Transcriptome sequencing of different narrow-leafed lupin tissue types provides a comprehensive uni-gene assembly and extensive gene-based molecular markers

Lars G. Kamphuis1,2,†, James K. Hane1,‡, Matthew N. Nelson2,¶, Lingling Gao1, Craig A. Atkins3 and Karam B. Singh1,2,*  

1CSIRO Plant Industry, Wembley, WA, Australia  
2The UWA Institute of Agriculture, University of Western Australia, Crawley, WA, Australia  
3The School of Plant Biology, University of Western Australia, Crawley, WA, Australia

Summary
Narrow-leafed lupin (NLL; Lupinus angustifolius L.) is an important grain legume crop that is valuable for sustainable farming and is becoming recognized as a human health food. NLL breeding is directed at improving grain production, disease resistance, drought tolerance and health benefits. However, genetic and genomic studies have been hindered by a lack of extensive genomic resources for the species. Here, the generation, de novo assembly and annotation of transcriptome datasets derived from five different NLL tissue types of the reference accession cv. Tanjil are described. The Tanjil transcriptome was compared to transcriptomes of an early domesticated cv. Unicrop, a wild accession P27255, as well as accession 83A:476, together being the founding parents of two recombinant inbred line (RIL) populations. In silico predictions for transcriptome-derived gene-based length and SNP polymorphic markers were conducted and corroborated using a survey assembly sequence for NLL cv. Tanjil. This yielded extensive indel and SNP polymorphic markers for the two RIL populations. A total of 335 transcriptome-derived markers and 66 BAC-end sequence-derived markers were evaluated, and 275 polymorphic markers were selected to genotype the reference NLL 83A:476 × P27255 RIL population. This significantly improved the completeness, marker density and quality of the reference NLL genetic map.

Keywords: lupin, indel, SNP, molecular marker, next-generation sequencing.

Introduction
Narrow-leafed lupin (NLL; Lupinus angustifolius L.) is a recently domesticated grain legume crop from the Mediterranean region (Cowling et al., 1998). NLL is a diploid (2n = 40 chromosomes), annual, self-pollinating crop with a haploid genome size of 924 Mb (Naganowska et al., 2003). NLL is well adapted to nitrogen- and phosphorus-deficient, acid, sandy soils and is therefore a valuable break crop in the fertile acid soils in the Mediterranean-like climates of Western Australia, Chile and South Africa as well as in cool temperate regions of Europe including Poland, Ukraine, Belarus and the Russian Federation (Wolko et al., 2011). Lupins have the ability to fix atmospheric nitrogen in symbiosis with Bradyrhizobia and are exceptionally effective in accessing soil phosphorus (Cheng et al., 2011; Lambers et al., 2013), making a valuable contribution to the sustainable management of infertile agricultural soils. Other rotational benefits are in the management of diseases and pests, as well as control of grass weeds (Seymour et al., 2011). Despite these multiple benefits, profitability of the crop has yet to reach its full potential and has declined in recent years, in part due to low prices and its primary use as animal feed (Berger et al., 2012b). However, NLL shows promise for wider use in aquaculture (Glencross et al., 2003) and is gaining recognition as a potential human health food. Its grain is high in protein and dietary fibre, gluten-free and low in fat and starch and thus has a very low glycaemic index (Johnson et al., 2003). Moreover, lupin seeds contain constituents that alter satiety (Lee et al., 2006) and some conglutins in the seed are able to bind insulin potentially providing a pharmaceutical role similar to the hypoglycaemic drug metformin (Magni et al., 2004); studies have associated lupins playing a positive role in important areas of health including combating obesity, diabetes and cardiovascular disease (Belski et al., 2010; Duranti and Morazzoni, 2011).

The genetic base of Australian and European NLL is very narrow compared with the broad genetic diversity available in wild germplasm collections, and this limits the yield and adaptive potential of the crop (Berger et al., 2012a). Introducing improvements of agronomically important and human health benefit-related traits from phenotypic and genetic screening of a diverse range of wild NLL material (Berger et al., 2013) represents an underutilized resource. To achieve this in an efficient manner, there is a need to develop a suite of genetic and genomic resources for NLL.

While some genomic resources are currently available for the genus Lupinus, these are far fewer than have been established in major legume genomic efforts such as those of Glycine max, Medicago truncatula and Lotus japonicus (reviewed in Kasprzak et al., 2006; Young and Bharti, 2012). Lupinus-specific genomic resources currently include: 14 605 short genome survey sequences totalling 14 455 for L. angustifolius and 150 for L. albus; expressed sequence tags (ESTs) totalling 9140 for L. albus, 416 for L. luteus and 388 for L. angustifolius; and 2.5 Gb of Roche 454 mRNA transcriptome data for L. luteus (Parra-
are to date no comprehensive transcriptomic datasets available for L. angustifolius; however, bacterial artificial chromosome (BAC) libraries for cultivars Sonet (Kasprzak et al., 2006) and Tanjil (Gao et al., 2011) have been developed and have led to early insights into the composition and organization of the NLL genome (Gao et al., 2011; Książkiewicz et al., 2013). The cv. Sonet BAC library comprises 55,296 clones with an average insert size of 100 kb, which is equivalent to 6x coverage of the NLL (haploid) genome. The cv. Tanjil BAC library comprises 111,360 clones with an average insert size of 99.7 kb, which represents an estimated 12X coverage of the NLL genome. Both ends of 9,600 randomly selected clones of the Tanjil BAC library were sequenced, producing 13,985 BAC-end sequences equivalent to ~1% of the NLL genome. In addition to these publicly available genomic resources for NLL, a draft genome has been recently reported (Yang et al., 2013b).

In addition to the genomic resources above, several studies have generated increasingly dense genetic maps for NLL, using a range of different molecular marker types in mainly two recombinant inbred line (RIL) populations: Tanjil × Unicrop and 83A:476 × P27255 (Boersma et al., 2005; Nelson et al., 2006, 2010; Yang et al., 2013b). Most research efforts have focused on the 83A:476 × P27255 RIL population, a wide cross, which has been adopted as the reference population by the research community. Also known as the ‘domesticated’ × ‘wild’ (DxW) RIL population, the most recent reference map release comprised a total of 1207 high-quality loci spread evenly over 20 linkage groups and three small clusters, including 708 locus-specific markers, 492 multilocus microsatellite fragment length polymorphism (MFLP) markers and seven single-gene trait loci (Kroc et al., 2014). A genetic map based on the narrower Tanjil × Unicrop RIL population was recently developed by Yang et al. (2013b) and comprised 8244 restriction site-associated DNA (RAD) markers. While this Tanjil × Unicrop map had a higher average marker density than the 83A:476 × P27255 reference map, there was a considerable marker distribution bias in the Tanjil × Unicrop map, possibly arising from extensive regions of low genomic diversity between the closely related founding parents of the Tanjil × Unicrop RIL population.

Next-generation sequencing technologies have made it possible to sample deeply from the mRNA transcriptome using RNA-Seq. Generating in-depth transcriptome datasets allows accelerated gene discovery in ‘genomic orphan’ species without the great expense and complexity of genome sequencing and assembly (Varshney et al., 2009). This approach can be used to identify gene exome sequences, highlight differential gene expression, determine phylogenetic relationships and facilitate the development of various types of gene-based molecular markers including insertion or deletion (indel) markers and single nucleotide polymorphism (SNP) markers. This approach has been successfully applied in other legumes, including chickpea (Cicer arietinum) (Hiremath et al., 2011; Kudapa et al., 2014), pea (Pisum sativum) (Kaur et al., 2012), pigeonpea (Cajanus cajan) (Dutta et al., 2011), common bean (Phaseolus vulgaris) (Blair et al., 2011), faba bean (Vicia faba) (Kaur et al., 2012), lentil (Lens culinaris) (Sharpe et al., 2013), peanut (Arachis hypogaea) (Zhang et al., 2012), soybean (Glycine max) (Deschamps et al., 2010), Bituminaria bituminosa (Pazos-Navarro et al., 2011) and yellow lupin (L. luteus) (Parra-González et al., 2012). It is an effective means of marker development as it targets nucleotide diversity within genic regions, allowing the possibility of generating ‘perfect’ diagnostic markers that reside within genes responsible for important traits relevant to breeding programmes.

