

A genomic toolkit for 'the soybean of the tropics' – winged bean (*Psophocarpus tetragonolobus*)

Wai Kuan Ho

University of Nottingham Malaysia <https://orcid.org/0000-0003-4622-1696>

Alberto Tanzi

University of Nottingham

Fei Sang

University of Nottingham <https://orcid.org/0000-0002-8746-5868>

Niki Tsoutsoura

University of Nottingham

Niraj Shah

University of Nottingham

Christopher Moore

University of Nottingham <https://orcid.org/0000-0002-3122-7502>

Victoria Wright

University of Nottingham

Festo Massawe

University of Nottingham Malaysia Campus <https://orcid.org/0000-0002-0744-4777>

Sean Mayes

Sean.Mayes@nottingham.ac.uk

University of Nottingham <https://orcid.org/0000-0003-4770-3662>

Article

Keywords:

Posted Date: February 18th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1355353/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Communications on March 1st, 2024. See the published version at <https://doi.org/10.1038/s41467-024-45048-x>.

Abstract

A sustainable supply of plant protein is critical for future generations. We also need to reduce carbon emissions from agriculture, while increasing resilience in the face of elevated temperatures and climate volatility. Agricultural diversification with more crop options is hampered by limited available genomic resources, lack of understanding of the genetic structure of breeding germplasm and important trait inheritance- needed to breed desired ideotypes.

Winged bean (*Psophocarpus tetragonolobus*), a high protein seed tropical legume, has been termed 'the soybean for the tropics' with potential to help reduce soybean imports in many tropical and semi-tropical countries.

Here we present the first chromosome level winged bean genome assembly (98.9% of 536.1MB scaffolds are assigned), an investigation of genetic diversity of 136 worldwide accessions for breeding, together with the first two genetic maps and a trait QTL analysis for genomic regions with desirable ideotype traits for breeding; namely architecture, protein content and phytonutrients.

Introduction

Winged bean (*Psophocarpus tetragonolobus* (L.) DC.), also known as goa bean, is a dicotyledonous ($2n = 2x = 18$) legume belonging to the Fabaceae family and Papilionoideae subfamily. It is the only domesticated species of genus *Psophocarpus* and has three pairs of short and six pairs of long chromosomes¹. It is cultivated mainly in hot, humid, equatorial countries of Southern Asia, Melanesia and the Pacific region. It is grown as a vegetable for its immature pods and for leaves and for tuberous roots, with consumption preferences differing geographically². Its ability to fix atmospheric nitrogen with a wide range of *Rhizobium*^{3,4} makes it an ideal horticultural crop for more sustainable agriculture particularly in the humid and sub-humid tropics. Nevertheless, its uncontrolled branching habit, indeterminate growth and inconsistent productivity all limit current large-scale utilisation. Our previous work has suggested that an optimal trade-off could be achieved between vegetative growth and yield (pod number and seed number), for example, by controlling branch number⁵. Individuals with fewer but longer branches could maintain high pod productivity, while potentially reducing the vegetative biomass.

Its description as 'a soybean of the tropics' was coined in the 1980s after work by the US National Academy suggested it had potential to partially replace soybean in tropical countries⁶. In particular, its mature seeds are a good source of protein and other nutrients. Mature dried winged bean seeds are reported to contain between 16.4 to 21.3% lipid and 34.3 to 40.7% crude protein with a favourable amino acid balance and micronutrient composition^{7,8}. However, its dietary diversification and nutritional potentials have been hampered by the crop architecture and the 'hard-to-cook' phenomenon, with the presence of anti-nutritional factors (ANFs), such as hemagglutinating lectins, trypsin and chymotrypsin inhibitors, and tannins⁹, in the seed. The crude protein content in the seeds of 25 accessions from the International Institute of Tropical Agriculture (IITA) ranged from 28.4 to 31.3% with tannin and phytate varying significantly¹⁰. Some of the issues caused by antinutritional factors can be overcome through different cooking and processing methods^{11,12}. Coupled with

architectural changes and breeding selection for high protein and/or high lipid, genetic improvement could widen winged bean production from a purely horticultural crop to an arable crop for its seed protein, also helping to reduce dependency on soybean imports for many countries and the environmental impact associated with soybean production expansion.

Genome-assisted breeding improvement strategies can lead to genetically improved winged bean to contribute to food and nutritional security. However, the lack of molecular resources has hampered a better understanding of the diversity and phylogenetic relationships of germplasm conserved *ex situ*. Using Genotype-by-Sequencing (GbS) we have evaluated the genetic structure of a wide range of global accessions, developed two controlled cross genetic maps and performed the first QTL analysis of architectural, protein and a nutritional trait. We also present the first chromosome-scale genome assembly through a combinatorial approach integrating Illumina and Nanopore reads with BioNano optical mapping and within species genetic maps.

Results

Genome assembly and annotation

Ma3, a cultivar (renamed from the previously reported M3⁵) released by Malaysian Agricultural Research and Development Institute (MARDI) was selected for genome sequencing. Our approach of using Oxford Nanopore Technology (ONT) reads corrected by Illumina reads generated 1,164 scaffolds with the largest being 19.99 Mb in length. An improvement was further made with the addition of BioNano optical mapping, yielding a total of 48 hybrid scaffolds with the largest scaffold being 38.64 Mb and a N50 at 23.88 Mb (Table 1 and Table S1). A total of 94.6% of Illumina reads were successfully remapped back to the assembly.

Subsequently, SNP markers in two genetic maps derived from Genotype-by-Sequencing (DArTSeq™) were used to assign scaffolds for the final assembly at a pseudo-chromosome level, spanning 530.28 Mb in total. This assembly accounts for 98.9% of the assembled scaffolds, with only 5.85 Mb of scaffolds left unassigned. The genetic maps derived from Cross XT ($F_2 = 184$) and Cross XB ($F_2 = 223$) have Ma3 as the common paternal parent and are reported for the first time here (Fig. 1a). The genetic map of Cross XT comprises of 692 high quality spaced SNP markers (with no missing data across the whole population) covering 1,102.1 cM (an average interval of 1.6 cM), the XB map comprises of 393 spaced SNP markers covering 1,336.9 cM (3.4 cM average distance) (Table S2). The nine anchored pseudo-chromosomes ranged from 24.61 to 84.79 Mb in length (Table S2). Of the SNP markers present in the linkage maps, 97.4% and 95.4% SNPs, respectively, could be mapped back to the genome assembly. From the full DArTseq™ dataset derived from the mapping and diversity analysis, 95.7% of SNP ($n = 21,412$) and 77.2% ($n = 40,557$) of presence-absence variants (PAV/*in silico* DArT), respectively, had their sequence tags mapped to the final genome (Fig. 1b). Repetitive elements represented 37.7% ($n = 242,355$) of the winged bean genome, with the majority being Terminal Inverted Repeat elements (TIR; 49.02%) followed by Long Terminal Repeat elements (LTR; 28.66%) (Fig. S1). The Class I/Class II transposable elements (TE) ratio was observed to be 0.81 (Table S3 and S4). Of the dominant Long Terminal Repeat (LTR) in Class I retrotransposons, the *Gypsy* superfamily was six times more abundant than *Copia*-type (Fig. S1).

Other than *de novo* gene prediction and homology comparison, initial transcriptomic resources from leaf, root, pod, and reproductive tissue (comprising of bud and flower)¹³ were also included, assisting in 26,370 protein coding genes being annotated. Of 1,614 Embryophyta larger group single-copy core genes from both Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis, 84.1% complete conserved genes were recovered with 9.5% as missing and 6.4% are fragmented (Table 1). Additionally, nearly 80% of the annotated genes have an Annotation Edit Distance (AED) score ≤ 0.3 ¹⁴ (Fig. S2a), indicating a high degree of quality and continuity for the current annotation. These gene models have an average length of 4,203 bp and an average coding-sequence length of 1,396 bp (Fig. S2b), with an average of five exons per gene. As many as 1,551 of them are transcription factors, accounting 5.9% of the total annotated genes. Together, transcription factor classes MYB and bHLH are dominant and accounted for 17.6% of the identified transcription factors (Table S5). Additionally, 10,303 small nucleolar RNAs and 446 tRNA genes were identified.

The previously reported genome size of winged bean varies widely; Vatanparast and colleagues¹⁵ have suggested it to be 1.22 Gbp/1C in size whereas previous flow cytometry has estimated it to be 782 Mb (or 0.8 pg)/1C¹⁶. Our *K*-mer distribution analysis has predicated a smaller genome size at 569 Mb although GC-rich regions and repetitive sequences might have been underestimated¹⁷ (Fig. S3). This would also explain a lower proportion of repetitive elements in the winged bean genome, as compared to 49.5% in cowpea genome¹⁸ and 41% of TE in *P. vulgaris* genome v2.1¹⁹. The genome sequenced material has around 5.4% heterozygosity detected from the 21,412 gene-enriched SNPs, with 49.9% of DArTseq SNPs located within coding regions. Overall, the current reference genome covers 93.2% of the *K*-mer estimated genome size of 569 Mb and is expected to provide comprehensive coverage of the coding regions of winged bean. By comparing the extent of protein-coding gene homology (at $e\text{-value} < 1e^{-10}$), it was observed that winged bean shares 87.2% with *Glycine max* (Wm82.a4.v1)²⁰, 86.5% with *Phaseolus vulgaris* (Pvulgaris v2.1)²¹ and 87.0% with *Vigna unguiculata* (Vunguiculata v1.2)¹⁸. The majority of the genes are conserved between the species, as illustrated in Fig. 2. for example. However, among highly conserved orthologues identified on Pt01 and Pt07 (at $e = 0$) 549 and 496, respectively, we did find genes distributed across all 11 chromosomes of common bean, suggesting extensive chromosomal rearrangement or gene translocation in winged bean compared to common bean (Fig. S4). Even with the smallest Pt09 pseudochromosome, blocks of gene collinearity were found on eight Pv chromosomes. Taken together, our approach has demonstrated a cost-effective means for generating species-specific genomic landscapes, particularly for crops lacking well annotated closely related species, such as winged bean. The strategic sequencing of individual species within clades where no current crop genome exists could allow progress to be made in multiple species.

Table 1
Statistics summary of reads and assembly

ONT corrected by Illumina,	
No. of scaffolds	1,164
Min read length (Mbp)	0.002
Median read length (Mbp)	0.019
Mean read length (Mbp)	0.443
N50 scaffold length (Mbp)	3.808
Max scaffold length (Mbp)	19.988
Total scaffold length (Mbp)	515.733
BioNano map,	
No. of scaffolds	150
Min read length (Mbp)	0.058
Median read length (Mbp)	0.989
Mean scaffold length (Mbp)	4.395
N50 scaffold length (Mbp)	23.344
Max scaffold length (Mbp)	39.793
Total scaffold length (Mbp)	659.317
Hybrid scaffolds,	
No. of scaffolds	48
Min scaffold length (Mbp)	0.123
Median scaffold length (Mbp)	6.699
Mean scaffold length (Mbp)	11.169
N50 scaffold length (Mbp)	23.875
Max scaffold length (Mbp)	38.637
Total scaffold length (Mbp)	536.132
Repeat region of assembly (bp)	202,036,684
Completeness of gene prediction (BUSCO)	
Complete genes (C)	84.1%
Complete and single copy (S)	82.2%
Complete and duplicated (D)	1.9%

ONT corrected by Illumina,	
Fragmented genes (F)	6.4%
Missing genes (M)	9.5%
Total BUSCO groups searched	1,614
Pseudochromosome, integrated with genetic maps,	
Scaffolds assigned to LGs	
Cross XB (FP15 x Ma3)	32
Cross XT (Tpt10 x Ma3)	38
Unassigned scaffolds to LGs	10
Length of assigned sequences to LGs	530,283,461 bp
Length of unassigned sequences to LGs	5,848,080 bp

Winged bean germplasm diversity

For any genomics-based breeding effort to attain maximum genetic gain, an understanding of the genetic relationships among the available breeding germplasm is needed. The genetic diversity of winged bean germplasm was investigated using 171 individuals originating from 16 countries (Table S6) based on GbS (DArTSeq) markers. Using 7,791 SNP [with minor allele frequency (MAF) > 0.01], the level of heterozygosity ranged from 0.4 to 56.6% with a median value at 6.1%, largely reflective of a self-pollinating species (Table S6). Nevertheless, the presence of individuals with high heterozygosity implies the possibility of outcrossing and is consistent with our observations of bees and other insects visiting flowers in open field conditions. As high as a 16.4% average out-crossing rate has been reported from 457 Thai accessions evaluated using 14 Simple Sequence Repeat (SSR) markers²².

