Transcriptome sequencing and marker development in winged bean (*Psophocarpus tetragonolobus*; Leguminosae)

Mohammad Vatanparast[1], Prateek Shetty[2], Ratan Chopra[3], Jeff J. Doyle[4], N. Sathyanarayana[5*] and Ashley N. Egan[1*]

[1]US National Herbarium (US), Department of Botany, Smithsonian Institution-NMNH, 10th and Constitution Ave, Washington DC, 20013, USA.

[2]Department of Plant Biology, Michigan State University, 612 Wilson Road, Room 166, East Lansing, MI, 48824, USA.

[3]United States Department of Agriculture, Agriculture Research Service, 3810 4th St., Lubbock, TX, 79415, USA.

[4]Section of Plant Breeding & Genetics, School of Integrative Plant Science, Cornell University, 412 Mann Library, Ithaca, NY, 14853, USA.

[5]Department of Botany, Sikkim University, 5th Mile, Tadong, Gangtok, Sikkim, 737102, India.

Email Addresses:

Mohammad Vatanparast: Vatanparastm@si.edu

Prateek Shetty: prateekshettys@gmail.com

Ratan Chopra: Ratan.Chopra@ARS.USDA.GOV

Jeff J. Doyle: jjd5@cornell.edu

N. Sathyanarayana: nsathyanarayana@cus.ac.in

Ashley N. Egan: egana@si.edu

*Corresponding authors:

Ashley N. Egan (primary)

N. Sathyanarayana (secondary)

Winged bean, *Psophocarpus tetragonolobus* (L.) DC, is similar to soybean in yield and nutritional value but more viable in tropical conditions. Here, we strengthen genetic resources for this orphan crop by producing a *de novo* transcriptome assembly and annotation of two Sri Lankan accessions (denoted herein as CPP34 [PI 491423] and CPP37 [PI 639033]), developing simple sequence repeat (SSR) markers, and identifying single nucleotide polymorphisms (SNPs) between geographically separated genotypes. A combined assembly based on 804,757 reads from two accessions produced 16,115 contigs with an N50 of 889 bp, over 90% of which has significant sequence similarity to other legumes. Combining contigs with singletons produced 97,241 transcripts. We identified 12,956 SSRs, including 2,594 repeats for which primers were designed and 5,190 high-confidence SNPs between Sri Lankan and Nigerian genotypes. The transcriptomic data sets generated here provide new resources for gene discovery and marker development in this orphan crop, and will be vital for future plant breeding efforts. We also analyzed the soybean trypsin inhibitor (STI) gene family, important plant defense genes, in the context of related legumes and found evidence for radiation of the Kunitz trypsin inhibitor (KTI) gene family within winged bean.

Winged bean (*Psophocarpus tetragonolobus* (L.) DC) is a promising legume crop of the world's tropical regions. It is predominantly self-pollinated and possesses a twining habit, tuberous roots, longitudinally winged pods, and both annual and perennial growth forms [1]. The genus *Psophocarpus* Neck. ex DC comprises 10 species. Excluding cultivated winged bean, all other species are wild and native to Africa, Madagascar and the Mascarene Islands in the Indian Ocean [2]. Winged bean is speculated to have originated from the progenitor species *P. grandiflorus* R. Wilczek and is now cultivated extensively in Papua New Guinea and Southeast Asia, and to a lesser extent in Africa [1,2]. Winged bean has a diploid genome (2n=2x=18) [3] and an estimated genome size of 1.22 Gbp/C (A.N. Egan, unpublished data).

Every part of the winged bean is edible, earning it the distinction of "*Supermarket on a stalk*" [4]. The exceptional nutritional quality of this plant, and the fact that it provides suitable human food sources at all stages of its life cycle, makes it a promising candidate for increased, widespread use in protein deficient tropical areas of the world. The young pods contain 2.4 grams (g) protein per 100g of edible portion; the dried tubers and seeds contain 8-20% and 34% protein, respectively, as well as a high oil contents (18%) - traits which have earned it the name "*soybean of the tropics*" [5]. If both seed and tuber yields are combined, winged bean can outperform many other legume crops that are conventionally grown in the tropics and thus offers a cheap nutritional food source. Consequently, it is projected as a promising alternative to soybean in areas where soybean cultivation success is marginal.

Since the 1975 publication by the US National Academy of Sciences of *The Winged Bean: A High Protein Crop for the Tropics* [6], considerable effort has been focused on studying the nutritional quality and climatic and ecological tolerances of the plant [7,8]. Winged bean reportedly possesses anti-nutritional factors such as phytoglutenins, cyanogenic glycosides,

tannins, lectins, flatulence factors, and saponins [9]. However, processing using moist heat or soaking has been shown to safely eliminate these substances . Research efforts concerning such anti-nutritional components have yielded significant knowledge concerning trypsin, a serine protease that acts to hydrolyze proteins as part of vertebrate digestion, and trypsin inhibitors, proteins that stop the action of trypsin, thereby interfering with digestion. It has been suggested that trypsin inhibitors play a role in protecting plant tissues against the action of bacterial proteases at the colonization site of pathogenic bacteria [10]. In addition, studies show involvement in defense against insects that suck the phloem sap and against bacteria that invade upon wounding [11]. In biomedical research, these modes of action have made trypsin and trypsin inhibitors vital components of molecular cell research where they are widely used in cell culture to detach cells from tissue culture plates. Since their first discovery in soybeans in 1945 [12], other Kunitz-type trypsin inhibitors have been discovered and characterized from winged bean [13,14], predominantly from seeds.

It is hard to find another high rainfall-adapted tropical legume with as many desirable characteristics as winged bean[1]. However, much needs to be done in terms of breeding efforts, especially to develop self-supporting, determinate cultivars bearing large numbers of relatively small pods having nutritious seeds and tubers, and cultivars resistant to biotic and abiotic stresses. Considerable variability for growth vigor and quantitative characters such as protein and oil content as well as photoperiodic responses has been recorded [15]. Several beneficial mutants were recovered during the 1970s and ‘80s through mutation breeding experiments [16]. However, a recent study using inter-simple sequence repeat (ISSR) markers reported low genetic diversity among the winged bean germplasm collected from different parts of the world [17]. With the advent of genomic tools such as molecular markers, genetic maps etc., the genetic improvement of underutilized crops has been greatly facilitated,

enabling the development of improved genotypes or varieties with enhanced trait values [18]. In the case of winged bean, studies on genomic resource development for enabling basic and applied research on genetics, evolution, ecology and molecular breeding programs are lacking, yet the advent of genomic technologies provides significant prospects for improvement [19]. Transcriptome sequencing is cost-effective and a valuable method for efficient and rapid identification of molecular markers in resource poor plant species [20].

The present study was undertaken with the following objectives: (a) to generate a set of expressed sequence tag (EST) resources through whole transcriptome analysis based on Roche 454-based transcriptomes for two winged bean accessions from Sri Lanka; (b) to develop a *de novo* assembly for these transcriptomes; (c) to annotate the transcriptome information; and (d) to discover microsatellite markers for future genetic studies. We also compared Sri Lankan accessions to a Nigerian winged bean transcriptome previously sequenced on the Illumina platform (e) to identify Single Nucleotide Polymorphisms (SNPs) evident between the geographically separated genotypes and (f) to present an analysis of the Kunitz trypsin inhibitor gene family in the context of related legumes.

Results

Sequencing and De novo assembly of winged bean transcriptomes

Pyrosequencing of two Sri Lankan accessions produced comparable sequence output, where genotype CPP34 produced a total of 369,820 single-end reads comprising 136,943,216 bp with an average read length of 574 bp and genotype CPP37 produced a total of 334,639 single-end reads comprising 92,126,948 bp with an average read length of 565 bp (Table 1).

