2014-12-01

Targeted Sequencing of Plant Genomes

Mark D. Huynh
Brigham Young University - Provo

Targeted Sequencing of Plant Genomes

Mark D. Huynh

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

Master of Science

Joshua A. Udall, Chair
Bryce A. Richardson
Peter J. Maughan

Department of Plant and Wildlife Sciences

Brigham Young University

December 2014

ABSTRACT

Targeted Sequencing of Plant Genomes

Mark D. Huynh
Department of Plant and Wildlife Sciences, BYU
Master of Science in Genetics and Biotechnology

Next-generation sequencing (NGS) has revolutionized the field of genetics by providing a means for fast and relatively affordable sequencing. With the advancement of NGS, whole-genome sequencing (WGS) has become more commonplace. However, sequencing an entire genome is still not cost effective or even beneficial in all cases. In studies that do not require a whole-genome survey, WGS yields lower sequencing depth and sequencing of uninformative loci. Targeted sequencing utilizes the speed and low cost of NGS while providing deeper coverage for desired loci. This thesis applies targeted sequencing to the genomes of two different, non-model plants, *Artemisia tridentate* (sagebrush) and *Lupinus luteus* (yellow lupine). We first targeted the transcriptomes of three species of sagebrush (*Artemisia*) using RNA-seq. By targeting the transcriptome of sagebrush we have built a resource of transcripts previously unmatched in sagebrush and identify transcripts related to terpenes. Terpenes are of growing interest in sagebrush because of their ability to identify certain species of sagebrush and because they play a role in the feeding habits of the threatened sage-grouse. Lastly, using paralogs with synonymous mutations we reconstructed an evolutionary time line of ancient genome duplications. Second, we targeted the flanking loci of recognition sites of two endorestriction enzymes in genome *of L. luteus* genome through genotyping-by-sequencing (GBS). GBS of yellow lupine provided enough single-nucleotide polymorphic loci for the construction of a genetic map of yellow lupine. Additionally we compare GBS strategies for plant species without a reference genome sequence.

Keywords: genotyping-by-sequencing, lupine, plant genomes, sequencing, sagebrush, transcriptome, terpenes

ACKNOWLEDGMENTS

This thesis represents a portion of the work that I was steeped in for two years of intense growth. None of which would have progressed without the care of my committee, friends and family. Thank you collectively for stretching my abilities beyond their previous limitations. Individually I would like to thank Dr. Joshua Udall for his overly-generous support and enduring patience. By his guidance I grew to be a more independent thinker and was challenged to innovate. I am grateful for his concern that I grow both academically and personally. Dr. Bryce Richardson provided me with opportunities to learn genetics, but also ecology. He taught me the necessity of being a steward for both the environment and the creatures and plants of this planet. Dr. Jeff Maughan, through his example as an excellent teacher helped me to improve my ability to learn and teach others. He encouraged me to always keep the big picture in mind even when the world of academia seemed small. Thank you Dr. Dan Fairbanks for first believing in me. Lastly, I cite the book of Proverbs chapter three verse six.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SUPPLEMENTAL DATA

CHAPTER 1

Sequencing Three Transcriptomes of Sagebrush

INTRODUCTION

The sagebrushes (*Artemisia* subgenus *Tridentatae*) are pivotal members and the most abundant and widespread vegetation of the semi-arid ecosystems of western North America. Sagebrush ecosystems cover vast areas of the western United States and Canada [36]. This study focuses on two species of subgenus *Tridentatae:* big sagebrush (*Artemisia tridentata)* and low sagebrush (*A. arbuscula* ssp. *arbuscula)*. *A. tridentata* occupies about 43 million ha of the United States and includes three major subspecies: *A. tridentata* ssp. *tridentata* and *A. tridentata* ssp. *vaseyana* exist as both diploids and tetraploids, while *A. tridentata* ssp. *wyomingensis* is exclusively tetraploid. In comparison, *A. arbuscula* occupies about 28 million ha of the United States with four described subspecies, including diploid, tetraploid and occasionally hexaploid cytotypes [5], [28]. The two species typically have parapatric occurrences, especially between *A. arbuscula* and *A. tridentata* ssp. *vaseyana.* The former species occupies ridgelines and uplands with shallow soils, whereas the later typically occupies deeper soils [26], [35].

The sagebrush ecosystems are habitat and forage for numerous sagebrush-dependent wildlife species. Most notably is the Greater sage-grouse (*Centrocercus urophasianus*), which is of concern due to a declining habitat and shrinking breeding populations. Since being listed in 2010 by the U.S. Fish and Wildlife Service as a candidate for the endangered species list, it is currently the subject of one of the largest conservation efforts in North America [12]. Sage-grouse eat sagebrush leaves exclusively in winter months and they remain a primary food source throughout the rest of the year. For sage-grouse, habitat selection and forage of sagebrush is

guided impart by avoidance of plants with higher concentrations of monoterpenes [15]. Sage-grouse may be especially sensitive to terpenes when selecting a food source because they lack the ability to process out and metabolize excess terpenes through mastication and a ruminant digestive system like sheep and cattle [33]. A greater understanding of the chemical components that affect sagebrush palatability is a critical goal for sage-grouse conservation. For example, when available, sage-grouse prefer *A. nova* or *A. tridentata* ssp. *wyomingensis* rather than other species and subspecies of sagebrush, a preference that is believed to be correlated with decreasing leaf terpene concentration [35], [40]. This discrimination appears to be deeper than selecting solely between species. Indeed, Frye et al. [15] demonstrated that feeding selection within a conspecific patch of sagebrush is specific to plants with lower monoterpenes, regardless of species. Because conservation will largely be based on restoration at the ecosystems level, a finely tailored effort is needed that considers both the types of terpenes produced and their expression profiles among and within species.

Terpenoids have also been shown to be important in inter- and intraspecies plant communication [22]. Plant volatiles, including terpenes, released from the leaves of injured sagebrush plants function in priming the defense of the surrounding plant community [23]. As a result of this functional versatility, a large amount of research has been geared towards the isolation and classification of terpenes produced by the genus *Artemisia* (see review by Turi et al. [42]), however, the identification of genes and alleles involved in terpene biosynthesis and differences among sagebrush species or populations has not been reported.

While much work has been done in characterizing sagebrush based on taxonomic characters and cytology, little has been done to describe their transcriptomes. An NCBI search for *Artemisia* nucleotide sequences returns 26 sequences for *A. arbuscula* and less than 600

sequences for *A. tridentata*. The only transcriptome study of sagebrush was reported by Bajgain et al. [3], where they identified single nucleotide polymorphism (SNP) data from transcript amplicons of three big sagebrush subspecies in an attempt to elucidate complex polyploid and hybrid relationships. However, the combined Illumina and 454 sequencing technologies used in the study may not have fully sampled the transcriptome. Here we attempt to more fully sample the transcriptome with deeper sequencing and by including more than one species of sagebrush. This data not only provides the basis for elucidating specific biosynthetic pathways, but also enables the study of gene duplication. Gene duplication drives plant evolution by creating duplicate genes that can mutate to acquire specialized or completely novel functions as well contribute to dosage effects [30], [34]. In addition to single gene duplications, whole-genome duplications (WGD) may occur. WGD are thought to drive evolution by creating a larger background for mutation that, in some ecological circumstances, may lead to a greater survivability of polyploids. These single gene duplications and WGD can be detected by the proxy use of synonymous mutations [7], [44]. The chronology of these duplications provides inferences about the origin of a particular species and divergent taxa.

In this paper we present the assembly of three transcriptomes representing two species of subgenus *Tridentatae*. We utilize the transcriptomes to identify and analyze a putative ortholog of a terpene synthase (TS) present in both *A. tridentata* and *A. arbuscula* and for detecting ancient duplication events using synonymous mutation rates between paralogs. These transcriptomes will undoubtedly be useful for further elucidating the complex evolutionary history of sagebrush through transcript identification and SNP detection. They may also serve as reference transcriptomes for subsequent transcriptome analyses within the genus and for gene expression analyses (RNA-seq experiments).

MATERIALS AND METHODS

RNA Sequencing

Five half-sib seedlings from each *A. tridentata* ssp. *wyomingensis* (UTW1, 38.3279 N, 109.4352 W) *A. tridentata* ssp. *tridentata* (UTT2, 38.3060 N, 109.3876 W) and *A. arbuscula* ssp. *arbuscula* (CAV-1, 40.5049 N, 120.5617 W) were grown in a petri dish on top of wetted filter paper for two days. No specific permissions were required for these locations and none of the species are endangered or protected. Seedlings were then flash frozen in liquid nitrogen and ground using a mortar and pestle. RNA was extracted using a Norgen RNA Purification Kit (Norgen Biotek Corp., Ontario, Canada). Sequencing libraries were prepared using an Illumina Tru-seq RNA Kit V2 (Illumina Inc., San Diego, California). Libraries were then pooled and multiplexed on an Illumina MiSeq lane and sequenced as 250 bp paired-end reads at the Center for Genome Research and Biocomputing, Oregon State University.

Transcriptome Assembly

Illumina reads were trimmed for quality using default settings in the program Sickle (github.com/najoshi/sickle). Reads were then assembled using the program SOAPdenovo-trans [26] at variable k-mer lengths ranging from 35 to 127 in increments of 4. The best assembly for each set was based on N50 and the number of scaffolds. Other modified parameters included the number of scaffolds > 800 base pairs (bp) and the number of bp in scaffolds > 800 bp.

Transcript Characterization

Assembled transcriptomes were uploaded to the program TRAPID [6] where transcripts could be identified by protein domains related to terpene synthases (IPR005630). Transcripts were also blasted on NCBI using blastx [1] with the NR database for putative orthologs. To

compare the different sagebrush groups, a three-way blast was also performed using a custom script to identify orthologs between sagebrush samples. The default settings in Geneious version 6.05 (Biomatters Ltd., Auckland, New Zealand) were used to align and call SNPs between putative orthologs.

Ancient Gene Duplication Detection

Because of its greater depth of coverage, paralogs in *A. arbuscula* were detected by a self-blast with a maximum e-value threshold of $1e^{-20}$. Reciprocal blast hits were considered as potential paralogs. The synonymous mutation rate (Ks) was calculated for each paralog pair. A histogram of pairwise of Ks values was plotted. The highest peak was taken as the best estimate of a duplication event. We then calculated the time of this event by using the estimated background mutation rate in dicots used by Blanc and Wolfe [7] of 1.5×10^{-8} substitutions per synonymous site per year. The location and number of peaks were detected using the program EMMIX [29] by selecting the model with lowest Bayesian information criterion from models predicting 1-10 peaks. Statistically significantly peaks were identified using SiZer [10] a program that determines peak significance ($p<0.05$) by detecting changes in the slope of a curve.

RESULTS

Transcriptome Assembly

Trimmed 250 bp paired-end reads were assembled *de novo* using SOAPdenovo-Trans at variable k-mer lengths for a total of 35 assemblies for each sagebrush sample. The best assembly was chosen based on number of scaffolds, number of scaffolds >800 bp, number of bp in scaffolds > 800 bp, and N50 (Fig 1). At short k-mer lengths (~35-47), the assembler was not able to sufficiently differentiate similar sequences, so they collapsed together. At moderate k-mer

lengths (~47-75), contigs were again broken—likely due to bubbles, assemblies split by polymorphisms, in the contig graph. At long k-mer lengths (~75-127), the assembler was able to differentiate similar regions and make a less error-prone assembly. In all cases the best assembly for all samples was with a k-mer length of 127. A larger k-mer length may have produced a more acceptable assembly; however, SOAPdenovo-Trans is currently limited to 127-mers. Assemblies of 127-mers had the least amount of scaffolds coupled with the greatest N50. A smaller number of scaffolds with a greater N50 indicate that the assembler was able to join together multiple scaffolds as contigs. This is also indicated by the decreasing number of scaffolds > 800 bps and the number of bps in scaffolds > 800 bp. The assemblies with highest quality are summarized in Table 1**.**

Transcriptome Characterization

The program TRAPID identified a total of 61,883 transcripts, representing 3,427 GO terms and a total of 6,067 gene families with the greatest number of transcripts (407) mapping to the 568_HOM000025 gene family, which is associated with ATP-binding. The transcripts are divided unevenly between the samples with the majority of transcripts detected in *A. arbuscula*, likely because of the increased read coverage from that sample (Fig. 2). More *A. tridentata* transcripts would likely be discovered with increasing read coverage.

Terpene Synthases

As an example of how these transcriptomes may be used, 16 transcripts related to terpene synthases (TS) were found by searching for protein domains associated with terpene synthases (IPR001906) or by the gene family HOM000066. The 16 transcripts—12 in *A. arbuscula*, 3 in *A. t.* ssp. *tridentata* and 1 in *A. t.* ssp. *wyomingensis*—are listed in Table S1. Blasting transcript

C44821 from *A. arbuscula* against the NR database showed a single match of 89% percent identity with an E-value of $1e^{-255}$ and query coverage of 100% for TPS5 (MrTPS5) identified in chamomile (*Matricaria chamomilla*). The putative TPS5 of *A. arbuscula* ssp. *arbuscula* was used to search the transcriptomes of *A. tridentata* ssp. *wyomingensis* and *A. tridentata* ssp. *tridentata*. A single hit was found for *A. tridentata* ssp. *tridentata* (C12295) with an E-value $1e^{-255}$ and 96% identity. A multiple alignment (Fig. 3) of the three transcripts revealed 38 SNP loci in chamomile, 12 SNP loci in big sagebrush and 5 SNP loci in low sagebrush compared to the consensus sequence of all three sequences. Both sagebrush species also possessed 2 tandem amino acid deletions when compared to chamomile. There were 19 shared non-synonymous mutations in both the sagebrushes. *A. arbuscula* ssp. *arbuscula* and *A. tridentata* had 5 and 9 unique non-synonymous sites, respectively. *A. annua* currently represents most of the available transcript data for the genus *Artemisia* on NCBI. Though *A. annua* is classified under the same genus, it is not a sagebrush and despite having the largest collection of published sequences for genus *Artemisia* we could not find an orthologous sequence for this putative TPS5.