Bayesian clustering and phylogenetic analysis [Neighbour-Joining (NJ) tree] and Principal Coordinate Analysis (PCoA) results all closely corresponded to each other and suggest the existence of three major gene pools (CP, OT, LT) (Fig. 3 and Fig. S5). Based on the membership coefficient ≥ 0.7 , approximately a quarter of these lineages show allelic admixture. As previously reported^{2,23}, the clustering does not correlate well with the country of origin, which might reflect seed exchange among genebanks with limited documentation or historical exchange with subsequent collection capturing the secondary collection sites, although broader regional clustering can be seen (Fig. S5b). Comparing among the countries having more than five accessions included in this study, the materials originating from Colombia and Thailand (44.4% out of nine accessions and 45.2% out of 31 lines, respectively, excluding biological replicates) show high levels of admixture. This suggests possible in-field hybridisation events between accessions from genetic exchanges of germplasm. This has further supported our previous hypothesis that the germplasm were most likely introduced from Asia to South America given that winged bean is not been reported as a native or traditional crop consumed in

South America². In contrast, potentially more limited diversity and possible founder effects might be reflected in the germplasm collected from India and the Philippines (Fig. 3d).

The CP subpopulation, being the smallest cluster, is composed of mainly Colombian and Papua New Guinean accessions whereas the LT subpopulation predominantly consists of accessions collected from Sri Lanka and Thailand. It is also interesting to note that for 57 accessions, replicate samples from the same breeding accession based on pedigree information were available. However, 44% (25 out of 57) of these were found not to be highly similar when using molecular markers. Although mislabelling at some stage cannot be ruled out, outcrossing might have also contributed to the differences between genotypes in or from the same accession. As such, our findings could help breeders by emphasising the need for quality assurance for breeding programmes and to test pedigree information, as well as providing insights into some of the publicly available germplasm.

Analysis of Molecular Variance (AMOVA) reveals that the genetic variance observed is not only attributed to the differences between individuals within population (45.7%) but also between populations (30.9%). Comparatively, genetic differentiation was lower within individual accessions, accounting for 23.4% of the total variance. Nevertheless, the population differentiation was significant at all levels (Table S7). The moderate, yet greatest, difference (0.4479) in pairwise Nei's G_{ST} between the CP and LT subpopulations is in line with the population structure analysis as the CP and OT population clustered together when $K = 2$ (Table S8). Similarly, the first principal component (Coordinate 1 = 39.0% variation explained) in the PCoA distinguishes CP and LT subpopulations whereas the second component (Coordinate 2 = 23.6%) reveals the divergence between OT and CP subpopulations (Fig. 3b). However, genetic differentiation among OT and LT subpopulations was relatively low at 0.1792 (Table S9).

Both Shannon-Wiener index (H) and Simpson's index (λ) suggested that the OT subpopulation ($n = 70$) had the highest genotypic richness or genetic diversity among all populations (Table S9). Consistent with this, the expected heterozygosity (H_e) from Nei's Index suggests a stronger selection signature or founder effect in the CP subpopulation ($H_e = 0.102$) in comparison with OT ($H_e = 0.247$) and LT ($H_e = 0.134$) subpopulations (Table S9). The genetic richness preserved within the OT gene pool might reflect different domestication selection criteria for the diverse traits in this crop, due to its various edible parts and their varying consumption preference in different countries (Tajima's $D = 1.0204$). For example, tubers are mainly consumed in Myanmar, whereas young pods are preferred in Malaysia. On the other hand, the observed Tajima's D value of -0.3394 indicates that the CP cluster could be derived from directional selection pressure thus driving the divergence, while balancing selection or population contraction could have happened in OT and LT gene pools (Tajima's $D = 1.0204$ and 0.6142, respectively) with OT being more severely affected (Table S9).

Linkage disequilibrium (LD) decay

The genome-wide linkage disequilibrium decay rate of winged bean was estimated to be in the range of 76 to 113 kb based on GbS SNPs when the MAF was set at > 0.01 and > 0.05 , respectively. As reported in a peanut (*Arachis hypogaea*) mini-core collection that consisted of 107 accessions and also in 632 lines of a temperate and tropical maize collection, LD is affected significantly by MAF selection^{18,24}. At MAF > 0.05 ,

6,149 SNP were distributed across the genome with approximately 683 SNP per chromosome on average, with Pt08 showing the fastest LD decay (Fig. S6 and Table S10). Only 0.8% of SNP pairs showed high LD ($r^2 \geq 0.8$). This would argue that while winged bean is predominantly autogamous, outcrossing is still a major contributor to current genetic diversity. This LD decay rate is similar to cultivated pigeonpea (*Cajanus cajan*). When compared to other closely related autogamous legumes such as soybean (250–375kb)^{25,26}, common bean (240–250kb)^{27,28} and cowpea (809–4,705 kb)²⁹, LD decay in winged bean is more rapid. High LD decay could indicate that phenotypic evaluation of this set of germplasm can identify close gene-trait associations through association analysis, although larger numbers of markers will also be needed.

Quantitative Trait Locus (QTL) analysis of XB population

As a first step in breeding and trait analysis we developed a number of controlled crosses with the Ma3 genotype as the common parent, focusing on architecture and pigmentation. Two of these populations were used for genetic map construction to assist in scaffold assignment and orientation to generate the physical map of winged bean (as mentioned above). The segregation for morphological traits in the XB F₂ between FP15 and Ma3 has previously been reported⁵. For this cross, a QTL analysis was also conducted (illustrated in Fig. 4) and is summarised in Table 2 and Table S11. To test the utility of the genome resources generated, we then investigated the inheritance of several traits of potential breeding interest in the XB population in addition to those previously reported⁵.

Table 2. QTL analysis results

Trait	QTL	Nearest marker	position (cM)	physical position (bp) (2-LOD interval)	LOD	PVE (%)	additive	dominance
total branch no.	<i>qTNoB-2</i>	22986_20:T > C	28.0	Chr01.1: 1,877,647 .. 24,357,512	3.7	17.8	-1.50	-0.21
internode length	<i>qInL-1</i>	18922_15:A > C	24.7	Chr02.1: 329,655 .. Chr02.5: 6,219,037	2.1	10.6	-1.06	-0.72
branch length	<i>qLoB</i>	13849_61:A > C	59.1	Chr02.1: 15,929,057 .. 27,979,125	3.6	17.5	25.78	-30.60
seed storage protein	<i>qSSP-3</i>			Chr02.1: 20,776,303 .. 27,979,125	3.5	6.7	-1.00	0.88
days to flowering	<i>qDtF</i>	15138_30:C > T	99.6	Chr02.1: 29,081,247 .. Chr02.5: 29,506,843	2.5	12.6	2.23	-1.78
pod no.	<i>qPPP</i>	7847_38:T > C	138.4	Chr02.5: 19,515,042 .. 25,398,143	3.0	15.5	4.11	5.06
seed storage protein	<i>qSSP-4</i>	24545_60:G > A	145.7	Chr02.5: 22,449,002 .. 29,506,843	3.0	5.6	0.97	-0.19
calyx colour	qCLX-3	14631_48:A > C	147.9	Chr02.5: 26,720,591 .. 29,506,843	5.5	25.8	-0.51	-0.52
pod pattern	qPP-3	22709_12:C > G	37.4	Chr03.1: 14,207,653 - 22,702,852	5.3	24.8	-0.85	-0.28
seed storage protein	qSSP-1	13560_30:G > T	59.1	Chr03.1: 20,722,534 .. 26,891,275	5.5	10.8	-1.30	1.00
seed storage protein	<i>qSSP-2</i>	15602_68:T > A	145.9	Chr03.5: 10,575,228 .. 23,809,662	3.7	6.4	-1.44	-0.91

Trait	QTL	Nearest marker	position (cM)	physical position (bp) (2-LOD interval)	LOD	PVE (%)	additive	dominance
branch no.	<i>qNoB</i>			Chr03.5: 13,353,743 .. 24,206,688	3.5	16.9	1.04	-0.15
total branch no.	qTNoB-1			Chr03.5: 10,575,228 .. 24,576,784	4.4	21.1	1.81	-0.35
internode length	<i>qInL-2</i>	24841_65:G > T	68.7	Chr04.1: 17,010,559 .. Chr04.2: 8,317,162	2.6	13.0	-0.54	1.75
stem length	qStL				4.4	21.6	-19.60	54.90
seed coat colour*	qSC-2	<i>P2</i> , 14758_43:A > C	0.0	Chr05.1: 673,767 .. 1,539,918	39.1	61.7	-0.45	0.31
pod pattern	qPP-2				17.7	61.7	-1.23	0.11
calyx colour	qCLX-2				20.2	66.5	-0.67	0.33
seed coat colour*	qSC-1	<i>P1</i> , 13546_10:C > G	49.6	Chr05.1: 5,869,029 .. 10,124,037	35.9	58.5	-0.45	0.30
pod pattern	qPP-1				14.5	54.5	-1.19	0.11
calyx colour	qCLX-1				17.3	60.8	-0.71	0.25
node no.	<i>qNoN</i>	12217_11:A > G	15.8	Chr05.1: 673,767 .. 14,461,456	2.6	12.9	-1.62	-0.82
pod length	<i>qPoL-1</i>	15537_67:A > C	198.1	Chr05.3: 4,466,417 .. 12,850,382	3.1	15.9	-1.20	0.65
100-seed weight	<i>qHSW</i>	8194_23:A > C	227.7	Chr05.3: 14,300,923 .. 20,126,305	3.2	16.4	-2.70	1.19
chlorophyll content	qCC-3	14275_32:C > G	213.5	Chr05.3: 21,060,745 .. 23,175,358	6.5	29.5	-6.00	4.10
chlorophyll content	qCC-2	19643_19:T > C	277.2	Chr05.3: 7,260,935 .. 16,578,779	6.7	30.0	-5.46	4.30

