ARTICLE

Sequence analysis of the arcelin-phytohaemagglutinin-α-amylase inhibitor (APA) locus in three phylogenetically arrayed *Phaseolus vulgaris* clones

Crop Breeding and Applied Biotechnology
22(4): e421122415, 2022
Brazilian Society of Plant Breeding.
Printed in Brazil
http://dx.doi.org/10.1590/1984-70332022v22n4a50

Juliano Lino Ferreira[1*], James Kami[2], Aluízio Borem[3] and Paul Gepts[2]

Abstract: *APA is a multigene locus that comprises the protein genes of arcelin (Arc), phytohemagglutinin (PHA) or lectin, and an α-amylase inhibitor (α-AI). These genes play essential roles in the defense responses of legume seeds to protect against seed weevils in common bean (P. vulgaris). During the evolution of this complex locus, PHA proteins appeared first, followed by α-AI and Arc proteins. This study compared and analyzed the sequences of three bacterial artificial chromosome (BAC) clones containing APA fragments of the genotypes G02771, BAT93 and DGD1962. These three genotypes are recognized as representative of the crucial steps of APA evolution in common bean. Our findings demonstrated that rearrangements during evolution can be used to characterize the APA locus, but concurrent adjacent gene regions on either side of the APA locus are highly conserved. Part of the instability of the APA locus may be due to the insertion of retroelements and gene conversion.*

Keywords: *Comparative genomics, genome, common bean*

INTRODUCTION

The common bean (*P. vulgaris* L.) is a member of Leguminosae that is largely adapted to environments with moderate growing temperatures. This species is categorized into the following two major types: dry edible beans and snap or garden beans. Globally, their crop economic valuation is not accurate because other species are regularly counted in the statistical data collected in different countries, resulting in a reported production of approximately 18.9 million T for all categories, which makes this the most extensively produced grain (Myers and Kmiecik 2017). In Brazil, given the importance of this legume, advancements made during the common bean breeding program of the Federal University of Lavras during the last 50 years and improvements made by the common bean breeding program from the Instituto Agronômico de Campinas since 1932 were reported. (Lemos et al. 2020, Bezerra et al. 2021).

Seeds of dry edible beans are vulnerable to predation by postharvest pests, including bean weevils (bruchids), which feed on proteins, carbohydrates, and lipids in the grains. The major pests of common bean seeds worldwide are the bruchids (Coleoptera), including *Acanthoscelides obtectus* (Say), commonly known as the bean weevil, and *Zabrotes subfasciatus* (Boh.), usually known as the Mexican bean weevil. *A. obtectus* is the insect that is most harmful to

***Corresponding author:**
E-mail: juliano.ferreira@embrapa.br
ORCID: 0000-0002-8502-4444

Received: 06 June 2022
Accepted: 13 December 2022
Published: 28 December 2022

[1] Embrapa Pecuária Sul, Rodovia BR 153, km 633, 96401-970, Bagé, RS, Brazil
[2] University of California, Davis, 1 Shields Avenue, Davis, CA 95616, United States of America
[3] Universidade Federal de Viçosa, Avenida Peter Henry Rolfs, s/n, Campus Universitário, 36570-900, Viçosa, MG, Brazil

stored grains in Brazil (Carneiro et al. 2015), and it causes losses of between 7 and 13% of the common bean yield in Latin America (Alvarez et al. 2005). Blair et al. (2010a) stated that, generally, bruchids account for approximately 13% of grain losses in common bean. According to a review by Tigist (2020), bean bruchid damage depends on the storage period and storage conditions. Compiled investigates have reported that losses vary between 10 and 40% of dry seed weight. It has also been reported that losses could reach up to 50-70% in on-farm storage facilities because of the lack of postharvest management practices.

In *P*. *vulgaris*, the Arcelin-Phytohaemagglutinin-α-Amylase inhibitor (APA) protein family is encoded by a complex, multigene locus on chromosome Pv04, which includes genes for arcelin (Arc), phytohemagglutinin (PHA), or lectin, and an α-amylase inhibitor (α-AI). In the evolutionary pathway leading to *P*. *vulgaris*, extensive evolution of the APA locus has occurred as follows: a lectin ancestor gene underwent a paralogous duplication event giving rise to the progenitor of the true lectin and the progenitor of the other lectin-related genes. The latter progenitor evolved, generating the gene coding for the biologically active form of α-amylase inhibitor and, through a second duplication event, the Arc genes found only in some wild accessions of *P*. *vulgaris*; this likely resulted from the domestication bottleneck that has characterized this crop species (Lioi et al. 2007). APA genes at the DNA level are between 732 and 825 bp long, share identity above 77% and are intronless. The differences in functionality among APA members are mainly derived from sequence variations, which cause structural changes by the elimination of one, two or three loops in the tertiary protein structure (Rougé et al. 1993).

In a comprehensive review, Duarte et al. (2018) reported the important role of APA genes in bruchid resistance and affirmed that these genes are expressed exclusively during the development of the bean seed in cotyledons and the embryonic axis. This review also notes that in a screening panel of 210 wild Mexican accessions, numerous accessions exhibited natural resistance against two important bruchid pests. The presence of Arc genes – the last evolutionary change in the APA locus – was associated with this resistance, evidencing the importance of a complete set of APA genes. According to Tigist et al. (2019), members of the APA family confer resistance against bean bruchids, affecting the survival and growth of *Z*. *subfasciatus*.