Detection of Ancient Gene Duplication

Excluding self-hits and hits that were too divergent for the Jukes-Cantor model of DNA substitution, we detected 4,383 viable paralog hits for peak detection. The maximum detected Ks value was 1.4640 and the minimum was 0.0011 with a median valued of 0.2062. We deliberately included multiple potential paralogs for each sequence in order to accurately detect historic genome duplications.

EMMIX detected seven peaks at K_S values of 0.01, 0.022, 0.05, 0.12 0.27, 0.51, and 0.91 (Fig. 4). The first four peaks were excluded because they were ≤ 0.1. The remaining three peaks we considered as evidence for ancient duplications. From our analysis of significance using

SiZer, only a single large peak from $K_S \approx 0.22$ to 0.60 was shown to be significant. For comparison we dated our duplications using the background mutation rate of 1.5×10^{-8} substitutions per synonymous site per year. We estimate these three duplication events that were in the predecessor of *A. arbuscula* ssp. *arbuscula* to be around 18 million years ago (mya), 34 mya and 60 mya.

DISCUSSION

We present three assembled transcriptomes of sagebrush now in the public domain (PRJNA258191, PRJNA258193, PRJNA258169). These transcriptomes add to the sparse amount of transcriptome data currently available for analysis in sagebrush. With a total of 61,883 transcripts identified by TRAPID, these transcriptome assemblies are a resource for advancing the characterization of species and subspecies and their chemical pathways related to defense, plant communication and a variety of other secondary compounds.

Sixteen transcripts with protein domains associated with terpene synthases (TSs) were identified, among them a putative Amorpha-4,11-diene synthase, the TS responsible for the malaria drug artemisin. Terpenoids like artemisin comprise the largest groups of natural products with over 30,000 distinct chemical structures [43]. They are involved in a series of biological processes such as formation of biological structures, defense and signaling [19].

Many TSs have been found to synthesize multiple products from a single substrate [8], [20]. Thus, a single TS is of great importance in discovering and understanding a variety of terpenoid products. While chemical pathways radiating from a single TS to a diversity of terpenes have fundamentally been explained, the mechanism that switches between the different pathways is still unknown. Degenhardt et al. [14] assert that one of the best ways to improve

understanding of TS function is to have more primary amino acid sequences in order to identify functional elements of TS. Transcripts allow for detection of protein functional groups that aid in detection of these elements.

A putative ortholog1 TS (*MrTPS5*) of *M. chamomilla* was found in both *A. tridentata* and *A. arbuscula.* In chamomile, *MrTPS5* has been found to produce mainly germacrene D, a volatile emission produced in response to herbivory [2]. However, demonstrating that TS produce multiple products, Irmish et al. [20] detected trace amounts of a variety of other terpenoids also produced by *MrTPS5*.

There were 38 SNPs between the chamomile transcript and the consensus sequence of *A. tridentata* and *A. arbuscula*—including 19 non-synonymous SNPs—which may contribute to a divergence of terpenoid products. Furthermore, the 6 bp deletion in sagebrush when compared to chamomile may indicate an autapomorphic feature derived within the tribe Anthemideae between the *Artemisia* and *Matricaria* genera. Whether these idiosyncrasies contribute to structural or functional differences in the resultant synthase proteins and subsequent terpenes has yet to be determined.

The sequences of loci involved with terpene products could be important in classification and phylogenetic analysis because it has been shown that terpenes exist in different quantities and types between species and subspecies of sagebrush [9], [27]. Exploiting these differences could bypass the subjective nature of morphology in favor of a genetic basis. This would be especially useful in the sagebrushes, where hybridization can make variable morphological characters difficult to interpret. The transcripts may prove more useful than their metabolic products because highly divergent TSs have been shown to produce the same product and highly similar TSs have been shown to produce different products [8], [41].

While our study focused primarily on TS transcripts, these transcriptomes possess a wealth of other research possibilities for studying sagebrush. For example, we also detected 39 transcripts related to the coumarin pathway. The coumarins are important for both the identification and ecological effect of sagebrush [27], [46]. Coumarins are a water-soluble class of chemicals that fluoresce blue when exposed to UV-light and present in the different taxa of sagebrush at varying levels [39]. Grinding sagebrush leaves in alcohol or water in the presence of UV-light can distinguish between different types of taxa such as *Artemisia arbuscula*—which fluoresces brightly—and *Artemisia tridentata* ssp. *wyomingensis*, which has little or no fluorescence. In addition, coumarin content can also predict the palatability of sagebrush; regardless of species, sage-grouse prefer sagebrush with greater fluorescence [35]. These transcriptomes provide a genetic basis for this important chemical pathway.

Polyploidy is an evolutionary process that creates genetic diversity, drives morphological complexity and may have afforded a greater resistance to extinction [13], [45]. At least one polyploid ancestor is suspected in all plant species [7]. These ancient duplications can be difficult to detect due to gene loss; however, analysis of existing paralogs can reveal a signal that lends to inference. In their study of ancient duplications in model plant species, Blanc and Wolfe [7] were unable to detect ancient duplications in any Asteraceae. Barker et al. [4] continued the work in Asteraceae from ESTs for 4 tribes of Asteraceae and found evidence for family level duplications in all samples as well tribe specific duplications in two samples. However, sequences for tribe Anthemidiae (which includes sagebrush) were not included in their study, and nearly all available sequences for genus *Artemisia* are from *A. annua,* a wormwood.

Our detection of ancient duplications revealed three secondary peaks with overlapping tails outside of the initial peak of recent gene duplications. The program SiZer was unable to

differentiate two peaks identified by EMMIX and called a single broad peak from a $K_S \approx 0.22$ to 0.60 as significant. We believe that the large overlap of these peaks obscures their identity. Evidence for two peaks is indicated by an additional sharp decline in the SiZer map at $K_S = 0.71$. Additional evidence for the peak at $K_S = 0.51$ is the replication of a similar peak by Barker et al. [4]. They were also unable to find the most ancient duplication ($K_S = 0.91$) as a significant peak using SiZer. However, our detection of this peak, as well as their detection of similar peaks in all four of their sampled tribes of Asteraceae suports makes us agree with their conclusion that the significance of this peak is obscured by its overshadowed positive slope by the negative slope of more prominent recent duplications.

The presence of two of our secondary peaks is congruent with a study by Barker et al. [4] that has demonstrated that tribes such as Cardueae and Cichorioideae within Asteraceae retain a detectable signal for the shared paleopolyploidization at K_S near 0.90, while others such as Mutisieae and Heliantheae possess signals for tribe specific paleopolyploidization events near a K_S of 0.50. Furthermore, based on their data they estimate that tribe-specific duplications should fall within the expected Ks range of 0.50-0.62; our detected K_S value of .51 falls within this range. We estimate a K_S value of 0.50 to correspond to 34 mya. This is near a previously estimated range (33-39 mya) for the radiation of the Asterodiae tribes [24], which includes the sagebrush tribe Anthemideae. The more ancient peak at $K_S = 0.91$ is likely an ancient paleopolyploidization shared by all members of the Asteraceae estimated 50 mya near the divergence of Asteraceae from its sister group Calyceraceae [16], [24].

The more recent peak at $K_S = 0.27$ corresponds to a time about 18 mya and was not detected in other tribes of Asteraceae sampled by Blanc and Wolfe [7] or Barker et al. [4]. This ancient duplication also occurred more recently than the estimates of Asteraceae tribe

differentiation near a K_S value of 0.50. Instead, this peak at 18 mya may be evidence of a duplication event unique to the divergence of genus *Artemisia*. Similar results have been been reported using the most reliable pollen fossil of *Artemisia* for a calibration point and genetic data (nrDNA, ITS and ETS) by Sanz et al. [37]. They estimated the divergence of *Artemisia* from its most closely related genera to have taken place around 19.8 mya in the Early Miocene.

While it is not certain that these putative WGD resulted in these species divergent events, Soltis et al. [38] have highlighted a positive correlation with the divergence of angiosperms in the recent aftermath of WGD. Furthermore, as we have shown, our estimated dates fall near other independently estimated dates for major events in the evolutionary history of sagebrush. Thus we believe this study lends genetic support to a divergence of the Asteraceae near 50 mya, the radiation of the Asterodieae tribes 33-39 mya and an ancient duplication event unique to genus *Artemisia* around 20 mya. This data also allows for future evolutionary and phylogenetic comparisons in the already described tribes of Asteraceae as well as more distantly related taxa.

REFERENCES

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.
2. Arimura G, Huber DP, Bohlmann J (2004) Forest tent caterpillars (*Malacosoma disstria*) induce local systemic diurnal emissions of terpenoid volatiles in hybrid populations (*Populus trichocarpa* x *deltoides*): cDNA cloning, functional characterization, and patterns of gene expression of (-)-germacrene D synthase, PtdTPS1. The Plant J 37: 603-616.
3. Bajgain P, Richardson BA, Price JC, Cronn RC, Udall JA (2011) Transcriptome characterization and polymorphism detection between subspecies of big sagebrush (*Artemisia tridentate*). BMC Genomics 12: 1-15.
4. Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH (2008) Multiple paleopolyploidizations during the evolution of the compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol Biol Evol 11: 2445-2455.
5. Beetle AA (1960) A study of sagebrush: The section *Tridentatae* of *Artemisia*. Bulletin 368. Laramie, WY: University of Wyoming, Agricultural Experiment Station 83: 416.
6. Bel MV, Proost S, Neste CV, Deforce D, Van de Peer Y, Vandepoele K (2013) TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes. Genome Bio 14: 1-10.
7. Blanc G, Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. The Plant Cell 16:1667-1678.

8. Bohlmann J, Steele CL, Croteau R (1997) Monoterpene synthases from grand fir (*Abies grandis*). cDNA isolation, characterization, and functional expression of myrcene synthase, (-)-(4S)-limonene synthase, and (-)-(1S,5S)-pinene synthase. J Biol Chem 272: 21784–21792.
9. Byrd DW, McArthur ED, Wang H (1998) Narrow hybrid zone between two subspecies of big sagebrush, *Artemisia tridentata* (Asteraceae). VIII. Spatial and temporal pattern of terpenes. Biochem Syst Eco 27: 11-25.
10. Chaudhuri P, Marron JS (2012) SiZer for exploration of Structures and Curves. J Am Stat Assoc 94: 807-823.
11. Connelly J, Schroeder MA, Sands AR, Braun CE (2000) Guidelines to manage sage grouse populations and their habitats. Wildlife Soc Bull 28: 967-985.
12. Copeland HE, Pocewicz A, Naugle DE, Griffiths T, Keinath D, Evans J, Platt J (2013) Measuring the effectiveness of conservation: a novel framework to quantify the benefits of sage-grouse conservation policy and easements in Wyoming. PLoS One: doi:10.1371/journal.pone.0067261.
13. Crow KD, Wagner GP (2006) What is the role of genome duplication in the evolution of complexity and diversity? Mol Biol Evol 23: 887-892.
14. Degenhardt J, Kollner TB, Gershenzon J (2009) Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. Phytochem 70: 1621-1637.
15. Frye GG, Connelly JW, Musil DD, Forbey JS (2013) Phytochemistry predicts habitat selection by an avian herbivore at multiple spatial scales. Ecology 94: 308-314.

16. Funk et al. (2005) Everywhere but Antarctica: using a supertree to understand the diversity and distribution of the Compositae. Biol Skr 55: 3403-3417.

17. Geneious version (6.05) created by Biomatters. Available from http://www.geneious.com/

18. Hecht S, et al. (2001) Studies of the nonmevalonate pathway to terpenes: The role of the GcpE (IspG) protein. PNAS 98: 14837-14842.

19. Hsieh MH, Goodman HM (2005) The Arabidopsis IspH homolog is involved in the plastid nonmevalonate pathway of isoprenoid biosynthesis. Plant Physiol 138: 641-653.

20. Irmish S, Krause ST, Kunert G, Gershenzon J, Degenhardt J, Köllner T (2012) The organ-specific expression of terpene synthase genes contribute to the terpene hydrocarbon composition of chamomile essential oils. BMC Plant Bio 12: 1-13.

21. Joshi N (2013). Sickle windowed adaptive trimming for fastq files using quality. Available from http://github.com/najoshi/sickle

22. Karban E, Shiojiri K, Huntzinger M, McCall AC (2006) Damage-induced resistance in sagebrush: volatiles are key to intra- and interplant communication. Ecology 87: 922-930.

23. Kessler A, Halitschke R, Diezel C, Baldwin IT (2006) Priming of plant defense responses in nature by airborne signaling between *Artemisia tridentata* and *Nicotiana attenuata*. Oecologia 148: 280-292.

24. Kim KJ, Choi KS, Jansen RK (2005) Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae). Mol Biol Evol 22: 1783-1792.

25. Luo et al. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1:18.

26. Mahalovich MF, McArthur ED (2004) Sagebrush (Artemisia spp.) Seed and Plant Transfer Guidelines. Native Plant J 5: 141-148.

27. McArthur ED, Welch BL, Sanderson, SC (1988) Natural and Artificial Hybridization between big sagebrush (*Artemisia tridentata*) subspecies. J Hered 79: 268-276.

28. McArthur ED, Sanderson SC (1999) Cytogeography and chromosome evolution of subgenus *Tridentatae* of *Artemisia* (Asteraceae). Am J Bot 86: 1754-1775.

29. Mclachlan G, Peel D, Basford K, Adams P. (1999) The EMMIX software for the fitting of mixtures of normal and t-components. J Stat Softw 4: 2.

30. Ohno S (1970) Evolution by gene duplication. New York: Springer.

31. Pellant M (1996) Cheatgrasss: the invader that won the west. BLM, Interior Columbia Basin Ecosystem Management Project.

32. Richardson BA, Page JT, Bajgain P, Sanderson SC, Udall JA (2012) Deep sequencing of amplicons reveals widespread intraspecific hybridization and multiple origins of polyploidy in big sagebrush (*Artemisia tridentata*; Asteraceae). Amer J Bot 99: 1962-1975.

33. Remington TE, Braun CE (1985) Sage grouse food selection in winter, North Park, Colorado. J Wildlife Manage 49: 1055-1061.