Trait	QTL	Nearest marker	position (cM)	physical position (bp) (2-LOD interval)	LOD	PVE (%)	additive	dominance
seed no./pod	<i>qSPP-1</i>	18511_24:G > T	7.9	Chr06.1: 2,443,143–4,835,406	3.4	17.5	1.88	-0.31
seed storage protein	<i>qSSP-5</i>	18371_20:A > G	90.8	Chr07.2: 1,912,623 .. 19,565,332	3.0	8.1	-1.00	0.79
seed no./pod	<i>qSPP-2</i>	15840_66:A > T	0.0	Chr08.1: 234,458 .. 4,081,805	3.2	16.2	1.31	0.63
pod maturing time	<i>qPMT</i>	22543_38:C > T	3.4	Chr09.1: 614,913 .. Chr09.2: 2,556,146	3.1	15.4	-0.73	19.47
pod pattern	<i>qPP-4</i>	15409_18:T > A	80.9	Chr09.3: 5,123,254 .. 8,234,733	4.2	20.3	-0.89	-0.58
pod length	<i>qPoL-2</i>	11154_46:G > T	93.2	Chr09.3: 6,579,464 .. 10,994,671	3.0	15.4	-1.38	-1.28

* only major peaks listed, refer to Table S12 for candidate genes in Chr05.1

** significant QTL are indicated in **bold**

Variation in seed storage proteins (SSP) compared to soybean

The review by Kadam et al in the 1980s reported that winged bean seeds has a comparable essential amino acid composition to soybean⁷. From this base, molecular breeding could enhance protein composition and content. It was also reported that protein components in the winged bean seed are predominantly vicilin-type 7S globulin occupying approximately 31.9% of the protein extract, with a lack of 11S legumin-type^{30,31}. In addition to the 18 genes putatively encoding 7S globulin identified from the genome sequence (Table S12), we also found three genes (Psote01G0072700, Psote03G0109900 and Psote03G0110000) homologous to the 11S globulin in soybean, *Medicago truncatula* and chickpea³². Of these, two are approximately 6Mb from the *qSSP-1* QTL. Given that legumin contains a higher proportion of sulphur-containing amino acids than 7S vicilin/convicilin³³, it is worth investigating whether these genes are functional or are being expressed at low levels, so were undetectable by previous methodologies. Additionally, 2S albumin was reported to vary among winged bean cultivars³⁴, another component rich in sulphur in common bean³⁵. It was reported to be the second most abundant protein fraction occupying 26.5% of total protein in winged bean. Quantifying sulphur content in each protein fraction could provide further targets for winged bean improvement for good nutritional quality and amino acid balance (Table S12).

In contrast to soybean, developmental stage characterisation has suggested that protein bodies in winged bean seeds accumulate at a late stage of maturation (45 days after flowering) when starch granules are decreasing rapidly^{36,37}. The protein profiles characterised by Makeri and colleagues further indicated that the N content in the globulin fraction is highest in winged bean seed, but lowest in soybean globulins, but also that N content is significantly different from other protein fractions in both crops³¹. These observations, in turn, might suggest that the targets for genetic improvement in seed protein content could be different between winged bean and soybean.

The paternal Ma3 genotype had a measured average of $34.0 \pm 1.8\%$ ($n = 14$) total seed protein content, whilst the maternal FP15 had $39.3 \pm 3.2\%$ ($n = 9$) although a wider difference has been observed between other lines from other studies⁶. Among the F₃ seed from the XB cross and assuming this represents the maternal genotype ($n = 161$ lines evaluated), transgressive segregation was observed in the population, ranging from 27.2 to 44.6% total protein per seed (Fig. 5). In other legumes such as pigeonpea (*Cajanus cajan*), soybean and pea (*Pisum sativum*) similar seed protein segregation was reported, although 24.6% in pigeonpea and 33% in pea were considered as high protein genotypes^{38,39,40} and are lower than seen here in winged bean. A significant QTL explaining 10.8% protein variation was detected on Pt03.1, and three globulin genes were identified within two of the three putative QTL, with the QTL cumulatively explaining 18.7% of the total phenotypic variation (Table 2). Given the high genetic complexity underlying the accumulation of seed protein - for example, 33 consensus amino acid content QTL were identified in soybean⁴¹ - it is not surprising that candidate gene identification is not straightforward (Table S13). Nevertheless, Psote03G0171400 (*qSSP-1*) and its paralogue Psote02G0416500 (approximately 2.3Mb from *qSSP-4*) are homologous to xylem bark cysteine peptidase 3 (P34, Glyma.08G116300), which is the second highest abundance transcript detected in all four stages of soybean seeds during storage protein accumulation⁴². Although this is not a well characterised gene, it is highly expressed at 40 days after flowering (DAF) in soybean and associated with enriched GO terms 'transmembrane transport' and 'nutrient reservoir activity'⁴³. Furthermore, using the same accessions, the authors compared their DEGs with the Harada-Goldberg soybean RNA-seq dataset (<http://seedgenenetwork.net>) comprising of 15 tissues with this gene found to be tissue specific and expressed in an early maturation of seed stage⁴². Although the soybean orthologues of Psote03G0156500, Psote03G0160400 and Psote03G0160600 (*qSSP-1*) are not tissue specificity, they were expressed highly in all four soybean seed developmental stages⁴². In addition, their orthologues in the common bean have been identified as the target genes of microRNAs expressed in the developing seed at 20 days after anthesis (DAA), considered to be the mid-filling/maturation stage. The genomic sequences surrounding the SNP of flanking markers detected which detected QTL for protein content (and other traits) could facilitate marker assisted selection (MAS) (presented in Table S14). Overall, from this initial work, it is clear that there is potential within winged bean germplasm to select for a high protein legume seed crop, either through conventional approaches or MAS.

At the moment, winged bean is primarily a horticultural crop, grown on trellis structures and manually harvested. For large-scale seed production, the crop could be grown as a creeping vine on the flat (assuming that flowering and pod development is not interrupted) or changes in the architecture of the plant could be

bred for, to facilitate mechanisation of sowing and harvesting or even to optimise intercropping with complementary amino acid sources, such as cereals.

Altering architectural traits to improve mechanisation

Our previous work⁵ suggested that fewer but longer branches could achieve the best trade-off between yield and vegetative biomass. Both *branch length* and *total number of branches* traits had a significant positive correlation with *pod number* ($r_s = 0.44$, $p < 0.001$ and $r_s = 0.28$; $p < 0.01$, respectively). Selection for these traits could therefore have major impact on plant architecture and yield of winged bean. There could have been selection for this trait during domestication, given the observed reduction in the higher orders of branches in the domesticated species, compared to closely related species like *P. scandens* (Tanzi AS, pers. comm.). Our QTL analysis has further suggested a putative *qNoB* (*branch number* QTL, measured from branches > 10cm in length within the first ten nodes) that co-localises with a significant *total branch number* QTL (*qTNoB-1*; LOD 4.4, 21.1% PVE, counted from the entire main stem) with the former having a narrower 2-LOD range (Chr03.5: 13,353,743 .. 24,206,688), implying that yield variation in this material may be contributed to by branches departing from upper stem nodes. Although lower *branch number* yet higher *branch length* could be a better selection criterion for yield, the *branch length* QTL (*qLoB*) is currently putative at LOD 3.6. The Arabidopsis *CUC2* orthologues (Psote01G0176400 and Psote03G0433200) both located within the QTL (*qTNoB-1/qNoB* and *qTNoB-2*) are promising candidate genes that could be examined further (Table S15). Li and colleagues have recently discovered the CUC2/3-DA1 UBP15 regulatory module and its role in controlling axillary meristem initiation⁴⁴. The *AtUBP15* orthologue (Psote03G0414300) is positioned within the *qTNoB-1/qNoB* region as well. CUC2 is a member of the NAC domain of transcription factors that defines organ boundaries and organ numbers⁴⁵.

Inheritance of phytonutrient content - purple pigmentation

The maternal parent of the XB population, FP15, has pods which are near-full or full purple when they approach maturity, while the seeds are purple or near black (Fig. S7). Being a member in the CP subpopulation with extremely low heterozygosity, this line might have undergone active selection by farmers for its distinctive pigmentation. The paternal parent Ma3 (OT subpopulation) has green stems and fully green pods, with the F₃ population (F₂ maternal tissue) derived having pigment segregation in the seed testa and for pod specks. As seed coat is a maternal tissue, it is not surprising that F₁ seeds had dark purple without segregation. Of the F₃ seed screened, 80 had a brown seed coat and 122 were purple. This ratio ($\chi^2 = 1.411$; $p = 0.24$) could be attributed to two dominant loci with a duplicate recessive epistasis interaction. Two major significant QTL (*P1* and *P2* alleles, Table 2) were detected explaining 61.7% and 58.5% of the phenotypic variation, respectively. Maternal tissue interaction and/or environmental stimulus were speculated to be contributing to the variation in seed coat colour intensity in an *Trifolium* interspecific cross⁴⁶. Similar mechanisms could be present in winged bean.

These pigmentation loci (Table 2) demonstrate the complexity in the regulatory network of anthocyanin biosynthesis pathway in a tissue-specific manner. The pigmentation in pod and calyx showed the same segregation patterns as the seed testa ($\chi^2 = 0.067$, $p = 0.80$; $\chi^2 = 0.229$, $p = 0.63$) and exhibited co-dominance and/or an incomplete dominance. Of these traits, the F₁ had purple-green calyx and purple specks on pods.

While the pod pigmentation varied in speck quantity and even expanded to purple ‘wings’ (pod fringes), no F₂ progeny were identified which had as full a purple pod as the maternal parent. Notably, lines with green calyx had brown seed testa, suggesting the lack of anthocyanin expression in both tissues (Fig. S8). In support of this notion, QTL for *calyx colour* and *pod pattern* were identified also on Pt02 and Pt03, respectively. In soybean, there have been six loci identified; *I*, *T*, *Wp*, *W1*, *R* and *O*, producing colours ranging from yellow, green buff and brown to black⁴⁷. Both brown (*irT*) and black (*iRT*) genotypes contain proanthocyanidins (PAs) but only black ones contain anthocyanin⁴⁷.