Recently, studies have demonstrated the importance and applicability of APA loci in resistance to seed weevils in subsistence agriculture. For example, Mukankusi et al. (2019) reported the association of one SNP marker with the APA locus in a review. Additionally, Kamfwa et al. (2018) identified one QTL related to resistance to *A*. *obtectus* associated with the APA locus, which indicated that the resistance to this weevil is somewhat complex and includes the interaction of the APA locus with other loci (Kamfwa et al. 2018). In parallel, Blair et al. (2010b) suggested that other genes out of the APA locus may be needed in combination with arcelin for resistance to develop. Zaugg et al. (2013) also reported that Arc is sufficient for resistance to bruchid beetles, especially *Z*. *subfasciatus* and, to a lesser extent, *A*. *obtectus*. According to Zaugg et al. (2013), seven Arc variants were identified in all wild *P*. *vulgaris* accessions that were reported before their study. However, resistance to both bruchid species only occurs in those containing Arc-4. These authors also describe a new *P*. *vulgaris* accession containing new APA variants, Arc-8 and ARL-8, which confer resistance to both common bean weevil species; this result provides a valuable genotype for the purpose of breeding bruchid resistance.

A bacterial artificial chromosome (BAC) is an engineered DNA molecule that clones DNA sequences into bacterial cells (e.g., *Escherichia coli*). The BAC cloning system can stably maintain large DNA fragments (100,000 to 300,000 base pairs) from an organism, with a low rate of chimerism and high clonal stability; furthermore, this system is easy to manipulate (Shizuya et al. 1992). Therefore, the BAC system is a valuable tool for various genome mapping, sequencing projects, and comparative genomics. Thus, studies on the comparison of genomes arose with the advent of genome sequencing and comparing genomes is a powerful, respected, and valuable method for understanding adaptation and evolution. The purpose of this study was to better understand the change in the APA locus in *P*. *vulgaris* using BAC sequences containing the locus from three well-known common bean genotypes that represent crucial steps in APA evolution by scanning for genes, mobile elements, and conserved sequences at this locus.

MATERIAL AND METHODS

Kami et al. (2006) isolated and constructed BAC libraries of the APA locus from the three carefully chosen *P*. *vulgaris* clones used in this investigation. The following methods were thoroughly described in that work. Specifically, BAC libraries were established in the following common bean genotypes: 1) G02771, a wild Mesoamerican accession with genotype Arc$^+$; PHA$^+$; αAI$^+$ (Goossens et al. 2000). 2) BAT93, a domesticated Mesoamerican breeding line (Vlasova et al. 2016) with genotype PHA$^+$; αAI$^+$; Arc$^-$ and 3) DGD1962, a wild accession of northern Peru containing ancestral phaseolin sequences in *P. vulgaris* (Rendón-Anaya et al. 2017) and with the genotypic formula PHA$^+$; αAI$^+$; Arc$^-$. The sequence of APA in G02771 - BAC clone 71F18 - has been deposited with GenBank (DQ323045) and was previously published by Kami et al. (2006). In this work, we performed bioinformatics analysis of the APA loci of DGD1962 and BAT93. These loci were not in GenBank, therefore, they were determined following the scheme used by Kami et al. (2006), i.e., sequencing of an APA gene family-containing the BAC clone from the accessions for BAT93 and DGD1962.

The BAC sequences of BAT93 and DGD1962 were then screened for open reading frames (ORFs) with AUGUSTUS (Stanke and Waack 2003), GeneScan (Burge and Karlin 1998), TwinScan (Gross and Brent 2006) and FGENESH (Solovyev et al. 2006). The gene prediction in AUGUSTUS was trained with the following options: *Arabidopsis thaliana* and *Zea mays*. Along the same lines, scanning with FGENESH used the following options: *Nicotiana tabacum*, *Medicago truncatula*, dicot or monocot. For this analysis, each application of the FGENESH program trained with one of the three types of sequence data was considered a different program. Finally, the ORF sequence of each BAC clone was screened through the best agreement from the output of AUGUSTUS, FGENESH, GeneScan and TwinScan. Consistent with the procedure of Kami et al. (2006), agreement with three or more programs in each region was the criterion to go to the next step, which was a BLAST search (Altschul et al. 1990) for both proteins and nucleotides. Based on the highest value in GenBank, not less than e^{-20}, the region was identified by the sequence name in this database. Prior to the comparison, the genes, mobile elements, and other fragments were manually entered in each BAC sequence using WebACT software (Abbott et al. 2005). After this step, the BAC library sequences of the three clones were compared using ACT software (Carver et al. 2012). This comparison was useful to verify rearrangements, conserved regions, sequence duplication, etc. One can choose the smallest amount of sequence homology with this software. In this study, 800 bp was chosen as the minimum size criterion for homology between sequences. In this manner, the sequences of the three APA-containing BAC clones were analyzed and compared.

The following is a list of abbreviations used in this study: Arc: arcelin; ARL: arcelin-like; Gag-pol integrase: gag-pol polyprotein (Integrase core domain); Gag-pol RT/RNase: gag-pol polyprotein (RT/Rnase H domain); PHA: phytohemagglutinin or lectin; SYP81: syntaxin of plants 81; PTI 2B: putative translation initial 2B beta subunit; AAI-1 precursor: alpha-amylase inhibitor 1 precursor; PHA-L: phytohemagglutinin - leucoagglutinin; PHA-E: phytohemagglutinin - erythroagglutinin; PHA Pdlec2: Leucoagglutinating phytohemagglutinin; α-AI or AAI: α-amylase inhibitor; En/Spm: enhancer/suppressor mutator; OsI: hypothetical protein OsI_028310 [*Oryza sativa* Indica Group]; Ser/Threo: serine/threonine specific protein phosphatase PP2A; FLD: flowering locus D; CDS: coding DNA sequence.