34. Rensing SA (2014) Gene duplication as a driver of plant morphogenetic evolution. Curr Opin Plant Bio 17: 43-48.

35. Rosentreter R (2005) Sagebrush identification, ecology, and palatability relative to sage-grouse. Sage-grouse habitat restoration symposium proceedings pp. 3-16.

36. Rowland M, Suring LH, Wisdom MJ (2005) Assessment of habitat threats to shrublands in the Great Basin: a case study. Advances in Threat Assess and Their App to Forest and Rangeland Manage. pp. 673-685.

37. Sanz M, Schneeweiss GM, Vilatersana R, Valles J (2011) Temporal origins and diversification of *Artemisia* and allies (Anthemideae, Asteraceae). Collectanea Botanica 30: 7-15.

38. Soltis DE et al. (2009) Polyploid and angiosperm diversification. Am J Bot 96: 336-348.

39. Stevens R, McArthur ED (1974) A simple field technique for the identification of some sagebrush taxa. J Range Manage 27: 325–326.

40. Thacker ET, Gardner DR, Messmer TA, Guttery MR, Dahlgren DK (2011) Using Gas Chromatography to Determine Winter Diets of Greater Sage-Grouse in Utah. J Wildlife Manage 76: 588-592.

41. Theis N, Lerdau M (2003) The evolution of function in plant secondary metabolites. Int J Plant Sci 164: S93-S102.

42. Turi CE, Shipley PR, Murch SJ (2014) North American *Artemisia* species from subgenus *Tridentatae* (sagebrush): a phytochemical, botanical and pharmacological review. Phytochem 98: 9-26.

43. Umlauf D, Zapp J, Becker H, Adam KP (2004) Biosynthesis of the irregular monoterpene *Artemisia* ketone, the sesquiterpene germacrene D and other isoprenoids in *Tanacetum vulgare* L. (Asteraceae). Phytochem 65: 2464-2470.

44. Vanneste K, Van de Peer Y, Maere S (2013) Inference of genome duplications from age distributions revisited. Mol Biol Evol 30: 177-90.

45. Vekemans D. et al. (2012) Gamma paleohexaploidy in the stem lineage of core eudicots: significance for MADS-Box gene and species diversification. Mol Biol Evol 29: 3793-3806.

46. Welch BL, McArthur ED (1986) Wintering mule deer preference for 21 accessions of big sagebrush. West N Am Naturalist 46.

TABLES

Table 1. Summary of 127 k-mer Assemblies Using SOAPdenovo-trans for *A. tridentata tridentate* (UTT2), *A. tridentata wyomingensis* (UTW1) and *A. arbuscula* (CAV-1).

	UTT2	UTW1	CAV-1
Scaffolds	16,276	9,741	35,866
Bps in Scaffolds > 800 bps	3,720,411	1,612,837	12,873,716
# of Scaffolds > 800 bps	3,310	1,448	10,131
N50 Scaffold Length	652	585	809

FIGURES

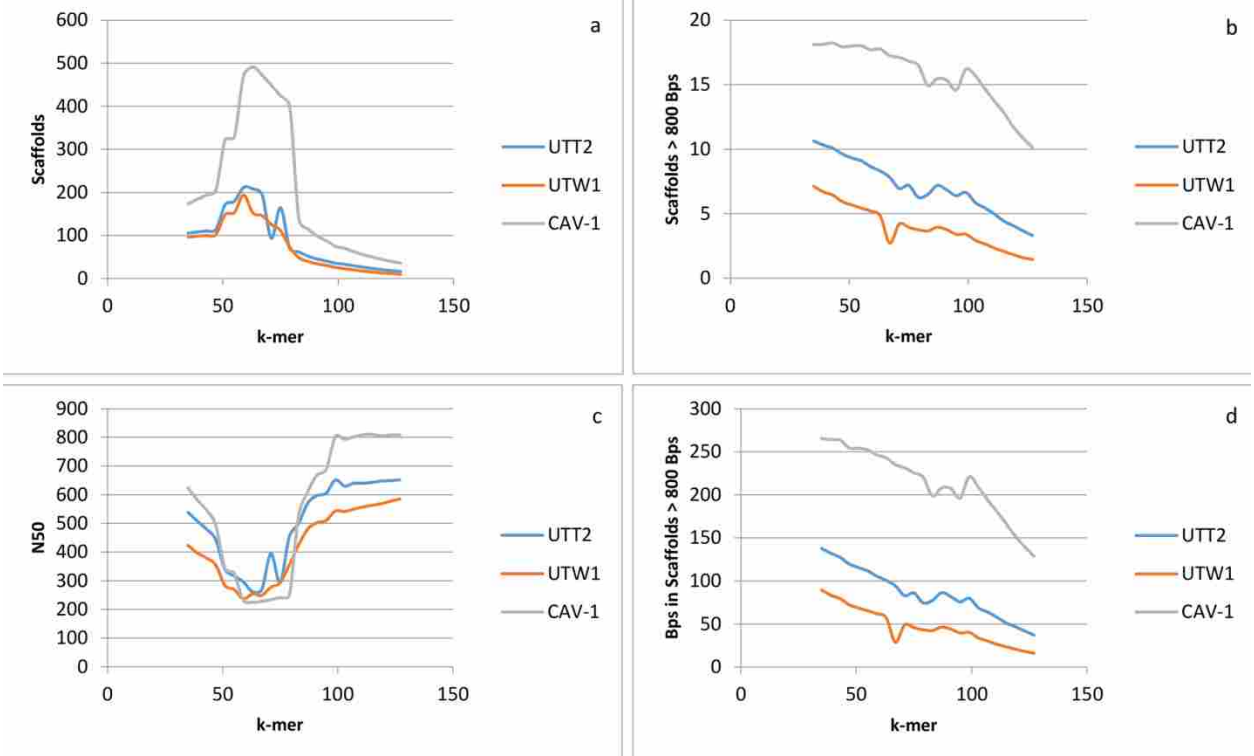

Figure 1. Assembly Statistics Based on Variable k-mer Lengths Assemblies were made based on variable k-mer lengths ranging from 35-127 in multiples of four. a) Thousands of scaffolds vs. k-mer length. b) Thousands of scaffolds > 800 bp vs. k-mer length. Scaffolds are divided by 1000. c) N50 vs. k-mer length. d) Mbps in scaffolds > 800 vs. k-mer length.

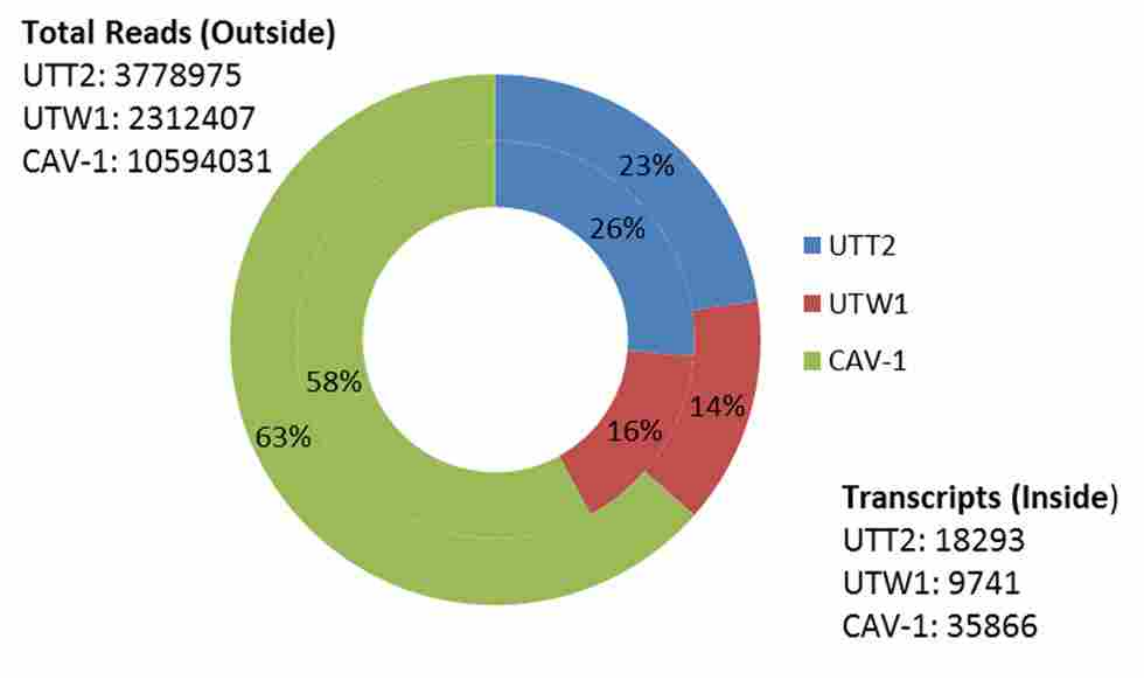

Figure 2. Distributions of Detected Reads and Transcripts The outside ring is the number of initial reads and the inside ring is the number of detected transcripts. Reflecting the number of reads, the majority of detected transcripts were from CAV-1, while both UTT2 and UTW1 had a similar number of reads between them.

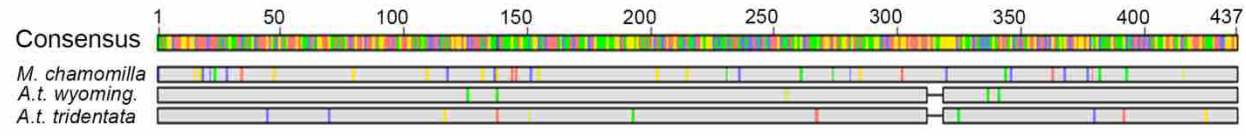

Figure 3. Multiple Sequence Alignment for a Partial Transcript of the MrTPS5 Gene A multiple sequence alignment of A. arbuscula, A. tridentata and M. chamomile showing the same 6 SNP base pair deletions present in the genus Artemisia. SNP loci are highlighted as blue for cytosine, green for thymine, yellow for guanidine and red for adenine. Geneious generated the consensus sequence by a majority vote consensus.

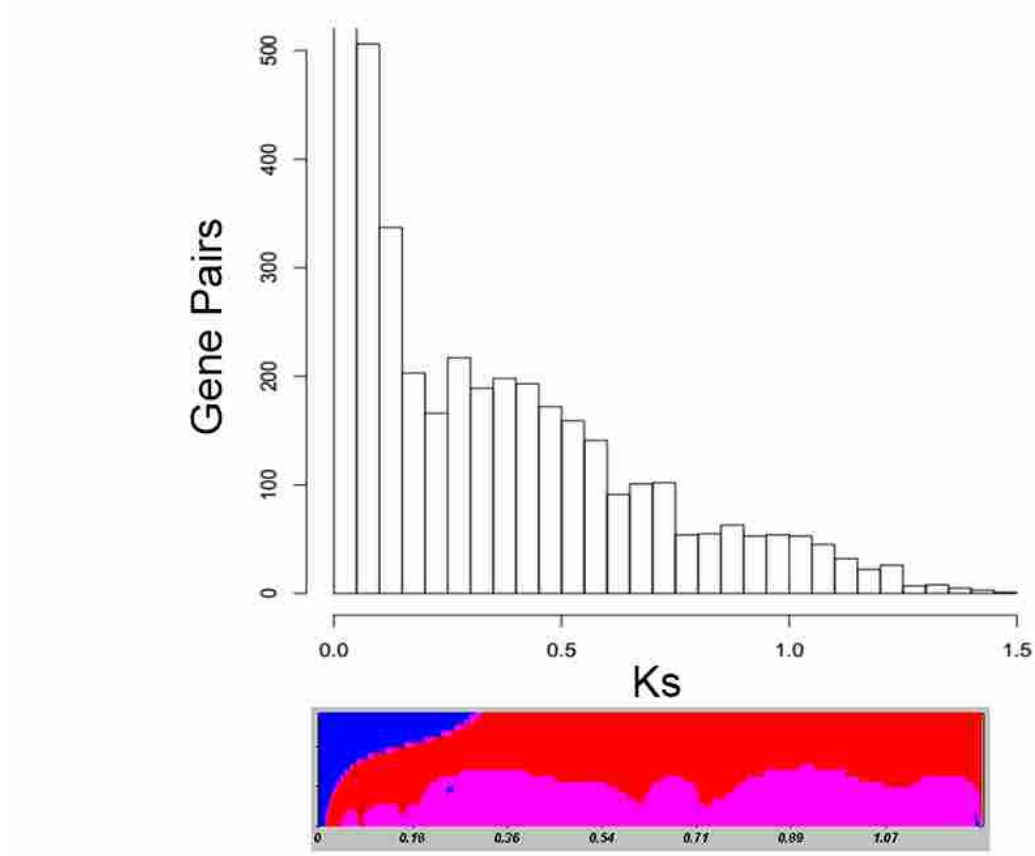

Figure 4. A histogram of K_S values with significant peaks identified in a SiZer graph below. Blue represents increases in slope; red indicates decreases; pink areas have no significant slope changes. A sharp increase at a $K_S \approx 0.22$ is indicated by a blue dot. This increase is followed by a broad pink peak of no changes with a decrease beginning at a K_S of 0.60. Additional sharp declines are identified at K_S of 0.71 and 1.34.

