

Identification of acyl transferases responsible for the diverse hydroxycinnamoyl chemical space in *Bidens pilosa*

By

Khuliso Mathatha

(14000441)

Dissertation

Submitted in fulfilment of the requirements

for the degree of

Master of Science (MSc)

In

Biochemistry

In the faculty of Science, Engineering and Agriculture

at the

University of Venda, South Africa

Supervisor: Professor N.E Madala

February 2022

Table of Contents

Declaration.....	vi
Dedication.....	vii
Acknowledgement.....	viii
Preface.....	ix
Executive summary.....	x
List of Abbreviations.....	xii
List of Units.....	xv
List of Figures and Tables.....	xvi
Chapter 1.....	1
Introduction and literature review.....	1
1.1. Medicinal plants.....	2
1.2. Asteraceae plants.....	3
1.3. <i>B. pilosa</i> is an underutilised plant species.....	4
1.3.1. Nutraceutical importance of <i>B. pilosa</i>	5
1.4. Plant secondary metabolites (SMs).....	7
1.5. The phenylpropanoid pathway.....	9
1.6. Chlorogenic Acids.....	12
1.7. Pharmacological importance of CGA.....	14
1.7.1. Anti-HIV properties of CGAs.....	14

1.7.2. Anti-diabetic Activity	15
1.8. Decoding the genetics behind the chemistry of <i>B. pilosa</i>	16
1.9. SMRT sequencing	17
1.10. Study Rationale	19
1.11. Aim	20
1.12. Objectives	20
References	21
Chapter 2	32
Application of SMRT sequencing approach for identification of putative hydroxycinnamoyl-CoA: quinate/shikimate acid hydroxycinnamoyl transferase genes from <i>Bidens pilosa</i> L. responsible for biosynthesis of different forms of chlorogenic acids	33
Abstract	34
2.1 Introduction	35
2.2. Methodology	38
2.2.1. Total RNA isolation	38
2.2.2. cDNA library construction and SMRT sequencing	38
2.2.3. Identification of potential HQT/HCT genes from <i>B. pilosa</i> L.	39
2.2.4. Percentage similarity	39
2.2.5. Integrity of Sequences	40
2.2.6. Multiple sequence alignment (MSA)	40

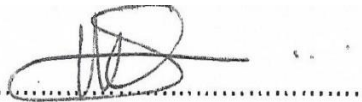
2.2.7. Phylogenetic Analysis	40
2.2.8. Metabolite extraction.....	41
2.2.9. Liquid chromatography mass spectrometry	41
2.3. Results and discussion	43
2.3.1. Characterization of putative Hydroxycinnamoyl-CoA: quinic hydroxycinnamoyl transferase gene 1	45
2.3.2. Characterization of putative Hydroxycinnamoyl-CoA: quinic hydroxycinnamoyl transferase gene 2	48
2.3.3. Characterization of putative Hydroxycinnamoyl-CoA: quinic hydroxycinnamoyl transferase gene 3	50
2.3.4. Characterization of putative Hydroxycinnamoyl-CoA: shikimate hydroxycinnamoyl transferase gene	52
2.3.5. Overview	Error! Bookmark not defined.
2.4. Metabolites profiling.....	58
2.4.1. Annotation of Mono-Acyl Chlorogenic Acids.....	58
2.4.2. Annotation of <i>p</i> -Coumaroyl-Caffeoylquinic Acids	59
2.4.3. Characterization of di-Caffeoylquinic Acids	60
2.4.4. Characterization of Feruloyl-Caffeoylquinic Acids	61
2.5. Conclusion.....	62
Acknowledgements	63
Conflict of Interest	63
References	64

Supplementary Data.....	70
Chapter 3.....	149
Identification of putative acyl transferase genes responsible for biosynthesis of homogenous and heterogenous hydroxycinnamoyl-tartaric acid esters from <i>Bidens pilosa</i>	150
Abstract.....	151
3.1 Introduction.....	152
3.2. Methodology.....	155
3.2.1. Total RNA isolation.....	155
3.2.2. cDNA preparation and sequencing.....	155
3.2.3. Identification of potential HTT genes from <i>B. pilosa</i>	156
3.2.4. Percentage similarity.....	156
3.2.5. Integrity of Sequences.....	156
3.2.6. Multiple sequence alignment (MSA).....	157
3.2.7. Phylogenetic Analysis.....	157
3.2.8. Metabolite extraction.....	157
3.2.9. Liquid chromatography mass spectrometry.....	157
3.3. Results and Discussion.....	159
3.4. Conclusion.....	168
References.....	169
Supplementary Data.....	176

Chapter 4	185
4. Conclusion	186

Declaration

I, **Khuliso Mathatha**, declare that this thesis submitted to the University of Venda for Master of Science degree in Biochemistry under the faculty of science has not been submitted to any other University. I can further declare that the work presented in this dissertation is my work, with exception of the referenced material which has been properly cited and acknowledged.

Signature: 

Date: 22/02/2022

Dedication

This work is dedicated to my mom, Tendani Tshisikule.

Ndi khou livhuwa u tikiwa nga dzi thabelo, maipfi a thuthuwedzo, u pfesesa na lufuno lwe vha ntsumbedza musi ndi tshi khou ita mushumo hoyu. Nga ndothe ndo vha ndi nga si kone u bveledza hoyu mushumo. Ndi a vha livhuwa nga u vha thikho ya vhutshilo hanga.

Ndaa!

Acknowledgement

First and foremost, I would like to thank the **Almighty God** for giving me the sufficient grace and strength to complete this project.

I also wish to extend my sincere gratitude to **Prof N.E Madala** for giving me an opportunity to be trained through this project by him. Most notably, his continuing support, enthusiasm, advice, patience, his supervision and mentoring. He always saw beyond what I could do and showed me that I can do much more if I focus and work hard. He has been a great support system in both academic and life in general, I have learnt a lot. *Zwe vha mpfunza na u gudisa, zwi do vha zwone zwi no amba kha lwendo lwanga uya phanda. Ndi a livhuwa.*

To **Dr I. Mwaba**, thank you for all the guidance you have given me through this project. I am very grateful for your patience with me and the time you invested in me. *Merci beaucoup.*

To **Dr L. Mathomu**, you are appreciated for going through my work, advice and ideas that saw this project through. *Ndo livhuwesa.*

I wish to further extend my acknowledgement to the following institutions for funding this project:

- **The National Research Foundation (NRF) of South Africa**
- **Shimadzu SA Bursary**

I am extremely grateful for my family and friends for their support, love, prayers and their encouragement which helped me in completion of this project.

Finally, I would love to thank my lab mates and the whole unit of Biochemistry for their motivation and moral support throughout this study.

Preface

This document has been prepared in a form of manuscripts submitted for publication.

The outline is as follows:

Executive summary

Chapter 1: Introduction and Literature review

Manuscript submitted:

Mathatha, K., Mwaba, I., Mathomu, L.M. and Madala, N.E. (2022). Putative Hydroxycinnamoyl-CoA: quinate/shikimate acid hydroxycinnamoyl transferase genes from *Bidens pilosa* responsible for biosynthesis of different forms of chlorogenic acids.

Presented in **Chapter 2**

Manuscript published:

Mathatha, K., Khwathisi, A., Ramabulana, A.T., Mwaba, I., Mathomu, L.M., Madala, N.E., 2022. Identification of putative acyltransferase genes responsible for the biosynthesis of homogenous and heterogenous hydroxycinnamoyl-tartaric acid esters from *Bidens pilosa*. *South African Journal of Botany* 149, 389-396.

<https://doi.org/10.1016/j.sajb.2022.06.008>

Presented in **Chapter 3**

Chapter 4: General conclusion

Executive summary

For decades, plants have been the backbone of complicated traditional herbal medicine system. Plants have been used by people and animals as a source of nutrients as well as medicine. Production of secondary metabolites by these plants is a characteristic that makes them attractive to both animals and humans. In plants, secondary metabolites play a role in defence mechanism and assist the plant to adapt to their immediate environment. Secondary metabolites are known to possess anti-diabetic, anti-malaria, anti-inflammatory and alleviate complications associated with obesity and cardiovascular diseases. Plants from the *Asteraceae* family are known to contain metabolites with nutraceutical properties. Plants such as *Helianthus annuus*, *Lactuca sativa*, *Chicorium intybus* and *Bidens pilosa* are some of the edible examples within the *Asteraceae* family known to exhibit interesting nutraceutical properties. *B. pilosa* adapt to almost every environmental condition, which makes it to be found in all parts of the world. As such, this plant has been used to manage and treat illnesses affecting humankind. Unique to this plant is the existence of large contingency of structurally diverse chlorogenic acids. *B. pilosa* is known to produce different structural hierarchies of chlorogenic acids (i.e., *mono-*, *di-*, *tri-* acyls). The biosynthetic pathway to produce the diverse array of chlorogenic acids in *B. pilosa* is not yet elucidated. It is known from other plants like *Helianthus annuus* that the production of chlorogenic acids is coded by hydroxycinnamoyl-CoA: quinate hydroxycinnamoyl transferase gene (HQT) and hydroxycinnamoyl-CoA: shikimate hydroxycinnamoyl transferase gene (HCT). Apart from the chlorogenic acids (quinic acid acyls), *B. pilosa* is known to also produce acyls of tartaric acid, also characterised by existence of structurally diverse isomers thereof. From other plants, the tartaric acid esters are coded by the hydroxycinnamoyl-CoA: tartaric hydroxycinnamoyl transferase (HTT) gene and, surprisingly, the gene encoding the tartaric acid esters/acyls from *B. pilosa* is also not known. It is therefore imperative that a study aimed at establishing/identification of the gene elements which are responsible for the diversification of chlorogenic acids and related compounds (such as tartaric acid acyls) in *B. pilosa* is conducted.

To achieve this, a Single Molecule Real Time (SMRT) sequencing approach was used to establish the full-length gene transcripts, in attempt to identify the acyl transferase

responsible for chlorogenic acids production in *B. pilosa*. The SMRT sequencing technique has brought a lot of improvement from the Sanger and Next generation sequencing such as generation of long reads which overcomes the challenges of sequence assembly synonymous with the short reads achieved by the former two sequencing approaches. Moreover, this technique allows detection of isoforms of a specific gene, caused by either inherited genetic code or alternative splicing events.

From the SMRT sequencing results, three HQT genes and one HCT gene responsible for production of wide array of chlorogenic acids and two HTT genes responsible for production of tartaric acid esters were identified through series of bioinformatics analyses of the sequences obtained through SMRT sequencing. All the identified genes contained the conserved regions that are found in already published acyltransferases, with the highly conserved DFGWG motif present in all transcripts identified herein. The second motif, the HXXXD motif showed a single amino acid variation from gene to gene, with HQT1 and HQT2 showing HTLSD motif and HQT3 having HTLAD motif, all of which are synonymous with the plants from the *Asteraceae* family. In HTT genes, the second motif identified in *B. pilosa* has never been recorded in literature. HTT1 and HTT2 showed to have HRVLD and HRVAD motif respectively. From these sequences, the open reading frames (ORFs) were computed, and these sequences can be used to design primers that can be used to amplify these genes in the future. Through multi sequence alignment and phylogenetic trees, the identified genes were also found to have similarities with the genes of other plants from the *Asteraceae* family.

In conclusion, the SMRT sequencing approach enabled identification of acyltransferases genes that plays a role in the biosynthesis of CGAs in *B. pilosa*. Bioinformatics tools were shown to be sufficient to annotate and characterise these genes. Through LC-MS analyses of randomly collected *B. pilosa* plants, the CGAs content of this plant were revisited, and these plants were found to produce structurally diverse CGAs compounds, suggesting that the identified genes are functional. Future studies should aim to clone these transcripts in plant systems that do not produce CGAs in attempt to enhance their nutraceutical attributes.

List of Abbreviations

AIDS	Acquired Immunodeficiency Syndrome
ARV	Antiretroviral
BLAST	Basic local alignment search tool
bp	Base Pairs
BPI	Base peak intensity
C'3H	3-cinnamate hydroxylase
C4H	Cinnamate-4-hydroxyase
CA	Cinnamic acid
CCS	Circular Consensus Sequencing
cDNA	Complementary deoxyribonucleic acid
CGA	Chlorogenic acid
CID	Collision induced dissociation
CQA	Caffeoylquinic acid
DDA	Data dependent acquisition
EICs	Extracted ion chromatograms
EI	Electron ionization
ESI	Electrospray ionization
ExPASy	Expert Protein Analysis System
FAO	Food and Agriculture Organization
FQA	Feruloylquinic acid
g	gram
HCA	Hydroxycinnamic acid

HCT	Hydroxycinnamoyl-CoA: shikimate hydroxycinnamoyl transferase gene
HIV	Human immunodeficiency virus
HQT	Hydroxycinnamoyl-CoA: quinic hydroxycinnamoyl transferase gene
HTT	Hydroxycinnamoyl-CoA: tartaric hydroxycinnamoyl transferase gene
INT	Integrase
LC-MS	Liquid chromatography mass spectrometer
MAFFT	Multiple Alignment Using Fast Fourier Transform
MRM	Multiple reaction monitoring
MSA	Multiple Sequence Alignment
MSI	Metabolomics standard initiative
MUSCLE	Multiple Sequence Comparison by Log- Expectation
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
NJ	Neighbour Joining
ORF	Open Reading Frame
Pac Bio	Pacific Biosciences
PAL	Phenylalanine ammonia lyase
PCR	Polymerase Chain Reaction
PMI	Percentage Matrix Index
PPP	Phenylpropanoid pathway
QA	Quinic acid
Rt	Retention time
RT	Reverse transcriptase
SGS	Second generation sequencing

SM	Secondary Metabolites
SMRT	Single molecule real time
TGS	Third generation sequencing
UHPLC-qTOF-MS/MS	Ultra high-performance liquid chromatography quadrupole time of flight
WHO	World Health Organisation
ZMW	Zero mode waveguide

List of Units

%	Percent
μl	Microlitre
$^{\circ}\text{C}$	degree Celsius
mM	millimolar
g/L	gram per litre
mL	millilitre
g	gram
$\mu\text{g/mL}$	microgram per millilitre
mg/mL	milligram per millilitre
μm	micrometre
mL/min	millilitre per minute
xg	times gravity
mm	millimetre
kV	kilovolt
v	voltage
L/h	litre per hour
pg/mL	picograms per millilitre
bw	band width

List of Figures and Tables

Table 1.1. Uses of *B. pilosa* and preparation method in Africa (Modified from (Arthur, 2012))

Table 1.2. Different plants that produce CGAs and the diseases they treat when prepared as traditional herbal medicine.

Figure 1.1. A flowering stem of *B pilosa*, showing leaves in green, flowers (White and Yellow) and sticky fruit (Black).

Fig 1.2. A picture of non-flavonoid polyphenols simple structures from **A.** phenols, **B.** benzoic acids, **C.** hydrolysable tannins, **D.** acetophenones, **E.** phenylacetic acids, **F.** cinnamic acids, **G.** coumarins, **H.** benzophenones and **I.** xanthenes.

Figure 1.3 Biosynthetic pathways of how CGAs are produced through mediation by HQT and HCT genes (Lepelley *et al.*, 2007).

Figure 1.4 Multiple sequence alignment of BAHD acyltransferases family genes including but not limited to *helianthus annuus* (QBM78938.1), *Cynara cardunculus var scolymus* (AFL93687.1), *Artemisia annua* (PWA39281.1), *Lactuca sativa* (XP_023733842.1), *mikania micrantha* (KAD5794970.1), *Lonicera japonica* (AEK80405.1), *Chicorium intybus* (ANN12610.1) and *Tanacetum cinerariifolium* (GEV77257.1). The first conserved motif which is HXXXG is circled in black and the second motif DFGWG is circled in red. Residues are grouped according to colours, for instance the same colour represent similar residues across all genes from different plants. The alignment was generated using MUSCLE. The position of the residue is shown by the number on the right.

Figure 1.5. Structural hierarchy of CGA.

Fig 1.6. Ribbon structure of HIV-1 INT with the Mg²⁺ cofactor and the (a) 3*trans*,5*trans*-diCQA, (b) 3*cis*,5*trans*-diCQA, (c) 3*trans*,5*cis*-diCQA, and (d) 3*cis*,5*cis*-diCQA ligand. Adapted from Makola *et al.*, (2016).

Figure 1.7. An Iso seq 2.0 workflow depicting how to go about full transcriptome sequencing from total RNA isolated from the subject of choice. The flow diagram shows how to do sequencing depending on what the desired outcome might be.

Table 2.1. This table highlights a summary of all the acyltransferases sequences identified in *B. pilosa*, the sequence length and the ORF for each gene. The sequences have been attached as supplementary files and an example of HQT1 ORF as determined through ExPASy is shown.

Table 2.2: List of chlorogenic acids (CGAs) molecules isolated from randomly sampled *B. pilosa* plants established through analysis by LC-QTOF-MS. The different colour shading indicates different structural hierarchy of the identified molecule.

Figure 2.1. Multiple sequence alignment of *B. pilosa* HQT1.

Multiple sequence alignment of *B. pilosa* HQT1 with its homologues from *Helianthus annuus* (QBM78938.1), *Cynara cardunculus* var *scolymus* (AFL93687.1), *Artemisia annua* (PWA39281.1), *Lactuca sativa* (XP_023733842.1), *Mikania micrantha* (KAD5794970.1), *Lonicera japonica* (AEK80405.1), *Chicorium intybus* (ANN12610.1) and *Tanacetum cinerariifolium* (GEV77257.1). Residues are grouped according to colours, for instance the same colour represents similar residues across all genes from different plants. The alignment was generated using the MUSCLE algorithm of the MEGA software. The position of the residue is shown by the number on the right.

Figure 2.2. The evolutionary history of HQT1 genes was computed using the Neighbor-Joining method (Saitou & Nei., 1987). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site.

Figure 2.3. Multiple sequence alignment of *B. pilosa* HQT2.

Multiple sequence alignment of *B. pilosa* HQT2 with its homologues from *Helianthus annuus* (XP_021990087.1), *Cynara cardunculus* var *scolymus* (P_024980016.1), *Artemisia annua* (PWA77292.1), *Lactuca sativa* (CAB4074763.1), *Mikania micrantha* (KAD2394232.1), *Chicorium intybus* (ANN12611.1), *Echinacea purpurea*

(QRI59127.1), *Cirsium arvense* (QQH14906.1) and *Crepidiastrum sonchifolium* (AZT78993.1). Residues are grouped according to colours, for instance the same colour represents similar residues across all genes from different plants. The alignment was generated using MUSCLE. The position of the residue is shown by the number on the right.

Figure 2.4. The evolutionary history of HQT2 genes was computed using the Neighbor-Joining method (Saitou & Nei., 1987). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site.

Figure 2.5. Multiple sequence alignment of *B. pilosa* HQT3.

Multiple sequence alignment of *B. pilosa* HQT3 with its homologues from *Helianthus annuus* (XP_0222026334.1), *Cynara cardunculus* var *scolymus* (P_024966573.1/ADL62855.1), *Artemisia annua* (PWA55118.1), *Lactuca sativa* (CAB4110182.1/CAB4074703.1), *Chicorium intybus* (ANN12811.1), *Crepidiastrum_sonchifolium* (AZT78993.1) and *Taraxacum_antungense* (QBQ52948.1). Residues are grouped according to colours, for instance the same colour represents similar residues across all genes from different plants. The alignment was generated using MUSCLE. The position of the residue is shown by the number on the right.

Figure 2.6. The evolutionary history of HQT3 genes was computed using the Neighbor-Joining method (Saitou & Nei., 1987). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site.

Figure 2.7. Multiple sequence alignment of *B. pilosa* HCT.

Multiple sequence alignment of *B. pilosa* HCT with its homologues from *Helianthus annuus* (XP_022018316.1), *Cirsium japonicum* (QQH14914.1), *Artemisia annua* (PWA37917.1), *Mikania micrantha* (KAD4889191.1), *Chicorium intybus*

(ANN12608.1) and *Echinacea purpurea* (QRI59128.1). Residues are grouped according to colours, for instance the same colour represents similar residues across all genes from different plants. The alignment was generated using MUSCLE. The position of the residue is shown by the number on the right.

Figure 2.8. The evolutionary history of HCT genes was computed using the Neighbor-Joining method (Saitou & Nei., 1987). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site.

Figure 2.9. Representative chemical structures of mono, di and tri- acylated CGAs compounds identified in *B. pilosa*.

Figure 2.10. Representative BPI of UHPLC-QTOF (ESI) chromatograms of methanol extracts of *B. pilosa* L.

Figure 3.1. HTT gene sequences from different *Asteraceae* family plants showing emphasis to the conserved motifs, (A) DFGWG and (B) HXXXD highlighted in red. The first two sequences (Plant_Black_Jack_HQ_transcript/109501 and Plant_Black_Jack_HQ_transcript/123562) are HTT1 and HTT2 genes respectively from *B. pilosa* L. obtained from Pacbio sequencing, and the other sequences were retrieved from NCBI (shown by NCBI accession numbers).

Figure 3.2. Neighbour joining phylogenetic analysis of HTT genes from *B. pilosa* and other HTT genes from *Asteraceae* family were retrieved from NCBI database. The evolutionary history was computed using the Neighbor-Joining method. The optimal tree with the sum of branch length = 4,11446114 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site. This analysis involved 15 amino acid sequences. Evolutionary analyses were conducted in MEGA X.

Figure 3.3. Representative UHPLC-qTOF-MS/MS chromatogram showing distribution patterns of tartaric acid derivatives in *B. pilosa*, with Y-axis showing peak intensity and X-axis showing retention time. The distribution pattern is as per the masses, at m/z 473 (A), at m/z 457 (B) and at m/z 487 (C).

Figure 3.4. Typical mass spectra of the fragmentation patterns of dicaffeoyltartaric acid (A), *p*-coumaroyl-caffeoyl tartaric acid (B), feruloyl caffeoyl tartaric acid (C).

Figure 3.5. Chemical structures of HTT derivatives from *B. pilosa*. (A) dicaffeoyltartaric acid, (B) *p*-coumaroyl caffeoyl tartaric acid, (C) Feruloyl-caffeoyl tartaric acid.

Chapter 1

Introduction and literature review

Introduction

1.1. Medicinal plants

Medicinal plants refer to a group of plants with medicinal properties (Jamshidi-Kia *et al.*, 2018). The World Health Organization (WHO) defines traditional medicinal plants as natural plant materials which are used at least or in the absence of industrial processing for the treatment of diseases at a local or regional scale (Tilburt and Kaptchuk., 2008). These plants have been used for disease management since time immemorial (Bartolome *et al.*, 2013; Rasool Hassan, 2012; Halberstein, 2005). Plants have been traditionally used in fighting against diseases such as diabetes, cancer and even those caused by viral infections (World Health Organization, 2013). These plants are a rich source of well sought-after compounds that can be used in the development of drugs (Jamshidi-Kia *et al.*, 2018). Countries such as China, Greece, Egypt, India, and Persia have commonly used medicinal plants as drugs and disinfectants (Jamshidi-Kia *et al.*, 2018). In fact, medicinal plants were the only available option in ancient times to help people recover from diseases (Jamshidi-Kia *et al.*, 2018; Halberstein, 2005). Demand for medicinal plants is increasing daily due to the progressive acceptance by communities. Different parts of these plants are used to bring about the desired effects i.e., leaves, stem and roots (Bartolome *et al.*, 2013; Halberstein, 2005).

The United Nations (UN) world health organisation estimated that about 5.6 billion people which is 80% of the human population, use plant-derived traditional medicine for primary health care (Shen *et al.*, 2012). Although synthetic drugs produced in laboratories are mainly used, the after-effects of some of these synthetic drugs are strenuous (Sehuda *et al.*, 2014). Therefore, medicinal plants are increasingly recognised, and the public is starting to trust these traditional herbs. Recently, biotechnological approaches have been applied to study the biosynthetic pathways responsible for production of pharmacological metabolites in plants (Kalakotla *et al.*, 2014). Several scientific reports have indicated plants from the *Asteraceae* family to produce structurally diverse chlorogenic acids (CGAs) compounds. This is an indication that plants from this family share common genetical information which is the hydroxycinnamoyl-CoA quinate/shikimate hydroxycinnamoyl transferase (HQT/HCT

gene in this case). HQT/HCT genes assist plants in *Asteraceae* family with the production of CGAs.

1.2. *Asteraceae* plants.

Asteraceae is the largest family of angiosperms and has a worldwide distribution that contains 25 000 to 30 000 identified plant species (Arthur, 2012). It is reported that in the families of weeds, *Asteraceae* shows the highest percentage (18.6%) of useful family members in traditional herbal medicine (Jayasundera *et al.*, 2021; Bonet and Valles., 2002). Plants from this family are used as food and medicine throughout the world (Koc *et al.*, 2015), hence they are regarded as economically important (Pza.sanbi.org, 2019; Tadesse, 2014). The production of structurally diverse secondary metabolites makes these plants nutraceutically important. Amongst other secondary metabolites, flavonoids and phenolic acid are biochemically important metabolites that are found in plants from this family (Sonnante *et al.*, 2010). As such, they are responsible for pharmacological attributes such as anticarcinogenic, anti-HIV (Nyamukuru *et al.*, 2017), anti-oxidative, cholesterol-lowering, bile expelling, hepatoprotective, and diuretic activities, as well as antifungal and antibacterial properties (Sonnante *et al.*, 2010).

Plants from this family can be used either as a nutrition source or for medical purposes. However, there are other plants in the same family that are used for both nutrition and medication. Examples of those that are used as a food source include sunflower (*Helianthus annuus*), artichoke (*Cynara cardunculus*), and lettuce (*Lactuca sativa*) (Heywood *et al.*, 2007). Interestingly, *Bidens pilosa* is a member of this family and is used as a food source and for medication (Pozharitskaya *et al.*, 2010). In the Venda region of South Africa, the leaves of *B. pilosa* are cooked and eaten as food. The very same leaves are crushed and applied to an open wound, and this could be a useful activity for people suffering from diabetes (Bartolome *et al.*, 2013).

1.3. *Bidens pilosa* is an underutilised plant species

B. pilosa is a member of the *Asteraceae* family which can be found all over the world due to its adaptive capabilities. It is a small, erect herb that is distributed throughout the year. It is characterised by its bright green leaves with serrated prickly edges. It produces small, yellow flowers and black fruit (Figure 1.1).

B. pilosa is a well-studied species amongst the *Asteraceae* family due to the biological activities reported from its extracts (Arthur, 2012). *B. pilosa* serves as a source of nutrition for both humans and animals (Gbashi *et al.*, 2016; Morton, 1962). The plant contains a diversity of interesting metabolites, including hydroxycinnamic acids, flavonoids and other compounds that are of great medicinal importance (Bartolome *et al.*, 2013). Beside its usage as a source of nutrition, it is also used as a resistance modifying agents against resistant bacteria and in treatment of over 40 diseases in folklore medicine (Borges *et al.*, 2013). Literature has reported that *B. pilosa* is potentially safe to use even at high dosages as a medicinal plant (Hong *et al.*, 2011). Some of its important roles include anti-microbial, anti-cancer (Shen *et al.*, 2018), anti-oxidative, anti-inflammatory, anti-allergic, anti-parasitic (Wink, 2012) and antidiabetic (Arthur, 2012; Mao *et al.*, 2010). More interestingly, *B. pilosa* has been shown to exhibit strong anti-HIV properties (Bartolome *et al.*, 2013), probably through its HCA (Hydroxycinnamic acid) derivatives which has been shown through computational studies to bind to HIV-1 integrase enzyme (Masike *et al.*, 2017; Makola *et al.*, 2016). Elsewhere, the preparations of this plant have been used synergistically with ARV for the treatment/management of HIV infection (Dhalla *et al.*, 2006; Furler *et al.*, 2003).

Preparation of this plant depends on the intended use and desired effect thereof. Each part of this plant i.e., leaves, stem, roots, seeds, and flowers are used as ingredients for traditional herbal medicine, either dry or fresh (Redl *et al.*, 1996). This plant can be taken as tea, juice or decoctions to manage or treat diseases in humans (Arthur, 2012). It can also be applied directly to wounds or burns. Animals use this plant through oral ingestion in order to treat and manage diseases affecting them (Arthur, 2012). The United Nations Food and Agriculture Organization (FAO) encouraged cultivation of this plant because it is easy to grow, safe, palatable, and edible (Mboya, 2019; Bartolome *et al.*, 2013).



Figure 1.1. A flowering stem of *B. pilosa*, showing leaves in green, flowers (White and Yellow) and sticky fruit (Black).

1.3.1. Nutraceutical importance of *B. pilosa*.

B. pilosa contains metabolites of nutraceutical significance as it possesses both nutritional and pharmaceutical values. It is used as herbal medicine worldwide to support and protect the liver, reduce inflammation, aid in weight loss, stimulate childbirth, increase urination and it is also used as anti-cancer (Shen *et al.*, 2018), anti-bacteria, anti-malarial, immunomodulatory agent (Bartolome *et al.*, 2013; Jimoh *et al.*, 2011; Mao *et al.*, 2010). This plant can be prepared as decoctions/infusions that a patient can take orally; however, it can also be prepared as a paste for external use such as treating wounds (Bartolome *et al.*, 2013). Moreover, pharmacological important properties have been discussed exhaustively in a study by Bartolome *et al.*, 2013. Literature has recorded extensive use of this plant in Africa to treat some conditions troubling people daily (Bartolome *et al.*, 2013). Table 1.1 below will show some of the conditions the plant is used to treat and part of the plant that is effective in treating a specific condition. The table will also show that each part of this plant has a respective function as nutraceutical components.

Table 1.1. Uses of *B. pilosa* and preparation method in Africa (Modified from Arthur, 2012)

Country	Plant part/Preparation	Treatment
South Africa	Concoction of leaf Suspension of powdered leaves	Abdominal pains Arthritis Malaria
Zimbabwe	Leaf tea	Hangover Headache Diarrhoea Stomach and mouth ulcers
Uganda	Crushed leaves Leaf decoction crushed leaves Decoction of leaf powder Herbal powder	Blood clotting agent Headache Ear infection Kidney problems Flatulence
Kenya	Ground leaves	Insecticides Colds/flu Urinary tract infections Infected wounds of skin Upper respiratory tract infections
Ivory coast	Crushed leaves	Jaundice/dysentery
Tanzania	Leaf sap	Burns
Nigeria	Powder from seeds Leaf extract	Anaesthetic Swollen spleens

1.4. Plant secondary metabolites (SMs)

Plant secondary metabolites are responsible for the pharmacological properties in plants and serve as important indicators to evaluate plants medicinal properties (Singh, 2015; Sehuda *et al.*, 2014). These metabolites are known to be the basis of many pharmaceutical drugs and plant derived drugs such as morphine (Yuan *et al.*, 2016; Tilburt and Kaptchuk, 2008). Secondary metabolites represent an interface between the plant and the surrounding environment by helping the plant to adapt to the forever changing environments (Yanqun *et al.*, 2020). Secondary metabolites play a vital role in plants' structural support, defence, survival, and adaptation (Mazid *et al.*, 2011). Upregulation and/or downregulation of these compounds in plants is affected by environmental factors and pathogen attack (Kundu and Vadassery, 2018; Soni *et al.*, 2015). Secondary metabolites are divided into classes following their structural diversity. Examples of such are phenylpropanoids, terpenoids, alkaloids, saponins, lipids and carbohydrates.

Secondary metabolites (SM) can be divided into two functional classes namely constitutive which is also called phytoanticipins (Morris *et al.*, 2020) and induced metabolites also known as phytoalexins. The SM classes are formed in response to biotic and abiotic factors (Mazid *et al.*, 2011). The phytoanticipins form a plant's first line of defence in response to pathogen attack while induced metabolites take at least 24 – 36 hours to form an active immune response (Maag *et al.*, 2015). Literature has reported that some metabolites exist as both phytoanticipins and phytoalexins. For instance, chlorogenic acids are found in healthy plants and can also be induced by environmental factors such as pathogen attack and therefore are classified as both phytoanticipins and phytoalexins (Mazid *et al.*, 2011). Phenols fall under the subclass of these secondary metabolites and are important antioxidants as part of human diet (Sehuda *et al.*, 2014). Phenols also possess anticancer, antidiabetic and many other properties.

The structure of phenols consists of an aromatic ring carrying one or more hydroxyl groups (Khadem and Marles, 2010). There are two main groups under the phenols which are flavonoids and non-flavonoids. The flavonoid group includes flavanones, flavones, dihydroflavonols, flavonols, flavan-3-ols, isoflavones, anthocyanidins,

proanthocyanidins and chalcones. The flavonoid group comprises compounds with a C₆-C₃-C₆ structural backbone (Khadem and Marles, 2010). The non-flavonoid polyphenols can be classified based on their carbon skeleton into the following subgroups: simple phenols, benzoic acids, hydrolysable tannins, acetophenones, phenylacetic acids, cinnamic acids, coumarins, benzophenones and xanthones shown in figure 1.2 below (Da Porto, 2021).

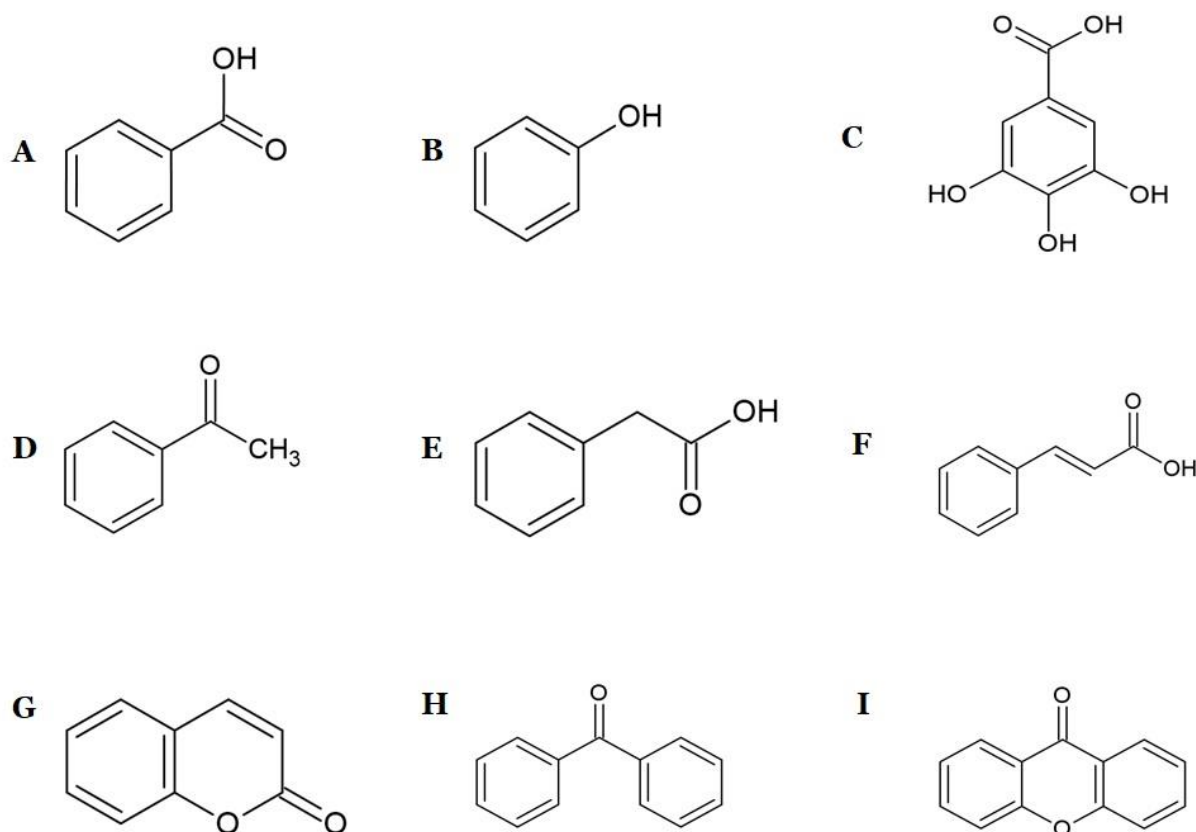


Figure 1.2. A picture of non-flavonoid polyphenols simple structures from **A.** phenol, **B.** benzoic acids, **C.** hydrolysable tannins, **D.** acetophenones, **E.** phenylacetic acids, **F.** cinnamic acids, **G.** coumarins, **H.** benzophenones and **I.** xanthones.

Phenolic acids have a carboxyl group attached or linked to a benzene ring. Two classes of phenolic acids can be distinguished through their structure: benzoic acid derivatives (i.e., hydroxybenzoic acids, C₆-C₁) and cinnamic acid derivatives (i.e., hydroxycinnamic acids, C₆-C₃) (Ramabulana *et al.*, 2020; Khadem and Marles, 2010). Hydroxycinnamic acids are carboxylic acids that belongs to the phenylpropanoids class.

1.5. The phenylpropanoid pathway

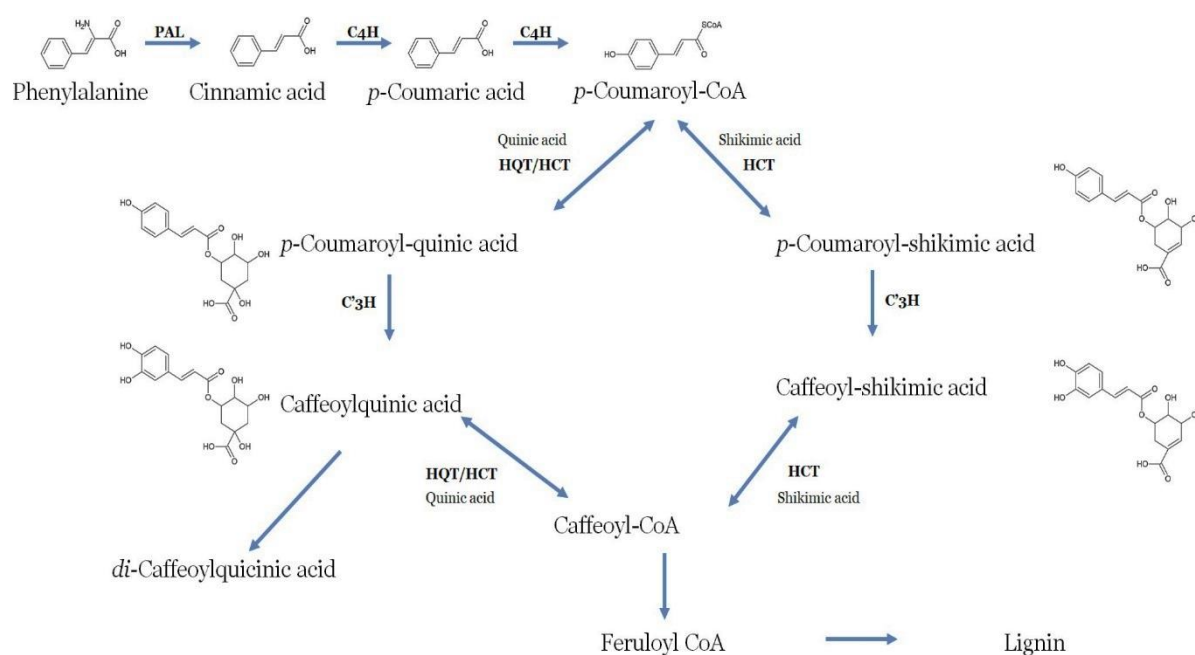


Figure 1.3. Biosynthetic pathways of how CGAs are produced through mediation by HQT and HCT genes (Lepelley *et al.*, 2007).

The phenylpropanoids serve as a rich source of secondary metabolites (Fraser and Chapple, 2011). The key enzymes in this pathway are hydroxycinnamoyl quinate/shikimate hydroxycinnamoyl transferase (HQT/HCT), coumaroyl-3-hydroxylase C₃H, phenylalanine ammonia-lyase (PAL) and cinnamic acid 4-hydroxylase (Fraser and Chapple, 2011). The genes encoding for these enzymes are referred to as acyltransferase genes and their role is to aid in attaching either caffeic acid, coumaroyl acid or ferulic acid to the quinic acid ring. CGAs are due to acylation of different cinnamic acids to a quinic acid. There are different biosynthetic routes in which CGAs are synthesised (Figure 1.3). However, the HQT gene mediated pathway is regarded as the primary pathway in which CGAs are synthesized (Liu *et al.*, 2018). Literature has showed that there is correlation between HQT gene and chlorogenic acid content because the silencing of HQT gene in tomato caused 98% decrease of chlorogenic acid yield in leaves (Niggeweg *et al.*, 2004). Contrary to the above, the overexpression of HQT gene in tomato was found to increase the yield of CGAs by 85%, an indication of the correlation between HQT gene expression and CGAs yield

(Liu *et al.*, 2018). Acyltransferases have also been shown to play a role in the biosynthesis of tartaric acid derivatives (e.g., Chicoric acid). These genes use a tartaric acid as a substrate in place of a quinic acid or shikimic acid. This phenomenon has been observed in purple coneflower (Fu *et al.*, 2021).

These transferases fall under the superfamily of acyl transferases which is known as BAHD acyltransferase family. The name of this family is deduced from the first letters of the first four genes that were biochemically characterised under this family i.e., Benzylalcohol-O-acetyltransferase (BEAT), Anthocyanin-O-hydroxycinnamoyltransferase (AHCT), anthrani-lateN-hydroxycinnamoyl/benzoyltransferase (HCBT) and deacetylindoline 4-O-acetyltransferase (DAT) (St-pierre and De luca, 2000). Enzymes of this family share several conserved amino acid sequences, the first one is the HXXXG motif, and it is located near the centre of the gene (St-pierre and De luca, 2000). The DFGWG is the second highly conserved motif of these genes, and it is located near the carboxyl terminal (St-pierre and De luca, 2000). All functionally characterised genes under this family contain these two conserved regions. Figure 1.4 shows an MSA of different acyltransferases and the conserved motifs are circled in a black and red colour. Sequences aligned herein are from NCBI as shown by their accession numbers in brackets. The sequences aligned are from the following plants, *Helianthus annuus* (QBM78938.1), *Cynara cardunculus var scolymus* (AFL93687.1), *Artemisia annua* (PWA39281.1), *Lactuca sativa* (XP_023733842.1), *mikania micrantha* (KAD5794970.1), *Lonicera japonica* (AEK80405.1), *Chicorium intybus* (ANN12610.1) and *Tanacetum cinerariifolium* (GEV77257.1).

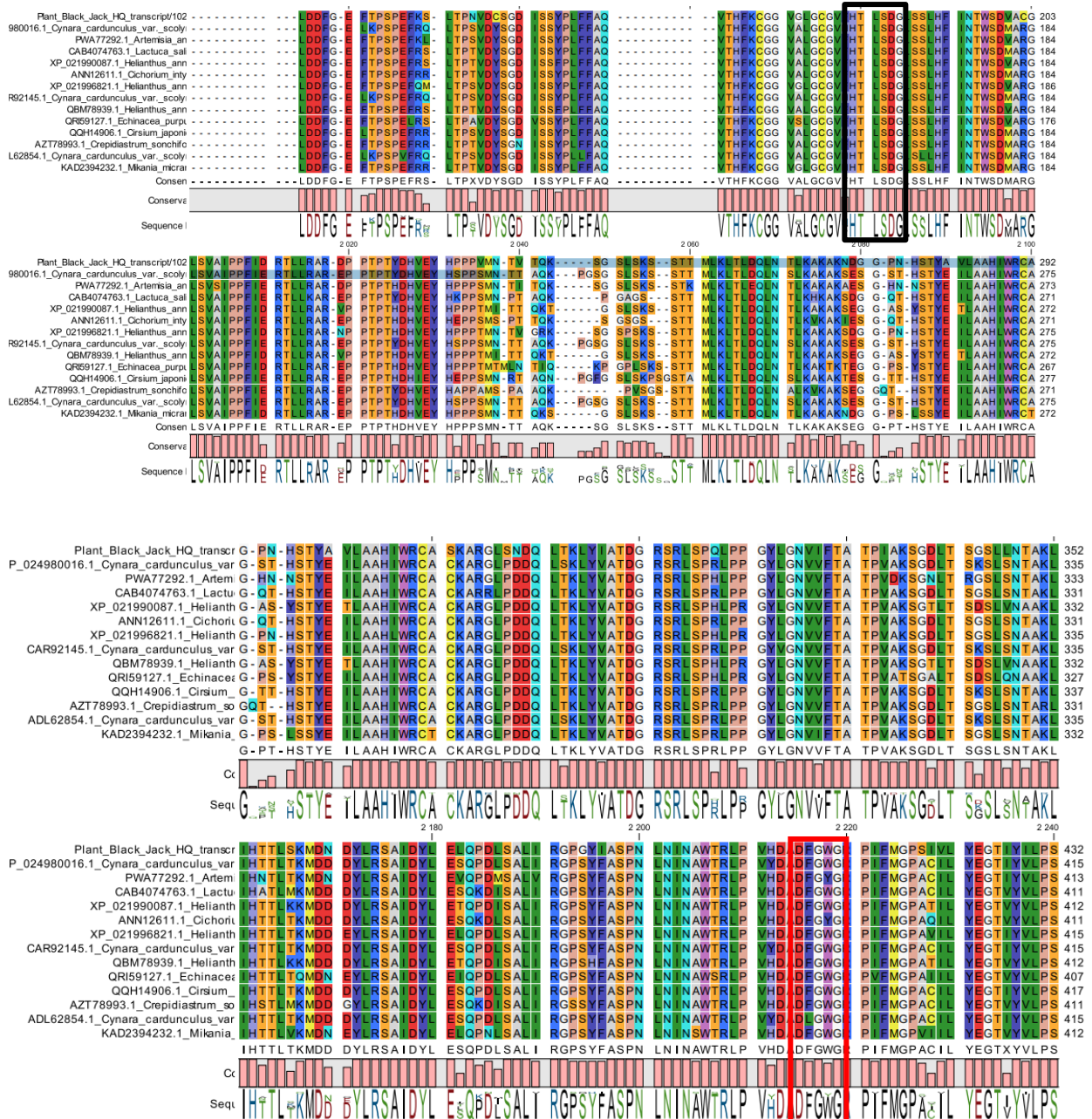


Figure 1.4. Multiple sequence alignment of BAHD acyltransferases family genes including but not limited to *helianthus annuus* (QBM78938.1), *Cynara cardunculus var scolymus* (AFL93687.1), *Artemisia annua* (PWA39281.1), *Lactuca sativa* (XP_023733842.1), *mikania micrantha* (KAD5794970.1), *Lonicera japonica* (AEK80405.1), *Chicorium intybus* (ANN12610.1) and *Tanacetum cinerariifolium* (GEV77257.1). The first conserved motif which is HXXXG is circled in black and the second motif DFGWG is circled in red. Residues are grouped according to colours, for instance the same colour represent similar residues across all genes from different plants. The alignment was generated using MUSCLE. The position of the residue is shown by the number on the right.

