Multiple post-domestication origins of kabuli chickpea through allelic variation in a diversification-associated transcription factor

R. Varma Penmetsa1, Noelia Carrasquilla-Garcia1, Emily M. Bergmann1, Lisa Vance1, Brenna Castro1, Mulualem T. Kassa1, Birinchi K. Sarma1,2, Subhojit Datta1,3, Andrew D. Farmer4, Jong-Min Baek1, Clarice J. Coyne5, Rajeev K. Varshney6, Eric J. B. von Wettberg7,8 and Douglas R. Cook1

1Department of Plant Pathology, University of California, One Shields Ave, Davis, CA 95616, USA; 2Department of Mycology and Plant Pathology, Banaras Hindu University, Pandit Madan Mohan Malviya Road, Varanasi, Uttar Pradesh 221005, India; 3Indian Institute of Pulses Research, Kanpur, Uttar Pradesh 208024, India; 4National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, NM 87505, USA; 5USDA-ARS, Western Regional Plant Introduction Station, Washington State University, 59 Johnson Hall, Pullman, WA 99164-6402, USA; 6International University, 11200 SW 8th Street, Miami, FL 33199, USA; 7Kushlan Institute for Tropical Science, Fairchild Tropical Botanic Garden, 10901 Old Cutler Road, Coral Gables, FL 33156, USA

Authors for correspondence: R. Varma Penmetsa
Tel: +1 530 752 6886
Email: rpenmetsa@ucdavis.edu

Douglas R. Cook
Tel: +1 530 754 6561
Email: drcook@ucdavis.edu

Received: 9 November 2015
Accepted: 24 March 2016

New Phytologist (2016)
doi: 10.1111/nph.14010

Key words: basic helix–loop–helix (bHLH) transcription factor, chickpea B locus, Cicer arietinum, diversification, domestication, legume ortholog, Mendel’s A locus.

Summary

• Chickpea (Cicer arietinum) is among the founder crops domesticated in the Fertile Crescent. One of two major forms of chickpea, the so-called kabuli type, has white flowers and light-colored seed coats, properties not known to exist in the wild progenitor. The origin of the kabuli form has been enigmatic.

• We genotyped a collection of wild and cultivated chickpea genotypes with 538 single nucleotide polymorphisms (SNPs) and examined patterns of molecular diversity relative to geographical sources and market types. In addition, we examined sequence and expression variation in candidate anthocyanin biosynthetic pathway genes.

• A reduction in genetic diversity and extensive genetic admixture distinguish cultivated chickpea from its wild progenitor species. Among germplasm, the kabuli form is polyphyletic. We identified a basic helix–loop–helix (bHLH) transcription factor at chickpea’s B locus that conditions flower and seed colors, orthologous to Mendel’s A gene of garden pea, whose loss of function is associated invariantly with the kabuli type of chickpea.

• From the polyphyletic distribution of the kabuli form in germplasm, an absence of nested variation within the bHLH gene and invariant association of loss of function of bHLH among the kabuli type, we conclude that the kabuli form arose multiple times during the phase of phenotypic diversification after initial domestication of cultivated chickpea.

Introduction

Crop domestication and subsequent diversification represent adaptations to human-built environments and offer insights into the evolutionary forces that shape phenotypic diversity. Chickpea (Cicer arietinum), a widely cultivated food legume, was domesticated in the Fertile Crescent together with several other so-called founder crops (Zohary et al., 2012). This is evidenced by the Neolithic archeological record (Tanno & Wilcox, 2006) and the prevalence of crop wild relatives in the eastern Mediterranean, particularly Cicer reticulatum, the wild annual Cicer species from which the cultigen is derived.

Both wild and cultivated chickpea are predominantly selfing (e.g. Singh & Diwakar, 1995) and diploid (2n = 16), with a genome size of 738 Mbp. Cultivated chickpea differs from its wild relatives in its erect plant growth habit, altered phenomenology, seed coat texture and reduced seed dormancy. These are ‘domestication syndrome’ traits that were probably favored by humans during crop domestication (Hammer, 1984). Further variation exists within cultivated chickpea, notably for seed features that reflect cultural preferences (size, shape, coat thickness and color), based on which cultivated chickpea is categorized as either ‘desi’ or ‘kabuli’ types. Seeds of desi are smaller in size, darker in color and with thicker seed coats, and are typically consumed after dehulling. In South Asia and sub-Saharan Africa, cultivation of desi predominated historically. The kabuli type is associated with light-colored, larger sized seeds and with thinner seed coats, and is typically consumed as whole seeds in west Asia and Mediterranean regions where kabuli is common.

The greater similarity of desi seed to that of wild C. reticulatum underlies the proposal that the desi type represents the early domesticated form (Ladizinsky & Adler, 1976), with kabuli being a subsequently derived type. These differences in market class partially correspond to differences in seed size and other morphological traits that have been used to draw a distinction between proposed microspermum and macrospermum subspecies
(Moreno & Cubero, 1978). Despite the widely held view that the desi and kabuli types are distinct subgroups within cultivated chickpea, molecular marker-based analyses have been equivocal in supporting such a distinction (Iruela et al., 2002; Upadhyaya et al., 2008; Roorkiwal et al., 2014). Thus, the origin of the kabuli type is currently ambiguous.

