Roche 454 sequencer is first used to generate transcriptome or genomic sequences that can be assembled and used as reference sequences (as in, e.g. Novaes *et al.* 2008). We then use this reference for subsequent alignment of Illumina short reads. This method gains maximum leverage from the longer read lengths of 454 sequencing and the deeper coverage of Illumina. Assembling 454 sequence reads is less problematic than Illumina reads, making it the high-throughput sequencing method of choice for species with few genomic resources, and it is particularly useful in transcriptome characterization (Cheung *et al.* 2006; Emrich *et al.* 2007; Cheung *et al.* 2008; Novaes *et al.* 2008). The 454 assemblies can therefore be used for gene annotation and the Illumina sequences used to identify SNPs and examine relative expression differences.

Once SNPs have been identified, a highly efficient way to validate them and carry out large-scale surveys of their frequencies is the Sequenom MassARRAY iPLEX genotyping platform (Gabriel *et al.* 2009). In this method, a short section of DNA containing a SNP is amplified from an individual by PCR. This is followed by a high-fidelity single-base primer extension reaction over the SNP being assayed, using nucleotides of modified mass. The different alleles therefore produce oligonucleotides with mass differences that can be detected using highly accurate Matrix-Assisted Laser Desorption/Ionization Time-

Of-Flight (MALDI-TOF) mass spectrometry. Up to 40 different SNPs can be multiplexed in one assay if primers are designed by custom software to give unique mass ranges for each SNP. This method is especially suited for detecting homeologs which differ in only a few SNPs as, unlike microarrays which rely on hybridization of oligonucleotides, it detects differences by single-nucleotide extension over SNPs.

In this paper we demonstrate the utility of hybrid next-generation sequencing and Sequenom genotyping for the study of homeolog evolution in *T. miscellus*. We report the transcriptome characterization of *T. dubius*, one of the diploid progenitors of *T. miscellus*, with 454 sequencing and the subsequent discovery of over 24,000 SNPs between *T. dubius* and the other parental diploid species, *T. pratensis*, using Illumina sequencing. We validated a subset of 98 SNPs that represent homeolog pairs in *T. miscellus* at the genomic level using Sequenom MassARRAY iPLEX genotyping. In addition, expression profiling of a *T. miscellus* individual using Illumina sequencing was performed. We assessed the utility of this profile for the selection of candidate genes for the investigation of loss from the genome. These methods could be applied to any organism, allowing efficient and cost-effective generation of genetic markers.

Materials and methods

Seeds were collected from natural populations of allotetraploid *T. miscellus* (Soltis and Soltis collection number 2671) and its diploid parent species, *T. dubius* (collection no.

2674) and *T. pratensis* (collection no. 2672), in Oakesdale, WA. The three species grow in sympatry in this location, and this fact, together with microsatellite data (Symonds *et al.* 2009), suggest that the diploid populations were the source of the progenitors of the allotetraploid population. These seeds were germinated and grown in an air-conditioned greenhouse with supplementary lighting at the University of Florida (Gainesville, FL, USA). *T. miscellus* from Oakesdale is the short-liguled form, with *T. pratensis* as the maternal parent (Soltis & Soltis 1989; Soltis *et al.* 1995).

RNA was extracted from leaf tissue of three individuals from Oakesdale: *T. dubius* 2674-4 (ID no. 3911), *T. pratensis* 2671-1 (ID no. 3912), and *T. miscellus* 2671-1 (ID no. 3912). Basal leaf tissue from each plant was flash frozen and ground in liquid nitrogen using a pestle and mortar. RNA extractions were performed following a portion of the CTAB DNA extraction protocol (Doyle & Doyle 1987) and subsequent use of the RNeasy Plant Mini Kit (Qiagen, Stanford, CA, USA) with on-column DNase digestion. This method was originally developed for the successful extraction of RNA from *Amborella* and *Nuphar* (Kim *et al.* 2004) and copes well with the latex produced by *Tragopogon* photosynthetic tissue. This was followed by an RNA cleanup using the protocol of the RNeasy Plant Mini Kit. These extractions were quality-checked using the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA).

454 EST sequencing and processing

Using the *T. dubius* RNA, a normalized cDNA library was produced via the following method. The Evrogen MINT cDNA synthesis kit (Evrogen, Moscow, Russia) was used to produce double-stranded cDNA following the manufacturer's protocol. This cDNA was cleaned using the Wizard® SV Gel and PCR Clean-Up System (Promega, Madison, WI, USA). The Evrogen TRIMMER cDNA normalization kit (Evrogen, Moscow, Russia) was used to normalize and amplify the cDNA library, following the manufacturer's instructions. In the normalization step, a 0.5 dilution of the duplex-specific nuclease (DSN) was found to be optimal. In the amplification step, 12 cycles were found to be optimal. The resulting normalized library was used for 454 sequencing.

454 sequencing was performed as described in the supplementary material and methods to Margulies *et al.* (2005) with slight modifications as specified by 454 Life Sciences. Briefly, cDNA was sheared by nebulization to a size range of 300 to 800 bp. DNA fragment ends were repaired and phosphorylated using T4 DNA polymerase and T4 polynucleotide kinase. Adaptor oligonucleotides "A" and "B" supplied with the 454 Life Sciences sequencing reagent kit were ligated to the DNA fragments using T4 DNA ligase. Purified DNA fragments were hybridized to DNA capture beads and clonally amplified by emulsion PCR (emPCR). DNA capture beads containing amplified DNA were deposited on a 70 x 75 mm PicoTiter plate, and DNA sequences determined using the GS-FLX instrument. This resulted in 822,594 EST sequences. The *T. dubius* 454 EST sequences were assembled with the Newbler assembler, a part of the software package distributed with 454 sequencing machines. Newbler is an assembler that takes into account the specifics of pyrosequencing errors to generate accurate contigs

(Chaisson & Pevzner 2008). Our assembly used the default directives and a vector trimming database including the Evrogen primer and 454 adapter sequences.