1.6. Chlorogenic Acids

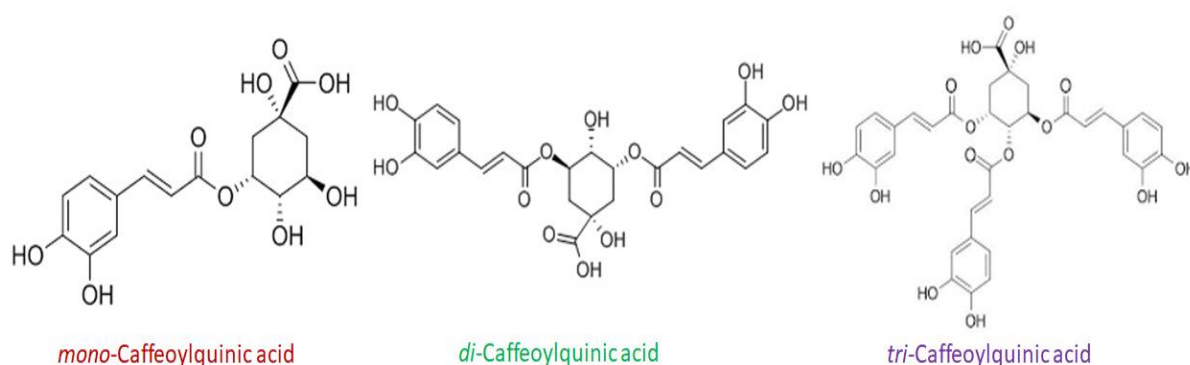


Figure 1.5. Structural hierarchy of CGA.

Chlorogenic acids are ester compounds formed between *trans*-hydroxycinnamic acids and quinic acids (Karaköse *et al.*, 2015; Jaiswal *et al.*, 2014; Jaiswal & Kuhnert, 2011). CGAs are a class of natural compounds which are significant in plants due to their heightened biological activities (Gbashi *et al.*, 2016; Arthur, 2012; Sonnante *et al.*, 2010). CGAs occur naturally and they act as plants defence mechanism and their role is to protect the plant from herbivores (Kutchan, 2001), fungi and bacteria (Lattanzio *et al.*, 2006). Elsewhere, CGAs producing plants have been shown to possess anti-diabetic (Sonnante *et al.*, 2010) and anti-HIV properties (Bartolome *et al.*, 2013). CGAs are absorbed by humans in their diet through the small intestine (Naveed *et al.*, 2017; Stalmach *et al.*, 2009). In the body, CGAs are known to help in the functioning of the liver, stomach, gallbladder and in lowering the secretion of glucose into the blood which in turn reduce the risk of heart disease and diabetes (type II) (Meng *et al.*, 2013). CGAs have different structural conformations (Figure 1.5), this is due to multi-acylation patterns on quinic acid, hence various structural hierarchy (Figure 1.5), from mono-acyl, di-acyl and tri-acyl (Ramabulana *et al.*, 2020).

To date, coffee has been shown to produce the largest and diverse composition of CGAs compounds (Lepelley *et al.*, 2007). However, other plants such as those from the *Asteraceae* family have been shown to contain large amounts of these compounds. There are many other plants in the *Asteraceae* family that are underutilised and have various medicinal properties (Table 1.2) such as *Taraxacum spp* that has been used to manage type 2 diabetes, hepatitis B and heart diseases

(Martinez *et al.*, 2015). *Vernonia fastigiata* has been shown to exhibit large amounts of CGAs (Masike *et al.*, 2017), and it has also been shown to be anti-bacterial (Erasto *et al.*, 2006), anti-viral (Bessong *et al.*, 2005) and anti-malarial (Njan *et al.*, 2008). *Cynara cardunculus* (Artichoke) and *Helianthus annuus* have been shown to also produce CGAs compounds and has been used for wound healing, anti-inflammatory, antihypertensive, cholesterol lowering, antifungal and anti-bacterial properties (Rauf *et al.*, 2020; Falã *et al.*, 2014). *Chicorium intybus* (chicory), another CGAs producing plant has been used for blood glucose lowering, anti-hepatotoxic, antioxidant (Bahmani *et al.*, 2015).

Table 1.2. Different plants that produce CGAs and the diseases they treat when prepared as traditional herbal medicine.

No.	Plants	Diseases	Reference
1.	<i>Taraxacum spp</i>	Diabetes II, Hepatitis B, Cardiovascular complications	(Martinez <i>et al.</i> , 2015)
2.	<i>Vermonia fasitigiata</i>	Anti-bacterial, anti-viral and anti-malaria	(Erasto <i>et al.</i> , 2006); (Bessong <i>et al.</i> , 2005); (Njan <i>et al.</i> , 2008)
3.	<i>Cynara cardunculus</i>	Anti-bacterial and antifungal	(Zhua <i>et al.</i> , 2005)
4.	<i>Helianthus annuus</i>	Wound healing, anti-inflammatory	(Poljšak <i>et al.</i> , 2020)
5.	<i>Chicorium intynus</i>	Antioxidant, glucose lowering	(Bahmani <i>et al.</i> , 2015)
6.	<i>Echinacea purpurea</i>	Respiratory infections	(Kumar and Ramaiah, 2011)

1.7. Pharmacological importance of CGA

1.7.1. Anti-HIV properties of CGAs

HIV/AIDS has infected more than 60 million people and caused at least 25 million deaths globally ever since its emergence in 1981 (Friedland, 2016). Developing countries are most prone to HIV/AIDS and have faced high rate of morbidity and mortality (Sharp and Hahn, 2011). The highest prevalence was recorded in the sub-Saharan Africa. To date, there is no cure for HIV/AIDS but anti-retroviral (ARV's) are used to suppress the viral load of this disease (UNAIDS, 2014). Some people suffer the side effects of these drugs (World Health Organization, 2013) and as consequence. alternative medicines are required to manage this disease. CGAs have been recorded to exhibit anti-HIV properties and they are safe for consumption, even at high dosages (Bartolome *et al.*, 2013; Hong *et al.*, 2011). HIV-1 has three important enzymes that are essential for its replication and subsequent infection. These enzymes are HIV-1 Protease (PROT), reverse transcriptase (RT), and integrase (INT) enzyme (Menéndez-Arias, 2010). The latter helps with the incorporation of viral DNA into the host cell genome after the reverse transcription of the viral RNA (Menéndez-Arias, 2010). This is a very important step in viral replication and makes anti-HIV-1 INT inhibitors a very interesting area of research. The HIV-1 INT has three domains, namely N terminal domain, catalytic core domain, and the C-terminal domain. The CCD has a conserved catalytic triad, the DDE motif with residues ASP64, ASP116, and GLU152. HIV-1 INT serves as the divalent metal (Mg^{2+} or Mn^{2+}) cofactor that deprotonates water for 3' end processing of viral cDNA (Menéndez-Arias, 2010). The di-caffeoylquinic acids (diCQAs) is a group of plants secondary metabolites, they have been found to have an irreversible inhibitory interaction with the HIV-1 INT catalytic core (Zhu *et al.*, 1999). The irreversible interaction that dicaffeoylquinic acids have with HIV-1 catalytic core makes them potential inhibitors (Makola *et al.*, 2016). Below is the HIV-1 INT enzyme docked with diCQA's (Figure 1.6).

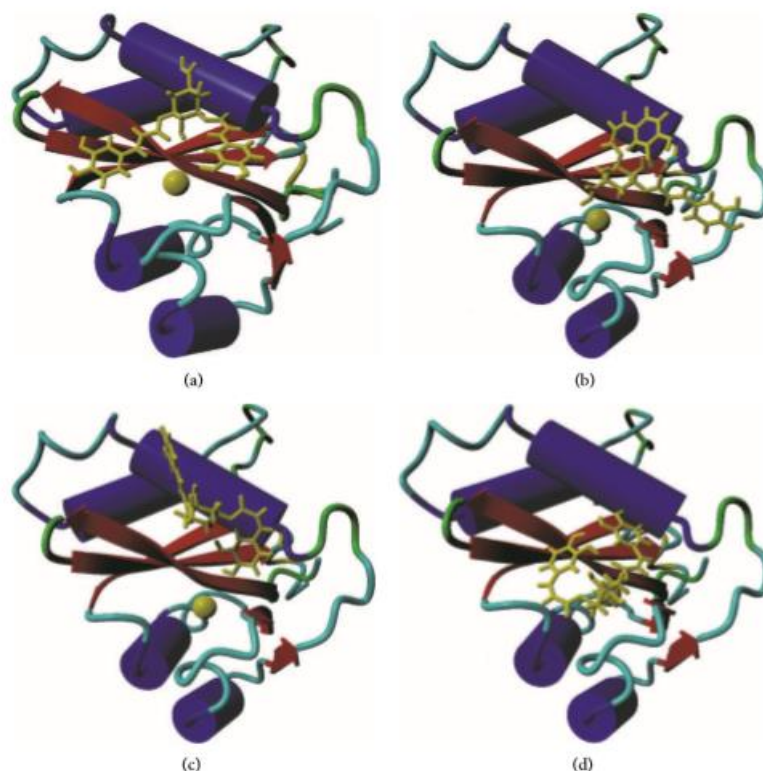


Fig 1.6. Ribbon structure of HIV-1 INT with the Mg²⁺ cofactor and the (a) *3trans,5trans*-diCQA, (b) *3cis,5trans*-diCQA, (c) *3trans,5cis*-diCQA, and (d) *3cis,5cis*-diCQA ligand. Adapted from Makola *et al.*, (2016).

1.7.2. Anti-diabetic Activity

Diabetes is now a global burden and, as HIV, it also has a huge impact on the economy. According to the International Diabetes Foundation, 382 million people were diagnosed with diabetes in 2013 and it has been predicted that by 2035 a number of affected people will rise to 592 million. Current oral antidiabetic drugs have unmet efficacy and undesirable side effects in patients often leading to lethal complications (Meier *et al.*, 2016)

B. pilosa have been shown to possess hypoglycemic activity in diabetic mice (Lai *et al.*, 2015). Polyynes from *B. pilosa* have been found to possess glucose-lowering activity (Lai *et al.*, 2015). Cytopyloyne which is also found in *B. pilosa* was found to have high glucose-lowering activity (Chang *et al.*, 2013). These cytopiloyne exert antidiabetic activities by regulating β -cell function (Lai *et al.*, 2015). Therefore,

continuing the search for new diabetes treatments is important. CGAs compounds have shown antidiabetic activity with no side effects (Lai *et al.*, 2015; Sonnante *et al.*, 2010). *B. pilosa* is known to be anti-HIV and anti-diabetic, these properties are due to the production of structurally diverse CGAs compounds, forms mono-, di- and tri-CQA (Ramabulana *et al.*, 2020; Mao *et al.*, 2010).

1.8. Decoding the genetics behind the chemistry of *B. pilosa*.

Although a lot has been reported about this plant, there is still a need to research and to understand the genetic makeup of this plant that enables it to produce different hierarchies of CGAs compounds. For instance, it is known that *B. pilosa* can produce a diverse array of CGAs, more than most plants even those that have been commercialised such as coffee. To understand the genetic makeup of this plant, high end throughput techniques such as Pacific biosciences (Pac Bio) Single Molecule Real Time (SMRT) sequencing can be used to decode the genetic codes of this plant. Advantage of this technique is that it does not need a reference genome to sequence the whole transcriptome of a plant. Oligo dT primers designed for this technique targets the poly A tail of the sequence. It is useful if there is not much information recorded in literature about the genetic makeup of a plant like *B. pilosa*.

1.9. SMRT sequencing

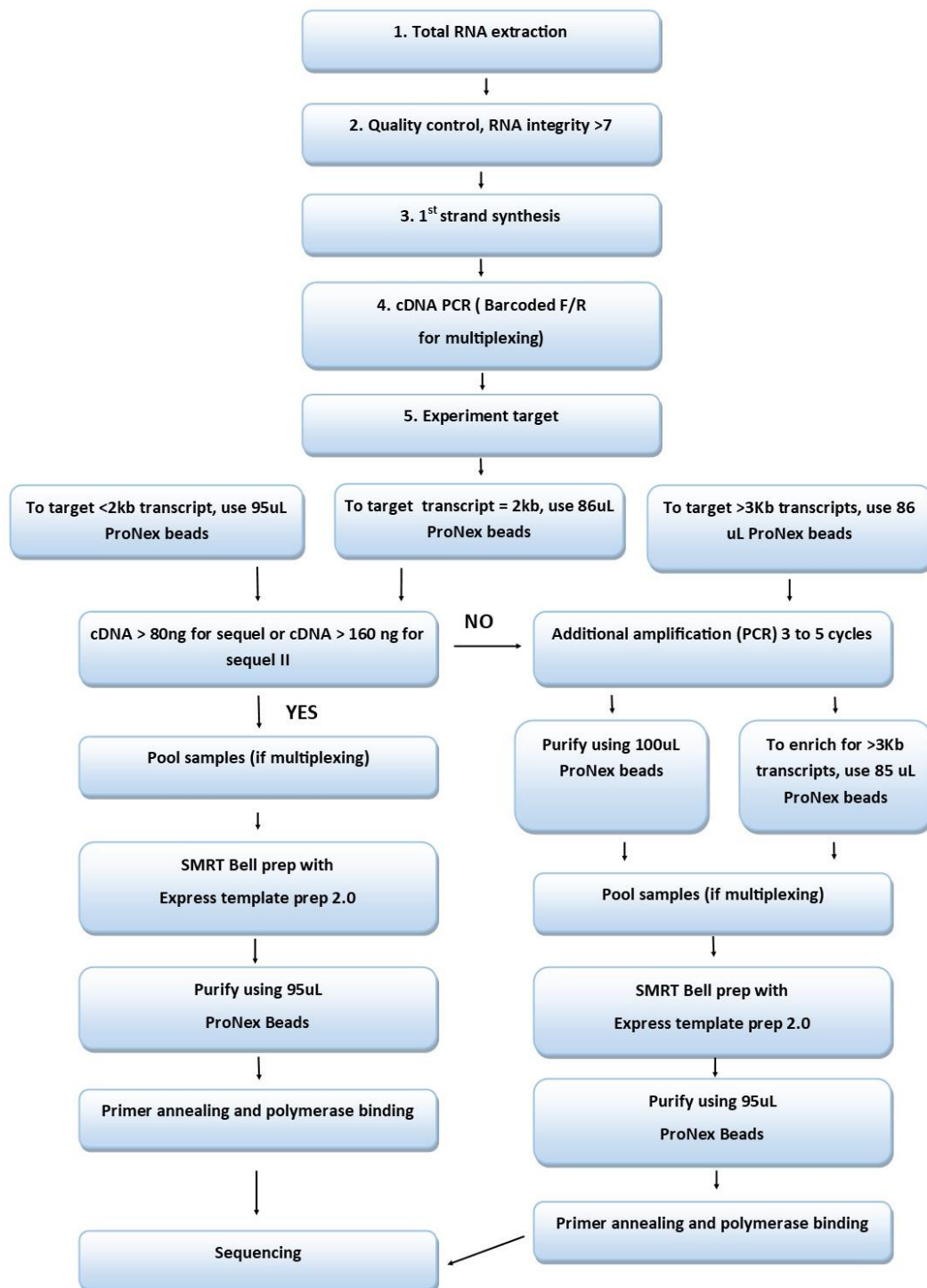


Figure 1.7. An Iso seq 2.0 workflow depicting how to go about full transcriptome sequencing from total RNA isolated from the subject of choice. The flow diagram shows how to do sequencing depending on what the desired outcome might be.

Second generation sequencing (SGS) offered vast improvements over Sanger sequencing (Ma *et al.*, 2019). However, they also have some shortfalls such as short read lengths which makes it difficult to assemble, detect isoforms and to determine complex genomic regions (Ma *et al.*, 2019; Rhoads and Au, 2015). Pac Bio developed the SMRT sequencing approaches which offers alternative means that overcome the SGS limitations. SMRT sequencing is regarded as the Third-generation sequencing (TGS) method because it overcomes most limitations of the second-generation sequencing (SGS) (Xu *et al.*, 2015). SMRT sequencing offers long read lengths and faster runs as compared to the SGS. Due to the ability of SMRT to produce long read length it is now possible to study larger genomes like that of humans (Dong *et al.*, 2015; Rhoads and Au, 2015). Also, isoforms can be identified using this very same technique. The process has four main components namely SMRT cell, zero mode waveguide (ZMW), polymerase and four fluorescent labelled nucleotides. Inside the cell is a ZMW which provides the smallest available volume for light detection (Rhoads and Au, 2015). Each ZMW contains a single polymerase embedded at the bottom to initiate replication. The fluorescent bases will be introduced into the SMRT cell and as they pass through the polymerase, they produce distinct colours that identify each base (Rhoads and Au, 2015). As the phosphate group is cleaved from the nucleotide, it takes away the fluorescing characteristic. The workflow of this technique is shown herein (Figure 1.7). In this study SMRT sequencing will be used to get the full-length transcriptome of *B. pilosa*, this will then be followed by usage of different bioinformatics tools such as NCBI (National Center for Biotechnology Information), BLAST (Basic Local Alignment Search Tool), ExPASy (Expert Protein Analysis System) etc. in order to identify all the acyltransferases in *B. pilosa* that are responsible for production of CGAs.

1.10. Study Rationale

Plants have been used to treat various diseases especially in developing countries. Recently, biotechnological approaches have been applied to study the biosynthetic pathways responsible to produce pharmacologically relevant metabolites in plants. For instance, HQT gene is known to code for the acyltransferase responsible for the production of CGAs and these compounds have been shown to exhibit anti-HIV and anti-diabetic activities. However, their composition varies depending on a plant species, and to date, coffee has been shown to produce the largest and most diverse composition of CGA compounds (Lepelley *et al.*, 2007). However, differences in CGAs composition between different coffee species has also been noted elsewhere (Lepelley *et al.*, 2007), an indication that these differences are genetically coded. Other plants such as those from the *Asteraceae* family such as *Vernonia fastigiata* and *B. pilosa* have been shown to contain large amounts of these compounds. Therefore, characterization of the HQT gene from these plants is significant to understand how they are able to produce different hierarchies of CGA compounds in large quantities. Thereafter, the identified genes can be cloned into expression vectors and large amount of these compounds can be produced through biotechnological means *in vitro*. The findings of this study may also help in enhancing nutraceutical value of other plants that do not produce chlorogenic acids.

1.11. Aim

To identify acyltransferases responsible for CGA production in *B. pilosa*.

1.12. Objectives

- i. To use SMRT sequencing technique to obtain the full-length transcriptome of *B. pilosa*.
- ii. To identify all acyltransferases from *B. pilosa* transcriptome that are found in *B. pilosa* using bioinformatics tools.
- iii. To identify HQT and HCT genes and their isoforms in *B. pilosa*
- iv. To Identify chlorogenic acids compounds produced by acyltransferases in *B. pilosa* through LC-MS
- v. To correlate specific acyltransferase gene with a class of metabolites identified through LC-MS.

References

Adedapo, A., Jimoh, F. and Afolayan, A. (2011). Comparison of the nutritive value and biological activities of the acetone, methanol and water extracts of the leaves of *Bidens pilosa* and *Chenopodium album*. *Acta Poloniae Pharmaceutica*, 68(1): 83-92.

Arthur, G.D., Naidoo, K.K. and Coopoosamy, R.M. (2012). *Bidens pilosa* L.: Agricultural and pharmaceutical importance. *Journal of Medicinal Plants Research*, 6(17): 3282-3281. <https://doi.org/10.5897/JMPR012.195>

Bahmani, M., Shahinfard, N., Rafieian-Kopaei, M., Saki, K., Shamsavari, S., Taherikalani, M., Ghafourian, S. and Baharvand-Ahmadi, B. (2015). Chicory: A review on ethnobotanical effects of *Cichorium intybus* L. *Journal of Chemical and Pharmaceutical Sciences*, 8(4): 672-682.

Bartolome, A.P., Villaseñor, I.M. and Yang, W.C. (2013). *Bidens pilosa* L. (Asteraceae): botanical properties, traditional uses, phytochemistry, and pharmacology. *Evidence-based complementary and alternative medicine*. <https://doi.org/10.1155/2013/340215>.

Bessong, P.O., Obi, C.L., Andréola, M., Rojas, L.B., Pouységu, L., Igumbor, E., Meyer, J.M., Quideau, S., Litvak, S. (2005). Evaluation of selected South African medicinal plants for inhibitory properties against Human Immunodeficiency Virus type 1 reverse transcriptase and integrase. *Journal of Ethnopharmacology*, 99: 83–91. <https://doi.org/10.1016/j.jep.2005.01.056>

Bonet, M.À. and Valles, J. (2003). Pharmaceutical ethnobotany in the Montseny biosphere reserve (Catalonia, Iberian Peninsula). General results and new or rarely reported medicinal plants. *Journal of Pharmacy and Pharmacology*, 55(2): 259-270. <https://doi.org/10.1211/002235702432>

Borges, C.C., Matos, T.F., Moreira, J., Rossato, A.E., Zanette, V.C. and Amaral, P.A. (2013). *Bidens pilosa* L. (Asteraceae): traditional use in a community of southern Brazil. *Revista Brasileira de Plantas Mediciniais*, 15: 34-40. <https://doi.org/10.1590/S1516-05722013000100004>

Chang, C.L.T., Liu, H.Y., Kuo, T.F., Hsu, Y.J., Shen, M.Y., Pan, C.Y. and Yang, W.C. (2013). Antidiabetic effect and mode of action of cytopiloyne. *Evidence-Based Complementary and Alternative Medicine*. <https://doi.org/10.1155/2013/685642>

Da Porto, A., Cavarape, A., Colussi, G., Casarsa, V., Catena, C. and Sechi, L.A. (2021). Polyphenols Rich Diets and Risk of Type 2 Diabetes. *Nutrients*, 13: 1445. <https://doi.org/10.3390/nu13051445>

Dhalla, S., Chan, K.J., Montaner, J.S. and Hogg, R.S. (2006). Complementary and alternative medicine use in British Columbia—a survey of HIV positive people on antiretroviral therapy. *Complementary Therapies in Clinical Practice*, 12(4): 242-248. <https://doi.org/10.1016/j.ctcp.2006.05.002>

Dong, L., Liu, H., Zhang, J., Yang, S., Kong, G., Chu, J.S., Chen, N. and Wang, D. (2015). Single-molecule real-time transcript sequencing facilitates common wheat genome annotation and grain transcriptome research. *BMC genomics*, 16(1):1-13. <https://doi.org/10.1186/s12864-015-2257-y>

Erasto, P.; Grierson, D.; Afolayan, A. (2006). Bioactive sesquiterpene lactones from the leaves of *Vernonia amygdalina*. *Journal of Ethnopharmacology*, 106: 117–120. <https://doi.org/10.1016/j.jep.2005.12.016>

Falã, P.L., Ferreira, C., Rodrigues, A.M. and Serralheiro, M.L., 2014. Studies on the molecular mechanism of cholesterol reduction by *Fraxinus angustifolia*, *Peumus boldus*, *Cynara cardunculus* and *Pterospartum tridentatum* infusions. *Journal of Medicinal plants research*, 8(1): 9-17. <https://doi.org/10.5897/JMPR2013.5273>

Fraser, C.M. and Chapple, C. (2011). The phenylpropanoid pathway in *Arabidopsis*. *The Arabidopsis Book/American Society of Plant Biologists*, 9. <https://doi.org/10.1199/tab.0152>

Friedland, G. (2016). Marking time in the global HIV/AIDS pandemic. *Jama*, 316(2): 145-146. <https://doi.org/10.1001/jama.2016.9006>

Fu, R., Zhang, P., Jin, G., Wang, L., Qi, S., Cao, Y., Martin, C. and Zhang, Y. (2021). Versatility in acyltransferase activity completes chicoric acid biosynthesis in purple coneflower. *Nature communications*, 12(1): 1-13. <https://doi.org/10.1038/s41467-021-21853-6>

Furler, M.D., Einarson, T.R., Walmsley, S., Millson, M. and Bendayan, R. (2003). Use of complementary and alternative medicine by HIV-infected outpatients in Ontario, Canada. *AIDS patient care and STDs*, 17(4), 155-168. <https://doi.org/10.1089/108729103321619764>

Gbashi, S., Njobeh, P., Steenkamp, P. and Madala, N. (2017). Pressurized hot water extraction and chemometric fingerprinting of flavonoids from *Bidens pilosa* by UPLC-tandem mass spectrometry. *CyTA-Journal of Food*, 15(2): 171-180. <https://doi.org/10.1080/19476337.2016.1230151>

Gbashi, S., Njobeh, P., Steenkamp, P., Tutu, H. and Madala, N. (2016). The effect of temperature and methanol–water mixture on pressurized hot water extraction (PHWE) of anti-HIV analogues from *Bidens pilosa*. *Chemistry Central Journal*, 10(1): 1-12. <https://doi.org/10.1186/s13065-016-0182-z>

Halberstein, R.A. (2005). Medicinal plants: historical and cross-cultural usage patterns. *Annals of epidemiology*, 15(9): 686-699. <https://doi.org/10.1016/j.annepidem.2005.02.004>

Heywood, V.H., Brummitt, R.K., Culham, A., & Seberg, O. (2007). *Asteraceae*. In: *Flowering Plant Families of the World*. New York, Firefly Books: 46-51.

Hong, C.E., Ji, S.T. and Lyu, S.Y. (2011). Absence of mutagenicity in three Nigerian medicinal plants-*Bidens pilosa*, *Cleistopholis patens* and *Tetrapleura tetraptera*. *Tropical Journal of Pharmaceutical Research*, 10(2). <https://doi.org/10.4314/tjpr.v10i2.66557>

Jaiswal, R., Kiprotich, J., & Kuhnert, N. (2011). Determination of the hydroxycinnamate profile of 12 members of the *Asteraceae* family. *Phytochemistry*, 72: 781 – 790. <https://doi.org/10.1016/j.phytochem.2011.02.027>

Jaiswal, R., Müller, H., Müller, A., Karar, M.G.E. and Kuhnert, N. (2014). Identification and characterization of chlorogenic acids, chlorogenic acid glycosides and flavonoids from *Lonicera henryi* L. (*Caprifoliaceae*) leaves by LC–MSn. *Phytochemistry*, 108: 252-263. <https://doi.org/10.1016/j.phytochem.2014.08.023>

Jamshidi-Kia, F., Lorigooini, Z. and Amini-Khoei, H. (2018). Medicinal plants: Past history and future perspective. *Journal of herbmed pharmacology*, 7(1). <https://doi.org/10.15171/jhp.2018.01>

Jayasundera, M., Florentine, S., Tennakoon, K.U. and Chauhan, B.S. (2021). Medicinal Value of Three Agricultural Weed Species of the *Asteraceae* Family: A Review. *Pharmacognosy Journal*, 13(1). <https://doi.org/10.5530/pj.2021.13.36>

Jimoh, F.O., Adedapo, A.A. and Afolayan, A.J. (2010). Comparison of the nutritional value and biological activities of the acetone, methanol and water extracts of the leaves of *Solanum nigrum* and *Leonotis leonorus*. *Food and Chemical Toxicology*, 48(3): 964-971. <https://doi.org/10.1016/j.fct.2010.01.007>

Kalakotla, S., Mohan, G.K., Rani, M.S., Divya, L. and Pravallika, P.L. (2014). Screening of *Saraca indica* (Linn.) medicinal plant for antidiabetic and antioxidant activity. *Der Pharmacia Lettre*, 6: 227-233.

Karaköse, H., Jaiswal, R., Deshpande, S. and Kuhnert, N. (2015). Investigation of the photochemical changes of chlorogenic acids induced by ultraviolet light in model systems and in agricultural practice with *Stevia rebaudiana* cultivation as an example. *Journal of Agricultural and Food Chemistry*, 63(13): 3338-3347. <https://doi.org/10.1021/acs.jafc.5b00838>

Khadem, S. and Marles, R.J. (2010). Monocyclic phenolic acids; hydroxy- and polyhydroxybenzoic acids: occurrence and recent bioactivity studies. *Molecules*, 15(11): 7985-8005. <https://doi.org/10.3390/molecules15117985>

Koc, S., Isgor, B.S., Isgor, Y.G., Shomali Moghaddam, N. and Yildirim, O. (2015). The potential medicinal value of plants from *Asteraceae* family with antioxidant defense

enzymes as biological targets. *Pharmaceutical biology*, 53(5): 746-751.
<https://doi.org/10.3109/13880209.2014.942788>

Kumar, K.M. and Ramaiah, S. (2011). Pharmacological importance of *Echinacea purpurea*. *International Journal of Pharma and Bio Sciences*, 2(4): 304-314.

Kundu, A. and Vadassery, J. (2019). Chlorogenic acid-mediated chemical defence of plants against insect herbivores. *Plant Biology*, 21(2): 185-189.
<https://doi.org/10.1111/plb.12947>

Kutchan, T.M. (2001). Ecological arsenal and developmental dispatcher. The paradigm of secondary metabolism. *Plant physiology*, 125(1): 58-60.
<https://doi.org/10.1104/pp.125.1.58>

Lai, B.Y., Chen, T.Y., Huang, S.H., Kuo, T.F., Chang, T.H., Chiang, C.K., Yang, M.T. and Chang, C.L.T. (2015). *Bidens pilosa* formulation improves blood homeostasis and β -cell function in men: a pilot study. *Evidence-based complementary and alternative medicine*. <https://doi.org/10.1155/2015/832314>

Lattanzio, V., Lattanzio, V.M. and Cardinali, A. (2006). Role of phenolics in the resistance mechanisms of plants against fungal pathogens and insects. *Phytochemistry: Advances in research*, 661(2): 23-67.

Lepelley, M., Cheminade, G., Tremillon, N., Simkin, A., Caillet, V. and McCarthy, J. (2007). Chlorogenic acid synthesis in coffee: An analysis of CGA content and real-time RT-PCR expression of HCT, HQT, C3H1, and CCoAOMT1 genes during grain development in *C. canephora*. *Plant Science*, 172(5): 978-996.
<https://doi.org/10.1016/j.plantsci.2007.02.004>

Liu, Q., Liu, Y., Xu, Y., Yao, L., Liu, Z., Cheng, H., Ma, M., Wu, J., Wang, W. and Ning, W. (2018). Overexpression of and RNA interference with hydroxycinnamoyl-CoA quinate hydroxycinnamoyl transferase affect the chlorogenic acid metabolic pathway and enhance salt tolerance in *Taraxacum antungense* Kitag. *Phytochemistry Letters*, 28: 116-123. <https://doi.org/10.1016/j.phytol.2018.10.003>

Ma, J.E., Jiang, H.Y., Li, L.M., Zhang, X.J., Li, H.M., Li, G.Y., Mo, D.Y. and Chen, J.P. (2019). SMRT sequencing of the full-length transcriptome of the Sunda pangolin (*Manis javanica*). *Gene*, 692: 208-216. <https://doi.org/10.1016/j.gene.2019.01.008>

Maag, D., Erb, M., Köllner, T.G. and Gershenzon, J. (2015). Defensive weapons and defense signals in plants: some metabolites serve both roles. *BioEssays*, 37(2): 167-174. <https://doi.org/10.1002/bies.201400124>

Makola, M.M., Dubery, I.A., Koorsen, G., Steenkamp, P.A., Kabanda, M.M., du Preez, L.L. and Madala, N.E. (2016). The effect of geometrical isomerism of 3, 5-dicaffeoylquinic acid on its binding affinity to HIV-integrase enzyme: A molecular docking study. *Evidence-Based Complementary and Alternative Medicine*. <https://doi.org/10.1155/2016/4138263>

Mao, D.J., Xie, J.F., Quan, G.M. and Zhang, J.E. (2010). Effects of *Bidens pilosa* aqueous extracts on germination and seedling growth of two pastures. *Journal of Foshan University (Natural and Science Edition)*, 28(5): 7-11. [https://doi.org/10.6165/tai.2009.54\(3\).255](https://doi.org/10.6165/tai.2009.54(3).255)

Martinez, M., Poirrier, P., Chamy, R., Prüfer, D., Schulze-Gronover, C., Jorquera, L. and Ruiz, G. (2015). *Taraxacum officinale* and related species—An ethnopharmacological review and its potential as a commercial medicinal plant. *Journal of Ethnopharmacology*, 169: 244-262. <https://doi.org/10.1016/j.jep.2015.03.067>

Masike, K., Khoza, B.S., Steenkamp, P.A., Smit, E., Dubery, I.A. and Madala, N.E. (2017). A metabolomics-guided exploration of the phytochemical constituents of *Vernonia fastigiata* with the aid of pressurized hot water extraction and liquid chromatography-mass spectrometry. *Molecules*, 22(8): 1200. <https://doi.org/10.3390/molecules22081200>

Mazid, M., Khan, T.A. and Mohammad, F., 2011. Role of secondary metabolites in defense mechanisms of plants. *Biology and medicine*, 3(2): 232-249. <https://doi.org/10.1111/j.1469-8137.1994.tb02968.x>

- Mboya, R.M. (2019). The Nutritional and Health Potential of Blackjack (*Bidens pilosa* L.): A Review—Promoting the Use of Blackjack for Food. *International Journal of Applied Research on Public Health Management (IJARPHM)*, 4(1): 47-66. <https://doi.org/10.4018/IJARPHM.2019010104>
- Meier, C., Schwartz, A.V., Egger, A. and Lecka-Czernik, B. (2016). Effects of diabetes drugs on the skeleton. *Bone*, 82: 93-100. <https://doi.org/10.1016/j.bone.2015.04.026>
- Menéndez-Arias, L. (2010). Retroviral Enzymes. *Viruses*, 2(5): 1181-1184. <https://doi.org/10.3390/v2051181>
- Meng, S., Cao, J., Feng, Q., Peng, J. and Hu, Y. (2013). Roles of chlorogenic acid on regulating glucose and lipids metabolism: a review. *Evidence-based complementary and alternative medicine*. <http://dx.doi.org/10.1155/2013/801457>
- Morris, H., Hietala, A.M., Jansen, S., Ribera, J., Rosner, S., Salmeia, K.A. and Schwarze, F.W. (2020). Using the CODIT model to explain secondary metabolites of xylem in defence systems of temperate trees against decay fungi. *Annals of botany*, 125(5): 701-720. <https://doi.org/10.1093/aob/mcz138>
- Morton, J.F., 1962. Spanish needles (*Bidens pilosa* L.) as a wild food resource. *Economic Botany*, 16(3): 173-179. <https://doi.org/10.1007/BF02860036>
- Mudau, S.P., Steenkamp, P.A., Piater, L.A., De Palma, M., Tucci, M., Madala, N.E. and Dubery, I.A. (2018). Metabolomics-guided investigations of unintended effects of the expression of the hydroxycinnamoyl quinate hydroxycinnamoyltransferase (hqt1) gene from *Cynara cardunculus* var. *scolymus* in *Nicotiana tabacum* cell cultures. *Plant Physiology and Biochemistry*, 127: 287-298. <https://doi.org/10.1016/j.plaphy.2018.04.005>
- Naveed, M., Hejazi, V., Abbas, M., Kamboh, A.A., Khan, G.J., Shumzaid, M., Ahmad, F., Babazadeh, D., FangFang, X., Modarresi-Ghazani, F. and WenHua, L. (2018). Chlorogenic acid (CGA): A pharmacological review and call for further research. *Biomedicine & Pharmacotherapy*, 97: 67-74. <https://doi.org/10.1016/j.biopha.2017.10.064>

Niggeweg, R., Michael, A.J. and Martin, C. (2004). Engineering plants with increased levels of the antioxidant chlorogenic acid. *Nature biotechnology*, 22(6): 746-754. <https://doi.org/10.1038/nbt966>

Njan, A.A., Adzu, B., Agaba, A.G., Byarugaba, D., Díaz-Llera, S., Bangsberg, D.R. (2008). The analgesic and antiplasmodial activities and toxicology of *Vernonia amygdalina*. *Journal of Medicinal Food*, 11(3): 574–581. <https://doi.org/10.1089/jmf.2007.0511>

Nyamukuru, A., Tabuti, J.R., Lamorde, M., Kato, B., Sekagya, Y. and Aduma, P.R. (2017). Medicinal plants and traditional treatment practices used in the management of HIV/AIDS clients in Mpigi District, Uganda. *Journal of Herbal Medicine*, 7: 51-58. <https://doi.org/10.1016/j.hermed.2016.10.001>

Poljšak, N., Kreft, S. and Kočevar Glavač, N. (2020). Vegetable butters and oils in skin wound healing: Scientific evidence for new opportunities in dermatology. *Phytotherapy research*, 34(2): 254-269. <https://doi.org/10.1002/ptr.6524>

Pozharitskaya, O.N., Shikov, A.N., Makarova, M.N., Kosman, V.M., Faustova, N.M., Tesakova, S.V., Makarov, V.G. and Galambosi, B. (2010). Anti-inflammatory activity of a HPLC-fingerprinted aqueous infusion of aerial part of *Bidens tripartita* L. *Phytomedicine*, 17(6): 463-468. <https://doi.org/10.1016/j.phymed.2009.08.001>

Pza.sanbi.org. (2019). "Asteraceae". *Plantz Africa*. [online] Available at: <http://pza.sanbi.org/asteraceae> [Accessed 21 Feb. 2019].

Ramabulana, A.T., Steenkamp, P.A., Madala, N.E. and Dubery, I.A. (2020). Profiling of altered metabolomic states in *Bidens pilosa* leaves in response to treatment by methyl jasmonate and methyl salicylate. *Plants*, 9(10): 1275. <https://doi.org/10.3390/plants9101275>

Rasool Hassan, B.A. (2012). Medicinal plants (importance and uses). *Pharmaceutica Analytica Acta*, 3(10): 2153-2435. <http://dx.doi.org/10.4172/2153-2435.1000e139>

Rauf, S., Ortiz, R., Shehzad, M., Haider, W. and Ahmed, I. (2020). The exploitation of sunflower (*Helianthus annuus* L.) seed and other parts for human nutrition, medicine and the industry. *Helia*, 43(73): 167-184. <https://doi.org/10.1515/helia-2020-0019>

Redl, K., Breu, W., Davis, B. and Bauer, R. (1994). Anti-inflammatory active polyacetylenes from *Bidens campylotheca*. *Planta Medica*, 60(1): 58-62. <https://doi.org/10.1055/s-2006-959409>

Rhoads, A. and Au, K.F. (2015). PacBio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 13(5): 278-289. <https://doi.org/10.1016/j.gpb.2015.08.002>

Sehuda, K., Belgin, S., Yasemin, G. and Ozlem, Y. (2014). The Potential medicinal value of plants from *Asteraceae* family with Antioxidant defence Enzymes as biological Target. *Pharmaceutical Biology*, 53(5). <https://doi.org/10.3109/13880209.2014.942788>

Sharp, P.M. and Hahn, B.H. (2011). Origins of HIV and the AIDS pandemic. *Cold Spring Harbor perspectives in medicine*, 1(1): 6841. <https://doi.org/10.1101/cshperspect.a006841>

Shen, T., Li, G.H., Wang, X.N. and Lou, H.X. (2012). The genus *Commiphora*: a review of its traditional uses, phytochemistry and pharmacology. *Journal of ethnopharmacology*, 142(2): 319-330. <https://doi.org/10.1016/j.jep.2012.05.025>

Singh, R. (2015). Medicinal plants: A review. *Journal of Plant Sciences*, 3(1): 50-55. <https://doi.org/10.11648/j.jps.s.2015030101.18>

Soni, U., Brar, S. and Gauttam, V.K. (2015). Effect of seasonal variation on secondary metabolites of medicinal plants. *International Journal of Pharmaceutical Sciences and Research*, 6(9): 3654-3662. <https://doi.org/10.4103/0257-7941.179869>

Sonnante, G., D'Amore, R., Blanco, E., Pierri, C.L., De Palma, M., Luo, J., Tucci, M. and Martin, C. (2010). Novel hydroxycinnamoyl-coenzyme A quinate transferase

genes from artichoke are involved in the synthesis of chlorogenic acid. *Plant Physiology*, 153(3): 1224-1238. <https://doi.org/10.1104/pp.109.150144>

Stalmach, A., Mullen, W., Barron, D., Uchida, K., Yokota, T., Cavin, C., Steiling, H., Williamson, G. and Crozier, A. (2009). Metabolite profiling of hydroxycinnamate derivatives in plasma and urine after the ingestion of coffee by humans: identification of biomarkers of coffee consumption. *Drug Metabolism and Disposition*, 37(8): 1749-1758. <https://doi.org/10.1124/dmd.109.028019>

St-Pierre, B. and De Luca, V. (2000). Origin and diversification of the BAHD superfamily of acyltransferases involved in secondary metabolism. *Recent advances in phytochemistry*, 34: 285-315.

Tadesse, M. (2014). How to study the *Asteraceae* (Compositae) with special reference to the *Asteraceae* of fee. *Ethiopian Journal of Biological Sciences*, 13: 91-101.

Tilburt, J.C. and Kaptchuk, T.J. (2008). Herbal medicine research and global health: an ethical analysis. *Bulletin of the World Health Organization*, 86: 594-599. <https://doi.org/10.2471/BLT.07.042820>

UNAIDS. (2014). "UNAIDS Fact Sheet". *Kampala: UNAIDS*.

Wink, M. (2012). Medicinal plants: a source of anti-parasitic secondary metabolites. *Molecules*, 17(11): 12771-12791. <https://doi.org/10.3390/molecules171112771>

World Health Organization. (2013). Consolidated Guidelines on the Use of Antiretroviral Drugs for Treating and Preventing HIV Infection: Recommendations for a Public Health Approach. *Geneva*.

Xu, Z., Peters, R.J., Weirather, J., Luo, H., Liao, B., Zhang, X., Zhu, Y., Ji, A., Zhang, B., Hu, S. and Au, K.F. (2015). Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of *Salvia miltiorrhiza* and tanshinone biosynthesis. *The Plant Journal*, 82(6): 951-961. <https://doi.org/10.1111/tpj.12865>

Yuan, H., Ma, Q., Ye, L. and Piao, G. (2016). The traditional medicine and modern medicine from natural products. *Molecules*, 21(5): 559.

<https://doi.org/10.3390/molecules21050559>

Zhu, K., Cordeiro, M.L., Atienza, J., Robinson Jr, W.E. and Chow, S.A., 1999. Irreversible inhibition of human immunodeficiency virus type 1 integrase by dicaffeoylquinic acids. *Journal of Virology*, 73(4): 3309-3316.

<https://doi.org/10.1128/jvi.73.4.3309-3316.1999>

Shen, Y., Sun, Z., Shi, P., Wang, G., Wu, Y., Li, S., Zheng, Y., Huang, L., Lin, L., Lin, X. and Yao, H. (2018). Anticancer effect of petroleum ether extract from *Bidens pilosa* L and its constituent's analysis by GC-MS. *Journal of ethnopharmacology*, 217: 126-133.