RESULTS AND DISCUSSION

The following three common bean BAC clones with the APA locus were analyzed: the APA clone from G02771 provided by Kami et al. (2006) and two additional sequenced clones reported for the first time here, one from DGD1962, a wild clone from northern Peru containing ancestral phaseolin sequences, and the other from BAT93, a Mesoamerican breeding line. The DGD1962 and BAT93 BAC clones were constructed and sequenced by Kami et al. (2006), and these sequences represent approximately 0.02% to 0.033% of the entire common bean genome (McClean et al. (2004). The BAC clone sequence of G02771 has six genes at the APA locus, numbered APA-1 to APA-6 from the 5' to the 3' end. Among these genes, the first two are unique to G02771, a wild Mesoamerican clone with resistance to bruchids. APA-1 and APA-2 belong to the Arc subfamily. APA-3, APA-4, and APA-5 belong to the phytohemagglutinin subfamily, and APA-6 belongs to the α-amylase inhibitor subfamily (Kami et al. 2006).

Screening of the BAT93 and DGD1962 BAC libraries led to the identification of several subclone sequences matched to APA sequences, as described by Kami et al. (2006) for the G02771 BAC library. The BAT93 and DGD1962 sequences were largely colinear (Figure 1) despite some rearrangements. In these BAC clones, in addition to the presence of APA genes, several other sequences of importance were identified in the clones DGD1962 (Table 1), BAT93 (Table 2), and G02771, the latter of which was presented and detailed in Kami et al. (2006).

All three BAC clones display sequences with homology to retrotransposons, which are used to separate the sequences of the three APA subfamilies (Figure 1). In G02771, the region between 78,000 bp and the BAC clone 3' end aligned with that of other BAC clones. In contrast, the upstream section was not well aligned with that of the other genotypes. This may be partially explained by the fact that the upstream region of G02771 is interrupted by a chloroplast DNA insertion (Kami et al. 2006). This question can be addressed by identifying new BAC clones 5' upstream of the current G02771

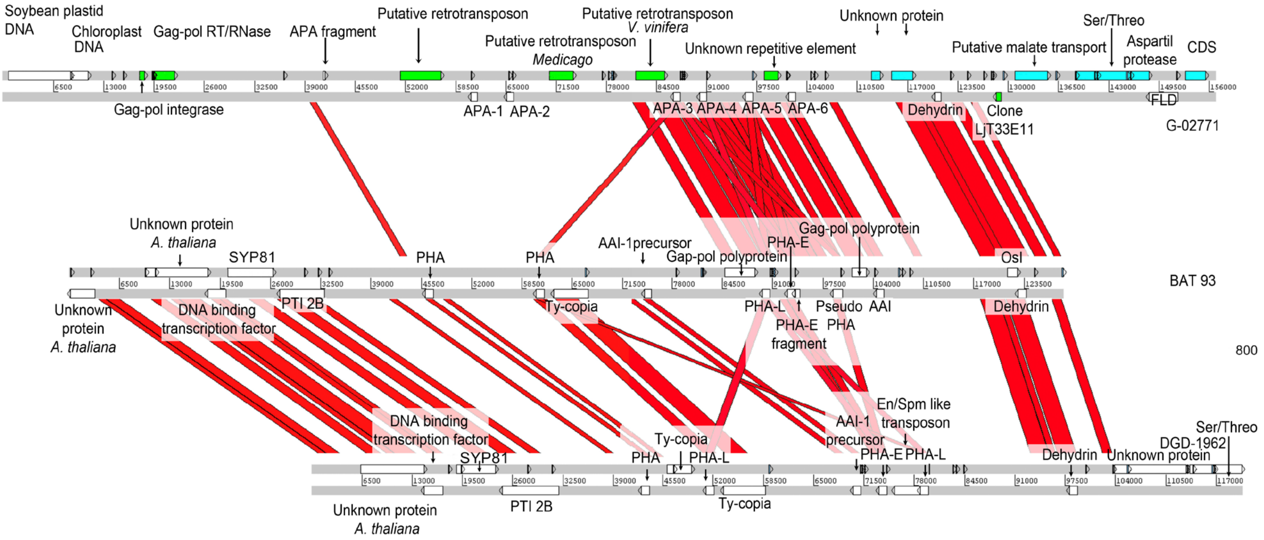

Figure 1. Comparison of the APA locus sequence between G02771 (156,768 bp), BAT93 (128,559 bp) and DGD1962 (120,380 bp). The red line indicates regions that share more than 800 bp. Abbreviations: Gag-pol integrase: gag-pol polyprotein (Integrase core domain); Gag-pol RT/RNase: gag-pol polyprotein (RT/Rnase H domain); SYP81: syntaxin of plants 81; PTI 2B: putative translation initial 2B beta subunit; PHA: phytohemagglutinin; AAI-1 precursor: alpha-amylase inhibitor 1 precursor; PHA-L: phytohemagglutinin - leucoagglutinin; PHA-E: phytohemagglutinin - erythroagglutinin; AAI: α-amylase inhibitor; En/Spm: enhancer/suppressor mutator; OsI: hypothetical protein OsI_028310 [*Oryza sativa* Indica Group] ; Ser/Threo: serine/threonine specific protein phosphatase PP2A; FLD: Lz-0 flowering locus D; CDS: coding DNA sequence