Comparisons of 454 ESTs to public sequence database (Annotation)

Assembled and annotated contig EST assemblies and singletons were obtained from the curated Gene Indices Project (Quackenbush *et al.* 2000); http://compbio.dfci.harvard.edu/tgi/) from three other species in the Asteraceae: *Lactuca sativa* (ver. 3.0), *Lactuca serriola* (ver. 1.0) and *Helianthus annus* (ver. 5.0). These sequences were pooled, formatted into a blastable database and aligned to the *T. dubius* 454 EST assemblies with WU-TBLASTX (ver 2.0), which translates both the query and subject sequences in all 6 potential reading frames prior to alignment, to identify the top hit for each *T. dubius* contig (pval <= 1e-05 and <= 1e-10). The *T. dubius* 454 EST contigs were also BLASTX-aligned to *Arabidopsis* CDS sequences (TAIR v. 8) because *Arabidopsis* represents the best curated plant genome available. Top hits for each *T. dubius* contig to the *Arabidopsis* protein set were identified (pval <= 1e-05 and <= 1e-10). Similarity search results are summarized in Table 1.

Illumina sequencing

The RNA extractions from *T. dubius* 2674-4 (ID no. 3911), *T. pratensis* 2672-5 (ID no. 3913), and *T. miscellus* 2671-1 (ID no. 3912) were used for Illumina sequencing. Poly A+ RNA was isolated from total RNA through two rounds of oligo-dT selection

(Dynabeads mRNA Purification Kit, Invitrogen Inc., Santa Clara, CA, USA). The mRNA was annealed to high concentrations of random hexamers and reverse transcribed. Following second strand synthesis, end repair, and A-tailing, adapters complementary to sequencing primers were ligated to cDNA fragments (mRNA-Seq Sample Prep Kit, Illumina, San Diego, CA, USA). Resultant cDNA libraries were size fractionated on agarose gels, and 250 bp fragments were excised and amplified by 15 cycles of polymerase chain reaction.   Resultant libraries were quality assessed using a Bioanalyzer 2100 and sequenced for 36 cycles on an Illumina GA II DNA sequencing instrument using standard procedures.

SNP discovery

All Illumina reads from the *T. dubius* and *T. pratensis* parents and the *T. miscellus* allotetraploid were labeled with species identifiers, pooled and aligned to the *T. dubius* 454 FLX contigs with the MosaikAligner package (Hillier *et al.* 2008) using the following MosaikAligner parameters: -a (alignment algorithm) all; -p (CPUs used) 8; -mm (maximum mismatch) 2 in a preliminary analysis and 1 in a final analysis; -m (alignment mode) unique; -hs (hash size) 15; -mhp (maximum number of hash positions to use) 100. These alignment parameters ensured that each Illumina sequence aligned to a unique position within the 454 *T. dubius* EST assembly reference sequences and with no more than one base-pair mismatch in the final analysis. Illumina reads that did not align with the 454 contigs under these stringent conditions were discarded from the analysis.

SNPs were identified within the alignments with the GigaBayes package (http://bioinformatics.bc.edu/marthlab/GigaBayes). GigaBayes is a reimplementation of the PolyBayes (Marth *et al.* 1999) SNP discovery tool that has been optimized for next-generation sequences. Arguments to GigaBayes were: --D (pairwise nucleotide diversity) 0.001; --ploidy (sample ploidy) diploid; --sample (sequence source) multiple; --anchor; --algorithm banded; --CAL (minimum overall allele coverage) 3; --QRL (minimum base quality value) 20. Custom PERL scripts were written to automate the SNP discovery process on all alignments to reference contigs and to parse the GigaBayes output files (GFF), which contain the site identification of each SNP, its representation within each of the three *Tragopogon* species (*T. dubius, T. pratensis* and *T. miscellus*) and its allele usage.

Any site where both the *T. pratensis* and *T. dubius* homeologs were evidenced in the *T. miscellus* data was flagged as a suitable SNP for the study of homeolog loss in *T. miscellus*. Where both homeologs were present in at least ten *T. miscellus* Illumina reads, and the observed allelic ratio was more balanced than 70:30 in either direction, we took this as preliminary evidence that both homeologs were equally expressed. In contrast, any site where either the *T. pratensis* or *T. dubius* parental homeolog was present at 10X while the other was absent, was identified as suggestive of either complete silencing of one parental homeolog or genomic homeolog loss.

SNP validation

A subset of SNPs identified using the above methods was analyzed using the Sequenom MassARRAY iPLEX platform at the Center for Plant Genomics, Iowa State University. Genomic DNA was extracted from leaf tissue of the three plants used for the transcriptome sequencing, using a modified CTAB protocol (Doyle & Doyle 1987). Multiplexed assays were designed using the Sequenom Assay Design 3.1 software for four plexes containing a total of 139 SNPs between *T. dubius* and *T. pratensis*. Of these, 42 were scored as “potential gene loss” using the Illumina read data, 77 were scored as “alleles balanced”, 19 were scored as “low coverage in *T. miscellus*” and one had no *T. miscellus* reads. This assay design was used to genotype a 384-well plate that included *T. dubius*, *T. pratensis* and *T. miscellus* genomic DNA samples (~ 20 ng/ul). The resulting data were analyzed using the MassARRAY Typer 4.0 Analyzer software. Using the manufacturer’s settings, the Sequenom software was used to call SNPs at “aggressive”, “moderate” and “conservative” degrees of confidence.