<https://doi.org/10.1016/j.jep.2018.02.019>

Chapter 2

Application of SMRT sequencing approach for identification of putative hydroxycinnamoyl-CoA: quinate/shikimate acid hydroxycinnamoyl transferase genes from *Bidens pilosa* L. responsible for biosynthesis of different forms of chlorogenic acids

K Mathatha¹, I Mwaba¹, LM Mathomu¹, AR Ndhlala², NE Madala^{1*}

¹Department of Biochemistry and Microbiology, Faculty of Science, Agriculture and Engineering, University of Venda, Private Bag X5050, Thohoyandou, Limpopo 0950, South Africa.

²Green Biotechnologies Research Centre of Excellence, School of Agricultural and Environmental Sciences, University of Limpopo, Republic of South Africa

*To whom correspondence should be addressed: NE Madala
(ntaka.madala@univen.ac.za)

Abstract

Bidens pilosa L. is an underutilised plant that serves both as an ingredient for traditional herbal medicine and as a food. This plant has received a lot of attention as of late because of its chemistry which comprises chemicals from various classes, with those from phenylpropanoid (e.g., chlorogenic acids) existing in abundance. *B. pilosa* L. produces chemically diverse chlorogenic acids (CGAs) compounds, characterised mainly by isomers thereof. The enzymes involved in their biosynthesis of chlorogenic acids have been decoded and a characterised in many plants. However, genes that play a role in the biosynthesis of chlorogenic acids in *B. pilosa* L. are not yet described and, as such, the aim of the current study is to identify different isoforms of the Hydroxycinnamoyl-CoA: quinate/shikimate acid hydroxycinnamoyl transferase (HQT/HCT) genes that play a role in the diversification of chlorogenic acids in this plant. Herein, a robust gene sequencing technology through Single Molecule Real Time (SMRT) approach was applied to establish the gene sequences encoding for transferases responsible for CGAs production in *B. pilosa* L. Sequence homology of the established genes was evaluated by means of multiple sequence alignment against already published orthologous genes from closely related plants. Thereafter, phylogeny trees were also constructed to graphically display the taxonomical relationship with other chlorogenic acids producing plants. To further demonstrate the functionality of these genes in plants, chlorogenic acids profiling was also carried out using Liquid chromatography quadrupole time of flight mass spectrometer (LC-QTOF-MS). The understudied *B. pilosa* L. plants were found to produce isomeric forms of *mono*, *di* and *tri-acylated* CGAs. Future work should involve expression of the identified genes in other non-chlorogenic acids producing plants to enhance their nutraceutical value.

Keywords: *Bidens pilosa* L., Chlorogenic acid, HQT gene, Isomers, LC-QTOF-MS, Phenylpropanoid, SMRT sequencing, PacBio.

2.1 Introduction

Bidens pilosa L. from the *Compositae* family is an underutilised, well-known species owing to nutraceutical activities reported from its extracts (Bartolome *et al.*, 2013; Arthur, 2012). *Bidens pilosa* grows all year round in most areas globally and has been shown to treat over 40 diseases in folklore medicine (Borges *et al.*, 2013). This plant has been shown to be effective in treating malaria (Tobinaga *et al.*, 2009), diabetes (Sonnante *et al.*, 2010), hypertension and obesity (Gökçen and Şanlıer, 2019), bacterial infection (Seca and Pinto, 2019), wounds healing and gastrointestinal sickness (Arthur, 2012). *Bidens pilosa* L. is known for its wide array of phytochemical constituents (Ramabulana *et al.*, 2020) which includes phenylpropanoids, hydroxycinnamic acids (HCA) derivatives, alkaloids, flavonoids, aliphatic (Xuan and Khanh, 2016). Amongst phenylpropanoids produced by this plant are chlorogenic acids (CGAs) which are of great medicinal value (Xuan and Khanh, 2016; Ramabulana *et al.*, 2020). Chlorogenic acids includes all the hydroxycinnamic acids derivatives including caffeoyl-, feruloyl-, dicaffeoyl- and coumaroylquinic acids (Cheynier *et al.*, 2010) and other cinnamic acid conjugates other than those of quinic acid (Ncube *et al.*, 2014).

Chlorogenic acids are produced through the phenylpropanoid pathway (Sonnante *et al.*, 2010), Production of CGA compounds has been associated with the activity of various acyltransferases of the phenylpropanoid pathway in other plants (Comino *et al.*, 2009; Moglia, 2016). Hydroxycinnamoyl-CoA: quinate/shikimate acid hydroxycinnamoyl transferase (HQT/HCT) have been identified in different plants such as dandelion (Liu *et al.*, 2019), chicory (Legrand *et al.*, 2016), sunflower (Cheevarungnapakul, 2019) and *Arabidopsis thaliana* (Hoffmann *et al.*, 2004). These enzymes are members of the superfamily called the **BAHD** acyl transferase family and they have been shown to play a role in the production of wide spectrum of metabolites (St-pierre and De luca, 2000). HQT genes expression has been correlated with CGA accumulation, for instance, overexpression of HQT gene in tomato correlated with the increased yield of CGA compounds (Niggeweg *et al.*, 2004). This then shows that these genes play a vital role in the production of CGA compounds.

Bidens pilosa L. produces mono-, di-, tri- caffeoyl- quinic acid, a characteristic that is unique in the plant's chemistry as most plants only produce one form (structural hierarchy) of these compounds (Ramabulana *et al.*, 2020). Production of these compounds by *B. pilosa* L. makes this plant very interesting in order to understand how all these forms of metabolites are produced in this plant. The production of the structurally diverse CGAs is an indication that the genetic makeup of *Bidens pilosa* L. might possess the genes with interesting regulatory and structural elements responsible for CGAs diversification observed in this plant. However, in *B. pilosa* L. the genes responsible for CGAs production have not yet been identified and characterised. Therefore, decoding the gene composition of *B. pilosa* L. might lead to unearthing of the novel genes encoding for the wide spectrum of CGA compounds that *B. pilosa* L. produces.

In other plants, the genes encoding acyltransferase genes responsible for CGAs production have been identified through the next generation sequencing (NGS) technique (Kim *et al.*, 2013). The NGS approach has multiple challenges such as generation of short read length, which leads to subsequent mistakes during sequence assembly and its inability to distinguish between isoforms (Malar *et al.*, 2019). Single molecule real time (SMRT) sequencing approach is a new technique that results in a full-length transcriptome and has been superior in overcoming challenges associated with NGS technique (Rhoads and Au, 2015). SMRT sequencing was used to get insights into disease resistance and CGA synthesis in *Solanum melongena* L (eggplant) (Li *et al.*, 2021). SMRT sequencing also revealed that the inability of *Coffea humblotiana* to produce caffeine is due to the absence of caffeine synthase gene that converts theobromine into caffeine (Raharimalala *et al.*, 2021). SMRT was used to reveal the transcriptome of *Eucommia ulmoides* during leaf growth and further correlate it to the metabolite content during that stage (Li *et al.*, 2019). The above few examples are a highlight that SMRT sequencing has been applied and shown to help understand sophisticated molecular phenomenon.

For identification of the CGA compounds in plants, liquid chromatography in combination with mass spectrometry (LC-MS) technique has been used successfully (Clifford *et al.*, 2003). However, the structural diversity of these metabolites poses a

serious analytical challenge because they produce very similar MS signals, which makes it impossible to distinguish between them. Regardless of these challenges, LC-MS is still a golden technique that is used in identifying the wide array of metabolites in plants (Zheng *et al.*, 2017; Ncube *et al.*, 2014). In this study, the full-length transcriptome of HQT/HCT genes from *B. pilosa* L. was achieved through SMRT sequencing. Bioinformatics tools were further used in identifying HQT/HCT genes in *B. pilosa* L. Optimised LC-MS based on collision induced dissociation approach was also used to identify the various structural hierarchies and isomers of the CGA compounds produced by *Bidens pilosa* plant in attempt to correlate the composition of these metabolites with the presence of the identified genes.

2.2. Methodology

2.2.1. Total RNA isolation

Bidens Pilosa L. plants were grown at the University of Venda greenhouse (22.9761° S, 30.4465° E). This area is regarded as semi-arid and is characterised by low rainfall as well as high temperatures. Four weeks old plants were taken to Inqaba biotech in Pretoria, South Africa in separate soil pots. Total RNA was extracted from 1 gram leaf material. *Quick* – RNA plant mini prep kit was used for extraction of total RNA following manufacturer’s protocol without any modification. Briefly, whole plants were cut with a sterile surgical blade to small pieces. Bashing beads with lysis buffer were used for extraction of total RNA. Zymo-spin IIICG columns were used to precipitate total RNA. The Zymo-spin IICR column was used to filter unwanted biological components in the tube. RNA wash buffer was used to wash RNA and DNase/Rnase free water was used to resuspend RNA. Total RNA was quantified using Nanodrop 2000c spectrophotometer (Thermofischer Scientific, USA).

2.2.2. cDNA library construction and SMRT sequencing

Complementary DNA (cDNA) Library preparation for sequencing was done using the Iso-Seq™ Express Template Preparation for Sequel® and Sequel II Systems’ procedure and checklist as recommended by the manufacturer. Briefly, total RNA (600ng) was reverse transcribed into cDNA using a NEBNext® Single Cell cDNA Synthesis and Amplification Module and Iso-Seq Express Oligo Kit that was optimized for generation of good quality full-length cDNAs. NEBNext Single Cell RT Primer Mix was used for first strand synthesis. NEBNext Single Cell RT Enzyme Mix (Oligo dT) was used in the reverse transcription process of cDNA targeting the poly(A) tail structure together with Iso-Seq Express Template Switching Oligo. The obtained full-length cDNA was amplified by PCR. The amplified product was purified by Pacific Biosciences (PB) magnetic beads and quantified. The amplified cDNA fragments were amplified by PCR again and the full-length cDNA was purified by PB magnetic beads. After the library was constructed, Qubit 2.0 was used for accurate DNA quantification. Then Agilent 2100 was used to detect the library size. Pacbio Sequel II platform was used to sequence *B. pilosa* transcriptome after the library was qualified to ensure the

library meets the minimum requirements for sequencing as reported elsewhere (Zhang *et al.*, 2020).

2.2.3. Identification of potential HQT/HCT genes from *B. pilosa* L.

Genes in phenylpropanoid pathway were retrieved from the National Centre for Biotechnology information (NCBI) database (<https://www.ncbi.nlm.nih.gov/biosystems/493811>), and in-house database was created using this dataset of genes. A BLAST+ (Basic Local Alignment Search Tool) command line was used to search for the same genes in the *B. pilosa* L. full transcriptome sequences. Results were then subjected to the filtering process as follows: *Helianthus annuus* nucleic acid sequences encoding for HQT1, HQT2, HQT3 and HCT genes were retrieved from NCBI. These sequences were submitted as reference genes to create a custom database from which the similar genes of the phenylpropanoid pathway (PPP) from *B. pilosa* L. data were retrieved/identified through the BLAST platform of the NCBI. Thereafter, an excel file was created and all the sequences with similarity to the phenylpropanoid genes from *B. pilosa* L. genes were saved, and the results were stored as follows: sequences that showed similarity either to HQT1, HQT2, HQT3 or HCT were marked as such on the excel file using specified numerical filters. This allowed the identification of only the genes of interest which are potential HQT1, HQT2, HQT3 and HCT. Thereafter, identified potential HQT/HCT sequences were subjected to further scrutiny such as percentage similarity.

2.2.4. Percentage similarity

Multiple Alignment using Fast Fourier Transform (MAFFT) CD-hit was used to compute percentage matrix index. This was done by submitting sequences of identified homologues for each isoform of HQT (HQT1, HQT2, HQT3) and HCT genes separately. The results were saved in a separate excel file where different colours were used to mark the most identical sequences (>95%). The identical sequences were eliminated, and preference was also given to sequences >1000 bp

2.2.5. Integrity of Sequences

To check the integrity of the sequences, the Expert Protein Analysis System (ExPASy) was used. All the identified sequences in each family (HQT1, HQT2, HQT3 and HCT) were subjected to this computational tool to check whether the sequence is truncated or not. The sequences were first mapped to the reference gene to establish the start and stop codon. Thereafter, the trimmed sequences were taken to ExPASy to check if they will produce an open reading frame (ORF) in an acceptable orientation. The results were then recorded in excel as truncated if the sequence could not generate a meaningful ORF.

2.2.6. Multiple sequence alignment (MSA)

Potential HQT/HCT genes were further aligned with those of *B. pilosa*'s homologues. The identified sequence homologues were from *Helianthus annuus* (QBM78938.1), *Cynara cardunculus var scolymus* (AFL93687.1), *Artemisia annua* (PWA39281.1), *Lactuca sativa* (XP_023733842.1), *Mikania micrantha* (KAD5794970.1), *Lonicera japonica* (AEK80405.1), *Chicorium intybus* (ANN12610.1) and *Tanacetum cinerariifolium* (GEV77257.1) as obtained from NCBI. These sequences were aligned using MUSCLE function built in MEGA software (Hall, 2013) and viewed using GLC workbench.

2.2.7. Phylogenetic Analysis

Identified HQT and HCT protein sequences from *B. pilosa* L. together with the sequences retrieved from NCBI were exported to MEGA version 10.1.7 (Hall, 2013) and the neighbour joining (NJ) phylogenetic tree was constructed with the bootstrap value of 1000 (Mao *et al.*, 2019).

2.2.8. Metabolite extraction

Two grams (2 g) of fine ground leaves of *B. pilosa* L. were dissolved in 20 mL of 80% aqueous methanol. The mixture was spun throughout the night in a digital rotisserie tube rotator at 70 rpm to enhance the metabolite extraction. Tubes were placed in a rack to allow separation of supernatant and pellet by gravity. The supernatant (1 mL) was transferred to a 2 mL tubes. The samples were filtered twice using 1 mL syringe fitted with a 0.22 μm nylon filter into a 2 mL vial fitted with 0.2 mL conical bottom glass insert. The samples were stored at 4 °C until analysis.

2.2.9. Liquid chromatography mass spectrometry

Bidens pilosa L. analysis were performed on an LC-qTOF-MS, model LC-MS 9030 instrument (Shimadzu, Kyoto Japan), fitted with a Shim Pack Velox C18 column (100 mm \times 2.1 mm with particle size of 2.7 μm) (Shimadzu, Kyoto, Japan), placed in a column oven thermo-stated at 55 °C. A binary solvent mixture consisting of solvent A: 0.1% formic acid in water and solvent B: 0.1% formic acid in acetonitrile (UHPLC grade, Romil SpS, Cambridge, UK) was used with a total flow rate of 0.4 mL/min. successful separation of analytes was achieved through a 53 min long gradient method consisting of the following steps: initial, 10% B for 3 min, followed by a steep gradient to 60% B over 40 min, constant at 60% for 3 min, increased to 90% in 2 min, kept at 90% over 3 min, and returned to 10% for 2 min and finally the initial conditions (10% B) were re-established and column was allowed to re-equilibrate for 3 min. Mass spectrometry detection parameters were set as follows: ESI negative ionization mode; interface voltage of 3.5 kV; nebulizer gas flow at 3 L/min; heating gas flow at 10 L/min; heat block temperature at 400 °C; CDL temperature at 250 °C; detector voltage at 1.70 kV; TOF tube temperature at 42 °C. Sodium iodide (NaI) was used as a mass calibration solution to ensure acquisition of high accurate masses (m/z) at a range of 100–1000 for both high-resolution MS and tandem MS (MS/MS) experiments. For MS/MS experiments, argon gas was used as collision gas, and MS^E mode using collision energy ramp of 15 to 25 eV was used to generate possible fragments. Lab solution software (Postrun Analysis) from Shimadzu was used for peak and fragments detection.

Table 2.1 A summary of all the acyltransferase genes identified in *B. pilosa* L and their length (bp), the full sequences of each of the genes and ORF are attached as supplementary files. HQT1 ORF as determined through ExPASy was included as an example in the table below.

No.	Name	Nucleotide Sequence (AA)	Length (bp)	ORF sequence (AA)
01.	HQT1	Supplementary file Nucleotide Seq 1	1471 bp	MKLTVKESSIIKPAKPTPVTRIWNNSNLDLVVGRIHILTVYFYRPNSSGFFDPGVMKEALA GVLVSFFPMAGRLAKDGNRIEINCNGEGVLVFEAEADCCIDDFGEITPSELRQLAPTVD YSGEIDSYPLVITQVTRFKCGGVSLGCGLH HTLSD GLSSLHFINTWSDKARGLSVAVPPFLD RTLLRARNPPTPMFNHVEYDQPPSMITPLENQRSPSHKSTSTVMLRSLDQLNDLKLKAK GDESAHHSTYDILAAHLWRCVCKSRGLLDDQPTKLYVATDGRSRLNPPLPPGYLGNVIFTAT PIMKVGEFKSESLGDTARRIHNELARMDDQYLRSAYDLETIADPSTLVRGPSYFASPNLNVN SWTRLPIYDS DFGWR PIFMGPASILYEGTIYIIPCPSGDRSVKLAVCLDSDHMTLFKECLYDF
02.	HQT2	Supplementary file Nucleotide Seq 2	1534 bp	Supplementary figure S7B
03.	HQT3	Supplementary file Nucleotide Seq 3	1558bp	Supplementary figure S7C
04.	HCT	Supplementary file Nucleotide Seq 4	1601 bp	Supplementary figure S7D

2.3. Results and discussion

The transcriptomic data of *B. pilosa* L. revealed 2535754 HiFi reads, and the HiFi read length mean per base pair for this data was found to be 1776 (Figure 2.1). To identify the acyltransferases from the whole transcriptome (2535754 HiFi reads) data of *B. pilosa* L., Basic Local Alignment Search Tool plus (BLAST+) command line was used to generate results timeously since the data was large and could take time to process using a normal online BLAST. The phenylpropanoid pathway gene sequences from other plants were obtained from NCBI, BLAST+ command line identified 312 acyltransferases genes in *B. pilosa* L. The phenylpropanoid pathway genes from other plants were used as reference to probe for the same genes that play a role in phenylpropanoid pathway, especially those responsible for CGAs biosynthesis in *B. pilosa* L. From the identified 312 acyltransferase gene sequences, the following outcomes were achieved: three potential HCT sequences (Figure S2), four potential HQT1 sequences (Figure S3), twenty-one potential HQT2 sequences (Figure S4) and twelve potential HQT3 (Figure S5). Multiple Alignment using Fast Fourier Transform Percentage Matrix Index (MAFFT PMI) was used to eliminate identical sequences out of these potential acyltransferase genes (Figure S6, A-D). Open reading frames were generated using Expert Protein Analysis System (ExPASy) to determine truncated and untruncated sequences and results were mapped in excel for easy navigation. All the identified sequences produced six open reading frames each, when they were processed using ExPASy translation tool and showed a very good ORF sequence on the first frame (5' to 3' frame) (Figure S7, A-D). The identified ORFs also contained the conserved regions (HXXXD and DFGWG motifs) which are synonymous with the BAHD acyl transferase (St-pierre and De luca, 2000). The size of the genes identified are as follows: HCT - 1601 bp, HQT1 – 1471 bp, HQT2 – 1534 bp and HQT3- 1558 bp (Table 1). HQT1 open reading frame was shown (Table 2.1) and the conserved regions were marked by a red colour. To characterize and annotate HQT/ HCT genes found in *B. pilosa* L., a MSA of the known genes from other plants including those within the *Compositae* family was generated (Figure 2.1, 2.3, 2.5 and 2.7). The MSA of acyltransferases (Hydroxycinnamoyl-CoA: quinate/shikimate acid hydroxycinnamoyl transferase (HQT/HCT)) from *B. pilosa* L. in comparison with other known genes was generated using MEGA X and visualised using CLC genomics

workbench software. The homologous sequences of the HQT gene of *B. pilosa* L. which were used in this study are from *Helianthus annuus* (sunflower), *Cynara cardunculus var scolymus* (Cardoon), *Artemisia annua* (Sweet wormwood), *Lactuca sativa* (Lettuce), *mikania micrantha* (Bittervine), *Lonicera japonica* (Honeysuckle), *Chicorium intybus* (Chicory), *Echinacea purpurea* (Purple cornflower), just to name the few. These homologous sequences together with others not mentioned here were used to generate a percentage matrix index (PMI) (Figure S6, A-D), through MSA. Furthermore, MSA allowed for the annotation of genes through evaluation of the level of similarities in the base pairs (bp) and the percentage matrix index was also found to be sufficient to establish the overall similarities of the genes understudied by giving a similarity percentage.

2.3.1. Characterization of putative Hydroxycinnamoyl-CoA: quinic hydroxycinnamoyl transferase gene 1

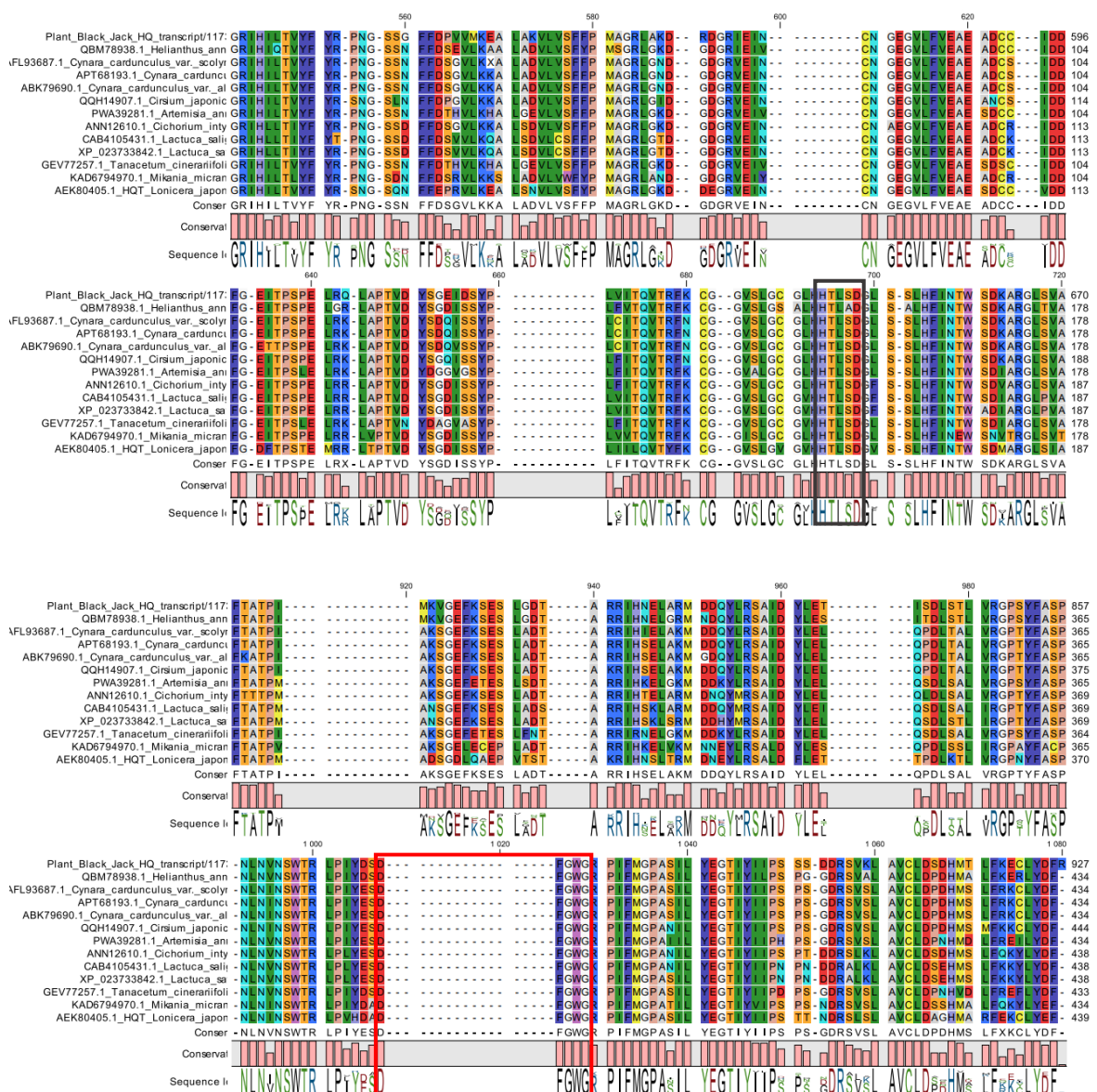


Figure 2.1 Multiple sequence alignment of *B. pilosa* L. HQT1.

Multiple sequence alignment of *B. pilosa* HQT1 with its homologues from *Helianthus annuus* (QBM78938.1), *Cynara cardunculus* var *scolymus* (AFL93687.1), *Artemisia annua* (PWA39281.1), *Lactuca sativa* (XP_023733842.1), *Mikania micrantha* (KAD5794970.1), *Lonicera japonica* (AEK80405.1), *Chicorium intybus* (ANN12610.1) and *Tanacetum cinerariifolium* (GEV77257.1). Residues are grouped according to colours, for instance the same colour represents similar residues across all genes from different plants. The alignment was generated using the MUSCLE algorithm of the MEGA software. The position of the residue is shown by the number on the right.

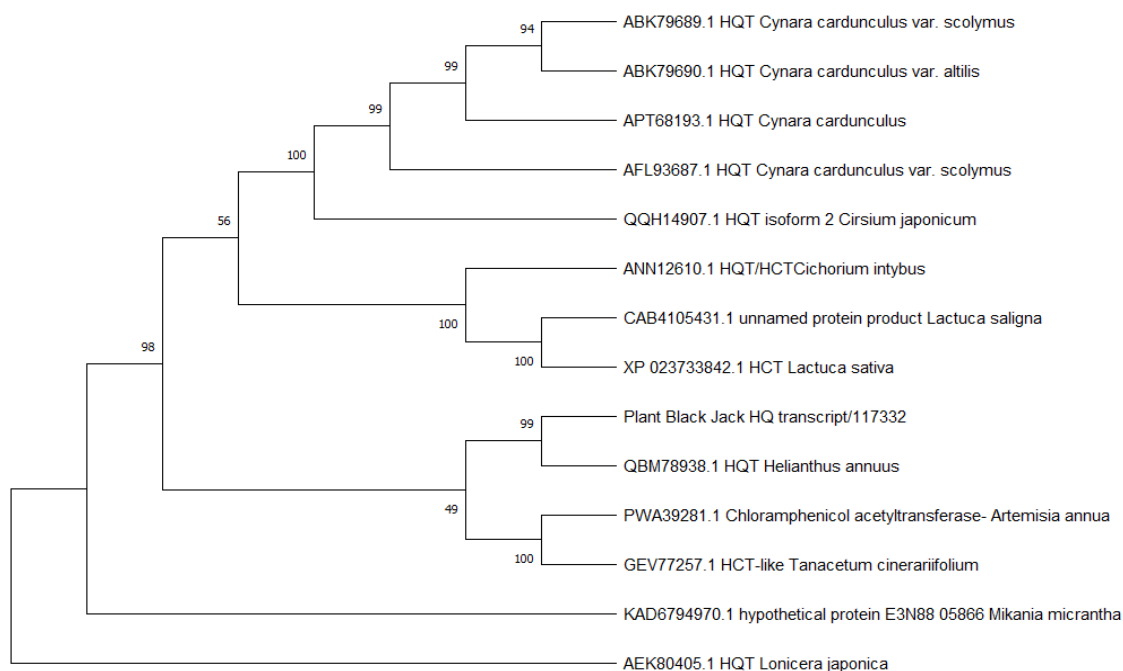


Figure 2.2 Phylogenetic tree of HQT1 genes was inferred using the Neighbor-Joining method (Saitou & Nei., 1987). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site.

All sequences which showed high similarity index were used to generate multiple sequence alignment to establish homology in these sequences. Figure 2.1 shows a multiple sequence alignment of sequences from *B. pilosa* L. and other plants from the same family. The alignment was generated using MUSCLE built in MEGAX and visualised using CLC workbench. The BLAST search results showed that the similarity index of HQT1 gene with its homologues is high (82% on average), the MSA further support the findings that BLAST results showed because colours are showing a lot of bases aligning in the MSA, which is a true reflection of what BLAST percentage identity was showing. The results indicated that the sequence from *Helianthus annuus* showed a very high similarity index which is 84.25% with a total query coverage of 88% against the HQT1 sequence from *B. pilosa* L. The query coverage is important, atleast more than half of the query sequence should align with the subject sequence to qualify the similarity index. For instance, if only less than 50% of the query sequence is aligned to the subject sequence and result in a high similarity index, probability is that the sequences that are being compared might not be the same since only a small part of the query sequence is aligning to the subject. The putative HQT1 gene from *B. pilosa*

L. showed a very high similarity with its homologues as shown by different colours, it has also shown conserved regions. The HQT1 gene has several conserved regions which include the HXXXD and DFGWG motifs (St-pierre and De luca, 2000). Figure 2.1 is showing a motif HTLSD across all the HQT1 genes from the plants computed, including the newly established sequence from *B. pilosa* L. The second motif is shown in red square which is the DFGWG motif and is also conserved in *B. pilosa* L. and all its homologues as shown in Figure 2.1. All the HQT1 genes aligned herein were from plants from the *Compositae* family and it is therefore safe to say that, unless stated elsewhere the HXXXD of the HQT genes from the *Compositae* is HTLSD, as shown in figure 2.1. Furthermore, a Neighbour-Joining phylogenetic tree was constructed using the same sequences and interesting enough, plants from *Compositae* plants were found to be grouped closer together which proves that these genes share an evolutionary history (Figure 2.2). This then positively helps in ascertaining that the gene identified from *B. pilosa* L. is putatively HQT1 gene of this plant. The HTLSD motif has been reported before in other plants such as *Echinacea purpurea* L. (Fu *et al.*, 2021), *Lonicera macranthoides* (Chen *et al.*, 2017), *A. spathulifolius* (Park *et al.*, 2021) and most importantly the HTLSD motif has been identified Hydroxycinnamoyl-CoA: tartaric acid hydroxycinnamoyl transferase genes from *Bidens pilosa* (Mathatha *et al.*, 2022).

2.3.2. Characterization of putative Hydroxycinnamoyl-CoA: quinic hydroxycinnamoyl transferase gene 2

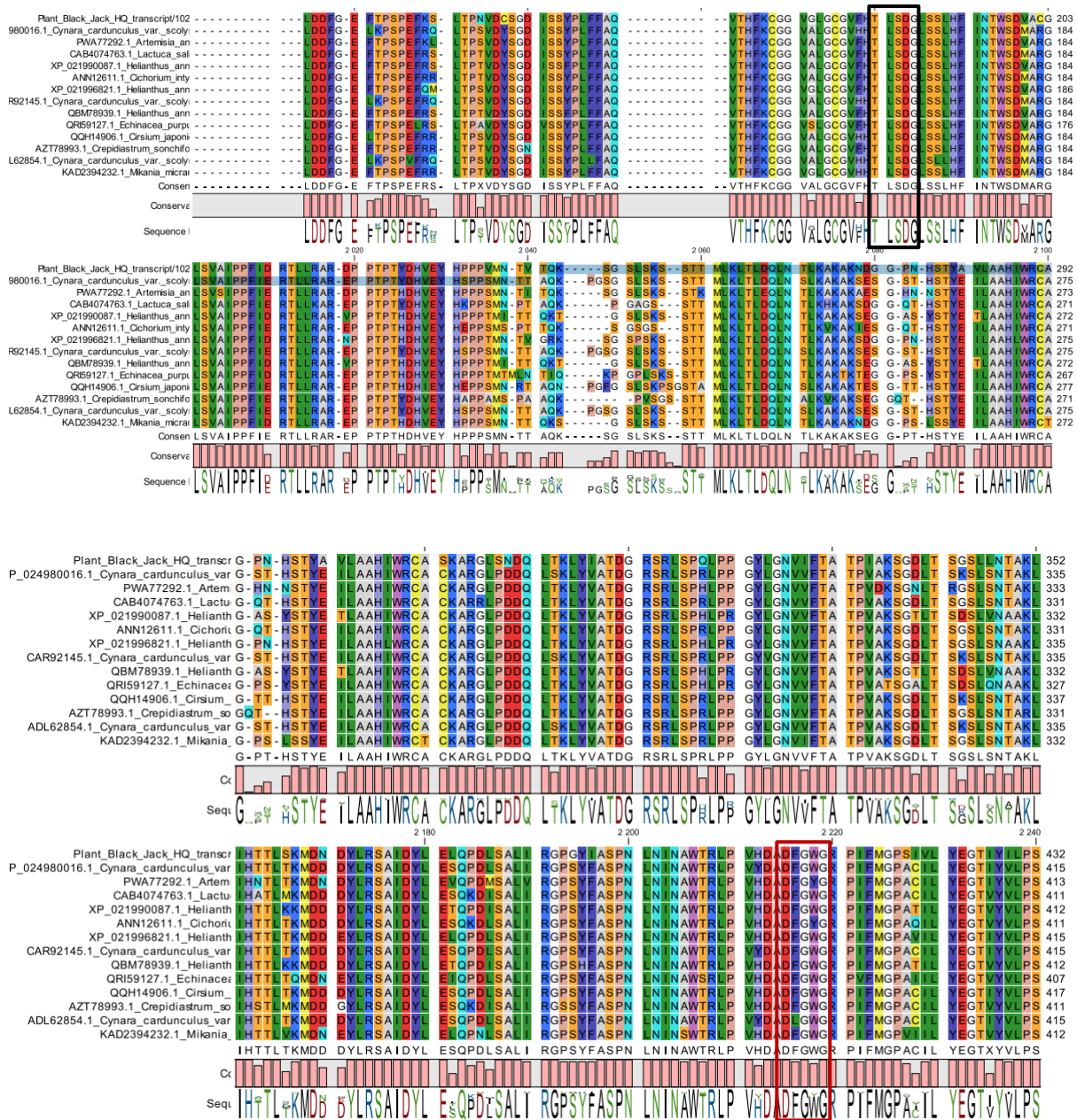


Figure 2.3 Multiple sequence alignment of *B. pilosa* L. HQT2.

Multiple sequence alignment of *B. pilosa* L. HQT2 with its homologues from *Helianthus annuus* (XP_021990087.1), *Cynara cardunculus* var *scolymus* (XP_024980016.1), *Artemisia annua* (PWA77292.1), *Lactuca sativa* (CAB4074763.1), *Mikania micrantha* (KAD2394232.1), *Chicorium intybus* (ANN12611.1), *Echinacea purpurea* (QRI59127.1), *Cirsium japonica* (QQH14906.1) and *Crepidiastrum sonchifolium* (AZT78993.1). Residues are grouped according to colours, for instance the same colour represents similar residues across all genes from different plants. The alignment was generated using MUSCLE. The position of the residue is shown by the number on the right.

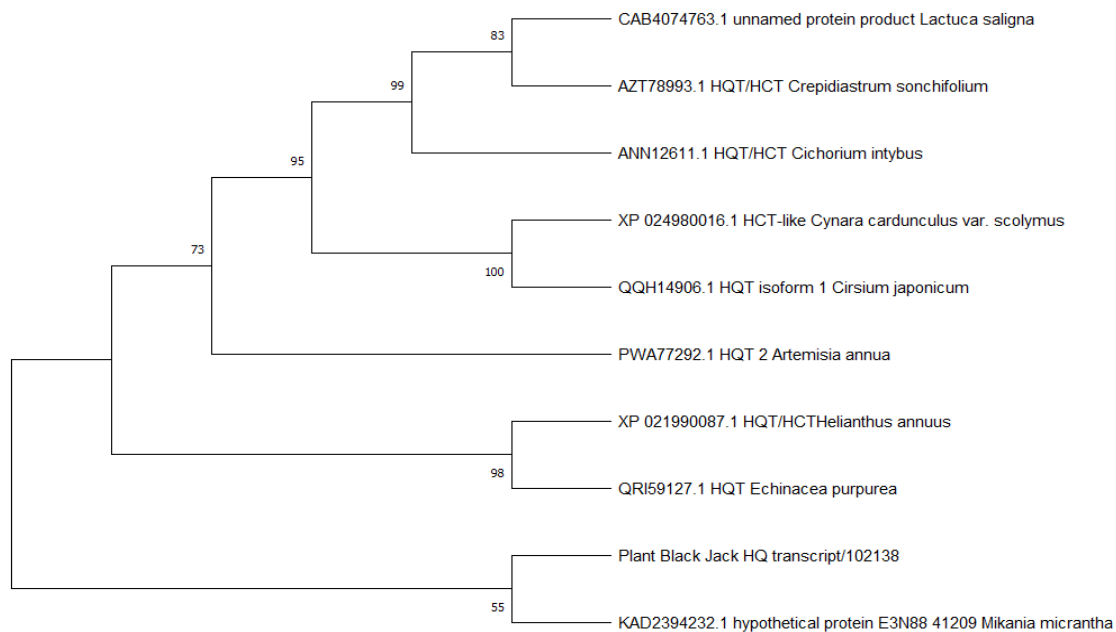


Figure 2.4 Phylogenetic tree of HQT2 genes was inferred using the Neighbor-Joining method (Saitou & Nei., 1987). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site.

Literature was used to validate the identified putative HQT2 gene of *B. pilosa* L. by looking for the conserved motif identified in other known HQT2 genes of other plants. Similarly, MSA and BLAST searches were also used in positively identifying this gene as HQT2 of *B. pilosa* L. A similar gene from *Mikania micrantha* showed a very high query coverage of 85% and similarity index (SI) of 88% followed by sequences from *Helianthus annuus* at SI of 87% with the same query cover percentage. Phylogenetic similarity was evaluated using the NJ method and it was found to support findings presented by NCBI and the confidence level percentage between *B. pilosa* L. sequence and *Mikania micrantha* was found to be 55% (Figure 2.4). The MSA also revealed that the two motifs found in the acyltransferase of the BAHD family were also conserved in all HQT2 (Figure 2.3). The first motif marked in black is HTLSD (similarly to the HQT1) and the second motif is DFGWG marked in red. All the sequences aligned here are from plants that fall under the *Compositae* family, same as *B. pilosa* L.

2.3.3. Characterization of putative Hydroxycinnamoyl-CoA: quinic hydroxycinnamoyl transferase gene 3

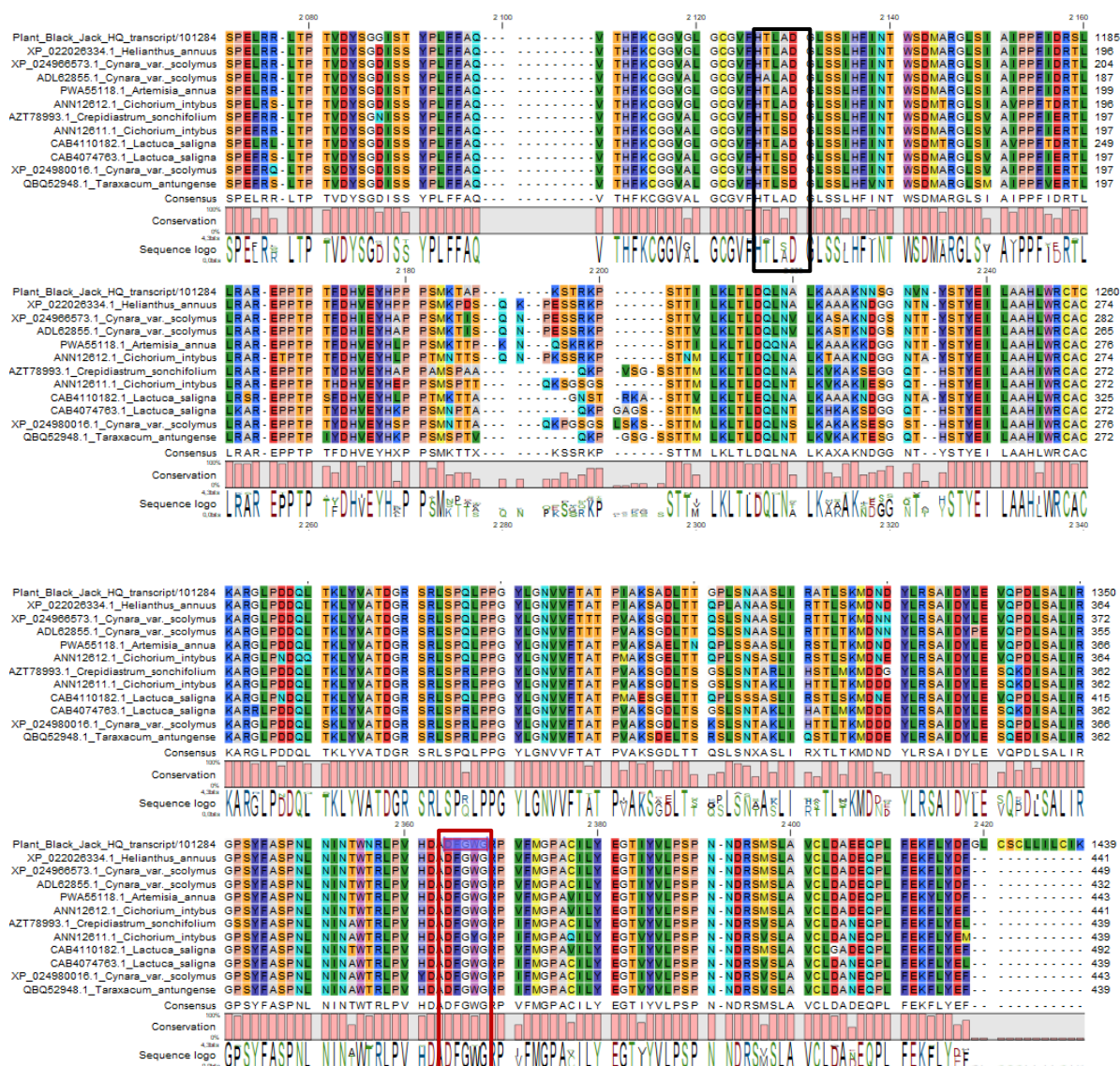


Figure 2.5 Multiple sequence alignment of *B. pilosa* L. HQT3.

Multiple sequence alignment of *B. pilosa* L. HQT3 with its homologues from *Helianthus annuus* (XP_022026334.1), *Cynara cardunculus* var *scolymus* (P_024966573.1/ADL62855.1), *Artemisia annua* (PWA55118.1), *Lactuca sativa* (CAB4110182.1/CAB4074703.1), *Chicorium intybus* (ANN12811.1), *Crepidiastrum sonchifolium* (AZT78993.1) and *Taraxacum antungense* (QBQ52948.1). Residues are grouped according to colours, for instance the same colour represents similar residues across all genes from different plants. The alignment was generated using MUSCLE. The position of the residue is shown by the number on the right.

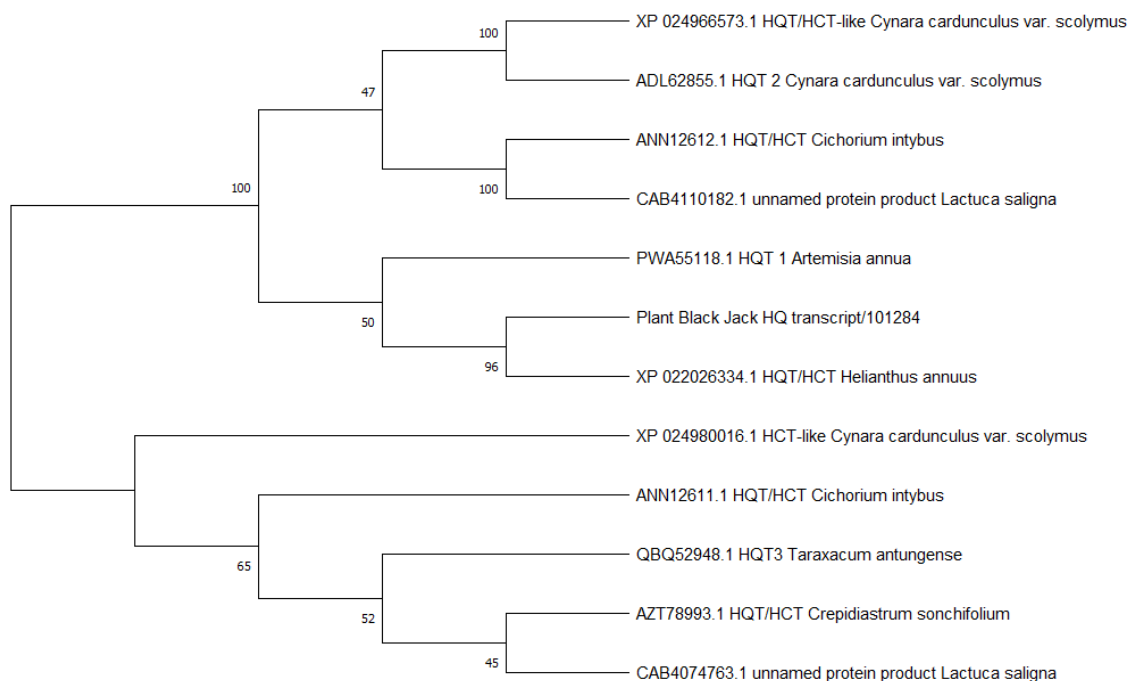


Figure 2.6 Phylogenetic tree of HQT3 genes was inferred using the Neighbor-Joining method (Saitou & Nei., 1987). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site.