Using read count as a proxy, the depth of sequencing across our contigs was similar for the independent *de novo* assemblies, ranging from one to 4,953 reads, with an average read depth of 25 reads per contig for CPP34 and ranging from one to 3,972 reads with an average read depth of 30 reads per contig for CPP37. Comparison of transcripts from the CPP34 and CPP37 independent assemblies (Supplementary file 1, inclusive of Tables S1-S3 and Figs. S1, online) found fewer than 200 high-confidence SNPs between them (data not shown), equating to approximately one SNP every 150,000 bp. Therefore, reads from the independently sequenced accessions were combined and co-assembled. For the combined assembly (CPP34-7), this translated to 704,459 reads comprising 229,070,164 bases from both accessions (Table 1). Because 454 pyrosequencing produces comparatively long reads (300-800 bp long), unassembled reads, here notated as singletons post-assembly, may potentially represent full-length mRNA transcripts. In order to not lose potential information, singletons of the CPP34-7 were extracted and appended to the final assembly of CPP34-7 and used in the GO and SNP analyses.

Functional annotation & legume sequence similarity

For the GO analysis, the combined assembly of CPP34-7 was used with inclusion of singletons (16,115 contigs plus 81,126 singletons, table 1). Using a total of 97,241 transcripts, TransDecoder could track 33,038 transcripts against BLAST and Pfam databases. Of these 33,038 transcripts, BLAST searches against NCBI's nr database retrieved 32,993 transcripts with hits (see Supplementary file 2 online), discarding 45 transcripts that had zero hits in NCBI. Therefore, 64,248 (66%) of our original 97,241 transcripts did not hit any known gene or DNA region in NCBI and Pfam databases, of which 62,783 were singletons. Thus, 79% of singletons were discarded in the BLAST searching steps due to a lack of

annotation. Of the 32,993 transcripts with BLAST hits, the GO analysis determined GO ID and enzyme code (EC) assignments for 16,561 (50.1%) with full or partial annotations (Fig. 1 in text, and see Supplementary file 2 online). Of the 16,561 annotated transcripts, 5,053 have predicted functions (EC codes). Overall, 2,829 transcripts were not functionally annotated by Blast2GO (zero hits) of which 1,932 (68%) corresponded to singletons. Participation of genes in a particular biological process and molecular function are shown in figure 2. Several transcripts were assigned to more than one GO term; therefore, the total number of GO terms obtained for our dataset was higher than the total number of transcripts. In total, 47,178 GO terms were retrieved, with 46.2%, 37% and 16.8%, corresponding to the MF, BP, and CC categories, respectively. In the MF category, nucleotide binding (number of sequences = 3,413), kinase activity (1,474) and DNA binding (1,200) had the highest number of assigned sequences. In the BP category, cellular protein modification (1,953), carbohydrate metabolic processes (1,080) and transport (908) were the majority and in the case of CC, genes involved in the plastid (319), cytoskeleton (288) and ribosome (281) activities were highly represented (Fig. 2).

A comparison of our assembled contigs against other legume NCBI protein sequence databases from chickpea, pigeon pea, soybean, common bean, *Medicago truncatula,* and *Lotus japonicus* using the BLASTX program from NCBI showed that 15,558 of 16,115 (96.5%) contigs from the CPP34-7 assembly had significant sequence similarity to sequences in one or more legume protein databases. About 90.5% of the 16,115 contigs had ≥ 80 % sequence identity (Fig. 3). The majority of the contigs (57.3%) were most similar to *Glycine max* (Fig. 4), a finding that, at first glance, seems to contradict that expected based on evolutionary relationships of legume lineages, but is likely due to the relative over-representation of genes within the soybean genome due to i) recent whole genome

duplication and ii) a much higher level and standard of annotation and gene discovery relative to other legume genomes. Differences in evolutionary rate across lineages may also impact this outcome. In relation to *Phaseolus vulgaris*, it is known that *Phaseolus* has a higher mutation rate than *Glycine* and related lineages, [21,22] which could increase the divergence, and thereby decrease the best-BLAST hits, of *Psophocarpus* against *Phaseolus* relative to *Glycine*. However, this explanation is invoked with caution given that it assumes similar relative rates between *Glycine* and *Psophocarpus*, information that is beyond the scope of this project.

Identification of transcription factors

In the overall GO analysis, 274 transcripts were annotated as transcription factors (Fig. 2). Of the 16,115 contigs, 176 putative winged bean transcription factor genes, distributed in at least ten families, were identified representing 1.1% of winged bean transcripts, which were assigned to different categories. Among these, basic leucine zipper (bZIP; 32), Teosinte-Branched1/Cycloidea/PCF (TCP; 19), MADS (17), MYB (11) and WRKY (9) were among the top five categories (Fig. 5.). The overall distribution of transcription factor encoding transcripts among the various known protein families is very similar to that of soybean and other legumes. However, almost all families showed minor species specific differences (for example, bZIP, MYB, WRKY etc.) with regard to TF gene families reported for *Lotus*, *Medicago* and *Glycine max.*

Identification of simple sequence repeats

The SSR analysis detected 10,984 perfect SSRs, 13 imperfect SSRs, and 1,959 compound

SSRs, for a total of 12,956 SSRs (see Supplementary file 3 online). Of the 10,984 primary SSRs, 57 were adenine (A: 30) or thymine (T: 27) monomers with at least 13 repeats. These were assumed to represent remnants of mRNA poly-A tails and were thus removed prior to primer prediction. No runs of 12 of more cytosine or guanine monomeric repeats were found. Nearly three-quarters of the remaining 10,927 perfect SSRs (7,933) were hexamers with only two repeats. Although these 12-mers may be useful as linkage markers, the low number of repeat units would likely take these out of the microsatellite category. The remaining 2,994 perfect SSRs were distributed across di-, tri-, tetra-, penta-, and hexamer SSRs (Fig. 6) and were used for primer creation. The majority (63%) of SSRs were detected in the tri- and hexamer categories (Fig 6a). In general, the number of SSRs detected in each size category decreased with increasing repeat number (Figure 6b-f). Primers were successfully created for 2,594 SSRs with product sizes ranging from 100 to 280 bp (see Supplementary file 4 online). Analysis of the primed SSRs showed bias towards certain di- and tri- repeat type motifs (Table 2).

Single nucleotide polymorphism discovery

GS Reference Mapper mapped 87.7% of reads from Chapman [23] onto the CPP34-7 reference 'genome' which consisted of the 97,241 transcripts. Of the 14,571,393 bp of mapped reads, we identified 113,757 SNPs with >95% confidence from the 454HCDiffs file (available upon request), suggesting a SNP frequency of one in 128 bp of coding regions. Interestingly, the majority of high-confidence SNPs were found within singletons (91,686; 80.6%) vs. contigs (22,071; 19.4%), a higher percentage than expected given that singletons make up 67.9% of total transcript length. As a conservative measure, we filtered SNPs based on allele frequency from >95% - 100% confidence levels and those having >20x coverage (Table 3), producing a

total of 13,091 SNPs distributed across 10,176 transcripts of which 5,196 (39.7%) were from contigs, representing 1 SNP every 1,113 bp. The subsequent increase in the proportion of SNPs within contigs is expected in this case given that more highly expressed genes will be more likely to be represented by >20x coverage and are most likely to assemble into contigs. Lastly, we removed all single nucleotide indels (7,665 of the 13,091) and those length variants that involved insertions or deletions of one or more nucleotides alone (i.e. those without point mutations involved in the length variants), resulting in a high-confidence set of 5,190 SNPs, 96% of which are one-to-one point mutations (see Supplemental file 5 online). Within the 5,190 SNPs, 151 unique SNP patterns were found and 211 (4%) SNPs were length variants involving one or more point mutation within the length variant. Of the 4,979 one-to-one polymorphisms, 3,433 (68.9%) were transitions and 1,546 (31.1%) were transversions, producing a transition:transversion ratio of 2.22.