Both putative *P1* and *P2* loci are on Pt05, approximately 50cM apart (Recombination Fraction = 0.32). *PtMYB113* (Psote05G0045200) is the most likely the candidate gene for the *P1* locus, despite being 1.2 Mb away from the current QTL range (Table S16), which might be partly attributable to our spaced genetic map. It is homologous to *MYB113* in Arabidopsis, the *R* locus gene (Glyma.09G235100) in soybean^{48,49} and *VuMYB90-1* in cowpea⁵⁰. *PtMYB114* (Psote05G0128800) gene is likely to be the gene for the adjacent QTL peak of the *P1* locus (Fig. S9). In contrast to the R2R3 MYB activators of the MBW complex in Arabidopsis, grape (*Vitis vinifera*), cowpea and cauliflower (*Brassica oleracea*) which have arisen from tandem duplication and the paralogues remain in very close linkage^{51, 52, 53}, *PtMYB113* (Chr05.1: 4,657,704..4,659,076) and *PtMYB114* (Chr05.1: 13,552,089..13,553,762) are not co-localised next to each but are still linked. It is deduced from a phylogenetic study that both belong to the Subgroup 6 which is the anthocyanin-related clade (Fig. 6a and Table S17)^{54, 55, 56}. Their orthologues (Fig. 6b and Table S18) are known to be the activators for late biosynthetic genes (LBGs) through interactions at their R3 repeats with a conserved motif ([D/E]L_{x2}[R/K]_{x3}L_{x6}L_{x3}R) in the N-terminal MYB-interacting region (MIR) of the bHLH TFs (Fig. 7)^{56,57}. The SG6 KPRPR[S/T]F motif is more conserved in *PtMYB113* than in *PtMYB114* but both are conserved in the anthocyanin-promoting (S/A)NDV motif (Fig. 7)^{58,59,54}. Although there is a G/A SNP at the 95th amino acid leading to a change from a hydrophobic glycine to hydrophilic glutamic acid in *PtMYB113* in the purple FP15 line, this same mutation was not observed in an additional four accessions, comprising of two purple-seeded and two cream-seeded lines. Similarly, the polymorphism of four SNPs and a C InDel in two introns were not consistently observed in these additional lines. No SNP variation was found at the 5' UTR region of *PtMYB113* as well as in the coding regions of *PtMYB114* (two SNPs were identified in introns) for both parental lines amplified with gene-specific primers (Table S19). A 13kb transposon *TgmR** was observed to be inserted into the Intron 2 of the soybean *R* locus⁴⁸. Despite the gene being expressed at low level, *TgmR** in the black seeded soybean is more methylated, preventing transposon binding and thus correct intron splicing could take place to upregulate anthocyanidin synthase genes⁵⁸. A similar scenario may be true in the winged bean *P1* locus. (Fig. 7). Instead of ANS, *PtANL2a* (Psote05G0083500) located at the *P1* locus is likely to be the target gene of this TF. In Arabidopsis, the loss-of-function *anl2* mutant had reduced anthocyanin levels in the subepidermal cells, in addition to generating an aberrant root structure⁵⁹. Studies from Wei et al. have suggested that the TTG1-dependent MBW complex could activate the expression of *ANL2* indirectly although the exact mechanism is yet to be unravelled⁶⁰. This is also supported by the previous finding that all four *ANL2* transcripts were up-regulated in winged bean leaf containing higher tannin with the fold-change ranging from 1.2 to 4.3, despite not reaching statistical significance⁶¹. In line with Wei et al.'s finding, the absence of *PtMYB113* expression in this transcriptomic study leads us to postulate that its expression could be at an earlier stage for *ANL2* activation before flavonoid begins to be accumulated. It is also worth noting that the

clustering of anthocyanin biosynthesis pathway genes correlates well with high LOD scores observed throughout Chr05.1 (Fig S10).

Discussion

The winged bean genome is of value for molecular breeding improvement as most plant species with sequenced genomes are taxonomically distant, limiting the cross-species use of positional information. Stefanović and colleagues⁶³ have suggested *P. tetragonolobus* diverged approximately 11 Mya from *Erythrina soursae*. The common ancestor between this clade and the clade consisting of Phaseoleae subtribe Glycininae and subtribe Phaseolinae which arose approximately 21 Mya, with an estimated divergence slightly earlier (approximately 23 Mya) from the Cajaninae clade⁶³. This first winged bean genome lays the foundation for crop genome guided breeding improvement programmes, to enhance the wider adoption of the species as a protein source for the future, given the limited conserved synteny with the well annotated *P. vulgaris*. The high levels of genetic diversity seen within the breeding material diversity study could be explained by: i) *ex-situ* conservation efforts by the gene banks might have been conducted early, ii) geographical preferences for different edible parts and cropping methods have limited the loss of genetic diversity overall, iii) seed phenotypes have not been as strongly visually distinctive compared to other legumes, particularly with limited variation in seed mottling and hilum pattern, with the major selection focus on the vegetative parts of the plant.

Thus far, our approach in integrating Illumina, ONT, BioNano and genetic mapping have resulted in a pseudo-chromosome level genome spanning 530.28 Mb. The genome annotation has facilitated the identification of a number of candidate genes associated with QTL relevant to ideotype development for breeding selection, in particularly, variability for seed protein content, plant architecture and pigmentation. In addition, it has been possible to identify the regions of the genome containing QTL flanking marker variation, potentially allowing the development of simple KASP markers for further selection of desirable architectural, protein and nutritional traits. Selection for alleles of genes associated with these candidates (if validated) could be subsequently utilised for varietal development or flanking markers around the QTL of interest could be used directly for breeding.

Materials And Methods

Library preparation of material for whole genome sequencing

The common paternal line Ma3 for both Cross XT and Cross XB was selected for whole genome sequencing, based on single-seed descent derived material from the individual plant used during controlled hybridisation. The plant was grown in Future Crop glasshouse under 28°C day/23°C night temperature cycle (Sutton Bonington, UK). High molecular weight (HMW) genomic DNA was isolated from 100 mg fresh leaves using DNeasy Mini Kit (Qiagen). DNA quantity and quality were assessed by Nanodrop, Qubit dsDNA BR assay (ThermoFisher) and the Genomic DNA ScreenTape assessed (Agilent) prior to library preparation.

ONT library preparation: A rapid library prep was prepared from 400 ng of DNA (ONT; SQK-RAD004) and run over a MinION flow cell (ONT; FLO-MIN106 R9.4) on the GridION X5 mk. A small-scale ligation sequencing

library (ONT; SQK-LSK109) was also prepared and run over a MinION flow cell on the GridION. A large-scale ligation sequencing library was subsequently prepared and the entire library prep was loaded onto a PromethION flow cell (ONT; FLO-PRO001) and sequenced on the PromethION platform.

Illumina library preparation: An Illumina-compatible sequencing library was prepared from 500 ng of DNA using the KAPA HyperPlus Kit (Roche) and the KAPA Single-Indexed Adapter Kit, Illumina Platforms, Set B (Roche). Genomic DNA was fragmented for 10 min and 2 cycles of PCR were used during the library amplification stage. Post-amplification SPRI-based size selection was performed using a 0.7/0.9 ratios of AMPure XP beads (Beckman Coulter). The library was quantified using the Qubit Fluorometer and the Qubit dsDNA HS Assay Kit (ThermoFisher) and the fragment length distribution was assessed using the Agilent TapeStation 4200 and the Agilent High Sensitivity D1000 ScreenTape Assay (Agilent). Final library QC was performed using the KAPA Library Quantification Kit for Illumina (Roche). The library was sequenced on the Illumina NextSeq500 (NextSeq Control Software v2.2.0) using a NextSeq500 Mid Output 300 cycle kit v2.5 (Illumina) to generate 150-bp paired-end reads.

Bionano mapping

High molecular weight genomic DNA was extracted from fresh young leaves using the Bionano Prep Plant Tissue DNA Isolation Kit (Bionano Genomics: 80003) according to the Isolation Base Protocol (Bionano Genomics: Document #30068-RevD). Due to the partially insoluble nature of the DNA quantification was not possible so 21 µl (including the bulk of DNA) was taken into labelling.

DNA was labelled with DLE-1 enzyme using the Bionano Prep DLS Kit (Bionano Genomics: 80005) and following the standard protocol (Bionano Genomics: Document #30206-RevF). The labelled sample was quantified by Qubit Fluorometer 4 (Thermo Fisher: Q33238) and the Qubit dsDNA HS Assay Kit (Thermo Fisher: Q32854). The average concentration of the labelled sample was 3.36 ng/µl (CV = 0.09). The labelled DNA reaction was run over one flow cell on a Bionano Saphyr Chip (Bionano Genomics: 20319) on the Bionano Saphyr to generate 540 Gbp of data. For analysis, the molecules file was used to generate a *de novo* assembly using the default Bionano Access settings (Bionano Access: 1.3.0; Bionano Tools: 1.3.8041.8044; Bionano Solve: Solve3.3_10252018; RefAligner: 7915.7989rel; HybridScaffold/SVMerge/VariantAnnotation: 10252018). This assembly was used to generate a hybrid scaffold from the winged bean NGS assembly. The hybrid scaffold was constructed using default settings; conflict resolution was set to 'Resolve Conflicts' for both the Bionano assembly and sequence assembly.

Genome Assembly and TE characterisation

Raw ONT reads were firstly filtered by quality score (< 7) and read length (< 1k), and then trimmed and corrected by Canu v1.7⁶⁴. Processed reads with high qualities were assembled by wtdbg2 with default settings⁶⁵. Assembled contigs were polished twice by Illumina PE reads using Racon⁶⁶ and further polished by Pilon v1.22⁶⁷. EDTA pipeline⁶⁸ with default settings was used for transposable elements characterisation with the classification in reference to Wicker et al.⁶⁹. The Illumina assembly was subjected to K-mer analyses using both Jellyfish⁷⁰ and GenomeScope⁷¹.

Gene Annotation

De novo transcriptome assembly was conducted by Trinity v2.11.0⁷² with previously published RNA-seq data¹⁴ (PRJNA374598). Raw assembled transcripts were clustered by CD-HIT⁷³ and corrected by CAP3⁷³. Protein coding regions were identified by TransDecoder. Genome annotation was done by MAKER2 v2.31.10⁷⁴ using transcriptome annotation. The completeness of annotation was estimated using BUSCO⁷¹ with settings -l embryophyta_odb9 -m genome -c 12 -long -species arabidopsis. The types of transcription factors were predicted from Plant Transcription Factor Database (PlantTFDB, <http://planttfdb.gao-lab.org/index.php>)⁷⁵. Collinear blocks are defined using default parameters in MCScanX⁷⁶ and visualised using SynVisio⁷⁷.

Genotype-by-Sequencing (GbS) using DArTseq

Genomic DNA was extracted from young leaves using Qiagen DNeasy® 96 Plant Kit and subsequently genotyped by DArTseq™ platform (Diversity Arrays Technology Pty Ltd) in which the DNA was fragmented through *Pst*I-*Mse*I double digestion, followed by adaptor ligation, amplification and then sequenced on Illumina HiSeq 2500. SNP and PAV calling were then carried out using proprietary DArT analytical pipelines⁷⁸.

Genetic mapping

SNP markers from DArTseq were filtered so as to be polymorphic between parental lines and to have no missing values across the populations for a total of 221 and 183 F₂ individuals, respectively in XB and XT populations. Maps were constructed using JoinMap® v5⁷⁹, first linkage groups (LGs) were obtained through a logarithm of odds (LOD) value of 30 using “grouping (tree)” function. Initial mapping was then performed using a maximum likelihood (ML) algorithm, in order to deal with computational load of more than 100 markers in each LG. After sufficient markers had been removed using iterative reduction of markers with ML, final LGs were generated using regression mapping with default parameters: recombination frequency ≤ 0.40 , LOD value ≥ 1.0 , goodness-of-fit jump threshold for removal of loci equal to 5.0, and rippling performed after every marker was added into the map. Recombination frequency was converted into map distance (cM) using the Haldane’s function and markers were checked for nearest neighbour (NN) fit values and mean χ^2 contribution for the goodness- of-fit within each LG map.

Markers showing significant zygotic segregation distortion (SD) were first removed from initial mapping calculations, and later re-introduced using regression mapping. This was monitored by comparing initial LGs (without SD markers) with newly obtained ones (with SD markers), in order to ensure the marker order between non-distorted markers was consistent. To improve the QTL analysis for *qCC*, *qSC*, and *qCLX*, markers on XB.lg05 were filtered out with a 0.95 threshold and the linkage group was mapped using a maximum likelihood algorithm.

Genetic diversity and population structure

Marker quality control was performed by filtering biallelic SNP loci with minor allele frequency (MAF) > 0.01 and missing value of not more than 10%. The ancestry of each individual was estimated using fastSTRUCTURE⁸⁰ with admixture as the ancestry model. fastSTRUCTURE was run using the structure.py script provided with the software. The software was run using multiple choices of K (K = 1 to K = 8) to investigate the structure of the data. ChooseK.py was used to choose the most likely number of clusters.

Maximum likelihood was used and genotypes were assigned to a subpopulation. The membership coefficient for clusters was 0.7 or else genotypes were considered as admixed.