The identification of the sequence corresponding to the HQT3 gene followed the same protocol used to identify HQT1 and HQT2. MSA and BLAST searches were used to check for homology. NCBI-based BLASTX search showed that the sequence retrieved from *Helianthus annuus* to have a similarity index of 92% from query coverage of 83%, followed by the sequence from *Artemisia annua* with 91% similarity index and 72% query coverage. HQT3 aligned very well with its orthologues and showed similarities in most base pairs but most importantly the conserved motifs were also seen in the alignments (Figure 2.5). Evolutionary history of these genes through sequences shows that *B pilosa* L. is 96% closely related to *Helianthus annuus* which in phylogeny means they are quite same (Figure 2.6). For the HXXXD motif, these results indicate that HQT 3 had HTLAD sequence as compared to its counterparts (HQT1 and HQT2) in this plant with the HTLSD motif. It was noted that the serine was substituted by alanine, and this has been recorded in other plants and showed that both these motifs are associated with HQT genes (Park *et al.*, 2021).

2.3.4. Characterization of putative Hydroxycinnamoyl-CoA: shikimate hydroxycinnamoyl transferase gene

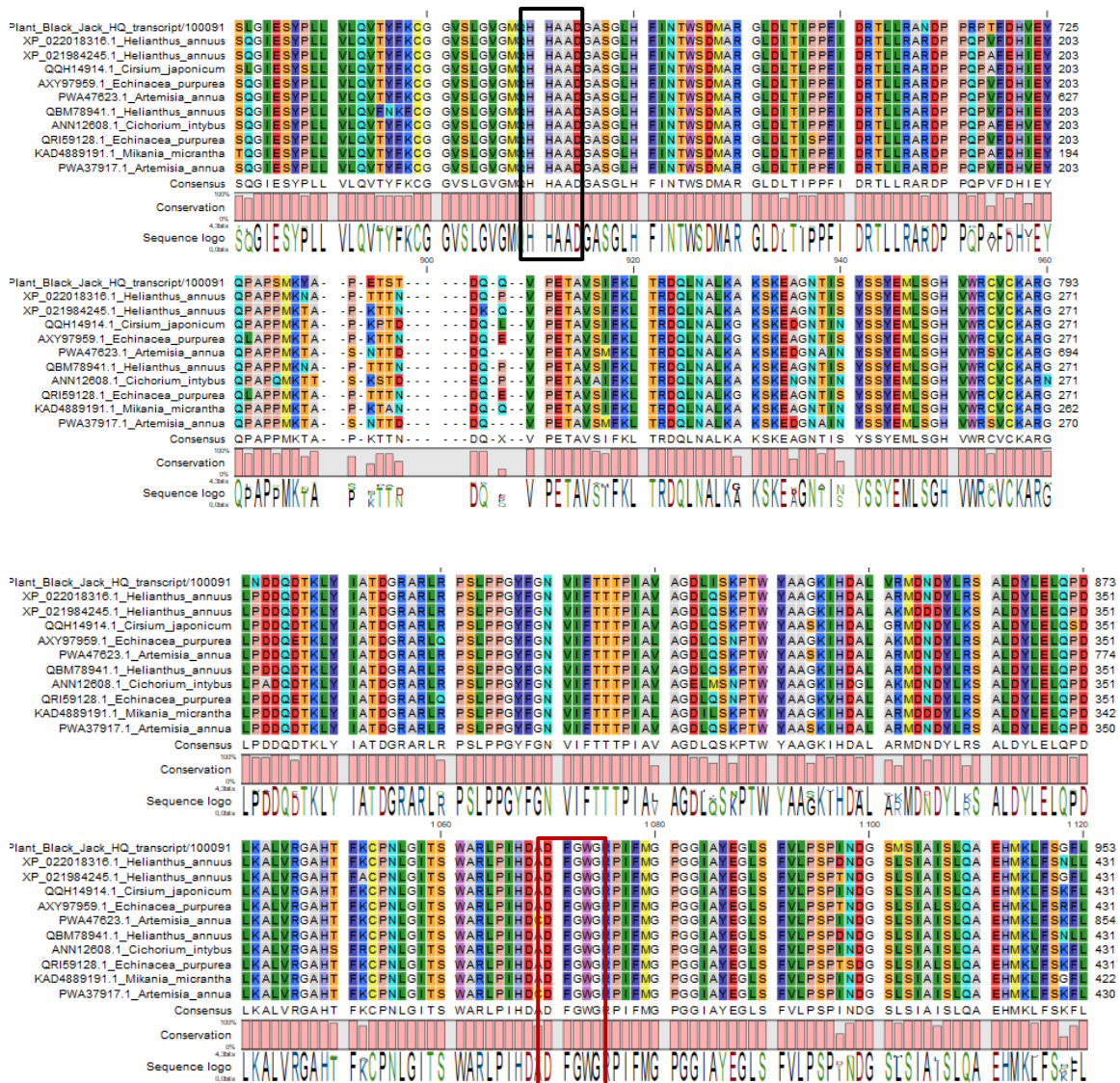


Figure 2.7 Multiple sequence alignment of *B. pilosa* L. HCT.

Multiple sequence alignment of *B. pilosa* L. HCT with its homologues from *Helianthus annuus* (XP_022018316.1), *Cirsium japonicum* (QQH14914.1), *Artemisia annua* (PWA47823.1), *Mikania micrantha* (KAD4889191.1), *Chicorium intybus* (ANN12608.1) and *Echinacea purpurea* (QRI59128.1). Residues are grouped according to colours, for instance the same colour represents similar residues across all genes from different plants. The alignment was generated using MUSCLE. The position of the residue is shown by the number on the right.

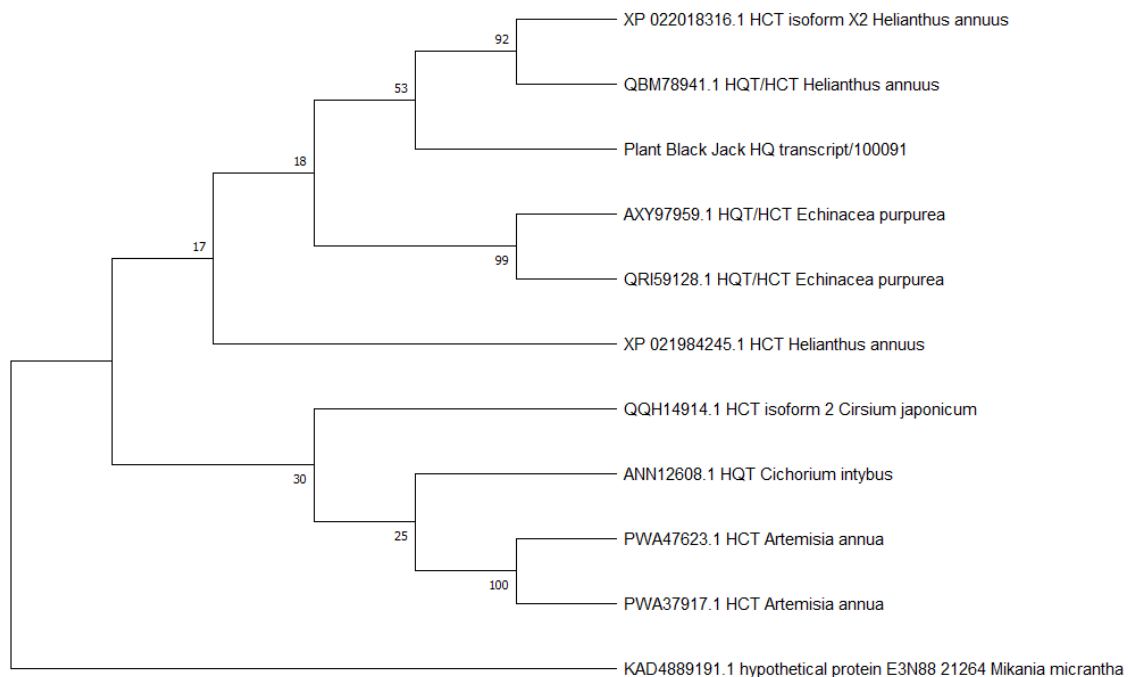


Figure 2.8 Phylogenetic tree of HCT genes was inferred using the Neighbor-Joining method (Saitou & Nei., 1987). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site.

For identification of the HCT sequence, the sequence retrieved from *Helianthus annuus* showed a high level of similarity index (94.47%) and a query cover of 81% when compared to a putative HCT gene identified from *B. pilosa* L. Putative HCT gene was found to be a BAHD acyltransferase family member, because it possesses HHAAD and DFGWG motifs respectively. These are very conserved regions across HCT genes of plants from the *Compositae* family. The HXXXD motif exhibits different amino acids in the XXX region, hence this motif is different from one gene to another from different plants (Park *et al.*, 2021). However, HHAAD motif has been noted in the HCT gene of *A. spathulifolius* that is in the same *Compositae* family as *B. pilosa* L. (Park *et al.*, 2021). Based on the bioinformatics analyses carried out herein, it can be said that the gene identified herein encodes for HCT. The phylogenetic tree of all the genes aligned herein is shown in Figure 2.8.

2.3.5. Discussion

HQT is a hydroxycinnamoyl-CoA quinate hydroxycinnamoyl transferase gene that facilitate the multi acylation of different cinnamic acids on the quinic acid to produce structurally different chlorogenic acids, mostly positional isomers (Ncube *et al.*, 2014). HQT gene has been well documented to be substrate specific and prefers quinate over shikimate (Comino *et al.*, 2007). These genes have been reported in different plants such as globe artichoke (Sonnante *et al.*, 2010), sunflower (Cheevarungnapakul, 2019), taraxacum and many others. The HQT gene has three isoforms which are HQT1, HQT2 and HQT3 (Sonnante *et al.*, 2010), as shown above for *B. pilosa* L. (Figure 1, 3, 5 and 7). These genes are part of a wide family of enzymes called the BAHD acyltransferases (St-pierre and De luca, 2000) and they are known to play a role in the biosynthesis of CGA compounds (Lepelley *et al.*, 2007). This study has reported for the first time three HQT genes from *B. pilosa* L., a plant which is known to produce a wide variety of chlorogenic acid compounds (Ramabulana *et al.*, 2020). This plant has been shown to produce mono-, di-, and tri- acylated CGAs, a unique phenomenon in plants (Ramabulana *et al.*, 2020), as other plants produce either one form or two of the three structural hierarchies. Although there are plants that have been commercialised that produce the entire hierarchy of these compounds such as coffee (Lepelley *et al.*, 2007), *B. pilosa* L. could serve as an alternative source of these compounds since it produces various CGAs of different hierarchies and structural diversity. Ironically, other plants such as chicory have been used to supplement commercial coffee, and accidentally these are plants that all contain large amounts of CGAs (Bahri *et al.*, 2012; Sonnante *et al.*, 2010), which makes *B. pilosa* L. a perfect candidate for this purpose too.

As stated above, the acyltransferases are responsible for biosynthesis of CGA metabolites and as such, the current study has overwhelmingly demonstrated that *B. pilosa* L. has at least three HQT and one HCT genes responsible for production of these compounds. Most importantly, all the three HQT genes identified in *B. pilosa* L. share very important conserved regions of acyltransferases which are HXXXD and DFGWG, and the sequences from *B. pilosa* L. matches very well with those from other plants from the *Compositae* family. As for the hydroxycinnamoyl CoA shikimate

hydroxycinnamoyl transferase (HCT), an acyltransferase gene which uses shikimate over quinate as a substrate (Comino *et al.*, 2007), only one gene was identified in *B. pilosa* L. herein, but it is important to mention that no single report exist of shikimate derivatives in *B. pilosa* L. This is interesting because in most cases, the presence of this gene has been shown to correlate with the existence of shikimate derivatives (Hoffman *et al.*, 2004). Elsewhere, HQT genes have been shown in *Arabidopsis thaliana*, but this plant does not produce CGA molecules (Niggeweg *et al.*, 2004), a phenomenon which cannot be explained based on the observations noted above.

To further validate the findings herein, phylogenetic analysis of HQT genes identified from *B. pilosa* L. shows that all the sequences aligned are similar with those from other plants from the *Compositae* family. Thus, phylogenetic trees show these genes to be in the same clade with other plants from the *Compositae* family (Figure 2.2, 2.4 and 2.6). The bootstrap values from these trees show that these genes are significantly similar and share a common ancestral origin. The constructed phylogenetic trees further validate that the sequences obtained from *B. pilosa* L. are indeed acyltransferases (Figure 2.2, 2.4 and 2.6).

Table 2.2 List of chlorogenic acids (CGAs) molecules isolated from randomly sampled *B. pilosa* L. plants established through analysis by LC-QTOF-MS. The different colour shading indicates different structural hierarchy of the identified molecule.

	Mass (<i>m/z</i>)	Rt (min)	Fragment ions	Molecular formula	Compound name	Abbreviation	Spectra
1.	353.088	2.975	135, 173, 191	C ₁₆ H ₁₈ O ₉	<i>cis</i> -3-Caffeoylquinic acid	<i>cis</i> -5-CQA	Fig. S17-A
2.		4.333	135, 179, 191	C ₁₆ H ₁₈ O ₉	<i>trans</i> -3-Caffeoylquinic acid	<i>trans</i> -3-CQA	Fig. S17-B
3.		7.992	173, 191	C ₁₆ H ₁₈ O ₉	<i>cis</i> -4-Caffeoylquinic acid	<i>cis</i> -4-CQA	Fig. S17-C
4.		8.783	173, 191	C ₁₆ H ₁₈ O ₉	<i>trans</i> -4-Caffeoylquinic acid	<i>trans</i> -4-CQA	Fig. S17-D
5.		10.025	191	C ₁₆ H ₁₈ O ₉	<i>trans</i> -5-Caffeoylquinic acid	<i>trans</i> -5-CQA	Fig. S17-E
6.		10.608	191	C ₁₆ H ₁₈ O ₉	<i>cis</i> -5-Caffeoylquinic acid	<i>cis</i> -5-CQA	Fig. S17-F
7.	367.103	8.710	134, 193	C ₁₇ H ₂₀ O ₉	<i>trans</i> -3-Feruloylquinic acid	<i>trans</i> -3FQA	Fig. S18-A
8.		13.277	134, 173, 193	C ₁₇ H ₂₀ O ₉	3-Feruloylquinic acid	3-FQA	Fig. S18-B
9.		13.985	191	C ₁₇ H ₂₀ O ₉	5-Feruloylquinic acid	5FQA	Fig. S18-C
10.		14.210	134, 191, 134, 173, 191	C ₁₇ H ₂₀ O ₉	5-Feruloylquinic acid	5FQA	Fig. S18-D
11.		15.635	135, 173, 179, 191	C ₁₇ H ₂₀ O ₉	Feruloylquinic acid isomer ***	iso-FQA	Fig. S18-E
12.		16.085	173, 179, 191, 335	C ₁₇ H ₂₀ O ₉	Feruloylquinic acid isomer ***	iso-FQA	Fig. S18-F
13.	499.124	22.248	135, 173, 179, 191, 353	C ₂₅ H ₂₄ O ₁₁	3-Coumaroyl-4-caffeoylquinic acid	3-Co-4-CQA	Fig. S19-A
14.		22.732	119, 163, 191, 337	C ₂₅ H ₂₄ O ₁₁	3-Coumaroyl-5-caffeoylquinic acid	3-Co-5-CQA	Fig. S19-B
15.		24.407	119, 173, 191, 337	C ₂₅ H ₂₄ O ₁₁	3-Coumaroyl-4-caffeoylquinic acid	3-Co-4-CAQ	Fig. S19-C
16.		25.151	119, 173, 337	C ₂₅ H ₂₄ O ₁₁	3-Coumaroyl-4-caffeoylquinic acid	3-Co-4-CQA	Fig. S19-D
17.		26.390	135, 173, 191, 353	C ₂₅ H ₂₄ O ₁₁	4-Coumaroyl-5-caffeoylquinic acid	4-Co-5-CQA	Fig. S19-E
18.		28.873	135, 179, 191, 353	C ₂₅ H ₂₄ O ₁₁	4-Coumaroyl-5-caffeoylquinic acid	4-Co-5-CQA	Fig. S19-F
19.	515.119	18.203	173, 179, 191, 335, 353	C ₂₅ H ₂₄ O ₁₂	3,4- <i>di</i> -Caffeoylquinic acid	3,4- <i>di</i> -CQA	Fig. S20-A

20.		19.395	135, 173, 179, 191, 335, 353	C ₂₅ H ₂₄ O ₁₂	3,4- <i>di</i> -Caffeoylquinic acid	3,4- <i>di</i> -CQA	Fig. S20-B
21.		20.245	135, 179, 191, 353	C ₂₅ H ₂₄ O ₁₂	3,5- <i>di</i> -Caffeoylquinic acid	3,5- <i>di</i> -CQA	Fig. S20-C
22.		20.945	179, 191, 353	C ₂₅ H ₂₄ O ₁₂	3,5- <i>di</i> -Caffeoylquinic acid	3,5- <i>di</i> -CQA	Fig. S20-D
23.		21.462	135, 173, 179, 191, 353	C ₂₅ H ₂₄ O ₁₂	4,5- <i>di</i> -Caffeoylquinic acid	4,5- <i>di</i> -CQA	Fig. S20-E
24.		22.995	135, 173, 179, 191, 353	C ₂₅ H ₂₄ O ₁₂	4,5- <i>di</i> -Caffeoylquinic acid	4,5- <i>di</i> -CQA	Fig. S20-F
25.		26.337	135, 173, 179, 191, 353	C ₂₅ H ₂₄ O ₁₂	4,5- <i>di</i> -Caffeoylquinic acid	4,5- <i>di</i> -CQA	Fig. S20-G
26.	529.135	22.797	135, 173, 179, 193, 367	C ₂₆ H ₂₆ O ₁₂	3-Caffeoyl-4-feruloylquinic acid	<i>iso</i> -3-C-4-FQA	Fig. S21-A
27.		23.147	135, 173, 367	C ₂₆ H ₂₆ O ₁₂	3-Caffeoyl-4-feruloylquinic acid	<i>iso</i> -3-C-4-FQA	Fig. S21-B
28.		23.822	135, 173, 367	C ₂₆ H ₂₆ O ₁₂	3-Caffeoyl-4-feruloylquinic acid	3-F-4-CQA	Fig. S21-C
29.		24.205	135, 179, 191, 353	C ₂₆ H ₂₆ O ₁₂	3-Feruloyl-5-caffeoylquinic acid	3-F-5-CQA	Fig. S21-D
30.		24.335	179, 191, 353, 367	C ₂₆ H ₂₆ O ₁₂	3-Feruloyl-5-caffeoylquinic acid	3-F-5-CQA	Fig. S21-E
31.		24.755	135, 161, 367	C ₂₆ H ₂₆ O ₁₂	Feruloyl-caffeoylquinic acid ***	FQA	Fig. S21-F
32.		25.338	135, 173, 191, 367	C ₂₆ H ₂₆ O ₁₂	4-Caffeoyl-5-feruloylquinic acid	4-C-5-FQA	Fig. S21-G
33.		26.230	134, 173, 193, 367	C ₂₆ H ₂₆ O ₁₂	4-Caffeoyl-5-feruloylquinic acid	4-C-5-FQA	Fig. S21-H
34.		26.905	135, 173, 191, 353	C ₂₆ H ₂₆ O ₁₂	4-Caffeoyl-5-feruloylquinic acid	4-C-5-FQA	Fig. S21-I
35.		27.572	135, 173, 179, 367	C ₂₆ H ₂₆ O ₁₂	Feruloyl-caffeoylquinic acid ***	FCQA	Fig. S21-J
36.		27.838	135, 173, 367	C ₂₆ H ₂₆ O ₁₂	Feruloyl-caffeoylquinic acid ***	FCQA	Fig. S21-K

2.4. Metabolites profiling

The methanol extracts of *B. pilosa* L. were also analysed using UHPLC-QTOF to check the metabolite content of this plant, with specific attention given to chlorogenic acids. A wide distribution of metabolites from *B. pilosa* L. was seen as presented by the chromatogram generated from the UHPLC-QTOF (Figure 2.9). Single ion chromatograms at m/z 353 (Figure S11), at m/z 367 (Figure S12), at m/z 499 (Figure S13), at m/z 515 (Figure 2.14) and at m/z 529 (Figure S15) were generated to show distribution of isomers for each compound. To further validate that indeed the identified genes have a function towards production of CGAs in *B. pilosa* L., LC-MS analyses using a targeted approach for CGAs was carried out. Below is a detailed characterization information showing how different forms of CGA were annotated from the methanolic extracts of *B. pilosa* L.

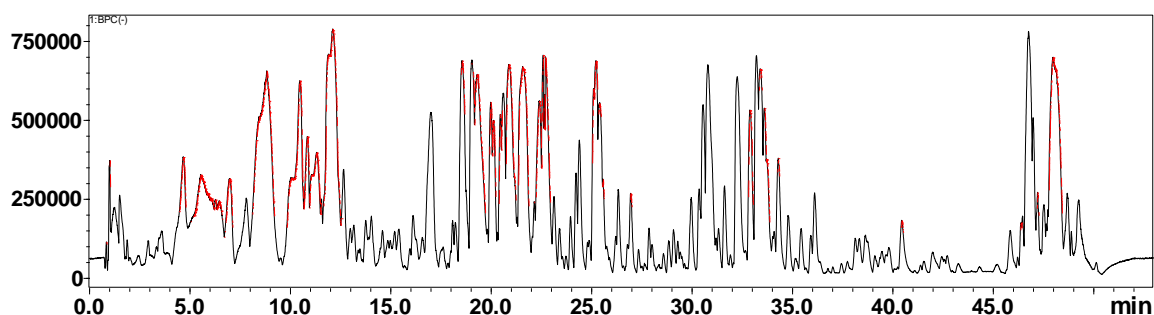


Figure 2.9 A representative base peak ion (BPI) of UHPLC-QTOF-MS chromatograms showing separation of metabolites from the methanol extracts of *B. pilosa* L.

2.4.1. Annotation of Mono-Acyl Chlorogenic Acids

Six molecules with the same precursor ion at m/z 353.088 $[M-H]^-$ were identified to represent caffeoylquinic acids in *B. pilosa* L (Table 2.2). Upon fragmentation, molecules at retention time 2.97 min and 4.33 min produced product ions at m/z 135 [Caffeic Acid- CO_2], at m/z 173 [Quinic Acid- $H-H_2O$] $^-$ and at m/z 191 [QA- H] $^-$. Through this fragmentation pattern and based on the information published elsewhere, these molecules were identified as *cis*-3-Caffeoylquinic acid (**1**) and *trans*-3-Caffeoylquinic (**2**) respectively (Ncube *et al.*, 2014). Molecules at retention time 7.9 min and 8.7 min similarly produced the product ions at m/z 173 [QA- $H-H_2O$] $^-$ and at m/z 191 [QA- H] $^-$.

These two molecules were putatively identified as *cis*-4-Caffeoylquinic acid (**3**) and *trans*-4-Caffeoylquinic acid (**4**) respectively (Ramabulana *et al.*, 2020). Lastly, two other molecules with retention time 10.0 min and 10.6 min produced a product ion at m/z 191 [QA-H]⁻. These two metabolites were positively identified as *trans*-5-Caffeoylquinic acid (**5**) and *cis*-5-Caffeoylquinic acid (**6**) respectively (Jaiswal *et al.*, 2014). Notably, as the case elsewhere, the *trans* isomers of both 3-CQA and 4-CQA eluted after their *cis* counterparts. However, the *trans* isomers of 5-CQA eluted before the *cis* form (Clifford *et al.*, 2008).

Similar approach was used in the identification of six molecules with a common precursor ion at m/z 367.103 [M-H]⁻ representing feruloylquinic acids. Molecules at retention time 8.7 min and 13.22 min, showed product ions upon fragmentation at m/z 134 [FA-H-CO₂-CH₃]⁻, at m/z 173 [QA-H-H₂O]⁻ and at m/z 193 [FA-H]⁻. Using this fragmentation pattern, these molecules were identified as *trans*-3-Feruloylquinic acid (**7**) and *cis*-3-Feruloylquinic acid (**8**) respectively. Molecule at 13.99 min showed a prominent peak at 191 [QA-H]⁻. Molecule at 14.21 min produced fragment ions at m/z 134 [CA-CO₂]⁻, at m/z 173 [QA-H-H₂O]⁻, at m/z 191 [QA-H]⁻. Both these molecules are geometrical isomers of each other and were identified as *trans*- and *cis*-5-Feruloylquinic acid (**9** and **10**) (Ncube *et al.*, 2014). Molecules at retention time 15.64 min and 16.09 min are geometrical isomers of each other and were identified as 4-Feruloylquinic acid (**11** and **12**). Upon fragmentation these molecules produced product ions at m/z 135 [CA-CO₂]⁻, at m/z 173 [QA-H-H₂O]⁻, at m/z 179 [CA-H]⁻, and at m/z 191 [QA-H]⁻. Molecule **12** also showed peaks at m/z 335 [CQA-H₂O-H]⁻.

2.4.2. Annotation of *p*-Coumaroyl-Caffeoylquinic Acids

Herein, multiple isomers of coumaroyl-caffeoylquinic acids were found to exist in *B. pilosa* L. as recorded elsewhere (Ramabulana *et al.*, 2020). Product ions produced during fragmentation served as diagnostic peaks for identification of acyl positions. Six peaks were detected with a common precursor ion at m/z 499.124 [M-H]⁻. Molecule **13** (3-Coumaroyl-4-caffeoylquinic acid) eluted at 22.25 min producing product ions at m/z 135 [CA-CO₂]⁻, at m/z 173 [QA-H-H₂O]⁻, at m/z 179 [CA-H]⁻, at m/z 191 [QA-H]⁻ and at m/z 353 [CQA-H]⁻. Two other molecules (**15** and **16**) were detected with similar

fragmentation pattern at retention times 24.41 min and 25.15 min respectively. Molecules 15 and 16 produced product ions at m/z 119 [p -Co-H-CO₂]⁻, at m/z 173 [QA-H-H₂O]⁻, at m/z 337. Molecule 15 also showed a prominent peak at m/z 191 [QA-H]⁻. Molecules 15 and 16 were identified as isomers of 3-Coumaroyl-4-caffeoylquinic acid. A molecule (14) at retention time 22.732 produced product ions at m/z 119 [p -Co-H-CO₂]⁻, at m/z 163 [p CoA-H]⁻, at m/z 191 [QA-H]⁻ and at m/z 337 [p CoQA-H]⁻ and was identified as 3-Coumaroyl-5-caffeoylquinic acid. The peak at m/z 163 is a characteristic of a coumaric acid at position 3 of a quinic acid. Two isomers (17 and 18) were also detected at retention times 26.39 min, 28.873 min respectively. Upon fragmentation, the following product ions were produced at m/z 135 [CA-CO₂]⁻, at m/z 173 [QA-H-H₂O]⁻, at m/z 191 [QA-H]⁻ and at m/z 353 [CQA-H]⁻. Molecule 18 also showed a peak at m/z 191 [QA-H]⁻. These two molecules were identified as 4-Coumaroyl-5-caffeoylquinic acid. Jaiswal *et al.* (2011) also identified p -Coumaroyl-Caffeoylquinic acids with the same fragmentation pattern reported herein.

2.4.3. Characterization of di-Caffeoylquinic Acids

In this study, seven *di*-CQA molecules were detected with a precursor ion at m/z 515.12 [M-H]⁻. Two isomers (19 and 20) eluted at retention times 18.2 min and 19.40 min. Upon fragmentation they produced product ions at m/z 173 [QA-H-H₂O]⁻, at m/z 179 [CA-H]⁻, at m/z 191 [QA-H]⁻, at m/z 335 [CQA-H₂O-H]⁻ and at m/z 353 [CQA-H]⁻. Molecule 20 also showed a prominent peak at m/z 135 [CA-CO₂]⁻. Therefore, molecules 19 and 20 were identified as 3.4-*di*-Caffeoylquinic acid, a peak at m/z 173 was used as a diagnostic peak since it represents 4-acyl substitution (Ncube *et al.*, 2014). Two other molecules (21 and 22) at retention times 20.25 min and 20.95 min were detected with fragmentation pattern showing product ions at m/z 135 [CA-CO₂]⁻, at m/z 179 [CA-H]⁻, at m/z 191 [QA-H]⁻ and at m/z 353 [CQA-H]⁻. These two molecules were identified as geometrical isomers of 3.5-*di*-Caffeoylquinic acid. Lastly, three isomers of 4.5-*di*-Caffeoylquinic acid were detected at retention times 21.46 min, 22.99 min and 26.34. Upon fragmentation these molecules (23, 24, 25) produced product ions at m/z 135 [CA-CO₂]⁻, at m/z 173 [QA-H-H₂O]⁻, at m/z 179 [CA-H]⁻, at m/z 191 [QA-H]⁻ and at m/z 353 [CQA-H]⁻. Notably the *di*-CQA have been showed previously

to appear in pairs of isomers due to UV induced geometrical isomers (Ramabulana *et al.*, 2020).

2.4.4. Characterization of Feruloyl-Caffeoylquinic Acids

Feruloyl-Caffeoylquinic acids were detected in *B. pilosa* L. with a precursor ion at m/z 529.135 [M-H]⁻. Eleven isomers of FCQA were identified as shown in table 1 (26 to 36). The first three molecules (26 to 28) were detected at retention times 22.80 min, 23.15 min and 23.82 min. MS fragmentation produced product ions at m/z 135 [CA-CO₂]⁻, at m/z 173 [QA-H-H₂O]⁻ and at m/z 367 [FQA-H]⁻. Molecule 26 also shows prominent peaks at m/z 173 [QA-H-H₂O]⁻ and at m/z 193. These molecules (26 through 28) are isomers of each other and were identified as geometrical isomers of 3-Caffeoyl-4-feruloylquinic acid. Two other molecules (29 to 30) at retention times 24.21 min and 24.34 min were detected and upon fragmentation they produced product ions at m/z 135 [CA-CO₂]⁻, at m/z 179 [CA-H]⁻, at m/z 191 [QA-H]⁻, at m/z 353 [CQA-H]⁻. Molecule 30 also shows another peak at m/z 367 [FQA-H]⁻. These molecules (29 and 30) were positively identified as 3-Feruloyl-5-caffeoylquinic acid. Three other molecules (32, 33 and 34) were detected at retention times 25.34 min, 26.23 min and 26.91 min, and upon fragmentation they produced product ions at m/z 135 [CA-CO₂]⁻, at m/z 173 [QA-H-H₂O]⁻, at m/z 191 [QA-H]⁻ and at m/z 367 [FQA-H]⁻. Molecule 34 showed a prominent peak at m/z 353 [CQA-H]⁻. This fragmentation pattern led to the identity of these molecules as geometrical isomers of 4-caffeoyl-5-feruloylquinic acid. Lastly, there other molecules were detected at retention times 24.76 min, 27.57 min and 27.84. Molecule 31 shows product ions at m/z 135 [CA-CO₂]⁻, at m/z 161 and at m/z 367, molecule 35 produced product ions at m/z 135 [CA-CO₂]⁻, at m/z 173 [QA-H-H₂O]⁻, at m/z 179 [CA-H]⁻, at m/z 367 [FQA-H]⁻ and lastly molecule 36 produced product ions at m/z 135 [CA-CO₂]⁻, at m/z 173 [QA-H-H₂O]⁻, at m/z 367 [FQA-H]⁻. All the three molecules were identified as isomers Feruloyl-caffeoylquinic acid but position of acylation could not be determined.

2.5. Conclusion

Secondary metabolites have been a subject of interest as of late, owing to the many health benefits they exhibit. This has led to more interest in plants that produces structurally diverse metabolites. Plants produce different forms of CGA compounds, with *B. pilosa* L. producing at least three structural hierarchies/forms (*mono*, *di* and *tri*-CQA). However, the biosynthetic/enzymatic machinery responsible for this wide spectrum of metabolites has not yet been established in many plants including in *B. pilosa* L. As such, this study reports for the first-time the sequences of the genes that play a role in the production of chlorogenic acids in *B. pilosa* L. Herein, SMRT sequencing approach in combination with bioinformatics methods allowed for identification of at least three HQT genes and one HCT gene. The identified genes were found to have sequence homology to the orthologous proteins encoded by the similar genes from other plants within the *Compositae* family. Furthermore, SMRT sequencing approach allowed for isolation and annotation of a full-length sequence which have sequence similarity of about 90 % when compared to homologues sequences from similar plants. The sequences of the genes obtained here were found to have convincing homology with the sequences retrieved from the NCBI encoding for similar genes in sunflower plants, which also falls within the same taxonomical family (*Compositae*). Interestingly, multi sequence alignment revealed that *B. pilosa* L. and other plants from the *Compositae* plants to contain a conserved HXXXD motif with the HTLSD sequence. This conserved motif together with the DFGWG motif in the HQT genes could provide a premise to design degenerative primers which can be used to isolate similar genes in other plants from the same family. Using extracts from *B. pilosa* L. plants, these plants were reaffirmed to contain structurally diverse CGA molecules through UHPLC-QTOF-MS analyses. The study also reaffirms this plant as an alternative source of these valuable nutraceutical metabolites. Future studies should aim at cloning the identified genes in other plants with an inferior metabolite composition to enhance their nutraceutical status.

Acknowledgements

The authors would like to thank the University of Venda and the National Research Foundation for the financial support. Ms Thifheli Kutama is also thanked for allowing us access to her field grown plants. Dr Khathutshelo Magwede is also thanked for helping with the authentication of plant species.

Conflict of Interest

The authors declare no conflict of interest

References

- Arthur, G.D., Naidoo, K.K., Cooposamy, R.M., 2012. *Bidens pilosa* L.: Agricultural and pharmaceutical importance. *Journal of Medicinal Plants Research* 6(17), 3282–3281. <https://doi.org/10.5897/JMPR12.195>
- Bahri, M., Hance, P., Grec, S., Quillet, M.C., Trotin, F., Hilbert, J.L., Hendriks, T., 2012. A “Novel” protocol for the analysis of hydroxycinnamic acids in leaf tissue of chicory (*Cichorium intybus* L., *Asteraceae*). *The Scientific World Journal*. <https://doi.org/10.1100/2012/142983>
- Bartolome, A.P., Villaseñor, I.M., Yang, W.C., 2013. *Bidens pilosa* L. (*Asteraceae*): botanical properties, traditional uses, phytochemistry, and pharmacology. *Evidence-based complementary and alternative medicine*. <https://doi.org/10.1155/2013/340215>
- Borges, C.C., Matos, T.F., Moreira, J., Rossato, A.E., Zanette, V.C., Amaral, P.A., 2013. *Bidens pilosa* L. (*Asteraceae*): traditional use in a community of southern Brazil. *Revista Brasileira de Plantas Mediciniais* 15, 34–40. <https://doi.org/10.1590/S1516-05722013000100004>
- Cheeverungnapakul, K., Khaksar, G., Panpetch, P., Boonjing, P., Sirikantaramas, S., 2019. Identification and functional characterization of genes involved in the biosynthesis of caffeoylquinic acids in sunflower (*Helianthus annuus* L.). *Frontiers in plant science* 10, 968. <https://doi.org/10.3389/fpls.2019.00968>
- Chen, Z., Liu, G., Liu, Y., Xian, Z., Tang, N., 2017. Overexpression of the LmHQT1 gene increases chlorogenic acid production in *Lonicera macranthoides* Hand-Mazz. *Acta Physiologiae Plantarum* 39(1), 1–10. <https://doi.org/10.1007/s11738-016-2310-8>
- Cheyrier, V., Schneider, R., Salmon, J.M., Fulcrand, H., 2010. Chemistry of wine. *Comprehensive Natural Products II* 3, 1119–1172. <https://doi.org/10.1016/B978-008045382-8.00088-5>

Chiang, Y.M., Chuang, D.Y., Wang, S.Y., Kuo, Y.H., Tsai, P.W., Shyur, L.F., 2004. Metabolite profiling and chemopreventive bioactivity of plant extracts from *Bidens pilosa*. *Journal of Ethnopharmacology* 95(2-3), 409–419. <https://doi.org/10.1016/j.jep.2004.08.010>

Comino, C., Hehn, A., Moglia, A., Menin, B., Bourgaud, F., Lanteri, S., Portis, E., 2009. The isolation and mapping of a novel hydroxycinnamoyltransferase in the globe artichoke chlorogenic acid pathway. *BMC Plant Biology* 9(1), 1–13. <https://doi.org/10.1186/1471-2229-9-30>

Comino, C., Lanteri, S., Portis, E., Acquadro, A., Romani, A., Hehn, A., Larbat, R., Bourgaud, F., 2007. Isolation and functional characterization of a cDNA coding a hydroxycinnamoyltransferase involved in phenylpropanoid biosynthesis in *Cynara cardunculus* L. *BMC Plant Biology* 7(1), 1–14. <https://doi.org/10.1186/1471-2229-7-14>

Felsenstein, J., 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39, 783–791. <https://doi.org/10.1111/j.1558-5646.1985.tb00420.x>

Fu, R., Zhang, P., Jin, G., Wang, L., Qi, S., Cao, Y., Martin, C., Zhang, Y., 2021. Versatility in acyltransferase activity completes chicoric acid biosynthesis in purple coneflower. *Nature communications* 12(1), 1–13. <https://doi.org/10.1038/s41467-021-21853-6>

Gökçen, B.B., Şanlıer, N., 2019. Coffee consumption and disease correlations. *Critical reviews in food science and nutrition* 59(2), 336–348. <https://doi.org/10.1080/10408398.2017.1369391>

Hall, B.G., 2013. Building phylogenetic trees from molecular data with MEGA. *Molecular biology and evolution* 30(5), 1229–1235. <https://doi.org/10.1093/molbev/mst012>

Hoffmann, L., Besseau, S., Geoffroy, P., Ritzenthaler, C., Meyer, D., Lapierre, C., Pollet, B., Legrand, M., 2004. Silencing of hydroxycinnamoyl-coenzyme A

shikimate/quininate hydroxycinnamoyltransferase affects phenylpropanoid biosynthesis. *The Plant Cell* 16(6), 1446–1465. <https://doi.org/10.1105/tpc.020297>

Jaiswal, R., Kiprotich, J., Kuhnert, N., 2011. Determination of the hydroxycinnamate profile of 12 members of the *Asteraceae* family. *Phytochemistry* 72, 781–790. <https://doi.org/10.1016/j.phytochem.2011.02.027>

Kim, Y.B., Thwe, A.A., Kim, Y.J., Li, X., Kim, H.H., Park, P.B., Suzuki, T., Kim, S.J., Park, S.U., 2013. Characterization of genes for a putative hydroxycinnamoyl-coenzyme A quinate transferase and p-coumarate 3'-hydroxylase and chlorogenic acid accumulation in tartary buckwheat. *Journal of agricultural and food chemistry* 61(17), 4120–4126. <https://doi.org/10.1021/jf4000659>

Legrand, G., Delporte, M., Khelifi, C., Harant, A., Vuylsteker, C., Mörchen, M., Hance, P., Hilbert, J.L., Gagneul, D., 2016. Identification and characterization of five BAHD acyltransferases involved in hydroxycinnamoyl ester metabolism in chicory. *Frontiers in plant science* 7, 741. <https://doi.org/10.3389/fpls.2016.00741>

Lepelley, M., Cheminade, G., Tremillon, N., Simkin, A., Caillet, V., McCarthy, J., 2007. Chlorogenic acid synthesis in coffee: An analysis of CGA content and real-time RT-PCR expression of HCT, HQT, C3H1, and CCoAOMT1 genes during grain development in *C. canephora*. *Plant Science* 172(5), 978–996. <https://doi.org/10.1016/j.plantsci.2007.02.004>

Li, D., Qian, J., Li, W., Yu, N., Gan, G., Jiang, Y., Li, W., Liang, X., Chen, R., Mo, Y., Lian, J., 2021. A high-quality genome assembly of the eggplant provides insights into the molecular basis of disease resistance and chlorogenic acid synthesis. *Molecular Ecology Resources* 21(4), 1274–1286. <https://doi.org/10.1111/1755-0998.13321>

Li, L., Liu, M., Shi, K., Yu, Z., Zhou, Y., Fan, R., Shi, Q., 2019. Dynamic changes in metabolite accumulation and the transcriptome during leaf growth and development in *Eucommia ulmoides*. *International journal of molecular sciences* 20(16), 4030. <https://doi.org/10.3390/ijms20164030>

Liu, Q., Yao, L., Xu, Y., Cheng, H., Wang, W., Liu, Z., Liu, J., Cui, X., Zhou, Y., Ning, W., 2019. In vitro evaluation of hydroxycinnamoyl CoA: quinate hydroxycinnamoyl transferase expression and regulation in *Taraxacum antungense* in relation to 5-caffeoylquinic acid production. *Phytochemistry* 162, 148–156.

<https://doi.org/10.1016/j.phytochem.2019.02.014>

Makola, M.M., Dubery, I.A., Koorsen, G., Steenkamp, P.A., Kabanda, M.M., du Preez, L.L., Madala, N.E., 2016. The effect of geometrical isomerism of 3, 5-dicaffeoylquinic acid on its binding affinity to HIV-integrase enzyme: A molecular docking study. *Evidence-Based Complementary and Alternative Medicine*.

<https://doi.org/10.1155/2016/4138263>

Malar, C.M., Yuzon, J.D., Panda, A., Kasuga, T., Tripathy, S., 2019. Updated assembly of *Phytophthora ramorum* pr102 isolate incorporating long reads from PacBio sequencing. *Molecular Plant-Microbe Interactions* 32(11), 1472–1474.

<https://doi.org/10.1094/MPMI-05-19-0147-A>

Mathatha, K., Khwathisi, A., Ramabulana, A.T., Mwaba, I., Mathomu, L.M., Madala, N.E., 2022. Identification of putative acyltransferase genes responsible for the biosynthesis of homogenous and heterogenous hydroxycinnamoyl-tartaric acid esters from *Bidens pilosa*. *South African Journal of Botany* 149, 389-396.

<https://doi.org/10.1016/j.sajb.2022.06.008>

Moglia, A., Acquadro, A., Eljounaidi, K., Milani, A.M., Cagliero, C., Rubiolo, P., Genre, A., Cankar, K., Beekwilder, J., Comino, C., 2016. Genome-wide identification of BAHD acyltransferases and in vivo characterization of HQT-like enzymes involved in caffeoylquinic acid synthesis in globe artichoke. *Frontiers in plant science* 7, 1424.

<https://doi.org/10.3389/fpls.2016.01424>

Ncube, E.N., Mhlongo, M.I., Piater, L.A., Steenkamp, P.A., Dubery, I.A., Madala, N.E., 2014. Analyses of chlorogenic acids and related cinnamic acid derivatives from *Nicotiana tabacum* tissues with the aid of UPLC-QTOF-MS/MS based on the in-source collision-induced dissociation method. *Chemistry Central Journal* 8(1), 1–10.