The dark-seeded desi and the light-seeded kabuli types also differ in pigmentation in vegetative tissues (stem) and in flowers. Pink to purple pigmented petals are typical of desi chickpea, whereas non-pigmented (white) petals are characteristic of kabuli chickpea. Additional variants exist for color of seed coats (green, black) and flowers (blue), but these occur at low frequencies (Pundir et al., 1985). Studies on the genetics of pigmentation in chickpea go back several decades (Pundir et al., 1985; and references therein). Three genetic loci, C, B and P, conditioning flower color in chickpea have been described, where recessive alleles at either C or B confer white petal color, whereas the recessive condition at P results in blue petals (Kumar et al., 2000).

Human selection for lighter seed color during crop domestication is well documented in pea (Hellens et al., 2010) and sorghum (Wu et al., 2012). Molecular genetic analyses of seed and flower color in soybean and garden pea (Hellens et al., 2010; Yang et al., 2010; Moreau et al., 2012) implicate genes in the phenylpropanoid pathway (Winkel-Shirley, 2001; Grotewold, 2006), in which a bifurcation typically leads either to the production of condensed tannins (proanthocyanidins) in seeds, giving them their dark appearance, or to the brightly colored anthocyanins characteristic of flowers. Both the biosynthetic enzymes and their transcriptional regulators have been well characterized in several plants, notably Arabidopsis, maize and Medicago, the latter of which is a close relative of chickpea (Xie et al., 2003; Baudry et al., 2004; Broun, 2005; Appelhagen et al., 2011; Dixon et al., 2013).

We performed molecular diversity analyses in a representative collection of chickpea germplasm, together with its wild annual relatives, to characterize the impact of domestication and to reveal relationships within chickpea germplasm that hitherto were masked by extensive admixture (intermating) introduced through crop breeding. In parallel, using a candidate pathway and gene approach, we identified a transcription factor whose molecular variants tightly associate with the seed coat and flower pigmentation differences characteristic of desi vs kabuli types. Our results indicate that these kabuli characteristics arose multiple times after domestication of the cultivated chickpea during the phase of post-domestication diversification, and that the varied and conflicting support for kabuli being a distinct subtype within cultivated chickpea is probably a consequence of demographic factors, such as admixture and isolation, that occurred in germplasm collection and crop breeding.

Materials and Methods

Plant genotypes

We assembled a diverse range of germplasm representing both desi and kabuli market types, the primary (Cicer reticulatum Ladiz.) and secondary (C. echinospermum P. H. Davis) gene pools, and annual Cicer accessions from the tertiary gene pools (see Supporting Information Table S1) to assess patterns of population genetic variation and correspondence with the kabuli form. Plants of the recombinant inbred population CRIL7 (Tekeoglu et al., 2000) and of the chickpea minicore were obtained from the US Department of Agriculture-Agricultural Research Service (USDA-ARS) in Pullman, WA, USA. Genomic DNA for the chickpea reference set (Upadhyaya et al., 2008) was obtained from ICRISAT, Hyderabad, India.

Single nucleotide polymorphism (SNP) assay development

Genotyping was based on SNPs (Table S2) discovered from a comparison of Sanger-sequenced PCR amplicons between C. arietinum accession ICC4958 (a desi type) and wild C. reticulatum accession PI489777. These two genotypes are the parents of an existing recombinant inbred line (RIL) population. Genomic DNA was isolated from young leaf tissue using a Qia-gen plant DNA isolation kit (Valencia, CA, USA) following the manufacturer’s protocols. Polymorphisms were identified by sequence alignment in Jalview among sequence pairs for a set of 1440 primarily single copy genes, with orthology inferred from legume expressed sequence tag (EST) data (i.e. the transcriptomes of Medicago truncatula, Lotus japonicus and Glycine max) and subsequently based on conserved genome location in a multispecies comparative genetic analysis (D. R. Cook et al., unpublished; and validated in chickpea, Hiremath et al., 2012). SNPs meeting the assay design criteria, determined by Illumina Inc. using their proprietary Assay Design Tool, were converted to a 768 Illumina GoldenGate genotyping assay performed at the UC Davis Genome Center DNA Technologies Core. The resulting data were curated using the Illumina Beadstudio software package (Illumina, San Diego, CA, USA). SNPs that failed in > 20 samples or when genotyping data failed to accurately identify parental genotypes among replicated samples were excluded. This filtering yielded 538 high-quality SNPs (of the 768 SNPs evaluated) representing polymorphisms in 538 genes.