Results

454 Sequencing, assembly and annotation of *T. dubius* cDNA sequences

454 FLX sequencing of the normalized *T. dubius* cDNA pool from *T. dubius* leaf tissue produced 822,594 reads (237 bp av. length) representing >195 MB of sequence. These reads have been uploaded to the NCBI Short Read Archive (Accession Number

SRA009218.1). Assembly of the 454 FLX reads with the Roche 454 Newbler assembler produced 33,515 contigs (14.7 Mb) with an average length of 439 bp (min = 96, Max = 3418), an average depth of 17.6 reads and N50 Contig Size of 626 bp (see Figure 1).

In comparison with other species in the Asteraceae, 21,498 (64 %) of the *T. dubius* 454 EST sequences matched previously characterized EST assemblies (TBLASTX) from *Lactuca sativa, L. serriola* and *Helianthus annus* with p-values of e-5 or better. This low percentage may reflect the low depth and coverage in many of our 454 contigs (Figure 1) or significant divergence among the species. Of the 21,498 hits, 18,526 (86 %) were to unique EST assemblies in this curated database. The 14 % of non-unique contigs may be due to paralogous sequences in *T. dubius*, or to non-overlapping assemblies of *T. dubius* sequence from the same cDNA template, as the 'shotgun' nature of 454 sequencing enables simultaneous sampling of discrete template regions. The majority of best matches occurred between *T. dubius* and *L. sativa* (Table 1). In comparison with *A. thaliana*, 18,923 *T. dubius* 454 EST contig assemblies match *A. thaliana* CDS sequences (TBLASTX) at p-values of e-5 or better, while a total of 22,946 *T. dubius* contigs hit at least one sequence in either the *A. thaliana* or the Asteraceae collection.

SNP discovery

Non-normalized cDNA pools sequenced on single lanes of an Illumina GAII analyzer resulted in 7,128,226, 6,840,425 and 6,729,215 reads from *T. dubius*, *T. pratensis* and *T. miscellus*, respectively. These reads have been uploaded to the NCBI Short Read Archive

(Accession Number SRA009218.1). Alignment of pooled Illumina reads to the *T. dubius* 454 assembled EST reference sequences with a mismatch tolerance of 2 bp followed by identification of polymorphic sites that were represented to a minimum of three-fold redundancy in both *T. dubius* and *T. pratensis* revealed >45,000 potential SNPs within 10,428 contigs. To reduce the risk of misaligning repetitive or highly paralogous sequences, parameters were adjusted to permit only a single mismatch over the length of the Illumina reads. Of the total pooled *T. dubius*, *T. pratensis* and *T. miscellus* Illumina reads, 11,050,022 (53.4%) aligned. The remaining reads were unaligned because they did not map a unique location in the 454 contig reference sequence collection, or they did not meet the single mismatch criterion. This higher confidence alignment, when parsed for polymorphic sites that were represented to a minimum of three-fold redundancy in both *T. dubius* and *T. pratensis*, resulted in the identification of 24,078 potential SNPs between *T. dubius* and *T. pratensis* within 7,837 unique 454 EST contig reference sequences. To identify an even higher-quality collection of potential SNP sites between *T. dubius* and *T. pratensis*, the aforementioned alignments were parsed for SNP sites that were represented to a minimum depth of 10X in both the *T. dubius* and *T. pratensis* data sets. This high-quality collection that maximizes the likelihood that discovered polymorphic sites represent true SNPs between *T. dubius* and *T. pratensis* consists of 7,782 SNPs within 2,885 unique contigs.

Of the 7,782 SNPs, 2,989 had sufficient *T. miscellus* Illumina reads for transcriptome analysis. Of these, 2,064 (69 %) appeared to show equal homeolog expression in *T. miscellus*, 671 (22 %) showed differential expression in *T. miscellus*, and 254 (8.5 %)

showed potential homeolog loss in *T. miscellus*. Interestingly, the cases of differential expression were mainly due to higher expression of the *T. dubius* homeolog than of the *T. pratensis* homeolog (454/671; 77%) and the majority of the apparent losses were also of the *T. dubius* homeolog (164/254; 65%) rather than the *T. pratensis* homeolog (90/254).

SNP validation

Sequenom MassARRAY iPLEX assays were designed for 139 of the putative SNPs (four plexes). These assays were used to analyze the genomic DNA of the two diploid plants whose transcriptomes were used for 454 and Illumina sequencing. For 19 of the assays, the Sequenom assay failed to call a SNP in both diploid species, and 22 assays only worked in one of the diploid species. This failure rate is comparable to those obtained by other groups (Dunstan *et al.* 2007). Of the 98 informative assays (Table 2), 92 (94%) confirmed the SNP calls. In five of the remaining assays, the correct polymorphism was present but there was an extra allele in the genome of one diploid (i.e. heterozygosity) that had not been detected by via transcriptome sequencing. In only one case did the base call differ between the sequencing and Sequenom methods: here Sequenom indicated the same base in both alleles.

We then examined the Sequenom data for the genomic DNA of the *T. miscellus* plant. Of the 139 SNP assays, 41 did not successfully call any bases within our confidence limits in this plant. In 28 of these 41 cases, the assay also failed to call a SNP in one or both diploid species, but in the remaining 13 cases, the assay called a SNP in both diploid species but not in *T. miscellus* (see Table 2). In another 13 cases, one or more SNPs were called in *T. miscellus*, but a base had only been successfully called in one of the diploids (not shown in Table 2). Thus, in total, 85 of the 139 Sequenom assays (61 %) provided a call. In no cases did we find a SNP homeolog present in *T. miscellus* that had not been found in either *T. dubius* or *T. pratensis* at that locus.