<https://doi.org/10.1186/s13065-014-0066-z>

Niggeweg, R., Michael, A.J., Martin, C., 2004. Engineering plants with increased levels of the antioxidant chlorogenic acid. *Nature biotechnology* 22(6), 746–754. <https://doi.org/10.1038/nbt966>

Park, S., Sivagami, J.C., Park, S., 2021. Transcriptome-wide identification and quantification of Caffeoylquinic acid biosynthesis pathway and prediction of their putative BAHGs gene complex in *A. spathulifolius*. *International journal of molecular science* 22, 633. <https://doi.org/10.3390/ijms22126333>

Raharimalala, N., Rombauts, S., McCarthy, A., Garavito, A., Orozco-Arias, S., Bellanger, L., Morales-Correa, A.Y., Froger, S., Michaux, S., Berry, V., Metairon, S., 2021. The absence of the caffeine synthase gene is involved in the naturally decaffeinated status of *Coffea humblotiana*, a wild species from Comoro archipelago. *Scientific reports* 11(1), 1–14. <https://doi.org/10.1038/s41598-021-87419-0>

Ramabulana, A.T., Steenkamp, P., Madala, N.E., Dubery, I.A., 2020. Profiling of chlorogenic acids from *bidens pilosa* and differentiation of closely related positional isomers with the aid of UHPLC-QTOF-MS/MS-based in-source collision-induced dissociation. *Metabolites* 10(5), 178. <https://doi.org/10.3390/metabo10050178>

Rhoads, A., Au, K., 2015. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics* 13(5), 278–289. <https://doi.org/10.1016/j.gpb.2015.08.002>

Saitou, N., Nei, M., 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4, 406–425. <https://doi.org/10.1093/oxfordjournals.molbev.a040454>

Seca, A.M., Pinto, D.C., 2019. Biological potential and medical use of secondary metabolites. *Medicines*, 6(2), 66. <https://doi.org/10.3390/medicines6020066>

Sonnante, G., D'Amore, R., Blanco, E., Pierri, C.L., De Palma, M., Luo, J., Tucci, M., Martin, C., 2010. Novel hydroxycinnamoyl-coenzyme A quinate transferase genes

from artichoke are involved in the synthesis of chlorogenic acid. *Plant Physiology* 153(3), 1224–1238. <https://doi.org/10.1104/pp.109.150144>

St-Pierre, B., De Luca, V., 2000. Origin and diversification of the BAHD superfamily of acyltransferases involved in secondary metabolism. *Recent Advances in Phytochemistry* 34, 285–315. [https://doi.org/10.1016/S0079-9920\(00\)80010-6](https://doi.org/10.1016/S0079-9920(00)80010-6)

Tobinaga, S., Sharma, M.K., Aalbersberg, W.G., Watanabe, K., Iguchi, K., Narui, K., Sasatsu, M., Waki, S., 2009. Isolation and identification of a potent antimalarial and antibacterial polyacetylene from *Bidens pilosa*. *Planta medica* 75(06), 624–628. <https://doi.org/10.1055/s-0029-1185377>

Xuan, T.D., Khanh, T.D., 2016. Chemistry and pharmacology of *Bidens pilosa*: an overview. *Journal of pharmaceutical investigation* 46(2), 91–132. <https://doi.org/10.1007/s40005-016-0231-6>

Zhang, T., Li, M., Zhan, Y.G., Fan, G.Z., 2020. Dataset of full-length transcriptome assembly and annotation of apocynum venetum using pacbio sequel II. *Data in brief* 33, 106494. <https://doi.org/10.1016/j.dib.2020.106494>

Zheng, X., Renslow, R.S., Makola, M.M., Webb, I.K., Deng, L., Thomas, D.G., Govind, N., Ibrahim, Y.M., Kabanda, M.M., Dubery, I.A., Heyman, H.M., 2017. Structural elucidation of *cis/trans* dicaffeoylquinic acid photoisomerization using ion mobility spectrometry-mass spectrometry. *The journal of physical chemistry letters* 8(7), 1381–1388. <https://doi.org/10.1021/acs.jpcllett.6b03015>

Supplementary Data

CCS Analysis Report

Value	Analysis Metric
2535754	HiFi Reads
4505879731	HiFi Yield (bp)
1776	HiFi Read Length (mean, bp)
Q38	HiFi Read Quality (median)
20	HiFi Number of Passes (mean)
269245	<Q20 Reads
569943945	<Q20 Yield (bp)
2116	<Q20 Read Length (mean, bp)
Q15	<Q20 Read Quality (median)

Figure S1. The circular consensus sequencing analysis report which shows the quality of the reads generated through the SMRT sequencing approach.

Sequenc	Sequence name	comment	similar	homolog	homolog 2
1	Plant_Black_Jack_HQ_transcript/100091	AACTCAAGACA	ACTTCACACCA	HCT1(MK598076)	
30	Plant_Black_Jack_HQ_transcript/109474	GAACTCA	Truncated Plant_Blac	HCT1(MK598076)	
89	Plant_Black_Jack_HQ_transcript/140168	GATCGAT	Truncated Plant_Blac	HCT1(MK598076)	

Figure S2. Excel snapshot showing identified three potential HCT gene sequences for *B. pilosa* during filtering process using excel custom made filters. Plant Blackjack HQ Transcript 100091 was identified as HCT gene for *B. pilosa*.

Sequenc	Sequence name	comment	similar	homolog	homolog 2
3	Plant_Black_Jack_HQ_transcript/100503	CCACCTTCTGGATA	AATACAACAC	HQT1(MK598073)	
39	Plant_Black_Jack_HQ_transcript/112031	ACCCACCTTCTGGAT	Plant_Blac	HQT1(MK598073)	
41	Plant_Black_Jack_HQ_transcript/113464	GGGTCGG	Truncated Plant_Blac	HQT1(MK598073)	
52	Plant_Black_Jack_HQ_transcript/117332	GACCACCTTCTGGATA	AAAACAAT	HQT1(MK598073)	

Figure S3. Excel snapshot showing identified four potential HQT1 gene sequences for *B. pilosa*.

Sequenc	Sequence name	comme	similar	homolo	homolog 2
6	Plant_Black_Jack_HQ_transcript/102138	GGTGAACAAAACCCAATCTCTT	HQT2(MK598074)		
10	Plant_Black_Jack_HQ_transcript/103089	GAACTCTA	Truncated	/sequence	HQT2(MK598074)
11	Plant_Black_Jack_HQ_transcript/103793	GTGAAAC	Truncated	Plant_Blac	HQT2(MK598074)
21	Plant_Black_Jack_HQ_transcript/106534	GTGTGCAACCCCAAACCAATAAC	HQT2(MK598074)		
46	Plant_Black_Jack_HQ_transcript/115403	GGTGCGC	Truncated	Plant_Blac	HQT2(MK598074)
49	Plant_Black_Jack_HQ_transcript/116155	GGTGAACAAAACCC	Plant_Blac	HQT2(MK598074)	
61	Plant_Black_Jack_HQ_transcript/121621	AACTCATC	Truncated	Plant_Blac	HQT2(MK598074)
69	Plant_Black_Jack_HQ_transcript/124097	AACCATG	Truncated	Plant_Blac	HQT2(MK598074)
84	Plant_Black_Jack_HQ_transcript/136685	GGTCATG	Truncated	Plant_Blac	HQT2(MK598074)
118	Plant_Black_Jack_HQ_transcript/17032	GGTGCGC	Truncated	Plant_Blac	HQT2(MK598074)
131	Plant_Black_Jack_HQ_transcript/20221	GGTGCAA	Truncated	Plant_Blac	HQT2(MK598074)
192	Plant_Black_Jack_HQ_transcript/52365	GGGTGAAACAAAAC	Plant_Blac	HQT2(MK598074)	
199	Plant_Black_Jack_HQ_transcript/55211	AGGTGAA	Truncated	/sequence	HQT2(MK598074)
263	Plant_Black_Jack_HQ_transcript/76946	GTGCGCAACCCCAAACCAATAAC	HQT2(MK598074)		
284	Plant_Black_Jack_HQ_transcript/90310	GTGCGCAACCCCAA	Plant_Blac	HQT2(MK598074)	
287	Plant_Black_Jack_HQ_transcript/90891	GGTGAACAAAACCC	Plant_Blac	HQT2(MK598074)	
298	Plant_Black_Jack_HQ_transcript/95595	GCGCAACCCCAAACC	Plant_Blac	HQT2(MK598074)	
299	Plant_Black_Jack_HQ_transcript/96170	GGTGAACAAAACCC	Plant_Blac	HQT2(MK598074)	
300	Plant_Black_Jack_HQ_transcript/96419	GCCCGAA	Truncated	/sequence	HQT2(MK598074)
305	Plant_Black_Jack_HQ_transcript/97943	GGTGAACAAAACCC	Plant_Blac	HQT2(MK598074)	
308	Plant_Black_Jack_HQ_transcript/98322	GGTGAAA	Truncated	Plant_Blac	HQT2(MK598074)

Figure S4. Excel snapshot showing identified twenty-one potential HQT2 gene sequences for *B. pilosa* during the filtering process using custom made filters in excel.

Sequenc	Sequence name	comme	similar	homolo	homolog 2
4	Plant_Black_Jack_HQ_transcript/101284	GTGTCTTCTTTAAACC	Plant_Blac	HQT3(MK598075)	
12	Plant_Black_Jack_HQ_transcript/103969	GGTGTCTTCTTTAAAC	Plant_Blac	HQT3(MK598075)	
24	Plant_Black_Jack_HQ_transcript/108057	GTGTCTTCTTTAAACC	Plant_Blac	HQT3(MK598075)	
27	Plant_Black_Jack_HQ_transcript/108970	GGTGTCTTCTTTAAAC	Plant_Blac	HQT3(MK598075)	
42	Plant_Black_Jack_HQ_transcript/113476	GGTGTCTTCTTTAAAC	Plant_Blac	HQT3(MK598075)	
58	Plant_Black_Jack_HQ_transcript/120049	GAACTCGTGTCTTCTT	Plant_Blac	HQT3(MK598075)	
71	Plant_Black_Jack_HQ_transcript/124745	GGGTGTC	Truncated	Plant_Blac	HQT3(MK598075)
250	Plant_Black_Jack_HQ_transcript/72570	GGTGTCTTCTTTAAAC	Plant_Blac	HQT3(MK598075)	
289	Plant_Black_Jack_HQ_transcript/91401	GTGTCTTCTTTAAACC	Plant_Blac	HQT3(MK598075)	
297	Plant_Black_Jack_HQ_transcript/95365	GGTGTCTTCTTTAAAC	Plant_Blac	HQT3(MK598075)	
302	Plant_Black_Jack_HQ_transcript/97320	GTGTCTTC	Truncated	Plant_Blac	HQT3(MK598075)
309	Plant_Black_Jack_HQ_transcript/98418	GTGTCTTCTTTAAACCTTCTTTCC	HQT3(MK598075)		

Figure S5. Excel snapshot showing identified twelve potential HQT3 gene sequences for *B. pilosa* during the filtering process using custom made filters on excel.

A

no	Sequence name						
1	MK598073,1	100	83,3	83,3	83,3	83,91	83,91
2	Plant_Black_Jack_HQ_transcript/100503	83,3	100	99,4	99,43	94,68	93,98
3	Plant_Black_Jack_HQ_transcript/112031	83,3	99,4	100	100	93,94	93,68
4	Plant_Black_Jack_HQ_transcript/97253	83,3	99,43	100	100	93,94	93,06
5	Plant_Black_Jack_HQ_transcript/117332	83,91	94,68	93,94	93,94	100	99,85
6	Plant_Black_Jack_HQ_transcript/113464	83,91	93,98	93,68	93,06	99,85	100

B

No.	Sequenc	Name				
1	MK598073,1	100	86,28	86,13	85,62	
2	Plant_Black_Jack_HQ_transcript/100503	86,28	100	99,74	100	
3	Plant_Black_Jack_HQ_transcript/112031	86,13	99,74	100	99,66	
4	Plant_Black_Jack_HQ_transcript/97253	85,62	100	99,66	100	

C

no.	Sequence ID																							
1	HQT2_MK598074_	100	66,22	65,75	66,46	64,77	66,31	65,91	66,56	66,4	66,43	57,14	57,45	65,91	66,2	66,54	66,54	65,91	66,54	65,91	65,91	66,2		
2	Plant_Black_Jack_HQ_transcript/102138	66,22	100	89,74	92,01	76,1	76,56	99,08	92,92	77,04	99,12	65,94	67,08	99,08	97	76,85	76,93	99,08	76,85	99,08	99,08	97		
3	Plant_Black_Jack_HQ_transcript/103099	65,75	89,74	100	97,76	75,83	75,46	88,97	96,87	75,92	90,52	64,77	66,41	88,97	88,8	75,96	76,04	88,97	75,96	88,97	88,97	88,8		
4	Plant_Black_Jack_HQ_transcript/103793	66,46	92,01	97,76	100	76,6	76,53	91,23	99,06	76,94	91,84	65,86	67,22	91,23	91,07	76,98	77,06	91,23	76,98	91,23	91,23	91,07		
5	Plant_Black_Jack_HQ_transcript/106534	64,77	76,1	75,83	76,6	100	87,56	75,45	77,17	87,71	79,5	77,46	82,65	75,45	75,83	88,2	88,12	75,45	88,2	75,45	75,45	75,83		
6	Plant_Black_Jack_HQ_transcript/115403	66,31	76,56	75,46	76,53	87,56	100	76,23	76,51	98,78	77,12	79,85	76,45	76,23	77,06	99,76	99,67	76,23	99,76	76,23	76,23	77,06		
7	Plant_Black_Jack_HQ_transcript/116155	65,91	99,08	88,97	91,23	75,45	76,23	100	92,14	76,57	100	65,55	66,56	100	96,08	76,46	76,54	100	76,46	100	100	96,08		
8	Plant_Black_Jack_HQ_transcript/121621	66,56	92,92	96,87	99,06	77,17	76,51	92,14	100	76,92	91,76	66,24	67,92	92,14	91,9	76,97	77,05	92,14	76,97	92,14	92,14	91,9		
9	Plant_Black_Jack_HQ_transcript/124097	66,4	77,04	75,92	76,94	87,71	98,78	76,57	76,92	100	77,27	80,42	77,05	76,57	77,5	99,07	99	76,57	99,07	76,57	76,57	77,5		
10	Plant_Black_Jack_HQ_transcript/136685	66,43	99,12	90,52	91,84	79,5	77,12	100	91,76	77,27	100	69,9	71,24	100	96,08	77,32	77,41	100	77,32	100	100	96,08		
11	Plant_Black_Jack_HQ_transcript/17032	57,14	65,94	64,77	65,86	77,46	79,85	65,55	66,24	80,42	69,9	100	84,64	65,55	65,83	81,19	81,19	65,55	81,19	65,55	65,55	65,83		
12	Plant_Black_Jack_HQ_transcript/20221	57,45	67,08	66,41	67,22	82,65	76,45	66,56	67,92	77,05	71,24	84,64	100	66,56	67,01	77,67	77,67	66,56	77,67	66,56	66,56	67,01		
13	Plant_Black_Jack_HQ_transcript/52365	65,91	99,08	88,97	91,23	75,45	76,23	100	92,14	76,57	100	65,55	66,56	100	96,08	76,46	76,54	100	76,46	100	100	96,08		
14	Plant_Black_Jack_HQ_transcript/55211	66,2	97	88,8	91,07	75,83	77,06	96,08	91,9	77,5	96,03	65,83	67,01	96,08	100	77,3	77,38	96,08	77,3	96,08	96,08	100		
15	Plant_Black_Jack_HQ_transcript/76946	66,54	76,85	75,96	76,98	88,2	99,76	76,46	76,97	99,07	77,32	81,19	77,67	76,46	77,3	100	99,92	76,46	100	76,46	76,46	77,3		
16	Plant_Black_Jack_HQ_transcript/90310	66,54	76,93	76,04	77,06	88,12	99,67	76,54	77,05	99	77,41	81,19	77,67	76,54	77,38	99,92	100	76,54	99,92	76,54	76,54	77,38		
17	Plant_Black_Jack_HQ_transcript/90891	65,91	99,08	88,97	91,23	75,45	76,23	100	92,14	76,57	100	65,55	66,56	100	96,08	76,46	76,54	100	76,46	100	100	96,08		
18	Plant_Black_Jack_HQ_transcript/95595	66,54	76,85	75,96	76,98	88,2	99,76	76,46	76,97	99,07	77,32	81,19	77,67	76,46	77,3	100	99,92	76,46	100	76,46	76,46	77,3		
19	Plant_Black_Jack_HQ_transcript/96170	65,91	99,08	88,97	91,23	75,45	76,23	100	92,14	76,57	100	65,55	66,56	100	96,08	76,46	76,54	100	76,46	100	100	96,08		
20	Plant_Black_Jack_HQ_transcript/97943	65,91	99,08	88,97	91,23	75,45	76,23	100	92,14	76,57	100	65,55	66,56	100	96,08	76,46	76,54	100	76,46	100	100	96,08		
21	Plant_Black_Jack_HQ_transcript/98322	66,2	97	88,8	91,07	75,83	77,06	96,08	91,9	77,5	96,03	65,83	67,01	96,08	100	77,3	77,38	96,08	77,3	96,08	96,08	100		

D

1	HQT3_MK598075_	100	85,68	85,68	85,87	85,41	85,41	85,87	85,9	85,87	85,41	85,87	86,41	85,41	85,68	85,4	85,87	85,92	85,68	86,43	85,68
2	Plant_Black_Jack_HQ_transcript/32620	85,68	100	100	96,05	94,83	94,83	96,05	96,03	95,97	94,83	96,05	95,51	94,83	100	95,51	96,05	96,42	100	95,49	100
3	Plant_Black_Jack_HQ_transcript/72570	85,68	100	100	96,05	94,83	94,83	96,05	96,03	95,97	94,83	96,05	95,51	94,83	100	95,51	96,05	96,42	100	95,49	100
4	Plant_Black_Jack_HQ_transcript/90541	85,87	96,05	96,05	100	97,65	97,65	100	100	99,92	97,65	100	98,13	97,65	96,05	98,41	100	99,32	96,05	98,12	96,05
5	Plant_Black_Jack_HQ_transcript/90548	85,41	94,83	94,83	97,65	100	100	97,65	97,8	97,73	100	97,65	99,92	100	94,83	99,32	97,65	98,25	94,83	99,92	94,83
6	Plant_Black_Jack_HQ_transcript/91401	85,41	94,83	94,83	97,65	100	100	97,65	97,8	97,73	100	97,65	99,92	100	94,83	99,32	97,65	98,25	94,83	99,92	94,83
7	Plant_Black_Jack_HQ_transcript/95365	85,87	96,05	96,05	100	97,65	97,65	100	100	99,92	97,65	100	98,13	97,65	96,05	98,41	100	99,32	96,05	98,12	96,05
8	Plant_Black_Jack_HQ_transcript/97320	85,9	96,03	96,03	100	97,8	97,8	100	100	99,92	97,8	100	98,28	97,8	96,03	98,56	100	99,32	96,03	98,28	96,03
9	Plant_Black_Jack_HQ_transcript/98418	85,87	95,97	95,97	99,92	97,73	97,73	99,92	99,92	100	97,73	99,92	98,21	97,73	95,97	98,48	99,92	99,24	95,97	98,2	95,97
10	Plant_Black_Jack_HQ_transcript/101284	85,41	94,83	94,83	97,65	100	100	97,65	97,8	97,73	100	97,65	99,92	100	94,83	99,32	97,65	98,25	94,83	99,92	94,83
11	Plant_Black_Jack_HQ_transcript/103969	85,87	96,05	96,05	100	97,65	97,65	100	100	99,92	97,65	100	98,13	97,65	96,05	98,41	100	99,32	96,05	98,12	96,05
12	Plant_Black_Jack_HQ_transcript/105693	86,41	95,51	95,51	98,13	99,92	99,92	98,13	98,28	98,21	99,92	98,13	100	99,92	95,51	99,35	98,13	98,62	95,51	99,84	95,51
13	Plant_Black_Jack_HQ_transcript/108057	85,41	94,83	94,83	97,65	100	100	97,65	97,8	97,73	100	97,65	99,92	100	94,83	99,32	97,65	98,25	94,83	99,92	94,83
14	Plant_Black_Jack_HQ_transcript/108970	85,68	100	100	96,05	94,83	94,83	96,05	96,03	95,97	94,83	96,05	95,51	94,83	100	95,51	96,05	96,42	100	95,49	100
15	Plant_Black_Jack_HQ_transcript/110829	85,4	95,51	95,51	98,41	99,32	99,32	98,41	98,56	98,48	99,32	98,41	99,35	99,32	95,51	100	98,41	98,94	95,51	99,35	95,51
16	Plant_Black_Jack_HQ_transcript/113476	85,87	96,05	96,05	100	97,65	97,65	100	100	99,92	97,65	100	98,13	97,65	96,05	98,41	100	99,32	96,05	98,12	96,05
17	Plant_Black_Jack_HQ_transcript/116876	85,92	96,42	96,42	99,32	98,25	98,25	99,32	99,32	99,24	98,25	99,32	98,62	98,25	96,42	98,94	99,32	100	96,42	98,61	96,42
18	Plant_Black_Jack_HQ_transcript/120049	85,68	100	100	96,05	94,83	94,83	96,05	96,03	95,97	94,83	96,05	95,51	94,83	100	95,51	96,05	96,42	100	95,49	100
19	Plant_Black_Jack_HQ_transcript/121790	86,43	95,49	95,49	98,12	99,92	99,92	98,12	98,28	98,2	99,92	98,12	99,84	99,92	95,49	99,35	98,12	98,61	95,49	100	95,49
20	Plant_Black_Jack_HQ_transcript/124745	85,68	100	100	96,05	94,83	94,83	96,05	96,03	95,97	94,83	96,05	95,51	94,83	100	95,51	96,05	96,42	100	95,49	100

Figure S6. This figure is showing a percentage matrix index of acyltransferases from *B. pilosa*, the colours herein represent the following: yellow=truncated sequence, green= 100% identity, Red = >95% similarity and the data is in the following order, HCT(A), HQT1(B), HQT2(C) and HQT3 (D).

GACCACCTTCTGGATAAAACAATACACCAATCCCCAACCTCACGTGTCCTTAAACTGATTCCTCAATTTACACCCCCAACTTGTCGGACAATCTTTTTTTTT
ATTTTTTGTCCGAGAATCTTAAAAAACATCCGATGAAGCTAATAGTGAAAGAATCATCAATTATAAAACCCGCTAAACCGACTCCGGTTACCCGGATATGG
AACTCGAACCTCGACTTAGTCGTGGGTGCAATCCATATCCTAACGTTTTACTTCTACCGACCGAATGGGAGTTCGGGTTTTTTTTGATCCGGTTGTTATGAA
GGAAGCTTTAGCCAAAGTTCCTTGTTCGTTTTTCTATGGCCGGACGGTTGGCAAAGACCGTGATGGCAGGATTGAAATTAATTGTAACGGTGAGGGT
GTTTTGTTTGTGCGAAGCGGAAGCAGATTGTTGCATTGATGATTTGGGGAGATTACGCCGTCGCCGGAGTTGAGGCAGTTGGCGCCTACGGTGGATTATT
CCGGAGAGATTGATTCGTATCCGCTTGTATTACACAGGTTACACGATTTAAATGTGGTGGGGTTTTCTTAGGGTGTGGACTACATCATACATTATCAGAT
GGACTCTCATCTCTTCATTTATCAACACATGGTCTGACAAAGCTCGAGTTTATCAGTCCGAGTCCCACCATTCTTGATCGTACTCTTATTTCGAGCGCG
GAACCCACCTACACCAATGTTTGACCATGTTGAGTATCACCAACCACCATCAATGATTGTCCATCGGAAAACCAAAAATCCCCATCTCACTCCAAGTCCA
CATCAACCGTGATGCTACGTCTCACACTTGATCAGTTAAATGATCTTAAACTAAAGGCAAAGGCGATGAAAGCGCACATCATAGCACATATGATATCCTA
GCCGCTCATCTATGGCGATGTGCGTGTAATCACGTGGACTCTTAGATGATCAACCAACTAAATTGTACGTGGCTACTGATGGACGGTCAAGATTGAACC
CGCCACTCCCTCCCGTTACCTTGGGAATGTCATTTTCACTGCCACCCCAATCATGAAAGTAGGCGAGTTTAAAGTCTGAGTCGTTAGGGGACACTGCAAG
GAGAATCCATAATGAGTTGGCTAGAATGGACGATCAATATCTTAGATCAGCTATTGACTACCTGGAGACAATATCTGATCTATCAACTCTTGTTCGTGGGC
CATCTTACTTTGCGAGTCCAAATCTGAATGTAACAGTTGGACTCGCTTACCCATCTATGACTCTGATTTCCGGTGGGGACGGCCCATTTTCATGGGACCT
GCGAGCATTCTCTATGAAGGCACGATTTATATCATACCGAGCTCGAGTGATGACCGGAGTGCAAGTTGGCGGTGTGCTTGGACTCGGATCATATGACTT
TGTTTAAGGAATGCTTGTATGATTTCTAGCGAAGTGATGAATTATAGATGTAATAAGGC

>|c||ORF

MKLIVKESSIIKPAKPTPVTRIWNNSLDL VVGRIHILTVYFYPNGSSGFFDPVVMKEALAKVLVSFFPMAGRLAKDRDGRIEINCNGEVLVFEAEADCCIDDFGEI
TPSPELRQLAPTVDYSGEIDSYPLVITQVTRFKCGGVSLGCGLHHTLSDGLSSLHFINTWSDKARGLSVAVPPFLDRTLIRARNPPTPMFDHVEYHQPPSMIVPS
ENQKSPSHSKSTSTVMLRLTL DQLNDLKLKAKGDESAAHSTYDILAAHLWRCACKSRGLLDDQPTKLYVATDGRSRLNPPLPPGYLGNVIFTATPIMKVGEFKSE
SLGDTARRIHNELARMDDQYLRSAIDYLETISDLSTLVRGPSYFASP NLNVNSWTRLPIYDSDFGWGRPIFMGPASILYEGTIYIIPSSSDDRSVKLAVCLDS DHMT
LFKECLYDF

Nucleotide sequence 1. HQT1 nucleotide sequence from *B. pilosa* as obtained from the SMRT sequencing technique. This sequence was used to generate ORF herein.

GGTGAACAAAACCCAATCTCTTTAGCTCCTTAAACTCTCCAATCACACACACATATCTTTGTGTGTAACCTTAAAAAACATGAAGATGAACATAACTATAA
CAAACATCAATCATTCTCCATCAAAAACACACCAGATGCTCCTAAACACCTATACACCTCTAACTTGGACCTTATTGTTGGCAGGATCCATCTCCTAA
CCGTTTACTTCTACCGACCGAACGGTTCCGGCTAACTTTTTCGACCCAAAGGTCATGAAAAAGGCGCTAGCCGATGTAAGTCTCGTTCTACCCAATGGC
AGGGCGGTTGGGTAGAGACGAGACGGGTAGGATCGTTATTAATTGTAATAACGAAGGCGCTTTTTCGTTGAAGCGGAATCGGATTCGAGTTTGGATGAT
TTTGGAGAGTTTACTCCGTCACCCGAGTTTAAAAGTCTTACGCCGAATGTTGACTGCTCGGGTGATATTTCTTCGTATCCGTTGTTTTTCGCACAGGTAAC
CATTTCAAATGTGGAGGAGTAGGTCTTGGTTGTGGTGTGTTCCATACATTATCAGATGGTTTATCCTCACTCCATTTTATAAACACATGGTCCGATGTGGCT
TGCGGGTTGTCTGTAGCCATCCCACCATTCAATTGACCGGACCTTACTACGTGCACGTGACCCACCCACCCCAACCTATGACCATGTGGAATACCACCCAC
CACCCGTCATGAACACGGTCACCCAAAAATCGGGTTCACCTTTCTAAATCATCAACCACCATGCTAAAACTCACACTAGACCAACTCAACACTCTCAAAGCT
AAAGCCAAGAATGACGGTGGGCCCAACCATAGCACGTACGCTGTCCTAGCCGCTCATATATGGCGGTGCGCGAGCAAGGCTCGTGGGCTCTCAAATGA
CCAGTTGACCAAACCTTTACATAGCCACAGATGGGAGGTCGAGACTTAGTCCCCAGCTCCCACCTGGCTACCTTGGGAATGTGATCTTTACAGCCACCCCA
ATTGCTAAATCAGGTGATCTGACATCAGGATCATTGTTGAACACTGCAAACTCATTACACCACGTTAAGCAAATGGACAATGATTATTTGAGATCAGCT
ATTGATTACCTTGAGTTACAACCCGACCTATCAGCTCTTATTCGTGGACCTGGCTACATTGCTAGCCGAATTTAAACATAAACGCTTGGACCAGACTTCTT
GTGCATGACGCGGATTTTGGGTGGGGTCGGCCGATCTTTATGGGGCCCTCGATTGTATTGTATGAGGGGACCATATATATTCTACCTAGCCCAAACAATG
ATAGGAGTGTGCTTTTGGCTGTTTGTTTAGATGCAAAGGAACAACCACTTTTTGAGAAGTACTTGTATGAGATTTAAGGTTTTTGAACCAAACAAGTTGTGG
GCTGTTTTGGTTTTTGGCTTGAATTTTCATTTTTGTGATAATAAGAGTGTGTTAGGACTTTAGGAGCCCATGATGATTGTAATTATAAGTTATTGAGTAGAAC
TTTATTAACCTCGTGTTT

>|c|ORF

MKMNITITNSSIIPPSKTPDAPKHLYSNLDLIVGRIHLLTVYFYRPNGSANFFDPKVMKKALADVLVSFYPMAGRLGRDETGRIVINCNNEGALFVEAESDSSLD
DFGEFTPSPEFKSLTPNVDCSGDISSYPLFFAQVTHFKCGGVGLGCGVFHTLSDGLSSLHFINTWSDVACGLSVAIPPFIDRTLLRARDPPTPTYDHVEYHPPPV
MNTVTQKSGSLSKSSTTMLKLTLDQLNLTAKAKNDGGPNHSTYAVLAAHIWRCASKARGLSNDQLTKLYIATDGRSRLSPQLPPGYLGNVIFTATPIAKSGDLT
SGSLLNTAKLIHTTSLKMDNDYLRSIDYLELQPDLSALIRGPGYIASPNLNINAWTRLPVHDADFGWGRPIFMGPSIVLYEGTIYILPSPNDRSVSLAVCLDAKEQ
PLFEKLYEI

Nucleotide sequence 2. HQT2 nucleotide sequence from *B. pilosa* as obtained from the SMRT sequencing technique. This sequence was used to generate ORF herein.

GTGTCTTCTTTAAACCTTCTTTTCCTTCATCAATGGGATCTTCTGATCACATGAAGCTGAACATAAACATCAAACACTCAACACTCATAACAACCATCCAAGCC
CACACAGGCTAATTCCACTAAGCAGCTATGGACGTCAAACCTGGACTTGGTGGTTGGCAGGATCCATATTCTGACGGTCTACTTCTACCGTCCGACCGGT
GCCGCCAATTCTTTGACCCGGTTGTCATGAAGAAGGCGTTGGCGGATGTGCTGGTTGCCTTTTATCCGATGGCTGGTGAATGGGTAAAGATGAGAAT
GGCAGAGTTGTGATTAATTGTAATGATGAGGGTGTGTTTGTGTTGTTGAAGCCGAGTCCGATTCCACGTTGGATGACTTCGGTGAGTTCACGCCGTGCGCCGG
AGCTAAGGCGGTTGACGCCACGGTTGACTATTCCGGTGGCATTCTACTTACCCTCTGTTTTTGTCTCAGGTGACACATTTCAAATGTGGAGGAGTTGGT
CTTGGTTGTGGTGTGTTTCATACATTAGCCGATGGTCTATCCTCAATACATTTTCATCAACACATGGTCCGACATGGCTCGGGGCCTTTCCATAGCCATCCC
ACCGTTCATTGACCGCTCCTTGTCTCGTGCACGTGAACCACCCACTCCCACGTTTGACCATGTGCAATACCACCCCCACCGTCGATGAAAACCGCTCCC
AAATCCACCCGAAAACCGTCCACCACGATCCTAAAGCTCACCCCTTGATCAACTCAATGCTCTCAAAGCCGCAGCCAAGAATAATAGTGGAAACGTCAACTA
TAGCACGTACGAGATCCTCGCGGCTCACTTATGGCGCTGCACGTGCAAGGCTCGTGGGCTCCCAGACGACCAACTAACCAAACCTTTACGTGGCTACTGA
TGGGCGGTCAAGACTGAGCCCCAGCTCCCCCAGGGTACCTAGGCAACGTGGTGTTCACCGCCACACCAATCGCCAATCAGCTGACCTCACAACCTG
GACCATTGTCCAATGCAGCATCTTTGATCCGGGCCACTTTGTCCAAAATGGACAACGACTATTTGAGATCCGCCATTGACTACCTCGAGGTGCAGCCCGA
TCTCTCGGCCTTGATTCGTGGCCCGAGTTACTTTGCAAGCCCAAACCTTGAACATAAACACGTGGAACCGGTTGCCTGTTTCATGATGCGGATTTGCGGTGG
GGTAGGCCCGTGTTCATGGGTCCGGCGTGTATATTATATGAGGGGACGATTTATGTTCTACCGAGCCCAAATAATGATAGGAGTATGTCACTCGCGGTGT
GTTTGGATGCTGAGGAGCAGCCATTGTTTGAAGAAGTTCTTATATGACTTCTAAGGATTGTGATGCAGCTGATGCTTGTTAATTCTGTGCATCAAATTTGTG
AAAATTGAGAAGGGCAATATTGGATTTTATTATTTTTTTTTTTCTTCTTTTTTTGGATTTTGAATGCAGGCAATTAGCTTTGTTAGTTGTGACAACTTTATAGTGT
ACTGGAAAAGGAGGCTGAGAATTATTATATGTACAAGTACAACATGTAC

>|c|ORF

MGSSDHMKLNINIKHSTLIQPSKPTQANSTKQLWTSNLDLVVGRIHILTVFYRPTGAANFFDPVVMKKALADVLVAFYPMAGRMGKDENGRVVINCNDEGVLFV
EAESDSTLDDFGEFTPSPELRRLTPTVDYSGGISTYPLFFAQVTHFKCGGVGLGCGVFHTLADGLSSIHFINTWSDMARGLSIAIPPFIDRSLLRAREPPTPTFDHV
EYHPPPSMKTAPKSTRKPSTTILKLTLDQLNALKAAAKNNSGNVNYSTYEILAAHLWRCTCKARGLPDDQLTKLYVATDGRSRLSPQLPPGYLGNVFTATPIAK
SADLTTGPLSNAASLIRATLSKMDNDYLRSIDYLEVQPDLSALIRGPSYFASPNLNINTWNRLPVHDA^{DFGWGR}RPVFMGPACILYEGTIYVLPSPNNDRSMSLAV
CLDAEEQPLFEKFLYDF

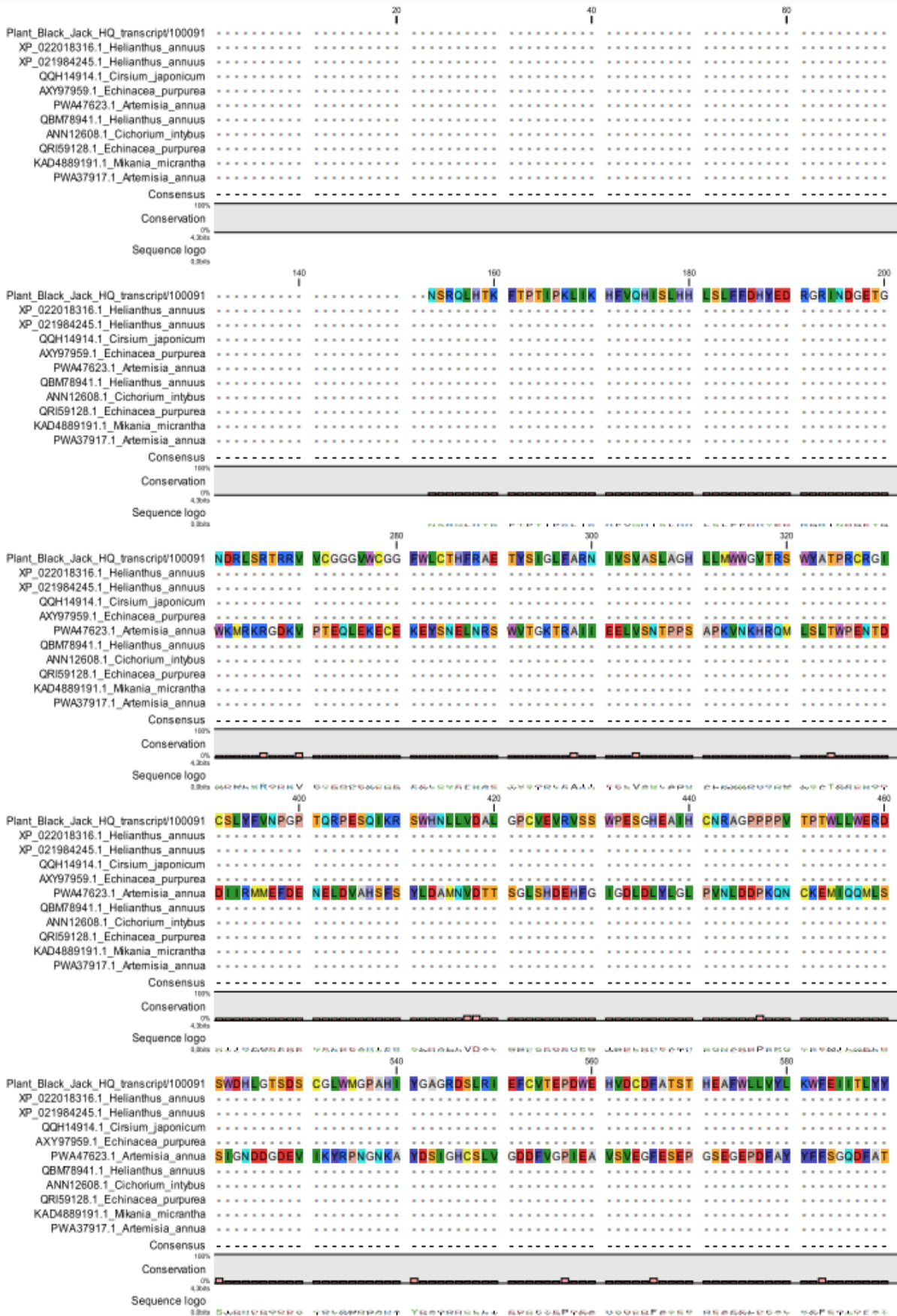
Nucleotide sequence 3. HQT3 nucleotide sequence from *B. pilosa* as obtained from the SMRT sequencing technique. This sequence was used to generate ORF in figure S6, C.

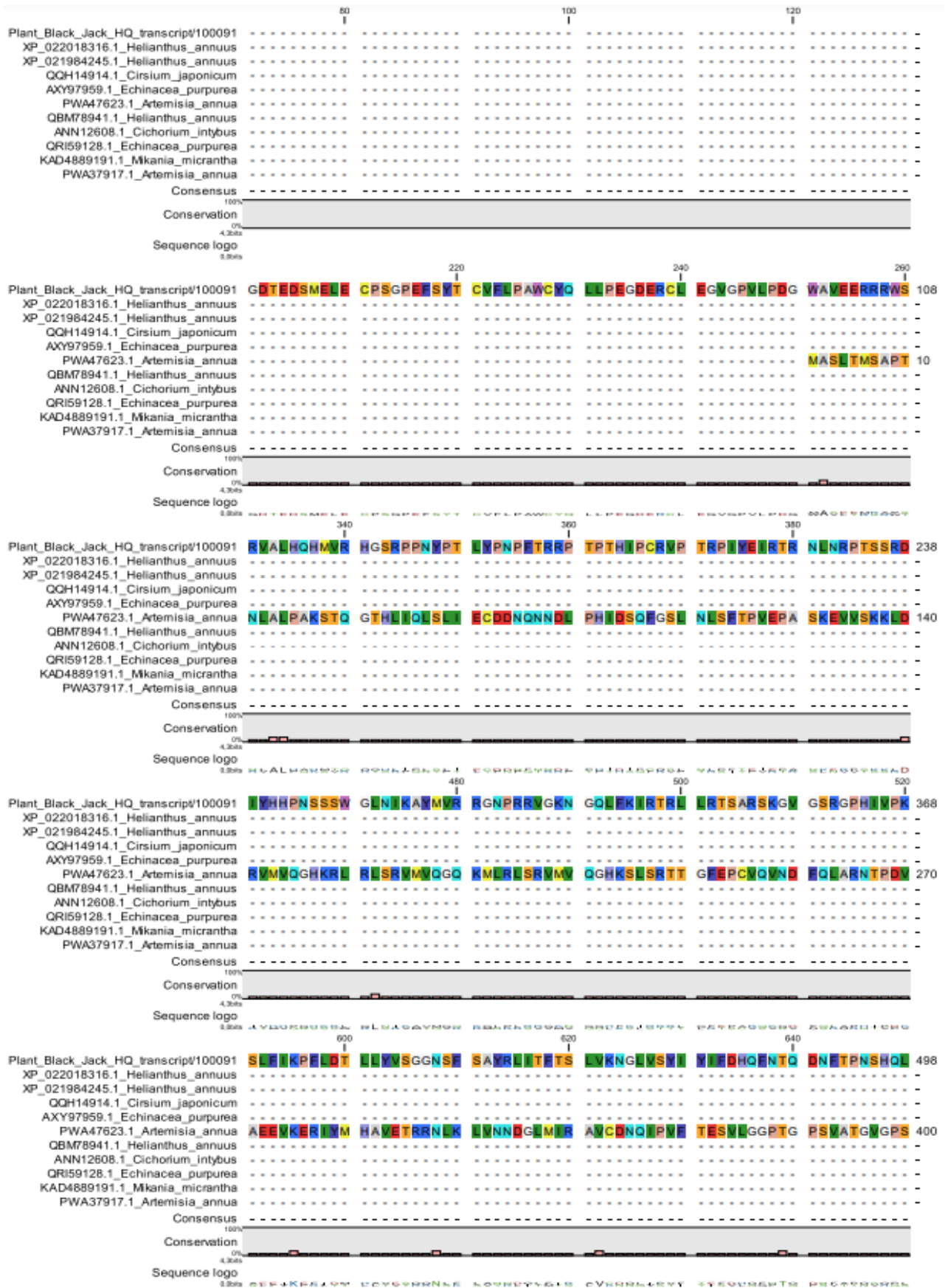
AACTCAAGACAACCTTCACACCAAATTCACACCAACTATCCCTAAATTAATCAAACATTTTGTCCAACACATTTCACTTCACCATTTGTCCTTGTTCTTTAAGA
TCATTAATATGAAGATCGAGGTTAGAGAATCAACGATGGTGAGACCGGCTGAGGAGACACCGAAGATTAATCTATGGAACCTCGAATGTTGACCTAGTGGT
CCCGAATTTTCATACACCTAGTGTGATTTTTACCGGCCTAATGGTGCTACCAACTTCTTTGACCCGAAGGTGATGAAAGATGCCTTGAGTAGGGCGTTGG
TCCCGTTTTACCCGATGGGTGGGCGGTTGAAGAAAGACGAAGATGGTCAATTGAGATCGATTGTCAAGGACAAGGCGTGTGTTTGTGGAGGGCGGAGT
CTGATGGTGTGGTGGATGATTTTGGTGACTTTGCACCCACTTTAGAGCTTAGGAACTTATTCCATCGGTTGATTATTCGCTAGGAATTGAATCGTATCCGT
TGCTAGTCTTGCAGGTCACTTACTTTAAATGTGGTGGGGTGTCACTAGGAGTTGGTATGCAACACCACGCTGCTGACGGGGCATCCGGGTTGCACTTCAT
CAACACATGGTCCGACATGGCTCGAGGCCTTGACCTAACTATCCCACCCTTTATTGACCGAACCTTTTACGCGCTAACGACCCACCCCGACCCACATTT
GACCATGTGAGTACCAACCCGCCCATCTATGAAATACGCACCCGAAACCTCAACCGACCAACAAGTTCGCGAGACTGCAGTCTCTATTTTTAAGTTAAC
CCGGGACCAACTCAACGCCCTGAAAGCCAAATCAAAGAAGCTGGTAACACAATTAGTTATAGCTCGTATGAGATGCTCTCGGGCCATGTGTGGAGGTGC
GTGTGTAAGCTCGTGGCCTGAATGATGATCAGGACACGAAGCTATACATTGCAACTGACGGGCGGGCCCGCCTCCGCCCGTCACTCCCACCTGGCTAC
TTTGGGAACGTGATATTTACCACCACCCCAATAGCAGTAGCTGGGGACTTAATATCAAAGCCTACATGGTACGCCGCGGGTAAAATCCACGACGCGTTGG
TAAGAATGGACAATGATTATTTAAGATCCGCACTCGATTACTTAGAACTTCAGCCGATCTAAAGGCGTTGGTTCGCGGGGGCCACACATTTAAGTGCCCA
AATCTTGGGATCACTAGTTGGGCTAGACTTCCGATTCATGATGCGGACTTTGGATGGGGCCGGCCATATTTATGGGGCCGGGAGGGATAGCTTATGAA
GGATTGAGTTTTGTGTTACCGAGCCCGATTAATGATGGGAGCATGTGCGATTGCGATTTGCTACAAGCTGAACACATGAAGCTTTTTAGTGGCTTCTTGTA
TGATATCTAAAATGGTTTGAGATTATAACTTTATATTACTCATTATTCATAAAACCTTTCTAGATACATTGTTGTATGTGAGTGGAGGCTGAAACAGTTTTTA
ATCAGCTTATCGGTTAATTACTTTTACATCATTGTAAGTAAAGAATGGTCTTGTATCTTATATATATATATTTGACCATCAATTTAATGT