Kunitz-type trypsin inhibitor gene family analysis

We identified 28 contigs from CPP34-7 and 20 contigs from the Chapman [23] transcriptome assembly corresponding to the KTI gene family within the *Psophocarpus* transcriptome (see Supplementary file 6 online). Due to the large number of paralogues in each species, there is no obvious criterion available for rooting this tree, so it was rooted with the largest clade of non-legume sequences, a clade of *Arabidopsis* sequences. Given this rooting, the Bayesian STI gene tree has a polytomous backbone and suggests six distinct subclades based on relatively high posterior probability and bootstrap support, here labeled as A-F (Fig. 7 in text; and see Supplementary file 7 online). The dominant feature of the tree (regardless of rooting) is a lack of clear orthologous relationships across taxa, with evidence of lineage-specific amplification of STI and KTI genes in each species. For example, subclade A comprises two

clades made up of only *Populus* sequences and an *Arabidopsis* STI member as well as a number of clades containing *Glycine, Phaseolus,* or *Medicago* gene family members, but with no *Psophocarpus* sequences included, whereas subclade C comprises *Populus* sequences only, illustrating a major intra-specific STI radiation (Fig. 7). The vast majority of *Psophocarpus* sequences cluster in clade F, along with many *Glycine* and a single *Phaseolus* sequence. Of the *Psophocarpus* sequences in subclade F, 15 contigs are paired between CPP34-7 and Chapman, forming sister groups that likely represent the same gene in each transcriptome, whereas 13 are unique (Fig. 7). Subclade F illustrates lineage-specific KTI expansion in both *Psophocarpus* and *Glycine*. All *Psophocarpus* sequences obtained from the Pfam or NCBI databases were nested within subclade F, where the majority of the Pfam sequences appeared as monophyletic clades with a contig each from CPP34-7 and Chapman nested therein (Fig. 7).

Discussion

The legumes represent the third largest family of the flowering plants, many of which are important sources of food, fodder, oil, fiber and medicines. However, with the exception of common pulse crops such as soybean, common bean, etc., a large number of legumes have remained underutilized due to poorly developed infrastructure, especially for genetic and genomic resources [24]. The advent of genomic technologies has brightened the prospects for such orphan crops [20,25,26], with recent research focusing on lentil (*Lens culinaris* Medik., [27]), chickpea [28], grass pea (*Lathyrus sativus* L., [29]), and a number of *Vigna* species ([30,31]), among others. Winged bean represents a promising alternative to protein-rich soybean for tropical regions of the world that house nearly 40% of the world population, of which nearly one third

is protein deficient, and many of whom are women and children [32]. Genomics assisted breeding and enabling biotechnologies that stem from it offer significant promise for targeted genetic improvement of nutritional and other quality traits in winged bean, thus aiding in the development of a low input, high quality legume-based protein diet for these parts of the world. Our combined assembly presents a genetic resource that can be mined for future genetic improvement and plant breeding initiatives. This paper reports development of genetic resources, including molecular markers, in winged bean, in addition to insights into the divergence of the Kunitz type trypsin inhibitors, which are important anti-nutritive agents in winged bean and other legumes.

In this study, we were able to annotate 32,993 (34%) transcripts from the winged bean combined assembly (CPP34-7; Fig. 2). Schmutz et al [21] annotated 27,197 protein-coding genes and 31,638 protein-coding transcripts from the *Phaseolus vulgaris* genome, suggesting that our annotated gene complement is reasonable, although it is likely that our transcriptomes do not provide a full gene complement due to low sequencing depth. Our level of unannotated transcripts is similar to results reported from other non-model organisms, including chickpea [33] and field pea [34]. These unidentified transcripts are likely due to: 1) correspondence to non-coding regions or pseudogenes, 2) short length of transcripts, or 3) novel coding genes that have yet to be described. Cellular, metabolic and transport processes were among the most highly represented groups in terms of GO analysis, as expected given that flower buds, young leaves and shoots are undergoing rapid growth and extensive metabolic activities.

Singletons (unassembled reads) in *de novo* transcriptome assemblies stem from such

293 phenomena as differences in assembly algorithms, sequencing errors, artifacts in cDNA
294 library construction, gene expression at low levels, or contamination from other organisms
295 such as bacteria or fungi [35]. Assessing the GC content of a transcriptome assembly can aid in
296 checking for possible contamination as different organisms have different genomic GC
297 content. We compared the GC content in our data against related legumes to check for
298 contamination and found no evidence of it (see Supplementary file 1 online). In the GO
299 analysis, ~80% of the singletons were discarded in the BLAST step, while the remaining
300 20% persisted, but only 10% proceeded through the GO annotation. Others have found
301 similar low levels of singleton annotation [36], yet, this low level of singleton annotation has
302 lead many to throw out unassembled reads. However, given the comparative length of 454
303 reads, these could easily represent full-length transcripts. Thus, we included singletons in the
304 GO and SNP analyses to evaluate their potential.

305

306 Because transcription factors play important roles in regulating plant functions, we paid
307 particular attention to their number and distribution within winged bean and in relation to
308 other legumes. Several TF gene families are preserved across different plant genera,
309 indicating conserved gene regulatory machinery in plants, as has been shown in legumes
310 previously [37]. In this study, we found that 2.4% of the total transcripts are putative
311 transcription factors according to GO analyses, a percentage much lower than the estimated
312 12% found in soybean (based on ~46,430 protein-coding genes) [38]. However, Libault et al [37]
313 estimated the number of TF-encoding genes across a number of species and found soybean to
314 have 3-4× the number of TFs relative to *Medicago truncatula* or *Lotus japonicus*, likely due
315 to recent whole genome duplication in soybean. If we compare the estimated number of TFs
316 in *M. truncatula* (1473) [37] against the number of putative protein-coding genes in *M.*
317 *truncatula* (~66,000; phytozome v11; https://phytozome.jgi.doe.gov/pz/portal.html), we

318 come out with a much more similar estimate (2.2%). However, this is likely an underestimate
319 and a shifting target as the annotation of *M. truncatula* is ongoing.
320
321 Our overall distribution of transcription factors in winged bean within the known TF families
322 is similar to that in soybean and other legume species, with bZIP, MYB, TCP, and WRKY
323 highly represented ([28,39]). The TF family most highly-represented in our data was the bZIP
324 family, which includes regulators of many central developmental and physiological processes
325 and abiotic and biotic stress responses [40,41]. In addition, elevated levels of expression were
326 also found for TCPs and MYB: TCPs have been characterized in other plant species for their
327 role in growth, development, and sex determination [42,43], whereas the MYB family has been
328 implicated in regulation of disease resistance and water loss regulation via stomatal
329 movement [44]. However, a significant portion of our transcripts comprised several smaller TF
330 families, here classified under the miscellaneous category for want of detailed
331 characterization. Also, we observed minor species-specific differences in the numbers and
332 proportion of our TFs relative to predicted TFs in *Lotus, Medicago* and *Glycine max* [37].
333 Further investigation is thus needed to elucidate the evolutionary and functional significance
334 of these events in winged bean.
335
336 Simple sequence repeat, or microsatellite, markers have long been used for genetic diversity
337 analyses and plant breeding efforts, largely due to their highly polymorphic, co-dominant
338 nature, prevalence throughout the genome, ease of use, and cost-effectiveness [45]. Because
339 they originate in coding regions, SSRs derived from genes have increased amplification
340 success in related species, are useful for assessing functional diversity and for marker-
341 assisted selection, and can act as anchor markers for evolutionary and comparative mapping
342 studies [46,47]. While some research has suggested that SSRs derived from coding regions are

less polymorphic than their anonymous counterparts [47], numerous population genetic, evolutionary, and plant breeding studies have found them to have adequate, if not higher, levels of polymorphism within legumes [48,49].

Within our data, we discovered nearly 5,000 perfect or compound genic-SSRs with three or more repeats. After filtering for perfect, simple SSRs, we discovered an unequal distribution across size-classes, with trinucleotide repeats making up the bulk (43%) of filtered SSRs. Given the coding nature of the transcriptome, this finding makes sense as proliferation of tri-nucleotide, in-frame repeats would be more tolerated [46]. The same trend has been noted in other plants, including for legumes *Medicago truncatula* [50] and peanut (*Arachis hypogaea* L.; [51]). Of the 2,594 SSRs for which primers were created, 1,928 (74.3%) were annotated in our Blast2GO analyses, 871 (45.2%) of which are putatively homologous to known proteins while 1,057 (54.8%) were similar to hypothetical, uncharacterized, unknown, or predicted proteins (mostly from *Phaseolus vulgaris* and *Glycine max* genome annotations).

