Kamphuis Lars G (Orcid ID: 0000-0002-9042-0513)
Bayer Philipp (Orcid ID: 0000-0001-8530-3067)

A pan-genome and chromosome-length reference genome of narrow-leafed lupin (*Lupinus angustifolius*) reveals genomic diversity and insights into key industry and biological traits

Gagan Garg[1,#], Lars G. Kamphuis[1,2,3,#], Philipp E. Bayer[4], Parwinder Kaur[5], Olga Dudchenko[6,7], Candy M. Taylor[2,5], Karen M. Frick[1,8], Rhonda C. Foley[1], Ling-Ling Gao[1], Erez Lieberman Aiden[5,6,7,9,10], David Edwards[2,4] and Karam B. Singh[1,2,3,*]

[1]CSIRO Agriculture and Food, Floreat, WA 6014, Australia

[2]UWA Institute of Agriculture, University of Western Australia, Crawley, WA 6009, Australia

[3]Centre for Crop and Disease Management, Curtin University, Bentley, WA, 6102, Australia

[4]The School of Biological Sciences, University of Western Australia, Crawley, WA 6009, Australia

[5]School of Agriculture and Environment, University of Western Australia, Crawley, WA 6009, Australia

[6]Center for Genome Architecture, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston TX 77030, USA

[7]Center for Theoretical Biological Physics, Rice University, Houston TX 77005, USA

[8]Section for Plant Biochemistry and Copenhagen Plant Science Centre, Department of Plant and Environmental Sciences, University of Copenhagen, Frederiksberg, Denmark

[9]Shanghai Institute for Advanced Immunochemical Studies, ShanghaiTech, Pudong, China

[10]Broad Institute of MIT and Harvard, Cambridge, MA, USA

[#]These authors contributed equally and are thus joint first authors.

[*]Author for correspondence: Karam B. Singh, karam.singh@csiro.au; +61893336320; Lars G. Kamphuis, Lars.Kamphuis@csiro.au; +61893336109.

Keywords: Legume, pan-genome assembly, quinolizidine alkaloids, seed storage proteins, resistance genes, Genistoids.

Author contact information and ORCID IDs:

Gagan Garg	0000-0002-9523-6360	Gagan.Garg@csiro.au
Lars G. Kamphuis	0000-0002-9042-0513	Lars.Kamphuis@csiro.au
Philipp E. Bayer	0000-0001-8530-3067	philipp.bayer@uwa.edu.au
Parwinder Kaur	0000-0003-0201-0766	Parwinder.kaur@uwa.edu.au

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/tpj.15885

Olga Dudchenko	0000-0001-9163-9544	Olga.Dudchenko@bcm.edu
Candy M. Taylor	0000-0003-2314-5520	candy.taylor@uwa.edu.au
Karen M. Frick	0000-0002-5635-6693	kmf@plen.ku.dk
Rhonda C. Foley	0000-0003-2942-394X	rfoleyperth@gmail.com
Ling-Ling Gao	0000-0002-4221-2816	Lingling.Gao@csiro.au
Erez Lieberman Aiden	0000-0003-0634-6486	erez.aiden@bcm.edu
David Edwards	0000-0001-7599-6760	Dave.edwards@uwa.edu.au
Karam B. Singh	0000-0002-2777-7448	Karam.Singh@csiro.au

Summary

Narrow-leafed lupin (NLL; *Lupinus angustifolius*) is a key rotational crop for sustainable farming systems, whose grain is high in protein content. It is a gluten-free, non-GM, alternative protein source to soybean and as such has gained an interest as a human food ingredient. Here, we present a chromosome-length reference genome for the species and a pan-genome assembly comprising 55 NLL lines including, Australian and European cultivars, breeding lines and wild accessions. We present the core and variable genes for the species and report on the absence of essential mycorrhizal associated genes. The genome and pan-genomes of NLL and its close relative white lupin (*L. albus*) are compared. Furthermore, we provide additional evidence supporting *LaRAP2-7* as the key alkaloid regulatory gene for NLL and demonstrate the NLL genome is underrepresented in classical NLR disease resistance genes compared to other sequenced legume species. The NLL genomic resources generated here coupled with previously generated RNA-seq datasets provide new opportunities to fast-track lupin crop improvement.

Introduction

Narrow-leafed lupin (NLL; *Lupininus angustifolius* L.) is a grain legume grown as a break crop in rotation with cereal crops, thereby reducing the need for fertilizers, increasing cereal yields and importantly providing disease breaks (Seymour *et al.*, 2012). NLL and other lupins thrive on nutrient poor soils due to their symbiosis with beneficial bacteria to fix atmospheric nitrogen and to efficiently mobilise phosphorus from soils (Lambers *et al.*, 2013). NLL is a relatively young pulse crop, having only begun the process of domestication in the early 20th Century in Germany, which was concluded by the release of the first fully domesticated cultivar with low alkaloid content, permeable seeds, early flowering and non-shattering pods in the 1960s in Australia (Gladstones, 1970). To date four lupin species have been domesticated, including white lupin (*L. albus*), yellow lupin (*L. luteus*), pearl lupin (*L. mutabilis*) and NLL, with NLL being the predominant lupin grown worldwide (85% of all lupins) (FAO, 2021). NLL is grown over 750,000 hectares predominantly in Australia, but also in other countries, such as Poland, Russia and Germany (FAO, 2021).

The lupin grain is mainly used for animal and aquaculture feed (White *et al.*, 2007), but in recent years has gained interest as a human health food and food additive. This is because lupin grain is rich in protein (30-40% of whole seeds), has low amounts of undesired starch compared to other pulses, is high in dietary fibre (25-30%), low in fat and carbohydrates and gluten-free (Kohajdová *et al.*, 2011). In human food products, lupin kernel flour is predominantly used as a food additive in bread and pasta (Kohajdová et al., 2011) and has been shown to reduce insulin resistance (Lee *et al.*, 2006). It's attractiveness as a gluten free, non-GM alternative to soybean has resulted in the production of a series of lupin-based gluten-free foods including pasta and meat replacement products for the vegetarian and vegan markets. Furthermore, specific lupin seed proteins have been demonstrated to reduce glycaemia to comparable levels as achieved with the predominately used hypoglycaemic drug metformin (Lee et al., 2006), and have additional nutraceutical properties, for example, that improve inflammatory related diseases, as well as anti-microbial properties (reviewed in Jimenez-Lopez, 2020).

While NLL is emerging as a human health food, toxic specialised metabolites of the quinolizidine alkaloids class are required to remain below a 0.02% threshold in the grain for it to be used for food or feed purposes (Frick *et al.*, 2017). A picture is emerging in which environmental stresses cause an increase in alkaloids in the NLL leaves and grain (Frick *et al.*,

2019, Frick *et al.*, 2018), but independent studies have identified different candidate genes for the key alkaloid regulatory locus *iucundus* (Kroc *et al.*, 2019b, Wang *et al.*, 2021).

All lupins belong to the genus *Lupinus* in the Genistoid clade of legumes, which diverged early in the evolution of Papilionoid legumes (Nevado *et al.*, 2016). There are over 267 species of lupin, some of which are ecological pioneers in impoverished conditions. The natural distribution of the genus is around the Mediterranean region (the 'Old World' lupins) and North and South America ('New World' lupins), where the Andean lupin species have shown speciation rates not seen elsewhere in the plant kingdom (Nevado et al., 2016). The genus contains both annual and perennial species that occur in a range of different habitats (Nevado et al., 2016), and thus possesses a wealth of information for adaptive traits useful in different agricultural and climatic zones. These may be used to develop a sustainable profitable lupin crop through the development of new lupin crop species or, in the case of NLL, through genetic base broadening (Berger *et al.*, 2012, Berger *et al.*, 2013). To this end a range of genetic and genomic resources have been developed for NLL in recent years. These include the generation of BAC-libraries (Gao *et al.*, 2011, Kasprzak *et al.*, 2006), transcriptome datasets (reviewed in Kamphuis *et al.*, 2020), cytogenetic maps (Bielski *et al.*, 2020) and various genetic maps. In addition, a survey genome sequence was generated in 2013 for NLL (Yang *et al.*, 2013), followed by the first comprehensive draft genome sequence coupled with a dense reference genetic map for the species (Hane *et al.*, 2017). More recently, an improved reference genome was released for NLL (Wang et al., 2021) and two independent reference genomes for its close relative white lupin (Hufnagel *et al.*, 2020, Xu *et al.*, 2020).

In the last decade, there has been a move to generate pan-genomes for plant species to capture the genomic diversity for a species and allow genes core to all genomes and those that are variable to be identified (Bayer *et al.*, 2020). Variable genes are often associated with processes involved in biotic and abiotic stress and could thus be important breeding targets in crop breeding programs. To date, two predominant approaches have been used to assemble a pan-genome. The first approach assembles a *de novo* genome for each accession, after which alignment of the genomes allows one to identify dispensable genomic regions. The second approach aligns all sequence reads from multiple accessions to a high-quality chromosome-length reference genome and subsequently assembles the unaligned reads into novel contigs. Following either or both approaches a pan-genome graph can be constructed, the variants called, and core and variable genes identified. The first approach has recently been used to generate a pan-genome for white lupin using 39 accessions, including 11 varieties, 1 landrace

and 2 wild accessions (Hufnagel *et al.*, 2021). This study identified 32,068 core and 14,822 variable genes and discovered candidate genes for alkaloid biosynthesis.

Here we report on the fourth iteration of a reference genome for NLL and the first pan-genome for NLL, which represents the second pan-genome for a genistoid legume (2n = 40). We use these NLL resources to perform a survey of the gene content and its conserved and variable gene sets. Furthermore, we demonstrate the absence of unique mycorrhizal associated genes in all NLL lines, confirm the identity of the *iucundus* locus for alkaloid biosynthesis and demonstrate that the NLL pan-genome is underrepresented in classical disease resistance genes compared to other sequenced legume species.

Results

Development of a chromosome-length *L. angustifolius* reference genome

We previously generated a draft reference assembly by combining short read Illumina sequencing reads with insert sizes from 170 bp to 40 Kb and BAC-end sequencing data (Hane et al 2017; Table S1). To improve this assembly, we generated ~98.5x coverage PacBio long read sequence data across 19 single molecule real-time (SMART) cells on a PacBio Sequel as well as ~ 50x coverage in situ Hi-C data generated across two lanes on an Illumina HiSeq2500. The draft genome was assembled using CANU (v1.8; Koren *et al.*, 2017) and scaffolded using Hi-C. The Hi-C reads were used to anchor, order, orient, and correct misjoins in the draft genome assembly created above using the 3D *de novo* assembly (3D-DNA) pipeline (Dudchenko *et al.*, 2017). The resulting assembly was then polished using the Juicebox Assembly Tools (Dudchenko *et al.*, 2018) and the contact maps were visualized using Juicebox visualization software (Durand *et al.*, 2016). Following the integration of Hi-C sequencing data, the new reference assembly contained 2,349 scaffolds assigned to 20 chromosomes with scaffold N50 length of 30.7 Mbp and the longest scaffold being 45.7 Mbp (Table 1, Fig 1).