SUPPLEMENTAL DATA

Transcript	Gene Family	GO annotation	InterPro annotation	Subsets
Ccav24930	568_HOM000066	GO:0000287,GO:0008152,GO:0016829,	IPR001906,IPR008930,IPR005630,	CAV,
Ccav26954	568_HOM000066	GO:0000287,GO:0008152,GO:0016829,	IPR001906,IPR008930,IPR005630,	CAV,
Ccav28248	568_HOM000066	GO:0000287,GO:0008152,GO:0016829,	IPR001906,IPR008930,IPR005630,	CAV,
Ccav28334	568_HOM000066	GO:0000287,GO:0008152,GO:0016829,	IPR001906,IPR008930,IPR005630,	CAV,
Ccav44821	568_HOM000066	GO:0000287,GO:0008152,GO:0016829,	IPR001906,IPR008930,IPR005630,	CAV,
Ccav49365	568_HOM000066	GO:0000287,GO:0008152,GO:0016829,	IPR001906,IPR008930,IPR005630,	CAV,
Ccav52791	568_HOM000066	GO:0000287,GO:0008152,GO:0016829,	IPR001906,IPR008930,IPR005630,	CAV,
Ccav56438	568_HOM000066	GO:0000287,GO:0008152,GO:0016829,	IPR001906,IPR008930,IPR005630,	CAV,
Ccav60306	568_HOM000066	GO:0000287,GO:0008152,GO:0016829,	IPR001906,IPR008930,IPR005630,	CAV,
Ccav62982	568_HOM000066	GO:0000287,GO:0008152,GO:0016829,	IPR001906,IPR008930,IPR005630,	CAV,

Ccav76367	568_HOM000066	GO:0000287,GO:0008152,GO:0016829, IPR001906,IPR008930,IPR005630,	CAV,
Ccav80259	568_HOM000066	GO:0000287,GO:0008152,GO:0016829, IPR001906,IPR008930,IPR005630,	CAV,
Cutt12295	568_HOM000066	GO:0000287,GO:0008152,GO:0016829, IPR001906,IPR008930,IPR005630,	UTT2,
Cutt25863	568_HOM000066	GO:0000287,GO:0008152,GO:0016829, IPR001906,IPR008930,IPR005630,	UTT2,
Cutt5751	568_HOM000066	GO:0000287,GO:0008152,GO:0016829, IPR001906,IPR008930,IPR005630,	UTT2,
Cutw15546	568_HOM000066	GO:0000287,GO:0008152,GO:0016829, IPR001906,IPR008930,IPR005630,	UTW1

Table S1. List of 16 Transcripts Associated with Terpene Synthases. Transcripts identified as being associated with terpene synthases by association with the HOM000066 gene family.

CHAPTER 2

Two Genetic Maps of Yellow Lupin

INTRODUCTION

Sustaining the world's food supply and textile industry relies heavily on crop improvements—especially in the face of a changing climate. Introduction of favorable variation such as increased yield and disease resistance relies on identifying robust genetic markers for these traits. Once markers are identified, genetic maps can be created and breeders can utilize marker-assisted selection (MAS) to produce desirable cultivars. Genetic maps can also be useful for *de novo* assembly of genome sequences.

Historically genotyping and genome mapping relied primarily on molecular markers such as RFLPs, SSRs and AFLPs. With SSRs and other PCR-based assays, *a priori* sequence information is needed to develop probes or primers for polymorphic loci. While AFLPs do not require any knowledge of the genome sequence, they are limited by their general ability to only detect dominant markers and thus unable to detect heterozygous loci. Additionally, technology based on indels or polymorphisms in restriction sites may not provide sufficient markers for the needed resolution of tightly linked markers within an ideal 5-10 cM of a trait locus [8]. A target enrichment technology that overcomes the need of *a priori* sequence information and marker density would be very useful to create efficient, high-density genetic maps.

Targeted enrichment of specific genomic regions within a population of individuals can provide thousands of useful single nucleotide polymorphisms (SNPs) while avoiding the costs of sequencing entire genomes [9], [31]. Genotyping-by-sequencing (GBS) is one such high-throughput method of targeted enrichment. GBS generates a high density of SNP markers in a

relatively inexpensive, efficient, and straightforward manner [12]. GBS does not require prior sequence information and it does utilize many thousands of molecular markers for high-resolution QTL mapping.

One particular method of GBS implementation enriches for targets with flanking restriction sites of two different enzymes [32], where one enzyme is typically methyl-sensitive to avoid targeting repeat regions. Targeted fragments have a barcoded adapter ligated on the 5' end; the barcode is later used for sample identification during demultiplexing. A non-barcoded Y-adapter is ligated on the 3' end of the digested fragments. This Y-adapter ensures that only fragments that have been cut by both enzymes will be amplified during PCR. After adapter ligation, PCR adds additional adapters complementary to Illumina sequencing primers. The fragments are then sequenced using Illumina sequencing technology.

In contrast to the relative simplicity of generating GBS data, downstream analysis of the sequencing data has proven to be more challenging, particularly in species without a reference genome sequence. This has catalyzed the production of a number of custom GBS solutions [19], [36]. The most widely used solution for species without a reference genome is currently UNEAK [16]—however one of the current weaknesses of UNEAK is that it trims all reads to 64 base pairs. This trimming causes a potential loss of polymorphic loci positioned in the read beyond the first 64 base pairs. A large amount of missing data resulting from both low and uneven coverage across samples [1], [10], [17] is a well-documented weakness of GBS. Another potential weakness is sequencing depth or coverage. Accurate genotyping of heterozygote individuals requires sufficient sequence coverage at targeted loci, potentially limiting GBS to either inbred populations or additional sequencing costs in large populations.

Yellow lupine (*Lupinus luteus*) is a native crop of the Mediterranean that has been domesticated in Africa, Australia and South America. Yellow lupine belongs to the legume family, Fabaceae, and is distantly related to other cultivated legumes (soybeans, pea, etc.). Yellow lupine has limited genomic resources. Its 2n=52 genome has not been sequenced and assembled. Its limited resources include an EST library of about 72,000 contigs [29]. A close relative blue lupine (*L. angustafolius)* has been more widely cultivated. Both a draft genome sequence and a genetic map have been created for blue lupine [45]. Despite this advancement, Berger et al. [2] argue that lupine global production is declining but its value could be improved by introducing genetic diversity from wild populations and by unlocking novel untapped genetic resources within existing cultivars.

Despite the relatively sparse genomic resources of yellow lupine, many phenotypic traits make it an increasingly desirable crop for rural areas that suffer from poor soil and limited access to protein-rich diets. For example, its evolution in dry, shallow, acidic and sandy soils [3] is a welcomed trait for environments of Western Australia which have at least 200,000 ha of acidic sands [6]. Yellow lupine also has highest protein content than other lupines. A remarkable average of approximately 45%, [34] protein content and 5.5% oil content [38] make yellow lupine a welcome candidate for supplementing human and livestock diets. However, its widespread implementation has been limited by factors such as high levels of alkaloids that make its consumption difficult for both humans and livestock. Identification of QTL for desirable and non-desirable traits would help growers to target and tune traits for better and more competitive crops.

In this study, we have used GBS to genotype two different populations of *L. luteus*—an eight generation recombinant inbred (RIL) and an F_2. We describe the methodologies we used for

GBS and compare the results from the two populations. We also offer a draft of a genetic map for yellow lupine and identify blocks of segregation distortion spread over several linkage groups.

MATERIALS AND METHODS

RIL Population

One hundred and fifty-seven samples from the Australian Woodjilx x P28213 population [2] were processed using the GBS protocol outlined by Poland et al. [32] with the addition of size selection step. First, sample genomic DNA was quantified using Quant-iT™ PicoGreen (Life Technologies, Carlsbad, California) and normalized to 40 ng/ul. Second, the DNA samples were digested with *two enzymes* PstI-HF and TaqαI. With a genome size of 980 and approximately 44% GC content, this produces a theoretical 683 fragments with a PstI-HF cut on the 5' end and a TaqαI cut on the 3' end. Third, 96 barcoded adapters for downstream identification were ligated to the 5' end of digested fragments. In concert, a Y-shaped adapter was ligated to the 3' end. Lastly, fragments were amplified with the addition of Illumina adapters through PCR. Amplified bands ranging from 400 to 700 bps were cut from a gel of the PCR products and eluted using a Promega SV Wizard Gel Clean-Up System (Promega Corporation, Madison, Wisconsin). Products were sent for 150 bp paired-end sequencing on 2 lanes of an Illumina HiSeq at BGI at UC Davis.

F_2 Population

One hundred and eighty-eight lupine samples of an F_2 generation from Centro de Genómica Nutricional Agroacuícola in Chile, including one parent, were sent to Cornell University. The samples were prepared by GBS using *a single enzyme* PstI and sequenced using

a modified version of the protocol (http://www.biotech.cornell.edu/.brc) by Elshire et al. [12]. The data were then sent to the Udall Lab at Brigham Young University for analysis.

Data Analysis

GBS reads were quality trimmed with sickle (http://github.com/najoshi/sickle/) and demultiplexed. Each pair of reads was categorized based on an exact match in the forward read to one of the barcodes, and barcodes were trimmed off. Using GSNAP [42], both the RIL and F_2 GBS reads were mapped to an unpublished SOAPdenovo [25] assembly of Illumina whole-genome shotgun reads of *L. luteus* called 9242X4. Sorted BAM files were prepared by SAMtools [22].

SNP Calling

Processing of BAM files—including SNP calling, imputation and phasing—was completed with the BamBam tool suite [27]. SNPs were called with a minimum coverage of 1, 2, or 3 reads. In order to capture all possible polymorphic loci, we used the 1x minimum coverage SNPs to build genetic maps while relying on the strictness of downstream filters. SNPs were then filtered by requiring less than 30% missing genotypes at a given locus with a minor allele frequency of 0.1 and a minimum coverage of 10 individuals for each minor allele. Missing genotypes were imputed by K-Nearest Neighbor with K = 10. Any loci with unknown genotypes for both parents were removed. F_2 haplotypes were coded based on similarity to known or imputed parental genotypes. Additional filtering was performed as part of marker mapping.

Marker Mapping

In constructing linkage groups for a genetic map, the question of whether to keep markers together or break them into separate linkage groups must be answered on a case-by-case basis. In

this study we chose a conservative approach and favored breaking groups to avoid creating artificial linkages.

Additional filtering of SNP loci was carried out by eliminating duplicate loci and loci that showed significant ($P \leq 0.0001$) segregation distortion (SD) from the expected Mendelian segregation ratios. Genetic markers from the RIL population were mapped using JoinMap 4.0 [40]. Markers were first assembled into large groups and then broken into smaller groups based on logarithm of the odds (LOD) scores.

LOD scores are a statistic used to show the odds that two or more loci are linked. It is calculated by taking the log of the likelihood of loci linked divided by the likelihood of loci being unlinked. A LOD score of 3.0 is generally considered minimum evidence that loci are linked. A LOD score of 3.0 means that the odds that two loci are linked is 1 out of a 1000, a LOD score of 4.0 means the odds are 1 out of 10,000, and so forth. Groups were initially formed at LOD score 4 and then divided at scores ranging from 7 to 20. Mapping and ordering of loci were completed using a maximum likelihood method. In order to ensure high quality, all weak linkages (recombination > 0.45 or LOD < 0.05) and suspect linkages (recombination > 0.50) were broken by forming another linkage group at the next highest LOD score or removal of certain loci.

The length of linkage groups can also guide their construction. The longer a linkage group becomes the more weak linkages (>35 cM) and suspect linkages (>50 cM) it likely contains. Many, but not all, suspect and weak linkages were filtered out because of LOD scores lower than 4. To ensure high quality linkage groups, we chose to break weak and suspect linkages rather than assuming that actual linkages existed (i.e. reduce false positives). This meant either excising a single locus or forming that linkage group at a higher LOD score.

RESULTS

Read Counts and SNP Calls

Read trimming and demultiplexing of reads from the RIL population containing 157 individuals yielded a total of 743M reads, with an average of 4.7M reads per sample (Table 1). 658M (88.6%) of the reads from the RIL population mapped to the 9242X4 reference. We selected this reference based on higher mapping percentage overall and per individual line when compared to mapping against a draft genome of blue lupine (data not shown). Our pipeline identified 4,448 marker loci for 157 individuals consisting of 197,619 (Woodjilx) genotypes, 411,654 (P28213) genotypes and 20,413 heterozygote genotypes (Table 2).

Read trimming and demultiplexing of reads from the larger F_2 population of 2 parents and 186 progeny yielded a total of 418M reads, with an average of 1.5M reads per sample (Table 1). 66% of the reads mapped from the F_2 population to the 9242X4 reference. We selected this reference based on higher mapping percentages overall and per individual line when compared to mapping against a draft genome of blue lupine (data not shown). Our pipeline generated 1,021 loci for the 186 progeny consisting of 64,136 Core 227 genotypes, 59,019 Core 98 genotypes and 51,611 heterozygote genotypes (Table 2).

The number of SNP markers decreased with increasing minimum coverage requirements, where the number of loci was compared before and after filtering at three different levels of coverage (Table 2). The number of loci mapped was very different in the two populations. In comparison to the F_2 population, the RIL population did not have an as dramatic decrease in loci when the minimum coverage threshold was raised. The RIL population had 4,448 loci at 1x coverage and retained 3,178 loci at 3x coverage. When the threshold for minimum coverage was raised in the F_2 population from 1x to 3x, the loci dropped from 1,021 to 2 loci. Based on these

results we decided to keep all loci at 1x coverage. This limited amount of coverage could be improved by additional sequencing. After duplicate markers were condensed the RIL population was left with 3428 loci and the F_2 population with 950.

Linkage Groups

Additional filtering of SD in JoinMap 4.0 yielded 1,101 markers for the RIL population. These markers were used to construct 31 linkage groups (Figure 5) for the expected 26 haploid chromosomes of *L. luteus*. The groups were formed with an average LOD score of 14.4. The size of the linkage map totaled 1,690.9 cM. Groups ranged from 16 to 105 cM with an average marker density of 0.46 markers per cM that ranged from 0.16 to 2.27 markers per cM (Table 3).

Using 950 SD filtered markers from the F_2 population, we constructed 20 linkage groups (Figure 6). The 20 linkage groups cover a total of 1,471 cM and were formed at an average LOD score of 17 (Table 4). The groups ranged from 30 to 135 cM with average size of 73.6 cM. Although the total sizes of both linkage maps were similar in size, the average marker density 0.13 of the F_2 population was much less than the RIL population. The range of marker density was also much lower at 0.09 to 0.22. Some F_2 linkage groups displayed SD which may have been the consequence of using 1x coverage for mapped markers (i.e. no heterozygotes).

DISCUSSION

Yellow lupine is a plant that possesses great nutritional potential, especially in protein content, yet its consumption is limited by an abundance of bitter and potentially poisonous alkaloids [21]. Genetic markers can provide a neutral genetic basis for phenotypic traits of yellow lupine such as alkaloid content. In a QTL analysis these traits can be linked to genetic loci by correlating the segregation of markers with phenotypic data. Once QTL are linked to

specific loci growers can use the information in MAS to select for or against particular traits. Over conventional breeding, MAS saves time and resources because traits can be screened for as early as the seedling stage and in single plants—opposed to large plots of plants whose phenotype may be masked or influenced by the environment [8]. To date, GBS has been used to generate millions of markers for future use in MAS (see the review by He et al. [18]).