In this study, we describe the generation, de novo assembly and annotation of transcriptome datasets derived from five different tissue types of the domesticated narrow-leaved lupin cv. Tanjil, which allowed tissue-specific expression patterns to be identified. The Tanjil transcriptome was compared to transcriptomes of cv. Unicrop and the wild accession P27255. Unicrop was the first fully domesticated Australian cultivar (with low alkaloid content, nonshattering pods, permeable seeds and early flowering) and was released in 1973 (Berger et al., 2012b). The cultivar Tanjil was released in 1998 and was a dominant variety for many years owing to its superior disease-resistance characteristics and has been chosen as the reference genotype for NLL (Berger et al., 2013; Gao et al., 2011). The Moroccan accession P27255 was selected as a representative wild accession as well as being a founding parent of the reference 83A:476 × P27255 RIL population. A draft genome survey assembly sequence for NLL cv. Tanjil was used to corroborate transcriptome-derived in silico predictions for gene-based length and single nucleotide (SNP) polymorphic markers. This yielded extensive indel and SNP polymorphic markers for the RIL populations developed in NLL. Transcriptome- and BAC-end sequence-derived simple-sequence repeat (SSR) markers (Gao et al., 2011) were evaluated, and polymorphic markers were used to genotype the reference RIL population, which has significantly improved the completeness, marker density and quality of the reference NLL genetic map.

Results

Transcriptome assembly of Lupinus angustifolius cv. Tanjil

Three NLL accessions were selected for in-depth transcriptome sequencing: the cultivars Tanjil and Unicrop and the wild accession P27255. RNA was isolated from five different tissue types (roots, stem, leaf, flower and seed) when individual plants started to set seed on the main stem inflorescence. Seed samples were mixed stages after anthesis ranging from early, mid, late to fully mature seeds. In addition, more limited transcriptome sequencing of 83A:476, a parent of one of the RIL population, was also conducted for leaf tissue as part of the 1000 Plants initiative (www.onekp.com). Transcriptome sequencing was performed by the Beijing Genome Institute (BGI) and resulted in quality-controlled RNASeq data ranging from 1.7 to 2.1 Gb per library (Figure S1). The Tanjil reference transcriptome was assembled into 104,766 sequences with a total length of 51.6 Mb (Table 1). Trinity transcriptome assembly predicted a total of 63,271 discrete loci, 53,761 of which had a single isoform (9150 had two or more isoforms). Herein, the complete set of 104,766 sequences is referred to as the Tanjil transcriptome assembly (Table 1, Figure S2).

Tissue-specific gene expression and function

The RNASeq data were interrogated to identify highly significant patterns of differential expression for transcripts in cv. Tanjil, between root, stem, leaf, flower and seed tissues that may indicate a strong association with a single tissue. Although as many as 10,240 transcripts (Trinity components) were constitutively expressed, with no significant differences in expression between all tissues, there was only a remarkably small number with highly specific expression limited to a single tissue type
(Table 2, Figure 1). There were no transcripts specific to stem or leaf tissues as these two RNAseq datasets showed highly similar expression patterns, whereas 28 were specific to flower tissue, 17 to seeds and 7 to roots. There were an additional two transcripts that were specific to stem and flower tissues combined and a single transcript specific to root and seed tissues combined.

The cv. Tanjil reference transcriptome was annotated with predicted gene functions by InterproScan, which was used to cross-reference with tissue-specific transcripts (Table S1). Examination of the Interpro terms associated with these tissue-specific genes depicted a series of biological processes consistent with the roles of each respective tissue. For example, specific expression in seed tissues was overwhelmingly the greatest overall of all tissues, with the majority belonging to seed-storage and lipid-transfer proteins. The flower tissues are sites of major morphological development and change, and transcripts encoding a high number of flower-specific degradative enzymes (in particular, pectinesterases) as well as enzymes involved in growth and development were identified. Root-specific transcripts included a major facilitator superfamily (MFS) membrane transporter that is likely to be involved in nutrient or ion import and two unknown proteins, whereas transcripts specific to either the root or the seed included those with predicted peroxidase and anthocyanin biosynthesis functions.

**Draft genome survey assembly of*** Lupinus angustifolius **cv. Tanjil**

A draft genome assembly for Tanjil was generated using 2 × 100 bp paired-end reads (insert sizes 170 and 500 bp) at approximately 25X haploid genome coverage. Total genome size was estimated by k-mer analysis of the raw data to be 1 036 553 800 bp. K-mer analysis also indicated negligible

| mRNA expression specificity | Count of Trinity components | Components with candidate markers | With functional annotation |
|-----------------------------|----------------------------|----------------------------------|---------------------------|
| Standard SNP: PxU only      | 1471                       | 1                                |
| Standard SNP: PdxW only     | 4736                       | 1                                |
| Standard SNP: PdxU + PdxW   | 1769                       | 1                                |
| Standard Length: PdxW only  | 194                        | 1                                |
| Standard Length: PdxU only  | 752                        | 1                                |
| Standard Length: PdxU + PdxW| 208                        | 1                                |
| Fluidigm SNP: PxU only      | 411                        | 1                                |
| Fluidigm SNP: PdxW only     | 1667                       | 1                                |
| Fluidigm SNP: PxU + PdxW    | 289                        | 1                                |

| Not conforming to above categories | Components with candidate markers | With functional annotation |
|-----------------------------------|----------------------------------|---------------------------|
| pxu                               | 52 976                           | 43                        |
| DdxW                              | 24 057                           | 12 853                    |
| PdxU                              | 4092                             | 462                       |
| PdxU + PdxW                       | 1730                             | 413                       |
| PdxU only                         | 413                              | 5845                      |
| PdxU + PdxW only                  | 1065                             | 1065                      |

© 2014 CSIRO. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 13, 14-25
heterozygosity and relatively low levels of repetitive DNA (Figure S4). The draft genome survey sequence was assembled into a total length of 521.2 Mb (Table 3).