Table 1. Putative gene regions (PGR) in the APA BAC clone of *Phaseolus vulgaris* accession DGD1962

PGR	Position (bp)	Direction*	DNA sequence similarity	Size (DNA)	BLAST result (ATGC)	Organism	E value	GenBank Accession #
1	6366-14464	F	60% (235/388)	460	unknown protein	*A. thaliana*	$2e^{-129}$	NP_683491
2	14566-16871	C	38% (61/158)	315	DNA-binding/transcription factor	*A. thaliana*	$6e^{-16}$	NP_177022
3	18769-23659	F	64% (224/350)	349	SYP81	*A. thaliana*	$9e^{-94}$	NP_564597
4	24627-31886	C	81% (336/414)	418	putative translation initiation factor 2B beta subunit	*N. tabacum*	$5e^{-175}$	AAD52847
5	42767-43725	C	94% (332/353)	959	PHA	*P. vulgaris*	$4e^{-148}$	X04659
6	46035-49147	F	41% (212/505)	491	Ty1-copia	*O. sativa* (japonica cultivar group)	$2e^{-101}$	ABA95205
7	51109-52034	C	95% (882/926)	926	PHA-L	*P. vulgaris*	0	X02409
8	53367-58664	C	85% (800/938)	5298	Ty1-copia retrotransposon	*S. melongena*	0	DQ644594
9	70125-70957	C	99% (728/730)	833	alpha-amylase inhibitor-1 precursor	*P. vulgaris*	0	EF087992
10	73439-74375	C	99% (935/938)	937	PHA-E	*P. vulgaris*	0	X02408
11	75437-78711	C	58% (66/112)	202	En/Spm-like transposon proteins	*M. truncatula*	$8e^{-27}$	ABE79785
12	78782-79704	C	99% (932/933)	933	PHA-L	*P. vulgaris*	0	X02409
13	98022-99035	C	90% (220/244)	1014	dehydrin	*V. unguiculata*	$7e^{-76}$	AF159804
14	103865-113068	F	62% (366/582)	593	unknown protein	*A. thaliana*	0	NP_188473
15	113569-120241	F	88% (220/249)	6673	Ser/Thr specific protein phosphatase PP2A	*M. sativa*	$4e^{-66}$	AF196287

* F – Forward strand; C – Complementary strand

Table 2. PGR in the APA BAC clone of *Phaseolus vulgaris* accession BAT93

PGR	**Position** (bp)	**Direction***	**DNA sequence similarity**	**Size** (DNA)	**BLAST result**	**Organism**	**E value**	**GenBank Accession #**
1	138-3280	C	42% (181/426)	406	unknown protein	*A. thaliana*	$7e^{-68}$	AAM13173
2	9809-17819	F	60% (236/388)	460	unknown protein	*A. thaliana*	$4e^{-130}$	NP_683491
3	17917-20206	C	39% (61/154)	315	DNA-binding/transcription factor	*A. thaliana*	$6e^{-16}$	NP_177022
4	20525-26242	F	64% (224/350)	349	SYP81	*A. thaliana*	$1e^{-93}$	NP_564597
5	27227-32916	C	75% (333/442)	446	putative translation initiation factor 2B beta subunit	*Nicotiana tabacum*	$3e^{-170}$	AAD52847
6	46015-47026	C	93% (372/400)	1012	PHA Pdlec2	*P. vulgaris*	$1e^{-157}$	X04659
7	60389-61314	C	100% (926/926)	926	PHA Pdlec2	*P. vulgaris*	0	X04659
8	62640-67006	C	84% (901/1063)	4367	Ty1-copia	*Solanum melongena*	0	DQ644594
9	74410-75242	C	100% (735/735)	833	alpha-amylase inhibitor-1 precursor	*P. vulgaris*	0	AY603476
10	84791-88590	F	42% (81/190)	485	gag-pol polyprotein	*P. vulgaris*	$8e^{-41}$	AAR13317
11	89656-90595	C	92% (763/828)	940	PHA-L	*P. coccineus*	0	AJ438774
12	92909-93847	C	98% (937/939)	939	PHA-E	*P. vulgaris*	0	X02408
13	93915-94428	C	93% (220/235)	514	PHA-E fragment	*P. vulgaris*	$3e^{-92}$	X04660
14	98846-99954	C	100% (975/975)	1109	pseudogene Pdlec1 for PHA	*P. vulgaris*	0	X04660
15	101226-103022	F	39% (119/299)	264	gag-pol polyprotein	*P. vulgaris*	$4e^{-47}$	AAR13317
16	104397-105278	C	99% (867/868)	882	alpha-amylase inhibitor	*P. vulgaris*	0	D49828
17	121369-122563	F	61% (43/70)	96	hypothetical protein OsI_028310	*O. sativa* (indica cultivar-group)	$3e^{-20}$	EAZ07078
18	122685-123698	C	90% (218/240)	1014	dehydrin	*Vigna unguiculata*	$3e^{-78}$	AF159804

* F – Forward strand; C – Complementary strand

clone using primers from the 5’ end of BAT93. The soybean plastid DNA, chloroplast, gag-pol integrase, and gag-pol rt/RNAse (reverse transcriptase/ribonuclease H) sequences were found only at the 5’ end of the G02771 BAC clone. The G02771 clone also had more retroelements than the BAT93 and DGD1962 clones. In plant genomes, retrotransposons are highly abundant, and they are frequently a major element of nuclear DNA (Li et al. 2004).