Using RepeatMasker we determined that the updated reference assembly has a high repetitive sequence content (57.7 % of the genome; Table S2 and S3) of which 54.4 % are known transposable elements (TEs). These TEs were mostly long terminal repeat retrotransposons (Table S2). Following repeat masking, gene annotation was performed and identified 38,545 genes (Data S1 and S2) with an average gene length of 3,665 bp, compared to 33,076 genes with an average gene length of 3,497 bp in the 2017 assembly by Hane et al (2017). To further corroborate completeness of the assembly and annotation we conducted a Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis (Simao *et al.*, 2015) using the land plants

(embryophyta) dataset, revealing that 1,598 of 1,614 BUSCOs (98.9 %) are present and 16 BUSCOs (1.1 %) are absent in the NLL assembly (Table 1). Using the plantae set of 425 BUSCOs revealed that 422 are present (99.3%), whereas 3 BUSCOs are missing (Table 1). Interestingly, the white lupin genome assemblies by Xu et al (2020) and Hufnagel et al (2020) had 11 and 4 plantae BUSCOs missing respectively, which led us to compare the missing BUSCOs between the two lupin species. This showed that two plantae BUSCOs (31855at33090 and 40630at33090) were absent in both lupin species and these encode a helicase and an ABC transporter. Further investigation into the plantae BUSCOs revealed that these BUSCOs are also absent in all published legume species sequenced to date (Bertioli *et al.*, 2019, Hirakawa *et al.*, 2016, Hufnagel et al., 2020, Kreplak *et al.*, 2019, Lonardi *et al.*, 2019, Sato *et al.*, 2008, Schmutz *et al.*, 2010, Schmutz *et al.*, 2014, Varshney *et al.*, 2012, Varshney *et al.*, 2013, Yang *et al.*, 2015, Young *et al.*, 2011, Shen *et al.*, 2020, Parween *et al.*, 2015, Kang *et al.*, 2014) except for 31855at33090 in chickpea and 40630at33090 in lotus (Table 2).

Selection of domestic and wild *L. angustifolius* accessions for creating a NLL pan-genome

To identify distinct groupings within the species, clustering using STRUCTURE analysis on DarTseq data was conducted for 194 NLL accessions (154 wild accessions, 26 Australian varieties and 17 European varieties) (Mousavi-Derazmahalleh *et al.*, 2018b). Clustering resulted in three groups: Cultivated varieties from Australia and Europe (n = 46), wild accessions from southern Europe (n = 43) and wild accessions from the Iberian Peninsula and North Africa (n = 59). There remained some wild mixed accessions (n = 58) that were less clearly defined and which originated across the whole geographic range of NLL. These showed evidence of admixture characterised by high levels of heterozygosity (5.5%), compared to the main groups (1.3%, 0.4% and 3.2%, respectively), and with heterozygosity as high as 20% in some lines]. For re-sequencing to generate the pan-genome, lines had heterozygosity < 2%, except Mandelup. Mandelup (2.5 % heterozygosity) was selected because of its commercial importance as a variety, its role as a parental line to increase genetic diversity from the wild in breeding efforts in Australia and as an important experimental line in NLL pre-breeding research.

We selected a total of 55 accessions, including Tanjil, to construct the NLL pan-genome with representatives from each of the major three groupings identified in the STRUCTURE analysis (Mousavi-Derazmahalleh *et al.*, 2018). Of these, four accessions from each of three main NLL groupings were sequenced at high coverage (~55x or more) and the remaining accessions were sequenced at a lower coverage (~10-30x for most accessions; Table S4), using short read

sequencing. *K*-mer clustering analysis of short read sequencing data for these 55 accessions also resulted in the formation of three major clades (Fig. 2), where a cluster of all cultivated Australian and European accessions grouped together, a second cluster contained two branches with Iberian and North African accessions and a third cluster contained most of the Southern European accessions.

NLL pan-genome assembly and validation

The NLL pan-genome was built using an iterative mapping and assembly approach described by Hu et al. (2020), using the improved reference assembly described herein as a starting point. The assembled pan-genome is approximately 975 Mb in size and includes scaffolds longer than 500 base pairs (bp) with an L50 value of 24.8 Mb. The pan-genome contains 39,339 gene models, compared with the Tanjil reference assembly of 653 Mb and 38,545 gene models (Table 1). The 975 Mb is very close to the estimated genome size for the species of 951Mb (Hane et al., 2017). A BUSCO analysis (Simao et al., 2015) was conducted as described earlier, where no additional missing BUSCOs from the reference assembly were identified. Paired end data for all accessions were mapped back to the pan-genome, with a range from 91.1-99.6% (Table S4). This data was used for gene presence/absence variation analysis and SNP discovery. The number of retro elements and other repeats increased in the pan-genome relative to the reference genome (Table S2 and S5).

Gene presence/absence discovery and characterization

A total of 39,339 genes were predicted in the pan-genome, which is an additional 794 genes compared to the reference assembly (Table 1; Data S3 and S4). The majority (95.2%, 37,459 genes) of the pan-genome is composed of core genes present in all lines, while 4.8 % (1,880 genes) of the genes are variable, including 30 genes present in one line only (Fig. 3A). The pan-genome size and annotated gene set expands with each added NLL line up to 39,339 genes (Fig. 3B). The size of the core genome diminishes with every added NLL line to 37,459 genes. Of the 1,880 variable genes, 1,192 (63.4%) were variable across the domesticated (24 lines/cultivars) and wild material (31 accessions), such that they were missing in at least one domesticated and one wild line. Of the remaining variable genes, 491 (26.1%) were present in all domesticated lines, but absent in at least one wild accession, with 37 of these absent in all wild accessions. Conversely, 197 variable genes were present in all wild accessions and absent in at least one domesticated line with 127 absent in all domesticated lines.

In total, 20,285,108 SNPs were identified in the pan-genome, creating an overall SNP density of 22.03 SNPs/Kbp. In contrast, when identifying SNPs in breeding lines and cultivars only we found a total of 7,165,473 SNPs resulting in density of 7.78 SNPs/Kbs. Thus the high SNP density in the pan-genome was predominantly contributed by the presence of many wild lines and was consistent with previous observations of high SNP density in a wild NLL line (P27255) when mapped to the Tanjil assembly (Hane et al., 2017). Private SNPs, which are present in only one sample but absent in others, are much more abundant in wild compared to domesticated lines (Fig. 3C) with the maximum number of private SNPs found in P26170 (602,327 SNPs).

Functional annotation of variable genes

Functional annotation of the 39,339 genes revealed that 73 % of the core genes (28,721 gene models) had a functional annotation, whereas 26 % (489 gene models) of the variable genes had a functional annotation. Among the most abundant protein domains found in the variable genes were reverse transcriptase, integrase, Myb-like DNA binding and pentatricopeptide repeat domains (Table S6). The most represented Pfam domains in the core genes also contained pentatricopeptide, reverse transcriptase, Myb-like DNA binding, whereas other highly abundant core Pfam domains also encoded leucine-rich repeat and protein kinase domains (Table S6). Subsequent analysis around gene content focused on three areas, including genes associated with symbiosis, alkaloid regulation and plant disease resistance.

Pan-genome gene content with industry-relevant phenotypes

(i) **Symbiosis genes**

Most legumes partake in symbiotic relationships with mycorrhizal fungi and/or rhizobia. Some of the plant genes required for a successful association are shared by both types of symbioses. It has been suggested that the evolutionary younger rhizobium–legume symbiosis recruited part of the genetic programme from the more ancient arbuscular mycorrhizal symbiosis (Parniske, 2008). Interestingly, NLL does not form symbiosis with mycorrhizae, but does with rhizobia. To expand our previous findings using the reference genome of (Hane et al., 2017), we compared the list of genes associated with arbuscular mycorrhizal and rhizobial associations identified in the *Medicago truncatula* genome with the core and variable genes of the NLL pan-genome. Thirty-eight out of 57 *M. truncatula* mycorrhizal and rhizobial associated genes were found in the NLL pan-genome core gene set (Table S7). None were found in the variable gene set. The identified genes were all confirmed to group with the *M. truncatula*

genes in the gene family trees on the Legume Information System genomic data portal (Dash *et al.*, 2016, Gonzales *et al.*, 2004; https://legumeinfo.org/). These included all genes currently known to be involved in rhizobium–symbiosis. However, NLL lacked nearly all genes examined that are required specifically for arbuscular mycorrhizal symbiosis (in italics in Table S7). Two exceptions were *PP2AB'1* (*Lupan_009748*) and *KIN2* (*Lupan_004177*) homologues, which encode serine/threonine protein phosphatase and a protein kinase.

(ii) Analysis of the alkaloid regulatory locus *iucundus*

Quinolizidine alkaloids are specialised metabolites toxic to animals and humans and are a significant industry issue for the production and use of NLL crops. In NLL a major locus controlling alkaloid grain content termed *iucundus* has been mapped to a small region in the genome (746 Kb) between molecular markers LaSSR_025 and LaSNP_509 (Hane et al., 2017). This region contains several candidate genes, including a *dihydrodipicolinate synthase-like* gene (*LaDHDPs-like; Lupan_013751*), an *APETALA2/ethylene response factor-like* gene (*LaRAP2-7; Lupan_013780*) and a leucine-rich repeat (LRR) receptor-like serine/threonine-protein kinase (*LaRSTK; Lupan_013738*) gene. Two independent publications have proposed different candidates as the causal gene underlying the *iucundus* low alkaloid phenotype. Kroc and associates (2019b) determined that *LaDHDPs* was closely linked to the *iucundus* locus, whereas *LaRAP2-7* co-segregated with the *iucundus* locus. Subsequently, a molecular marker in the coding sequence of *LaRAP2-7* was developed for marker assisted selection in breeding programs and demonstrated to correlate perfectly for 199 of the 202 NLL accessions tested and it was speculated the alkaloid regulation of the remaining three accessions was under a different regulatory mechanism (Kroc *et al.*, 2019a). However, independent analysis by Wang and colleagues (2021) suggested that *LaRAP2-7* is not located in the *iucundus* region and these authors proposed that *LaRSTK* is the key regulatory gene for alkaloid biosynthesis and that this gene activates *LaRAP2-7* gene expression.