Linkage Mapping

Using our RIL population we produced 31 linkages groups to represent the 26 chromosomes of yellow lupine. In comparison the F_2 population produced 20 linkage groups. The RIL population had both a higher number of markers and a higher density of markers. The initial numbers of unfiltered SNPs were similar between the two populations (Table 2) after filtering and assigning genotypes, but the F_2 population only retained about a fourth of the markers of the RIL population. One explanation for this is that both parents were not included on the GBS plate of the F_2 population. This was a mistake in GBS library preparation. We attempted to supplement the data of the missing GBS parent by including previously generated shotgun WGS reads of the same parent. However, the shotgun WGS library did not undergo the GBS protocol and did not have all of the loci represented. This limited the number of known alleles that could be genotyped and mapped. Another explanation is that there are fewer reads per individual in the F_2. This results in a decreased capacity to detect markers that pass minor allele frequency. Lastly, the heterozygous nature of the F_2 population makes imputation less effective.

Lack of Heterozygotes in the F_2 Population

Our F_2 data show a significant reduction in usable markers with increasing coverage requirements. At 3x coverage, and our filtering only 2 markers were identified on a MAF of 0.10

and missingness of 30% (Table 2). A majority of potential markers also suffered from severe segregation distortion—1:1:1 or 2:1:2 rather than the expected 1:2:1—both ratios suggest false homozygous calls because of the lack of heterozygotes. We suspect this distortion to be an artifact of the sequencing—especially the low coverage—rather than an actual 1:1:1 ratio in the population. However, in the end many we filtered out many of the distorted markers based on a p-value of < 0.0001

Creating a map from an F_2 population has inherent challenges when compared to a RIL population because of the need to accurately genotype heterozygotes. Uneven and low coverage—both typical of GBS studies—can affect the ability to call or impute heterozygotes. With low coverage sequencing, there is an inherent bias against identifying a heterozygote because there is a high probability of only sampling one allele from a diploid genome [15], [26]. At a locus with 1x coverage, it is impossible to detect heterozygosity. At a locus with 2x coverage, there's only a 50% probability. With 3x coverage, that probability increases to 75%. It is generally desirable to require more—sometimes much more—than a single read to recognize an allele and thus avoid erroneous genotype calls from sequencing errors, in which case the probability of confidently observing both genotypes of a heterozygote is much worse. With this increased difficulty to detect heterozygotes, ratios such as 1:1:1 or even 2:1:2 can be found within an F_2 population. These non-Mendelian ratios are not due to anomalies of transmission genetics (i.e. selection, meiotic drive, drift, etc.); rather they are based on a lack depth of sequencing coverage. Because RIL populations are not expected to have high-levels of heterozygosity, they require less sequencing depth in order to confidently call a genotype. This is perhaps why most GBS studies to date have focused on inbred populations [15]. Many of the F_2 studies have focused on crops with well-developed genetic resources such as maize.

Indeed Zhang et al. [33] in a GBS study found that within 11 F_2 populations of maize only 5% of their SNP loci were called as heterozygous. Similarly, in their methods paper Heffelfinger et al. [19] also used an F_2 maize population and warn of the prevalence of false homozygous calls in heterozygous regions due to low coverage. Because many of the tools designed for GBS experiments are for low expected heterozygosity [37], they suggest the next advancements in GBS studies should be an imputation program that can accurately call heterozygotes. One way that can improve heterozygote calls in an F_2 population is to have a reference genome. Using a reference genome, Chen et al. [7] generated a high density map with an F_2 maize population. This was completed by looking for recombinant breaks along the chromosome. First, each of their SNPs was mapped to their physical position on the reference genome. They then scanned the genome in windows of 18 SNPs where any window with less than 15 parental genotypes was deemed heterozygous. In spite of this novel method of assigning heterozygous genotypes, many plants do not yet have a reference genome and the problem of using heterozygous populations for GBS still needs to be properly addressed.

Segregation Distortion of RIL Markers

As part of our filters we removed loci that deviated significantly ($p<0.0001$) from the expected 1:1 ratio of a RIL8 population. While it is possible that these loci represent actual segregation distortion inherent in the yellow lupine genome, there is also a possibility that some of these loci represent false positives such as distortions based on less confident genotype calls resulting from low coverage. In comparison, SD did not appear as linkage group-wide blocks in the F_2 population. This is consistent with a report by Zhang et al. [46] that suggest SD is more prevalent in RIL populations than F_2 populations. Though the mechanisms for SD are not yet

fully understood, unintentional selection of some degree usually accompanies RIL population development.

There is evidence that including a large number of distorted markers can be either detrimental or beneficial for downstream QTL mapping [14], [31]. At least part of the effect of these markers likely depends on where the distortion is occurring, *i.e.* the distorted markers may not be in the proximity of effect for a given QTL. Our previous attempt at constructing a genetic map without filtering markers showing high levels of SD yielded overly large linkage groups (>200cM) with many weak and suspect linkages—even at LOD scores as high as 20. In these cases SD may have artificially inflated the degree of linkage between actually unlinked groups of markers because of 'missing' alleles from one of the parents. Thus we decided to limit the amount of SD in our markers by dropping loci with a p-value < 0.0001.

With the remaining markers we plotted the significance of their SD by position on each linkage group (Figure 5). Large blocks of SD are present in at least 5 of the 31 linkage groups. Localized blocks of SD called segregation distortion regions (SDR) have been described previously in species such as wheat [13] and barley [10]. The cause of SDR is hypothesized to their proximity to genes that are under gametic or zygotic distortion [43]. Prezygotic mechanisms are expected to cause a deviation from a 1:1 ratio of the allele frequencies, while postygotic mechanisms (*i.e.* unintentional selection by researchers) favor a particular genotype [35].

In a recent communication with the producers of our yellow lupine lines, we have discovered that SD is indeed present in the population. Segregation distortion has been noted in both flower and seed color from the expected phenotype frequencies. Because of this we also expect a higher rate of SD in our RIL population. Whether prezygotic, postzygotice or both modes of SD are involved in the distortion of our RIL population is yet to be determined. In

order to determine the precise mechanisms of SD both gametes and zygotes of our RIL population would need to be genotyped.

Improving GBS Read Quality

Beissinger et al. [1] have shown that uneven coverage of markers in the corn genome from GBS drastically reduced the number of usable markers. For example, they had coverage as high as 2,369 times expected read coverage at some loci and at other loci mapped reads were completely absent. In order to determine the amount of sequencing needed to in the population, they recommended a prescreening of a single individual. This individual would be sequenced a deep level in order to determine the amount of sequencing for adequate loci. However, this task may be more difficult than it seems: a doubling of the coverage can require a surprising nine times more sequencing [1], [11]. This was partially because random sampling of the genome does not result in even coverage across the genome, and random sampling of multiplexed samples doesn't yield equal coverage from those samples. Such sequencing is not practical in association studies because of the high cost to sequence a large number of individuals at deep coverage. Indeed, a primary benefit of GBS studies is that they avoid the cost of deeply sequencing a large number of individuals.

In order to achieve the required breadth and depth of coverage needed in genotyping, especially in heterozygous populations, some precautions must be taken in GBS [14], [15], [24]. Here we present three suggestions for improving coverage and thus genotyping in the GBS method. First, use the two-enzyme approach described by Poland et al. [20], [32], [33] that performs better than the one-enymze approach by Elshire et al. [12] when targeting unique sites in genomes >1,000 MB. The second enzyme is both methylation sensitive and a rare cutter. Requiring the more common cut site on one end of the fragment and a rare cutter on the other

end ensures that a smaller number of unique loci are targeted. Requiring that one of the enzymes is methylation sensitive ensures that they are generally in euchromatic regions of the genome. The effectiveness of a two-enzyme system is demonstrated in part by the coverage of our two populations. Our RIL population that was digested with two enzymes retains more loci at higher coverage stringencies. Theoretically, our double digest of the RIL population resulted in 683 fragments with one cut end from each enzyme. Compared to the 1.1M fragments of the ApeKI digestion of the F_2, this is a theoretical 1,500x reduction in fragments. Also, if different enzymes are used than those listed by Poland et al. [32], adapters and software must be modified appropriately to compliment different cut sites.

Second, we suggest using a size selection step to prevent sequencing of overly small or large fragments. This increases sequencing depth for a narrower size fraction of target loci. Size selection from an agarose gel for limiting the number of targeted loci has been shown to increase coverage in lodgepole pine [28]. Size selection also allows researchers to verify that the double digest provides a desired cutting profile of your DNA when selecting two enzymes on an agarose gel. Putative repeat regions represented as overly bright bands on a gel can limit the number of loci actually sequenced. If repeat regions are sequenced, they will not segregate in expected Mendelian ratios. In fact, they may not even be present in the reference sequence (depending on the assembly).

Lastly, calling and imputing accurate genotypes in a biparental population relies heavily on the quantity of data from the parents. Including each parent multiple times within a GBS lane of Illumina sequencing increases the coverage of the parental genotypes and the number of loci where one or both parents have a genotype. This results in more loci where the progeny genotypes can be phased and, consequently, more loci that can be included in a linkage map.

The 31 linkage groups of the RIL population and 20 linkage groups of the F_2 represent a significant contribution to the genomic resources of *L. luteus*. Future work in *L. luteus* that provides additional markers and/or individuals will provide the resolution needed to collapse and form groups into the expected 26 haploid chromosomes of yellow lupine. Mapping phenotypic data to these loci can identify QTL for use in MAS for a number of yellow lupine's agriculturally interesting traits such as its protein content and ability to grow in dry, saline environments.

REFERENCES

1. Beissinger et al. (2013) Marker density and read depth for genotyping populations using genotyping-by-sequencing. Genetics 193: 1073-1081.
2. Berger et al. (2013) The essential role of genetic resources in narrow-leafed lupine improvement. Crop and Pasture Sci 64:363-373.
3. Berger JD, Ludwig C (2014) Contrasting adaptive strategies to terminal drought-stress gradients in Mediterranean legumes: phenology, productivity, and water relations in wild and domesticated *Lupinus luteus L.* J Exp Bot doi:10.1093/jxb/eru006
4. Boersma et al. (2005) Construction of a genetic linkage map using MFLP and identification of molecular markers linked to domestication genes in narrow-leafed lupine (*Lupinus angustifolius L.*). Cell Mol Biol Lett 10:331–44.
5. Buirchell B, Berlandier FA (1999) Breeding for aphid resistance in yellow lupins (*L. luteus. L.*) Lupine, an ancient crop for the new millennium: Proceedings of the 9th International Lupine Conference: 154-155.
6. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635.

7. Chen et al. (2014) An ultra-high density bin-map for rapid QTL mapping for tassel and ear architecture in a large F2 maize population. BMC Genomics 15:433.
8. Collar BCY, Mackill DJ (2007) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos Trans R Soc Lond B Biol Sci 363: 557–572.
9. Deschamps S, Llaca V, May GD (2012) Genotyping-by-sequencing in plants. Biology 1: 460-483.
10. Devaux PA, Kilian A, Kleinhofs A (1995) Comparative mapping of the barley genome iwht male and female recombination-derived, doubled haploid populations. Mol Gen Genet 249:600-608.
11. Donato et al. (2013) Genotyping-by-Sequencing (GBS): A novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. PLoS One 8: e62137.
12. Elshire J et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379.
13. Faris JD, Haen KM, Gill BS (2000) Saturation mapping of a gene-rich recombination hot spot region in wheat. Genetics 154: 823-835.
14. Fu YB (2014) Genetic Diversity Analysis of Highly Incomplete SNP Genotype Data with Imputations: An Empirical Assessment. G3 4:891-900.
15. Gardner et al. (2014) Fast and cost-effective genetic mapping in apple using next-generation sequencing. G3: doi:10.1534/g3.
16. Glaubitz et al. (2014) TASSEL-GBS: A high capcity genotyping by sequencing analysis pipeline. PLoS One 9: e90346.

17. Hackett CA, Broadfoot LB (2003) Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. Heredity 90:33-38.

18. He et al. (2014) Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front Plant Sci 5: doi: 10.3389/fpls.2014.00484.

19. Heffelfinger et al. (2014) Flexible and scalable genotyping-by-sequencing strategies for population studies. BMC Genomics 15:979.

20. Hosdorf et al. (2014) Evaluation of juvenile drought stress tolerance and genotyping by sequencing with wild barley introgression lines. Mol Breeding 34: 1475-1495.

21. Księżak J, Bojarszczuk J (2014) Evaluation of the variation of the contents of anti-nutrients and nutrients in the seeds of legumes. Biotech Anim Husbandry 30:153-166.

22. Li et al. (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25:2078-2079.

23. Liu et al. (2010) Progress of segregation distortion in genetic mapping of plants. Research J Agronomy 4:78-83.

24. Lu et al. (2013) Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. PLoS ONE 9: e1003215.

25. Luo et al. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1:18.

26. Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12:443–451.

27. Page JT, Liechty ZS, Huynh MD, Udall JA: BamBam: genome sequence analysis tools for biologists. BMC Research Notes. BMC Research Notes 7:829.

28. Parchman et al. (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. Mol Eco 21:2991-3005.

29. Parra-Gonzalez et al. (2012) Yellow lupine (*Lupinus luteus L.)* transcriptome sequencing: molecular marker development and comparative studies. BMC Genomics 13:1-15.

30. Peterson GW, Dong Y, Horbach C, Fu Y (2014) Genotyping-by-sequencing for plant genetic diversity analysis: a lab guide for SNP genotyping. Diversity 6:665-680.

31. Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. The Plant Genome 5:92-102.

32. Poland J et al. (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7: e32253.

33. Russel et al. (2014) The use of genotyping by sequencing in blackcurrant (*Ribes nigrum*): developing high-resolution linkage maps in species without reference genome sequences. Mol Breeding 33:835-849.

34. Schumacher H et al. (2011) Seed protein amino acid composition of important local grain legumes *Lupinus angustifolius L.*, *Lupinus luteus L.*, *Pisum sativum L.* and *Vicia faba.* Plant Breeding 120:156-1664.