Missing values were numerically imputed before calculating allelic frequency using TASSEL v5⁸¹ with default settings. Principal Coordinate Analysis (PCoA) was performed using TASSEL. The Neighbour-joining (NJ) tree was permuted with 1,000 bootstrap values using R, *ape* v5.4⁸², *ade4* v1.7⁸³, *ade4* v2.1.3⁸⁴ and *poppr* v2.8.7⁸⁵. The NJ tree was viewed using iTOL⁸⁶. The global map was extracted from <https://mapchart.net/>. Biological replicates were confirmed from the pedigree record if they were i) clustered into same node in the NJ tree, ii) at a distance not more than 6% (Table S6).

Non biological replicates were identified and one replicate excluded, SNP were further filtered with no missing values allowed across the germplasms, resulting in 5,581 SNPs for calculating molecular diversity indices. The average pairwise divergence among genotypes, which represents the nucleotide diversity per bp, π (pi) and the expected number of polymorphic sites per nucleotide, θ (theta), were estimated using TASSEL. The normalized measure of difference between the observed (π) and expected (θ) nucleotide diversity, Tajima's D, was also computed. The population genetic analysis was carried out in R; *dartR* v1.8.3⁸⁶, *poppr* v2.8.7⁸⁵ and *mmod* v1.3.3⁸⁷. Permutation of Monte-Carlo test was run at 999.

Linkage disequilibrium (LD) decay

The correlation coefficient (r^2) among pairwise comparisons of loci from SNPs having 10% missing values were calculated using TASSEL v5 with the default sliding window setting. The LD decay analysis was conducted with two levels of MAF thresholds; > 0.01 and > 0.05. The LD decay rate was calculated using an R script written by F. Marroni available at <https://fabiomarroni.wordpress.com/2011/08/09/estimate-decay-of-linkage-disequilibrium-with-distance/>

Protein profiling

Ground winged bean seeds were passed through a 0.5mm sieve (Ultra-Centrifugal Mill ZM 200, Retsch). The samples were stored at -80°C overnight and then freeze-dried. The protein content was calculated based on modified a Dumas method using a Protein Analyzer (FlashEA® 1112 N/Protein, Thermo Scientific) with the conversion factor 6.25.

QTL mapping

F₂ individuals of the XB Cross were assessed under field conditions together with the parental lines. They were grown on ridges with trellis (1m x 2m) and 1.5m intra-spacing and 1m inter-spacing at the Field Research Centre of Crops For the Future (Semenyih, Malaysia). Seeds were scarified using sandpaper. The seedlings were then transplanted 14 days after emergence (DAE). Fertilizer (NPK 15:15:15) was applied 4 times during the growing season (estimated to be 15kg/ha), while pesticide was sprayed approximately every 10 days to prevent pest damage of the main stems (Karate – Syngenta, Switzerland) at 0.5mL/L concentration.

For the materials used for nutritional profiling, five replicates of each parental genotype were grown in a complete randomized design (CRD) from October 2016 to March 2017 in a sandy loam soil (pH 5.3). The XB

F₂ lines were assessed in three blocks between June to November 2017 in a complete randomized block design (CRBD), each with five replicates of each parent. The soil had a sandy loam profile with pH 5.0, while the day/night temperatures recorded were 32 ± 0.9°C and 23 ± 0.3°C (on site weather station; DeltaT). Traits measurements from 87 individuals (which were not border plants) were recorded as reported⁵. In addition, the average chlorophyll concentration was obtained from the middle leaflets of the top three fully opened leaves, each measured 3 times at 28 DAE using a SPAD-502 (Konica Minolta). For *qTNoB*, the number of branches along the entire stem were calculated, as opposed to *qNoB* in which represents the number of branches within the first ten nodes.

QTL analyses were conducted using MapQTL v6⁸⁸ using both non-parametric and parametric tests. Each trait was first analysed through a Kruskal-Wallis (KW) test, in order to establish single marker-trait associations (at $p < 0.01$). This was followed by Interval Mapping (IM) analysis with LOD threshold calculated through 10,000 permutation tests, taking the genome-wide (GW) significant threshold at $\alpha = 0.05$. Only QTLs that were consistent between these two tests were reported. A QTL was considered significant when equal to or above the GW-LOD threshold and explained $\geq 10\%$ of phenotypic variance (PVE). Multiple-QTL model (MQM) mapping was utilised for seed protein content QTL analysis. All QTL have been mapped onto their corresponding LG using MapChart v2.32⁸⁹, including markers as supporting intervals using a 2-LOD drop.

Phylogenetic analysis of R2R3 MYB genes

For the MYB subgroup characterisation, the amino acid sequences (sequence accessions listed in Table S17 and S18) were aligned using MUSCLE⁹⁰ in MEGA X⁹¹ using default settings. The phylogeny was constructed as a NJ tree with 5,000 bootstrap replicates.

Data availability:

The reference genome sequence and annotation files are available at ORCAE, <https://bioinformatics.psb.ugent.be/orcae/>. [underway] Raw reads were deposited in the NCBI Sequence Read Archive under the BioProject accession numbers xxxxxxxx.

Declarations

Acknowledgements

This work was supported by Crops For the Future, University of Nottingham Malaysia, Future Food Beacon and Deep Seq, University of Nottingham. We are grateful to East West Seed for their contribution of lines from their breeding programme in the Philippines. We are grateful to Dr Jon Stubberfield for technical assistance and method development.

Conflict of interests

All authors declare no conflict of interest.

Author Contributions

A.T.S. generated mapping populations, managed the field work and phenotypic measurement, developed genetic maps and performed QTL analysis. F.S. assembled the genome, characterised TE and gene annotation, performed K-mer and BUSCO analyses. C.M. generated Bionano mapping and assembly. N.T. performed protein profiling. W.K.H. performed data analysis, research coordination and prepared the manuscript. A.T.S., F.S., C.M. and N.T. contributed to the writing of the manuscript. N.S. wrote scripts for data format conversion and performed population structure analysis. S.M. supervised the scope of this study and revised the manuscript. F.M. edited the manuscript. S.M. and F.M. are the project leaders. All authors read and approved the manuscript.

References

1. Harder, D. K. & Smartt, J. Further evidence on the origin of the cultivated winged bean, *Psophocarpus tetragonolobus* (L.) DC. (Fabaceae): Chromosome numbers and the presence of a host-specific fungus. *Econ. Bot.* **46**, 187–191 (1992).
2. Tanzi, A. S. *et al.* Winged bean (*Psophocarpus tetragonolobus* (L.) DC.) for food and nutritional security: synthesis of past research and future direction. *Planta* **250**, 911–931 (2019).
3. Hildebrand, D. F., Blevins, D. G. & Hymowitz, T. Cross-inoculation specificity of *Psophocarpus tetragonolobus* (L.) DC. *Plant Soil* **60**, 139–142 (1981).
4. Iruthayathas, E. E. & Vlassak, K. Promiscuity of some legume species with winged bean *Rhizobium* strains. *Sci. Hortic. (Amsterdam)*. **31**, 1–9 (1987).
5. Tanzi, A. S., Ho, W. K., Massawe, F. & Mayes, S. Development and interaction between plant architecture and yield-related traits in winged bean (*Psophocarpus tetragonolobus* (L.) DC.). *Euphytica* **215**, 36 (2019).
6. National Research Council. 1981. Winged Bean: A High-Protein Crop for the Tropics. Washington, DC: The National Academies Press. <https://doi.org/10.17226/19754>
7. Gross, R. Composition and protein quality of winged bean (*Psophocarpus tetragonolobus*). *Qual. Plant. Plant Foods Hum. Nutr.* **32**, 117–124 (1983).
8. Garcia, V. V. & Palmer, J. K. Proximate analysis of five varieties of winged beans, *Psophocarpus tetragonolobus* (L.) DC*. *Int. J. Food Sci. Technol.* **15**, 469–476 (2007).
9. Kotaru, M. Investigations of antinutritional factors of the winged bean (*Psophocarpus tetragonolobus*). *Food Chem.* **24**, 279–286 (1987).
10. Adegboyega, T. T. *et al.* Nutrient and Antinutrient Composition of Winged Bean (*Psophocarpus tetragonolobus* (L.) DC.) Seeds and Tubers. *J. Food Qual.* **2019**, 1–8 (2019).
11. Tan, N. H., Rahim, Z. H. A., Khor, H. T. & Wong, K. C. Winged bean (*Psophocarpus tetragonolobus*) tannin level, phytate content and hemagglutinating activity. *J. Agric. Food Chem.* **31**, 916–917 (1983).

12. Sambudi, H. & Buckle, K. A. Characteristics of winged bean (*Psophocarpus tetragonolobus*) seeds during soaking and boiling. *J. Sci. Food Agric.* **57**, 585–595 (1991).
13. Wong, Q. *et al.* Development of Gene-Based SSR Markers in Winged Bean (*Psophocarpus tetragonolobus* (L.) DC.) for Diversity Assessment. *Genes (Basel)*. **8**, 100 (2017).
14. Hunt, S. P. *et al.* A Chromosome-Scale Assembly of the Garden Orach (*Atriplex hortensis* L.) Genome Using Oxford Nanopore Sequencing. *Front. Plant Sci.* **11**, (2020).
15. Vatanparast, M. *et al.* Transcriptome sequencing and marker development in winged bean (*Psophocarpus tetragonolobus*; Leguminosae). *Sci. Rep.* **6**, 29070 (2016).
16. Bennett, M. D. & Smith, J. B. Nuclear dna amounts in angiosperms. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **274**, 227–274 (1976).
17. Xia, Q. *et al.* The genome assembly of asparagus bean, *Vigna unguiculata* ssp. *sesquipedialis*. *Sci. Data* **6**, 124 (2019).
18. Lonardi, S. *et al.* The genome of cowpea (*Vigna unguiculata* [L.] Walp.). *Plant J.* **98**, 767–782 (2019).
19. DOE-JGI and USDA-NIFA. *Phaseolus vulgaris* v2.1. <http://phytozome.jgi.doe.gov>.
20. Schmutz *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
21. Schmutz, J. *et al.* A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* **46**, 707–713 (2014).
22. Laosatit, K., Amkul, K., Chankaew, S. & Somta, P. Molecular Genetic Diversity of Winged Bean Gene Pool in Thailand Assessed by SSR Markers. *Hortic. Plant J.* (2021) doi:10.1016/j.hpj.2021.05.001.
23. Yang, S., Grall, A. & Chapman, M. A. Origin and diversification of winged bean (*Psophocarpus tetragonolobus* (L.) DC.), a multipurpose underutilized legume. *Am. J. Bot.* **105**, 888–897 (2018).
24. Yan, J. *et al.* Genetic Characterization and Linkage Disequilibrium Estimation of a Global Maize Collection Using SNP Markers. *PLoS One* **4**, e8451 (2009).
25. Li, M. *et al.* Identification of genetic loci and candidate genes related to soybean flowering through genome wide association study. *BMC Genomics* **20**, 987 (2019).
26. Wen, Z. *et al.* Genome-wide association mapping of quantitative resistance to sudden death syndrome in soybean. *BMC Genomics* **15**, (2014).
27. Delfini, J. *et al.* Population structure, genetic diversity and genomic selection signatures among a Brazilian common bean germplasm. *Sci. Rep.* **11**, 2964 (2021).
28. Raggi, L., Caproni, L., Carboni, A. & Negri, V. Genome-Wide Association Study Reveals Candidate Genes for Flowering Time Variation in Common Bean (*Phaseolus vulgaris* L.). *Front. Plant Sci.* **10**, (2019).