We used the pan-genome resource to further explore the *iucundus* locus and resolve the identity of the gene underlying low alkaloid phenotypes. Fifty-four accessions constituting the NLL pan-genome for which we had seed, were evaluated for grain alkaloid content using the Dragendorff test. This is a semi-quantitative test whereby alkaloids, if present, react with the Dragendorff reagent to produce an orange to orange-red pigment. All cultivars and breeding lines, except for Polish cultivar Krasnolistny have a 'sweet' alkaloid phenotype as they didn't react with the reagent (Table 3). In contrast all wild accessions displayed a 'bitter' phenotype with the Dragendorff reagent producing the orange to orange-pink reactions on the

Whatmann filter paper. This phenotypic data was subsequently correlated with the proposed causal SNP mutations for *LaRAP2-7* and *LaRSTK*. The comparison clearly demonstrated a perfect correlation of the SNP genotype data for the 54 accessions with the grain alkaloid content phenotype for *LaRAP2-7* and not *LaRSTK.* Our phenotype data combined with the NLL pan genome resource therefore provide additional evidence to support the studies of Kroc et al (2019a, 2019b) and hypothesis that *iucundus* is an ethylene response factor that regulates grain alkaloid content in NLL.

(iii) Classical disease resistance genes are underrepresented in the NLL pan-genome

Classical disease resistance (*R*-) genes of the nucleotide binding site, leucine rich repeats (NLRs) family are key genes at the frontline of plant defence to numerous pathogens and pests. NLR genes were identified in the reference genome using two approaches. Firstly, conserved consensus sequences from plant extended nucleotide binding site (NBS) domains using the sequences described in Ameline-Torregrosa et al. (2008) were used to identify genes in the pan-genome annotation. Secondly, genes containing the PFAM domain PF00931 (NB-ARC) identified using Interproscan were identified as classical NLR genes. This approach identified a total of 67 NLRs in the NLL pan-genome, a surprisingly low number compared to other legume crops. Thirdly, RGAugury (Li *et al.*, 2016) was used, which identified 67 NLRs and 48 receptor-like proteins. We compared the NLR domains identified above with 32 NLR domain sequences that we derived from PCR amplification of the NLL genome using degenerate primers derived from *M. truncatula* conserved NBS/NB-ARC sequences. All 32 experimentally verified genes were contained within the 68 predicted NLR gene set and no additional NLR homologues were identified.

In the pan-genome contigs only one additional NLR was identified. Of the 68 NLRs identified in the pan-genome, 62 NLRs were found in all 55 NLL accessions, whereas 6 were found to be members of the variable gene set. Only one variable NLR gene came from a new pan-genome contig (*Lupan_039653*), with the remainder present in the Tanjil reference genome. All but two NLRs (*Lupan_039653* and *Lupan_026540*) were anchored in pseudochromosomes and were distributed across all pseudochromosomes except NLL-14 and NLL-16. Clusters of three or more NLR homologues on a pseudochromosome were identified on NLL-02 and NLL-06, which also harboured the largest number of NLRs with 13 and 10 genes, respectively (Fig. 4; Table S8).

Discussion

Comparison of different lupin reference genomes

As long read sequencing technologies have become more affordable the number of more complete and updated plant genome assemblies has increased. For NLL, the first draft assembly of 523 Mb was published in 2013 (Yang et al., 2013), followed by the first reference assembly of 609 Mb (Hane et al., 2017). Both assemblies used short read sequencing data and were fragmented with scaffold N50 values of 7,319 and 232, respectively. More recently, Wang and colleagues (2021) published an improved assembly incorporating PacBio data, which resulted in a 616 Mb assembly and N50 and L50 values of 9 and 30.8 Mb, respectively, which are comparable to the 653 Mb assembly with N50 and L50 values of 9 and 30.7 presented herein (Table 1). The mean coding sequence length for the new assembly compared to the Hane et al (2017) assembly increased from 1,289 to 1,337 bp, while the GC content of all the various NLL assemblies was between 33.27 and 33.46 %. The latter is similar to the GC content observed in the white lupin genome assembly of Hufnagel and colleagues (2020) of 33.79%, whereas another white lupin genome assembly had a GC content of 36.82 % (Xu et al., 2020).

Compared to previous assemblies the number of annotated protein-coding sequences increased from 33,076 and 33,097 in the Hane et al (2017) and Wang et al (2021) assemblies respectively, to 38,545 in the current assembly. The gene content of NLL identified here is like that reported for white lupin by Hufnagel and colleagues (2020) who predicted 38,258 genes, whereas the white lupin assembly by Xu and associates (2020) was predicted to contain 47,603 genes.

To validate the completeness of gene coding content we conducted BUSCO analysis of the different NLL and white lupin genomes (Table 1). Comparing the different NLL assemblies in both the embryophyte and plantae BUSCO analyses the improved assembly presented herein had the most complete and fewest fragmented BUSCOs. Alignment of RNA sequencing reads from different tissues further corroborated that between 89.1-99.0 % mapped back to the improved assembly, indicating it has captured most of the gene-rich space. Differences in completeness of the gene content has also been observed between the different white lupin assemblies (Table 1), with the Hufnagel assembly having a similar number of missing and fragmented BUSCOs. Two of the missing plantae BUSCOs (31855at33090 and 40630at33090) were absent in both lupin species and were subsequently found to also be absent in other legume genomes except for 31855at33090 in chickpea and 40630at33090 in lotus (Table 2). It is thus possible that some conserved plantae genes have been lost in Fabaceae species and can therefore affect the completeness scores for legumes. This is supported by the re-sequencing

of the additional 54 NLL accessions herein, which didn't uncover any of the missing BUSCOs (Table 1).

Features of the NLL pan-genome

To capture as much of the genetic diversity in the NLL pan-genome, wild and domestic lines from distinct clades identified in a previous diversity study were selected as representatives for each clade (Mousavi-Derazmahalleh et al. 2018b). To assemble the NLL pan-genome an iterative read mapping approach to the reference was used whereafter the unaligned reads were *de novo* assembled. This added 300,705 contigs to the reference assembly and produced a pan-genome size of 975 Mb. This assembly size is slightly larger than the estimated diploid genome size for cultivar Tanjil of 951 Mb by *k*-mer analysis or 924 Mb by flow cytometry (Hane et al., 2017). The *k*-mer based phylogenetic tree also showed a similar relationship as the previous analyses using molecular markers (Mousavi-Derazmahalleh *et al.*, 2018a, Mousavi-Derazmahalleh et al., 2018b) with three clear groupings of: i) domesticated varieties; ii) wild accessions from the Iberian peninsula and North Africa; iii) wild accessions from Southern Europe.

In terms of gene content, 39,339 genes were predicted, with the 794 additional genes identified in the new pan-genome contigs having a shorter average gene length than those genes identified in the reference genome. This could be due to small contigs that might contain partial/fragmented genes or due to expanding/evolving gene families. Of the variable genes, 37 were absent from all wild accessions, but present in all domesticated material, while 127 were present in all wild accessions and absent in all domesticated lines. Of the 127 genes only found in wild accessions, only five had a Pfam domain, while 29 of the 37 only found in domesticated lines, had a Pfam domain, with no obvious over-representation of specific Pfam domains.

Comparison of the NLL and white lupin pan-genomes

At this stage it is difficult to conduct a direct comparison of the NLL and white lupin pan-genomes. The first reason for this is because two different assembly approaches have been used, with the white lupin assembly adopting a *de novo* assembly approach of each of the 39 genomes and aligning these, whereas for NLL we aligned the sequence reads to the reference and *de novo* assembled the unaligned reads in an iterative approach. Both approaches have their own advantages and disadvantages as described in a recent review (Bayer et al., 2020). For example, we observed that the NLL pan-genome size increased 322 Mb in size relative to

the reference genome, whereas the white lupin pan-genome assembly approach added 11.7 Mb. This could be due to the approach taken or the second large difference, which is the ratio of wild versus domesticated accessions used in the pan-genome assemblies. The NLL pan-genome utilised many wild accessions (31 out of 55) while the white lupin pan-genome was more focused on domesticated accessions that are closely related, and only used four wild accessions out of a total of 39 accessions. As high-quality long read sequencing data generation becomes cheaper, a pan-genome graph approach will likely be more common and overcome some of the difficulties observed in the NLL and white lupin pan-genome assembly approaches (Bayer et al., 2020).

Despite these differences an interesting observation was that in both the NLL and white lupin pan-genomes (Hufnagel et al., 2021) a relatively low number of variable genes were identified (4.8 and 3.1 % respectively) compared to other crop pan-genomes such as *Brassica oleracea* (20%) (Golicz *et al.*, 2016)and wheat (35.7%) (Montenegro *et al.*, 2017) . Specifically, 1,195 variable genes were identified for white lupin (1,132 present in the reference assembly and 63 identified from the new pan-genome contigs) and 1,880 variable genes were identified for NLL, including 1,086 that reside in the reference genome and 794 from contigs produced through the pan-genome assembly. The additional pan-genome contigs relative to the reference assembly are also more sparsely populated with genes (an average of one gene per 16.9 kb vs one gene per 405.5 kb). This suggests that for both lupin species the level of diversity in wild and domesticated material is quite low compared to many other crop species. In both lupins, wild accessions have a higher number of variable genes than domesticated accessions. In both pan-genomes, genes specifically involved in mycorrhizal symbiosis were absent and classical disease resistance genes of the NLR type were significantly underrepresented compared to other legume genomes as outlined in more detail below.

NLL lacks genes specifically associated with arbuscular mycorrhizal symbiosis

Most legumes undergo beneficial interactions with both rhizobia and mycorrhizal fungi to help fix nitrogen and acquire phosphorous, respectively. The legume genes associated with these interactions are well understood and include genes that overlap between rhizobial and mycorrhizal associations and those that are unique to each interaction. NLL only interacts with *Rhizobium* and an analysis of the reference and pan-genome revealed that NLL lacked nearly all genes that are specifically required for arbuscular mycorrhizal symbiosis. There were two exceptions, and in both cases the genes were found to be core genes in the pan-genome. The genes were *PP2AB'1* (*Lupan_009748*) and *KIN2* (*Lupan_004177*) homologues, which encode

serine/threonine protein phosphatase and a protein kinase (Table S7). Genes with such annotations are known to have additional roles beyond symbiosis and thus could be involved in other NLL biological processes. Nevertheless, these findings conclusively demonstrate that essential mycorrhizal associated genes are absent in the NLL pan-genome as also reported for the white lupin (pan)-genome (Hufnagel et al., 2020, Hufnagel et al., 2021).

Analysis of grain alkaloid content of pan-genome lines identifies a candidate for the *iucundus* locus

Quinolizidine alkaloids in NLL are regulated by a major locus termed *iucundus* (Frick et al., 2017). In the literature two different genes have been proposed as (Kroc et al., 2019a, Kroc et al., 2019b) (Wang et al., 2021). Assessment of the relationship between grain alkaloid content (as determined via Dragendorff assays) and the proposed causal SNP mutations in these two candidates showed strong correlation between alkaloid content and causal SNP genotype for *LaRAP2-7* as opposed to *LaRSTK* (Table 3). Our results thus support the previous findings from Kroc and colleagues (2019a, 2019b), that *LaRAP2-7* is a strong candidate for regulation of grain alkaloid content. However, final validation through cloning and complementation or targeted gene-editing studies is required to conclusively prove that *RAP2-7* is the causal gene.