Only the Sequenom data for 81 assays were used to infer homeolog loss in *T. miscellus*. Of the 85 assays that worked in all three plants, three were excluded due to heterozygosity in *T. dubius* and a fourth because of an identical call in both diploids. Of the 81 assays used, 47 gave evidence in *T. miscellus* of both *T. dubius* and *T. pratensis* SNP homeologs, and 34 gave evidence of only one SNP homeolog. Thus, 41% of the SNP loci give evidence for homeolog loss. Of these, 25 (74%) showed loss of the *T. dubius* homeolog, and nine (26%) showed loss of the *T. pratensis* homeolog. If we increase stringency by omitting “aggressive” calls (i.e. less confident Sequenom calls), we find that 69 assays gave a call; of these, 44 gave evidence of both SNP homeologs in *T. miscellus*, and 26 gave evidence of only one SNP homeolog.

We then compared the Sequenom and Illumina sequence data for the *T. miscellus* plant, to discover how often Illumina expression data had successfully identified a candidate for genomic loss (Table 3). Illumina read counts were correct in 89 % of the cases where there was depth of coverage above 10X per SNP homeolog, and the Sequenom calls were at conservative, moderate and aggressive levels of confidence (listed in descending order). Where Illumina read data had predicted “potential gene loss”, this was shown by

Sequenom analysis in 23 of 27 cases (85%). In four cases, homeologs were detected in the genomic DNA by Sequenom but not in the transcriptome by Illumina (i.e. they were scored as “potential gene loss”). This may be due to homeolog silencing. In contrast, four SNP homeologs were detected in the transcriptome by Illumina (i.e. they were scored “alleles balanced”) but were not found in the genome by Sequenom. Manual examination of the mass spectrometer traces for these calls suggested that three of them, which had all been called at the “aggressive” (lowest) level of confidence, did in fact have both homeologs present in the gDNA. In no cases did a contradiction occur where Illumina showed no expression of one homeolog and Sequenom loss of the other homeolog.

Discussion

Genomic resources are scarce for many organisms that are studied in a natural ecological or evolutionary context (Ellegren 2008; Hudson 2008). Here, we demonstrate a protocol that uses next-generation technologies to rapidly develop SNP markers in many hundreds of genes in a species which is a good evolutionary model but which until now has not been a genetic model organism. These SNP markers have many potential uses. We have used them to distinguish between homeologous genes in the recent natural allopolyploid *T. miscellus*. Using transcriptome profiling and Sequenom genotyping, we have detected many cases of gene loss. Below we discuss the biological implications of our findings in *T. miscellus*, and the general utility of the methods described in this paper.

Biological implications of findings in *T. miscellus*

This paper provides the first large-scale analysis of homeologous gene loss in a recent (~40-generation-old) natural allopolyploid. In a single *T. miscellus* individual we found 254 cases of putative homeolog loss or silencing by transcriptome profiling with Illumina sequencing (3% of all SNPs). Sequenom analysis confirmed that in a sample of 27 of these SNPs, 23 (85%) were cases of genomic homeolog loss. The remaining 15% are likely to be homeologs that are present in the genome but were not being expressed at the time of sampling in the leaf tissue subject to transcriptome analysis. Homeolog loss therefore appears to be more common than homeolog silencing (i.e. lack of expression of a gene found in the genome) in this species.

We found preferential loss of *T. dubius* homeologs over *T. pratensis* homeologs in the allopolyploid *T. miscellus* in this study. Illumina read data on the transcriptome suggested loss or silencing of the *T. dubius* homeolog in 164 of 254 SNPs (64%) showing homeolog loss or silencing, and Sequenom analysis of the genome suggested loss of the *T. dubius* homeolog in 25 of 34 SNPs (73%) showing homeolog loss. In earlier studies a similar bias was found: combined results from Buggs *et al*. (2009), Tate *et al*. (2006), and Tate *et al*. (2009a) gave 56 *T. dubius* homeolog losses, and 27 *T. pratensis* homeolog losses across multiple populations. Interestingly, we also found a bias in gene expression in our Illumina read data, with *T. dubius* homeologs tending to be expressed more than *T. pratensis* homeologs in 77% of the SNPs where we detected differential expression.

Because *T. dubius* ESTs were used as the reference sequence we might expect a bias towards the alignment of Illumina reads derived from *T. dubius* homeologs. This possible bias may have contributed to the apparent higher expression of the *T. dubius* homeolog at many loci, but also suggests that the finding of a higher rate of loss of *T. dubius* homeologs is a robust result.