Population genetic analyses

Genetic relationships among genotypes were analyzed with the program STRUCTURE using an admixture model and the default settings (Pritchard et al., 2000). Genotypes were from the primary, secondary and tertiary gene pools assayed for 538 SNPs in orthologous genes. In the initial analysis, we examined K = 2 to K = 10, using 20 iterations of 25 000–50 000 samples with a 25 000–50 000 burn-in phase. To select K that best fits the data, we used the inflection and plateaus of the log probability curves on various STRUCTURE runs, together with genetic, biological and geographical coherence of grouping. This indicated K = 8 as the maximal number of groups beyond which further subgroupings were poorly supported. K = 3 to K = 8 were analyzed in greater detail using three iterations of 1 000 000 samples with a 50 000 burn-in phase. The Evanno method, implemented in STRUCTURE HARVESTER, was used to select K that best fits the data (following Evanno et al., 2005; Earl, 2012). Principal coordinate
analysis (PCoA) in GenAIEx 6.5 (Peakall & Smouse, 2012) was used as a further step to visualize the clustering of genetic variation (Peakall & Smouse, 2006, 2012).

To further assess genetic diversity within and among groups, a set of 93 low (<10%) admixed accessions identified from STRUCTURE over K=5–9 was analyzed using DARwin5 software (Perrier & Jacquemoud-Collet, 2006) to calculate the pairwise dissimilarity by simple matching according to the method of Saitou & Nei (1987). The resulting dissimilarity matrix was used to derive a weighted neighbor-joining tree (Saitou & Nei, 1987) with 1000 bootstraps. This weighted neighbor-joining analysis employs a likelihood-based criterion that models distance between genotypes as random variables that obey a Gaussian distribution (Bruno et al., 2000).

GenAIEx 6.5 (Peakall & Smouse, 2012) was used to estimate the observed heterozygosity (H_e), expected heterozygosity (H_e), inbreeding coefficient (F_is) and percentage of polymorphic loci among groups that were circumscribed by geographical origin and population genetic analyses. Accessions were clustered by geographical origin hierarchically, first considering 10 geographical source regions (C. reticulatum, C. echinospermum, Indian subcontinent 1, 2 and 3, Iran 1 and 2, Syria, USSR and Ethiopia). Comparisons among groups were made using GenODive 2.0 (Meirmans & Van Tienderen, 2004). Analysis of molecular variance (AMOVA) was implemented in GenAIEx 6.5 with codominant genotypic data.

Characterization of basic helix–loop–helix (bHLH) candidate and transcriptional targets

Plants were surface sterilized in bleach, planted in potting soil and grown in glasshouses at the UC Davis core plant growth facility. Flower colors of plants were recorded at maturity. For gene expression analysis, flower petals and pools of seed coats from developing seeds were harvested. Young developing leaves served as the source of genomic DNA. Genomic DNA and RNA were extracted with the Qiagen DNEasy and Qiagen RNEasy total RNA extraction kits using the manufacturer’s standard procedures and quantified in a NanoDrop instrument (Wilmington, DE, USA).

For real-time reverse transcription polymerase chain reaction (rtPCR), equal amounts of total RNA were used in parallel rtPCRs analyzed on an ABI7500 instrument. The expression of the target genes BANYULS (BAN) and LDOX was quantified, together with that of the control gene Elongation Factor-1 (EF1) (Garg et al., 2010), and the relative transcript abundance was calculated by the cycle threshold (CT) method (Schefe et al., 2006) from triplicate samples relative to EF1. The PCR conditions are listed in Table S3.

For the determination of sequence variants, genomic DNA and cDNA were used as template in PCR amplification with ExTaq polymerase, and the resulting amplicons were purified by an ExoSAP kit (Affymetrix, Santa Clara, CA, USA) before Sanger sequencing (for bHLH allele B (wild-type) and the missense and short deletion variants, b1, b2, b3 and b5), or analyzed by agarose gel electrophoresis (for the large deletion variant b4). Sequence trace files were examined in Sequencer software to curate for quality and to identify nucleotide variants. The oligonucleotides used for amplification and sequencing are given in Table S3.

Results

Patterns of molecular diversity in cultivated chickpea and its close wild annual relatives

With the goal of describing the molecular diversity of cultivated chickpea relative to wild annual Cicer species, we conducted multi-locus genotyping in a representative set of 322 accessions, including 224 cultivated chickpea and 98 wild annual genotypes which, together, span the primary, secondary and tertiary gene pools (Table S1). Genotyping was based on 538 high-quality, bi-allelic SNPs that distinguish alleles of low-copy conserved orthologous genes of cultivated C. arietium ICC4958 and wild C. reticulatum PI489777 (Table S2).