Classical disease resistance genes of the NLR class are underrepresented in the lupin genome

Analysis of NLR genes in the previous reference Tanjil genome assembly (Hane et al., 2017) and the updated version herein was conducted and revealed only 67 classical disease resistance genes of the NLR class, which is a relatively low number compared to other crops that typically contain several hundred of such genes. We were therefore interested to see if other domesticated and wild accessions had additional NLRs that weren't present in the reference assembly. However, examination of the pan-genome only identified a total of 68 NLR genes, of which 67 were in common with Tanjil, demonstrating both a surprisingly low number and high conservation in the NLR gene repertoire of NLL among both wild and domesticated accessions. NLRs can have a high sequence similarity and it is thus possible that additional NLRs were not identified in the pan-genome assembly as their reads aligned to other NLRs in the assembly as a short-read sequencing approach was used.

We also mined the reference genome of white lupin (Hugnagel et al 2020) and identified 44 NLRs. The 67 and 44 NLRs in the reference NLL and white lupin genomes, respectively, are considerably less than those observed in other closely related legumes such as *M. truncatula* (764), soybean (506), pigeonpea (406) and chickpea (187), or other key crops such as wheat,

which contains 604 NB-ARC genes (Chandra *et al.*, 2017). In the other sequenced legume species, NLRs are often found in clusters, where members of these clusters have a high sequence similarity as they have duplicated and diverged. Surprisingly, we only found two such clusters (on chromosomes NLL-02 and NLL-06), while other NLRs were distributed as single loci across the NLL genome. The reasons for this relatively low number of NLRs in the two examined lupin species remain unclear, but the inherent high levels of alkaloid content in wild lupin accessions which provide protection from pathogens and pests could be a possible explanation.

Conclusion

Legumes, including domesticated lupin species, are an important part for future sustainable agricultural production systems, especially with the growing demand for plant-based protein for not only animal feed but also human consumption. Two lupin species (NLL and white lupin) now have well-advanced genomic resources including transcriptome datasets and reference/pan-genomes. This will likely be further augmented by ongoing efforts to generate similar resources for the other two domesticated lupin species (yellow and pearl lupin) and genomic datasets for other New World lupin species. The combination of these resources will allow a better understanding of lupin biology, diversity and evolution and help accelerate lupin crop improvement. For example, the NLL pan-genome resources have contributed to the generation of a 30k multispecies pulse SNP chip (Kaur *et al.*, 2020), which includes 5,425 evenly distributed SNPs across the 20 chromosomes of NLL. In addition, the SNPs identified in the pan-genome can help identify candidate genes for important agronomic traits such as the alkaloid gene identified herein and a novel flowering time gene (Taylor *et al.*, 2021). The improvement in yield and other agronomic traits of interest to lupin breeding programs will help provide viable lupin crops for substantial parts of Mediterranean grain growing regions around the globe.

Experimental procedures

Sequencing of NLL lines

Narrow-leafed lupin seed was provided by the Department of Primary Industries and Regional Development (DPIRD) of Western Australia. Seeds were vernalised in Jiffy pellets for two weeks at 4°C in temperature-controlled growth cabinets in the dark before being planted in pots in temperature-controlled glasshouses to collect tissue for DNA extraction and seed for future experiments. DNA was isolated from frozen leaf material that was ground to a fine

powder under a stream of liquid nitrogen, using the CTAB method as described by Kamphuis *et al.* (2008). DNA quality was assessed on a 1%-agarose gel and DNA concentration measured using a Qubit. Samples were submitted to the Australian Genome Research Facility (AGRF) to generate Paired-End and Mate-Pair libraries, which were sequenced on an Illumina HiSeq. Raw data yield for the different lines varied from low coverage lines (13.8-25.9x) to high coverage lines (56.0-72.3x) based on estimated genome size of 951 Mb for NLL assembly (Hane *et. al*, 2017). High molecular weight DNA was extracted as above and send to the Ramaciotti Centre for PacBio sequencing on 19 SMRT cells, which yielded 91 Gb (~98x coverage). *In* situ Hi-C was performed as described previously (Kaur *et al.*, 2021), where frozen leaf tissue was crosslinked, ground and lysed to obtain permeabilized intact nuclei. DNA from the nuclei was digested with *Mbo*I and the overhangs filled in with a biotinylated base. Free ends were subsequently ligated together. Crosslinks were reversed, the DNA sheared to 300–500 bp and then biotinylated ligation junctions were recovered with streptavidin beads. This sample was then subjected to a standard Illumina Paired End library preparation and subjected to sequencing on an Illumina Hiseq.

Reference genome and pan-genome assembly and validation

The Tanjil reference genome was assembled using CANU (v1.8; Koren et al., 2017) and scaffolded with the Hi-C data using the 3D *de novo* assembly (3D-DNA) pipeline (Dudchenko et al., 2017). The resulting assembly was polished using the Juicebox Assembly Tools (Dudchenko et al., 2018) and the contact maps visualized using the Juicebox visualization software (Durand et al., 2016). The pan-genome was assembled using an iterative mapping and assembly approach described by Hu et al., (2020). The improved reference genome assembly of Tanjil presented herein was used as a reference for pan-genome construction. The procedure involved three main steps: mapping of the reads to the reference sequence; assembly of the unmapped reads; and production of a new reference sequence by updating the old one with the newly assembled contigs. The high coverage accessions were assembled first in increasing order of minimum percentage of reads mapped to the Tanjil genome assembly. Unmapped data from the low coverage accessions along with Tanjil and 83A476 was combined for assembly. The mapping and assembly were performed in the following order: Unicrop; Bolero; P26625; P22872; P27255; P25136; P26167; P26464; P22660; P26170; low coverage accessions. Mapping was performed using Bowtie2 v2.3.4 (--end-to-end --sensitive -X 1000) and assembly was performed using MaSuRCA v3.3.3 (cgwErrorRate=0.15, JF_SIZE = 1200000000) (Zimin *et al.*, 2013). The assembled contig sequences were compared with the NCBI nt database (downloaded 18 May, 2021) using BLAST v2.5.0. Contigs with best hits to non-green plants,

chloroplast or mitochondrial sequences were removed. The completeness of core eukaryotic genes in the pan-genome assembly was evaluated using the default parameters of BUSCO v4.1 for plantae and embryophyte (Simao et al., 2015).

Reference genome and pan-genome annotation

Repetitive DNA regions were first predicted in both the reference and pan-genome for both transposable elements (TEs) and tandem repeats as described previously by Hane et al (2017). Briefly, annotation of TEs was conducted using RepeatMasker v4.1.0 (Smit *et al.*, 2013-2015) to identify known repeat sequences in Repbase v26.10 (Jurka *et al.*, 2005). Non-coding RNA was predicted using Rfam 14.6 (July 2021, 4,070 families) and Infernal v 1.1.3 (Nawrocki & Eddy, 2013).

Following repeat masking the reference genome and newly assembled pan-genome contigs greater than 1 Kb in length were annotated using BRAKER version 2.1.2 (Hoff *et al.*, 2019). *De novo* gene prediction was conducted using Genemark-ES (Lomsadze *et al.*, 2005) and Augustus (Stanke *et al.*, 2004). Proteins from previously published lupin and soybean genomes (Hane et al., 2017, Hufnagel et al., 2020, Schmutz et al., 2010) were used as external validation. Annotations were filtered for incomplete models and longest isoforms using AGAT version 0.5.1 (Dainat *et al.*, 2021). The pan-genome was functionally annotated using Blast2GO CSIRO server (Conesa *et al.*, 2005). To identify protein domains and map protein families, Interproscan v 5.24-63.0 (Quevillon *et al.*, 2005) was used. Both programs were run separately for core and variable genes.

Gene presence/absence variation analysis and SNP discovery

Gene presence/absence variation was characterized using the SGSGeneLoss package (Golicz *et al.*, 2015), using the default parameters. Reads from all the NLL lines were mapped to the pan-genome using Bowtie2 v2.3.4 (--end-to-end --sensitive –X 1000) (Langmead & Salzberg, 2012). Based on presence/absence variation, gene models were divided into core and variable genes. A gene was considered core if it was present in all lines and variable if it was absent in at least one line. Curves describing pan-genome size, core genome size and new gene number for individual genes were drawn using PanGP program (Algorithm: Distance guide; Sample size: 1000; Sample repeat: 100; Amplification coefficient: 50) (Zhao *et al.*, 2014). Following the read mapping using Bowtie2 as described above, duplicate sequences were marked using Picard toolkit (2019) MarkDuplicates (https://broadinstitute.github.io/picard/) and SNPs were

discovered and filtered for low quality using SAMtools v1.10., mpileup (Li *et al.*, 2009) and BCFtools v1.3 (Danecek *et al.*, 2021).

Symbiosis analysis

Fifty-seven *M. truncatula* genes associated with mycorrhizal association that were previously used to identify homologues in the *L. angustifolius* genome (Hane et al., 2017) were used to identify annotated homologous genes in the pan-genome by BLASTP (Altschul *et al.*, 1997). Best hits were subsequently used to identify the old reference gene annotation identifiers. The legume gene family trees on the legume information system portal (https://legumeinfo.org/; Gonzales et al., 2004, Dash et al., 2016) for the *M. truncatula* and *L. angustifolius* genes were used to validate that these were true homologues.

Alkaloid analysis

Seed of 54 accessions of the NLL pan-genome were evaluated for grain alkaloid content using the Dragendorff test. Dragendorff reagent was prepared by dissolving 32 g of potassium iodide in 80 mL sterile water and subsequently mixed in a solution containing 3.4 g bismuth nitrate, 40 mL glacial acetic acid and 160 mL of sterile water to make a basal solution. One part basal solution was mixed with one part glacial acetic acid and four parts sterile water to make working solution. Whatman filter papers were soaked in this solution and allowed to dry overnight. Prepared Dragendorff filter papers were stored in the dark. Alkaloid tests were carried out by removing the seed coat and soaking crushed cotyledons overnight in 200 µL of sterile water. Subsequently, 5 µL was placed on the Dragendorff paper. Bitter seeds were identified by an orange/pink reaction with the reagent, while sweet seeds do not react with the reagent. Seeds of each of the accessions were analysed in triplicates.

NLR resistance gene analysis

Candidate genes containing nucleotide binding site (NBS/ NB-ARC) domains were identified in the NLL pan-genome using blastp similarity (Altschul et al., 1997) with 1e-5 E-value threshold to consensus coiled-coil (CC)-NBS-LRR and Toll and mammalian INTERLEUKIN1 receptor-NBS-LRR sequences from plant extended NBS domains using the sequences described in Ameline-Torregrosa et al (2008). In addition, RGAugury (v1) (Li et al., 2016) was used to identify resistance gene analogues. Furthermore, Interproscan (v 5.24-63.0) (Quevillon et al., 2005) was used to identify genes annotated with the PF00931 (NB-ARC) domain, using the annotated NLL pan-genome protein file (Data S5) and the white lupin genome annotated protein file (GCA_009771035.1) from Hufnagel et al (2020).