It is notable that a similar bias towards loss of *T. dubius* genetic material and higher expression of *T. dubius* genes has been found for rDNA in both *T. miscellus* and *T. mirus*, an allopolyploid that has *T. dubius* as the paternal parent and *T. porrifolius* as the maternal parent (Kovarik *et al.* 2005), In both species, concerted evolution has reduced the copy numbers of rDNA units derived mainly from the *T. dubius* diploid parent but, paradoxically, repeats of *T. dubius* origin dominate transcription in most populations studied (Matyasek *et al.* 2007). *Tragopogon mirus* also shows a bias toward loss of *T. dubius* homeologs using CAPS markers (Koh *et al*., submitted)

What causes the bias towards higher rates of gene loss and increased expression of *T. dubius* homeologs? One possibility might be maternal effects due to cytoplasmic-nuclear interactions. The *T. miscellus* plant in the current study, as well as all *T. mirus* plants and the majority of *T. miscellus* plants included in other studies, has *T. dubius* as the paternal parent. Perhaps selection favors maintaining ancestral similarity in the cytoplasmic and nuclear genomes. Another explanation might be the higher genetic variability of *T. dubius* populations (Soltis *et al.* 1995); it is possible that the *T. dubius* individual that we examined from Oakesdale was not genetically identical to the actual *T. dubius* progenitor of *T. miscellus* from Oakesdale. However, it seems unlikely that the bias is due to the selection of an inappropriate *T. dubius* genotype in this study as the other studies cited above as showing the same pattern have examined multiple *T. dubius* individuals. Our results also agree with those found in other species. In synthetic allopolyploids of *Brassica*, genomic changes occur more often in the paternal genome (Song *et al.* 1995). In natural *Gossypium hirsutum* (Flagel *et al.* 2008) and synthetic *Arabidopsis* allopolyploids (Wang *et al.* 2006), homeolog expression biases also tend to be in favor of the paternal genome. In maize, it has recently been shown that paternal genomic imprinting influences gene expression patterns in hybrids (Swanson-Wagner *et al.* 2009).

One mechanism by which homeolog loss may occur in *T. miscellus* is homeologous recombination, in which fragments of chromosomes can be lost. Ownbey (1950) observed multivalent formation in early generations of natural *T. miscellus*, and rare patterns of isozyme variation in *T. miscellus* are consistent with homeologous recombination (Soltis *et al.* 1995). More recently, Lim *et al*. (2008) and Tate *et al.* (2009b) report multivalent formation in both natural and synthetic *Tragopogon* allopolyploids, along with unisomy, trisomy, and reciprocal translocations in natural *Tragopogon* allopolyploids. Homoeologous recombination appears to have caused loss of chromosome fragments in re-synthesised *Brassica* allopolyploids (Song *et al.* 1995; Gaeta *et al.* 2007). Another possible mechanism of homeolog loss is gene conversion, as has been found for rRNA genes in both *T. miscellus* and *T. mirus* (Kovarik *et al.* 2005; Matyasek *et al.* 2007).

High-throughput SNP discovery together with the genotyping of many natural *T. miscellus* plants of independent origin and F_1 hybrids will enable us to examine genome-wide patterns of homeolog loss in this species. As SNPs are abundant in many species and easily detected (Gut 2001; Kwok 2001), they are excellent genetic markers for the generation of dense genetic maps that can support marker-assisted selection (MAS) and association genetics programs, as well as inform on genome organization and function (Pavy *et al.* 2008; Slate *et al.* 2009). In *T. miscellus*, application of these markers will enable us to understand further the causes of homeolog loss in this allopolyploid, showing us whether or not homeolog losses occur in linkage groups – implying the loss of large fragments of chromosomes – or in small fragments scattered throughout the genome.

Utility of methods

In the space of a few months, we have been able to identify at high stringency 7,782 homeolog-specific SNP markers within 2,885 unique contigs in *T. miscellus* using next-generation sequencing. The number of homeologous genes available for study has therefore been increased by two orders of magnitude compared to previous studies using a "one gene at a time" approach (Tate *et al.* 2006; Buggs *et al.* 2009; Tate *et al.* 2009a). The number of actual SNPs discovered is likely to be much higher than this, as we were likely over-stringent. We have developed working assays for 85 of these SNPs using Sequenom MassARRAY iPLEX technology. This high-throughput approach transforms our ability to study molecular evolution in *T. miscellus*.

The use of transcriptome sequencing with polyA purification is valuable for targeting functional genes for SNP discovery, as clearly shown by this study. However, there is the possibility that when these markers are then used to study the genome, polymorphisms will be discovered due to the presence of silent homeologs. In a few cases we found this: six of 139 Sequenom SNP assays found polymorphisms in genomic DNA of diploid plants that had not been detected by Illumina sequencing in the transcriptome. This was an acceptably low level of polymorphism that was undiscovered by transcriptome sequencing. However, it should be noted that *T. dubius* and *T. pratensis* are mostly selfing species (Cook & Soltis 1999; Cook & Soltis 2000) with limited polymorphism in their introduced ranges in North America (Soltis *et al.* 1995; Symonds *et al.* 2009). Outcrossing species with high heterozygosity may pose more difficulties in analysis.

Sequenom MassARRAY Typer 4.0 Analyzer software uses a three-parameter model to calculate the significance of each putative genotype. This compares the size of peaks for the possible bases at each SNP site and the peak for the unextended primer. Where an assay is not working well, the non-extended primer will be found in greater abundance than the extended oligonucleotides. For genotypes which are called, the degree of confidence that can be placed on the call is described as “conservative”, “moderate”, or “aggressive” in the software output. We found that four calls (three called at the “aggressive” (lowest) level of confidence and one at the “moderate” level) were not reliable due to failure to detect a base that was in fact present (a false negative). Manual examination of the mass-spectrometer trace in most cases allowed the call to be corrected.

This "false negative" problem is likely to be due to the malfunction of these specific assays, rather than the reliability of "aggressive" calls in general. Certain assays can function well in calling different bases in homozygotes, but in a heterozygote the primers bind preferentially to one allele, resulting in a false homozygote call. One reason why this occurs is if there is another SNP close to the SNP site that is being assayed (Liu *et al.* 2009). Preferential binding of primers can be assessed by genotyping more individuals that are expected to be heterozygous. If they all appear to be homozygous, then the Sequenom assay for that SNP should be rejected. We did this (see below) and found that these assays did not work correctly in multiple individuals. In addition, if we discard all aggressive Sequenom calls, we find that the correspondence between the Illumina and Sequenom data rises only slightly from 89% to 93%. This also suggests that there is not a general problem with the reliability of "aggressive" calls.