An additional feature of chloroplast DNA is the introduction of apparently random sequence fragments (called ‘NUPTS’) into the nuclear genome (Rousseau-Gueutin et al. 2018). This phenomenon has been observed in several plants, including *Arabidopsis*, maize, and tobacco. This accounts for the presence of the chloroplast DNA sequence in the G02771 APA clone (Kami et al. 2006). The control of the switch to flowering and the degree of the resistance response is strongly regulated by flowering locus D-like protein (FLD) to ensure that the plant continues to reproduce. Plants can respond to biotic stress by shifting their flowering time (Korves and Bergelson 2003). Figure 1 shows the presence of the FLD gene only in the G02771 APA clone, due to the shortness of the clone reads from BAT93 and DGD-1962. All three clones contained a dehydrin sequence toward their 3’ end. Dehydrin proteins play a major role in plant feedback and adaptation to abiotic stresses. These proteins are usually stored in aging seeds or are manufactured in plant tissues as a result of salinity, dehydration, cold and freezing stress. It is worth noting that plant mechanical damage, which is a common biotic stress applied by insects or herbivores, can also be interpreted as dehydration stress since it causes cellular damage leading to water loss (Hanin et al. 2011).

As observed by Kami et al. (2006), a dehydrin gene flanks the APA locus. Dehydrins constitute a family of proteins termed late-embryogenesis-abundant D11 [LEA]; these proteins are normally activated by environmental stress, pressures linked with low temperature and/or dehydration, and dehydration during seed maturation. In the three BACs, this gene is located at the 3’ end of the APA sequences, acting as a landmark. A region with a sequence matching the protein phosphorylase 2A (PP2A) regulatory subunit was detected downstream of the APA gene array in G02771 (Kami et al. 2006). The same coding region was present in DGD1962. Figure 1 shows that the fragment, comprising both the dehydrin and PP2A genes at the 3’ end, is strongly aligned and well-preserved between the three common bean lineages. Hence, the absence of PP2A in the BAT93 sequence is probably due to the short length of the BAT93 clone, and this idea is supported by the truncation of the clone immediately downstream of the dehydrin gene.

The DGD1962 sequence included four APA genes of the phytohemagglutinin subfamily and one APA gene of the α-amylase inhibitor subfamily (Figure 1). The α-AI precursor present in DGD1962 aligned well with the α-AI precursor of BAT93, whereas in G02771, this gene appears to have been lost (Figure 1). However, similar to G02771, BAT93 possesses an α-AI gene next to the dehydrin gene in a well-conserved region. Therefore, this last AI gene, APA-6 in G02771 or just α-AI in BAT93, seems to have been conserved.

When the BAT93 APA sequence was analyzed, seven APA genes were identified, including a pseudo-PHA sequence. Among them, two belonged to the α-amylase inhibitor subfamily, while the others belonged to the phytohemagglutinin subfamily. The PHA-L on BAT93 shares similarity with PHA-L, PHA-E and the second PHA-L on DGD-1962.

In the BAT93 and DGD1962 sequences, the regions comprising the first APA gene and its 5' region were also colinear (Figure 1). This segment contained the following four putative coding sequences: a DNA-binding transcription factor, a PTI 2B, a SPY81, and an unknown protein of *A. thaliana*. Bubeck et al. (2008) showed that extremely high expression of the SYP81 protein triggered a dosage-dependent restriction on α-amylase secretion; this function is similar with its role in protection against bruchids. Whether this close linkage between the two bruchid resistance loci is a coincidence remains to be determined.

Between G02771 and DGD1962, there was an inversion of APA-3 and APA-5 from G02771 based on their respective positions on DGD1962 (Figure 2). The APA-3 sequence of G02771 had three homologies in DGD-1962, including the first PHA-L, the next PHA-E, and the second PHA-L sequence. The APA-4 sequence of G02771 was homologous to that of PHA-E on DGD1962. The APA-5 sequence of G02771 had sequence similarity with PHA-E of DGD1962. Interestingly, the clone LjT33E11 and a putative malate transport sequence appeared to separate the unknown protein and the serine/threonine (Ser/Threo) gene in G02771, but these two sequences were absent in DGD1962 (Figure 2).