Accession numbers

Genome sequence assembly and annotation data can be found in GenBank under BioProject IDs: PRJNA797109, PRJNA512907 and PRJNA299755. In addition, the chromosome length and pan-genome assemblies and annotated gene and protein sequence can be accessed from the downloads section of our webpage: http://lupinexpress.org/. Furthermore, a JBrowse instance has been put up to browse the genomes and a BLAST sever to query the genomes and associated annotated gene/protein sets. The interactive Hi-C contact map for the genome is also available at www.dnazoo.org.

Acknowledgements

The authors are grateful to CSIRO, Curtin University, the University of Western Australia and the Grains Research and Development Corporation (GRDC, Project code CSP1806-009RTX) for their financial support. E.L.A. was supported by the Welch Foundation (Q-1866), a McNair Medical Institute Scholar Award, an NIH Encyclopedia of DNA Elements Mapping Center Award (UM1HG009375), a US-Israel Binational Science Foundation Award (2019276), the Behavioral Plasticity Research Institute (NSF DBI-2021795), NSF Physics Frontiers Center Award (NSF PHY-2019745), and an NIH CEGS (RM1HG011016-01A1). The authors thank Hayley Casarotto, Natalie Fletcher, Nick Pain and Daniel Lim for technical assistance on the project and Drs Jonathan Anderson and Jana Sperschneider for helpful comments on the manuscript. Hi-C data were created in collaboration with the DNA Zoo Consortium (www.dnazoo.org). DNA Zoo is supported by Illumina, Inc.; IBM; and the Pawsey Supercomputing Center.

Conflict of Interest

The authors declare no conflict of interest or competing interests.

Author contributions

LGK, RF, DE and KBS conceived the project. LGK, CMT and PK isolated the DNA from the NLL accessions for sequencing. PK, CMT and LGK generated the Hi-C library. PK, OD and ELA generated the Hi-C data and Hi-C guided assembly. GG, PB and PK generated the reference and pan-genome assemblies and associated annotations. LGK, LG, KF and RF conducted the NLR, alkaloid and conglutin experiments/analysis. LGK wrote the first draft of the manuscript with input from GG and KBS. All authors reviewed the manuscript and agreed on the final version of the manuscript.

Short legends for Supporting Information

Table S1. Summary of the total amount of sequence data generated for the *Lupinus angustifolius* cv. Tanjil genome assembly and the average coverage per library, assuming an estimated genome size of 951 Mb based on *k*-mer analysis (Hane *et.al*, 2017).

Table S2. Overview of the repeat content in the 4th generation of the narrow-leafed lupin (*Lupinus angustifolius*) reference genome assembly of cultivar Tanjil.

Table S3. Summary of non-coding RNA predicted within the narrow-leafed lupin (*Lupinus angustifolius*) cv. Tanjil genome assembly.

Table S4. Overview of the sequence coverage for the 55 *Lupinus angustifolius* accessions that were used to assemble the pan-genome.

Table S5. Summary of non-coding RNA predicted within the narrow-leafed lupin (*Lupinus angustifolius*) pan genome assembly.

Table S6. Top 10 most represented Pfam protein domains found in variable and core genes of the narrow-leafed lupin (*Lupinus angustifolius*) pan-genome using Interproscan.

Table S7. Overview of genes associated with arbuscular mycorrhizal and rhizobial associations in the genomes of *Medicago truncatula* and narrow-leafed lupin (*Lupinus angustifolius*).

Table S8. Overview of the distribution of NBS-LRR genes across the 20 linkage groups of NLL cv. Tanjil.

Data S1. Annotated genes in the *Lupinus angustifolius* reference genome.

Data S2. Protein coding sequences in the *Lupinus angustifolius* reference genome.

Data S3. Annotated genes in the *Lupinus angustifolius* pan-genome.

Data S4. Protein coding sequences in the *Lupinus angustifolius* pan-genome.

Data S5. Functional annotations assigned to gene annotations of the *Lupinus angustifolius* genome.

Data S6. References for Table S7.

References

(2019) Picard Toolkit. *Broad Institute, GitHub Repository*.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W.*, et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* **25,** 3389-3402.

Ameline-Torregrosa, C., Wang, B.-B., O'Bleness, M. S., Deshpande, S., Zhu, H., Roe, B.*, et al.* (2008) Identification and characterization of nucleotide-binding site-leucine-rich repeat genes in the model plant *Medicago truncatula*. *Plant Physiology,* **146,** 5-21.

Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J. and Edwards, D. (2020) Plant pan-genomes are the new reference. *Nature Plants,* **6,** 914-920.

Berger, J., Buirchell, B., Luckett, D. and Nelson, M. (2012) Domestication bottlenecks limit genetic diversity and constrain adaptation in narrow-leafed lupin (*Lupinus angustifolius* L.). *Theoretical and Applied Genetics,* **124,** 637-652.

Berger, J. D., Clements, J. C., Nelson, M. N., Kamphuis, L. G., Singh, K. B. and Buirchell, B. (2013) The essential role of genetic resources in narrow-leafed lupin improvement. *Crop and Pasture Science,* **64,** 361-373.

Bertioli, D. J., Jenkins, J., Clevenger, J., Dudchenko, O., Gao, D., Seijo, G.*, et al.* (2019) The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. *Nature Genetics,* **51,** 877-884.

Bielski, W., Ksiazkiewicz, M., Simonikova, D., Hribova, E., Susek, K. and Naganowska, B. (2020) The puzzling fate of a lupin chromosome revealed by reciprocal oligo-FISH and BAC-FISH mapping. *Genes,* **11,** 1489.

Chandra, S., Kazmi, A. Z., Ahmed, Z., Roychowdhury, G., Kumari, V., Kumar, M.*, et al.* (2017) Genome-wide identification and characterization of NB-ARC resistant genes in wheat (*Triticum aestivum* L.) and their expression during leaf rust infection. *Plant Cell Reports,* **36,** 1097-1112.

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

Dainat, J., Hereñú, D. and Git, P. (2021) AGAT: Analysis Toolkit to handle annotations in any GTF/GFF format. In: *Zenodo*. pp. https://doi.org/10.5281/zenodo.5336786.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O.*, et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience,* **10,** giab008.

Dash, S., Campbell, J. D., Cannon, E. K. S., Cleary, A. M., Huang, W., Kalberer, S. R.*, et al.* (2016) Legume information system (LegumeInfo.org): a key component of a set of federated data resources for the legume family. *Nucleic Acids Research,* **44,** D1181-D1188.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C.*, et al.* (2017) De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science,* **356,** 92-95.

Dudchenko, O., Shamim, M. S., Batra, S., Durand, N. C., Musial, N. T., Mostofa, R.*, et al.* (2018) The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. *bioRxiv* 254797.

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S.*, et al.* (2016) Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems,* **3,** 95-98.

FAO (2021) *Statiscical year book 2021: World Food and Agriculture*. Rome: Food and Agriculture Organization of the United Nations.

Frick, K. M., Foley, R., Siddique, K. H. M., Singh, K. B. and Kamphuis, L. G. (2019) The role of jasmonate signalling in quinolizidine alkaloid biosynthesis, wounding and aphid predation response in narrow-leafed lupin. *Functional Plant Biology,* **46,** 443-454.

Frick, K. M., Foley, R. C., Kamphuis, L. G., Siddique, K. H. M., Garg, G. and Singh, K. B. (2018) Characterisation of the genetic factors affecting quinolizidine alkaloid biosynthesis and its response to abiotic stress in narrow-leafed lupin (*Lupinus angustifolius* L.). *Plant Cell and Environment,* **41,** 2155-2168.

Frick, K. M., Kamphuis, L. G., Siddique, K. H. M., Singh, K. B. and Foley, R. C. (2017) Quinolizidine alkaloid biosynthesis in lupins and prospects for grain quality improvement. *Frontiers in Plant Science,* **8,** 87.

Gao, L. L., Hane, J. K., Kamphuis, L. G., Foley, R., Shi, B. J., Atkins, C. A.*, et al.* (2011) Development of genomic resources for the narrow-leafed lupin (*Lupinus angustifolius*): construction of a bacterial artificial chromosome (BAC) library and BAC-end sequencing. *BMC Genomics,* **12,** 521.

Gladstones, J. (1970) Lupins as crop plants. *Field Crop Abstracts,* **23,** 123-147.

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A.*, et al.* (2016) The pangenome of an agronomically important crop plant Brassica oleracea. *Nature Communications,* **7,** 13390.

Golicz, A. A., Martinez, P. A., Zander, M., Patel, D. A., Van De Wouw, A. P., Visendi, P.*, et al.* (2015) Gene loss in the fungal canola pathogen *Leptosphaeria maculans*. *Functional and Integrative Genomics,* **15,** 189–196.

Gonzales, M. D., Archuleta, E., Farmer, A., Gajendran, K., Grant, D., Shoemaker, R.*, et al.* (2004) The Legume Information System (LIS): An integrated information resource for comparative legume biology. *Nucleic Acids Research,* **33,** D660-D665.

Hane, J., Ming, Y., Kamphuis, L. G., Nelson, M. N., Garg, G., Atkins, C. A.*, et al.* (2017) A comprehensive draft genome sequence for lupin (*Lupinus angustifolius*), an emerging health food: insights into plant–microbe interactions and legume evolution. *Plant Biotechnology Journal,* **15,** 318-330.

Hirakawa, H., Kaur, P., Shirasawa, K., Nichols, P., Nagano, S., Appels, R.*, et al.* (2016) Draft genome sequence of subterranean clover, a reference for genus *Trifolium*. *Scientific Reports,* **6,** 30358.

Hoff, K. J., Lomsadze, A., Borodovsky, M. and Stanke, M. (2019) Whole-Genome Annotation with BRAKER. *Methods in Molecular Biology,* **1962,** 65-95.

Hu, H., Yuan, Y., Bayer, P. E., Fernandez, C. T., Scheben, A., Golicz, A. A.*, et al.* (2020) Legume pangenome construction using an iterative mapping and assembly approach. In: *Legume Genomics: Methods and Protocols, Methods in Molecular Biology*. (Garg, M. J. a. R., ed.). New York: Springer Nature, pp. 35-47.

Hufnagel, B., Marques, A., Soriano, A., Marquès, L., Divol, F., Doumas, P.*, et al.* (2020) High-quality genome sequence of white lupin provides insight into soil exploration and seed quality. *Nature Communications,* **11,** 492.

Hufnagel, B., Soriano, A., Taylor, J., Divol, F., Kroc, M., Sanders, H.*, et al.* (2021) Pangenome of white lupin provides insights into the diversity of the species. *Plant Biotechnology Journal,* **19,** 2532-2543.