Genetic relationships inferred from allele frequencies (i.e. STRUCTURE: Pritchard et al., 2000; Falush et al., 2003; Hubisz et al., 2009) revealed the separation of cultivated from wild accessions (Fig. 1). Moreover, analyses involving increased division into subgroups (i.e. K = 2, 3, ..., n) effectively resolved the wild progenitor species (i.e. C. reticulatum) and its nearest wild neighbor (C. echinospermum) from other wild species, and revealed numerous subgroups within cultivated material, including the bifurcation of cultivated material into accessions primarily from the Indian subcontinent and from West/Central Asia and the Middle East (Fig. 1a–c). The mean log probability of the data (ln P(D)) calculations (following Evanno et al., 2005; Earl, 2012) indicated that cultivated material is best described as six subgroups.

Estimates of genetic diversity among all groups, cultivated and wild, revealed the impact of domestication and breeding. AMOVA partitioned diversity among hierarchical sets of individual genotypes (two groups (cultivated and wild) and 10 subgroups) that were circumscribed by STRUCTURE. In an AMOVA, the majority of variation was attributed to differences between cultivated and wild groups (93%), with low levels of differentiation among cultivated groups (4%) and low levels of variation among individuals within groups (3%) (Table S4). Cultivated germplasm had significantly lower heterozygosity (P = 0.028) than either C. reticulatum or C. echinospermum (Table S4). Similarly, the cultivated group had a significantly lower level of gene diversity (H_e = 0.065; Table S5; P = 0.028 for the comparison) when compared with the wild relatives, C. reticulatum (H_e = 0.332; Table S5) or C. echinospermum (H_e = 0.301; Table S5). Although relatively small sample sizes may bias gene diversity estimates (Nei, 1978), genetic diversity within and differentiation among all cultivated groups were low.

Interestingly, none of the genetic groups identified based on allele frequency data corresponded to the often-asserted dichotomy of a desi and kabuli division within cultivated chickpea. This was most apparent at K = 3 (Fig. 1a), but remained a feature as additional subgroups were partitioned (Fig. 1b). To further test the hypothesis that kabuli has a single origin, not
reticulating with desi, a dissimilarity matrix was constructed and visualized as a neighbor-joining dendrogram (Fig. S1). At the species level, the resulting topology agreed with the prevailing taxonomic relationships among annual Cicer species (Javadi et al., 2007; Fig. S1). At a finer scale (Fig. 2), it readily identified the misclassified nature of accession PI 593709, a C. echinospermum genotype that was categorized as C. reticulatum at its collection time and in extant germplasm repositories. Accession PI 593709 grouped within a cluster that was exclusively composed of C. echinospermum accessions, as has been suggested previously (Iruela et al., 2002). Importantly, and in agreement with the results of STRUCTURE, among cultivated accessions the dissimilarity matrix did not yield a bipartite topology for desi and kabuli types. The same was true of analyses using the alternative minimum evolution method.

The analysis of population structure also revealed a high rate of admixture (Fig. 1). Among cultivated genotypes, such admixture is probably the consequence of breeding history. Because intermating during breeding might obscure inherent genetic stratification, we used membership coefficients obtained from STRUCTURE to identify 93 accessions with low levels (< 10%) of admixture: 81 of 222 C. arietinum, five of 16 C. reticulatum and all seven C. echinospermum accessions met this criterion. A dissimilarity matrix constructed from these 93 low admixed accessions separated domesticated from wild, and partitioned the 83 cultivated accessions into eight groups (Fig. 2). PCoA (Fig. 3) yielded congruent groupings, dividing clusters of genotypes along discriminant axes 1 and 2. The first axis (which accounted for 84.49% of the variation) separated cultivated from wild groups, whereas the second axis (5.75% of the variation) revealed four groups among the cultivated accessions.

In marked contrast with the result for all 322 accessions, groupings obtained with the low admixed set, either by neighbor joining or PCoA, reflected their geographical origins. Minimally,
Fig. 2 Weighted neighbor-joining dendrogram of the subset of 93 accessions with low levels of admixture. Eight groups within cultivated accessions and two additional groups, one each of the wild annual relatives *C. reticulatum* (light brown) and *C. echinospermum* (dark brown), are discernible. The dendrogram is rooted in three accessions from the distantly related wild species *C. yamashitae* (light gray). Bootstrap values are shown when > 40 (blue numbers). Names of the groups are to the right, with lowercase letters in parentheses corresponding to those in Fig. 1(c). ISC, accessions predominantly from the Indian subcontinent (largely from India and occasionally from Pakistan, Nepal, Bangladesh, Myanmar and Iran). In the dendrogram, desi accessions are in black text. The colors of the kabuli accessions correspond to their basic helix–loop–helix (bHLH) gene variants: light green, variant b1; turquoise, variant b2; dark green, variant b3; magenta, variant b4; red, variant b5.
Research

we identified a basal group consisting of accessions from Ethiopia, a subtending group composed of four geographically related subgroups from West/Central Asia and the Middle East, and a coherent subgroup predominantly from the Indian subcontinent that was further divisible into two or three subgroups based on neighbor-joining analyses or PCoA, respectively (Figs 2, 3). The fact that groups that formed from low admixed genotypes reflect geographical origins of accessions is probably a consequence of local breeding practices and regional cultivar preferences. Interestingly, even among these low admixed individuals, we observed kabuli and desi flower and seed coat phenotypes distributed throughout multiple groups (Figs 1c, 2), suggestive of introgression and extensive backcross selection for these traits, and/or of the possibility of their multiple origins.