Identification and distribution of polymorphism between NLL cultivars relative to a published genetic map

Reads of the remaining three parents (Unicrop, P27255 and 83A:476) of the two NLL RIL populations were aligned to the reference Tanjil transcriptome assembly in order to identify sequence polymorphisms. These included the datasets generated herein for Unicrop and P27255, as well as reads from 83A:476 leaf transcriptome sequencing by the 1000 Plant initiative. This identified a total of 2,076,890 and 101,853 conservatively filtered SNPs and insertion–deletion (indel) polymorphisms, respectively, across the four NLL lines. The wild cultivar P27255 was more polymorphic relative to 83A:476 than the more closely related cultivars Tanjil and Unicrop (Figure S3). Sequence-based markers from the most recent reference genetic map (Kroc et al., 2014) along with indels and SNPs identified here were matched by sequence similarity to the genomic scaffolds of the Tanjil survey assembly. This examination allowed the inference of the overall distribution of polymorphisms across the NLL genome. We observed that SNP and indel polymorphisms were distributed relatively evenly (Figure 2) with no strong biases to specific regions of the genome.
Development and validation of gene-based NLL molecular markers

With the knowledge that the identified SNPs and indels were relatively evenly distributed across the NLL genome, a total of 63,269 standard PCR-based candidate markers were designed with predicted polymorphism in either the Tanjil × Unicrop or 83A:476 × P27255 RIL populations using stringent primer design criteria (see Experimental procedures). Of these 63,269 polymorphisms, 57,615 were SNPs (Figure 3a) and 5,654 were length polymorphisms (Figure 3b). Of these 5,654 length polymorphisms, 838 were polymorphic between both parent-pairs, of which 73 were associated with linkage groups in the reference genetic map (via matches to mappable draft genome scaffolds) (Figure 3b; Table S2). A total of 11,149 SNPs were polymorphic between both parent-pairs (Figure 3a), with 890 of these being associated with linkage groups in the 83A:476 × P27255 reference genetic map (Table S2). These 11,149 SNPs were therefore classed as robust candidates for SNP marker development. A further 13,517 bi-allelic, genic SNPs were selected for Fluidigm assay design by Millennium Science (Figure 3c). Of these, 1,648 were polymorphic between both parent-pairs, with 114 of these associated with linkage groups (Figure 3c; Table S2).

A total of 10,955 draft genome assembly scaffolds were supported by in silico PCR mapping of one or more standard PCR candidate markers, with 3,198 scaffolds supported by length polymorphism markers and 10,673 scaffolds supported by SNP markers. Fluidigm candidate SNP markers supported 6,326 draft genome scaffolds, of which 5,682 overlapped with the in silico standard PCR markers for the same SNP site. There was a marked reduction in the number of markers polymorphic between Tanjil and Unicrop, in particular with linkage group NLL-04.

| Table 3 Lupinus angustifolius cv. Tanjil draft survey assembly statistics |
|-----------------|-----------------|
|                 | Contig          | Scaffold        |
| N90             | 698,252         | 47,079          |
| N50             | 93,792          | 10,137          |
| L90             | 133 bp          | 1,837 bp        |
| L50             | 1,088 bp        | 13,762 bp       |
| Max             | 33,271 bp       | 206,256 bp      |
| Total length    | 453,133 600 bp  | 521,188 753 bp  |
| Number ≥100 bp  | 1,096,703       | 191,701         |
| Number ≥2 kb    | 43,040          | 45,210          |

Figure 2 Comparison of the distribution and density of RNAseq-derived markers to the previously published genetic map (Kroc et al., 2014), using the draft genome assembly scaffolds as a point of reference to relate to the existing genetic map. Each series of 3 black concentric rings indicate different marker types. (A) Fluidigm-compatible bi-allelic SNP markers. (B) Standard PCR indel markers. (C) Standard PCR-compatible SNP markers. Marker polymorphism relative to the cv Tanjil reference transcriptome is indicated by colour as 83A:476 × P27255 (red), Tanjil × Unicrop (green) and both Tanjil × Unicrop and 83A:476 × P27255 (blue). Normalized marker density per Mb is indicated within each ring from 0 to 100+ along the vertical axis.
To estimate the distribution and density of predicted markers across the NLL genome, SNP and indel polymorphisms between cv. Tanjil, cv. Unicrop and cv. P27255 were placed upon the reference genetic map (Kroc et al., 2014). Draft genome scaffold sequences which contained mapped sequence-based markers were used as a point of reference to assign a location on the genetic map and to estimate normalized marker density (mutations per Mbp) (Figure S3). This indicated that among mapped scaffolds with at least one marker, the estimated density of both SNP and indel mutations between the closely related cv. Tanjil and cv. Unicrop was generally in the range of 10–100/Mbp, with SNPs being more frequent than indels and SNP density increasing to 100–1000/Mbp in a few regions. The estimated density of indel mutations between the domesticated cv. Tanjil and the wild accession P27255 was also in the range of 10–100/Mbp; however, SNP density was consistently high in the range of 100–1000/Mbp. Aside from the previously mentioned reduction in polymorphism within NLL04 between cv. Tanjil and cv. Unicrop, we did not observe strong biases in marker distribution or density within any linkage groups, demonstrating that our predicted markers are relatively evenly distributed throughout the NLL genome.

Validity of the polymorphism detection pipeline was tested by polymorphism screening and then linkage mapping of selected transcriptome-derived SNP and indel markers. A set of 96 SNPs, identified in both conventional and Fluidigm pipelines, was selected for Fluidigm assay synthesis, of which 88 (92%) were confirmed to be polymorphic (Table 4). A set of 239 transcriptome-derived indel markers with predicted polymorphisms ≥2 bp were selected for Multiplex Ready PCR analysis (Gao et al., 2011; Hayden et al., 2008), of which 194 (81%) were confirmed to be polymorphic (Table 4). An additional 66 BAC-end sequence-derived simple-sequence repeat (SSR) markers were tested on the two parent-pairs of the RIL population with a 37.8% polymorphism rate (Table 4). Subsequently, a total of 275 markers, 17 BAC-end SSRs, 170 indels and 88 SNPs were successfully genotyped on the 83A:476 × P27255 RIL population (n = 152) and subjected to linkage analysis.

Linkage analysis was conducted using the 275 new markers reported in this manuscript along with 492 MFLP, 156 RFLP, 298 DARt and 247 PCR-based STS markers and seven single-gene trait loci genotyped in 112 RILs from the current reference genetic map (Kroc et al., 2014). In total, 1475 markers and 152 RILs were used for linkage analysis. This resulted in the formation of 20 NLL linkage groups and one small cluster (Cluster-2 from Kroc et al. (2014) comprising 783 skeleton, 340 redundant and 352 attached markers (Table 5). The total length of the linkage map was 2263.9 centiMorgans (cM), with an average of 2.90 cM between nonredundant skeleton loci. In the new map, two of the three small clusters presented in the 2010 and 2014 reference maps (Kroc et al., 2014; Nelson et al., 2010) have been incorporated into main linkage groups: Cluster-1 mapped to the bottom of linkage group NLL-11 and Cluster-3 to the top of NLL-15 (Figure S5). The distribution of the 275 new markers per linkage group is presented in Table 6, and the marker order for each linkage group and the genotype data for 1475 markers in 161 RILs are presented in Table S3. Linkage group NLL-04 was remarkable in only having one new marker across the entire 145.7 cM length (Tables 5, 6). Two molecular markers developed in this study are linked to important traits in the lupin breeding programme, where LalND155 is linked to anthracnose resistance.
genes or conglutins in NLL (Foley et al., 2011) had identified a total of 16 seed-storage lipid-transfer proteins. Previous work using 2395 EST sequences tissue-specific transcripts encoding putative seed-storage and expression was most pronounced in the seed, with the majority of genes that are strongly associated with certain tissues. Tissue-specific transcripts also highlighted genes and biological process RNAseq data was to facilitate gene-based marker design, an Table S1). Although the primary purpose of generating the developed, followed by genotype and tissue-type comparisons (Kosambi cM). Redundant markers mapped to the same location as skeleton markers. Attached markers mapped to approximate intervals on the genetic map.