Jimenez-Lopez, J. C. (2020) Narrow-leafed lupin (*Lupinus angustifolius* L.) β-conglutin: A multifunctional family of proteins with roles in plant defence, human health benefits and potential uses as functional food. *Legume Science,* **2,** e33.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, j. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research,* **110,** 462-467.

Kamphuis, L. G., Foley, R., Frick, K. M., Garg, G. and Singh, K. B. (2020) Transcriptome resources paving the way for lupin crop improvement. In: *Compendium of Plant Genomes, The Lupin Genome.* (Singh, K. B., Kamphuis, L. G. and Nelson, M. N., eds.). Cham, Switzerland: Springer Nature.

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

Kang, Y. J., Kim, S. K., Kim, M. Y., Lestari, P., Kim, K. H., Ha, B.-K*., et al.* (2014) Genome sequence of mungbean and insights into evolution within *Vigna* species. *Nature Communications,* **5,** 5443.

Kasprzak, A., Safar, J., Janda, J., Dolezel, J., Wolko, B. and Naganowska, B. (2006) The bacterial artificial chromosome (BAC) library of the narrow-leafed lupin (*Lupinus angustifolius* L.). *Cellular and Molecular Biology Letters,* **11,** 396-407.

Kaur, P., Lui, C., Dudchenko, O., Nandety, R. S., Hurgobin, B., Pham, M*., et al.* (2021) Delineating the *Tnt1* insertion landscape of the model legume *Medicago truncatula* cv. R108 at the Hi-C resolution using a chromosome-length genome assembly. *International Journal of Molecular Sciences,* **22,** 4326.

Kaur, S., Keeble-Gagnere, G., Pasam, R. K. and Hayden, M. (2020) Pulse pipeline integration to become more efficien. In: *GRDC Groundcover.* Canberra, Australia: GRDC.

Kohajdová, Z., Karovičová, J. and Schmidt, Š. (2011) Lupin composition and possible use in bakery – a review. *Czech Journal of Food Sciences,* **29,** 203-211.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergham, N. H. and Phillippy, A. M. (2017) Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Research,* **27,** 722-736.

Kreplak, J., Madoui, M.-A., Cápal, P., Novák, P., Labadie, K., Aubert, G*., et al.* (2019) A reference genome for pea provides insight into legume genome evolution. *Nature Genetics,* **51,** 1411-1422.

Kroc, M., Czepiel, K., Wilczura, P., Mokrzycka, M. and Swiecicki, W. (2019a) Development and validation of a gene-targeted dCAPS marker for marker-assisted selection of low-alkaloid content in seeds of narrow-leafed lupin (*Lupinus angustifolius* L.). *Genes,* **10,** 428.

Kroc, M., Koczyk, G., Kamel, K. A., Czepiel, K., Fedorowicz-Strońska, O., Krajewski, P*., et al.* (2019b) Transcriptome-derived investigation of biosynthesis of quinolizidine alkaloids in narrow-leafed lupin (*Lupinus angustifolius* L.) highlights candidate genes linked to iucundus locus. *Scientific Reports,* **19,** 2231.

Lambers, H., Clements, J. C. and Nelson, M. N. (2013) How a phosphorusacquisition strategy based on carboxylate exudation powers the success and agronomic potential of lupines (*Lupinus*, Fabaceae). *American Journal of Botany,* **100,** 263-288.

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

Lee, Y. P., Mori, T. A., Sipsas, S., Barden, A., Puddey, I. B., Burke, V*., et al.* (2006) Lupin-enriched bread increases satiety and reduces energy intake acutely. *The American journal of clinical nutrition,* **84,** 975-980.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N*., et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics,* **25,** 2078-2079.

Li, P., Quan, X., Jia, G., Xiao, J., Cloutier, S. and You, F. M. (2016) RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics,* **17,** 852.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. and Borodovsky, M. (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research,* **33,** 6494-6506.

Lonardi, S., Muñoz-Amatriaín, M., Liang, Q., Shu, S., Wanamaker, S. I., Lo, S*., et al.* (2019) The genome of cowpea (*Vigna unguiculata* [L.] Walp.). *The Plant Journal,* **98,** 767-782.

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C.-K. K*., et al.* (2017) The pangenome of hexaploid bread wheat. *The Plant Journal,* **90,** 1007-1013.

Mousavi-Derazmahalleh, M., Bayer, P. E., Nevado, B., Hurgobin, B., Filatov, D., Kilian, A., *et al.* (2018a) Exploring the genetic and adaptive diversity of a pan-Mediterranean crop wild relative: narrow-leafed lupin. *Theoretical and Applied Genetics,* **131,** 887-901.

Mousavi-Derazmahalleh, M., Nevado, B., Bayer, P. E., Filatov, D. A., Hane, J. K., Edwards, D., *et al.* (2018b) The western Mediterranean region provided the founder population of domesticated narrow-leafed lupin. *Theoretical and Applied Genetics,* **131,** 2543-2554.

Nawrocki, E. P. and Eddy, S. R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics,* **29,** 2933-2935.

Nevado, B., Atchison, G. W., Hughes, C. E. and Filatov, D. A. (2016) Widespread adaptive evolution during repeated evoluntionary radiations in New World lupins. *Nature Communications,* **8,** 12384.

Parniske, M. (2008) Arbuscular mycorrhiza: the mother of plant root endosymbioses. *Nature Reviews Microbiology,* **6,** 763-775.

Parween, S., Nawaz, K., Roy, R., Pole, A. K., Suresh, B. V., Misra, G., *et al.* (2015) An advanced draft genome assembly of a desi type chickpea (*Cicer arietinum* L.). *Scientific Reports,* **5,** 1280.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., *et al.* (2005) InterProScan: protein domains identifier. *Nucleic Acids Research,* **33,** W116-W120.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Kato, T., Nakao, M., *et al.* (2008) Genome structure of the legume, *Lotus japonicus*. *DNA Research,* **15,** 227-239.

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *nature,* **463,** 178-183.

Schmutz, J., McClean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., *et al.* (2014) A reference genome for common bean and genome-wide analysis of dual domestications. *Nature genetics,* **46,** 707-713.

Seymour, M., Kirkegaard, J. A., Peoples, M. B., White, P. F. and French, R. J. (2012) Break-crop benefits to wheat in Western Australia—insights from over three decades of research. *Crop and Pasture Science,* **63,** 1-16.

Shen, C., Du, H., Chen, Z., Lu, H., Zhu, F., Chen, H., *et al.* (2020) The chromosome-level genome sequence of the autotetraploid alfalfa and resequencing of core germplasms provide genomic resources for alfalfa research. *Molecular Plant,* **13,** 1250-1261.

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. and Zdobnov, E. M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics,* **31,** 3210-3212.

Smit, A. F. A., Hubley, R., Green, P. and 2013-2015 (2013-2015) RepeatMasker Open-4.0. *http://www.repeatmasker.org*.

Stanke, M., Steinkamp, R., Waack, S. and Morgenstern, B. (2004) AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research,* **32,** W309-W312.

Taylor, C. M., Garg, G., Berger, J. D., Ribalta, F. M., Croser, J. S., Singh, K. B., *et al.* (2021) A *Trimethylguanosine Synthase1-like* (*TGS1*) homologue is implicated in vernalisation and flowering time control. *Theoretical and Applied Genetics,* **134,** 3411-3426.

Varshney, R. K., Chen, W., Li, Y., Bharti, A. K., Saxena, R. K., Schlueter, J. A., *et al.* (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nature biotechnology,* **30,** 83-89.

Varshney, R. K., Song, C., Saxena, R. K., Azam, S., Yu, S., Sharpe, A. G., *et al.* (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nature biotechnology,* **31,** 240-246.

Wang, P., Zhou, G., Jian, J., Yang, H., Renshaw, D., Aubert, M., *et al.* (2021) Whole-genome assembly and resequencing reveal genomic imprint and key genes of rapid domestication in narrow-leafed lupin. *The Plant Journal,* **105,** 1192-1210.

White, C. L., Staines, V. E. and Staines, M. H. (2007) A review of the nutritional value of lupins for dairy cows. *Australian Journal of Agricultural Research,* **58,** 185-202.

Xu, W., Zhang, Q., Yuan, W., Xu, F., Muhammad Aslam, M., Miao, R.*, et al.* (2020) The genome evolution and low-phosphorus adaptation in white lupin. *Nature Communications,* **11,** 1069.

Yang, H., Tao, Y., Zheng, Z., Zhang, Q., Zhou, G., Sweetingham, M. W.*, 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.

Yang, K., Tian, Z., Chen, C., Luo, L., Zhao, B., Wang, Z.*, et al.* (2015) Genome sequencing of adzuki bean (*Vigna angularis*) provides insight into high starch and low fat accumulation and domestication. *Proceedings of the National Academy of Science, United States of America,* **112,** 13213-13218.

Young, N., Debellé, F., Oldroyd, G., Geurts, R., Cannon, S. B., Udvardi, M. K.*, et al.* (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature,* **480,** 520-524.

Zhao, Y., Jia, X., Yang, J., Ling, Y., Zhang, Z., Yu, J.*, et al.* (2014) PanGP: A tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics,* **30,** 1297-1299.

Zimin, A. V., Marcais, G., Puiu, D., Roberts, M., Salzberg, S. L. and Yorke, J. A. (2013) The MaSuRCA genome assemble. *Bioinformatics,* **29,** 2669-2677.

Tables

Table 1. Overview of the published reference and pan-genome assemblies for narrow-leafed lupin (*Lupinus angustifolius*) and comparison to the two published white lupin (*L. albus*) genomes, including BUSCO statistics using version 4.1.2 for embryophyte (1,614) and plantae (425).

	Narrow-leafed lupin reference genomes			White lupin reference genomes		NLL pan-genome
	Hane et al 2017	Wang et al 2021	Garg et al 2022	Hufnagel et al 2020	Xu et al 2020	Garg et al 2022
Sequencing technologies used						
Illumina	√	√	√	√	√	√
PacBio	-	√	√	√	√	√
Hi-C	-	-	√	-	√	√
BioNano optical map	-	-	-	√	-	-
Genome statistics						
Genome size	609 Mb	616 Mb	653 Mb	451 Mb	559 Mb	975 Mb
N50	11	9	9	12	14	15
N50 length	21.3 Mb	30.8 Mb	30.7 Mb	17.4 Mb	18.7 Mb	24.8 Mb
GC content (%)	33.46	33.27	33.46	33.79	36.82	38.06
Annotated protein-coding sequences	33,076	33,097	38,545	38,258	47,603	39,339
Mean coding sequence length	1,289 bp	1,258 bp	1,337 bp	1,110 bp	1,194 bp	1,321 bp
BUSCO statistics (embryophyte)						
BUSCOs complete	1,586	1,556	1,590	1,579	1,518	1,590
BUSCOs fragmented	12	17	8	13	9	8
BUSCOs missing	16	41	16	22	87	16
BUSCO statistics (plantae)						
BUSCOs complete	420	417	421	418	412	421
BUSCOs fragmented	1	2	1	3	2	1
BUSCOs missing	4	6	3	4	11	3

Table 2. BUSCO analysis of sequenced reference legume genomes identified two plantae BUSCOs that are absent in most legume species sequenced to date.