29. Muñoz-Amatriaín, M. *et al.* The UCR Minicore: a valuable resource for cowpea research and breeding. *Legum. Sci.* (2021) doi:10.1002/leg3.95.
30. Yanagi, S. O. Properties of Winged Bean (*Psophocarpus tetragonolobus*) Protein in Comparison with Soybean (*Glycine max*) and Common Bean (*Phaseolus vulgaris*) Protein. *Agric. Biol. Chem.* **47**, 2273–2280 (1983).
31. Makeri, M. U., Mohamed, S. A., Karim, R., Ramakrishnan, Y. & Muhammad, K. Fractionation, physicochemical, and structural characterization of winged bean seed protein fractions with reference to soybean. *Int. J. Food Prop.* 1–17 (2017) doi:10.1080/10942912.2017.1369101.
32. Verma, S. & Bhatia, S. Analysis of genes encoding seed storage proteins (SSPs) in chickpea (*Cicer arietinum* L.) reveals co-expressing transcription factors and a seed-specific promoter. *Funct. Integr. Genomics* **19**, 373–390 (2019).
33. Warsame, A. O., Michael, N., O’Sullivan, D. M. & Tosi, P. Identification and Quantification of Major Faba Bean Seed Proteins. *J. Agric. Food Chem.* **68**, 8535–8544 (2020).
34. Ochiai Yanagi, S., Yoshida, N. & Saio, K. Generality and Diversity of Winged Bean (*Psophocarpus tetragonolobus*) Protein in Eight Various Lines. *Agric. Biol. Chem.* **47**, 2267–2271 (1983).
35. Yin, F. *et al.* Analysis of common bean expressed sequence tags identifies sulfur metabolic pathways active in seed and sulfur-rich proteins highly expressed in the absence of phaseolin and major lectins. *BMC Genomics* **12**, 268 (2011).
36. Saio, K., Nakano, Y. & Uemoto, S. Microstructure of Winged Beans. *Food Microstructure* **2**, 175–181 (1983).
37. YANAGI, S. Winged bean protein in comparison with other legume protein. *JARQ* **18**, 53–59 (1984).
38. Obala, J. *et al.* Seed protein content and its relationships with agronomic traits in pigeonpea is controlled by both main and epistatic effects QTLs. *Sci. Rep.* **10**, 214 (2020).
39. Zhang, Y. H. *et al.* Marker-assisted breeding for transgressive seed protein content in soybean [*Glycine max* (L.) Merr.]. *Theor. Appl. Genet.* **128**, 1061–1072 (2015).
40. Krajewski, P. *et al.* QTL for yield components and protein content: A multienvironment study of two pea (*Pisum sativum* L.) populations. *Euphytica* **183**, 323–336 (2012).
41. Gong, Q. *et al.* Meta-analysis of soybean amino acid QTLs and candidate gene mining. *J. Integr. Agric.* **17**, 1074–1084 (2018).
42. Zhang, H. *et al.* Transcriptome profiling reveals the spatial-temporal dynamics of gene expression essential for soybean seed development. *BMC Genomics* **22**, 453 (2021).

43. Parreira, J. R., Cappuccio, M., Balestrazzi, A., Fevereiro, P. & Araújo, S. de S. MicroRNAs expression dynamics reveal post-transcriptional mechanisms regulating seed development in *Phaseolus vulgaris* L. *Hortic. Res.* **8**, 18 (2021).
44. Li, Y., Xia, T., Gao, F. & Li, Y. Control of Plant Branching by the CUC2/CUC3-DA1-UBP15 Regulatory Module. *Plant Cell* **32**, 1919–1932 (2020).
45. Zheng, G. *et al.* Conserved and novel roles of miR164- CUC 2 regulatory module in specifying leaf and floral organ morphology in strawberry. *New Phytol.* **224**, 480–492 (2019).
46. Malaviya, D. R., Roy, A. K., Kaushal, P., Yadav, A. & Pandey, D. K. Complementary gene interaction and xenia effect controls the seed coat colour in interspecific cross between *Trifolium alexandrinum* and *T. apertum*. *Genetica* **147**, 197–203 (2019).
47. Ha, J. *et al.* Transcriptomic variation in proanthocyanidin biosynthesis pathway genes in soybean (*Glycine* spp.). *J. Sci. Food Agric.* **98**, 2138–2146 (2018).
48. Zabala, G. & Vodkin, L. O. Methylation affects transposition and splicing of a large CACTA transposon from a MYB transcription factor regulating anthocyanin synthase genes in soybean seed coats. *PLoS One* **9**, (2014).
49. Gao, R. *et al.* MYB transcription factors GmMYBA2 and GmMYBR function in a feedback loop to control pigmentation of seed coat in soybean. *J. Exp. Bot.* **72**, 4401–4418 (2021).
50. Li, Y. *et al.* Integrated Metabolomics and Transcriptomics Analyses Reveal the Molecular Mechanisms Underlying the Accumulation of Anthocyanins and Other Flavonoids in Cowpea Pod (*Vigna unguiculata* L.). *J. Agric. Food Chem.* **68**, 9260–9275 (2020).
52. Lin-Wang, K. *et al.* An R2R3 MYB transcription factor associated with regulation of the anthocyanin biosynthetic pathway in Rosaceae. *BMC Plant Biol.* **10**, 50 (2010).
53. Chiu, L.-W. *et al.* The Purple Cauliflower Arises from Activation of a MYB Transcription Factor. *Plant Physiol.* **154**, 1470–1480 (2010).
53. Herniter, I. A., Muñoz-Amatriáin, M., Lo, S., Guo, Y.-N. & Close, T. J. Identification of Candidate Genes Controlling Black Seed Coat and Pod Tip Color in Cowpea (*Vigna unguiculata* [L.] Walp). *G3 Genes/Genomes/Genetics* **8**, 3347–3355 (2018).
54. Stracke, R., Werber, M. & Weisshaar, B. The R2R3-MYB gene family in *Arabidopsis thaliana*. *Curr. Opin. Plant Biol.* **4**, 447–456 (2001).
55. Deng, J. *et al.* Multiple MYB Activators and Repressors Collaboratively Regulate the Juvenile Red Fading in Leaves of Sweetpotato. *Front. Plant Sci.* **11**, (2020).
56. Liu, J., Osbourn, A. & Ma, P. MYB Transcription Factors as Regulators of Phenylpropanoid Metabolism in Plants. *Mol. Plant* **8**, 689–708 (2015).

57. Yan, F. *et al.* Loss-of-Function Mutation of Soybean R2R3 MYB Transcription Factor Dilutes Tawny Pubescence Color. *Front. Plant Sci.* **10**, (2020).
58. Muñoz-Gómez, S., Suárez-Baron, H., Alzate, J. F., González, F. & Pabón-Mora, N. Evolution of the Subgroup 6 R2R3-MYB Genes and Their Contribution to Floral Color in the Perianth-Bearing Piperales. *Front. Plant Sci.* **12**, (2021).
59. Plunkett, B. J. *et al.* MYBA From Blueberry (*Vaccinium* Section *Cyanococcus*) Is a Subgroup 6 Type R2R3MYB Transcription Factor That Activates Anthocyanin Production. *Front. Plant Sci.* **9**, (2018).
60. Kubo, H., Peeters, A. J. M., Aarts, M. G. M., Pereira, A. & Koornneef, M. ANTHOCYANINLESS2, a Homeobox Gene Affecting Anthocyanin Distribution and Root Development in *Arabidopsis*. *Plant Cell* **11**, 1217–1226 (1999).
61. Wei, Z. *et al.* Genome-Wide Identification of Direct Targets of the TTG1–bHLH–MYB Complex in Regulating Trichome Formation and Flavonoid Accumulation in *Arabidopsis thaliana*. *Int. J. Mol. Sci.* **20**, 5014 (2019).
62. Singh, V., Goel, R., Pande, V., Asif, M. H. & Mohanty, C. S. De novo sequencing and comparative analysis of leaf transcriptomes of diverse condensed tannin-containing lines of underutilized *Psophocarpus tetragonolobus* (L.) DC. *Sci. Rep.* **7**, 44733 (2017).
63. Stefanović, S., Pfeil, B. E., Palmer, J. D. & Doyle, J. J. Relationships among phaseoloid legumes based on sequences from eight chloroplast regions. *Syst. Bot.* **34**, 115–128 (2009).
64. Jason R. Miller, N. H. B. and A. M. P. S. K. B. P. W. K. B. Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
65. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **17**, 155–158 (2020).
66. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Fast accurate novo genome Assem. from long uncorrected reads* 068122 (2016) doi:10.1101/068122.
67. Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
68. Ou, S. *et al.* Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline. *bioRxiv* (2019) doi:10.1101/657890.
69. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
70. G, M. & C., K. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).

71. Vurture, G. W. *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
72. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
73. Li, W. & Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
74. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
75. Jin, J. *et al.* PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* **45**, D1040–D1045 (2017).
76. Wang, Y. *et al.* MCSanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49–e49 (2012).
77. Bandi, V. & Gutwin, C. Interactive Exploration of Genomic Conservation. in *Proceedings of the 46th Graphics Interface Conference on Proceedings of Graphics Interface 2020 (GI'20)* (Canadian Human-Computer Communications Society, Waterloo, CAN, 2020).
78. Kilian, A., Sanewski, G. & Ko, L. The application of DArTseq technology to pineapple. *Acta Hort.* 181–188 (2016) doi:10.17660/ActaHortic.2016.1111.27.
79. Van Ooijen, J. W. JoinMap 5 ® Software for the calculation of genetic linkage maps in experimental populations of diploid species. *Kyazma B.V., Wageningen, Netherlands* 1–61 (2019).
80. Raj, A., Stephens, M. & Pritchard, J. K. FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589 (2014).
81. Bradbury, P. J. *et al.* TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
82. Paradis, E. & Schliep, K. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
83. Thioulouse, J. *et al.* *Multivariate Analysis of Ecological Data with ade4*. (Springer New York, 2018). doi:10.1007/978-1-4939-8850-1.
84. Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405 (2008).
85. Kamvar, Z. N., Tabima, J. F. & Grünwald, N. J. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2**, e281 (2014).

86. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
87. Winter, D. J. <scp>mmod</scp>: an R library for the calculation of population differentiation statistics. *Mol. Ecol. Resour.* **12**, 1158–1160 (2012).
88. Van Ooijen, J. W. MapQTL 6 - Software for the mapping of quantitative trait loci in experimental populations of diploid species. *Kyazma B. V.* 64 (2009).
89. Voorrips, R. E. MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs. *J. Hered.* **93**, 77–78 (2002).
90. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
91. S., K., G., S., M., L., C., K. & K., T. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).