Certain repeat motif types were more prevalent than others in our set of primed SSRs (Table 2), a not-uncommon finding that has been documented in legumes previously [52]. Zhang et al [53] first documented the bias of microsatellites to AG and AAG motifs in *Arabidopsis*, also noting differences of SSR distributions between 5' and 3' untranslated and coding regions, and correlation between trinucleotide repeat motifs and codon usage. In winged bean, the SSR repeat motif type $(AG/GA/TC/CT)_n$ represented the majority (77.7%) of all dinucleotide repeats, while motif types $(AT/TA)_n$ and $(AC/CA/GT/TG)_n$ comprised 13.7% and 8.6%, respectively. Our distribution and ranking of dinucleotide repeat motifs mirrors that in *Arabidopsis* (Table 2). The bias towards the repeat type containing AG and against that of GC has also been found in other plants, including *Phaseolus* [54], *Myrciaria dubia* (Kunth)

McVaugh [55] and across eukaryotes [56]. Past research has suggested that AG motifs are most prevalent in 5' untranslated regions [52,57] and possibly are involved in transcription and regulation [53]. As mentioned earlier, 25.7% of primed genic-SSR transcripts are unannotated, some of which may correspond to 5' untranslated regions where AG motif types are more prevalent. The frequency of trinucleotide repeat motif types was biased towards AAG in our set of primed SSRs, with this motif type comprising 29.1% of the 10 trinucleotide types, followed by the ATC motif type, comprising 13.9%. The ranking of these and other motifs closely resembles that of *Arabidopsis* (Table 2), with the first two most prevalent motifs the same [53].

SNPs provide another means of assessing genetic variation and, although less polymorphic than SSRs, are abundant and easily obtained via high-throughput sequencing. For example, Rajesh and Muehlbauer [58] estimated SNP frequency to be one in 66 bp in coding regions and one in 71 bp in genomic regions of chickpea [58]. In another study, Hyten et al [59] reported 7,000-25,000 predicted SNPs through deep resequencing of soybean by a whole genome sequence approach. In this study, we discovered more than 5,190 high-confidence SNPs between our Sri Lankan samples and the geographically separated Nigerian genotype [23]. SNP markers identified in this study can be used in quantitative trait loci (QTL) mapping, generating linkage maps, genotyping and breeding studies. Validation of SNPs determined herein is beyond the scope of this paper, nevertheless, this list presents a significant resource for future work in plant breeding and genetic diversity assessment [60] and marks the first SNP markers discovered to date in *Psophocarpus*.

Our high-confidence SNP set included 4,979 one-to-one SNPs (those without length variants and involving changes between a single nucleotide position), equating to a

transition:transversion (ts:tv) ratio of 2.22. This bias is commonly observed across SNPs throughout a genome, resulting from rampant methyl cytosine to uracil mutations [61]. Similar ratios were found across SNPs in other legumes [62,63]. In total, 44% of SNPs identified were found in singletons, a proportion not unexpected given that 68% of transcript read length is in singletons. But the very fact that alignable and putatively homologous singletons were found across the geographically separated and independently sequenced genotypes provides vindication for their inclusion in transcriptome characterizations, at least for 454 data. However, caution is warranted due to the 'singleness' of the unassembled read acting as a reference sequence.

Trypsin inhibitors play important roles in plant development and defense systems and have been studied from various aspects like biotic stress and wounding. These compounds inhibit activity of proteases and are induced by mechanical wounding in leaves, suggesting a strong role as anti-herbivory agents [64]. Trypsin inhibitors present in legumes include KTIs, the Bowman-Birk trypsin inhibitor, and Cowpea trypsin inhibitor. The KTIs were first discovered from soybean in 1945 [12]. Since then, a number of trypsin inhibitors have been discovered and characterized from winged bean [14,65], predominantly from seeds, where they are shown to act as insecticidal agents, preventing seed loss during development [66]. The soybean trypsin inhibitor (STI) gene superfamily has been well studied among *Populus* species, where it has been identified as a rapidly evolving gene family and shown to play multiple roles in anti-herbivory and other stress responses [67]. Philippe et al [64] suggest that the STI gene family has expanded due to repeated gene duplications within poplar relative to other plant species because poplars need strong anti-herbivory actions to maintain their long-lived life cycle. Our study also discovered several *Populus*-specific radiations (subclades A, C, & D; Figure 7).

In this study, we characterized 28 STI sequences from our CPP34-7 transcripts as well as 20 from the Nigerian winged bean transcriptome [23]. The vast majority of our STI sequences clustered with *Glycine* in subclade F (Fig. 7), which includes 28 of 32 overall, distinct *Psophocarpus* lineages, 15 of which are corroborated between the CPP34-7 and Chapman transcriptomes. Subclade F includes those proteins originally characterized as KTIs. Expansion of gene family members in *Psophocarpus* can be characterized as lineage (species)-specific or gene-specific (e.g., via tandem gene duplications). The lineage-specific radiations within *Psophocarpus* and *Glycine* may be inflated due to the presence of multiple alleles or alternatively spliced transcripts. Gene-specific amplification of STI family members in poplar is in part due to tandem duplication [64]. KTI genes (with highest sequence similarity to subclade F) are tandemly duplicated within soybean, with at least eight KTI loci linked within 68 kbp on chromosome 8 (between positions 44850000..44918000). Lineage-specific amplification of *Psophocarpus* KTI sequences is evident, and, given the expectation of conserved synteny between soybean and *Psophocarpus*, gene-specific amplification of *Psophocarpus* KTI sequences may be due to recent tandem duplications.

Besides the classically described KTI genes, several other prominent STI genes are present in our gene tree. Subclade B includes a single contig from our *Psophocarpus* transcriptome, with high sequence similarity to miraculin, a glycoprotein that strongly binds to human taste receptors in the presence of acidic compounds, modifying sour tastes into sweet ones [68]. Miraculin is classified into the STI family and encodes the Kunitz motif but differs from other STI or KTI family members in that it forms a homodimer instead of monomers [69]. Subclade D includes a single *Psophocarpus* contig that has high similarity to alpha-amylase/subtilisin inhibitor proteins known to inhibit the activity of insect a-amylase in

443 *Vigna* species, thus protecting against insect attack [70]. Subclade E comprises only legume STI
444 gene sequences, including a single paralog from *Psophocarpus* with high sequence similarity
445 to Kunitz-*type* trypsin inhibitor-*like* 2 proteins.
446
447 The unequal distribution of *Psophocarpus* STI sequences across the six subclades may be due
448 to tissue specificity, depth of transcriptome reads, amplification of certain gene subfamilies
449 or gene loss over time. As mentioned earlier, most of the winged bean KTIs currently known
450 were characterized from seeds, yet all of these are present in subclade F, in spite of the fact
451 that the CPP34-7 transcriptome did not include seed transcripts, but was sequenced from
452 young leaves, shoots, and buds. This subclade also includes a *Psophocarpus* nodulin
453 (Ptet_Q43325) expressed in nodules of winged bean, likely as a delayed response of the host
454 plant to *Rhizobium* infection [71]. Expression levels of STI genes in winged bean likely differ
455 across plant tissues, as demonstrated in poplar [64], and this may be one explanation for
456 unequal distribution of *Psophocarpus* sequences across the gene tree, although inclusion of
457 such tissue-specific genes as nodulins argues against that. Unfortunately, we cannot
458 determine tissue-specific expression of our winged bean STIs due to the fact that our
459 transcriptomes were sequenced from pooled tissue samples. But, given the roles of KTIs in
460 wound and herbivory defense, radiation of KTI genes would be evolutionarily beneficial to
461 large-leaved, highly nutritious plants such as winged bean. Deeper sequencing of the
462 transcriptomes across more tissue types would likely yield other STI gene family members in
463 *Psophocarpus* and provide a more holistic view of STI gene family evolution in the winged
464 bean.
465

466 **Materials and methods**

467

Plant material

Seeds of two winged bean (*Psophocarpus tetragonolobus*) genotypes were selected from the United States Department of Agriculture (USDA) Germplasm Resources Information Network (GRIN) seed bank. PI 639033 (CPP37) was field collected in 1999 while PI 491423 (CPP34) was donated in 1984, both from Sri Lanka. Seeds were grown to maturity in the greenhouse at Cornell University (Ithaca, NY, USA) for 3 years. Flowering and fruiting were induced by imposing a day length of less than 8 hours. For comparative purposes and to aid in the development of genetic resources for the winged bean, we compared our transcriptomes to an Illumina-based *P. tetragonolobus* transcriptome (SRR1772344) recently published and originally sourced from Nigeria [23].