This study also demonstrates that transcriptome profiling using Illumina sequencing is a useful method for identifying candidate homeologs for the study of homeolog loss in an allopolyploid species. This allows us to target these genes for developing SNP-typing assays, saving both time and money. The major cost in using Sequenom genotyping is the production of primers. Each SNP requires three primers: two for an initial amplification of the target region and one for the SNP-typing reaction. Once these primers have been synthesized, many samples can be SNP-typed at relatively low cost. We made use of this fact by screening an additional 94 individuals: a total of 87 diploid and *T. miscellus* plants from five natural populations, two 50-year-old herbarium specimens and five artificial crosses. Preliminary analyses of this survey allowed us to identify polymorphisms in the diploid plants and calculate allelic diversity. This data set showed repeatability of some homeolog losses in natural *T. miscellus* populations of different origins. Finally, we also found the first evidence for rare loss of alleles in F_1 hybrids between *T. dubius* and *T. pratensis*. Robust analysis of this data set is ongoing.

Broader applicability

Transcriptome sequencing by 454 has many potential applications in ecology (Ellegren 2008; Wang *et al.* 2009; Wheat 2008). It has been used for the *de novo* characterization of the transcriptome of the Glanville fritillary butterfly (Vera *et al.* 2008) and the *Eucalyptus grandis* genome (Novaes *et al.* 2008). Recent work in model organisms has used short-read sequencing to study differences in expression of SNP-containing alleles, for example in micro-RNAs in mice (Kim & Bartel 2009). Sequenom MassARRAY genotyping has been used to study allelic expression in hybrid maize (Stupar & Springer 2006) and levels of homeolog expression in allopolyploid cotton (Flagel *et al.* 2008; Chaudhary *et al.* 2009; Flagel *et al.* 2009). This study demonstrates the effectiveness of a hybrid Illumina and 454 sequencing approach and Sequenom MassARRAY iPLEX genotyping to increase dramatically our ability to study the evolution of duplicated genes in natural allopolyploids such as *T. miscellus.* These methods could be applied to any organism, allowing efficient and cost-effective generation of SNP markers.

Acknowledgments

We thank William Farmerie and Regina Shaw of the Interdisciplinary Center for Biotechnology Research at the University of Florida for 454 sequencing, and Michael Sandford for technical support. This work was supported by the University of Florida, and NSF grants MCB-034637, DEB-0614421, DEB-0919254, and DEB-0919348.

References

Adams KL, Cronn R, Percifield R, Wendel JF (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 4649-4654.

Adams KL, Wendel JF (2004) Exploring the genomic mysteries of polyploidy in cotton. *Biological Journal of the Linnean Society*, **82**, 573-581.

Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology*, **8**, 135-141.

Ainouche ML, Baumel A, Salmon A (2004) *Spartina anglica* C. E. Hubbard: a natural model system for analysing early evolutionary changes that affect allopolyploid genomes. *Biological Journal of the Linnean Society*, **82**, 475-484.

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. *Plant Journal*, **51**, 910-918.

Bottley A, Xia GM, Koebner RMD (2006) Homoeologous gene silencing in hexaploid wheat. *Plant Journal*, **47**, 897-906.

Buggs RJA (2008) Towards natural polyploid model organisms. *Molecular Ecology*, **17**, 1875-1876.

Buggs RJA, Doust AN, Tate JA*, et al.* (2009) Gene loss and silencing in *Tragopogon miscellus* (Asteraceae): comparison of natural and synthetic allotetraploids. *Heredity*, **103**, 73-81.

Buggs RJA, Pannell JR (2007) Ecological differentiation and diploid superiority across a moving ploidy contact zone. *Evolution*, **61**, 125 - 140.

Chaisson MJ, Pevzner PA (2008) Short read fragment assembly of bacterial genomes. *Genome Research*, **18**, 324-330.

Chaudhary B, Flagel L, Stupar RM*, et al.* (2009) Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (*Gossypium*). *Genetics*, **182**, 503-517.

Chen M, Ha M, Lackey E, Wang JL, Chen ZJ (2008) RNAi of met1 reduces DNA methylation and induces genome-specific changes in gene expression and centromeric small RNA accumulation in *Arabidopsis* allopolyploids. *Genetics*, **178**, 1845-1858.

Chen ZJ, Wang J, Tian L*, et al.* (2004) The development of an *Arabidopsis* model system for genome-wide analysis of polyploidy effects. *Biological Journal of the Linnean Society*, **82**, 689-700.

Cheung F, Haas B, Goldberg S*, et al.* (2006) Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. *BMC Genomics*, **7**, 272.

Cheung F, Win J, Lang J*, et al.* (2008) Analysis of the *Pythium ultimum* transcriptome using Sanger and Pyrosequencing approaches. *BMC Genomics*, **9**, 542.

Cook LM, Soltis PS (1999) Mating systems of diploid and allotetraploid populations of *Tragopogon* (Asteraceae). I. Natural populations. *Heredity*, **82**, 237-244.

Cook LM, Soltis PS (2000) Mating systems of diploid and allotetraploid populations of Tragopogon (Asteraceae). II. Artificial populations. *Heredity*, **84**, 410-415.