Figures

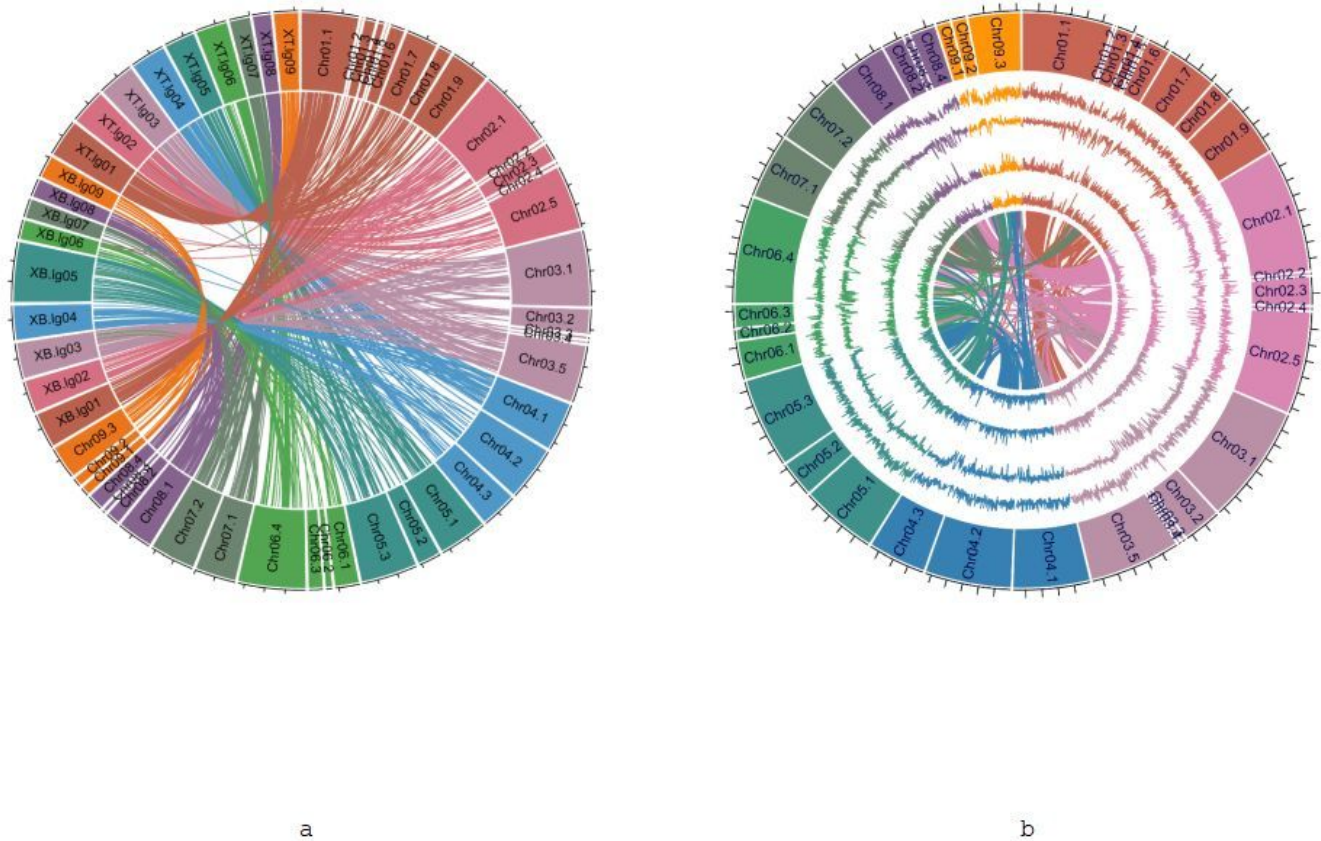


Figure 1

Overview of the winged bean [*Psophocarpus tetragonolobus* (L.) DC] genome assembly (a) anchored by two genetic maps (Cross XT and Cross XB) and its (b) genome characterisation. The blocks represent nine pseudomolecules featured in 5 Mbp intervals. Tracks in (b) displayed are as follow: outer most (i) chromosome name, (ii) gene density, (iii) TE elements density, (iv) SNP density, (v) presence-absence variation density (PAV), (vi) intra-genomic syntenic blocks. The density was calculated using 100 kb non-overlap window.

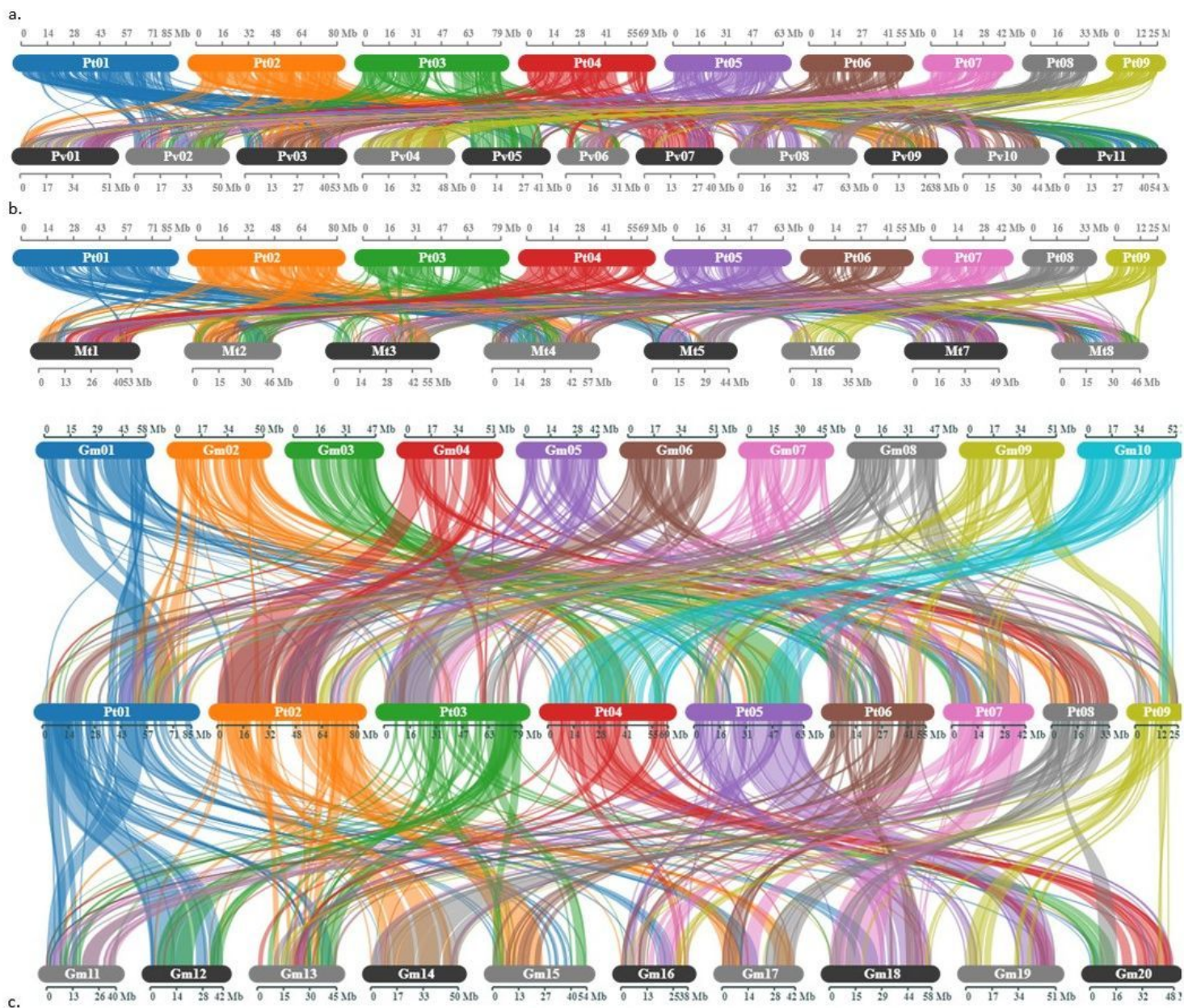


Figure 2

The syntenic blocks found between winged bean with (a) *P. vulgaris*; (b) *M. truncatula* and (c) *G. max*. A block of 1000 Ns was added between scaffolds for simplified chromosomal level visualisation.

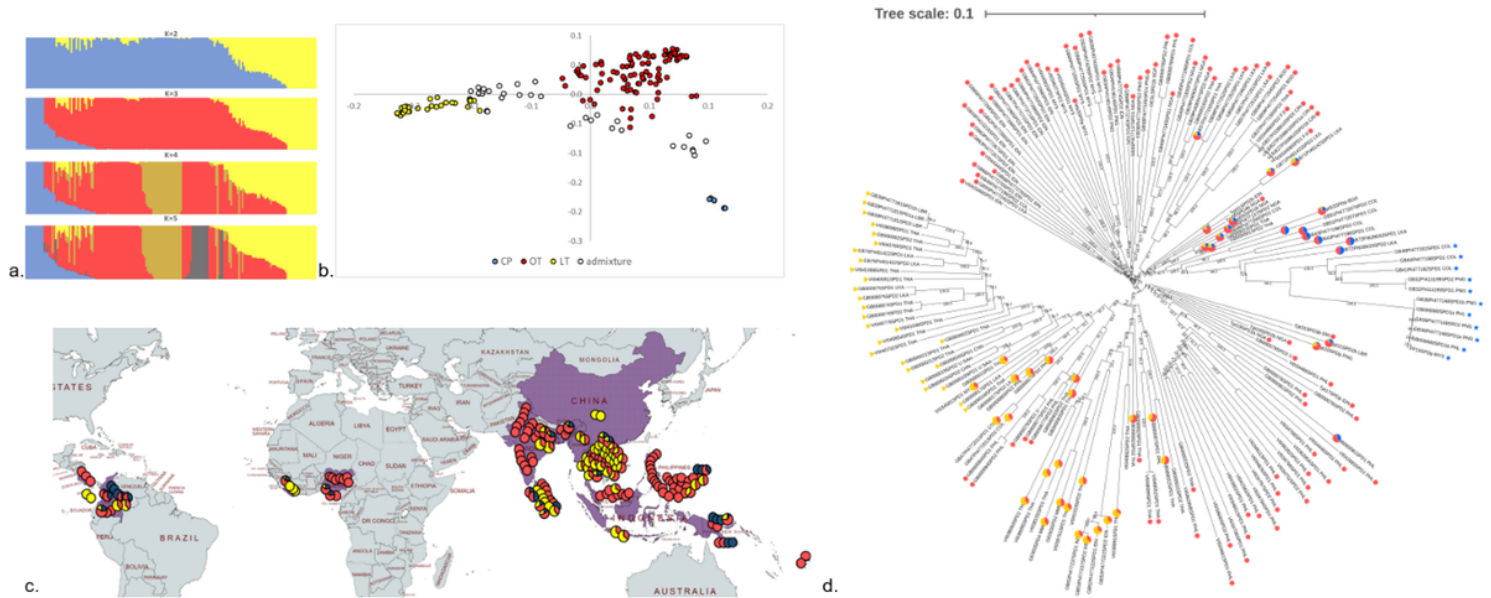


Figure 3

(a) A membership probability plot from CLUSTAL for K = 2 to K = 5. (b) Clustering from PCoA analysis; suggesting the most likely number of subpopulations is three. (c) An overview of the germplasm collected from different countries coloured by subpopulation. (d) Phylogenetic relationship of germplasm included in this study. blue star: CP subpopulation, red dot: OT subpopulation, yellow triangle: LT subpopulation.