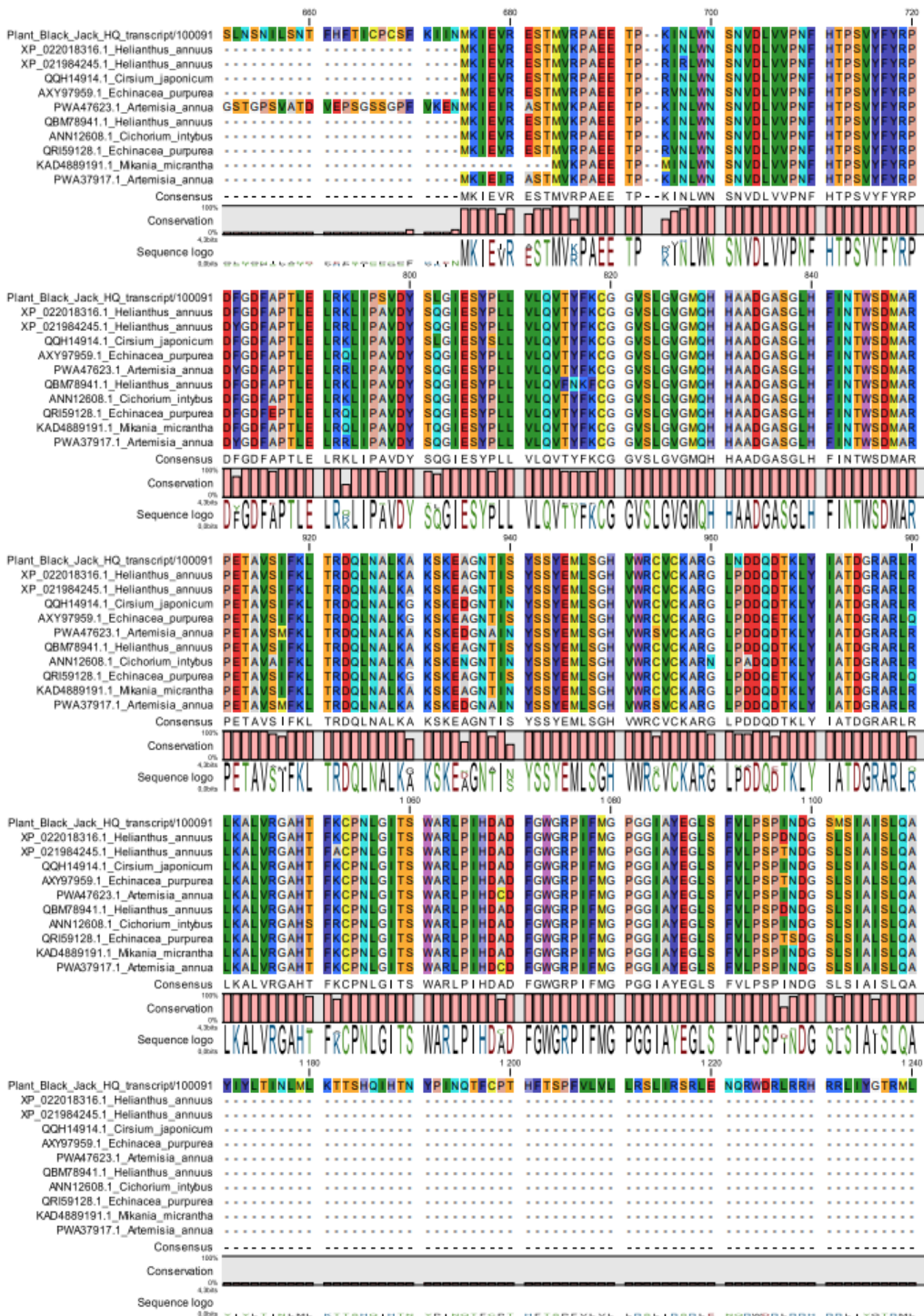
>|c|ORF1

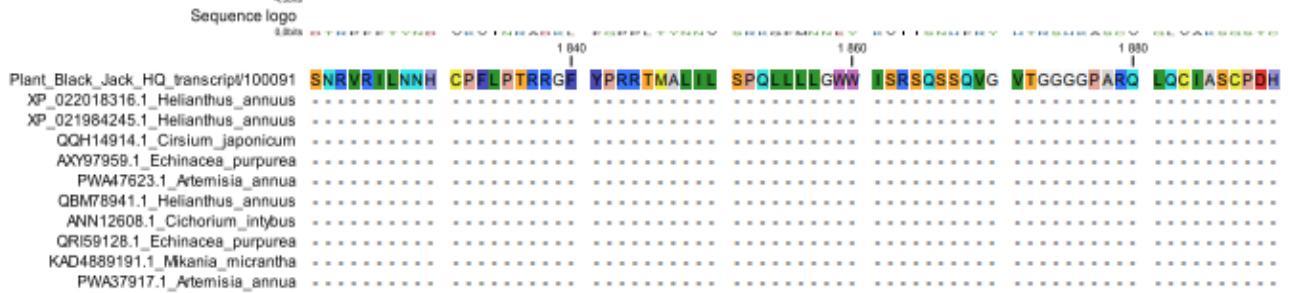
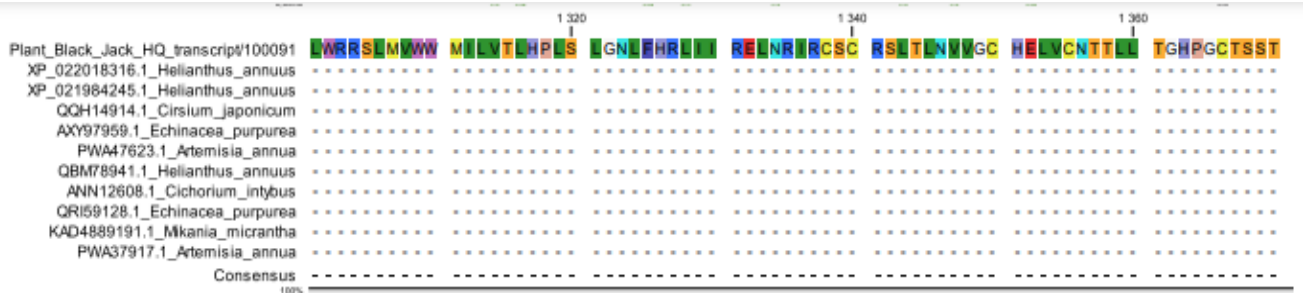
MKIEVRESTMVRPAEETPKINLWNSNVDLVVFNHFTPSVYFYRPNNGATNFFDPKVMKDALSRALVPFYPMGGRLKKDEDGRIEIDCQQGQVLFVEAESDGVVD
DFGDFAPTELELRKLIPSDYSLGIESYPLLVLQVTFYKCGGVSLGVGMQ**HHAAD**GASGLHFINTWSDMARGLDLTIPPFIDRTLRLRANDPPRPTFDHVEYQPAPS
MKYAPETSTDQQVPETAVSIFKLTRDQLNALKAKSKEAGNTISYSSYEMLSGHVWRCVCKARGLNDDQDTKLYIATDGRARLRPSLPPGYFGNVIFTTTTPIAVAG
DLISKPTWYAAGKIHDLVRMDNDYLRSDLYLELQPDALKALVRGAHTFKCPNLGITSWARLPIHDA**DFGWG**RPIFMGPGGIAIEGLSFVLPSPINDGSMISIAISL
QAEHMKLFSGFLYDI

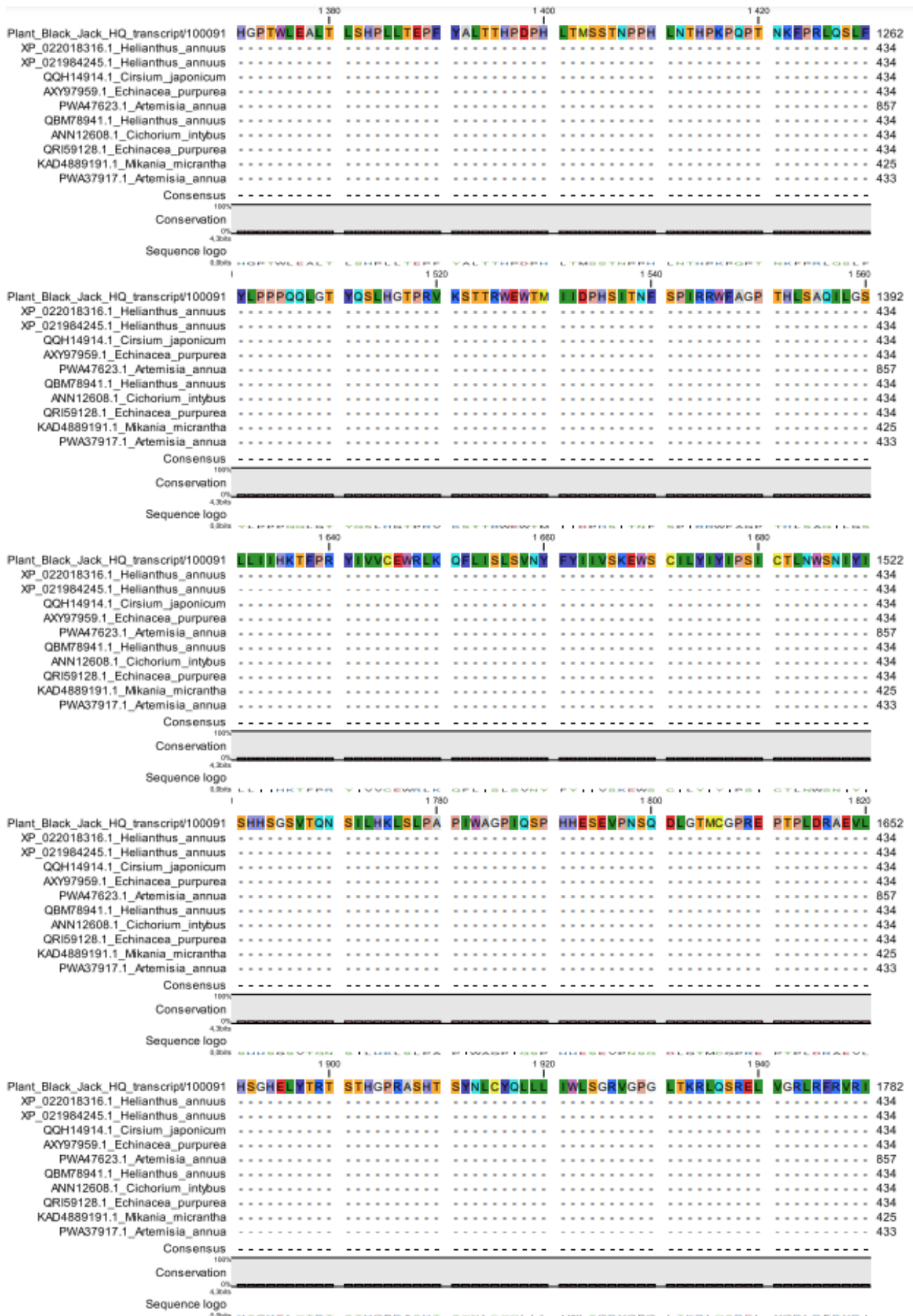
Nucleotide sequence 4. HCT nucleotide sequence from *B. pilosa* as obtained from the SMRT sequencing technique. This sequence was used to generate ORF herein.

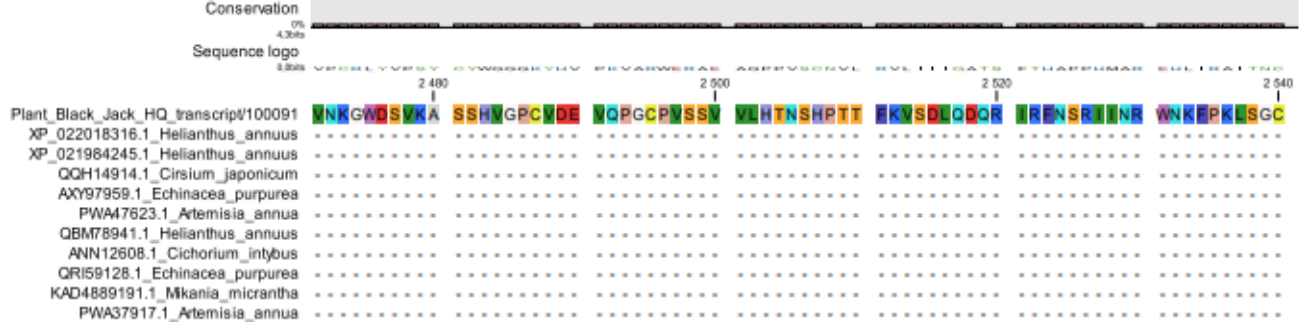
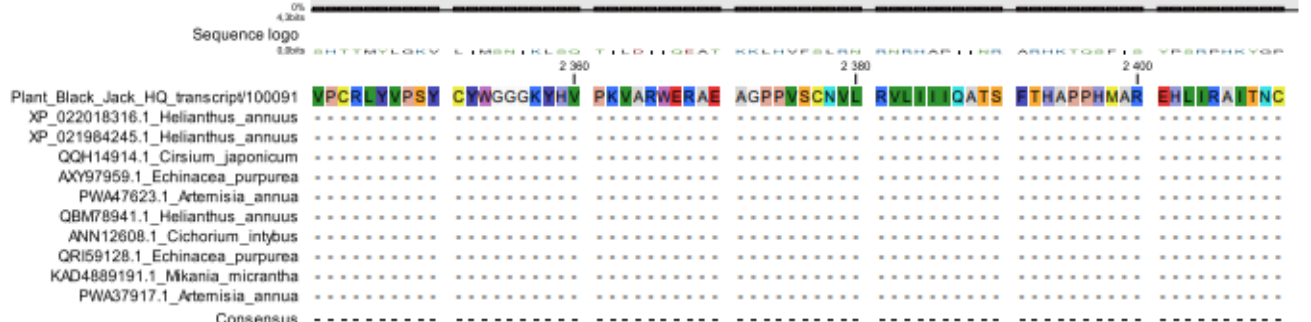
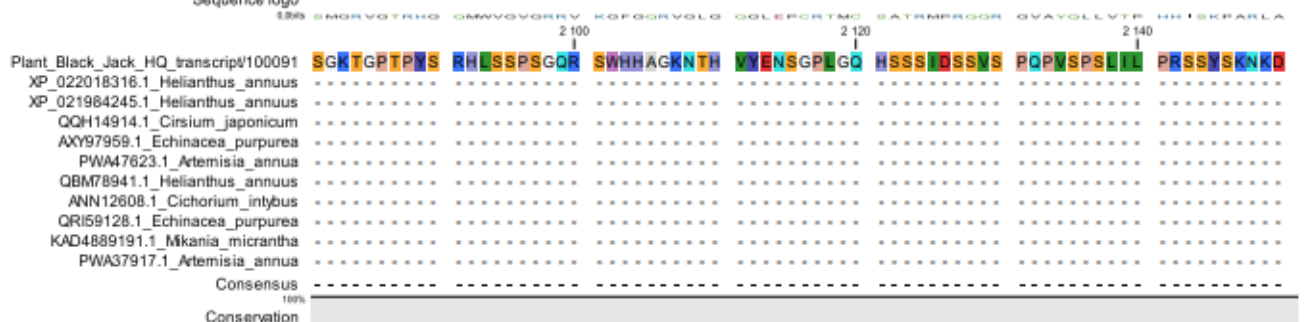
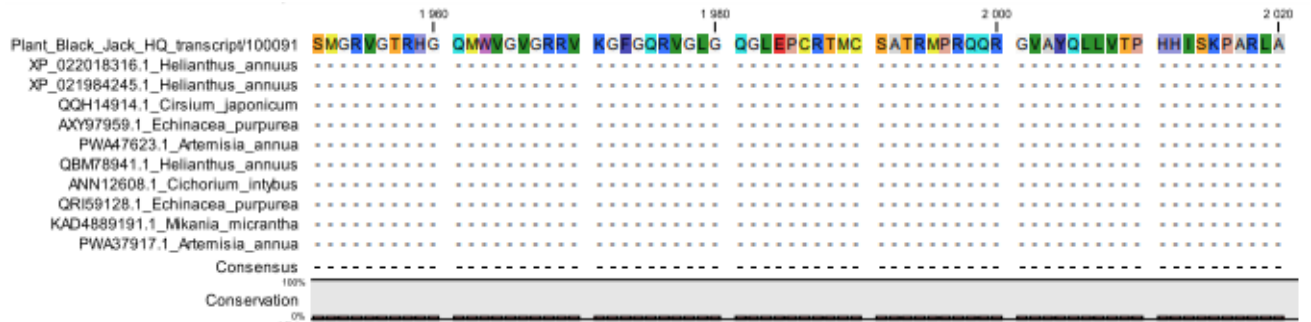


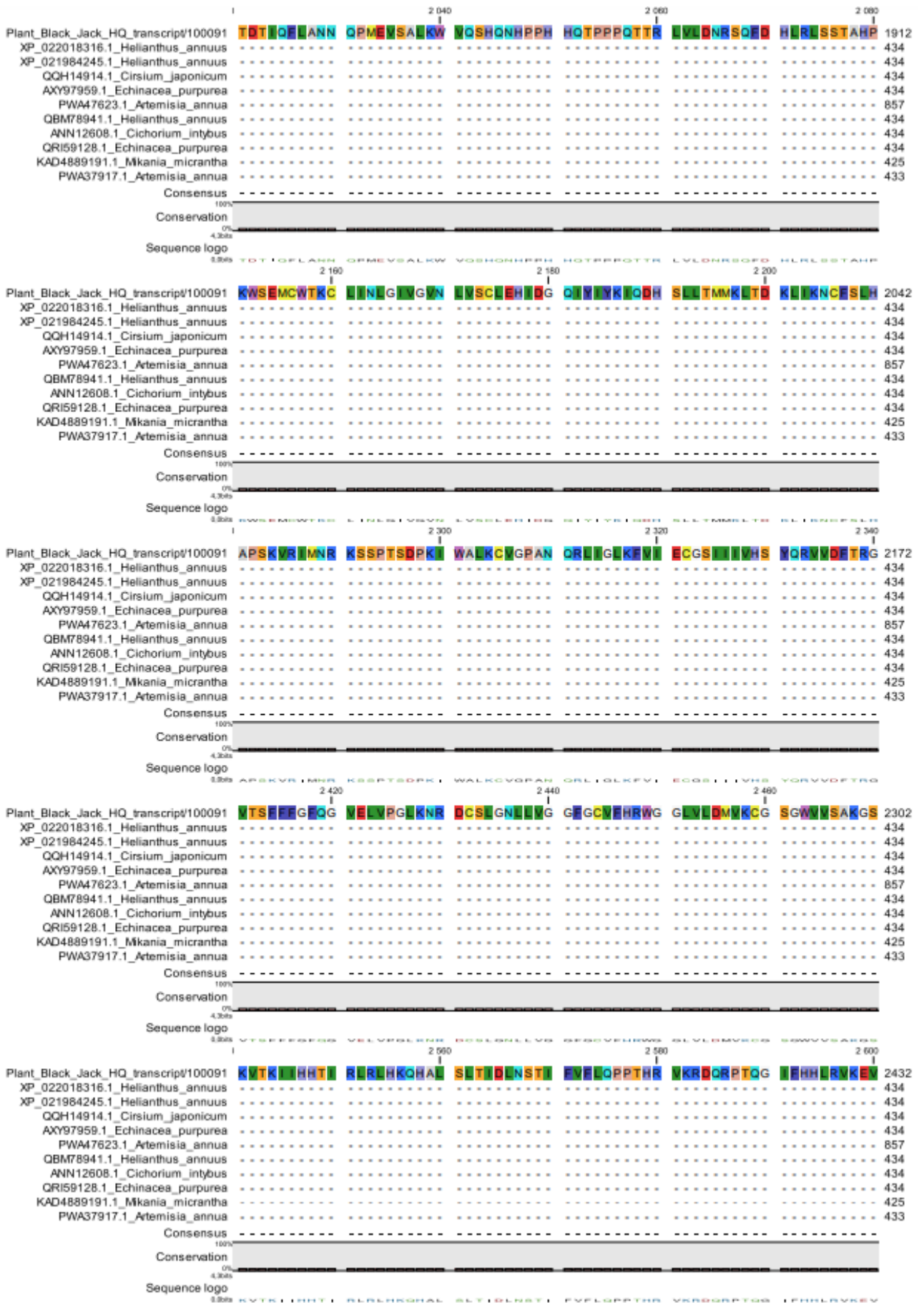


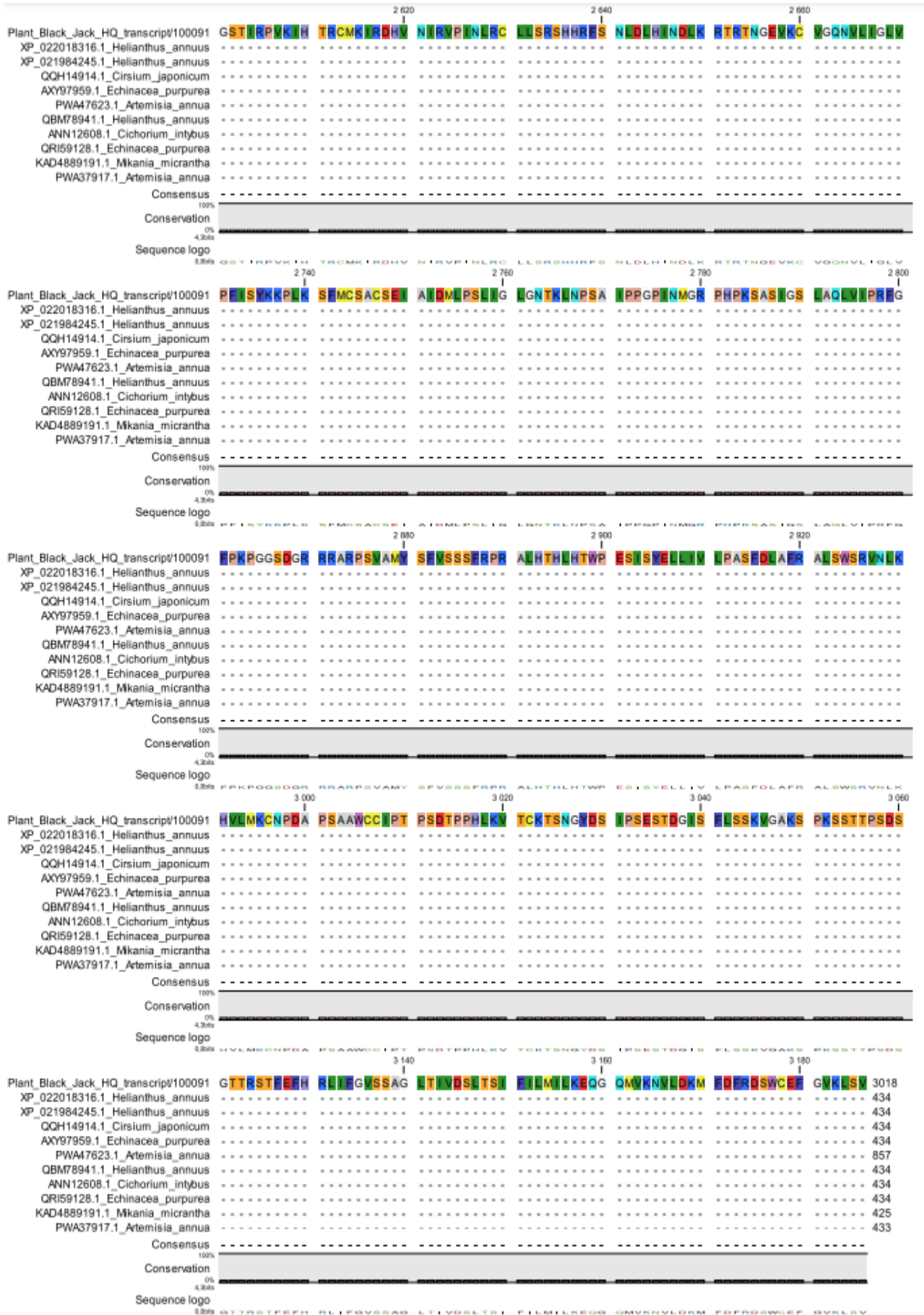












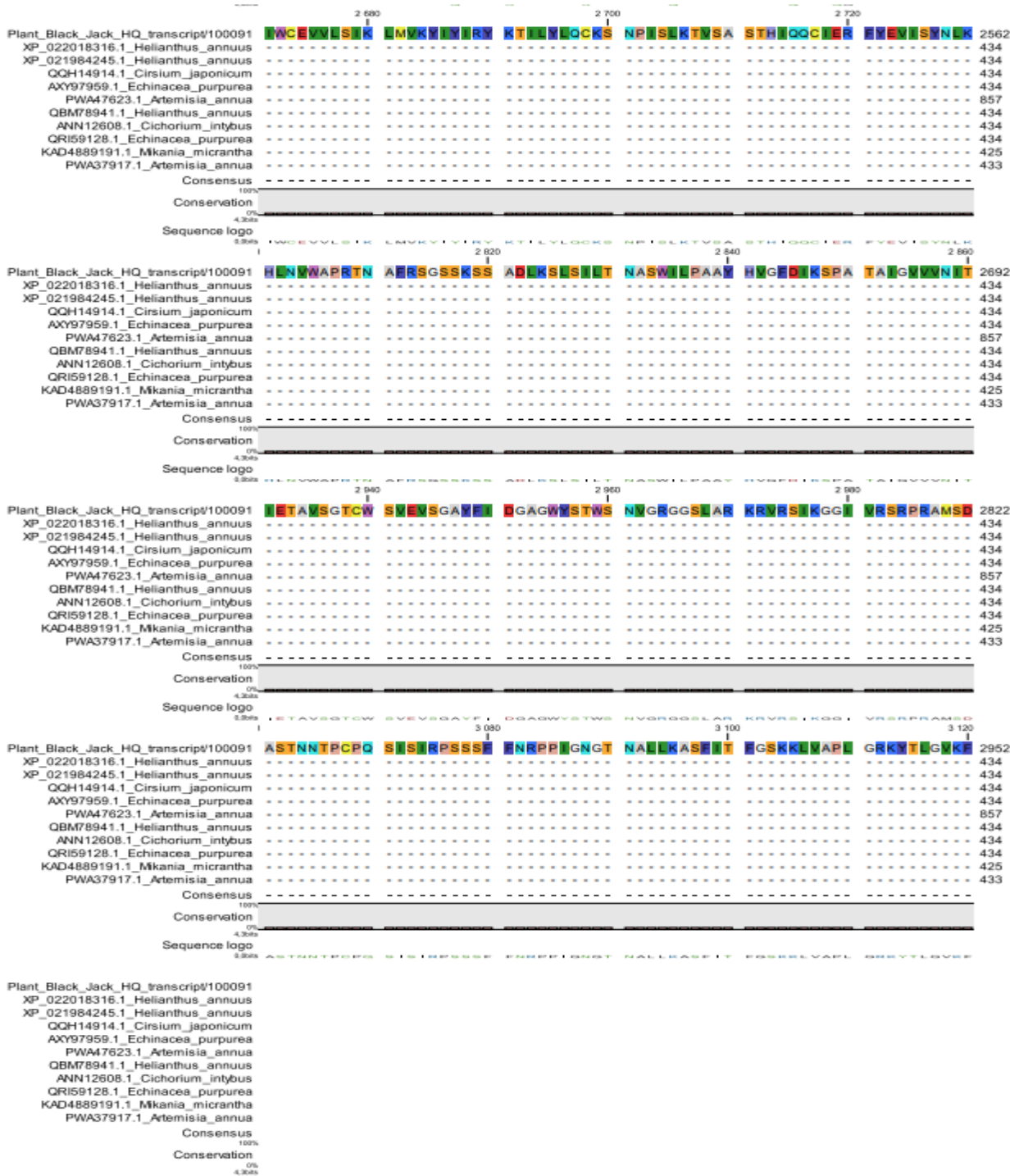
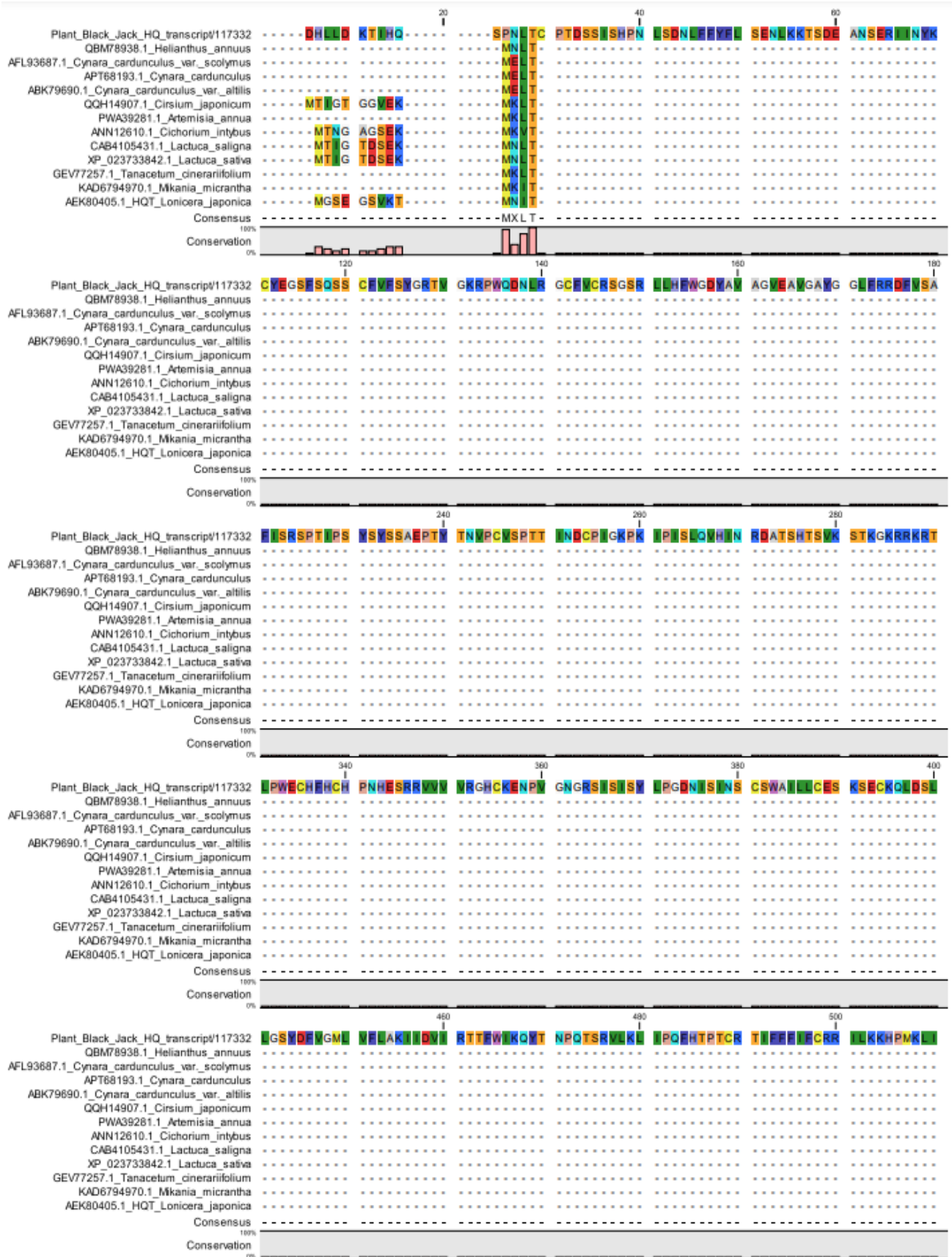
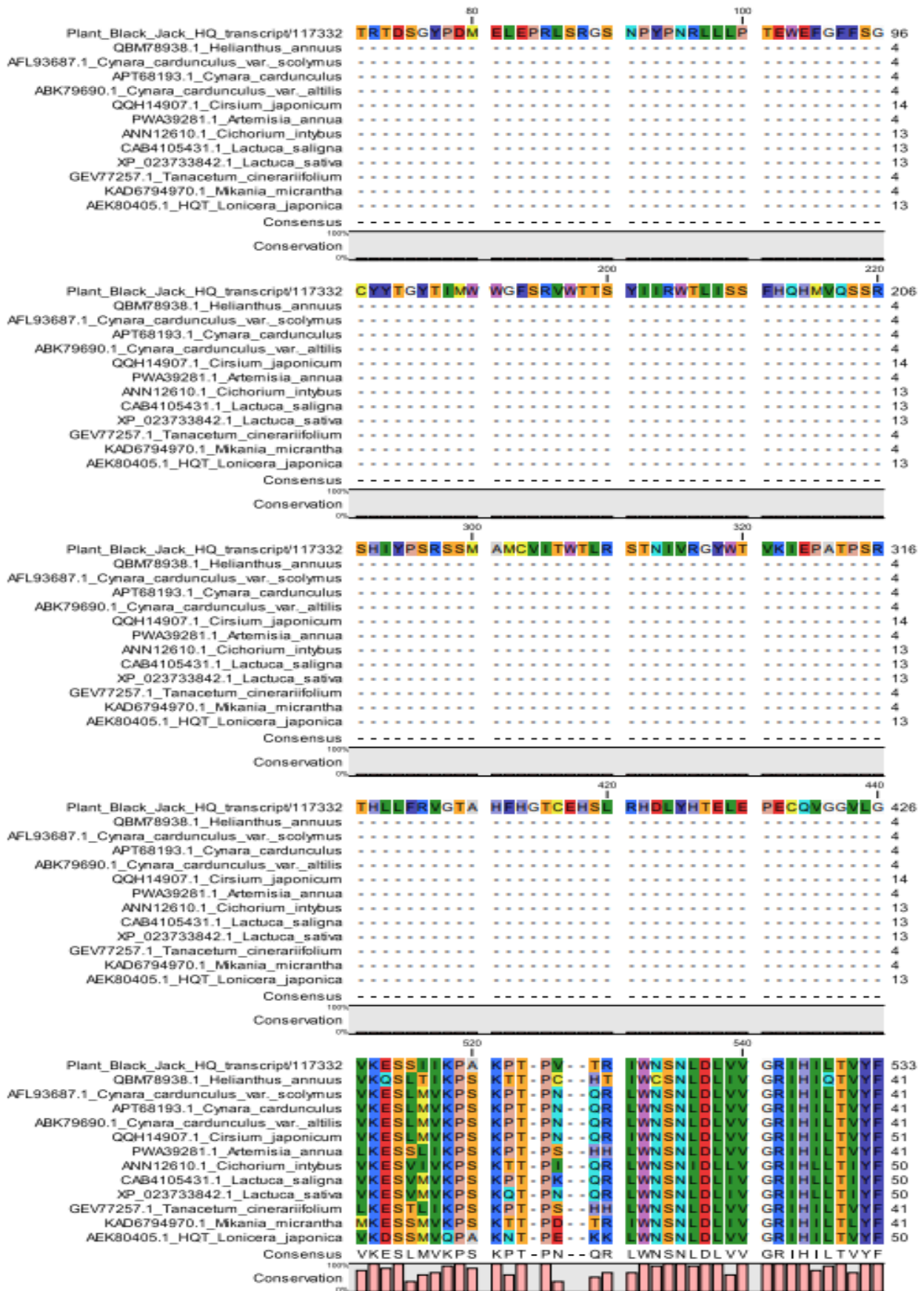
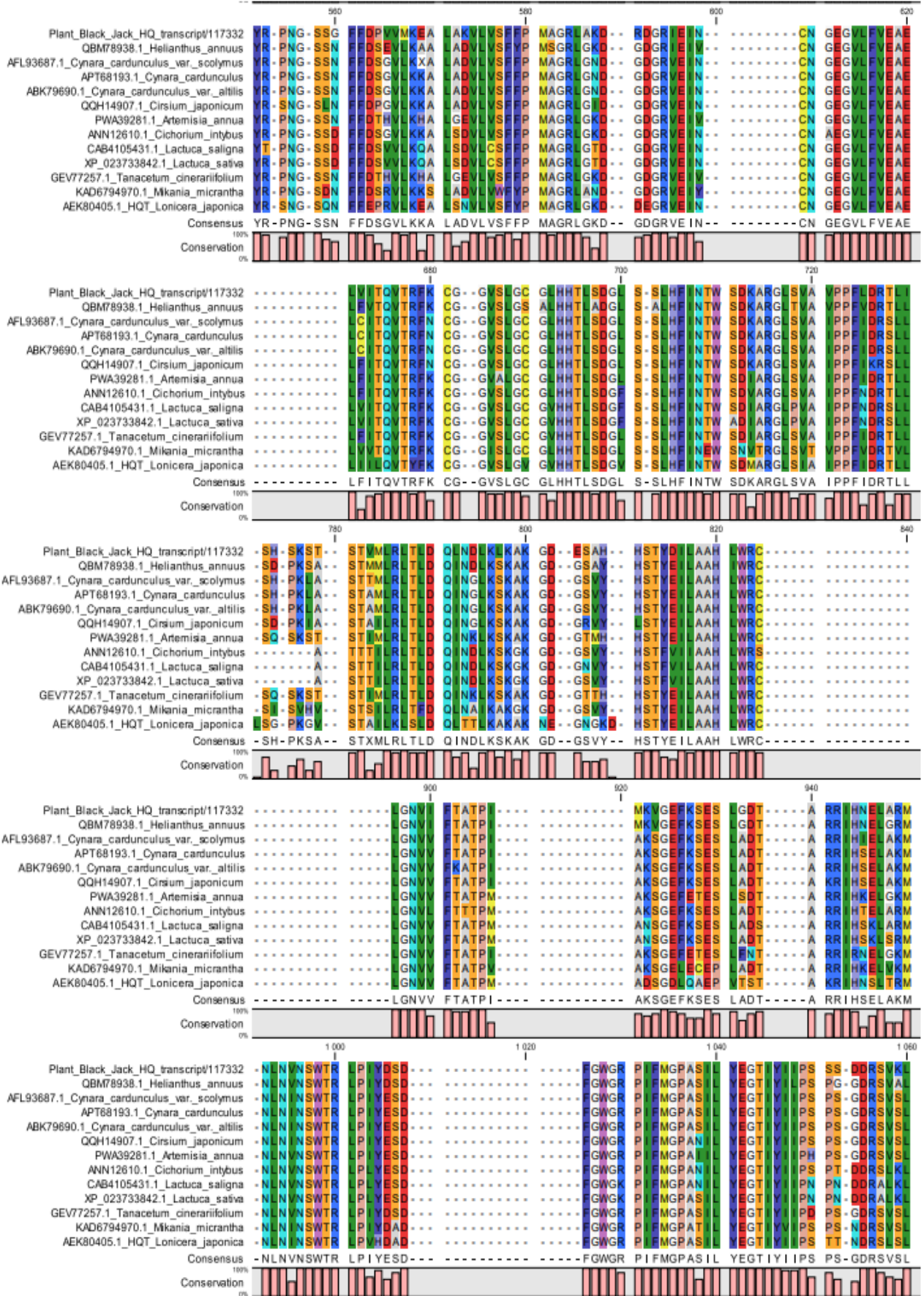


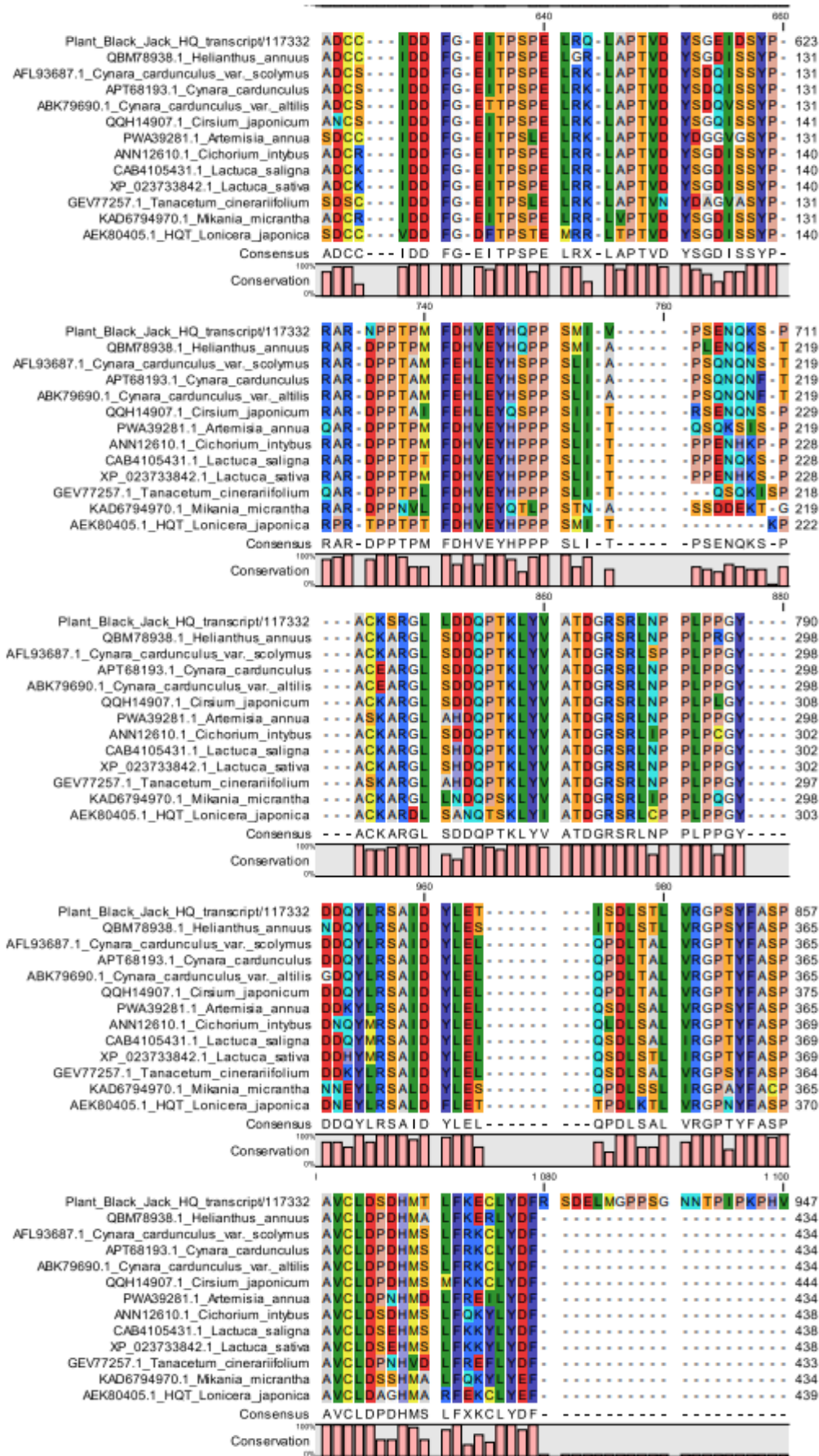
Figure S8. Multiple sequence alignment of *B. pilosa* HCT gene.