Functional variation in a transcriptional regulator assorts with the kabuli type

The differences in pigmentation of flowers, stems and seed coats between desi and kabuli chickpea (Fig. 4) are suggestive of differences in flavonoid biosynthesis. Because transcriptional changes often accompany crop domestication (e.g. Doebley et al., 2006; Olsen & Wendel, 2013), we examined components of the ternary ‘BMW’ transcriptional complex, composed of a bHLH, a Myb-domain and a WD40-repeat protein, which is a key regulatory module for flavonoid biosynthesis (Baudry et al., 2004; Broun, 2005; Grotewold, 2006; Appelhagen et al., 2011; Hichri et al., 2011). Our analysis of previous low-resolution genetic maps between chickpea (Millan et al., 2010) and pea (Hellens et al., 2010), using the Medicago genome as a bridge, determined that the phenotypically defined chickpea ‘B’ locus is syntenic and thus potentially orthologous to the A locus of pea. The A locus is Mendel’s flower color gene A and is conditioned by the BMW bHLH protein (Hellens et al., 2010).

Degenerate PCR primers, based on the pea and Medicago orthologs of A (bHLH), were used to amplify and then sequence cDNAs from petals and seed coats of desi (colored) chickpea genotype ICC4958 and wild C. reticulatum genotype PI599072 (also colored petals and dark seed coats), yielding identical sequence homologs of the bHLH protein. Phylogenetic analysis (Fig. 4f) positioned the chickpea candidate as orthologous to the authentic BMW bHLH proteins of Medicago and pea, whereas comparison with the whole-genome sequence of the reference kabuli accession CDC Frontier (Varshney et al., 2013) placed the chickpea ortholog on CaLG04, consistent with previous linkage mapping studies that position the B locus of chickpea to the top arm of CaLG4 (Millan et al., 2010). Recombinants in a RIL population derived from C. reticulatum PI599072 and kabuli FLIP94-92C (Tekeoglu et al., 2000) and segregating for flower color delimited a physical interval of 741 kbp containing 61 genes, including the bHLH, in a region syntenic to Medicago chromosome 8 and pea LG-II, where the functionally characterized orthologs of bHLH are located. Sequencing of this bHLH revealed a c. 1 kbp deletion specific to the kabuli parent FLIP84-92C that removes 223 bp of the coding region, spanning all of exon 4 and a portion of exon 5 (Fig. 4g). This deletion is shared in common with the genome sequenced reference kabuli genotype ‘CDC Frontier’ (Varshney et al., 2013) and results in coding region variation, including a premature stop codon that is predicted to eliminate protein function.

Multiple altered functional alleles of the bHLH transcription factor are correlated with absent target gene expression

Sequencing of full-length bHLH transcripts from flower and seed coat tissues in a set of 35 accessions (11 desi and 23 kabuli C. arietinum accessions and one accession of the wild relative C. echinospermum), encompassing the known seed color variation of chickpea (dark, green, dark tan and light tan), identified five coding region variants exclusive to and characteristic of kabuli individuals (Fig. 4g). We designated the allele of the desi type as ‘B’ and the variant alleles of kabuli types as ‘b’1–5 to reflect their known or predicted recessive nature relative to B (Fig. 4g). Conceptual translation of the three deletion alleles ‘b’1, ‘b’3 and ‘b’4 indicates that these result in frameshifts and premature termination of translation (Fig. 4g). In particular, allele b4, which is a 223-bp deletion in the transcript, corresponds to the c. 1 kbp length variant identified in kabuli genomic DNAs and used for co-segregation analysis, as described earlier. The missense variant allele b2 converts glutamine to proline (Q71P) within the Myb interaction domain of the bHLH protein (Hichri et al., 2011). This glutamine residue is invariant among bHLH orthologs, including proteins from basal vascular lineages (mosses), suggesting that it may be critical for protein–protein interactions between the Myb and bHLH components of the BMW regulatory complex, a prediction that is also supported by SIFT (Kumar et al., 2009). Missense variant allele b5 is restricted to and characteristic of...
accessions that share an intermediate flower color phenotype, wherein only portions of petals, particularly vasculature, are pigmented (Fig. 4c). In the deduced b5 protein, serine replaces leucine (S497L) in a residue that is conserved in bHLH proteins throughout vascular plants. This particular leucine residue is one of 19 conserved residues that are hallmarks of the bHLH domain (Toledo-Ortiz et al., 2003) and among 11 residues that form the dimerization interface, suggesting that the weak flower color phenotype of the b5 allele may derive from reduced hetero- or homodimerization capacity in the b5 variant.