### Table 5 Lupinus angustifolius linkage group details. Skeleton markers were used for locus ordering and interval size calculation groups (prefixed NLL) and Cluster-2

| Linkage group | Skeleton | Redundant | Attached | Total loci | Length (cM) |
|---------------|----------|-----------|----------|------------|-------------|
| NLL-01        | 58       | 24        | 0        | 122        | 158.3       |
| NLL-02        | 34       | 28        | 18       | 80         | 119.9       |
| NLL-03        | 39       | 14        | 34       | 87         | 128.1       |
| NLL-04        | 41       | 14        | 13       | 68         | 145.7       |
| NLL-05        | 43       | 21        | 13       | 77         | 119.1       |
| NLL-06        | 60       | 20        | 35       | 115        | 137.2       |
| NLL-07        | 56       | 28        | 16       | 100        | 132.3       |
| NLL-08        | 38       | 20        | 17       | 75         | 140.4       |
| NLL-09        | 26       | 10        | 20       | 56         | 103.3       |
| NLL-10        | 42       | 19        | 9        | 70         | 91.6        |
| NLL-11        | 52       | 34        | 21       | 107        | 144.7       |
| NLL-12        | 24       | 9         | 18       | 51         | 95.4        |
| NLL-13        | 34       | 9         | 8        | 51         | 104.7       |
| NLL-14        | 27       | 7         | 18       | 52         | 91.7        |
| NLL-15        | 37       | 10        | 12       | 59         | 102.2       |
| NLL-16        | 29       | 14        | 14       | 57         | 93.8        |
| NLL-17        | 35       | 30        | 16       | 81         | 102.4       |
| NLL-18        | 38       | 8         | 8        | 54         | 93.1        |
| NLL-19        | 24       | 6         | 10       | 40         | 74.2        |
| NLL-20        | 41       | 14        | 12       | 67         | 82.8        |
| Cluster-2     | 5        | 1         | 0        | 6          | 3.0         |
| Total         | 783      | 340       | 352      | 1475       | 2263.9      |

(Lanr1) and LaSSR_025 to the bitterness locus (lucundus) (Table S4).

### Discussion

This is the most comprehensive transcriptome study yet reported for a lupin species and incorporates datasets for five different tissue types across three NLL lines. These three NLL lines are highly relevant to the breeding programme as Unicrop was the first fully domesticated cultivar released in Australia, P27255 represents a wild accession that has been introduced in the breeding programme and Tanjil, a recent NLL cultivar, has been chosen by the lupin research community as the reference accession.

### Narrow-leaved lupin transcriptome characterization

A reference Tanjil transcriptome from a wide range of tissues was developed, followed by genotype and tissue-type comparisons using high-coverage short-read sequencing (Figure 1, Table 2, Table S1). Although the primary purpose of generating the RNAseq data was to facilitate gene-based marker design, an analysis of differential expression and prediction of ‘tissue-specific’ transcripts also highlighted genes and biological process that are strongly associated with certain tissues. Tissue-specific expression was most pronounced in the seed, with the majority of tissue-specific transcripts encoding putative seed-storage and lipid-transfer proteins. Previous work using 2395 EST sequences from mature seed tissue had identified a total of 16 seed-storage genes or conglutins in NLL (Foley et al., 2011), and this in-depth transcription study did not identify any additional members, perhaps reflecting the high levels of expression of these important genes in mature seeds. Also of note was a seed-specific type 2 fatty acid desaturase, which is contrasted by a type 1 fatty acid desaturase that was strongly associated with flower tissues. The type 1 desaturase is likely to synthesize oleic acid, whereas the type 2 desaturase is likely to convert other saturated fatty acids to unsaturated as part of the synthesis of stored vegetable oils. Notably, unlike most tissue-specific transcripts, the flower-specific type 1 desaturase had reasonably high baseline levels of expression in other tissues, particularly the seed.

The flower tissues are sites of major morphological development and change. This includes structural separation of whorls, internal structural changes during embryogenesis and seed formation. These processes were reflected by a high number of flower-specific degradative enzymes (in particular pectinesterases), growth and development regulators and a small number of lipid-storage proteins that may be involved in very early seed development.

A small number of transcripts were found to be specific to the combination of two tissues, such as transcripts which encode predicted peroxidase and anthocyanin biosynthesis functions found in both the root and seed tissues. As both root and seed tissues exist at some point below ground, this may hint at reactive oxygen species (ROS) providing a basal plant defence against soil pathogenic microbes and corresponding antioxidant protection of these tissues by anthocyanins. Furthermore, root, flower and seed tissues all exhibited tissue-specific expression of different PR10-like transcripts, further suggesting basal pathogen defences that may be safeguarding these tissues.

The identification of tissue-specific transcripts will aid in the design of tissue-specific expression constructs by identifying their promoter sequences in the survey assembly. Furthermore, these
data may prove to be directly useful in the detection of certain traits of interest in future breeding efforts.

**Increasing the repertoire of genic markers in NLL and improvement of the reference genetic map**

SNP and length polymorphic markers were developed for the parents of two RIL populations, using the three transcriptome datasets generated in this study as well as the transcriptome dataset for cv. 83A:476 obtained through the 1000 Plant initiative. This resulted in the identification of large new marker sets for NLL, including gene-based SNP and length (indel) polymorphic markers. Transcript-derived markers have been widely used in generating high-density linkage maps (Varshney et al., 2005) including other less studied, so-called orphan legumes (Dutta et al., 2011; Hiremath et al., 2011; Kaur et al., 2012; Parra-González et al., 2012; Pazos-Navarro et al., 2011; Sharpe et al., 2013; Zhang et al., 2012). As these markers are derived from coding parts of the genome, this database of markers could potentially contain ‘perfect markers’ for traits of interest. To date, no perfect markers for NLL have been reported, but rather tightly linked markers to traits, primarily key domestication traits identified in the 83A:476 × P27255 population and disease-resistance genes in the Tanjil × Unicrop population. These include markers tightly linked to genes controlling alkaid production and bitterness, seed water permeability, pod shattering, early flowering/vernalization, responsiveness and pigment production in the seeds (Boersma et al., 2005, 2007a,b, 2009), anthracnose resistance (Yang et al., 2004, 2013a; You et al., 2005) and Phomopsis stem blight resistance (Yang et al., 2002, 2013a). However, there remains a pressing need to increase genic marker density in relevant NLL breeding populations that segregate for key traits including yield, early vigour, drought tolerance and phenological adaptation (Berger et al., 2013).

Of the 239 selected markers identified in silico as being length polymorphic markers, 194 were confirmed as being polymorphic between the parents of both RIL populations by PCR (Table 6). Nine markers failed to generate amplicons and one had multiple amplicons. These primers either failed to amplify because they were designed of a misassembled transcript or the primer pairs were not amenable to the multiplex PCR technique used. Thirty-five markers were monomorphic by PCR for the 83A:476 × P27255 RIL population, whereas they were polymorphic for the Tanjil × Unicrop population. This may be because 83A:476 is genetically heterogeneous due to it being an F₂ derived line. In contrast, the other lines are inbred varieties (Unicrop and Tanjil) or a wild accession from an isolated population in Morocco (P27255). Nevertheless, the in silico predictions were 100% accurate for the Tanjil × Unicrop population, thus resulting in high confidence, good quality markers. Of the 96 SNP markers tested on the Fluidigm platform, 88 were successfully converted and genotyped on the 83A:476 × P27255 RIL population. Four markers had poor amplification across the population, three were polymorphic for the Tanjil × Unicrop population (but not the 83A:476 × P27255 population, for reasons discussed earlier), and one marker was heterozygous in all individuals consistent with it amplifying two distinct monomorphic loci. It thus appears that the stringent in silico generation of length and SNP markers resulted in high conversion rates (81% and 92%, respectively) of gene-based molecular markers using high-throughput genotyping platforms.