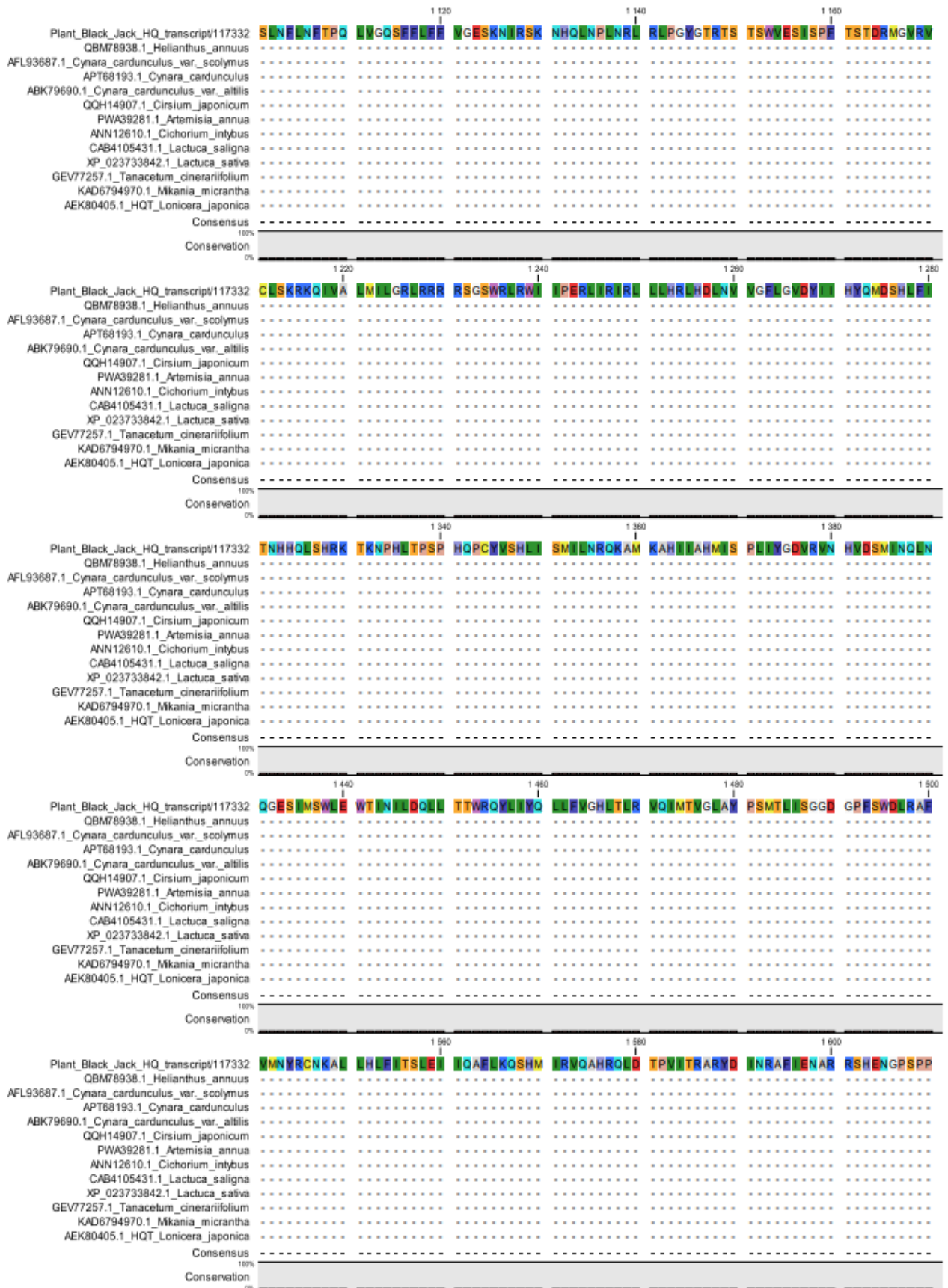
Multiple sequence alignment of *B. pilosa* HCT with its homologues from *Helianthus annuus* (XP_022018316.1), *Cirsium japonicum* (QQH14914.1), *Artemisia annua* (PWA37917.1), *Mikania micrantha* (KAD4889191.1), *Chicorium intybus* (ANN12608.1) and *Echinacea purpurea* (QRI59128.1). Residues are grouped according to colours, for instance same colour represent similar residues across all genes from different plants. The alignment was generated using MUSCLE. The position of the residue is shown by the number on the right.

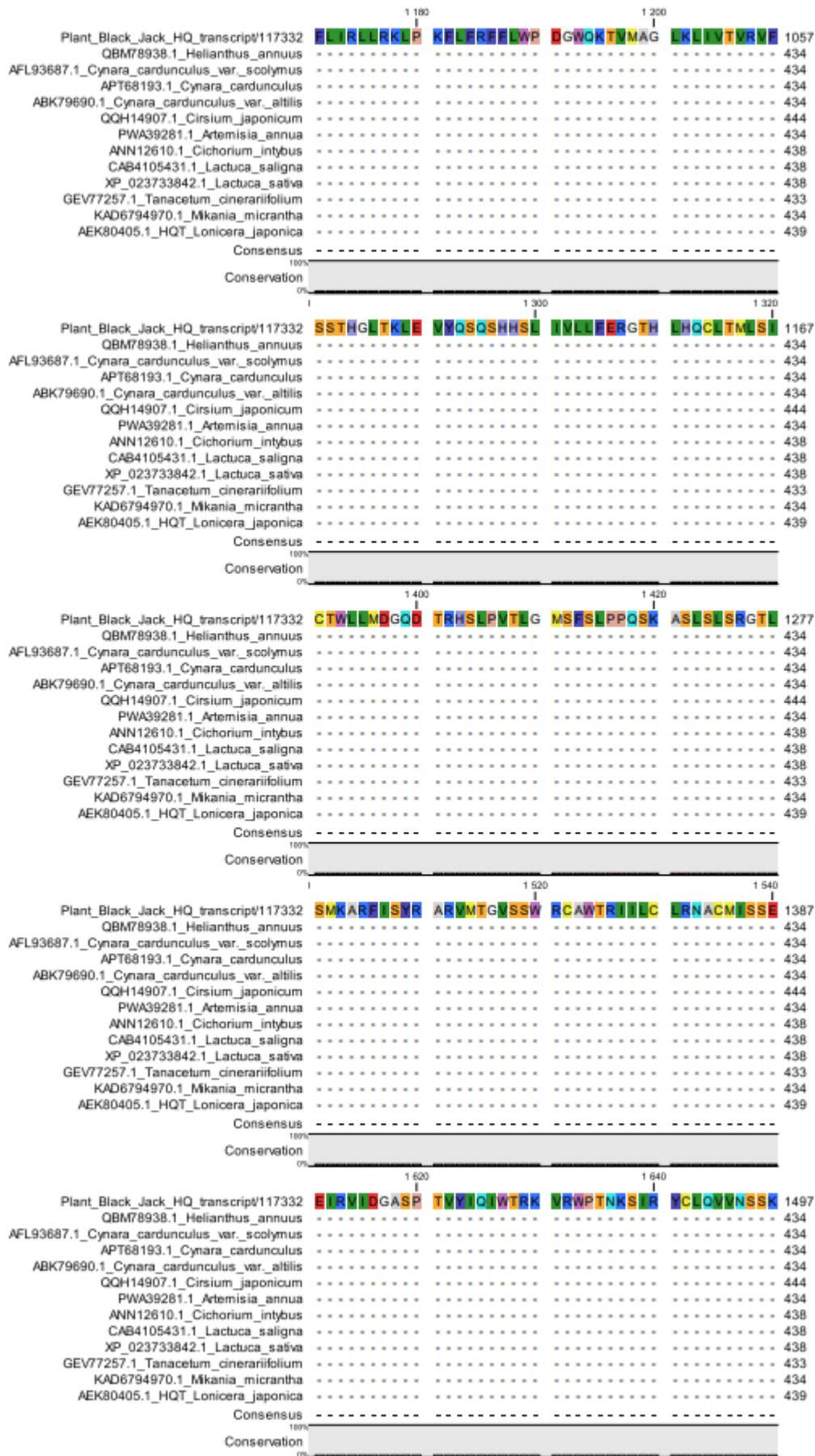


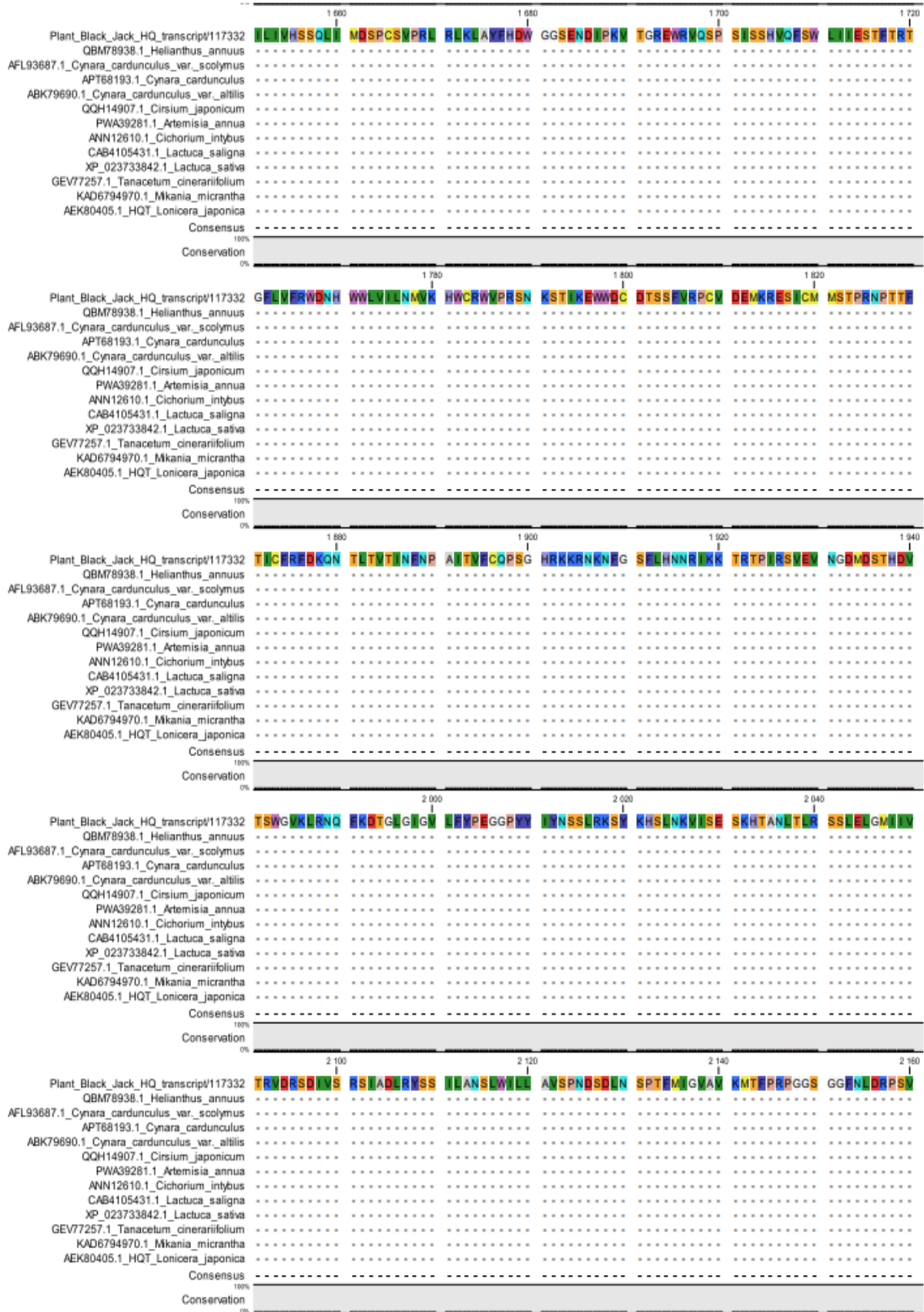


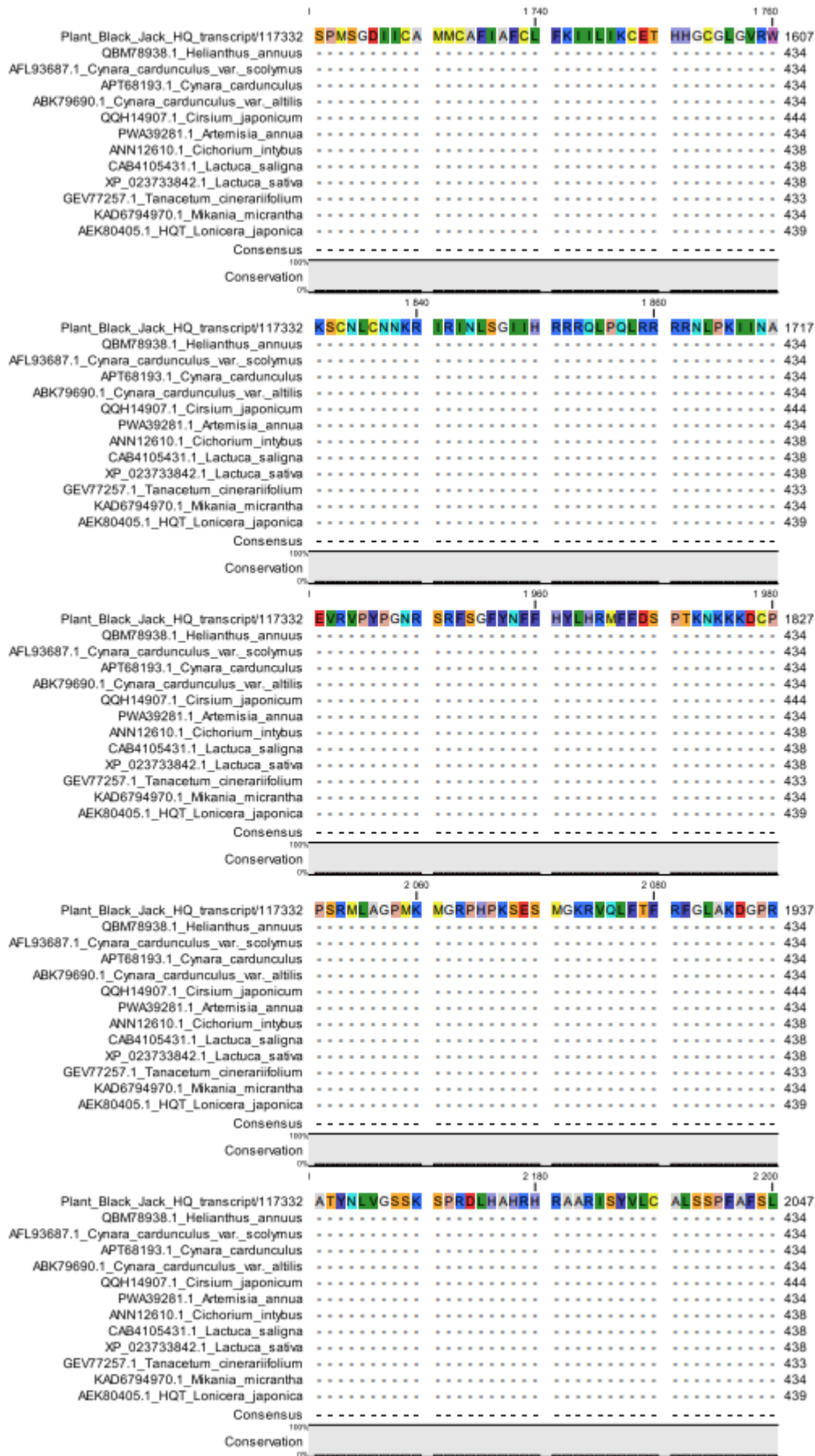


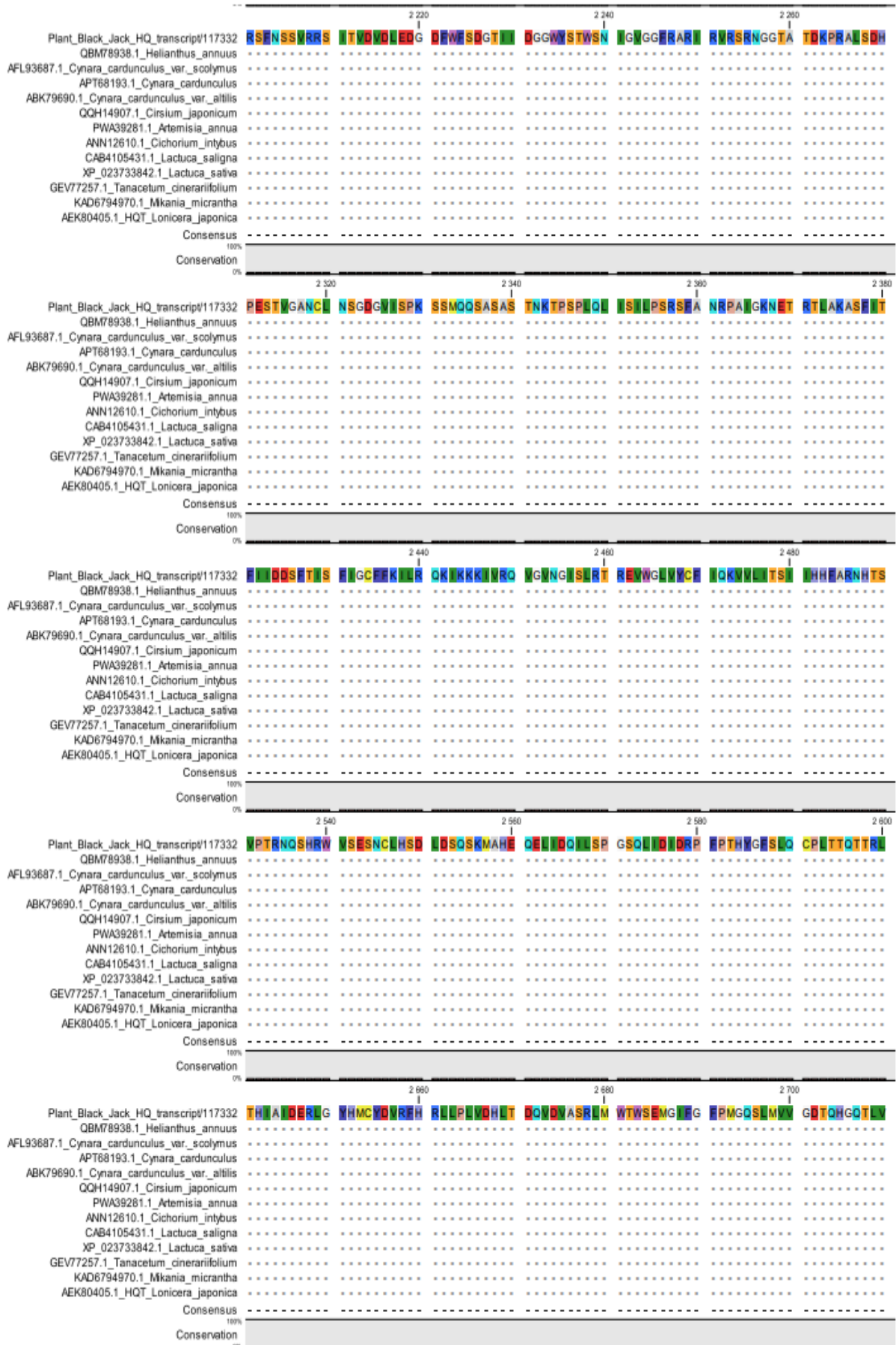


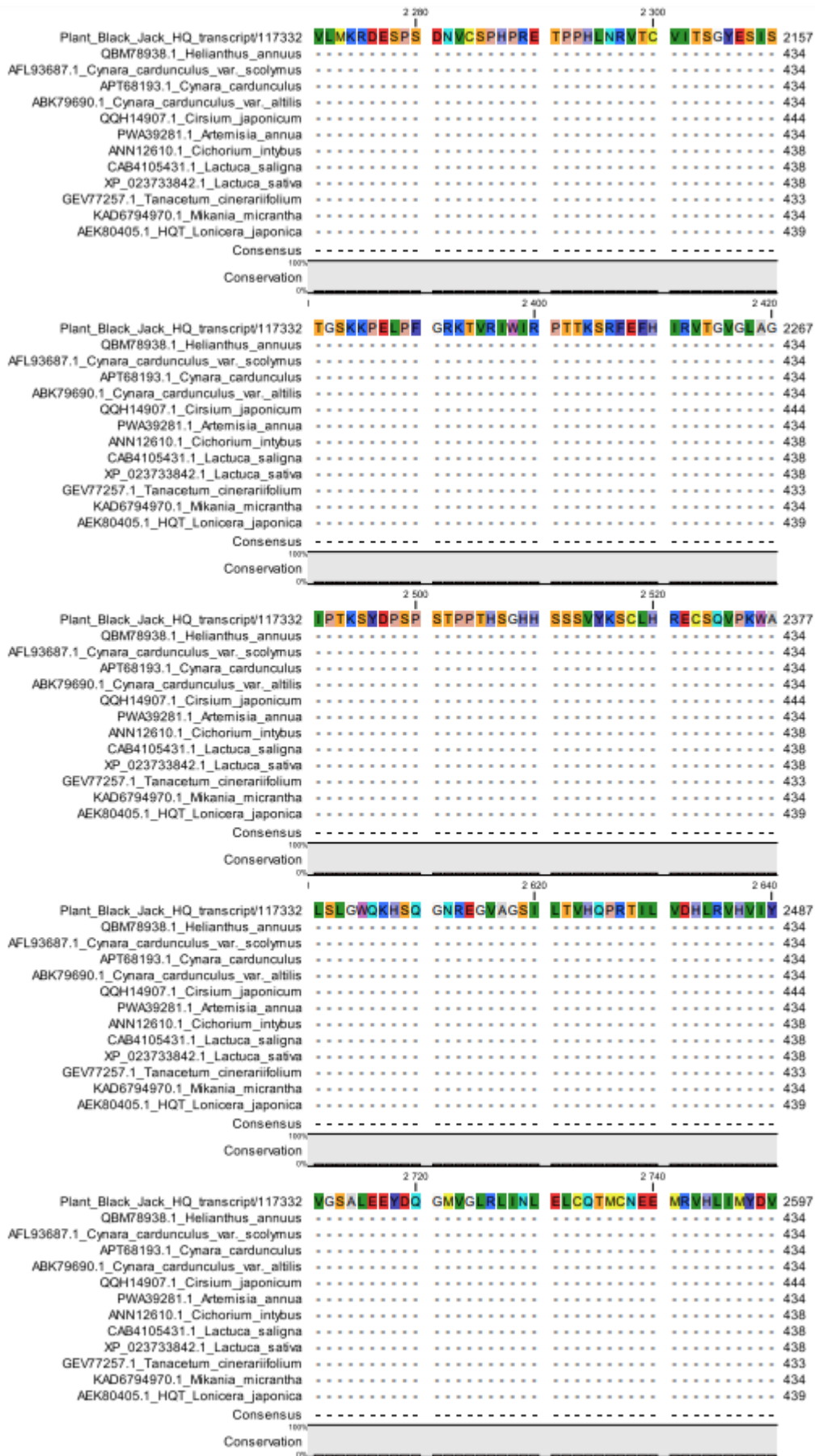












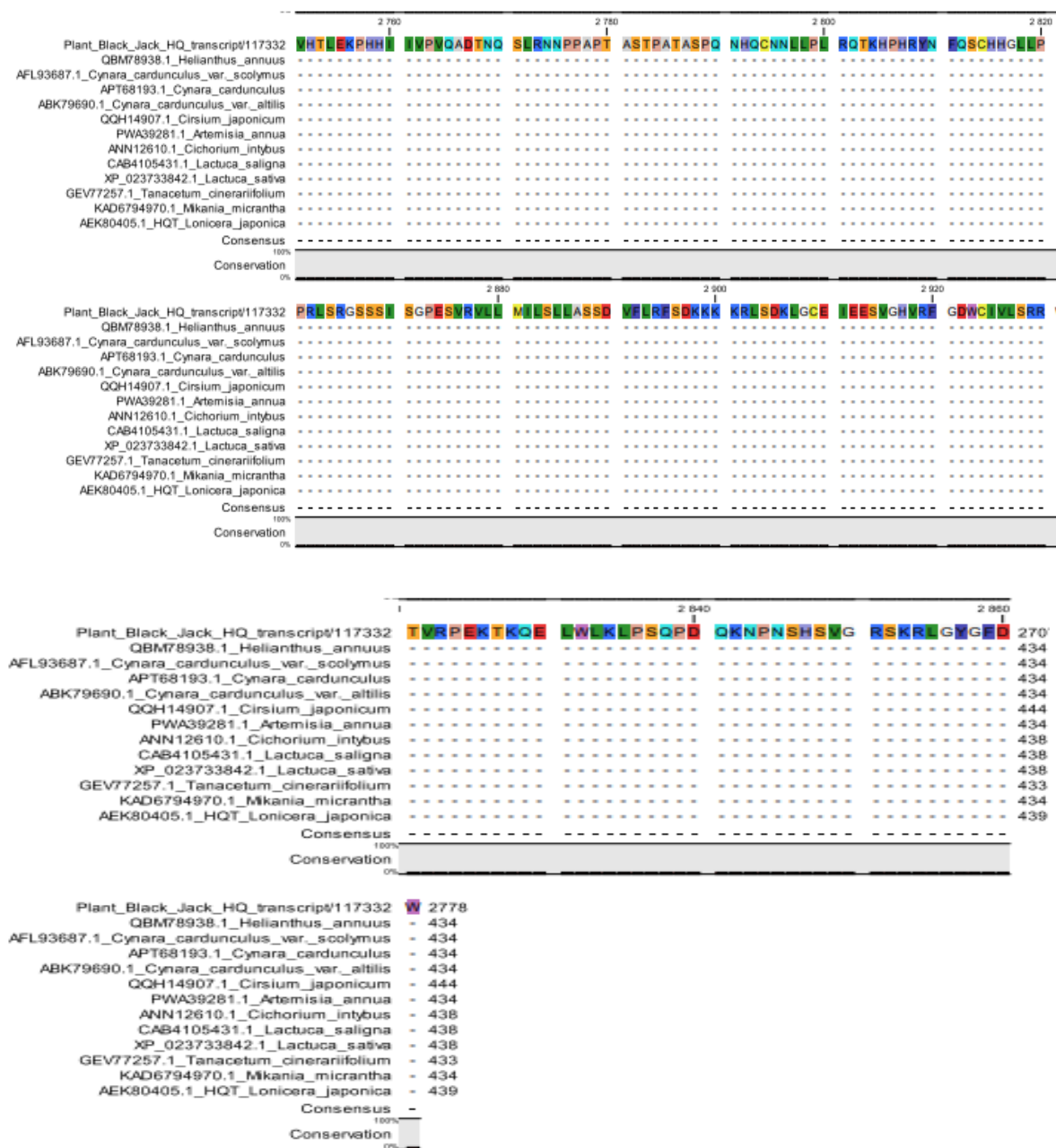
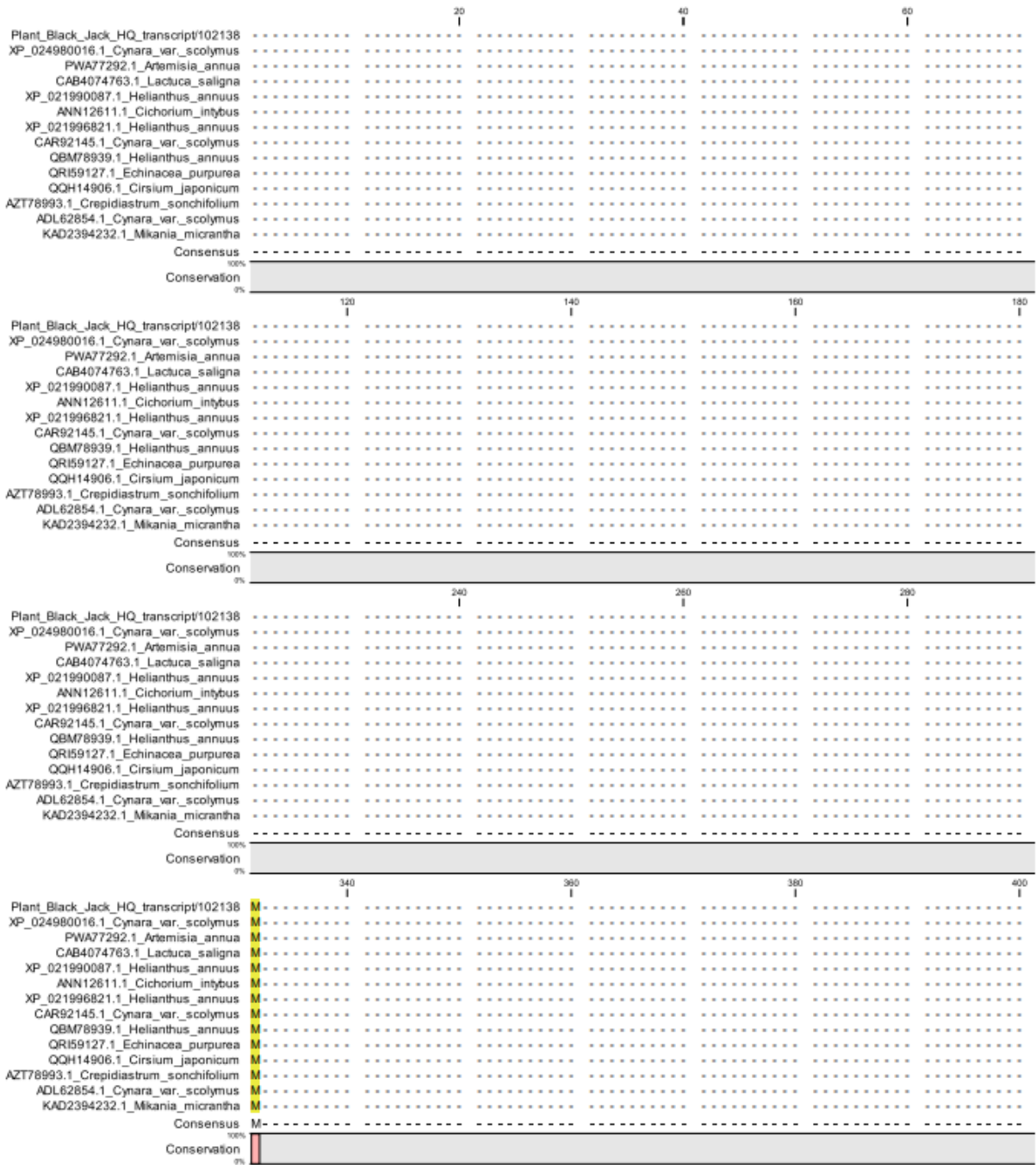
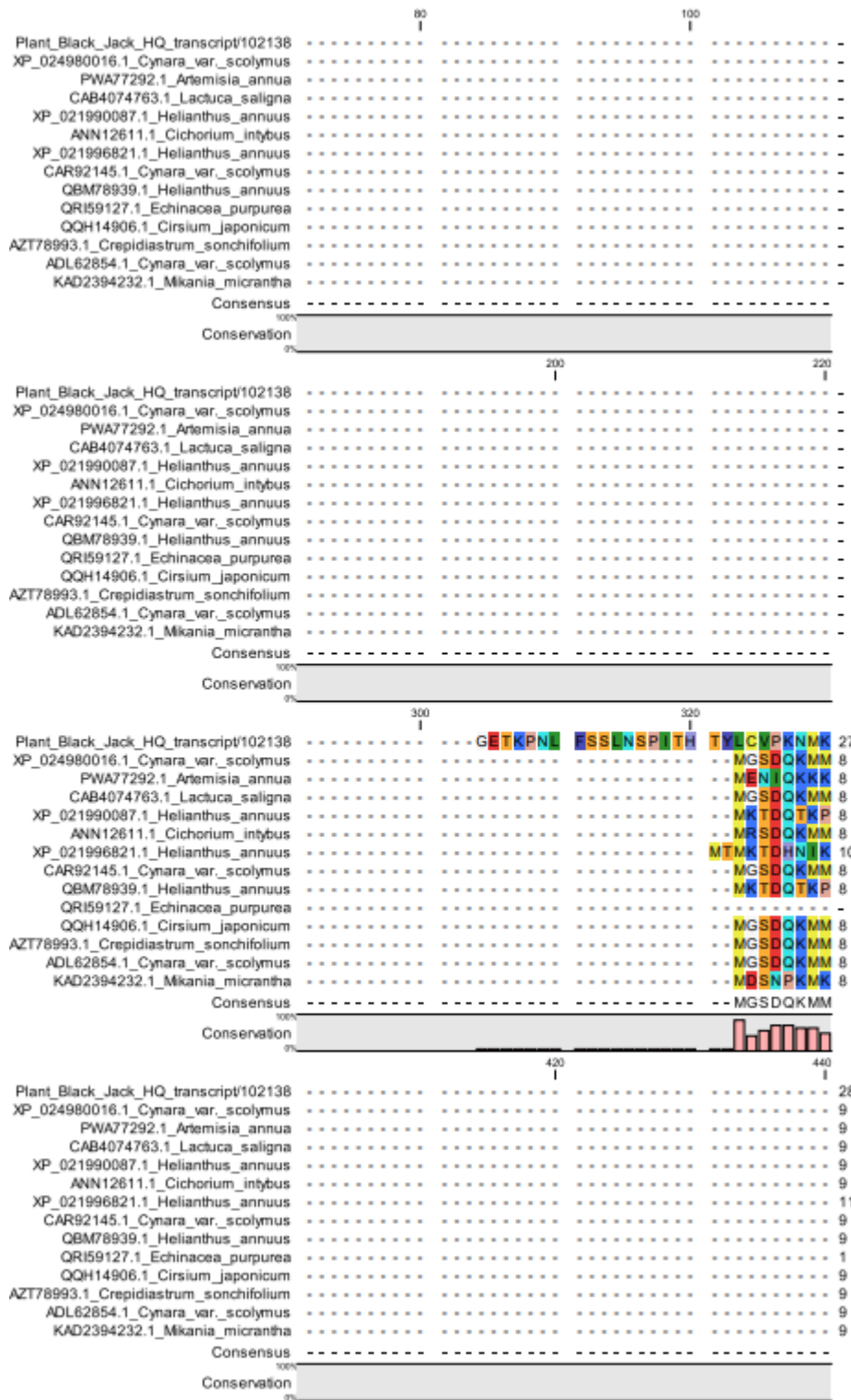
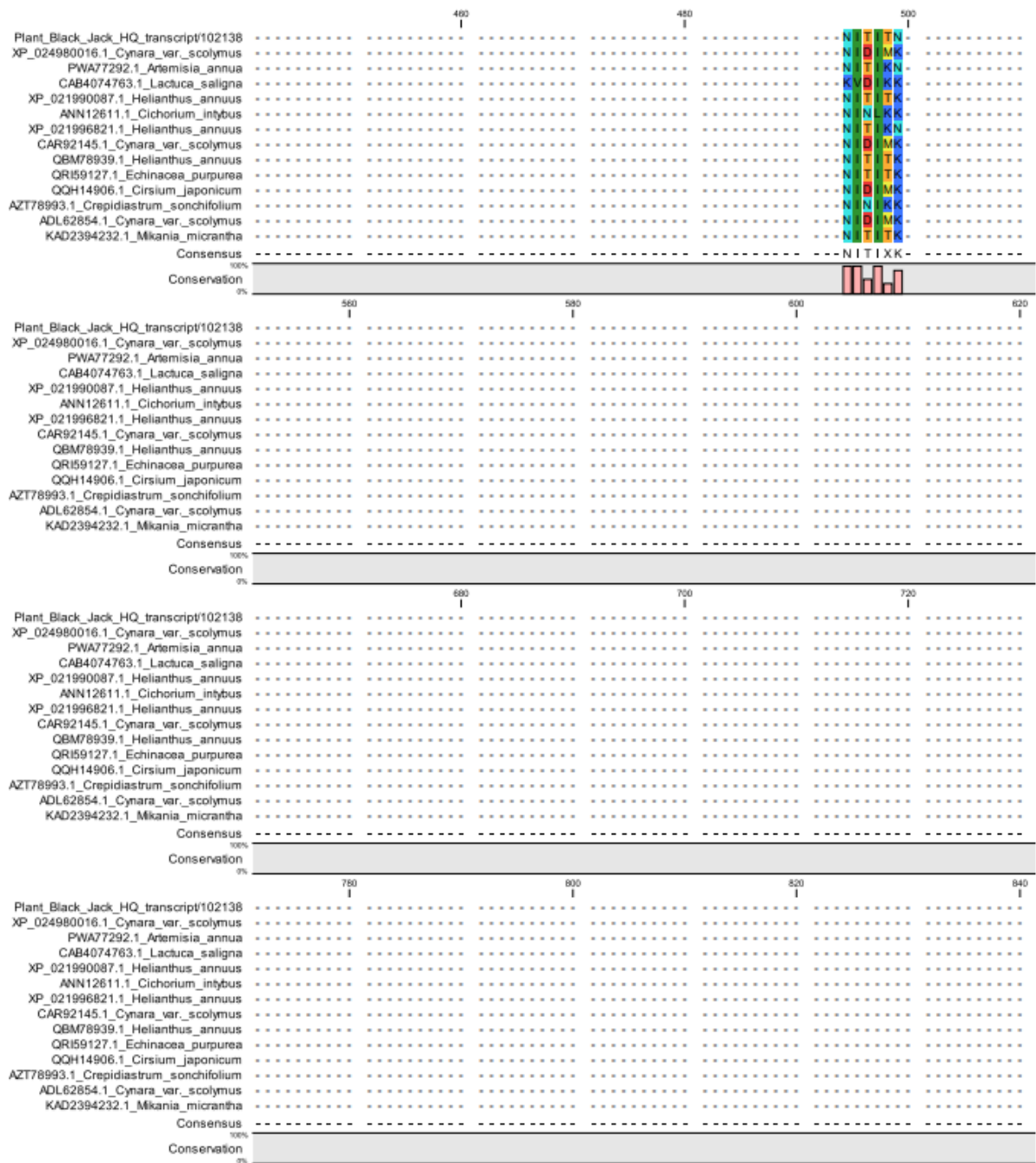


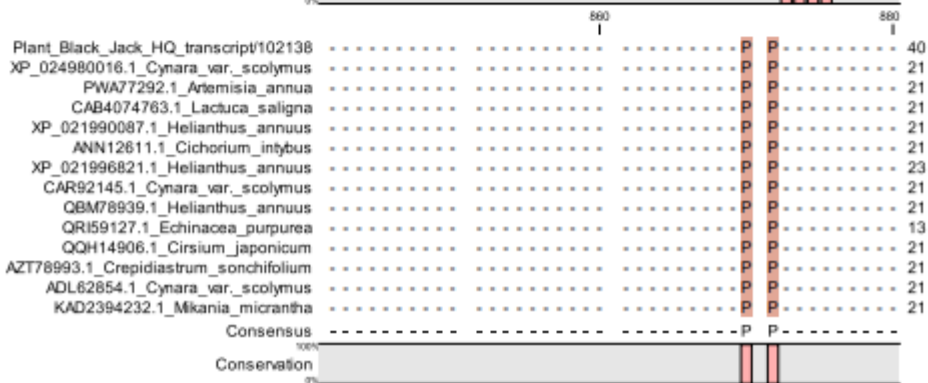
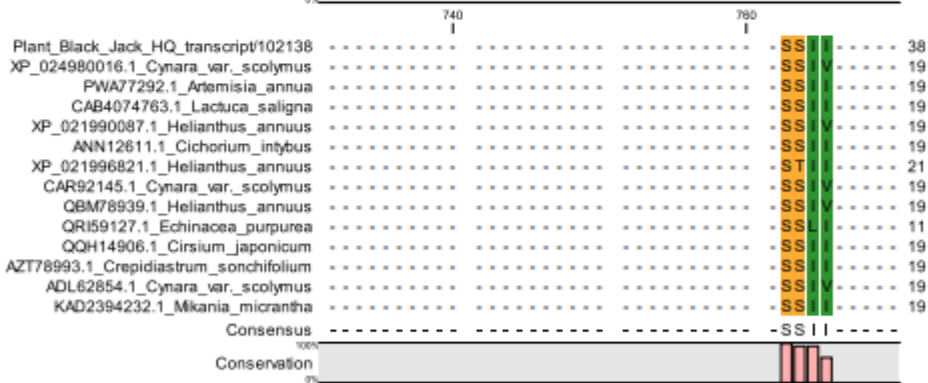
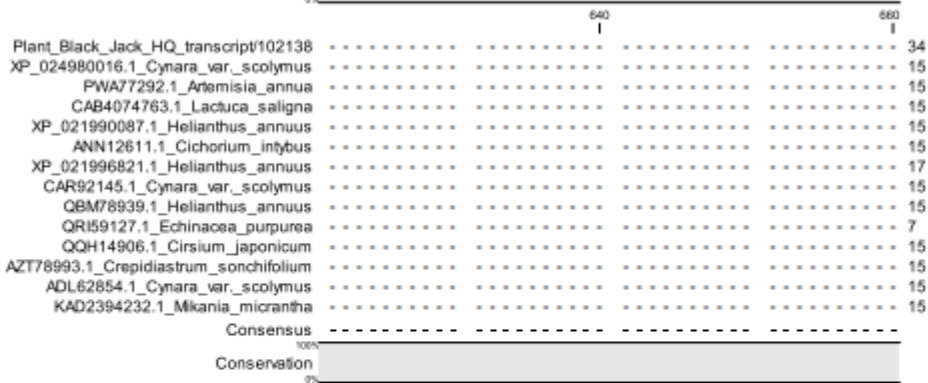
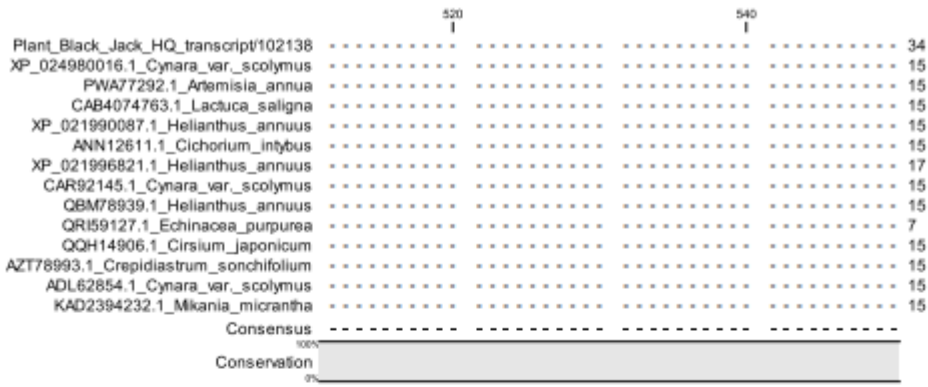
Figure S9. Full multiple sequence alignment of *B. pilosa* HQT1 gene.

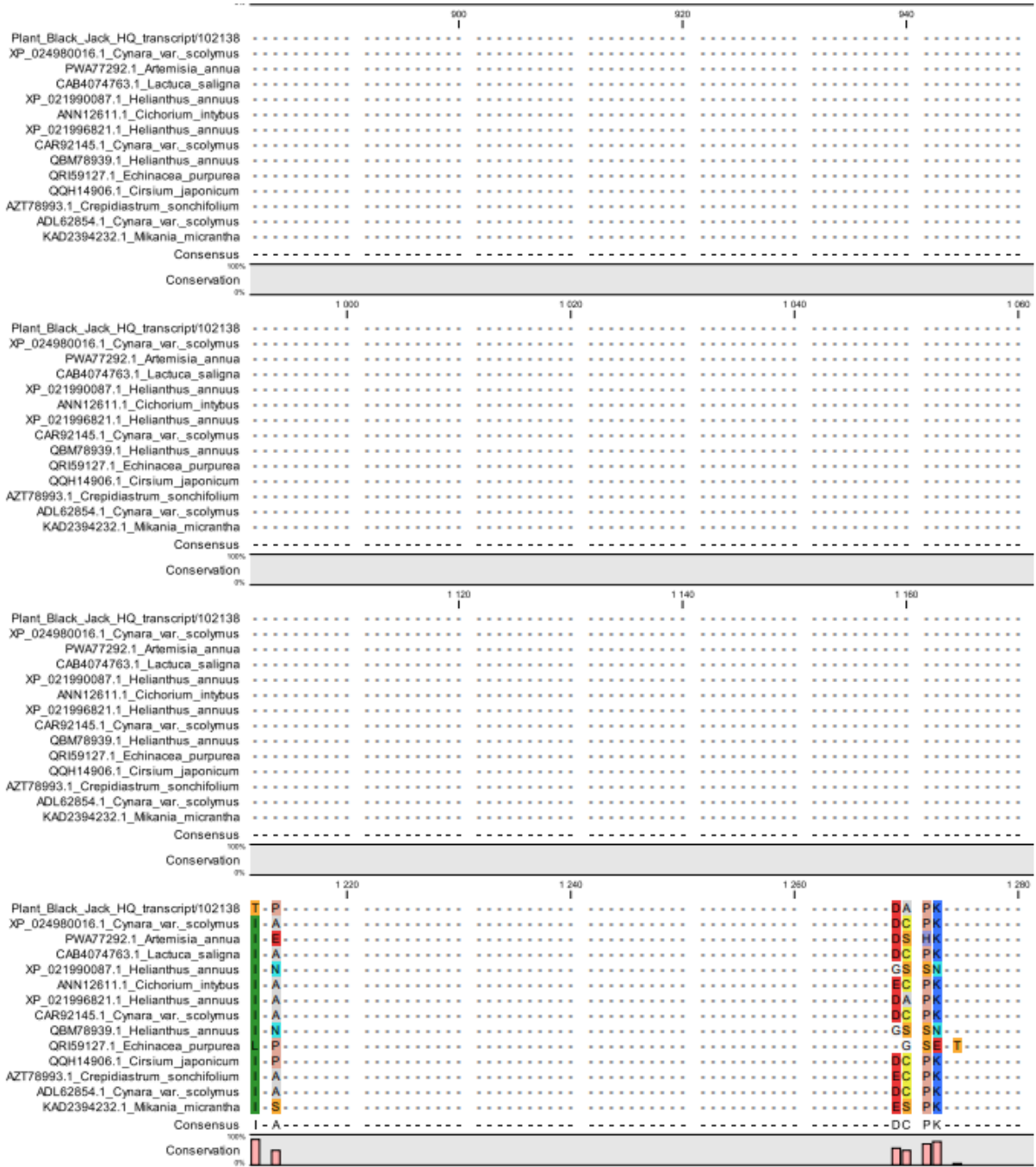
Multiple sequence alignment of *B. pilosa* HQT1 with its homologues from *Helianthus annuus* (QBM78938.1), *Cynara cardunculus var scolymus* (AFL93687.1), *Artemisia annua* (PWA39281.1), *Lactuca sativa* (XP_023733842.1), *Mikania micrantha* (KAD5794970.1), *Lonicera japonica* (AEK80405.1), *Chicorium intybus* (ANN12610.1) and *Tanacetum cinerariifolium* (GEV77257.1). Residues are grouped according to colours, for instance same colour represent similar residues across all genes from different plants. The alignment was generated using MUSCLE algorithm of the MEGA software. The position of the residue is shown by the number on the right.