To test the deduced effects of these variants on bHLH protein function, we quantified the expression of chickpea orthologs of leucoanthocyanin dioxygenase (LDOX/ANS) and anthocyanidin reductase (ANR/BANYULS), as these enzymes are known transcriptional targets of the BMW ternary complex in other plants (Appelhagen et al., 2011), and their expression is indicative of anthocyanin production in flowers and pro-anthocyanidin...
Table 1 Basic helix–loop–helix (bHLH) transcript haplotypes in desi and kabuli accessions in cultivated chickpea (Cicer arietinum)

<table>
<thead>
<tr>
<th>Type</th>
<th>No. of samples</th>
<th>Allele frequency (overall) (%)</th>
<th>Allele frequency (within type) (%)</th>
<th>Nucleotide position &amp; sequence within coding region</th>
<th>Predicted effect on protein</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>As homozygous</td>
<td>As heterozygous</td>
<td></td>
<td>199–200 (2 bp) 215 238–244 (7 bp) 461–683 (223 bp) 1490</td>
<td></td>
</tr>
<tr>
<td>desi</td>
<td>8</td>
<td>107</td>
<td>2</td>
<td>60.8 99.1</td>
<td>WT/functional protein</td>
</tr>
<tr>
<td>kabuli</td>
<td>b1</td>
<td>1</td>
<td>0</td>
<td>0.6 1.4</td>
<td>2-bp deletion; frameshift</td>
</tr>
<tr>
<td></td>
<td>b2</td>
<td>36</td>
<td>0</td>
<td>20.1 51.4</td>
<td>Mis-sense CAA → CCA; glutamine → proline</td>
</tr>
<tr>
<td></td>
<td>b3</td>
<td>1</td>
<td>0</td>
<td>0.6 1.4</td>
<td>7-bp deletion; frameshift</td>
</tr>
<tr>
<td></td>
<td>b4</td>
<td>26</td>
<td>1</td>
<td>14.8 37.9</td>
<td>Large internal 223-bp deletion; frameshift</td>
</tr>
<tr>
<td></td>
<td>b5</td>
<td>5</td>
<td>1</td>
<td>3.1 7.9</td>
<td>Mis-sense TTA → TCA; leucine → serine</td>
</tr>
</tbody>
</table>

The nature of allelic variants, their frequencies among all cultivated germplasm, and within desi and kabuli seed types, together with the predicted effects on the bHLH protein, are shown. CT, here is for the nucleotides ‘C’ and ‘T’ that occupy positions 199 and 200 within this sequence. Only variant nucleotide positions are listed. WT, wild-type.
group, where variant \( b_4 \) was also found. Variant \( b_4 \) was exceptional in being one of two frequent alleles among \( k\)abuli overall (Table 1), but occurring only once in the low admixed set, in the USSR clade (Fig. 2; accession IG74036 from Moldova, formerly USSR). This suggests that most \( k\)abuli accessions containing variant \( b_4 \), including the reference \( k\)abuli accession ‘CDC Frontier’, have received the \( b_4 \) allele through breeding introgression. Indeed, the average admixture rate of accessions containing allele \( b_4 \) was 34%, with 21 of 28 accessions with admixture rates in excess of 20%.

**Discussion**

The domestication of plants is shaped by human preference for traits, ranging from altered growth habit to ease of harvest, enlargement and flavor of edible parts, adjusted phenology and visual appearance of flowers or seeds. Artificial selection for preferred traits has continued since domestication. The selection and recombination of light-colored chickpea seed (i.e. the \( k\)abuli type of chickpea) represent the manifestation of such human preference. Indeed, condensed tannins can reduce protein digestibility and impart an astringent character to seed (Aw & Swanson, 1985; Dixon et al., 2013), which may have been factors in the selection of light-colored seed and, coincidentally, white-flowered plants. Consistent with this suggestion, contemporary chickpea culinary practices mimic genetics in that dark-seeded \( d\)esi, but not light-seeded \( k\)abuli, types are typically consumed after dehulling. Here, we provide genetic and transcriptional evidence that flower and seed coat color in chickpea derive from multiple, independent alleles of the same gene, selected independently to create the same light-colored phenotype throughout the existing chickpea gene pool. As is common to many other artificially selected agricultural traits (e.g. Doebley et al., 2006; Olsen & Wendel, 2013), the targets of selection were loss-of-function mutations in a transcriptional regulator, the chickpea ortholog of Mendel’s A gene in pea. With the exception of its heterogeneous origins, the genetic architecture underlying the loss of flower and seed coat color in cultivated chickpea germplasm is thus relatively simple and not unlike growth habit and flowering time differences between wild and cultivated chickpea, which are also controlled by large-effect loci (Millan et al., 2010). Indeed, major effect mutations swept to high frequency by human selection are common in the domestication history of most crops, although increasing numbers of minor loci are also being identified (reviewed in Doebley et al., 2006; Zachary et al., 2007; Olsen & Wendel, 2013).