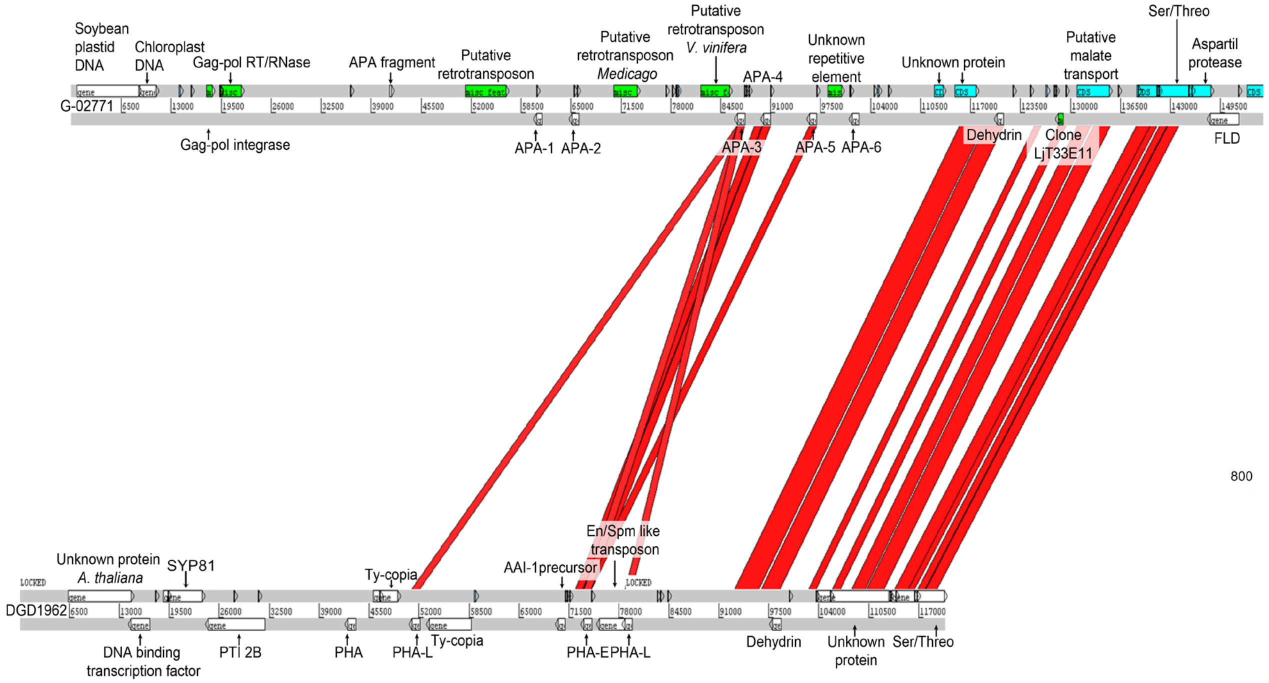

Figure 2. Comparison of the APA locus sequence between G02771 (156,768 bp) and DGD1962 (120,380 bp). The red line indicates regions that share more than 800 bp. Abbreviations: Gag-pol integrase: gag-pol polyprotein (Integrase core domain); Gag-pol RT/RNase: gag-pol polyprotein (RT/Rnase H domain); SYP81: syntaxin of plants 81; PTI 2B: putative translation initial 2B beta subunit; PHA: phytohemagglutinin; AAI-1 precursor: alpha-amylase inhibitor 1 precursor; PHA-L: phytohemagglutinin - leucoagglutinin; PHA-E: phytohemagglutinin - erythroagglutinin; AAI: *α*-amylase inhibitor; En/Spm: enhancer/suppressor mutator; Ser/Threo: serine/threonine specific protein phosphatase PP2A; FLD: Lz-0 flowering locus D; CDS: coding DNA sequence

Figure 1 shows that in G02771, APA-3 shared homology with PHA and PHA-E in BAT93, but in the opposite direction. In the same direction, APA-4 of G02771 shared similarity with PHA-L and pseudo-PHA in BAT93. The G02771 sequence fragment had more retroelements than the BAT93 and DGD1962 APA locus sequences. In G02771, the APA-3 to APA-6 sequences were more closely spaced than those in BAT93 and DGD1962. BAT93 and DGD1962 had four complete APA genes. One PHA-E sequence and another pseudo-PHA were present in BAT93. This pseudo-PHA of BAT93 had sequence similarity with the PHA-E gene on DGD1962 (Figure 1). Among the three BAC sequences, PHA-L showed sequence similarity with PHA-E as well as a PHA-E fragment. APA-5 shared similarity with PHA-L and PHA-E.

The PTI 2B gene was present in BAT93 and DGD1962 but was absent in G02771. Correspondingly, the DNA-binding transcription factor was present in BAT93 and DGD1962. One reason for this discrepancy is that this sequence was not present in the G02771 fragment alignment. The En/Spm-like retrotransposon appeared only in DGD-1962. The complete AAI gene was present only in BAT93. While the AAI-1 precursor was found in BAT93 and DGD1962, the LjT33E11, a putative malate transporter, and FLD sequences were only found in the G02771 BAC clone. Only the BAT93 sequence included the hypothetical protein OsI_028310.

Between the APA-2 and APA-3 sequences in G02771, a putative retrotransposon of *Medicago* was observed. This element was not present in BAT93 or in DGD1962. The APA-1 and APA-2 sequences in G02771 represented a clear expansion of the APA gene family compared to the BAT93 and DGD1962 APA sequences. The homology of the region between the APA-3 and APA-6 sequences was stronger between G02771 and BAT93 and between G02771 and DGD1962 (Figures 1 and 2).

Comparing the DGD1962 and BAT93 sequences, one can see a contradiction in the region between the ty1-copia and the α-AI precursor sequences. However, there was a large expansion between the α-AI precursor and the next APA gene, which is, in part, explained by the insertion of the Gag-pol polyprotein. The region comprising the last APA gene and PHA-L at 84,500 bp in DGD1962 (Figure 1) also indicated an expansion in BAT93. This expansion, in this case, was likely due to the appearance of a Gag-pol polyprotein and the generation of an additional APA gene (α-AI) by duplication. These two expansions are conserved in the genotype G02771. Additionally, the area between the first (PHA) and second (PHA-L) APA genes in DGD1962 presents another expansion in BAT93, but in this case, it is not associated with a retroelement insertion.