RNA isolation and library preparation

Young leaves, young buds, and young shoots were collected from 3-year old plants into liquid nitrogen to preserve RNA. Total RNA was extracted from each tissue (leaves, shoots, and buds) separately using the Qiagen RNAeasy mini kit according to manufacturer instructions. The quality and quantity of each RNA tissue extract was assessed using a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). All RNA samples had RIN (RNA integrity number) greater than 9.0 and were used for the analysis. RNA concentration was also quantified using the nanodrop 2000c spectrophotometer (NanoDrop Technologies, Inc., Montchanin, DE, USA). Before cDNA library construction, RNA from tissues for each accession was combined in equal molar amounts so as to allow each tissue equal representation in the final library construct. One microgram (ug) of the pooled tissue total RNA extracts were used for subsequent cDNA library construction of each accession using

493 the Clontech SMARTer cDNA synthesis kit (Clontech Laboratories, Inc., Mountain View,
494 CA, USA) according to manufacturer's instructions but using a 3' SMART CDS Primer IIA
495 modified to 5' --
496 AAGCAGTGGTATCAACGCAGAGTACTTTTTTGTTTTTTTCTTTTTTTVN -- 3' which
497 was purchased from IDT (Integrated DNA Technologies, Inc., CA, USA). cDNA libraries
498 were then purified using the PureLink PCR purification kit (Life Technologies, (Invitrogen),
499 Carlsbad, CA, USA) with Buffer HC which removed all fragments less than 300 bp.
500
501 *Transcriptome sequencing*
502
503 Samples were sequenced using single-end 454 pyrosequencing on the Roche 454 Genome
504 Sequencer FLX (Titanium chemistry) at the Brigham Young University Sequencing Center
505 (Provo, UT, USA). Libraries were tagged with multiplex identifier (MID) barcodes to allow
506 multiplexing of four species together over one titer plate. After sequencing, MID adaptors
507 and primers were removed from reads during pre-processing. Preliminary visualization of
508 data was done in FASTQC v. 0.11.3 [72].
509
510 *De novo assembly*
511
512 For the CPP34 transcriptome we used the standard flowgram file (SFF) originally generated
513 by the 454 GS FLX sequencer. However, for CPP37, we started with fasta (FNA) and quality
514 value (QUAL) files. We converted the FNA and QUAL files for CPP37 into a single FASTQ
515 file using a python script. We used the FASTX-Toolkit
516 (http://hannonlab.cshl.edu/fastx_toolkit/index.html) to trim and clean the CPP37 reads: we
517 discarded sequences shorter than 50 bp (-l 50) using FASTX CLIPPER and setting of first

base of 15 (-f 15) and last base of 800 (-l 800) using FASTA TRIMMER. Finally, the FASTQ Quality Filter was used with minimum quality score of 20 (-q 20) and minimum percent of included bases of 80 (-p 80). In all steps we used the quality score ASCII offset command (-Q33) to denote 454 file format. The quality of output reads after cleaning steps was inspected using FASTQC software v. 0.11.3 [72].

To determine the extent of divergence between our two independently sequenced 454-based transcriptomes, CPP34 and CPP37, we initially assembled each transcriptome independently and explored several contemporary assembly strategies, including Trinity [73], Velvet [74], MIRA [75], and GS *De Novo* Assembler (aka Newbler, Roche, USA) (see methods and results in Supplementary file 1 online). Our initial findings found fewer than 200 high confidence SNPs between assemblies of CPP34 and CPP37 (SNPs were detected between CPP34 and CPP37 the same way they were assessed between Sri Lankan and Nigerian accessions; see SNP methods below), suggesting a high degree of similarity between these two Sri Lankan accessions. Therefore, for subsequent assemblies and analyses we combined the reads from our two Sri Lankan accessions and produced a single assembly, notated as CPP34-7. Ultimately, we chose to use GS *De Novo* Assembler over the other programs because of the reliable output, comparable contig length, the fact that it considers alternative splicing [76], and that it is a program specifically designed for 454 data. Comparisons for several programs in the past showed that it performed best among other de novo assemblers for 454 transcriptome data [77]. Raw reads from CPP34 and CPP37 were combined by co-assembly within GS *De Novo* Assembler v. 2.9 with default settings using a minimum read length of 20, minimum overlap length of 40, minimum overlap identity of 90%, and Isotig threshold of 100.

Functional annotation

Prior to functional annotation, we identified candidate coding regions and filtered sequences based on a minimum amino acid length of 100 using the TransDecoder program (https://transdecoder.github.io) v. 2.0.1 applied to CPP34-7 contigs plus singletons, using the TransDecoder.LongOrfs command. To identify open reading frames (ORFs) with homology to known proteins and to maximize sensitivity for capturing ORFs that may have functional significance, Blastp and Pfam searches were conducted. The Blastp search was done using the Swissprot database with the E-value of 1E-5 and Pfam search was done using HMMER (http://hmmer.janelia.org), a biosequence analysis program using profile hidden Markov models and the Pfam database (http://pfam.xfam.org). Output files that were generated from the Blastp and Pfam database searches were leveraged by TransDecoder to ensure that peptides with BLAST or domain hits were retained in the set of reported likely coding regions by running the TransDecoder.Predict command. Finally, output of the TransDecoder analysis was used as input for functional annotation using the Blast2GO program [78]. First, we conducted a BLAST search on the output from Transdecoder against the NCBI's nonredundant (nr) database with the E-value of 1E-5 on the Smithsonian Hydra clusters. These BLAST results were then used as input to Blast2GO to assign Gene Ontology (GO) terms to our DNA regions, including biological processes (BP), molecular functions (MF), and cellular components (CC).

Sequence similarity with other legumes

To compare our complement of genes characterized from our winged bean transcriptome assembly against typical gene assemblies in other legumes, legume species' protein sequences ((*Medicago truncatula* Gaertn., *Glycine max* (L.) Merr. (soybean), *Lotus japonicus*

(Regel) K.Larsen, *Phaseolus vulgaris* L. (common bean), *Cicer arietinum* L. (chickpea), and *Cajanus cajan* (L.) Millsp. (pigeonpea)) along with *Populus trichocarpa* Torr. & A.Gray ex Hook. and *Arabidopsis thaliana* (L.) Heynh. protein sequences were downloaded from NCBI. BLASTX searches were performed against the CPP34-7 contigs with E-value of 1E-4, and the top hit for each contig was used for further analysis.

Transcription factor identification

CPP34-7 transcripts were translated to protein sequences for prediction of transcription factors in the assembly. Translated protein sequences were subjected to prediction using PlantTFDB (http://planttfdb.cbi.pku.edu.cn/), with further linking the prediction to best hits in Arabidopsis. Since not all transcription factors (TFs) could be predicted in the CPP34-7 assembly, we utilized the annotation results of BLASTX searches against legume databases. All identified and predicted transcription factors were further classified into categories.

Simple sequence repeat identification

To retrieve simple sequence repeat (SSRs; microsatellite) markers and also to design primers, SSR Locator v.1 [79] program was used to detect SSRs across contigs from CPP34-7. A SSR site was defined as a monomer occurring at least 12x with a dimer at least 6x, trimers at least 4x, tetra- and pentamers at least 3x, and hexa- to decamers occurring at least 2x. The space between compound SSRs was set to 100 bp and the space between imperfect SSRs to 5 bp. Primers were produced and reported for primary SSRs only.