The finding that all \( k\)abuli genotypes contain only single bHLH variants without nesting of mutations or evidence of interallelic recombination argues against a stepwise mutation accumulation model, and instead suggests that the five alleles arose independently, a conclusion that is further supported by the largely non-overlapping distribution of bHLH mutant alleles among genetically resolved groups. The preponderance in cultivated chickpea of multiple mutant isoforms in only one component of the BMW transcriptional complex is curious, given that loss of function in other components of the BMW complex leads to comparable phenotypes in other legumes. Thus, loss-of-function mutations are the source of breeding alleles in cultivated pea at the bHLH ‘A’ locus and at the WDR ‘A2’ locus (Hellens et al., 2010), and in soybean at the Myb ‘W2’ locus (Takahashi et al., 2012). It is possible that the Myb and WDR components of the chickpea BMW complex confer additional adaptations and thus are under purifying selection for function, or that there is (partial) functional redundancy for these components. Indeed, multiple Myb paralogs in grape have the capacity to regulate structural genes in the anthocyanin and condensed tannin pathways (Appelhagen et al., 2011; and references therein).

Although the significant majority of characterized natural variation in anthocyanin and/or proanthocyanidin content in seeds and flowers derives from altered transcriptional control (as in the cases of sorghum, Wu et al., 2012; rice, Sweeney et al., 2006; apple, Espley et al., 2009; cauliflower, Chiu et al., 2010), flavonoid biosynthetic genes are also potential targets of human selection in crops (e.g. soybean, Yang et al., 2010) and possibly also in chickpea. Thus, in chickpea, a tandem cluster of MATE transporters, orthologous to TT12 that acts in proanthocyanidin biosynthesis in legumes (Zhao & Dixon, 2009), occurs in a narrow genomic interval that has unusually low diversity among \( d\)esi but not \( k\)abuli genotypes (Varshney et al., 2013), consistent with a selective sweep and purifying selection at these TT12 orthologs in dark-seeded genotypes. Likewise, variation in flavonoid biosynthetic loci could provide the range of hues of almost white through very dark brown that exist in chickpea, and indeed allelic variation in these biosynthetic genes accounts for flower color variation in some species (Broun, 2005).

Although white-flowered \( k\)abuli genotypes are common in cultivated chickpea, they are rare in wild \( C\)icer, being known only in perennial \( C\)icer species in central Asia (Uzbekistan) in the tertiary gene pool (Van Oss et al., 2015), or derived from induced mutagenesis (Toker, 2009). We surveyed 23 natural populations of the cultigen’s progenitor (\( C\). \( r\)eticulatum) from the primary gene pool, and its sister species \( C\). \( e\)chinaspernum from the secondary gene pool, and did not encounter white-flowered individuals among >1000 wild annual \( C\)icer individuals observed throughout southeastern Anatolia. This suggests strong natural selection for colored flowers and/or seed in natural settings. Indeed, flower color can be an important determinant of pollinator activity (Stanton et al., 1986, 1989), even in self-pollinated legumes, in which pollinator activity can increase seed set (Palmer et al., 2009). Moreover, the products of anthocyanin and proanthocyanidin metabolism function as antioxidants and defensive chemicals and have been speculated to play a variety of ecological roles (Gould, 2004; Strauss & Whittall, 2006; Yuan et al., 2012). Although it is possible that pre-existing null alleles of bHLH were introgressed into cultivated chickpea from co-occurring wild relatives in which the alleles may have been at low frequencies, we favor the hypothesis that the existing chickpea bHLH loss-of-function alleles arose from novel post-domestication mutation events during the phase of crop diversification. Indeed, the maintenance of multiple, independent loss-of-function \( k\)abuli bHLH alleles within cultivated germplasm is consistent with the observation that many traits associated with crop improvement, as
opposed to domestication, remain variable in modern germplasm (e.g. Pundir et al., 1985).

Our results strengthen and extend the inferred phylogeny of annual Cicer of previous studies (Javadi et al., 2007), identifying C. reticulatum as the likely progenitor of cultivated chickpea. Moreover, the identification of genetic groups, resolved by multiple criteria that largely correlate with geographical distribution, adds an additional dimension to our understanding of chickpea germplasm. For example, at K=3 (Fig. 1), cultivated chickpea bifurcates into two groups whose origins are predominantly from the Indian subcontinent, and West/Central Asia and the Middle East, respectively. This separation may reflect the historically distinct breeding mandates of regional chickpea breeding centers (i.e. the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) in India and the International Center for Agricultural Research in the Dry Areas (ICARDA) in Syria), although such patterns also mirror known events in early secondary domestication. Thus, the distinctiveness of Indian and Ethiopian genotypes (Figs 1, 2) agrees with the archeological record, indicating that chickpea cultivation spread from the Middle East to South Asia at least 4000 yr ago and to Ethiopia at least 2300 yr ago (Redden & Berger, 2007); indeed, Harlan (1992) recognized India and Ethiopia as centers of chickpea diversity. In contrast with the idea of species distinctiveness, however, we identified many C. reticulatum accessions and several cultigen accessions that are admixtures (see Table S1), indicating gene flow between cultivated and wild forms and highlighting the potentially protracted, even ongoing, contribution of wild diversity to cultivated chickpea. Although chickpea and its relatives are predominantly selfing, some outcrossing is possible. Alternatively, these may represent contamination in collections, consistent with the lack of admixture recently found by Van Oss et al. (2015) in a working collection of wild material. Irrespective of their origins, from a practical point of view, hybrid accessions have the potential to serve as genetic bridges for the introgression of agronomically desired traits (e.g. biotic and abiotic stress tolerance, phenological variation) from more distantly related wild species into cultivated chickpea, which has thus far enjoyed only sporadic and unpredictable success (Hajjar & Hodgkin, 2007; Warschefsky et al., 2014). Even more extensive admixture is observed with cultivated germplasm among groups with distinct geographical origins (evident in Fig. 1a–c at multiple values of K, and in Fig. 2); this is probably the consequence of crop breeding, and may underpin the success of crop breeding in exploiting what diversity there is within cultivated germplasm.