The newly generated gene-based length polymorphism and SNP markers as well as BAC-end sequence-derived SSR markers identified previously (Gao et al., 2011) were used to genotype the 83A:476 × P27255 RIL population and resulted in a significant improvement of the overall marker density of the reference genetic map. This resolved two of the three small clusters of the previous reference map, which could now be assigned to two linkage groups (Figure S5). The incorporation of Cluster-1 into the top of NLL-15 was made possible by seven new marker loci, while the incorporation of Cluster-3 into the bottom of NLL-11 was made possible by the addition of one new marker and the increased statistical power afforded by the larger population size, which increased from 112 to 161 RILs.

Markers were distributed on all 20 main linkage groups and Cluster-2 (Table 6 and Figure S5). However, some linkage groups were under-represented especially NLL-04, which had only one new marker in the entire 145.7 cM length of the linkage group. This uneven distribution may have arisen from one or more of the selection criteria used to identify SNP and indel variants in the transcriptomes used in this study, most notably the requirement to be polymorphic between parents of both RIL populations, viz the 83A:476 × P27255 population used in this study and the Tanjil × Unicrop population used by Yang et al. (2013b). The latter population is a relatively narrow cross between Australian domesticated varieties, and the recently developed map of Yang et al. (2013b) showed extremely skewed distribution of markers ranging from 23 to 1498 markers per linkage group. In contrast, markers (including previously reported markers) in the map reported here were distributed much more evenly (40 to 122 markers per linkage group). It is therefore possible that new markers in this current map were unevenly distributed due to skewed distribution of DNA polymorphism between the closely related Tanjil and Unicrop varieties. In future work to develop a comprehensive reference genetic map for NLL, the focus will be on the more evenly polymorphic 83A:476 × P27255 population.

Efforts to develop a reference genome assembly of the cultivar Tanjil by improving the survey assembly presented herein with mate-pair datasets as well as the BAC-end sequence data generated (Gao et al., 2011) are ongoing. Additional SNP markers identified either in this project or in SNP markers already developed by Yang and colleagues (Yang, personal communication) for the Tanjil × Unicrop population that are also polymorphic on the 83A:476 × P27255 population will then be selected to assign as many scaffolds as possible to a linkage group, resulting in a dense reference genetic map of the 83A:476 × P27255 population. This will allow cross-comparison of the two dense genetic maps for NLL by having anchored markers from both projects on each genetic map. The molecular markers presented here and by Yang and colleagues (Yang et al., 2013b) will thus aid lupin breeding programmes, particularly when markers linked to traits can be associated with a scaffold. These scaffolds can be mined for candidate genes for that trait, and this will be facilitated by the extensive number of gene-based molecular markers generated in this study. Future work using the Tanjil transcriptome as a reference and comparison to other wild, early domesticated cultivars and current elite cultivars is underway to identify key genes underlying the domestication traits as well as to identify candidate genes for important traits in the wild germplasm currently absent in elite cultivars. The NLL transcriptome datasets presented herein will thus speed up breeding efforts for this recently domesticated grain legume.
Experimental procedures

Plant material, growth conditions, DNA and RNA isolation

Narrow-leafed lupin (NLL) seed was provided by the Department of Agriculture and Food, Western Australia (DAFWA). Seeds included four parents of two recombinant inbred line (RIL) populations, being Tanjil, Unicrop, 83A:476 and P27255 as well as 152 F₈ of individuals of the 83A:476 × P27255 RIL population, also known as the domestic × wild (D × W) population. The parental lines and the RILs were grown out at the Shenton Park Field Station of the University of Western Australia during the winter/spring season of 2012, and leaf material collected. DNA was isolated from frozen leaf material using the CTAB method as described previously by Kamphuis et al. (2008).

The domesticated NLL accessions Unicrop (early domesticated cultivar released in 1973) and Tanjil (more recent domesticated cultivar released in 1998) and wild accession P27255 were grown in temperature-controlled growth cabinets at 22 °C day for 16 h and 20 °C night for 8 h. Once plants started to flower and set seed on the main stem inflorescence, tissue was harvested from roots, stem, leaf, flower and mixed developmental stages of seed for subsequent RNA extraction.

Approximately 150 mg plant material from each tissue type was ground to a fine powder under a stream of liquid nitrogen. The ground powder was homogenized in 1.5 mL Eppendorf tubes using two-times 500 μL TRIzol reagents (Invitrogen, Carlsbad, CA), followed by a 15-min incubation period at room temperature (RT). After centrifugation at 12 000 g and 4 °C for 10 min, the supernatant was mixed with 200 μL chloroform, followed by another centrifugation step at 12 000 g and 4 °C for 15 min. The upper phase is mixed with 300 μL of high salt precipitation buffer (0.8 M sodium citrate/1.2 M NaCl) and 300 μL of isopropanol and incubated on ice for at least 10 min to selectively precipitate total RNA. Precipitated RNA is centrifuged (10 min 12 000 g and 4 °C) and washed twice in 75% ethanol. RNA was dissolved in diethylpyrocarbonate (DEPC)-treated water. Quality and quantity was assessed by both BioAnalyzer (Agilent, Santa Clara, CA) and Qubit assays (Invitrogen). RNA samples were provided to the Beijing Genome Institute (BGI) who generated and sequenced TruSeq libraries (Illumina, San Diego, CA) on a HiSeq2000 (Illumina) for each tissue type of cv. Tanjil, Unicrop and P27255. For each accession and tissue type, the RNAseq datasets were submitted to GenBank as short-read archives under BioProject PRJNA248164.

Tanjil de novo reference transcriptome

Additional sequence data, prepared from leaf tissue material of NLL domesticated cultivar 83A:476, were sourced from the 1KP initiative (www.onekp.com, dataset ‘TTRG’). Raw RNAseq data for all cultivars, including 83A:476, were trimmed for low quality (<Q30) and Illumina sequencing primer and adaptor sequences (Truseq v1 and v2) via Cutadapt 1.1 (Martin, 2011) (overlap 10, times 3, minimum length 25). Reads trimmed to less than 25 bp were discarded, and reads with a discarded pair were retained as singleton reads.

A de novo transcriptome was assembled from the combination of mRNA from all 5 tissues of cv Tanjil using Trinity r2012-01-25 (Haas et al., 2013) with glue and k-mer coverage thresholds of 5. Assembled transcripts were functionally annotated, initially by blastx to the NCBI NR protein database (top 20 hits at e-value threshold 1e-5). Functional annotations were assigned by Interproscan v 4.8 (Zdobnov and Apweiler, 2001). The Tanjil transcriptome assembly has been submitted to GenBank as a gene expression omnibus (GEO) under BioProject PRJNA248164.

Tanjil draft genome survey

Tanjil genomic DNA samples were sequenced by the BGI using the Illumina HiSeq 2000 platform. Two Illumina paired-end libraries were obtained with 100 bp read lengths, 170 bp and 500 bp insert sizes and ‘clean’ data yields of 31.3 and 13.6 Gb, respectively, and were submitted to GenBank as short-read archives under BioProject PRJNA248164. Genome size was estimated using a k-mer distribution analysis using k = 17 and a peak depth of 25 X coverage. The draft genome was assembled de novo with SOAPdenovo v1.05 (Luo et al., 2012). The assembled survey sequences were used in this study for the in silico elimination of poor marker candidates, as described below. The draft genome was also used as a diagnostic tool to assess the distribution of new markers designed in this study across the genetic map, by providing a point of reference to link marker sequence and genetic map data. In the future, these draft sequences will also be incorporated into a high-quality and comprehensive genome assembly of NLL cv Tanjil.