Dong Y, Liu Z, Shan X*, et al.* (2005) Allopolyploidy in wheat induces rapid and heritable alterations in DNA methylation patterns of cellular genes and mobile elements. *Russian Journal of Genetics*, **41**, 890-896.

Doyle J, Doyle JL (1987) Genomic plant DNA preparation from fresh tissue-CTAB method. *Phytochemical Bulletin*, **19**, 11-15.

Dunstan S, Hue N, Rockett K*, et al.* (2007) A TNF region haplotype offers protection from typhoid fever in Vietnamese patients. *Human Genetics*, **122**, 51-61.

Ellegren H (2008) Sequencing goes 454 and takes large-scale genomics into the wild. *Molecular Ecology*, **17**, 1629-1631.

Emrich SJ, Barbazuk WB, Li L, Schnable PS (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research*, **17**, 69-73.

Eveland AL, McCarty DR, Koch KE (2008) Transcript profiling by 3'-untranslated region sequencing resolves expression of gene families. *Plant Physiology*, **146**, 32-44.

Feldman M, Liu B, Segal G*, et al.* (1997) Rapid elimination of low-copy DNA sequences in polyploid wheat: A possible mechanism for differentiation of homoeologous chromosomes. *Genetics*, **147**, 1381-1387.

Flagel LE, Chen L, Chaudhary B, Wendel JF (2009) Coordinated and fine-scale control of homoeologous gene expression in allotetraploid cotton. *Journal of Heredity*, **100**, 487-490.

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

Gabriel S, Ziaugra L, Tabbaa D (2009) SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Current Protocols in Human Genetics*, **60**, 2.12.11-12.12.18.

Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC (2007) Genomic changes in resynthesized *Brassica napus* and their effect on gene expression and phenotype. *Plant Cell*, **19**, 3403-3417

Guo M, Rupe MA, Zinselmeier C*, et al.* (2004) Allelic variation of gene expression in maize hybrids. *Plant Cell*, **16**, 1707-1716.

Gut IG (2001) Automation in genotyping of single nucleotide polymorphisms. *Human Mutation*, **17**, 475-492.

Hegarty M, Barker G, Wilson I*, et al.* (2006) Transcriptome shock after interspecific hybridization in *Senecio* is ameliorated by genome duplication. *Current Biology*, **16**, 1652-1659.

Hegarty M, Jones J, Wilson I*, et al.* (2005) Development of anonymous cDNA microarrays to study changes to the *Senecio* floral transcriptome during hybrid speciation. *Molecular Ecology*, **14**, 2493-2510.

Hillier LW, Marth GT, Quinlan AR*, et al.* (2008) Whole-genome sequencing and variant discovery in *C. elegans*. *Nature Methods*, **5**, 183-188.

Holt RA, Jones SJM (2008) The new paradigm of flow cell sequencing. *Genome Research*, **18**, 839-846.

Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, **8**, 3-17.

Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497-1502.

Joly S, Rauscher JT, Sherman-Broyles SL, Brown AHD, Doyle JJ (2004) Evolutionary dynamics and preferential expression of homeologous 18S-5.8S-26S nuclear ribosomal genes in natural and artificial Glycine allopolyploids. *Molecular Biology and Evolution*, **21**, 1409-1421.

Judd WS, Campbell CS, Kellogg EA, Stevens PF, Donoghue MJ (2007) *Plant Systematics: A Phylogenetic Approach*, Third Edition edn. Sinauer, Sunderland, MA.

Kashkush K, Feldman M, Levy AA (2002) Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics*, **160**, 1651-1659.

Kihara H, Ono T (1927) Chromosomenzahlen und systematische Gruppierung der *Rumex*-Arten. *Zeitschrift für Zellforschung und mikroskopische Anatomie*, **4**, 475-481.

Kim J, Bartel DP (2009) Allelic imbalance sequencing reveals that single-nucleotide polymorphisms frequently alter microRNA-directed repression. *Nature Biotechnology*, **27**, 472-477.

Kim S, Yoo M-J, Albert VA*, et al.* (2004) Phylogeny and diversification of B-function MADS-box genes in angiosperms: evolutionary and functional implications of a 260-million-year-old duplication. *American Journal of Botany*, **91**, 2102-2118.

Kovarik A, Pires JC, Leitch AR*, et al.* (2005) Rapid concerted evolution of nuclear ribosomal DNA in two *Tragopogon* allopolyploids of recent and recurrent origin. *Genetics*, **169**, 931-944.

Kwok PY (2001) Methods for genotyping single nucleotide polymorphisms. *Annual Review of Genomics and Human Genetics*, **2**, 235-258.

Levy AA, Feldman M (2004) Genetic and epigenetic reprogramming of the wheat genome upon allopolyploidization. *Biological Journal of the Linnean Society*, **82**, 607-613.

Lim KY, Matyasek R, Kovarik A, Leitch AR (2004) Genome evolution in allotetraploid *Nicotiana*. *Biological Journal of the Linnean Society*, **82**, 599-606.

Lim KY, Soltis DE, Soltis PS*, et al.* (2008) Rapid chromosome evolution in recently formed polyploids in *Tragopogon* (Asteraceae). *PLoS ONE*, **3**, e3353.

Liu S, Chen D, Makarevitch I*, et al.* (2009) High-throughput genetic mapping of mutants via quantitative SNP-typing. *Genetics*, **submitted**.

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151-1155.

Marth GT, Korf I, Yandell MD*, et al.* (1999) A general approach to single-nucleotide polymorphism discovery. *Nat Genet*, **23**, 452-456.