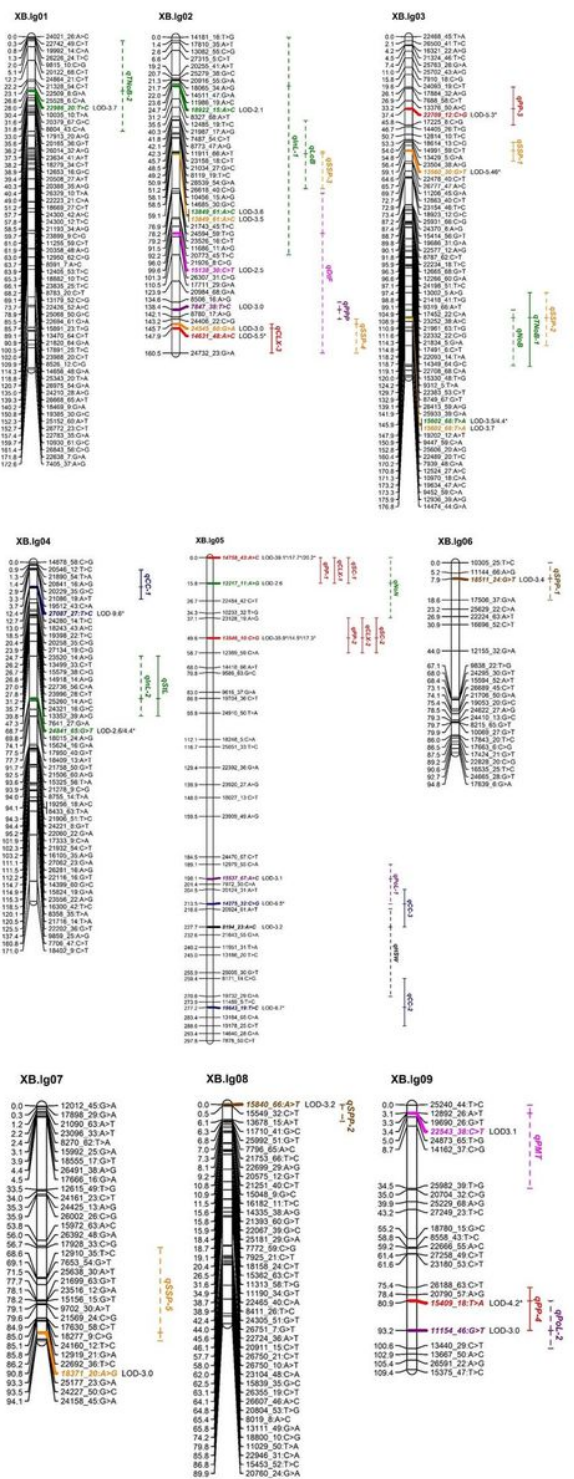


Figure 4

Linkage groups and QTLs for *TNoB*: total branch number per plant; *InL*: internode length; *LoB*: length of branch; *DtF*: days to first open flower; *PPP*: mature pods per plant; *CLX*: calyx colour; *PP*: specks presence/pod pigmentation; *NoB*: branch number per plant; *StL*: stem length; *SC*: seed coat colour; *NoN*: number of nodes; *PoL*: dry pod length; *HSW*: 100- seed weight; *SPP*: seeds per pod; *PMT*: pod maturing time; *SSP*: seed protein percentage. Vertical lines represent the 2-LOD support interval for significant QTL (LOD $\geq$ GW-threshold). Dashed lines represent QTL support intervals for putative QTL.

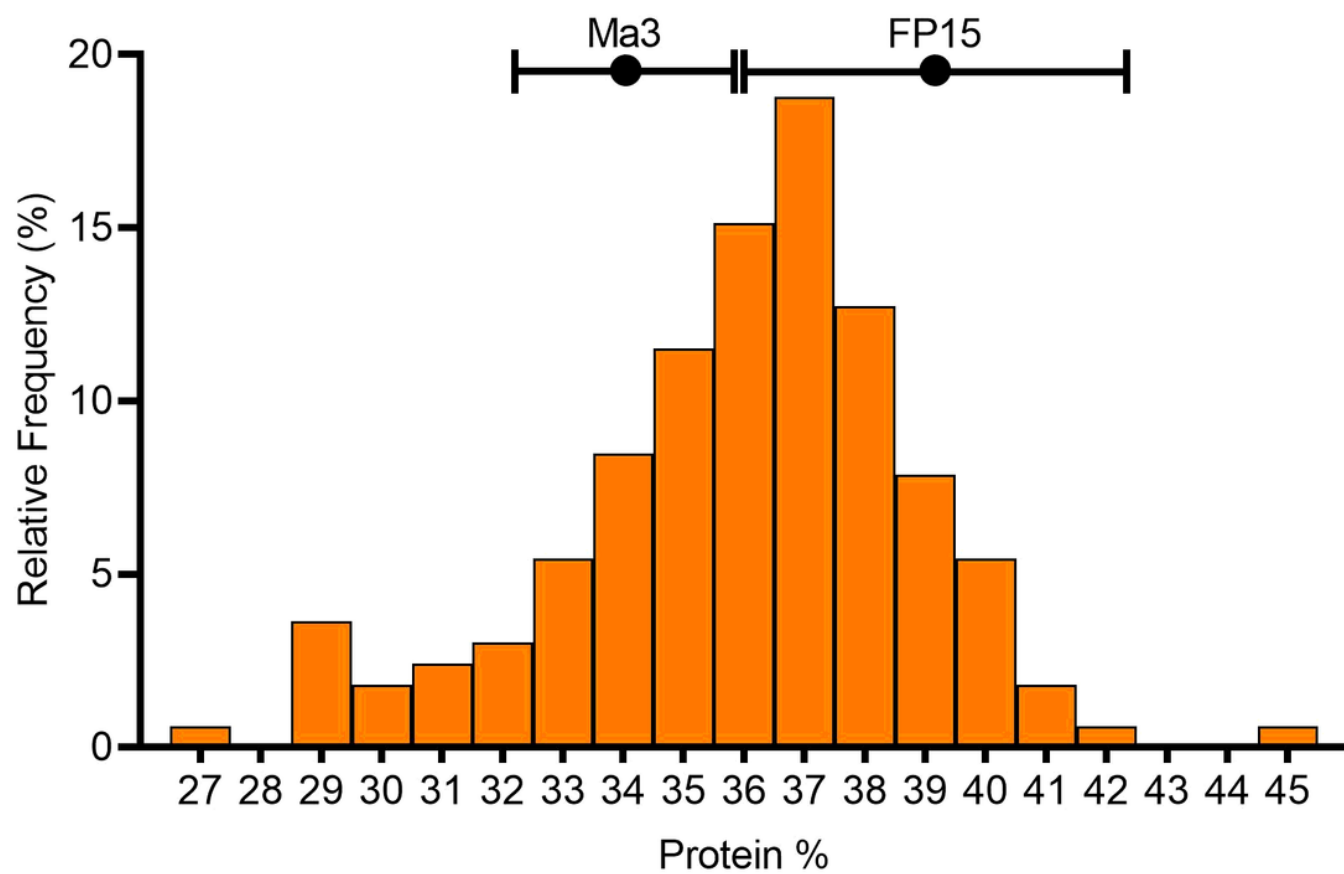


Figure 5

Transgressive segregation in seed protein content (%) observed in the XB population (n=161).

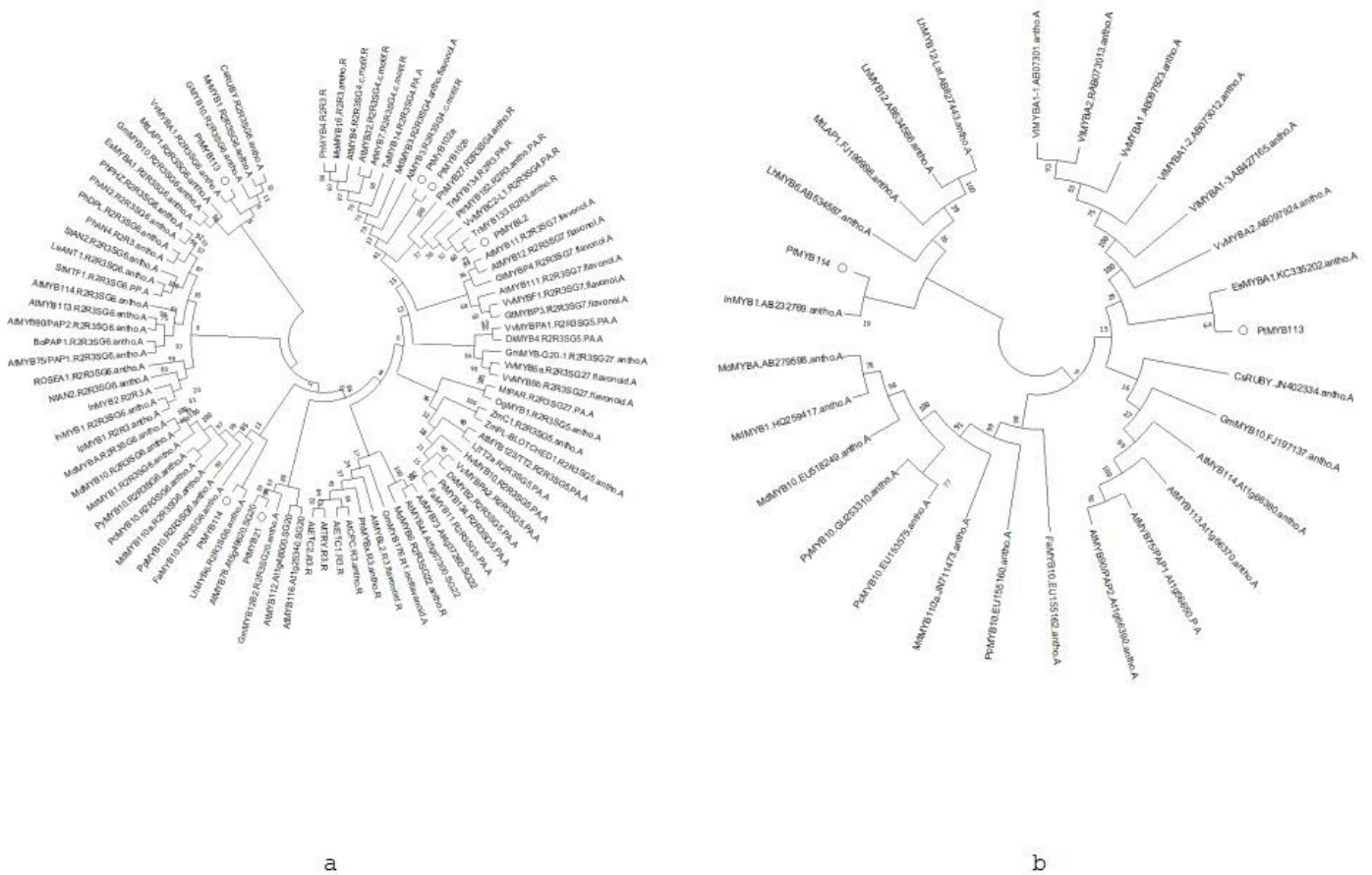


Figure 6

(a) Characterisation of MYB genes (*PtMYB113*, *PtMYB114*, *PtMYB21*, *PtMYB102a*, *PtMYB102b* and *PtMYBL2*, marked with circles) near the pigmentation loci (Table S16) with designated R2R3-MYB subgroups as previously reported using NJ method (bootstrap values with 5000 replicates are shown as percentage next to the branches) in MEGA X. (b) Characterisation of *PtMYB113* and *PtMYB114* with designated R2R3-MYB Subgroup 6 anthocyanin activators. The evolutionary distances were computed using the Poisson correction method. The subgroups are designated as previously reported⁵⁴⁻⁵⁷. The species and accession numbers of labelled MYB proteins are listed in Table S17 and S18.

This is a list of supplementary files associated with this preprint. Click to download.

- [WBSuppFigsfinal.docx](#)
- [WBSuppTablefinal.xlsx](#)