35. Schranz et al. (2007) Comparative Genetic Mapping in Boechera stricta, a Close Relative of Arabidopsis. Plant Physiol 144:286-298.

36. Sonah et al. (2013) An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS One 8: e54603.

37. Spindel et al. (2013) Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. Theor Appl Genet 126:2699-2716.

38. Sukaj A, Kotlarz A, Strobel W (2006) Compositional and nutritional evaluation of several lupine seeds. Food Chem 98:711-719.

39. Törjék et al. (2006) Segregation distortion in Arabidopsis C24/Col-0 and Col-0/C24 recombinant inbred line populations is due to reduced fertility caused by epistatic interaction of two loci. Theor Appl Genet 113:1551-1561.

40. Van Ooijen, JW (2011) Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. Genet Res 93:343-349.

41. Ward et al. (2013) Saturated linkage map construction in *Rubus idaeus* using genotyping by sequencing and genome-independent imputation. BMC Genomics 14:2.

42. Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26:873-881.

43. Xu S. (2008) Quantitative trait locus mapping can benefit from segregation distortion. Genetics 180:2201-2008.

44. Xu Y, Zhu L, Xiao J, Huang N, McCouch SR (1997) Chromosomal regions associated with segregation distortion of molecular markers in F_2 backcross, doubled haploid and recombinant inbred populations in rice (*Orzya sative L.*). Mol Gen Genet 253: 535-545.

45. Yang et al. (2013) Draft genome sequence, and a sequence-defined genetic linkage map of the legume crop species *Lupinus angustifolius L.*PLoS One 8:e64799.

46. Zhang et al. (2014) Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. Heredity 1-9. doi:10.1038/hdy.2014.99

47. Zhou L, Holliday JA (2012) Targeted enrichment of black cottonwood (*Populus trichocarpa*) gene space using sequence capture. BMC Genomics 13:703.

TABLES

Table 2. Summary Statistics of Read Mapping for RIL and F_2 Populations

RIL	Mapped	Total	%Mapped
Total	535,165,787	7.44E+08	
Average	6,774,250.5	9,413,213	73.3
F2			
Total	280,123,865	4.19E+08	
Average	1,490,020.6	2,228,178	66.4

Table 3 Summary of SNP Calls at Varying Coverages

RIL	Coverage	1	2	3
	SNP Loci	3381464	367989	79308
	After Filtering and Imputation	4448	3591	3178
	a	197619	159699	142543
	b	411654	346894	309298
	h	20413	8323	4948
F2				
	SNP Loci	3028444	36828	8896
	After Filtering and Imputation	1021	23	2
	a	64136	1343	142
	b	59019	1144	70
	h	51611	1602	106

Table 4 Summary of Segregation Distortion in RIL and F_2 Populations

Significance	RIL	F_2
-	619	172
*	86	26
**	163	23
***	69	11
****	178	24
*****	73	9
******	115	17
*******	2125	668
Total	3428	950

Table 5 Summary Statistics of 31 Linkage Groups of the RIL Population

Group	Length	Markers	LOD	Excluded
1	105	28	12	
2	101.6	39	10	
3	102.1	64	18	
4	93.4	53	14	
5	90.3	18	20	1
6	89.2	19	12	
7	92.2	52	11	1
8	79.6	44	15	
9	64.8	14	9	
10	50.5	29	20	
11	61.1	10	9	
12	56.7	9	7	
13	56.2	11	11	
14	56.3	20	11	
15	50.4	35	13	
16	45	11	15	1
17	41.4	26	20	
18	41.3	60	20	
19	40.8	35	12	
20	37.9	11	15	
21	38	14	12	
22	36.3	8	13	
23	35.1	11	15	
24	34.4	10	16	
25	34.2	10	20	
26	32.3	25	15	
27	32.3	17	11	
28	30	19	20	
29	23.8	54	20	
30	22.7	16	11	
31	16	11	20	
Total	1690.9	783		3
Avg.	54.54516	25.25806	14	
Dens.	0.463067			

Table 6 Summary Statistics of 20 Linkage Groups of the F_2 Population

Group	Length	Markers	LOD	Excluded
1	135.3	19	20	3
2	115.9	12	18	
3	106.4	11	20	2
4	93.3	15	15	
5	92.5	13	20	3
6	92.3	9	11	
7	80.3	11	13	1
8	79.8	13	20	1
9	79.3	9	20	
10	73.6	8	16	
11	72.7	8	11	
12	64.1	7	11	
13	62.3	8	15	1
14	54.8	5	20	1
15	54.3	7	20	
16	51	11	20	
17	50.9	8	13	
18	48.7	6	20	
19	34	9	18	
20	30	9	20	
Total	1471.5	198		12
Avg.	73.575	9.9	17	
Dens.	0.134557			

FIGURES

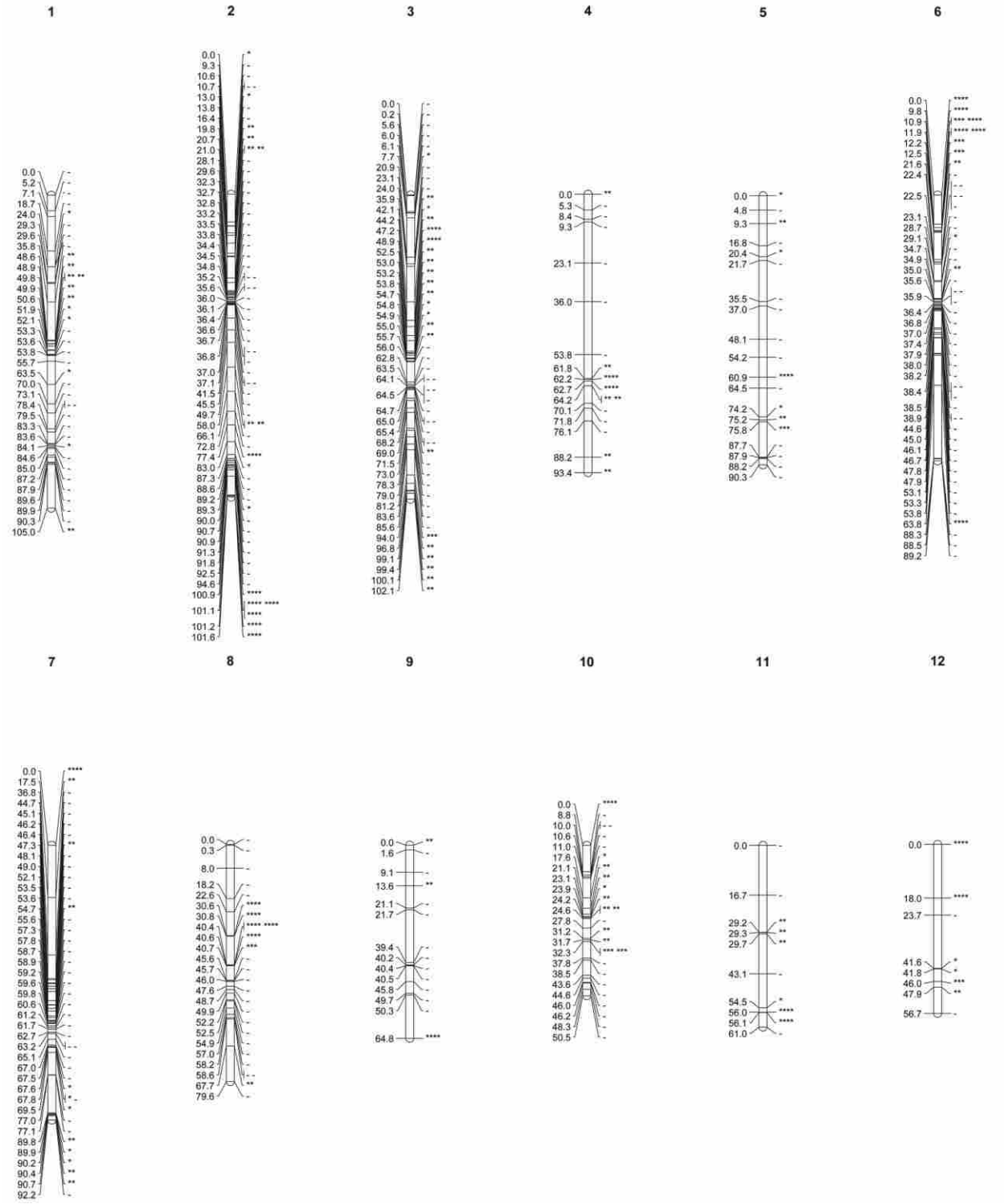

Figure continued on the next page.

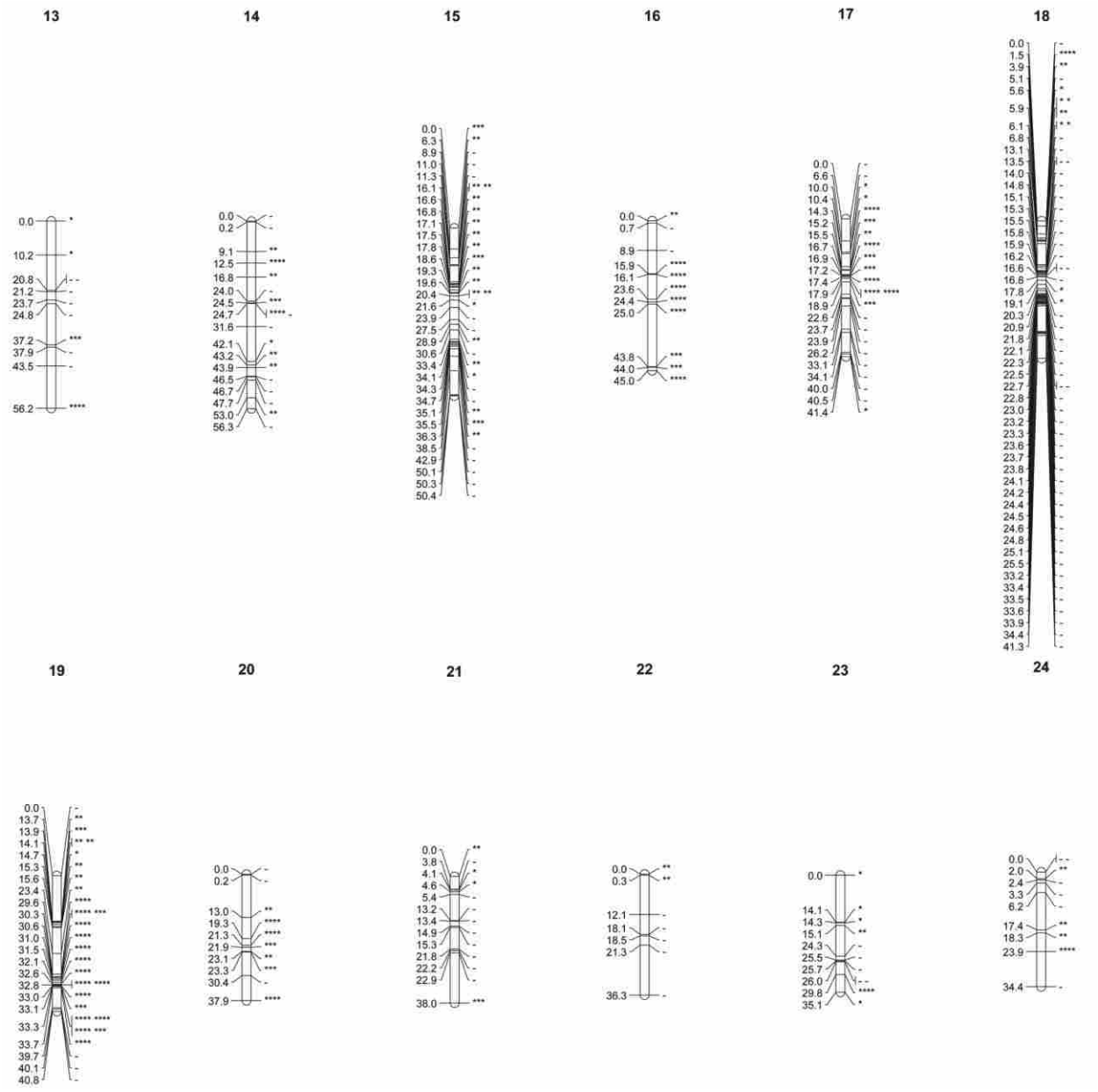

Figure continued on the next page.