Matyasek R, Tate JA, Lim YK*, et al.* (2007) Concerted evolution of rDNA in recently formed *Tragopogon* allotetraploids is typically associated with an inverse correlation between gene copy number and expression. *Genetics*, **176**, 2509-2519.

Mavrodiev EV, Tancig M, Sherwood AM*, et al.* (2005) Phylogeny of *Tragopogon* L. (Asteraceae) based on internal and external transcribed spacer sequence data. *International Journal of Plant Sciences*, **166**, 117-133.

Novaes E, Drost D, Farmerie W*, et al.* (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics*, **9**, 312.

Novak SJ, Soltis DE, Soltis PS (1991) Ownbey *Tragopogons* - 40 Years Later. *American Journal of Botany*, **78**, 1586-1600.

Ownbey M (1950) Natural hybridization and amphiploidy in the genus *Tragopogon*. *American Journal of Botany*, **37**, 487-499.

Pavy N, Pelgas B, Beauseigle S*, et al.* (2008) Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: application to the large and unsequenced genomes of white spruce and black spruce. *BMC Genomics*, **9**, 21.

Petit M, Lim K, Julio E*, et al.* (2007) Differential impact of retrotransposon populations on the genome of allotetraploid tobacco (*Nicotiana tabacum*). *Molecular Genetics and Genomics*, **278**, 1-15.

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

Sadava DE, Heller HC, Orians GH, Purves WK, Hillis DM (2008) *Life: the science of biology*, Eighth edn. Macmillan, New York.

Slate J, Gratten J, Beraldi D, *et al.* (2009) Gene mapping in the wild with SNPs: guidelines and future directions. *Genetica*, **136**, 97-107.

Soltis DE, Soltis PS (1989) Allopolyploid speciation in *Tragopogon*: insights from chloroplast DNA. *American Journal of Botany*, **76**, 14-18.

Soltis DE, Soltis PS (1999) Polyploidy: recurrent formation and genome evolution. *Trends in Ecology & Evolution*, **14**, 348-352.

Soltis DE, Soltis PS, Pires JC, *et al.* (2004a) Recent and recurrent polyploidy in *Tragopogon* (Asteraceae): cytogenetic, genomic and genetic comparisons. *Biological Journal of the Linnean Society*, **82**, 485-501.

Soltis DE, Soltis PS, Tate JA (2004b) Advances in the study of polyploidy since Plant Speciation. *New Phytologist*, **161**, 173-191.

Soltis PS, Plunkett GM, Novak SJ, Soltis DE (1995) Genetic variation in *Tragopogon* species: additional origins of the allotetraploids *T. mirus* and *T. miscellus* (Compositae). *American Journal of Botany*, **82**, 1329-1341.

Song K, Lu P, Tang K, Osborn TC (1995) Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 7719-7723.

Stebbins GL (1942) Polyploid Complexes in Relation to Ecology and the History of Floras. *American Naturalist*, **76**, 36-45.

Stupar RM, Springer NM (2006) *Cis*-transcriptional variation in maize inbred lines B73 and Mo17 leads to additive expression patterns in the F-1 hybrid. *Genetics*, **173**, 2199-2210.

Swanson-Wagner R, DeCook R, Jia Y*, et al.* (2009) Widespread paternal genomic imprinting of trans-eQTL influences gene expression patterns in maize hybrids. *Science*, **submitted**

Symonds VV, Soltis PS, Soltis DE (2009) The dynamics of polyploid formation in *Tragopogon* (Asteraceae): recurrent formation, gene flow, and population structure. *Evolution*, **submitted**.

Tate J, Joshi P, Soltis K, Soltis P, Soltis D (2009a) On the road to diploidization? Homoeolog loss in independently formed populations of the allopolyploid *Tragopogon miscellus* (Asteraceae). *BMC Plant Biology*, **9**, 80.

Tate JA, Ni ZF, Scheen AC*, et al.* (2006) Evolution and expression of homeologous loci in *Tragopogon miscellus* (Asteraceae), a recent and reciprocally formed allopolyploid. *Genetics*, **173**, 1599-1611.

Tate JA, Symonds VV, Doust AN*, et al.* (2009b) Synthetic polyploids of *Tragopogon miscellus* and *T. mirus* (Asteraceae): 60 Years after Ownbey's discovery. *Am. J. Bot.*, **96**, 979-988.

Udall JA, Wendel JF (2006) Polyploidy and crop improvement. *Crop Science*, **46**, S3-S14.

Van Tassell CP, Smith TPL, Matukumalli LK*, et al.* (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, **5**, 247-252.

Vera JC, Wheat CW, Fescemyer HW*, et al.* (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636-1647.

Wang J, Tian L, Lee H-S*, et al.* (2006) Genomewide nonadditive gene regulation in *Arabidopsis* allotetraploids. *Genetics*, **172**, 507-517.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**, 57-63.

Wheat C (2008) Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. *Genetica*, **Online First 18 October 2008**

Table 1: Results of *Tragopogon dubius* similarity searches (BLASTX).

Sequence collection used for similarity searches	**Number of 454 contigs with similarity @ 1e-05**	**Number of Annotation sequences hit @ 1e-05**	**Number of 454 contigs with similarity @ 1e-10**	**Number of Annotation sequences hit @ 1e-10**
Lactuca sativa, Lactuca serriola and Helianthus annus Gene Index	21,498 *(Lactuca sativa*: 11,080 *Lactuca serriola*: 6078 *Helianthus annus*: 4340)	16,611	18,526 (*Lactuca sativa*: 9,731 *Lactuca serriola*: 5264 *Helianthus annus*: 3531)	14,914
Arabidopsis annotated peptides	18,923	11,086	16,412	10,180