Detection of sequence variation across cultivars relative to the cv Tanjil reference transcriptome

RNAseq reads from all cultivars and tissue types were aligned to the cv. Tanjil reference transcriptome assembly via bowtie v2.0.5 (Langmead and Salzberg, 2012) (sensitive end-to-end, insert size range 0–500 bp). For the purposes of comparing sequence variations between cultivars (relative to the cv. Tanjil reference transcriptome), all alignment data for multiple tissue types were merged into a single dataset per cultivar, via SAMtools v1.18 (Li et al., 2009). Sequence polymorphisms were detected using GATK v2.2.16 (McKenna et al., 2010). This process involved local realignment with GATK’s RealignerTargetCreator (windowsize 20, minReadsAtLocus 2) and IndelRealigner (consensusDeterminationModel USE_SW, LODThresholdForCleaning 2, maxConsensuses 100, maxReadsForRealignment 100 000, maxReadsInMem 300 000) and variant calling with GATK HaplotypeCaller (minPruning 5). Raw RNAseq reads of cv. Tanjil were also aligned to the Tanjil reference transcriptome in the same way, to identify sites of heterozygosity or uncertain base calling.

Prediction of tissue-specific transcript expression

Relative expression of the Trinity-assembled transcripts of cv. Tanjil, between RNAseq libraries corresponding to root, stem, leaf, flower and seed tissues, was calculated from Bowtie alignments using Cuffdiff (fragment bias correction, multiple read correction, minimum alignment count ≥5, upper quartile normalization, compatible hits normalization) of Cufflinks v2.0.1 (Trapnell et al., 2010). Tissue-specific expression was predicted by identifying transcripts with significantly higher expression in one tissue relative to all other tissues. Significant differences in expression were required to involve a ≥ twofold change and q-value ≤0.05. Tissue-specific transcripts were also predicted for all combinations of two tissues, relative to the remaining three tissues.

Gene-based molecular marker design

Variants for all cultivars relative to the Tanjil reference transcriptome were combined into a single haplotype table with VCFtools.
v0.1.7 (vcf-merge, --collapse any --remove duplicates) (Danecek et al., 2011) from which regions suitable for the design of primer pairs flanking length and SNP polymorphisms were predicted. Known heterozygous sites in Tanjis and polymorphic other cultivars were excluded from potential oligo-hybridizing sequences, as well as the first and last 20 bp of a transcript. Candidate PCR markers were also filtered for single amplicons by in silico PCR in the Tanji reference transcriptome and for zero or one amplicon in the draft genome. In silico PCR was performed with e-PCR v2.3.12 (Schuler, 1997), allowing for two mismatches and two gaps in the primer pair and testing for potential amplicons between 10 and 10 000 bp in length. Candidate markers with single-copy in silico amplification in the draft genome were filtered for nontron spanning markers, requiring amplicons that did not differ in length by more than 5% from the transcript amplicon length. Length polymorphism markers were required to have at least a 2-bp difference in length between alleles. Marker alignments to the draft genome assembly derived from in silico PCR tests were also used to associate new markers with those of the reference genetic map (Kroc et al., 2014), where draft genomic scaffold sequences were long enough to contain matches to both new and mapped markers.

Regions containing a biallelic SNP suitable for the Fluidigm platform (Fluidigm, San Francisco, CA) were conservatively selected for bi-allelic SNPs that were predicted to be present between either the Tanjil and Unicon cultivars or the P27255 and 83A:476 cultivars, as this would allow maximum utility across the two important populations resulting from crosses between these respective pairs of parents. Regions flanking bi-allelic SNP sites were also required to be between 60 to 300 bp in length, with less than 60% G:C content, containing no other predicted SNP or indel polymorphisms between any of the four cultivars and be nontron spanning (single HSP for best BLASTN hit of flanking region to draft genome). Ninety-six Fluidigm markers were selected for testing, 48 belonging to those which aligned to draft assembly scaffolds or in some cases, to scaffolds which were able to be placed on the published genetic map (Nelson et al., 2010). Fluidigm markers mapping to genomic scaffolds were selected for initial testing in descending order of priority: (1) proximity to mapped traits, (2) predicted to be near the top or bottom of linkage groups on the genetic map, (3) co-mapped to a scaffold with existing marker and (4) matching draft assembly scaffold.

Genotyping and linkage analysis

A total of 239 new length-based polymorphic markers derived from the transcriptome datasets and 17 BAC-end-derived SSR markers were used to genotype 152 RILs using the high-throughput multiplex ready PCR as described by Gao et al. (2011). The simple-sequence repeat (SSR) markers used in this study were previously designed from a library of L. angustifolius cv Tanjil BAC-end sequences (Gao et al., 2011). A total of 96 SNP markers were used to genotype 152 RILs on the Fluidigm platform by Millennium Science, using 50 ng/µL DNA concentrations of which 2.5 µL was delivered to each microfluidic chamber.

Of the new markers used to genotype the RIL population, 170 transcriptome-derived length polymorphic (indel) markers, 88 polymorphic Fluidigm SNP markers and 17 BAC-end-derived SSR markers were polymorphic and high quality. These genotyping datasets were integrated with the most recently released reference genetic map comprising 492 MFLP, 156 RFLP, 298 DArT and 247 PCR-based STS markers and seven single-gene trait loci genotyped in up to 112 RILs (Kroc et al., 2014). Combining all datasets gave a total of 1475 markers in a population of 161 RILs, which were used for linkage mapping.

Linkage mapping was conducted with the aid of Multipoint v2.1 software (MultiQTL Ltd, Haifa, Israel), which uses the ‘evolutionary optimization strategy’ to perform multilocus ordering of linkage groups (Mester et al., 2003). We used the approach described in detail by Kroc et al. (2014) with some modifications. Briefly, redundant markers were set aside before clustering. Iterative clustering analysis was conducted at recombination frequency: rf = 0.14, 0.16, 0.18 and then increased at 0.01 increments until a maximum of rf = 0.27. At each clustering iteration, multipoint analysis was conducted and resulting groups merged as rf was incrementally increased. Jack-knife analysis was performed on the rf = 0.27 linkage groups to identify markers that had a destabilizing effect on locus order, which were then temporarily set aside. The remaining ‘skeleton’ markers were used to construct the framework map with genetic intervals size transformed to account for multiple meioses involved in the development of the RIL population and expressed in Kosambi centiMorgans (cM). Redundant markers were then assigned to their representative framework markers, and destabilizing markers were ‘attached’ to the most likely intervals between skeleton markers.

A comparison of the new genetic map to the previously published map (Kroc et al., 2014) was visualized with Circos v0.52 (Krzywinski et al., 2009), with new and old linkage groups linked by the relative cM coordinates of their shared sequence-based markers. Genetic marker density was also calculated in both the old and new maps, for sequence-based markers only, in terms of markers/10 cM. Marker densities were calculated within a 1-cM sliding window across the length of each linkage group.

Acknowledgements

We thank Natalie Fletcher for technical support and James Millar and Paul Lacaze from Millennium Science for support with the genotyping and development of Fluidigm SNP markers. We also thank Bevan Buichell and DAFWA for kindly providing the narrow-leaved lupin seeds and the 1000 Plant initiative for access to the narrow-leaved lupin 83A:476 transcriptome dataset. We thank Rhonda Foley and Rohit Mago for helpful comments on the manuscript. This research was undertaken with the assistance of resources provided at the NCI National Facility systems at The Australian National University and The University of Queensland through the National Computational Merit Allocation Scheme supported by the Australian Government. This project was funded by the Grains Research and Development Corporation (GRDC) and the Commonwealth Scientific and Industrial Research Organisation (CSIRO).

Conflict of interest

The authors declare they have no conflict of interest.