	Plantae BUSCO Statistics					
	Complete	Fragmented	Missing	31855at33090	40630at33090	Reference
Adzuki bean	417	5	3	Absent	Absent	Yang et al., 2015
Alfalfa	420	2	3	Absent	Absent	Shen et al., 2020
Barrel medic	420	1	4	Absent	Absent	Young et al., 2011
Chickpea (desi)	414	6	5	**Present**	Absent	Parween et al., 2015
Chickpea (kabuli)	419	2	4	**Present**	Absent	Varshney et al., 2013
Common bean	420	3	2	Absent	Absent	Schmutz et al., 2014
Cowpea	420	2	3	Absent	Absent	Lonardi et al., 2019
Lotus	386	19	20	Absent	**Present**	Sato et al., 2008
Mungbean	421	2	2	Absent	Absent	Kang et al., 2014
Pea	409	11	5	Absent	Absent	Kreplak et al., 2019
Peanut	418	3	4	Absent	Absent	Bertioli et al., 2019
Pigeonpea	414	6	5	Absent	Absent	Varshney et al., 2012
Soybean	420	1	4	Absent	Absent	Schmutz et al., 2010
Subclover	417	5	3	Absent	Absent	Hirakawa et al., 2016
Narrow-leafed lupin	421	1	3	Absent	Absent	This manuscript
White lupin	418	3	4	Absent	Absent	Hufnagel et al., 2020

Table 3. Identification of the true low-alkaloid regulatory gene in the *Lupinus angustifolius* pan-genome.

Accession ID	Cultivar/alternate name	Origin	Line information	Dragendorff test#	Alkaloid phenotype	*RAP2-7* (Lupan_013780)* allele	*RSTK** (Lupan_013738)* allele
-	Coromup	AUS	Cultivar		Sweet	A	T
-	Jindalee	AUS	Cultivar		Sweet	A	T
-	Krasnolistny	POL	Cultivar		Bitter	C	T
-	Oskar	POL	Cultivar	ND	Bitter	C	G
-	PBA Barlock	AUS	Cultivar		Sweet	A	T
-	PBA Jurien	AUS	Cultivar		Sweet	A	T
P20672	Uniwhite	AUS	Cultivar		Sweet	A	T
P20681	Unicrop	AUS	Cultivar		Sweet	A	T
P24138	Marri	AUS	Cultivar		Sweet	A	T
P24141	Chittick	AUS	Cultivar		Sweet	A	T
P25795	Emir	POL	Cultivar		Sweet	A	T
P26971	Danja	AUS	Cultivar		Sweet	A	T
P26973	Geebung	AUS	Cultivar		Sweet	A	T
P26975	Yorrel	AUS	Cultivar		Sweet	A	T
P28137	Myallie	AUS	Cultivar		Sweet	A	T
P28317	Sonet = WTD894	POL	Cultivar		Sweet	A	T
P28324	Kalya	AUS	Cultivar		Sweet	A	T
P28364	Quilinock	AUS	Cultivar		Sweet	A	T
P28578	Tanjil	AUS	Cultivar		Sweet	A	T
P28806	Bolero	POL	Cultivar		Sweet	A	T
P29086	Mandelup	AUS	Cultivar		Sweet	A	T
WALAN4536	Coyote	AUS	Cultivar		Sweet	A	T
P30000	83A:476	AUS	Breeding line		Sweet	A	T
P29039	BSKHA-640	BLR	Breeding line		Bitter	C	T
P20720	G111	ITA	Wild accession		Bitter	C	G
P21624	1111.3	TUR	Wild accession		Bitter	C	G
P22660	Q031	ISR	Wild accession		Bitter	C	G
P22687	GS037	ESP	Wild accession		Bitter	C	G
P22744	GS123	ESP	Wild accession		Bitter	C	G
P22810	NS009	ESP	Wild accession		Bitter	C	G
P22831	GP021	PRT	Wild accession		Bitter	C	ND, low coverage
P22839	GP051	PRT	Wild accession		Bitter	C	G
P22872	GM120	MAR	Wild accession		Bitter	C	G
P25016	QS212	ESP	Wild accession		Bitter	C	G
P25055	GF001	FRA	Wild accession		Bitter	C	G
P25136	NS030	ESP	Wild accession		Bitter	C	T
P26167	G84-038	ESP	Wild accession		Bitter	C	G
P26170	G84-040	ESP	Wild accession		Bitter	C	T
P26239	G84-091	ESP	Wild accession		Bitter	C	T
P26297	G84-159	PRT	Wild accession		Bitter	C	G
P26423	L.O.-1756	ESP	Wild accession		Bitter	C	G
P26464	GRC5011A	GRC	Wild accession		Bitter	C	G
P26559	GRC5038A	GRC	Wild accession		Bitter	C	G
P26603	GRC5054A	GRC	Wild accession		Bitter	C	G
P26625	GRC5066A	GRC	Wild accession		Bitter	C	G
P26668	ITA5269A	ITA	Wild accession		Bitter	C	G
P26676	CY06	CYP	Wild accession		Bitter	C	G
P26710	MAR5292A	MOR	Wild accession		Bitter	C	G
P27221	MAR6009A	MAR	Wild accession		Bitter	C	G
P27255	MAR6045A	MAR	Wild accession		Bitter	C	G
P27436	SYR6259A	SYR	Wild accession		Bitter	C	G
P27893	MAR6781A	MAR	Wild accession		Bitter	C	T
P27895	MAR6783A	MAR	Wild accession		Bitter	C	G
P28038	MJS373	GRC	Wild accession		Bitter	C	G
P28195	AN03	DZA	Wild accession		Bitter	C	G

[#] Alkaloid content in the seeds were determined using the Dragendorff test, where bitter seeds interacts with the Dragendorff reagent and produces orange/pink pigments, whereas sweet seeds do not react. [*]Causal SNP according to Kroc et al 2019 *Genes* 4, 428; Scaffold coordinate with causal A/C SNP: scaffold_8: 10,935,926. [**]Causal SNP according to Wang et al 2021 *The Plant Journal* 5, 1192-1210; Scaffold coordinate with causal T/G SNP: scaffold_8: 8,743,239.

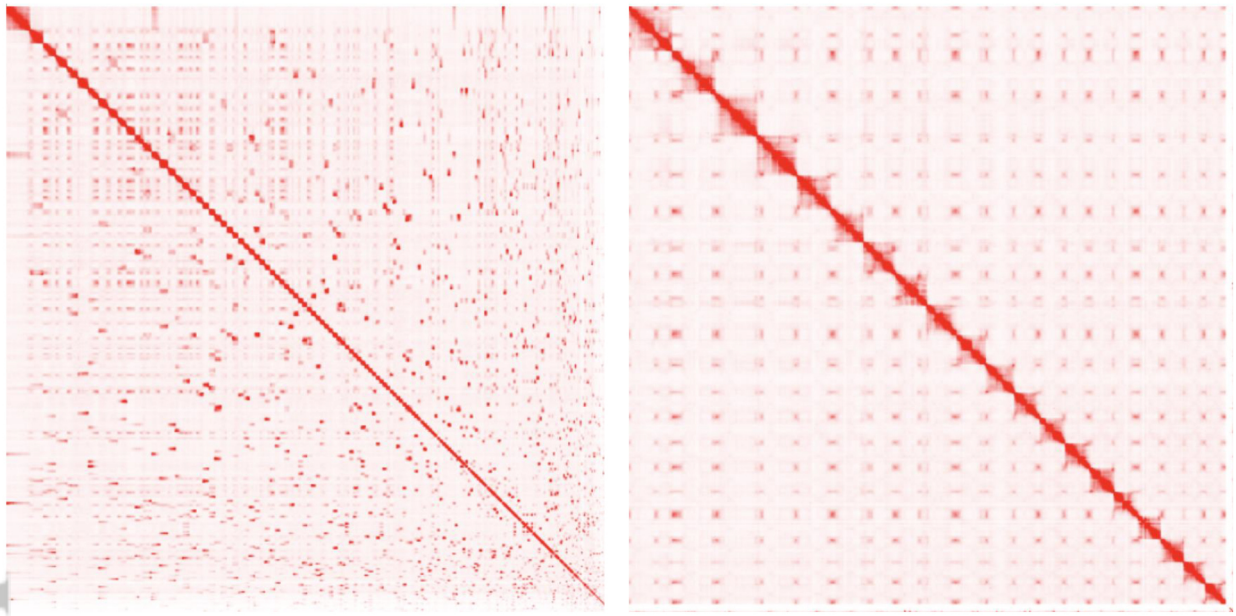
Figure legends

Figure 1. Hi-C map of the draft and chromosome-length assemblies of of narrow-leafed lupin (*Lupinus angustifolius*) cultivar Tanjil genome. Contact matrices were generated by aligning the same Hi-C data set to the Tanjil_Pacbio _v2.0 draft genome (left) and Tanjil_Pacbio _v2.0_HiC genome assembly generated using Hi-C (right). Pixel intensity in the matrix indicates how often a pair of loci co-locate in the nucleus. Correspondence between loci in the draft and final assemblies is illustrated using chromograms. The chromosome-length scaffolds in Tanjil_Pacbio_v2.0_HiC are assigned a linear color gradient; the same colors are then used for the corresponding loci in the Tanjil_Pacbio_v2.0 draft genome (left). The draft scaffolds are ordered by sequence name. Gridlines highlight the boundaries of eight chromosome-length scaffolds in Tanjil_Pacbio_v2.0_HiC (right). The interactive Hi-C contact map for the genome is also available at www.dnazoo.org.

Figure 2. *K*-mer based phylogeny of narrow-leafed lupin (*Lupinus angustifolius*) pan-genome accessions.

Figure 3. *Lupinus angustifolius* pan-genome statistics A) Number of absent variable genes across different narrow-leafed lupin (NLL; *Lupinus angustifolius*) lines. Here, 30 variable genes are absent in 54 lines and present in only one line; B) Model describing the sizes of the core (green) and pan-genome of NLL (blue); C) The number of private single nucleotide polymorphisms (SNPs) for each NLL pan-genome accession.