Analysis of Figure 1 shows that the putative rearrangements in the APA region primarily affect the APA locus itself, whereas the adjacent regions are highly colinear. Although we cannot exclude issues with sequence assembly based on relatively high sequence conservation, we propose here an alternative explanation for the increased frequency of rearrangements at the APA locus. Interspersed among the APA sequences were several retrotransposons or transposons. Unequal crossing-over among repetitive sequences such as these transposable elements and homologous APA sequences could lead to segmental duplications (Mieczkowski et al. 2006). Alternatively, gene conversion between APA member sequences could alter the collinearity as well (Cossu et al. 2017). While sequencing of complex loci such as the APA locus presented difficulties in the past because of the shortness of sequence reads, increased read lengths are now greatly facilitating sequencing and the assembly of complex loci (Vollger et al. 2019).

CONCLUSION

This analysis demonstrated that rearrangements along the evolutionary path can be used to characterize the APA locus; however, concurrent adjacent gene regions on each side of the APA locus were highly conserved. Part of this instability may be due to the insertion of retroelements and gene conversion.

The present study highlights the importance of the complete set of APA loci and suggests that the development of markers associated with Arc genes discovered in wild Mesoamerican accession clone G02771 would benefit breeding. Looking ahead, we designed a set of 47 new microsatellite-like molecular markers (SSRs) to help us track the APA locus in common bean germplasm. This upcoming work found two markers were positioned on the Arc 5 gene, and one was located 4500 bp downstream from Arc-5 in G02771 clone. These important and relevant results will be published as a continuation of this study.

ACKNOWLEDGMENTS

The project was supported by the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service, grant number #00–01972, to PG. This work was conducted during a scholarship supported by the International Cooperation Program CAPES/COFECUB at the University of California, Davis, and financed by CAPES – Brazilian Federal Agency for Support and Evaluation of Graduate Education within the Ministry of Education of Brazil.

REFERENCES

Abbott JC, Aanensen DM, Rutherford K, Butcher S and Spratt BG (2005) WebACT - an online companion for the Artemis Comparison Tool. **Bioinformatics 21**: 3665-3666.

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. **Journal of Molecular Biology 215**: 332-333.

Alvarez N, Hossaert-McKey M, Rasplus J-Y, McKey D, Mercier L, Soldati L, Aebi A, Shani T and Benrey B (2005) Sibling species of bean bruchids: a morphological and phylogenetic study of *Acanthoscelides obtectus* Say and *Acanthoscelides obvelatus* Bridwell. **Journal of Zoological Systematics and Evolutionary Research 43**: 29-37.

Bezerra LMC, Fredo CE, Chiorato AF and Carbonell SAM (2021) The research, development, and innovation trajectory of the IAC Common Bean Breeding Program. **Crop Breeding and Applied Biotechnology 21**: e36872124.

Blair MW, Muñoz C, Buendía HF, Flower J, Bueno JM and Cardona C (2010a) Genetic mapping of microsatellite markers around the arcelin bruchid resistance locus in common bean. **Theoretical and Applied Genetics 121**: 393-402.

Blair MW, Prieto S, Díaz LM, Buendía HF and Cardona C (2010b) Linkage disequilibrium at the APA insecticidal seed protein locus of common bean (*Phaseolus vulgaris* L.). **BMC Plant Biology 10**: 79.

Bubeck J, Scheuring D, Hummel E, Langhans M, Viotti C, Foresti O, Denecke J, Banfield DK and Robinson DG (2008) The syntaxins SYP31 and SYP81 control ER–Golgi trafficking in the plant secretory pathway. **Traffic 9**: 1629-1652.

Burge CB and Karlin S (1998) Finding the genes in genomic DNA. **Current Opinion in Structural Biology 8**: 346-354.

Carneiro JES, Paula Junior T and Borém A (2015) **Feijão: do plantio a colheita**. UFV, Viçosa, 384p.

Carver T, Harris SR, Berriman M, Parkhill J and McQuillan JA (2012) Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. **Bioinformatics 28**: 464-469.

Cossu RM, Casola C, Giacomello S, Vidalis A, Scofield DG and Zuccolo A (2017) LTR retrotransposons show low levels of unequal recombination and high rates of intraelement gene conversion in large plant genomes. **Genome Biology and Evolution 9**: 3449-3462.

Duarte MAG, Cabral GB, Ibrahim AB and Aragão FJL (2018) An overview of the APA locus and arcelin proteins and their biotechnological potential in the control of bruchids. **Agri Gene 8**: 57-62.

Goossens A, Quintero C, Dillen W, De Rycke R, Valor JF, De Clercq J, Van Montagu M, Cardona C and Angenon G (2000) Analysis of bruchid resistance in the wild common bean accession G02771: no evidence for insecticidal activity of arcelin 5. **Journal of Experimental Botany 51**: 1229-1236.

Gross SS and Brent MR (2006) Using multiple alignments to improve gene prediction. **Journal of Computational Biology 13**: 379-393.

Hanin M, Brini F, Ebel C, Toda Y, Takeda S and Masmoudi K (2011) Plant dehydrins and stress tolerance. **Plant Signaling & Behavior 6**: 1503-1509.

Kamfwa K, Beaver JS, Cichy KA and Kelly JD (2018) QTL mapping of resistance to bean weevil in common bean. **Crop Science 58**: 2370-2378.

Kami J, Poncet V, Geffroy V and Gepts P (2006) Development of four phylogenetically-arrayed BAC libraries and sequence of the APA locus in *Phaseolus vulgaris*. **Theoretical and Applied Genetics 112**: 987-998.