References

Belski, R., Mori, T.A., Puddey, I.B., Sipsas, S., Woodman, R.J., Ackland, T.R., Beilin, L.J., Dove, E.R., Carlyon, N.B., Jayaseena, V. and Hodgson, J.M. (2010) Effects of lupin-enriched foods on body composition and cardiovascular disease risk factors: a 12-month randomized controlled weight loss trial. Int J Obesity, 35, 810–819.
Berger, J.D., Buurichell, B., Luckett, D.J. and Nelson, M.N. (2012a) Domestication bottlenecks limit genetic diversity and constrain adaptation in narrow-leaved lupin (Lupinus angustifolius L.). Theor. Appl. Genet. 124, 637–652.

Berger, J.D., Buurichell, B., Luckett, D.J., Palta, J.A., Ludwig, C. and Liu, D. (2012b) How has narrow-leaved lupin changed in its 1st 40 years as an industrial, broadacre crop? A G x E-based characterization of yield-related traits in Australian cultivars. Field Crop Res. 126, 152–164.

Boersma, J.G., Buirchell, B.J., Sivasithamparam, K. and Yang, H. (2007a) Development of a sequence-specific PCR marker linked to the Ku gene which removes the vernalization requirement in narrow-leaved lupin. Plant Breed. 126, 306–309.

Boersma, J.G., Buirchell, B.J., Sivasithamparam, K. and Yang, H. (2007b) Development of two sequence-specific PCR markers linked to the le gene that reduces pod shattering in narrow-leaved Lupin (Lupinus angustifolius L.). Genet. Mol. Biol. 30, 623–629.

Boersma, J.G., Nelson, M.N., Sivasithamparam, K. and Yang, H. (2009) Development of sequence-specific PCR markers linked to the Tardus gene that reduces pod shattering in narrow-leaved lupin (Lupinus angustifolius L.). Mol. Breed. 23, 259–267.

Cheng, L., Buccarello, B., Shen, J., Allan, D. and Vance, C.P. (2011) Update on lupin cluster roots. Update on white lupin cluster root acclimation to phosphorus deficiency. Plant Physiol. 156, 1025–1032.

Cowling, W.A., Huynhe, C. and Swiecicki, W. (1998) Lupin breeding. In Lupins as Crop Plants: Biology, Production and Utilization (Gladstone, J.S., Atkins, C.A. and Hamblin, J., eds), pp. 93–120. Wallingford: CAB International.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G. and Durbin, R. (2011) The variant call format and VCFtools. Bioinformatics 27, 2556–2558.

Deschamps, S., Rota, M., Ratashak, J.P., Biddle, P., Thureen, D., Farmer, A., Durbin, R. (2011) The variant call format and VCFtools. Genet. Mol. Biol. 10, 331–344.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., Macnair, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., Williams, T., Dewey, C.N., Hirschel, R., Leduc, R.D., Friedman, N. and Regev, A. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8, 1494–1512.

Hayden, M.J., Nguyen, T.M., Waterman, A. and Chalmers, K.J. (2008) Multiplex-ready PCR: a new method for multiplexed SSR and SNP genotyping. BMC Genomics, 9, 80.

Hiremath, P.J., Farmer, A., Cannon, S.B., Woodward, J., Kudapa, H., Tuteja, R., Kumar, A., BhanupRakash, A., Mulaosmanovic, B., Gujaria, N., Krishnamurthy, L., Gaur, P.M., KaviKishor, P.B., Shah, T., Srivinasan, V., Lohse, M., Xia, Y., Town, C.D., Cook, D.R., May, G.D. and Varshney, R.K. (2011) Large-scale transcriptome analysis in chickpea (Cicer arietinum L.), an orphan legume crop of the semi-arid tropics of Asia and Africa. Plant Biotech. J. 9, 922–931.

Johnson, S.K., McQuillan, P.L., Sin, J.H. and Ball, M.J. (2003) Sensory acceptability of white bread with added Australian sweet lupin (Lupinus angustifolius) kernel fibre and its glycemic and insulinaemic responses when eaten as a breakfast. J. Sci. Food Agric. 83, 1366–1372.

Kamphuis, L.G., Lichtenzweig, J., Oliver, R.P. and Elwood, S.R. (2008) Two alternative recessive quantitative trait loci influence resistance to spring black stem and leaf spot in Medicago truncatula. BMC Plant Biol. 8, 30.

Kaparzak, A., Safar!, J., Janda, J., Dolez!, J., Wolko, B. and Naganowska, B. (2006) The bacterial artificial chromosome (BAC) library of the narrow-leaved lupin (Lupinus angustifolius L.). Cell. Mol. Biol. Lett. 11, 396–407.

Kaur, S., Pembleton, L.W., Ol Cogan, N., Savin, K.W., Leontorfe, P., Pauli, J., Materne, M. and Forster, J.W. (2012) Transcriptome sequencing of field pea and faba bean for discovery and validation of SSR genetic markers. BMC Genomics, 13, 104.

Kroc, M., Koczyk, G., Swiecicki, W., Kilian, A. and Nelson, M.N. (2014) New evidence of ancestral polyploidy in the Genistoid legume Lupinus angustifolius L. (narrow-leaved lupin). Theor. Appl. Genet. 127, 1237–1249.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. Genome Res. 19, 1639–1645.

KuakiviecKiewicz, M., Wyrwa, K., Szczepanik, A., Rychel, S., MajcherKiewicz, K., PrzySiLecka, L., KarloWSki, W., Wolko, B. and Naganowska, B. (2013) Comparative genomics of Lupinus angustifolius gene-rich regions: BAC library exploration, genetic mapping and cytogenetics. BMC Genomics, 14, 79.

Kudapa, K., Azam, S., Sharpe, A.G., Taran, B., Li, R., Deonovic, C., Cameron, C., Farmer, A.D., Cannon, S.B. and Varshney, R.K. (2014) Comprehensive transcriptome assembly of chickpea (Cicer arietinum L.) using Sanger and next generation sequencing platforms: development and applications. PLoS ONE, 9, e86039.

Lammers, H., Clements, J.C. and Nelson, M.N. (2013) How a phosphorus-acquisition strategy based on carbohydrate exudation powers the success and agronomic potential of lupines (Lupinus, Fabaceae). Am. J. Bot. 100, 263–288.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. Nat. Methods, 9, 357–359.

Lee, Y.P., Mori, T.A., Sipsas, S., Barden, A., Puddsey, I.B., Burke, V., Hall, R.S. and Hodgson, J.M. (2006) Lupin-enriched bread increases satiety and reduces energy intake acutely. Am. J. Clin. Nutr. 84, 975–980.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The sequence alignment/map format and SAMtools. Bioinformatics, 25, 2078–2079.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Sh, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D.W., Yi, S.M., Peng, S., Xiaoqian, Z., Li, G., Xiao, X., Li, Yang, H., Wang, J., Lam, T.W. and Wang, J. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience, 1, 18.

Magni, C., Sessa, F., Accardo, E., Vanoni, M., Morazgoni, P., Scarafoni, A. and Duranti, M. (2004) Conglutin gamma, a lupin seed protein, binds insulin in vitro and reduces plasma glucose levels of hyperglycemic rats. J. Nutr. Biochem. 15, 646–650.
Lupinus angustifolius transcriptomics and markers

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *F1000Res*. **journal**, 17, 10–12.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M.A. (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303.

Mester, D., Ronin, Y., Minkov, D., Nevo, E. and Korol, A. (2003) Constructing large-scale genetic maps using an evolutionary strategy algorithm. *Genetics*, **165**, 2269–2282.

Naganowski, B., Wolk, B., Slivinska, E. and Kaczmarek, Z. (2003) Nuclear DNA content variation and species relationships in the genus Lupinus (Fabaceae). *Ann. Bot.* **92**, 849–355.