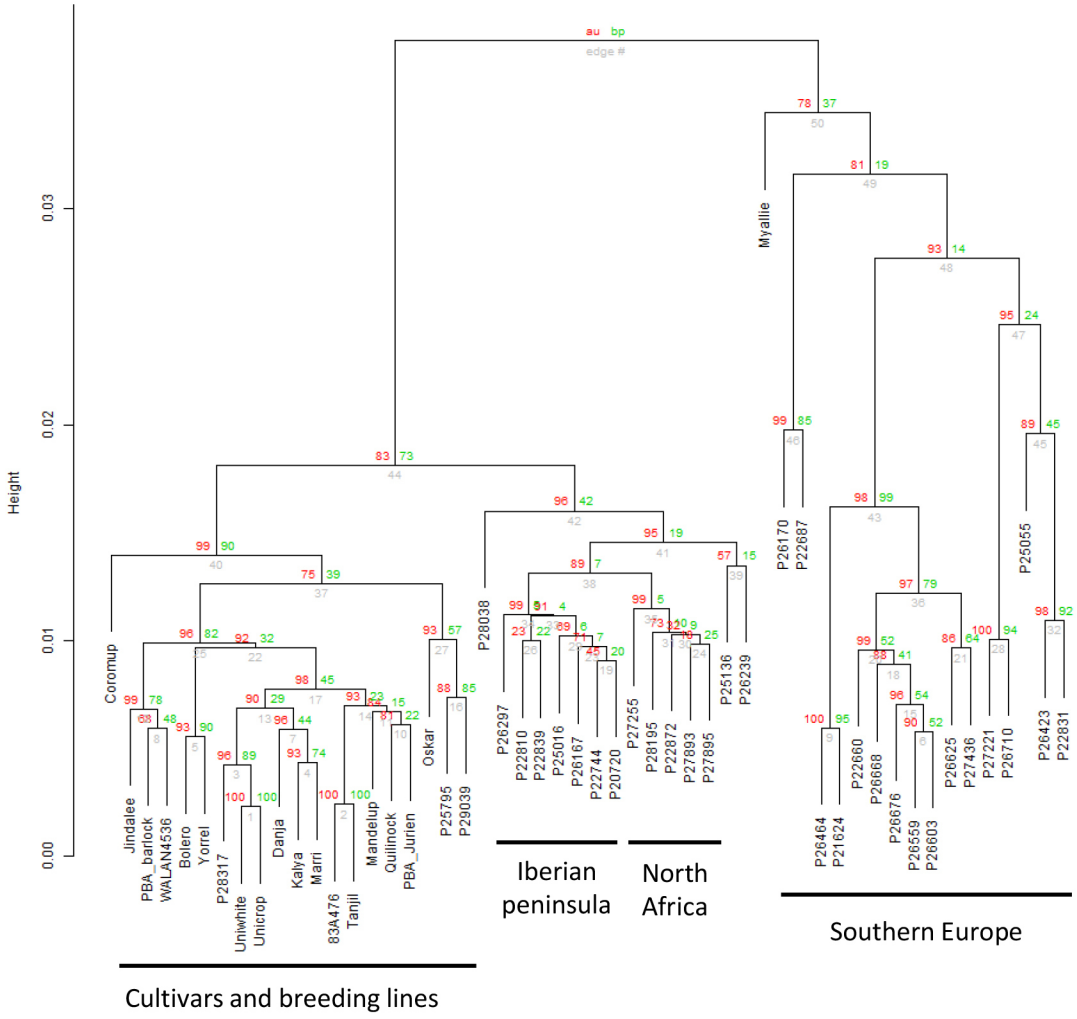
Figure 4. Schematic presentation of the 20 *Lupinus angustifolius* chromosomes and the physical location of classical disease resistance genes of the nucleotide binding-site leucine-rich repeat class termed NLRs.



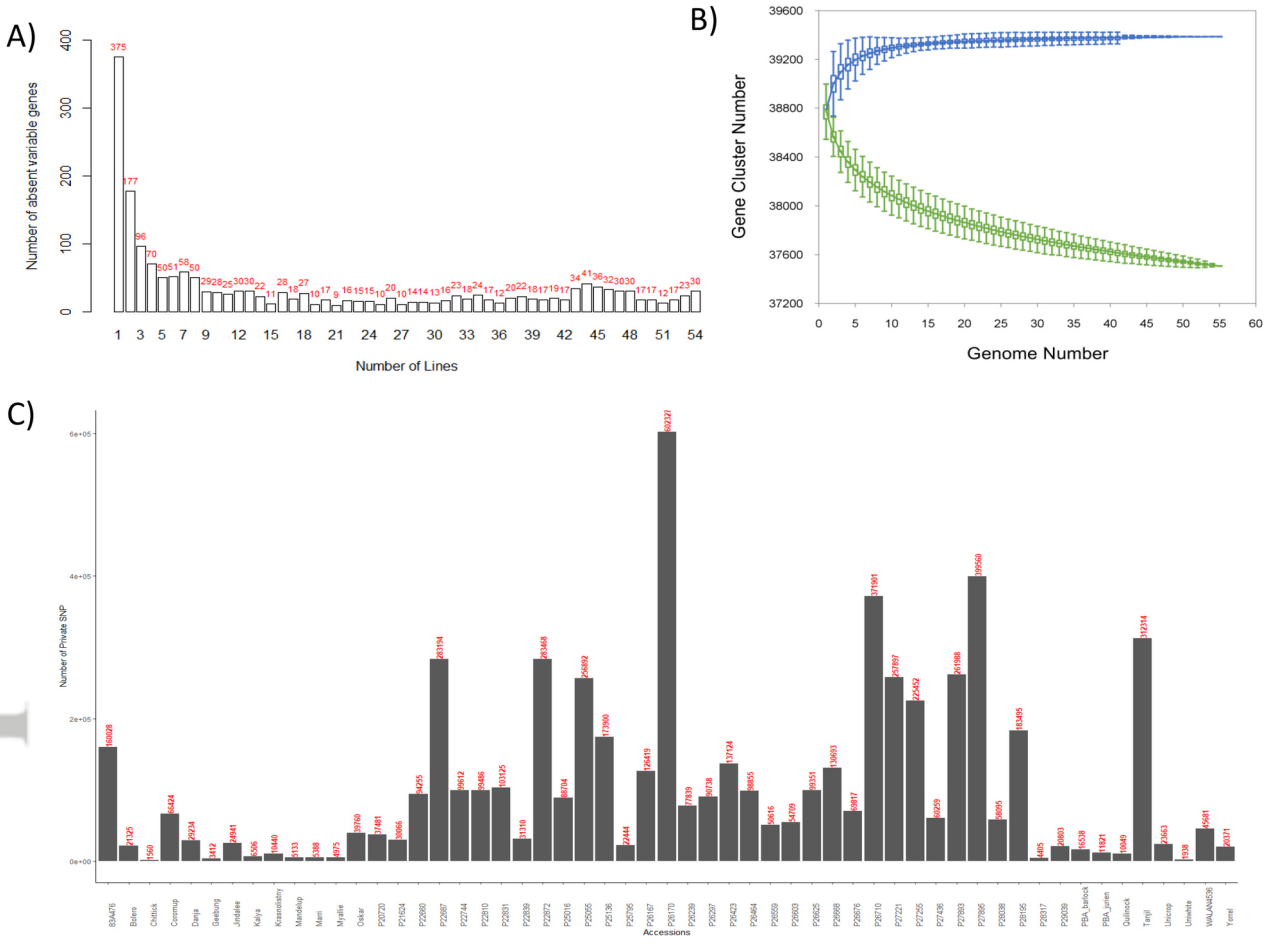
Draft assembly: Tanjil_v2.0 draft genome (PacBio & DNA-Seq)

Chromosome-length assembly: Tanjil_v2.0 _hic

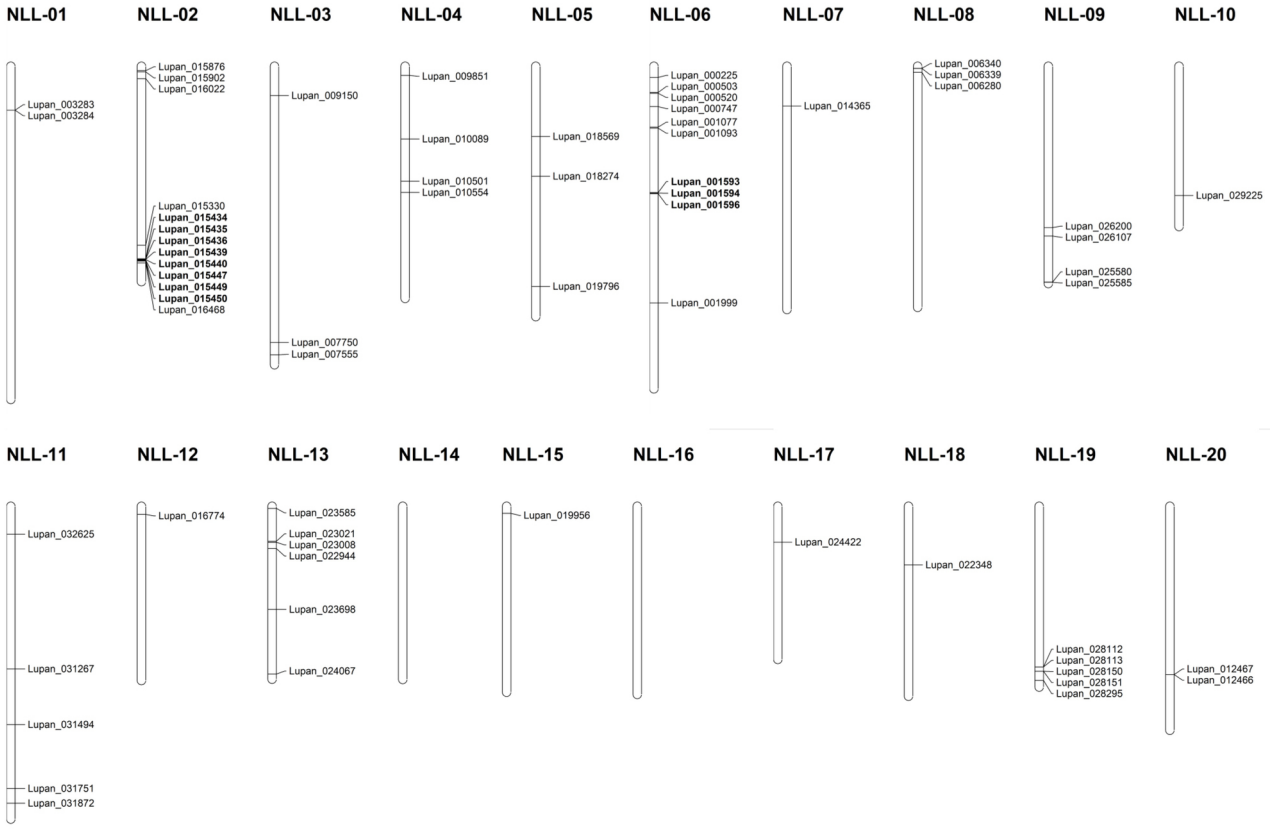
TPJ_15885_Figure 1.jpg



TPJ_15885_Figure 2.jpg



TPJ_15885_Figure 3.jpg



TPJ_15885_Figure 4.jpg