Single nucleotide polymorphism identification

To identify SNPs between *Psophocarpus* transcriptomes, we used the transcripts (contigs + singletons) from our combined assembly CPP34-7 as a reference 'genome'. We extracted singletons from the original reads and concatenated them with the contigs produced by our CPP34-7 assembly. We queried Chapman's Nigerian, Illumina-based transcriptome [23] against our CPP34-7 reference 'genome' using the GUI interface of GS Reference Mapper v. 2.9 (454 Life Sciences, Roche, USA) under default settings. We used only high-confidence variants to the reference sequence (454HCDiffs) and further filtered these to those having 20x or greater coverage. Lastly, to ensure the highest SNP call quality for use in future research, we followed the method of Schmutz *et al* [21] and discarded any SNPs where i) the reference or variant involved one or more N's; and or ii) the reference or variant allele was a single nucleotide insertion or deletion or did not involve a point mutation in the length variant.

Kunitz-type trypsin inhibitor gene family analysis

To reconstruct a gene tree of the STI superfamily, particularly the KTI gene families, and to understand the evolutionary diversification of this gene superfamily in *Psophocarpus* related to other legumes, we obtained available STI sequences for selected legumes and other angiosperms from the Pfam database (http://pfam.xfam.org). In total, 214 accessions were retrieved across *Arabidopsis*, *Populus*, *Medicago*, *Phaseolus*, *Glycine,* and *Psophocarpus*. We downloaded a reference alignment from the Pfam database and used this alignment as a scaffold upon which to align contigs garnered from our transcriptome (see Supplementary file 6 online). We extracted putative *Psophocarpus* STI regions from our transcriptome (CPP34-7) and Chapman's [23] after blasting against a local BLAST database [80] based on

618 available STI gene sequences obtained from the Pfam database.

619

620 We combined our extracted contigs with the Pfam STI sequences, converted the open reading
621 frames to amino acid sequences, and aligned in MAFFT v. 7. 245 [81]. Phylogenetic analysis
622 was conducted in RAxML v. 8.1.24 [82] with 1000 rapid bootstrap inferences and using the
623 best substitution model (LG+G) as determined by Prottest v. 3.4 [83]. Additionally, Bayesian
624 analysis was conducted using MrBayes v. 3.2.6 [84] under the JTT amino acid model
625 "aamodelpr = fixed(jones)" and gamma rates. Two independent Markov Chain Monte Carlo
626 (MCMC) analyses with 12 simultaneous chains and 25 million generations were run for each
627 analysis. Trees were sampled every 10,000 generations and the first 25% of trees were
628 discarded as burn-in. The convergence of MCMC chains was confirmed with Tracer version
629 1.6 [85]. All runs and parameters were checked to ensure proper mixing as evidenced by
630 effective sample size (ESS) scores being above 200 and the standard deviation of the split
631 frequencies having dropped below 0.01 [84].

632

633 **Acknowledgements**

634

635 Thanks go to Sue Sherman-Broyles and Jane L. Doyle for help in the greenhouse. Thanks to
636 Matthew Kweskin and Vanessa Gonzalez from the Smithsonian National Museum of Natural
637 History for help with data analysis. Computations were completed on the Smithsonian
638 Institution High Performance Cluster (SI/HPC). This research was supported by grants from
639 the US National Science Foundation to JJD (DEB-0948800) and ANE (DEB-1352217). We
640 acknowledge the support of Sir M Visvesvaraya Institute of Technology (Sir MVIT),
641 Bangalore and Sikkim University, Gangtok for facilities.

642

Authors' contributions

All authors contributed to various aspects of this work (ordered by degree of contribution): conceived the study (ANE, NS), aided in study design (ANE, NS, MV, PS), obtained funds for the research (JJD, ANE), coordinated activities (ANE, NS), obtained and grew plants from seed (ANE), extracted RNA and prepared cDNA libraries for 454 sequencing (ANE), conducted bioinformatic analyses (MV, PS, SS, ANE, RC), and contributed to preparation of the manuscript (MV, ANE, NS, PS, JJD, RC). All authors reviewed the manuscript.

Availability of Supporting Data

Transcriptome datasets supporting the conclusions of this article are available in the NCBI SRA repository under the accession number SRP067662 (raw 454 reads). In addition, several large datasets stemming from analyses of these data are available in Supplementary files online.

Competing Financial Interests

The authors declare that they have no competing financial interests.

References

1 Hymowitz, T. & Boyd, J. Origin, ethnobotany and agricultural potential of the winged bean—*Psophocarpus tetragonolobus*. *Econ. Bot.* **31**, 180-188 (1977).

2 Klu, G. Induced mutations for accelerated domestication - a case study of winged bean (*Psophocarpus tetragonolobus* (L.) DC). *West African Journal of Applied Ecology* **1**, 47-52 (2000).

3 Harder, D. K. Chromosome counts in *Psophocarpus*. *Kew Bulletin* **47**, 529-534 (1992).

4 Smith, O., Ilori, J. & Onesirosan, P. The proximate composition and nutritive value of the winged bean *Psophocarpus tetragonolobus* (L.) DC for broilers. *Anim. Feed Sci. Technol.* **11**, 231-237 (1984).

5 Amoo, I., Adebayo, O. & Oyeleye, A. Chemical evaluation of winged beans (*Psophocarpus tetragonolobus*), Pitanga cherries (*Eugenia uniflora*) and orchid fruit (Orchid fruit myristica). *African Journal of Food, Agriculture, Nutrition and Development* **6**, 1-12 (2006).

6 Bean, N. R. C.-P. o. t. W. *The winged bean: a high-protein crop for the tropics*. (National Academies, 1975).

7 Harder, D., Lolema, O. P. M. & Tshisand, M. Uses, nutritional composition, and ecogeography of four species of *Psophocarpus* (Fabaceae, Phaseoleae) in Zaire. *Econ. Bot.* **44**, 391-409 (1990).

8 Ruegg, J. Effects of temperature and water stress on the growth and yield of winged bean (*Psophocarpus tetragonolobus* (L.) DC.). *J. Hortic. Sci.* **56**, 331-338 (1981).

9 Tan, N. H., Rahim, Z. H., 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).

10 Ryan, C. A. Protease inhibitors in plants: genes for improving defenses against insects and pathogens. *Annu. Rev. Phytopathol.* **28**, 425-449 (1990).

11 Habu, Y., Fukushima, H., Sakata, Y., Abe, H. & Funada, R. A gene encoding a major Kunitz proteinase inhibitor of storage organs of winged bean is also expressed in the phloem of stems. *Plant Mol. Biol.* **32**, 1209-1213 (1996).

12 Kunitz, M. Crystallization of a trypsin inhibitor from soybean. *Science* **101**, 668-669 (1945).

13 Peyachoknagul, S. *et al.* Sequence and expression of the mRNA encoding the chymotrypsin inhibitor in winged bean (*Psophocarpus tetragonolobus* (L.) DC.). *Plant Mol. Biol.* **12**, 51-58 (1989).

14 Giri, A. P. *et al.* Identification of potent inhibitors of *Helicoverpa armigera* gut proteinases from winged bean seeds. *Phytochemistry* **63**, 523-532 (2003).

15 Harding, J., Martin, F. & Kleiman, R. Seed protein and oil yields of the winged bean *Psophocarpus tetragonolobus* in Puerto Rico. *Tropical Agriculture (Trinidad and Tobago)* **55**, 307 (1978).

16 Klu, G., Quaynor-Addy, M., Dinku, E. & Dikumwin, E. in *Joint FAO/IAEA Division of Nuclear Techniques in Food and Agriculture, Vienna (Austria)* 15-16 (International Atomic Energy Agency, 1989).

17 Chen, D. *et al.* Genetic diversity evaluation of winged bean (*Psophocarpus tetragonolobus* (L.) DC.) using inter-simple sequence repeat (ISSR). *Genet. Resour. Crop Evol.* **62**, 823-828 (2015).

18 Sharma, K. K., Dumbala, S. R. & Bhatnagar-Mathur, P. in *Plant Biotechnol.* 193-207 (Springer, 2014).

19 Egan, A. N., Schlueter, J. & Spooner, D. M. Applications of next-generation sequencing in plant biology. *Am. J. Bot.* **99**, 175-185 (2012).

20 Varshney, R. K., Close, T. J., Singh, N. K., Hoisington, D. A. & Cook, D. R. Orphan legume crops enter the genomics era! *Curr. Opin. Plant Biol.* **12**, 202-210 (2009).