Nelson, M.N., Phan, H.T., Ellwood, S.R., Moolhuijzen, P.M., Hane, J., Williams, A., O’Lone, C.E., Fossu-Nyarko, J., Scobie, M., Cakir, M., Jones, M.G., Belgard, M., Ktiazkiewicz, M., Wolk, B., Barker, S.J., Oliver, R.P. and Cowling, W.A. (2006) The first gene-based map of *Lupinus angustifolius* L.-location of domestication genes and conserved synteny with *Medicago truncatula*. *Theor. Appl. Genet.* **113**, 225–238.

Nelson, M.N., Moolhuijzen, P.M., Boersma, J.G., Chudy, M., Lesniewska, K., Belgard, M., Oliver, R.P., Siewicki, W., Wolk, B., Cowling, W.A. and Ellwood, S.R. (2010) Aligning a new reference genetic map of *Lupinus angustifolius* with the genome sequence of the model legume, *Lotus japonicus*. *DNA Res.*, **17**, 73–83.

O’ourke, J.A., Yang, S.S., Miller, S.S., Bucciarelli, B., Liu, J., Rydeen, A., Bozozki, Z., Udhe-Stone, C., Tu, Z.J., Allan, D., Gronwald, J.W. and Vance, C.P. (2013) An RNA-seq transcriptome analysis of orthophosphate-deficient white lupin reveals novel insights into phosphorus acclimation in plants. *Plant Physiol.* **161**, 705–724.

Parra-González, L., Aravena-Abarzuza, G., Navarro-Navarro, C., Udall, J., Maughan, J., Peterson, L., Salvo-Garindo, H. and Maureira-Butler, I. (2012) Yellow lupin (*Lupinus Áureus*) transcriptome sequencing: molecular marker development and comparative studies. *BMC Genomics*, **13**, 425.

Pazos-Navarro, M., Dabauza, M., Correia, E., Hanson, K., Teakle, N., Real, D. and Nelson, M.N. (2011) Next generation DNA sequencing technology delivers valuable genetic markers for the genomic orphan legume species, *Bituminaria bituminosa*. *BMC Genet.* **12**, 104.

Schuler, G.D. (1997) Sequence mapping by electronic PCR. *Genome Res.* **7**, 541–550.

Seymour, M., Kirkegaard, J.A., Peoples, M.B., White, P.F. and French, R.J. (2011) Break-crop benefits to wheat in Western Australia – insights from over three decades of research. *Crop and Pasture Sci.* **63**, 1–16.

Sharpe, A.G., Ramsay, L., Sanderson, L.-A., Fedoruk, M.I., Clarke, W.E., Li, R., Kagale, S., Vijayam, P., Vandenberg, A. and Bett, K.E. (2013) Ancient orphan crop joins modern era: gene-based SNP discovery and mapping in lentil. *BMC Genomics*, **14**, 192.

Trappel, C., Williams, B.A., Perea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515.

Vardhaney, R.K., Graner, A. and Sorrells, M.E. (2005) Genic microsatellite markers in plants: features and applications. *Trends Biotechnol.* **23**, 48–55.

Vardhaney, R.K., Nayak, S.N., May, G.D. and Jackson, S.A. (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol.* **27**, 522–530.

Wolk, B., Clements, J.C., Naganowska, B., Nelson, M.N. and Yang, H. (2011) Wild crop relatives: genomic and breeding resources. In *Lupinus* (Kole, C., ed), pp. 153–206. Heidelberg: Springer.

Yang, H., Shankar, M., Buichell, B.J., Sweetingham, M.W., Caminero, C. and Smith, P.M.C. (2002) Development of molecular markers using MFLP linked to a gene conferring resistance to *Diaporthe toxica* in narrow-leaved lupin (*Lupinus angustifolius* L.). *Theor. Appl. Genet.* **105**, 265–270.

Yang, H., Boersma, J.G., You, M., Buichell, B.J. and Sweetingham, M.W. (2004) Development and implementation of a sequence-specific PCR marker linked to a gene conferring resistance to anthracnose disease in narrow-leaved lupin (*Lupinus angustifolius* L.). *Mol. Breed.* **14**, 145–151.

Yang, H., Tao, Y., Zheng, Z., Shao, D., Lo, Z., Sweetingham, M.W., Buichell, B.J. and Li, C. (2013a) Rapid development of molecular markers by next-generation sequencing linked to a gene conferring phomopsis stem blight disease resistance for marker-assisted selection in lupin (*Lupinus angustifolius* L.) breeding. *Theor. Appl. Genet.* **126**, 511–522.

Yang, H., Tao, Y., Zhang, Z., Zhang, Q., Zhou, G., Sweetingham, M.W., Howieson, J.G. and Li, C. (2013b) Draft genome sequence, and a sequence-defined genetic linkage map of the legume crop species *Lupinus angustifolius* L. *PLoS ONE*, **8**, e64799.

You, M., Boersma, J.G., Buichell, B.J., Sweetingham, M.W., Siddique, K.H.M. and Yang, H. (2005) A PCR-based molecular marker applicable for marker-assisted selection for anthracnose disease resistance in lupin breeding. *Cell. Mol. Biol. Lett.* **10**, 123–134.

Young, N.D. and Bharti, A.K. (2012) Genome-enabled insights into legume biology. *Ann. Rev. Plant Biol.* **63**, 283–305.

Zdobnov, E.M. and Apweiler, R. (2001) InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848.

Zhang, J., Liang, S., Duan, J., Wang, J., Chen, S., Cheng, Z., Zhang, Q., Liang, X. and Li, Y. (2012) De novo assembly and characterisation of the transcriptome during seed development, and generation of genic-SSR markers in peanut (*Arachis hypogaea L.)*. *BMC Genomics*, **12**, 90.

**Supporting information**

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

**Figure S1** *Lupinus angustifolius* ‘clean’, quality-controlled RNA-seq data across the various tissues and cultivars/accessions used in this study.

**Figure S2** Frequency distribution (y-axes) for sequences of various lengths (x-axes) for the *L. angustifolius* cv. Tanji reference transcriptome assembly (A) and draft genome survey assembly (B).

**Figure S3** RNASeq-derived polymorphisms detected between *L. angustifolius* cv. Tanji, cv. Unicrop and wild accession P27255 overlaid onto the previously published genetic map (*Kroc et al., 2014*).

**Figure S4** K-mer frequency distribution in Illumina raw genomic reads used for the *L. angustifolius* cv. Tanji draft genome survey assembly, indicating a main k-mer coverage peak at K = 25 and relatively low levels of heterozygosity and repetitive DNA.

**Figure S5** Comparison of the new (outer ring) and last-published (inner ring) (*Kroc et al., 2014*) genetic maps for *Lupinus angustifolius*, generated using the 83A:476 × P27255 (‘Domestic vs Wild’) recombinant inbred line population.

**Table S1** Relative expression across tissues (FPKM) of tissue-specific mRNA transcripts (Trinity components) and their assigned Interpro functional annotations and putative roles.

**Table S2** Distribution of newly designed gene-based standard PCR SNP and indel and Fluidigm bi-allelic SNP markers over the 20 *L. angustifolius* linkage groups from *Kroc et al., 2014*.

**Table S3** Marker order for each linkage group and the genotype data for 1475 markers in the 161 recombinant inbred lines (RILs) of the 83A:476 × P27255 population.

**Table S4** Overview of molecular markers linked to important traits in the narrow-leaved lupin breeding program and the closest markers to these traits derived from this marker study.

© 2014 CSIRO. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 13, 14–25