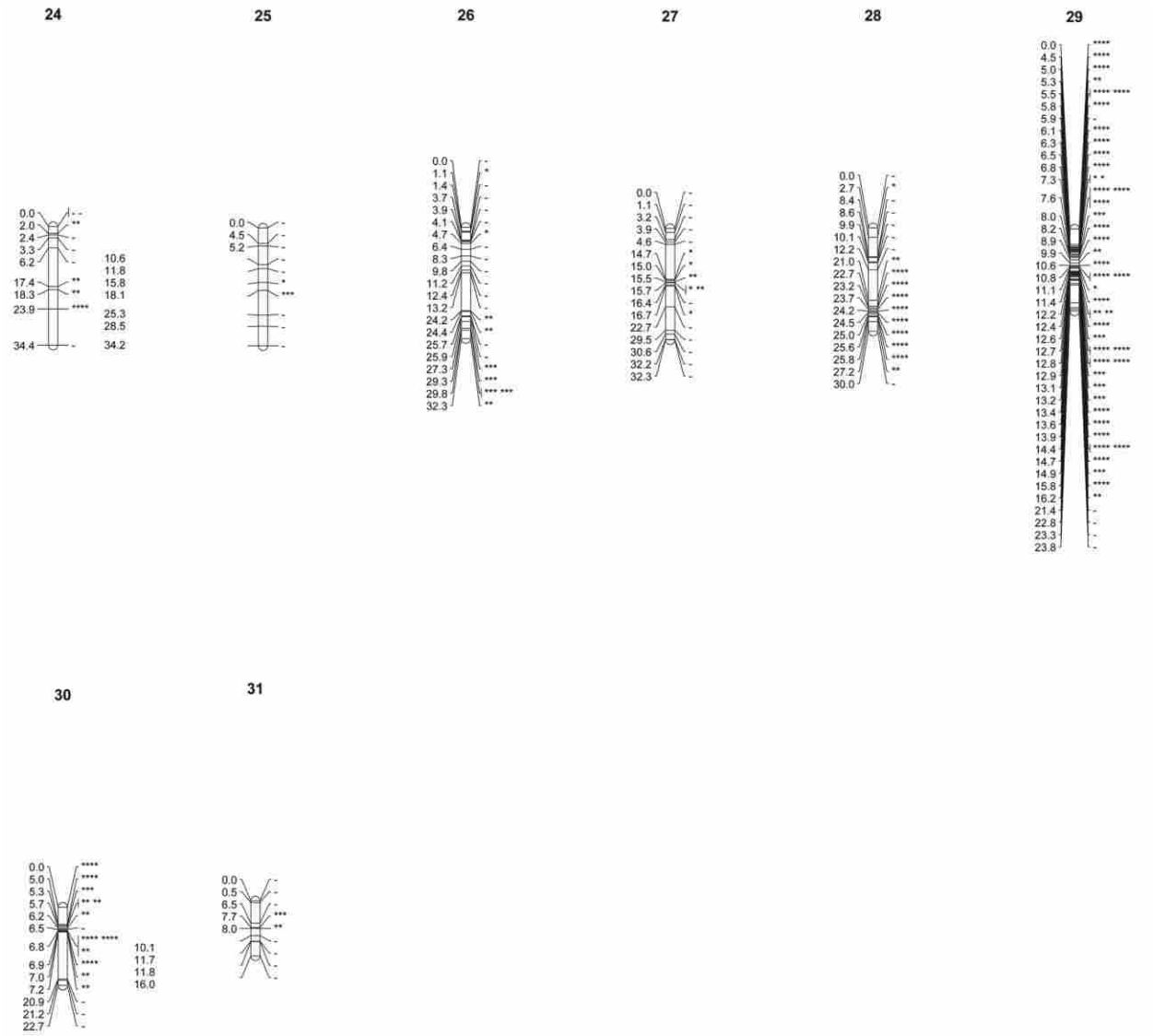

Figure 5. 31 linkage groups formed for the expected 26 linkage groups of haploid yellow lupine. The significance of segregation distortion is marked at each loci. Where - = (p>0.05), * = (p<0.01), ** = (p<0.001) and * = (p<0.00001).**

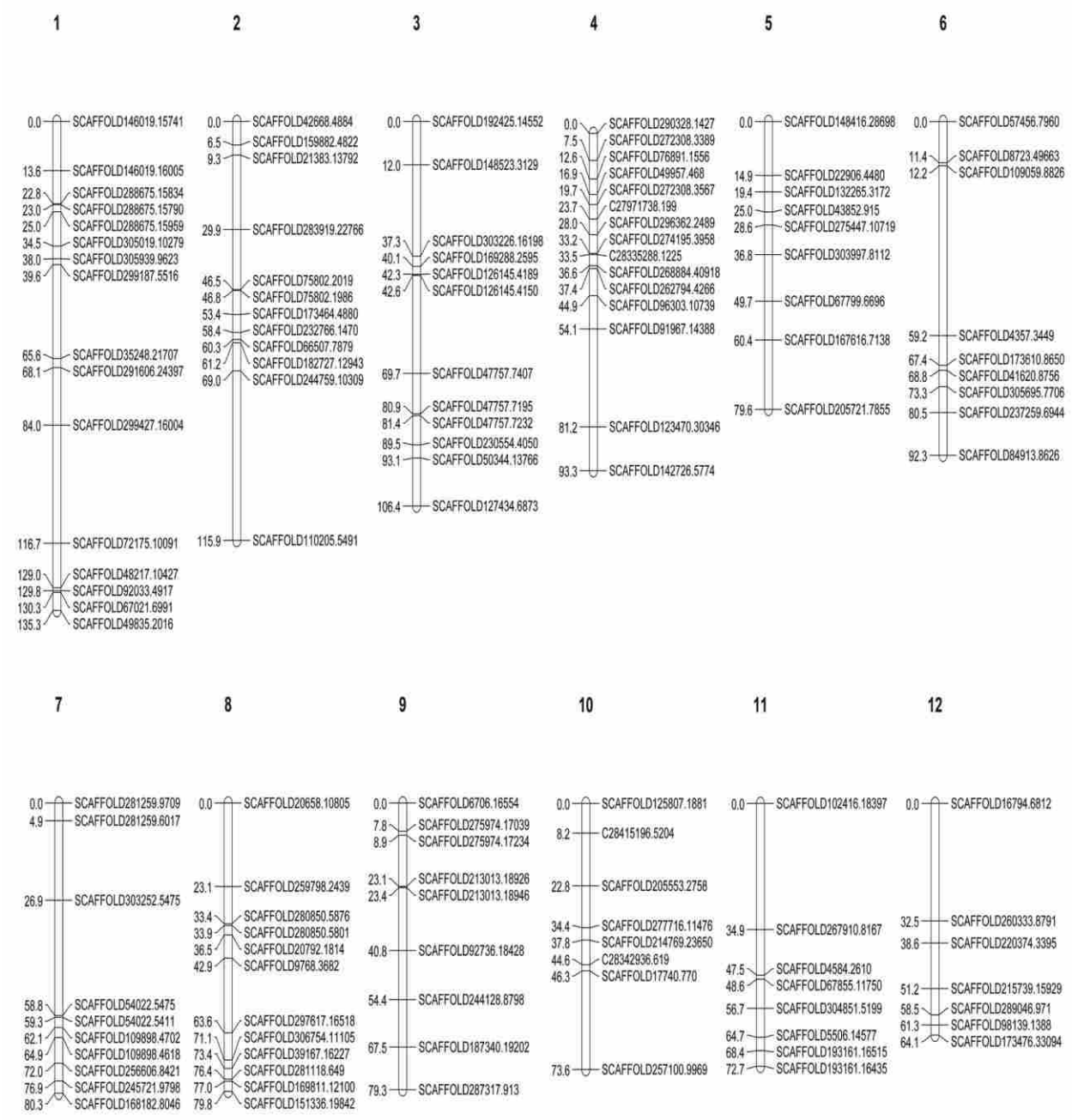

Figure continued on the next page.

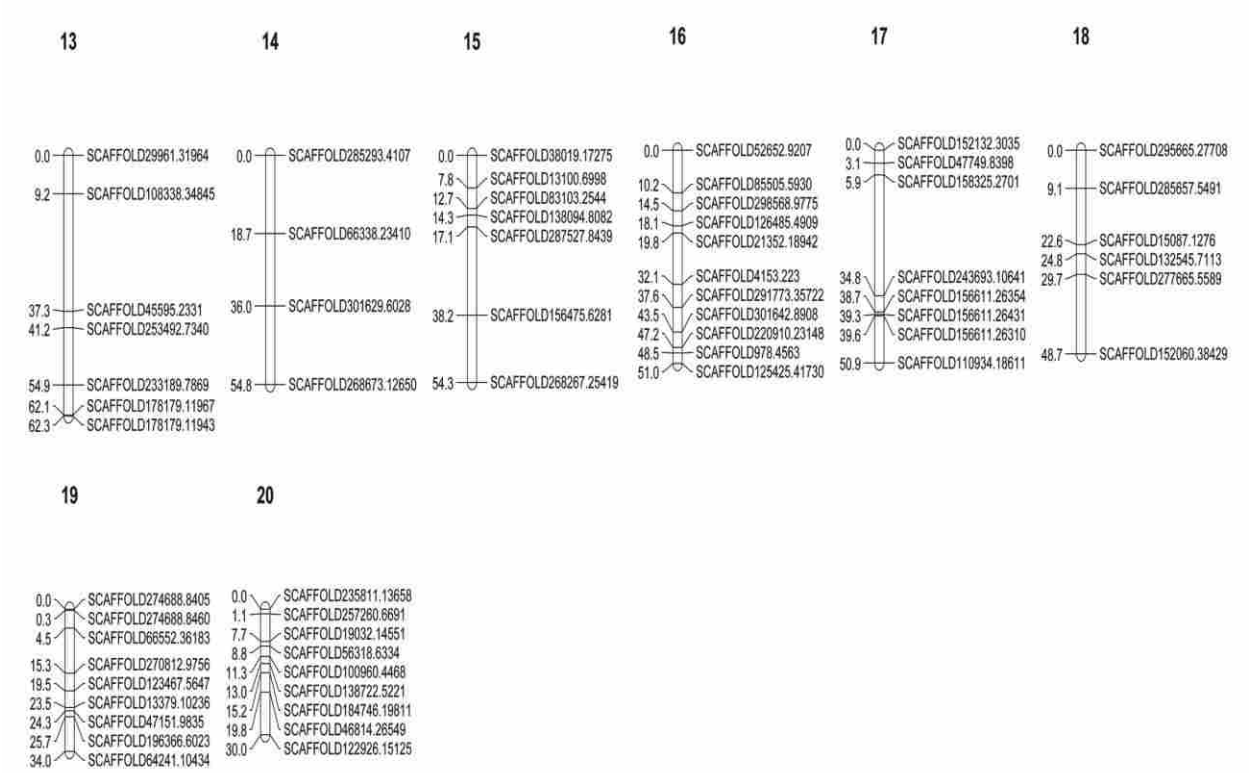

Figure 6. 20 linkage groups from the F_2 formed for the expected 26 linkage groups of haploid yellow lupine.

SUPPLEMENTAL DATA

RIL				F_2			
name	mapped	total	% mapped	name	Mapped	total	% mapped
100_.bam	2588984	3104002	0.834079359	9242X1	4120859	5317643	0.774941
101_.bam	3198422	3795474	0.842693693	9242X3	4055243	5186465	0.78189
102_.bam	3274057	4451564	0.735484652	CO_001_	1440899	2035213	0.707984
103_.bam	3724375	4837804	0.769848262	CO_002_	1038615	1585765	0.654961
104_.bam	2367262	2640358	0.896568571	CO_003_	787890	1208230	0.652103
105_.bam	3153891	3977438	0.792945358	CO_004_	573570	893870	0.64167
106_.bam	2747201	3286430	0.835922566	CO_005_	1223213	1705054	0.717404
107_.bam	3442120	3885926	0.885791443	CO_006_	912472	1405965	0.649001
108_.bam	3119592	3575744	0.872431583	CO_007_	668792	1053421	0.634876
109_.bam	5331134	6626154	0.804559327	CO_008_	717795	1121466	0.640051
10_.bam	1618736	2523628	0.641432097	CO_009_	1165265	1813456	0.642566
110_.bam	3179065	3806260	0.835220137	CO_010_	1405665	2155925	0.652001
111_.bam	3280443	4222580	0.776881196	CO_011_	1260133	1914208	0.658305
112_.bam	2739325	3455484	0.792747123	CO_012_	1229599	1912886	0.642798
113_.bam	2487202	2905840	0.855932192	CO_013_	459694	738339	0.622606
114_.bam	4704045	5585538	0.842182973	CO_014_	1154200	1716161	0.672548
115_.bam	2392567	2865414	0.834981263	CO_015_	2374821	3544220	0.670055
116_.bam	1857029	2084296	0.890962224	CO_016_	1575241	2357858	0.668081
117_.bam	3869997	4625364	0.836690258	CO_017_	1079737	1594099	0.677334
118_.bam	1426969	1801652	0.792033645	CO_018_	1181522	1782401	0.662882
119_.bam	3104954	3932518	0.789558751	CO_019_	2190008	3261586	0.671455
11_.bam	3335644	4885174	0.68280966	CO_020_	2849594	4191274	0.679887
120_.bam	1216311	1892306	0.64276655	CO_021_	2422403	3575338	0.677531
121_.bam	2704703	3357436	0.80558587	CO_022_	2995589	4400561	0.680729
122_.bam	3235619	3536278	0.914978687	CO_023_	1646158	2490638	0.660938
123_.bam	2719874	3495126	0.778190543	CO_024_	1851982	2733788	0.677442
124_.bam	3831314	4451728	0.860635241	CO_025_	1228053	1864625	0.658606
125_.bam	3654306	4623654	0.790350229	CO_026_	1420658	2127943	0.66762
126_.bam	4195274	5765742	0.727620834	CO_027_	1197026	1749311	0.684284
127_.bam	4718747	5478934	0.861252755	CO_028_	515046	856940	0.601029
128_.bam	3066404	3522058	0.870628479	CO_029_	650214	980155	0.663379
129_.bam	3499587	3926794	0.891207178	CO_030_	923803	1407686	0.656256
12_.bam	1805902	2274946	0.793821919	CO_031_	1540161	2319120	0.664114
130_.bam	3824031	4369388	0.875186868	CO_032_	265166	454504	0.583418
131_.bam	3984721	4447226	0.896001462	CO_033_	1730435	2446273	0.707376
132_.bam	2946606	3358790	0.877281997	CO_034_	2148281	3221824	0.66679
133_.bam	4042528	4700770	0.859971451	CO_035_	2187767	3261552	0.670775
134_.bam	3083351	3758800	0.820301958	CO_036_	364576	517100	0.70504