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 Lavin, M., Herendeen, P. S. & Wojciechowski, M. F. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. *Syst. Biol.* **54**, 575-594 (2005).

23 Chapman, M. A. Transcriptome sequencing and marker development for four underutilized legumes. *Applications in Plant Sciences* **3**, apps. 1400111, doi:doi:10.3732/apps.1400111. (2015).

24 Nelson, R. J., Naylor, R. L. & Jahn, M. M. The role of genomics research in improvement of "orphan" crops. *Crop Sci.* **44**, 1901-1904 (2004).

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

26 Graham, I. in *Successful Agricultural Innovation in Emerging Economies: New Genetic Technologies for Global Food Production* (eds D.J. Bennet & R.C. Jennings) 95-106 (2013).

27 Sharpe, A. G. *et al.* Ancient orphan crop joins modern era: gene-based SNP discovery and mapping in lentil. *BMC Genomics* **14**, 192, doi:DOI: 10.1186/1471-2164-14-192 (2013).

28 Hiremath, P. J. *et al.* Large- scale transcriptome analysis in chickpea (*Cicer arietinum* L.), an orphan legume crop of the semi- arid tropics of Asia and Africa. *Plant Biotechnol. J.* **9**, 922-931 (2011).

29 Yang, T. *et al.* Large-scale microsatellite development in grasspea (*Lathyrus sativus* L.), an orphan legume of the arid areas. *BMC Plant Biol.* **14**, 65, doi:DOI: 10.1186/1471-2229-14-65 (2014).

30 Chen, H. *et al.* Transcriptome sequencing of mung bean (*Vigna radiate* L.) genes and the identification of EST-SSR markers. *PloS One* **10**, e0120273, doi:DOI: 10.1371/journal.pone.0120273 (2015).

31 Souframanien, J. & Reddy, K. S. De novo assembly, characterization of immature seed transcriptome and development of genic-SSR markers in black gram [*Vigna mungo* (L.) Hepper]. *PloS One* **10**, e0128748 (2015).

32 FAO, I. *WFP. 2015*. (2015).

33 Kudapa, H. *et al.* Comprehensive transcriptome assembly of chickpea (*Cicer arietinum* L.) using Sanger and next generation sequencing platforms: development and applications. *PLos One* **9**, e86039 (2014).

34 Sudheesh, S. *et al.* De novo assembly and characterisation of the field pea transcriptome using RNA-Seq. *BMC Genomics* **16**, 611 (2015).

35 Pop, M. & Salzberg, S. L. Bioinformatics challenges of new sequencing technology. *Trends Genet.* **24**, 142-149 (2008).

36 Meyer, E. *et al.* Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics* **10**, 219 (2009).

37 Libault, M. *et al.* Legume transcription factor genes: what makes legumes so special? *Plant Physiol.* **151**, 991-1001 (2009).

38 Schmutz, J. *et al.* Genome sequence of the paleopolyploid soybean. *Nature* **463**, 178-183, doi:10.1038/nature08670 (2010).

39 Wang, Z. *et al.* SoyDB: a knowledge database of soybean transcription factors. *BMC Plant Biol.* **10**, 14 (2010).

40 Guimarães, P. M. *et al.* Global transcriptome analysis of two wild relatives of peanut under drought and fungi infection. *BMC Genomics* **13**, 387 (2012).

764 41 Llorca, C. M., Potschin, M. & Zentgraf, U. bZIPs and WRKYs: two large
765 transcription factor families executing two different functional strategies. *Front. Plant*
766 *Sci.* **5**, 10-3389 (2014).
767 42 Martín-Trillo, M. & Cubas, P. TCP genes: a family snapshot ten years later. *Trends*
768 *Plant Sci.* **15**, 31-39 (2010).
769 43 Ma, J. *et al.* Genome-wide identification and expression analysis of TCP transcription
770 factors in *Gossypium raimondii*. *Scientific Reports* **4**, 6645 (2014).
771 44 Yanhui, C. *et al.* The MYB transcription factor superfamily of Arabidopsis:
772 expression analysis and phylogenetic comparison with the rice MYB family. *Plant*
773 *Mol. Biol.* **60**, 107-124 (2006).
774 45 Wang, M. L., Barkley, N. A. & Jenkins, T. M. Microsatellite markers in plants and
775 insects. Part I: Applications of biotechnology. *Genes, Genomes and Genomics* **3**, 54-
776 67 (2009).
777 46 Varshney, R. K., Graner, A. & Sorrells, M. E. Genic microsatellite markers in plants:
778 features and applications. *Trends Biotechnol.* **23**, 48-55 (2005).
779 47 Ellis, J. & Burke, J. EST-SSRs as a resource for population genetic analyses. *Heredity*
780 **99**, 125-132 (2007).
781 48 Chankaew, S. *et al.* Detection of genome donor species of neglected tetraploid crop
782 *Vigna reflexo-pilosa* (creole bean), and genetic structure of diploid species based on
783 newly developed EST-SSR markers from azuki bean (*Vigna angularis*). *PLoS One* **9**,
784 e104990, doi:doi:10.1371/journal.pone.0104990 (2014).
785 49 Sun, X. *et al.* SSR genetic linkage map construction of pea (*Pisum sativum* L.) based
786 on Chinese native varieties. *The Crop Journal* **2**, 170-174 (2014).
787 50 Eujayl, I. *et al. Medicago truncatula* EST-SSRs reveal cross-species genetic markers
788 for *Medicago* spp. *Theor. Appl. Genet.* **108**, 414-422 (2004).

51 Bosamia, T. C., Mishra, G. P., Thankappan, R. & Dobaria, J. R. Novel and stress relevant EST derived SSR markers developed and validated in peanut. *PloS One* **10**, e0129127 (2015).

52 Mun, J.-H. *et al.* Distribution of microsatellites in the genome of *Medicago truncatula*: a resource of genetic markers that integrate genetic and physical maps. *Genetics* **172**, 2541-2555 (2006).

53 Zhang, L. *et al.* Preference of simple sequence repeats in coding and non-coding regions of *Arabidopsis thaliana*. *Bioinformatics* **20**, 1081-1086 (2004).

54 Blair, M. W., Torres, M. M., Giraldo, M. C. & Pedraza, F. Development and diversity of Andean-derived, gene-based microsatellites for common bean (*Phaseolus vulgaris* L.). *BMC Plant Biol.* **9**, 100 (2009).

55 Castro, J. C. *et al.* De novo assembly and functional annotation of *Myrciaria dubia* fruit transcriptome reveals multiple metabolic pathways for L-ascorbic acid biosynthesis. *BMC Genomics* **16**, 997 (2015).

56 Tóth, G., Gáspári, Z. & Jurka, J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* **10**, 967-981 (2000).

57 Morgante, M., Hanafey, M. & Powell, W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* **30**, 194-200 (2002).

58 Rajesh, P. & Muehlbauer, F. J. Discovery and detection of single nucleotide polymorphism (SNP) in coding and genomic sequences in chickpea (*Cicer arietinum* L.). *Euphytica* **162**, 291-300 (2008).

59 Hyten, D. L. *et al.* High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics* **11**, 38 (2010).

60 Mammadov, J., Aggarwal, R., Buyyarapu, R. & Kumpatla, S. SNP markers and their impact on plant breeding. *International Journal of Plant Genomics* **2012**, Article ID 728398, doi:doi:10.1155/2012/728398 (2012).

61 Coulondre, C., Miller, J. H., Farabaugh, P. J. & Gilbert, W. Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* **274**, 775-780 (1978).

62 Agarwal, G. *et al.* Comparative analysis of kabuli chickpea transcriptome with desi and wild chickpea provides a rich resource for development of functional markers. *PLoS One* **7**, e52443 (2012).

63 Leonforte, A. *et al.* SNP marker discovery, linkage map construction and identification of QTLs for enhanced salinity tolerance in field pea (*Pisum sativum* L.). *BMC Plant Biol.* **13**, 161 (2013).