Korves TM and Bergelson J (2003) A developmental response to pathogen infection in *Arabidopsis*. **Plant Physiology 133**: 339-347.

Lemos RC, Abreu AFB, Souza EA, Santos JB and Ramalho MAP (2020) A half century of a bean breeding program in the South and Alto Paranaíba regions of Minas Gerais. **Crop Breeding and Applied Biotechnology 20**: e295420211.

Li W, Zhang P, Fellers JP, Friebe B and Gill BS (2004) Sequence composition, organization, and evolution of the core Triticeae genome. **The Plant Journal 40**: 500-511.

Lioi L, Galasso I, Lanave C, Daminati MG, Bollini R and Sparvoli F (2007) Evolutionary analysis of the APA genes in the Phaseolus genus: wild and cultivated bean species as sources of lectin-related resistance factors? **Theoretical and Applied Genetics 115**: 959-970.

McClean P, Gepts P and Kami J (2004) Genomics and genetic diversity in common bean. In Wilson R, Stalker H and Brummer E (eds) **Legume crop genomics**. AOCS Press, Champaign, p. 60-82.

Mieczkowski PA, Lemoine FJ and Petes TD (2006) Recombination between retrotransposons as a source of chromosome rearrangements in the yeast *Saccharomyces cerevisiae*. **DNA Repair 5**: 1010-1020.

Mukankusi C, Raatz B, Nkalubo S, Berhanu F, Binagwa P, Kilango, M, Williams M, Enid K, Chirwa R and Beebe S (2019) Genomics, genetics and breeding of common bean in Africa: A review of tropical legume project. **Plant Breeding 138**: 401-414.

Myers JR and Kmiecik K (2017) Common bean: Economic importance and relevance to biological science research. In Pérez de la Vega M, Santalla M and Marsolais F (eds) **The Common bean genome**. Compendium of plant genomes. Springer, Cham, p. 1-20.

Rendón-Anaya M, Montero-Vargas JM, Saburido-Álvarez S, Vlasova A, Capella-Gutierrez S, Ordaz-Ortiz JJ, Aguilar OM, Vianello-Brondani RP, Santalla M, Delaye L, Gabaldón T, Gepts P, Winkler R, Guigó R, Delgado-Salinas A and Herrera-Estrella A (2017) Genomic history of the origin and domestication of common bean unveils its closest sister species. **Genome Biology 18**: 60.

Rougé P, Barre A, Causse H, Chatelain C and Porthé G (1993) Arcelin and α-amylase inhibitor from the seeds of common bean (*Phaseolus vulgaris* L.) are truncated lectins. **Biochemical Systematics and Ecology 21**: 695-703.

Rousseau-Gueutin M, Keller J, Carvalho JF, Aïnouche A and Martin G (2018) The intertwined chloroplast and nuclear genome coevolution in plants. In Ratnadewi D and Hamim H (eds) **Plant growth and regulation: Alterations to sustain unfavorable conditions**. IntechOpen, London, p. 61-84.

Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y and Simon M (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. **Proceedings of the National Academy of Sciences USA 89**: 8794-8797.

Solovyev V, Kosarev P, Seledsov I and Vorobyev D (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. **Genome Biology 7**: 1-12.

Stanke M and Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. **Bioinformatics 19** (Supplement 2): ii215-ii225.

Tigist SG (2020) Common bean (*Phaseolus vulgaris* L.) and the bean bruchid (*Zabrotes subfasciatus*): A Review. **International Journal of Research in Agriculture and Forestry 7**: 21-31.

Tigist SG, Melis R, Sibiya J, Amelework AB, Keneni G and Tegene A (2019) Population structure and genome-wide association analysis of bruchid resistance in Ethiopian common bean genotypes. **Crop Science 59**: 1504-1515.

Vlasova A, Capella-Gutiérrez S, Rendón-Anaya M, Hernández-Oñate M, Minoche AE, Erb I, Câmara F, Prieto-Barja P, Corvelo A, Sanseverino W, Westergaard G, Dohm JC, Pappas GJ Jr, Saburido-Alvarez S, Kedra D, Gonzalez I, Cozzuto L, Gómez-Garrido J, Aguilar-Morón MA, Andreu N, Aguilar OM, Garcia-Mas J, Zehnsdorf M, Vázquez MP, Delgado-Salinas A, Delaye L, Lowy E, Mentaberry A, Vianello-Brondani RP, García JL, Alioto T, Sánchez F, Himmelbauer H, Santalla M, Notredame C, Gabaldón T, Herrera-Estrella A and Guigó R (2016) Genome and transcriptome analysis of the Mesoamerican common bean and the role of gene duplications in establishing tissue and temporal specialization of genes. **Genome Biology 17**: 32.

Vollger MR, Dishuck PC, Sorensen M, Welch AE, Dang V, Dougherty ML, Graves-Lindsay TA, Wilson RK, Chaisson MJP and Eichler EE (2019) Long-read sequence and assembly of segmental duplications. **Nature Methods 16**: 88-94.

Zaugg I, Magni C, Panzeri D, Daminati MG, Bollini R, Benrey B, Bacher S and Sparvoli F (2013) QUES, a new *Phaseolus vulgaris* genotype resistant to common bean weevils, contains the Arcelin-8 allele coding for new lectin-related variants. **Theoretical and Applied Genetics 126**: 647-661.