A recent study (Agarwal et al., 2012) suggested that the origin of kabuli as a distinct entity (i.e. its separation from desi forms) dated to no less recently than 0.20–0.22 million yr ago (Mya), with evidence (mis)extrapolated from a molecular clock applied to sequence differences between single kabuli and desi genotypes. Agarwal’s estimate requires the existence of distinct populations of kabuli genotypes > 0.2 Mya, well before the generally accepted onset of agriculture, a conjecture unsupported by evidence. Although we do not exclude the possibility that some kabuli alleles existed at low frequency in wild Cicer populations since well before the origin of agriculture, we see it as far more likely that kabuli genotypes arose and gained favor with farmers c. 6000 yr ago, potentially as a means of distinguishing crop–wild hybrids or to improve flavor by reducing the astringency of seed.

Breeders have traditionally viewed kabuli forms of chickpea as a genetically distinct group, indeed as a homogenous entity. Moreno & Cubero (1978) grouped large-seeded and often light-colored chickpeas as distinct from smaller seeded and dark-colored forms based exclusively on phenotypic data. Although their distinction is not one of kabuli vs desi, per se, there is large overlap among these two market classes and their proposed megasperma and microspermum forms. Ladizinsky & Adler’s (1976) suggestion that desi forms represent the early domesticate, with kabuli being a subsequently derived type, seems more accurate. Our results provide an alternative model and one that is supported by both genetic and molecular data; namely, that kabuli is not a genetic entity at all, but rather a set of independently derived traits, including multiple loss-of-(or reduced)-function alleles of a transcription factor that block anthocyanin and anthocyanidin production in flowers and seeds, respectively. We suggest that the causal alleles for altered flower color and seed coat tannins arose post-domestication, in distinct subsets of the cultivated desi germplasm, and were then introgressed by breeders into today’s preferred cultivars. This shift in viewpoint has important implications for chickpea breeding, because it dispels the notion that kabuli forms are a genetically distinct group and compels a shift from the common practice of focusing germplasm collections and breeding along a desi/kabuli partition. White-flowered, light-seeded chickpeas are simply mutant forms of domesticated desi, with the various alleles embedded within genomes that are more closely related to desi forms than they are to one another.

Acknowledgements
This work was supported by the following awards: National Science Foundation DBI-0605251 to D.R.C.; National Science Foundation IOS-0965531 to D.R.C. and R.K.V.; United States Department of Agriculture-Agricultural Research Service, Specific Cooperative Agreement no. 58-5348-1-220 to D.R.C. and R.V.P. We thank David L. Remington and Justen B. Whittall for critical reading of earlier drafts of the manuscript, and George Vandemark for kindly providing CRIL2 and CRIL7 populations.

Author contributions

References


**Supporting Information**

Additional Supporting Information may be found online in the Supporting Information tab for this article:

**Fig. S1** Neighbor-joining dendrogram of 322 annual \textit{Cicer} accesses from analysis of 538 SNPs (single nucleotide polymorphisms), recapitulates the taxonomic relationships among the primary, secondary and tertiary gene pools.

**Table S1** Summary of annual \textit{Cicer} germplasm used in this study, and basic helix–loop–helix (bHLH) allelic form among genotypes of cultivated chickpea

**Table S2** Sequences flanking bi-allelic single nucleotide polymorphisms used for multi-locus genotyping of germplasm

**Table S3** Oligonucleotides used in this study, together with the analysis for which they were used

**Table S4** Summary results of hierarchical analyses of molecular variance across the wild (\textit{Cicer reticulatum} and \textit{C. echinospermum}) and cultivated (\textit{C. arietinum}) species, and among cultivated groups

**Table S5** Genetic diversity estimates in cultivated chickpea and its near wild relatives \textit{Cicer reticulatum} and \textit{C. echinospermum}

Please note: Wiley Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the \textit{New Phytologist} Central Office.