64 Philippe, R. N., Ralph, S. G., Külheim, C., Jancsik, S. I. & Bohlmann, J. Poplar defense against insects: genome analysis, full- length cDNA cloning, and transcriptome and protein analysis of the poplar Kunitz- type protease inhibitor family. *New Phytol.* **184**, 865-884 (2009).

65 Yamamoto, M., Saburo, H. & Ikenaka, T. Amino acid sequences of two trypsin inhibitors from winged bean seeds (*Psophocarpus tetragonolobus* (L) DC.). *J. Biochem.* **94**, 849-863 (1983).

66 Gatehouse, A. M., Hoe, D. S., Flemming, J. E., Hilder, V. A. & Gatehouse, J. A. Biochemical basis of insect resistance in winged bean (*Psophocarpus tetragonolobus*) seeds. *J. Sci. Food Agric.* **55**, 63-74 (1991).

67 Major, I. T. & Constabel, C. P. Functional analysis of the Kunitz trypsin inhibitor family in poplar reveals biochemical diversity and multiplicity in defense against herbivores. *Plant Physiol.* **146**, 888-903 (2008).

68 Theerasilp, S. & Kurihara, Y. Complete purification and characterization of the taste-modifying protein, miraculin, from miracle fruit. *J. Biol. Chem.* **263**, 11536-11539 (1988).

69 Takai, A. *et al.* Secretion of miraculin through the function of a signal peptide conserved in the Kunitz-type soybean trypsin inhibitor family. *FEBS Lett.* **587**, 1767-1772 (2013).

70 Kokiladevi, E., Manickam, A. & Thayumanavan, B. Characterization of alpha-amylase inhibitor in *Vigna sublobata*. *Botanical Bulletin of Academia Sinica* **46** (2005).

71 Manen, J.-F. *et al.* A nodulin specifically expressed in senescent nodules of winged bean is a protease inhibitor. *Plant Cell* **3**, 259-270 (1991).

72 Andrews, S. FastQC: A quality control tool for high throughput sequence data. *Reference Source* (2010).

73 Grabherr, M. G. *et al.* Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **29**, 644 (2011).

74 Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821-829 (2008).

75 Chevreux, B., Wetter, T. & Suhai, S. in *German conference on bioinformatics.* 45-56.

76 Mundry, M., Bornberg-Bauer, E., Sammeth, M. & Feulner, P. G. Evaluating characteristics of de novo assembly software on 454 transcriptome data: a simulation approach. *PloS One* **7**, e31410 (2012).

77 Kumar, S. & Blaxter, M. L. Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics* **11**, 571, doi:DOI: 10.1186/1471-2164-11-571 (2010).

78 Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674-3676, doi:10.1093/bioinformatics/bti610 (2005).

79 Da Maia, L. C. *et al.* SSR Locator: Tool for simple sequence repeat discovery integrated with primer design and PCR simulation. *International Journal of Plant Genomics* (2008).

80 McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32**, W20-W25 (2004).

81 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780 (2013).

82 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).

83 Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164-1165 (2011).

84 Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539-542 (2012).

85 Rambaut, A., Suchard, M. A., Xie, D. & Drummond, A. J. Tracer v1.5, available from http://beast.bio.ed.ac.uk/Tracer. (2013).

Figure 1: Summary of Gene Annotation analysis. Zero Hit refers to those in BLAST step without hits.

Figure 2: Gene Ontology classifications of Winged bean annotated transcripts. Numbers indicate the number of sequences associated with the particular GO term in each category.

885

886 **Figure 3: % Identity of CPP34-7 contigs against legume protein databases.**

887

888 **Figure 4: Legume sequence similarity analysis.** Relative numbers of contigs that had
889 significant sequence similarity by species for CPP34-7 contigs.

890

891 **Figure 5: Transcription factor family analysis.** Number of transcription factors determined
892 within the CPP34-7 assembly by transcription factor family.

893

894 **Figure 6: Results of microsatellite SSR analyses.** (A) Distribution of the 2,994 perfect
895 SSRs across different repeat size classes. Distribution of the number of repeats for (B) dimers
896 (C) trimers (D) tetramers (E) pentamers and (F) hexamers.

897

898 **Figure 7: Gene Tree of Kunitz Trypsin Inhibitor Gene Family.** Non-legume sequences:
899 *Arabidopsis thaliana* (At; black), *Populus trichocarpa* (Ptri; black). Legume sequences:
900 *Medicago truncatula* (Mt; pink), *Phaseolus vulgaris* (Pv; orange), *Glycine max* (Gm; green),
901 and *Psophocarpus tetragonolobus* from Pfam database (Ptet; navy blue), Chapman (2015)
902 transcriptome (Pt_c; royal blue), and CPP34-7 (Pt_isotig/contig; aqua blue). Sequence
903 notation is species abbreviation followed by Pfam accession number, contig/isotig number, or
904 read, followed by the range of amino acids used in the alignment. Numbers at the nodes are
905 posterior probability values and bootstrap supports. Subclades A-F are labeled. *Psophocarpus*
906 clades are indicated by arrows or blue banding. Tree rooted arbitrarily at an *Arabidopsis*
907 clade.

908

909 **Table 1.** Sequencing and assembly metrics for independent and combined assemblies using
910 GS *De Novo* Assembler.

Accessions	**Genotype CPP34**	**Genotype CPP37**	**Combined Assembly (CPP34-7)**
Number of raw reads	371,271	433,486	804,757
Number of bases (bp)	191,598,691	213,386,165	404,984,856
Number of reads post-filtering	369,820 (99.6%)	334,639 (77.2%)	704,459 (87.53%)
Number of bases post-filtering	136,943,216 (71.47%)	92,126,948 (43.17%)	178,911,104 (44.17%)
Number of reads aligned	277,351 (50.42%)	259,324 (63.04%)	435,897 (61.88%)
Number of contigs / bp	10,675 / 6,142,297	8,465 / 5,070,585	16,115 / 13,552,130
Avg. contig size (bp)	837	823	875
N50 (bp)	836	842	889
Longest contig (bp)	4,902	3,014	4,667
Number of singletons / bp	62,602 / 22,081,798	63,795 / 23,540,672	81,126 / 28,663,213
Number of transcripts / bp (contigs + singletons)	73,277 / 28,224,095	72,260 / 28,611,257	97,241 / 42,215,343

911

912

913 **Table 2.** Distribution of di- and trinucleotide repeat motif types in winged bean and
914 comparison with *Arabidopsis*.

Dinucleotide Repeat Composition	Number of transcripts	Percentage of Winged bean di- Repeats	Winged bean Rank	Percentage of Arabidopsis di- Repeats
AC/CA/GT/TG	22	8.6	3	8
AG/GA/CT/TC	199	77.7	1	83
AT/TA	35	13.7	2	8.8
CG/GC	0	0	4	0.14
Total	256	100		100

Trinucleotide Repeat Composition	Number of transcripts	Percentage of Winged bean tri- Repeats	Winged bean Rank	Arabidopsis Rank
AAC/ACA/CAA/GTT/TGT/TTG	134	11.4	4	3
AAG/AGA/GAA/CTT/TCT/TTC	343	29.1	1	1
AAT/ATA/TAA/TTA/TAT/ATT	58	4.9	8	5
ACC/CAC/CCA/GGT/GTG/TGG	118	10.0	5	4
ACG/CGA/GAC/CGT/GTC/TCG	35	3.0	9	9
ACT/CTA/TAC/AGT/TAG/GTA	13	1.1	10	8
AGC/CAG/GCA/TGC/CTG/GCT	136	11.5	3	7
AGG/GGA/GAG/TCC/CTC/CCT	118	10.0	6	6
ATC/CAT/TCA/GAT/ATG/TGA	164	13.9	2	2
CCG/CGC/GCC/GGC/GCG/CGG	59	5.0	7	10
Total	1,178	100		

915

916 **Table 3.** Results of single nucleotide polymorphism (SNP) detection between Sri Lankan and
917 Nigerian genotypes by degree of confidence.

918

Reads	95%	96%	97%	98%	99%	100%	Total
Contigs	43	68	76	94	50	2,552	2,883
Singletons	74	88	93	52	28	1,972	2,307
Total	117	156	169	146	78	4,524	5,190

919

920