135_.bam	2200988	2737128	0.804123154	CO_037_	543427	829600	0.655047
136_.bam	1341480	1775274	0.755646734	CO_038_	1114868	1679368	0.663862
137_.bam	1313270	1530638	0.85798863	CO_039_	2804387	4110809	0.682198
138_.bam	2818860	3530602	0.79840775	CO_040_	2191957	3215522	0.68168
139_.bam	2549596	2922392	0.872434636	CO_041_	1787941	2602478	0.687015
13_.bam	4307512	6094524	0.706783992	CO_042_	2199673	3279168	0.670802
140_.bam	712423	991334	0.718650828	CO_043_	1986722	2894617	0.686351
141_.bam	2437509	3173540	0.768072563	CO_044_	1680523	2537499	0.662275
142_.bam	3270583	3752434	0.871589747	CO_045_	1309769	1952529	0.670806
143_.bam	2542664	3322492	0.765288223	CO_046_	1769503	2620915	0.675147
144_.bam	1976166	2206292	0.895695583	CO_047_	2062383	3086928	0.668102
145_.bam	2727828	3167876	0.861090522	CO_048_	1927514	2950622	0.653257
146_.bam	3053335	3600270	0.848085005	CO_049_	452555	724627	0.624535
147_.bam	2573511	2946868	0.873303792	CO_050_	725255	1141386	0.635416
148_.bam	2283832	2643804	0.863843159	CO_051_	1100704	1699387	0.647706
149_.bam	3149446	3961626	0.794988219	CO_052_	2208171	3289959	0.671185
14_.bam	2551643	4128090	0.618117095	CO_053_	526928	826214	0.637762
150_.bam	2336992	2646146	0.8831682	CO_054_	801913	1252402	0.6403
151_.bam	4441276	5048038	0.87980241	CO_055_	2066568	3155587	0.654892
152_.bam	2254817	2640460	0.853948554	CO_056_	2500714	3661945	0.682892
153_.bam	3119444	4108764	0.759217127	CO_057_	2244811	3339589	0.672182
154_.bam	3547541	4661704	0.760996623	CO_058_	685636	1072537	0.639266
155_.bam	2934527	3355094	0.874648221	CO_059_	2143733	3155760	0.679308
156_.bam	1293905	1617662	0.799861158	CO_060_	500726	825473	0.606593
157_.bam	4394384	5624444	0.781301049	CO_061_	1109824	1714222	0.647421
15_.bam	4846986	6727012	0.720525844	CO_062_	2020617	3026901	0.667553
16_.bam	1715739	2287102	0.750180359	CO_063_	1741617	2632138	0.661674
17_.bam	3354606	5197718	0.64539977	CO_064_	1756274	2645910	0.663769
18_.bam	3574770	5257928	0.679881885	CO_065_	1988344	2917069	0.681624
19_.bam	3275092	4431988	0.738966802	CO_066_	2069728	3089408	0.669943
1.bam	3707025	5474696	0.677119789	CO_067_	2278606	3378507	0.674442
20_.bam	4348673	6112456	0.711444467	CO_068_	2002547	2868522	0.698111
21_.bam	4260427	5544120	0.768458655	CO_069_	2428287	3560715	0.681966
22_.bam	3880970	5781412	0.671284108	CO_070_	1484208	2182733	0.679977
23_.bam	3771030	5867476	0.642700541	CO_071_	500555	804579	0.622133
24_.bam	2521433	3715138	0.678691613	CO_072_	1840518	2754387	0.668213
25_.bam	3523281	5351792	0.658336684	CO_073_	1802777	2748281	0.655965
26_.bam	4007545	6816612	0.587908627	CO_074_	1691886	2421870	0.698587
27_.bam	6138042	8922116	0.687958103	CO_075_	1321818	1998235	0.661493
28_.bam	6945392	10263432	0.676712429	CO_076_	1518145	2320561	0.654215
29_.bam	4135941	5710362	0.724287007	CO_077_	2496850	3715152	0.672072
2.bam	4427597	6291040	0.703794126	CO_078_	2124292	3139545	0.676624
30_.bam	4274028	6010260	0.711121981	CO_079_	745303	1127766	0.660867
31_.bam	3724590	5528768	0.673674497	CO_080_	3804360	5636023	0.675008
32_.bam	1812661	2891804	0.62682706	CO_081_	1555789	2390626	0.650787

33_.bam	4372467	6381252	0.685205192	CO_082_	1211115	1859956	0.651153
34_.bam	4782858	6753880	0.708164492	CO_083_	1649716	2502783	0.659153
35_.bam	3796262	5072476	0.748404132	CO_084_	1138125	1747749	0.651195
36_.bam	6097852	9554372	0.638226353	CO_085_	789115	1255202	0.628676
37_.bam	2172734	3328392	0.652787893	CO_086_	1731212	2662181	0.650298
38_.bam	5746471	8466086	0.678763599	CO_087_	820504	1225462	0.669547
39_.bam	4546174	6754104	0.673098016	CO_088_	1241528	1836318	0.676096
3_.bam	3311067	4784916	0.691980173	CO_089_	1152981	1733995	0.664928
40_.bam	3859352	5503748	0.701222512	CO_090_	976055	1463509	0.666928
41_.bam	3174870	4813802	0.659534813	CO_091_	805350	1266605	0.635834
42_.bam	3601227	5229270	0.688667252	CO_092_	196661	360313	0.545806
43_.bam	4314164	6519944	0.661687278	CO_093_	177339	252839	0.701391
44_.bam	3912647	5822662	0.671968766	CO_094_	935883	1372008	0.682126
45_.bam	6926729	9685780	0.71514416	CO_095_	1305603	1929724	0.676575
46_.bam	5962596	8310982	0.717435798	CO_096_	1147420	1741233	0.65897
47_.bam	4227664	6085946	0.694660124	CO_097_	2264579	3299782	0.686281
48_.bam	4204205	6048686	0.695060878	CO_098_	703827	1010202	0.696719
49_.bam	3862637	5804750	0.665426935	CO_099_	1014868	1551236	0.654232
4_.bam	3809826	5515496	0.690749481	CO_100_	1533216	2297400	0.66737
50_.bam	3846299	5057254	0.760550884	CO_101_	1711393	2508085	0.68235
51_.bam	2044390	3414924	0.598663396	CO_102_	2082620	3084512	0.675186
52_.bam	3800767	6235150	0.609571061	CO_103_	1990490	2965307	0.671259
53_.bam	2362777	3595424	0.657162271	CO_104_	645855	995744	0.648616
54_.bam	3282537	4490660	0.730969835	CO_105_	1009257	1533662	0.65807
55_.bam	3836455	5721092	0.670580896	CO_106_	2294229	3347434	0.685369
56_.bam	2351138	3546414	0.662962079	CO_107_	800057	1222590	0.654395
57_.bam	5774322	8340378	0.692333369	CO_108_	1644329	2510032	0.655103
58_.bam	2245337	3494158	0.642597444	CO_109_	1938827	2924804	0.662891
59_.bam	1951707	3341134	0.584145084	CO_110_	2280623	3286808	0.693872
5_.bam	2641341	4346978	0.607626954	CO_111_	1685615	2485448	0.678194
60_.bam	4430247	6701428	0.661089995	CO_112_	2038596	3014561	0.67625
61_.bam	5178438	7948656	0.651485987	CO_113_	2178249	3266346	0.666876
62_.bam	3438900	4913830	0.699841061	CO_114_	2070873	3092139	0.669722
63_.bam	2843964	4032344	0.705288041	CO_115_	2146237	3187656	0.673296
64_.bam	4262540	6103170	0.698414103	CO_116_	2108302	3121410	0.675433
65_.bam	1437410	2424550	0.592856406	CO_117_	2234522	3251615	0.687204
66_.bam	3837386	5590974	0.686353755	CO_118_	788742	1205454	0.654311
67_.bam	3778356	5267260	0.717328554	CO_119_	1411453	2200878	0.641314
68_.bam	4332169	6240372	0.694216467	CO_120_	1170865	1704980	0.686732
69_.bam	2666760	4631058	0.575842496	CO_121_	1169383	1790902	0.652958
6_.bam	1871395	3067950	0.609982236	CO_122_	1073401	1562615	0.686926
70_.bam	5072808	8055622	0.629722696	CO_123_	981071	1467571	0.6685
71_.bam	5069634	7515838	0.674526779	CO_124_	1013377	1528591	0.662948
72_.bam	3060583	5000970	0.611997872	CO_125_	1356220	1999438	0.678301
73_.bam	2962509	5205696	0.569089897	CO_126_	966877	1407308	0.68704

74_.bam	5049472	7868758	0.641711437	CO_127_	361522	591747	0.61094
75_.bam	3942817	5723568	0.688873968	CO_128_	2364360	3527374	0.670289
76_.bam	3583420	5456736	0.656696604	CO_129_	2117668	2873005	0.737092
77_.bam	4921310	7371872	0.667579415	CO_130_	1384393	2039005	0.678955
78_.bam	5820446	9100846	0.639549993	CO_131_	994707	1492962	0.666264
79_.bam	2481504	3326520	0.745975975	CO_132_	2118340	3115672	0.679898
7_.bam	4242563	6256860	0.678065835	CO_133_	1821461	2704297	0.673543
80_.bam	2490098	3671788	0.678170417	CO_134_	1520682	2245324	0.677266
81_.bam	1369266	2045376	0.66944464	CO_135_	996599	1530333	0.65123
82_.bam	4000614	5901520	0.677895525	CO_136_	600073	893592	0.671529
83_.bam	4842468	6568560	0.737219117	CO_137_	1134254	1726915	0.656809
84_.bam	3422598	4962960	0.689628367	CO_138_	1883286	2766046	0.680859
85_.bam	3203000	4524912	0.70785907	CO_139_	1619299	2395341	0.67602
86_.bam	4195585	6365248	0.659139283	CO_140_	1805466	2679718	0.673752
87_.bam	3570877	5065532	0.704936224	CO_141_	342877	539668	0.635348
88_.bam	3866176	6234828	0.620093449	CO_142_	258868	439123	0.589511
89_.bam	3626000	5189924	0.698661483	CO_143_	1161457	1778959	0.652886
8_.bam	2685751	4149896	0.647185134	CO_144_	914113	1408271	0.649103
90_.bam	2108925	3172870	0.664674254	CO_145_	1465121	2217045	0.660844
91_.bam	5183693	7886220	0.65731022	CO_146_	1596189	2405019	0.663691
92_.bam	2672174	3841242	0.695653645	CO_147_	519324	775943	0.669281
93_.bam	1965908	3107094	0.632715972	CO_148_	1749584	2607011	0.671107
94_.bam	3927845	5827752	0.673989731	CO_149_	1892149	2840372	0.666162
95_.bam	2919050	4508488	0.647456531	CO_150_	1411876	2240311	0.630214
96_.bam	3783632	5431466	0.696613401	CO_151_	1168683	1788507	0.653441
97_.bam	2014771	2520780	0.79926491	CO_152_	2060121	3033751	0.679067
98_.bam	2677554	3284118	0.815303835	CO_153_	506585	812436	0.623538
99_.bam	2723421	3084016	0.883076158	CO_154_	1085318	1671657	0.649247
9_.bam	2961226	4268928	0.693669699	CO_155_	1805316	2696380	0.669533
Total	535165787	7.44E+08		CO_156_	1290970	1951761	0.661439
Average	6774250.5	9413213	0.733832277	CO_157_	2145272	3161363	0.678591
				CO_158_	1790229	2610067	0.685894
				CO_159_	953940	1463489	0.651826
				CO_160_	1989860	2929993	0.679135
				CO_161_	990487	1444491	0.6857
				CO_162_	1379209	2048235	0.673365
				CO_163_	1617934	2313171	0.699444
				CO_164_	1997397	2894748	0.690007
				CO_165_	1742786	2547448	0.68413
				CO_166_	1120296	1758016	0.63725
				CO_167_	1497459	2133943	0.701733
				CO_168_	1367511	2060235	0.663765
				CO_169_	1815599	2726972	0.665793
				CO_170_	1400089	2121864	0.659839
				CO_171_	1995097	5403733	0.369207

CO_172_	1489626	2190526	0.680031
CO_181_	2955153	4346567	0.679882
CO_182_	1520161	2287485	0.664556
CO_189_	1291568	2017714	0.640115
CO_190_	956399	1519504	0.629415
CO_191_	2202712	3235607	0.680772
CO_192_	715416	1139285	0.627952
CO_193_	547434	874386	0.626078
CO_194_	1143911	1644693	0.695516
CO_195_	1070608	1601047	0.668692
CO_196_	1171617	1763279	0.664454
CO_197_	719449	1093228	0.658096
CO_198_	2801371	4162079	0.67307
CO_199_	2631266	3912078	0.672601
CO_200_	1395917	1789262	0.780164
Total	280123865	4.19E+08	
Average	1490020.6	2228178	0.66449

Table S2. Mapping Results for RIL and F_2 Populations. The name of each line, total number of reads, total number of mapped reads and mapping percent for both populations of yellow lupine.

Demultiplex of Reads

java demultiplexGBS_se cornellbarcodes8.txt C4WMNACXX.8.fastq

java demultiplexGBS_se cornellbarcodes8.txt C4WMNACXX.8.fastq

Read Counts

for i in $(ls *.fastq | sed "s/.fastq//g"); do echo $i.fastq >> names.txt;done;

for i in $(ls *.fastq | sed "s/.fastq//g"); do cat $i.fastq | grep -c "@HWI" >> counts.txt;done;

paste names.txt counts.txt > readcounts.txt

Trim Read Qualities

for i in $(ls *.1.fastq | sed "s/.1.fastq//g"); do sickle se -t sanger -f $i.1.fastq -o $i.1t.fastq; done;

Build a reference of Parent 9242X4

gmap_build –D ./ -d myIndex 92424X4.scafSeq

Map the reads using GSNAP

for i in $(ls *.t.fastq | sed 's/.t.fastq//g'); do export BASE=$i; sbatch gsnap.sh; done;

ALSO

gsnap -n1 -Q -B4 -D ./ -d myIndex -A sam ./9242X3.fasta > ./samfiles/9242X3.sam

Count the % Mapped reads

for i in $(ls *.bam | sed 's/.bam//g'); do samtools flagstat $i.bam | sed -n 5p | awk ‘{print $1}’; done;

Use InterSnp to call SNPs…

/fslhome/jtpage/bambam/bin/interSnp -t 16 -m 2 ./9242X3.bam ./CO_304.bam *.bam > interSnpIvan2.txt

Use Pebbles for imputation and collapsing …

/fslhome/jtpage/bambam/bin/pebbles -k 10 -W 100 -m 1 -f .01 -F .9 interSnpIvanedit.txt > pebblesfilter.txt

Reformat the SNPs to JoinMap format (Snp2joinmap)

/fslhome/jtpage/bambam/scripts/snp2joinmap.pl pebblesfilter.txt > pebblesfilterjoinmap.txt

Condense duplicate markers

/fslhome/jtpage/bambam/scripts/condenseMarker.pl pebblesfilterjoinmap.txt > pebblesfiltercondense.txt

List S3. GBS Pipeline Commands. Commands for running the downstream processing of GBS reads.