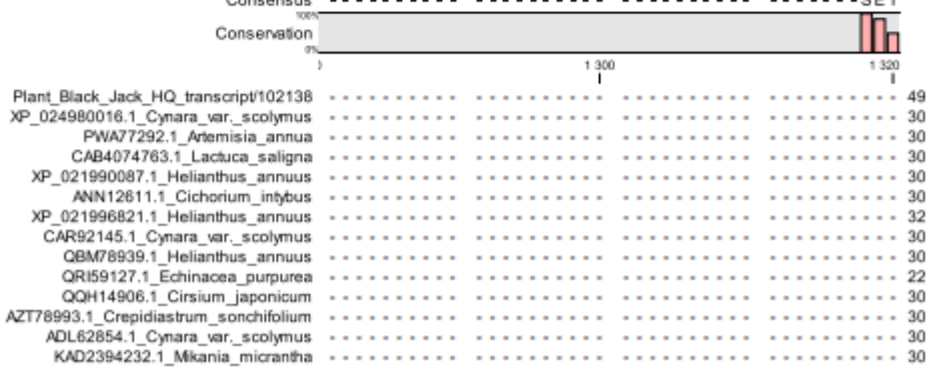
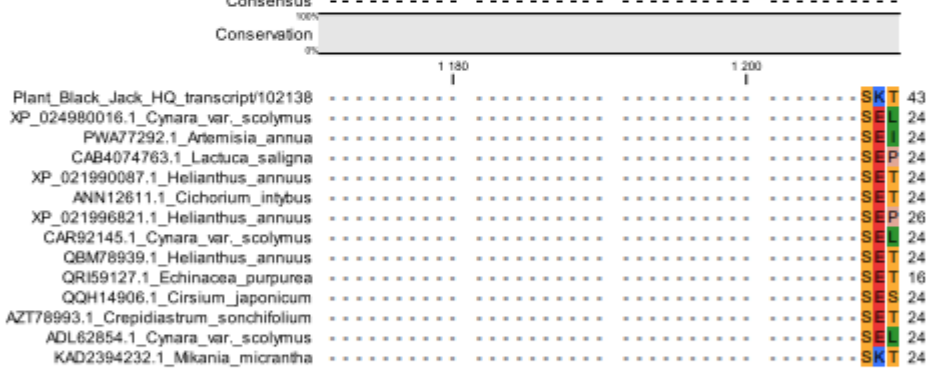
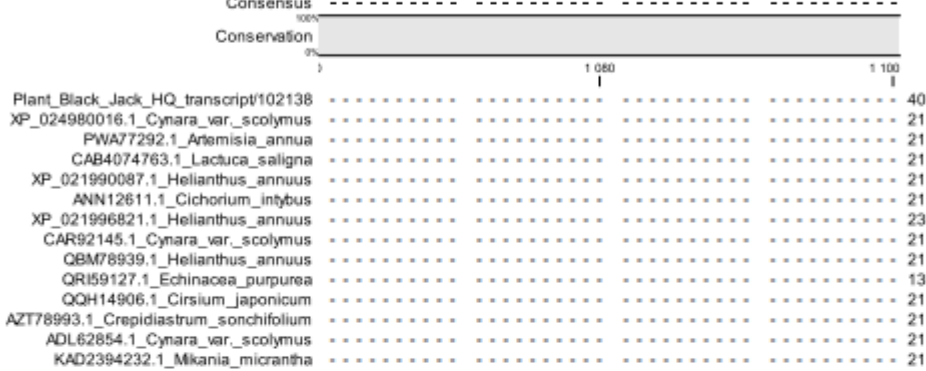
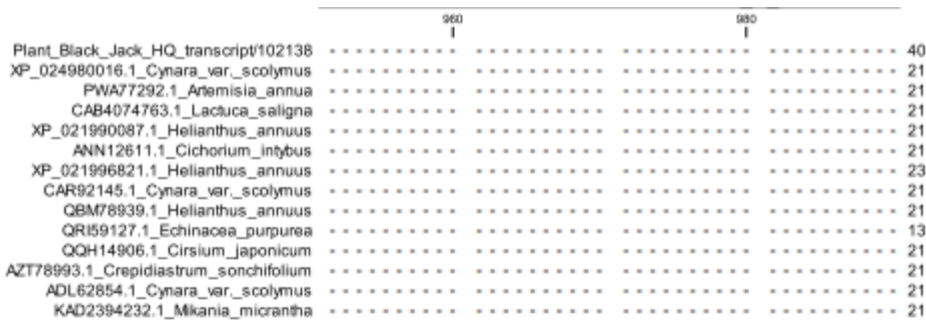


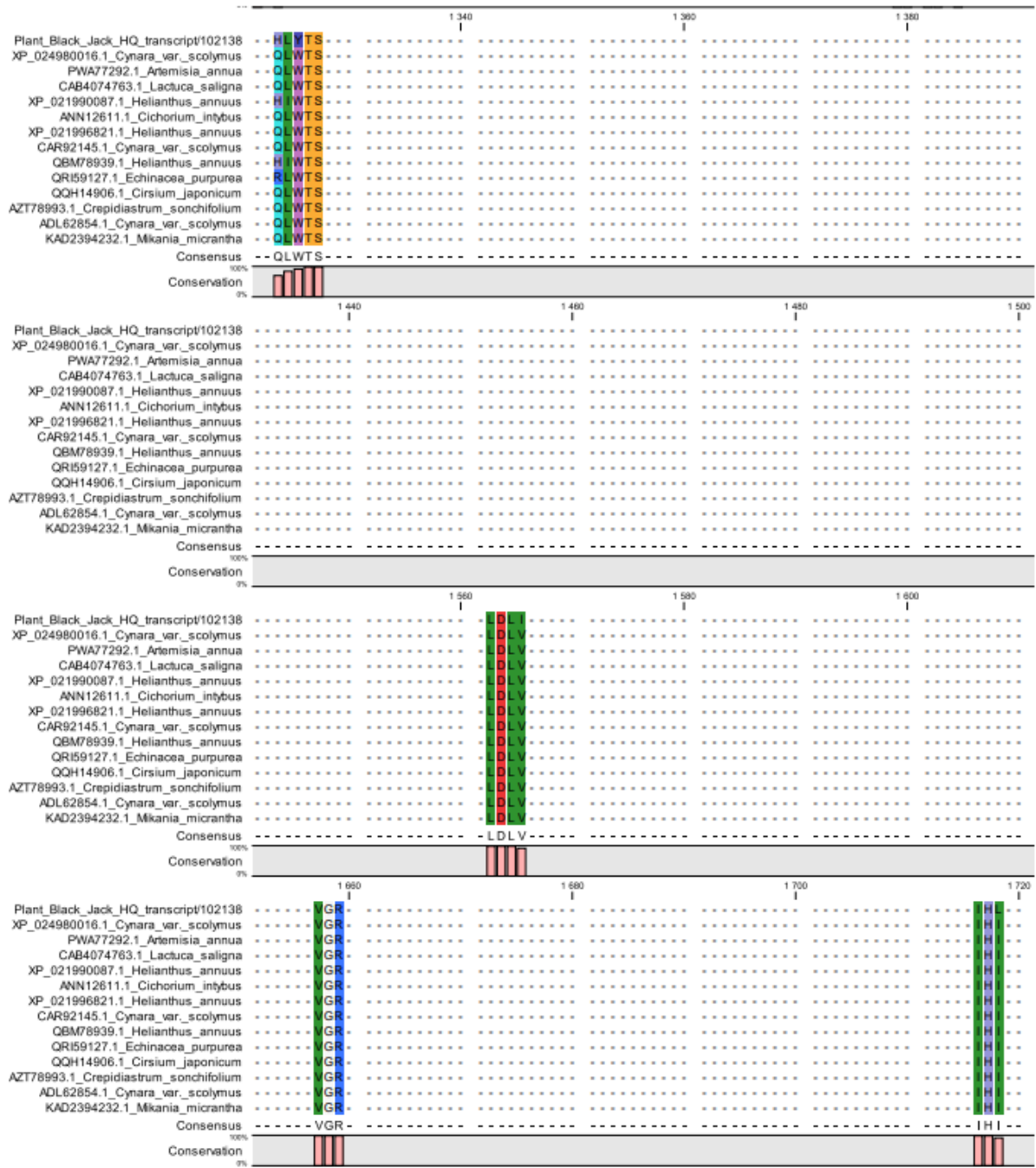


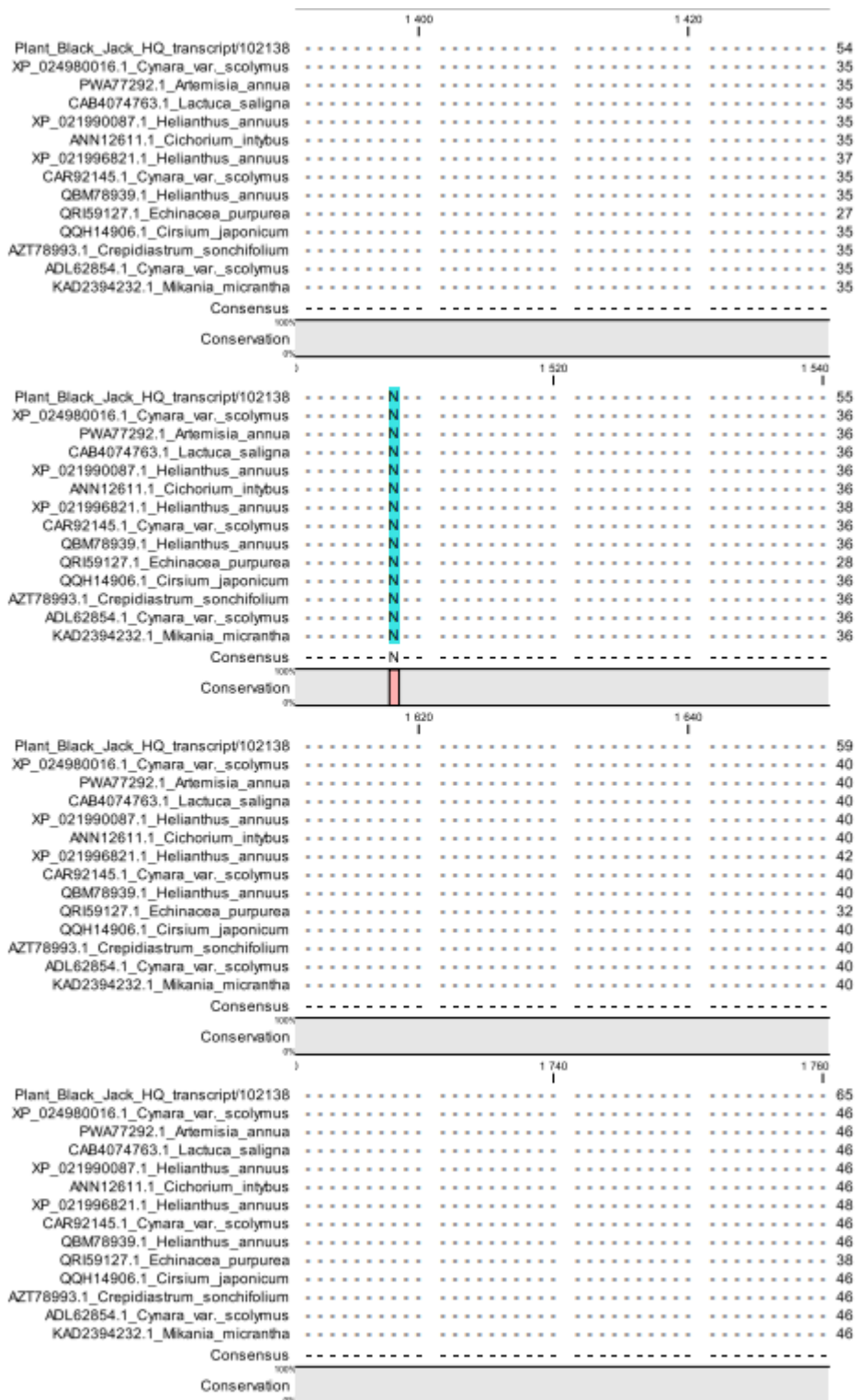


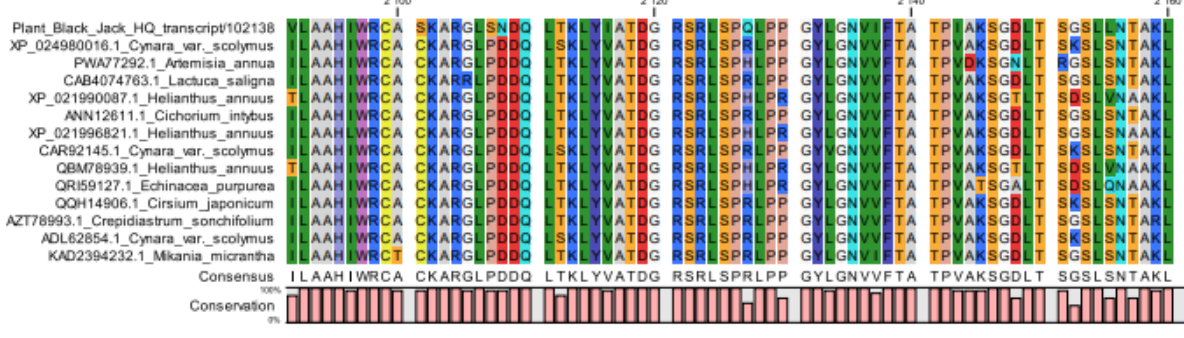
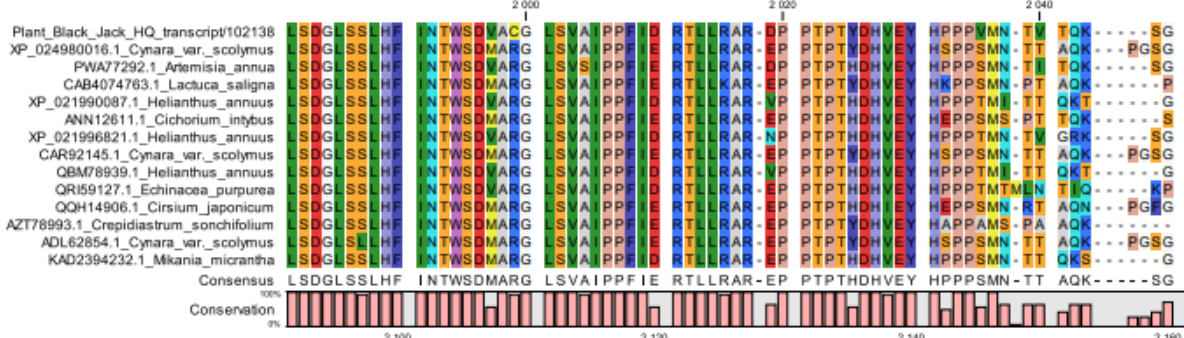
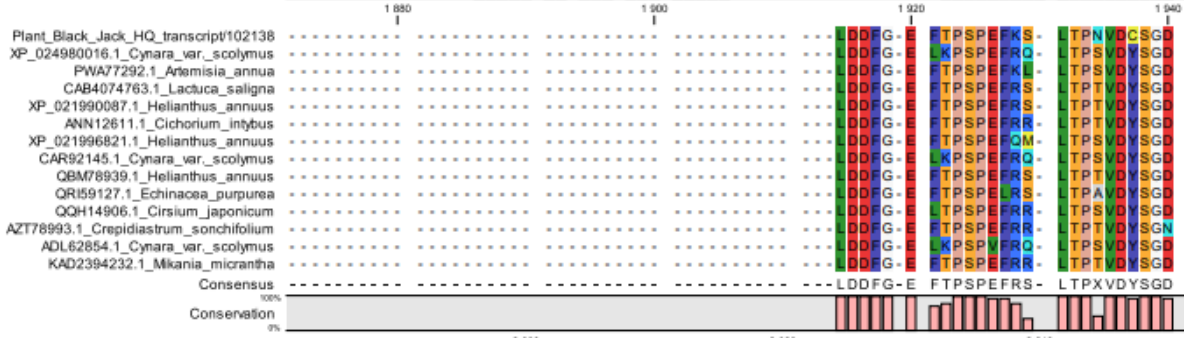
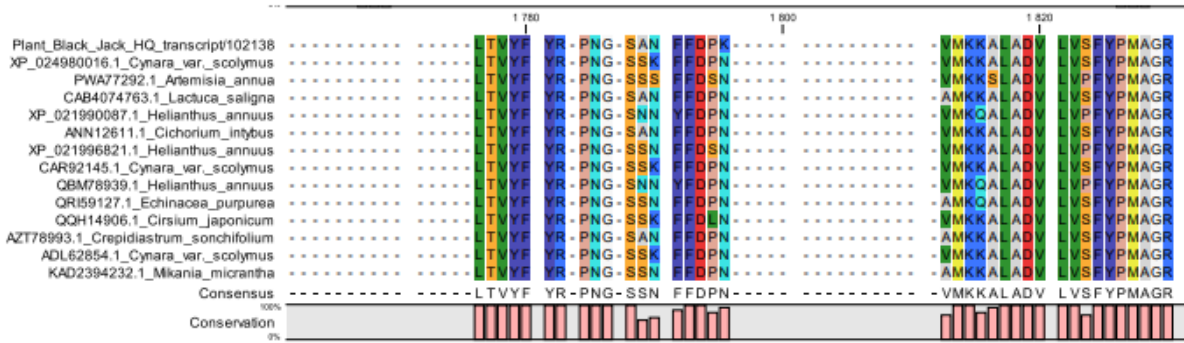


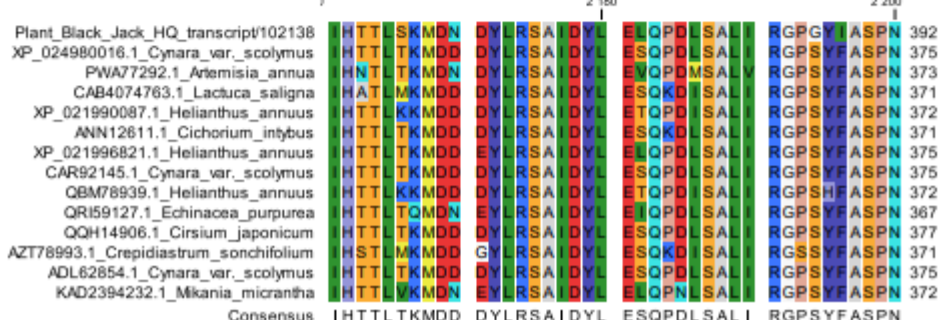
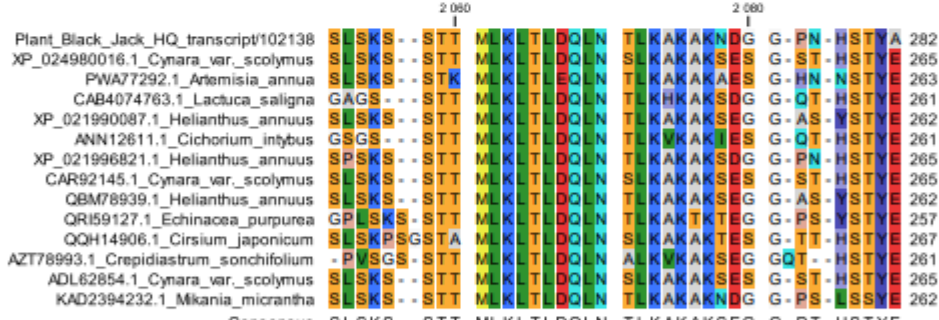
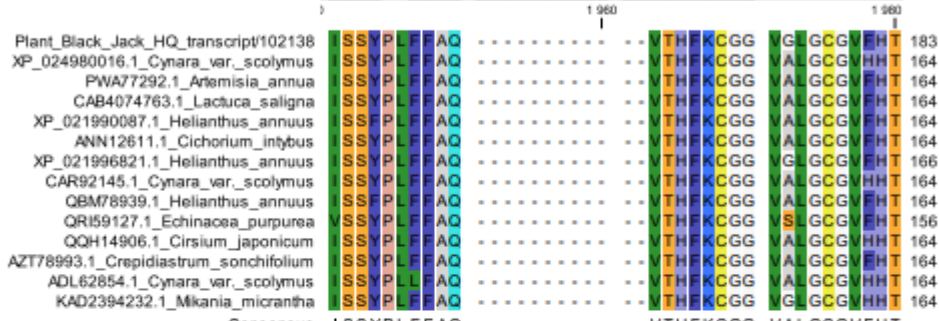
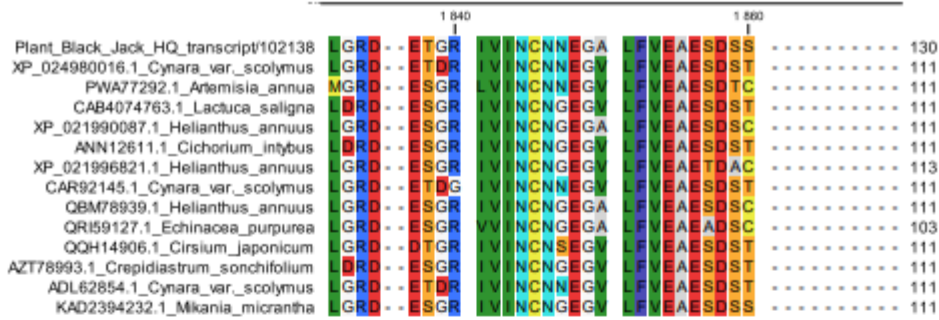


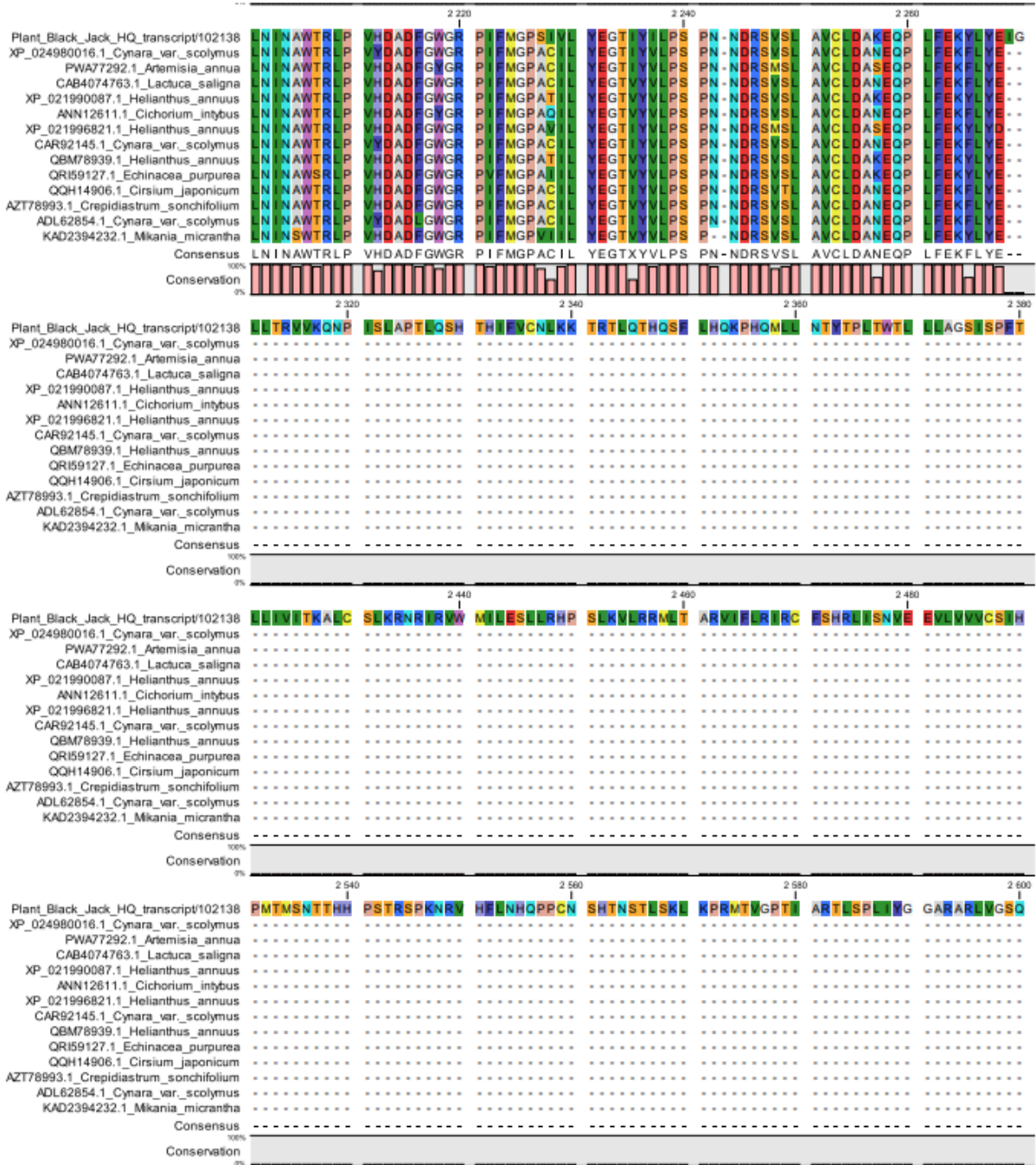


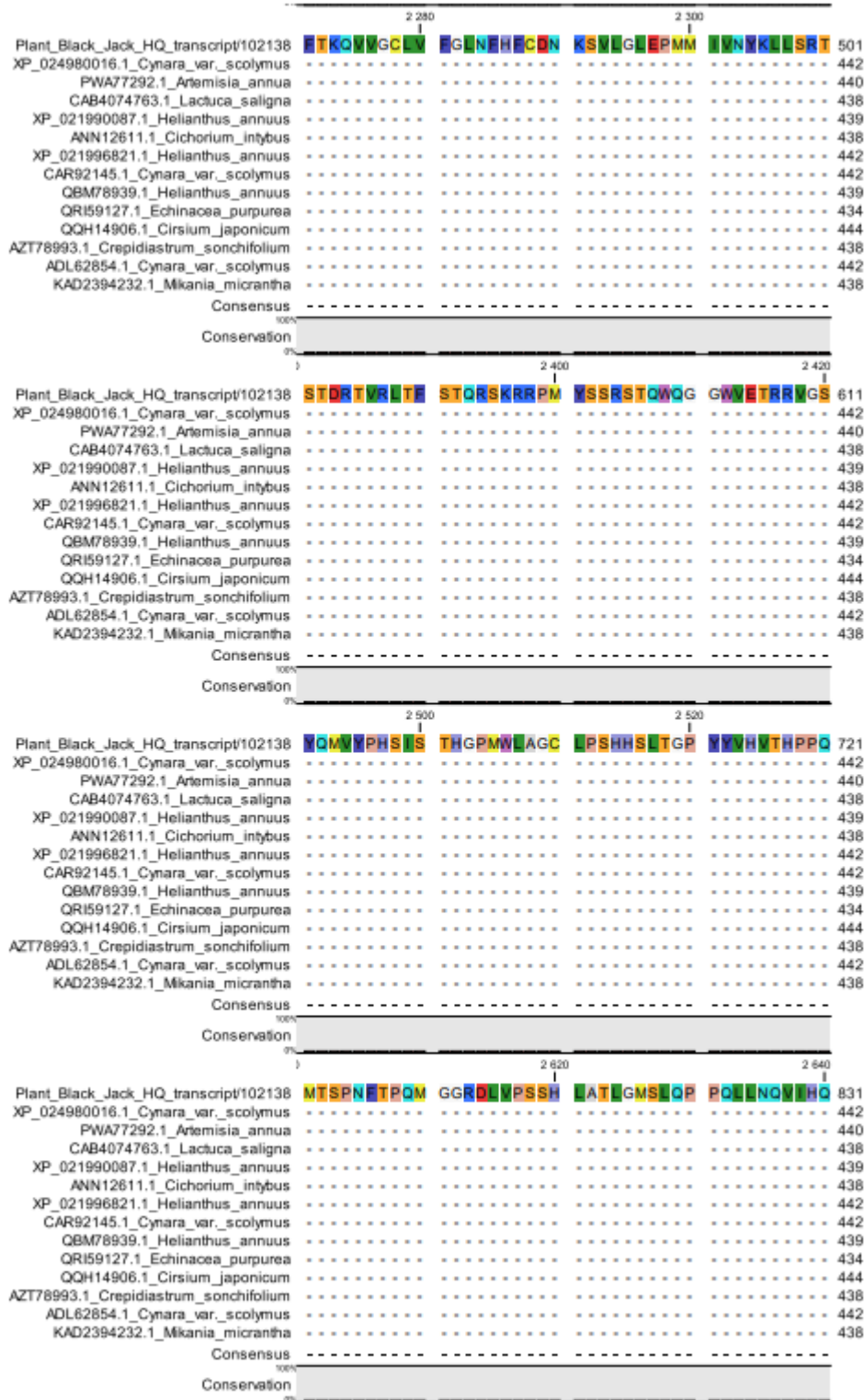


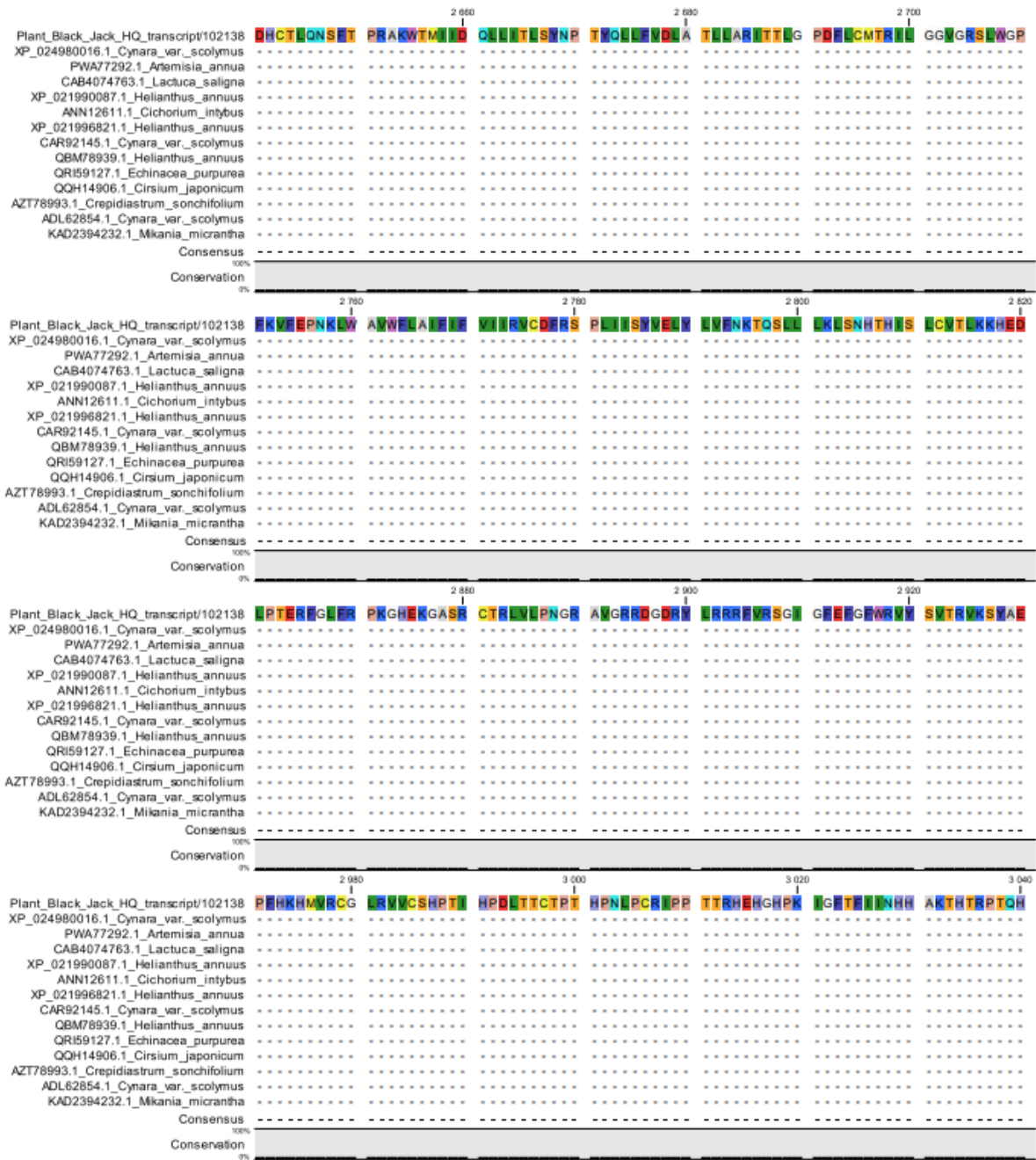


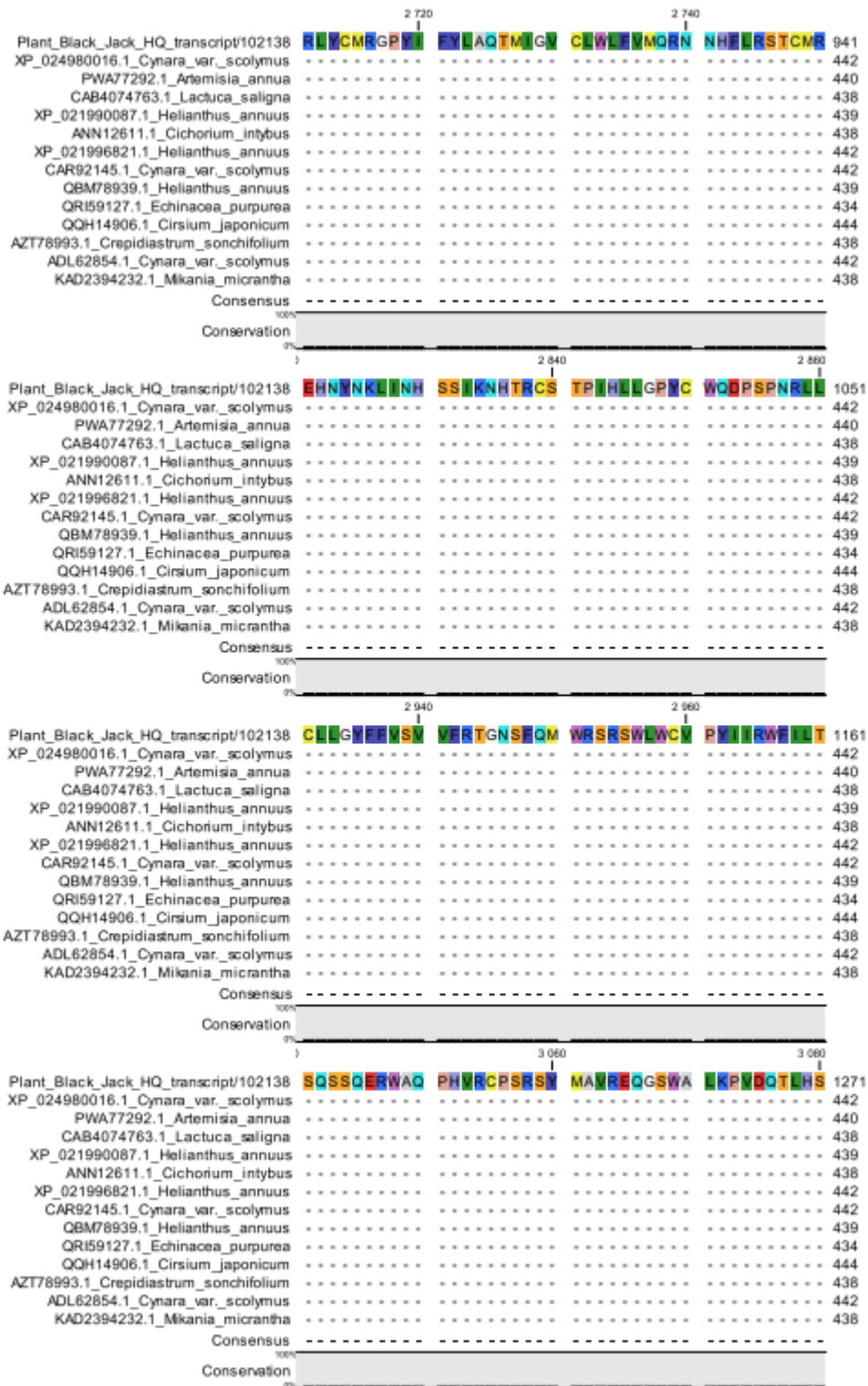


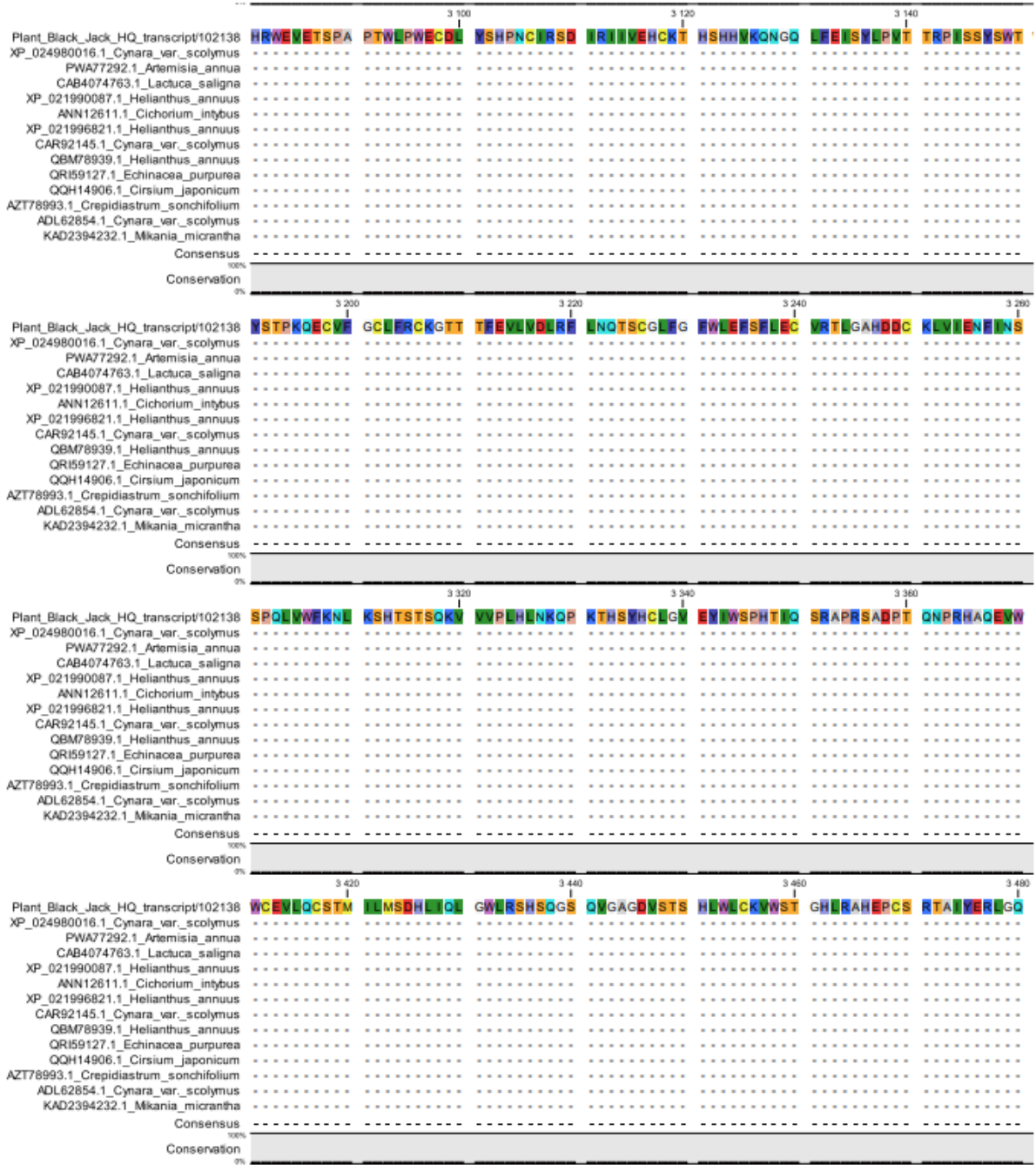


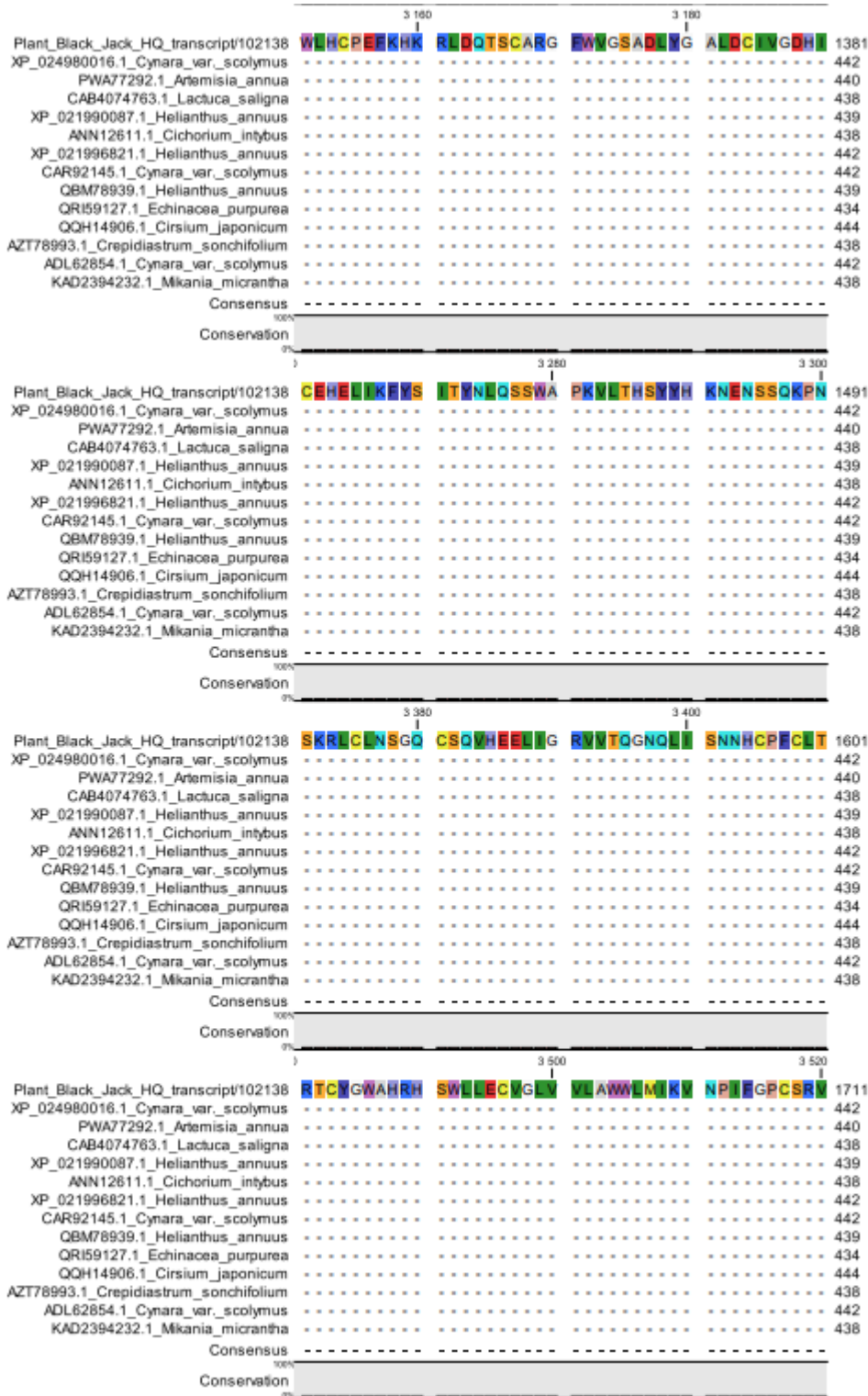


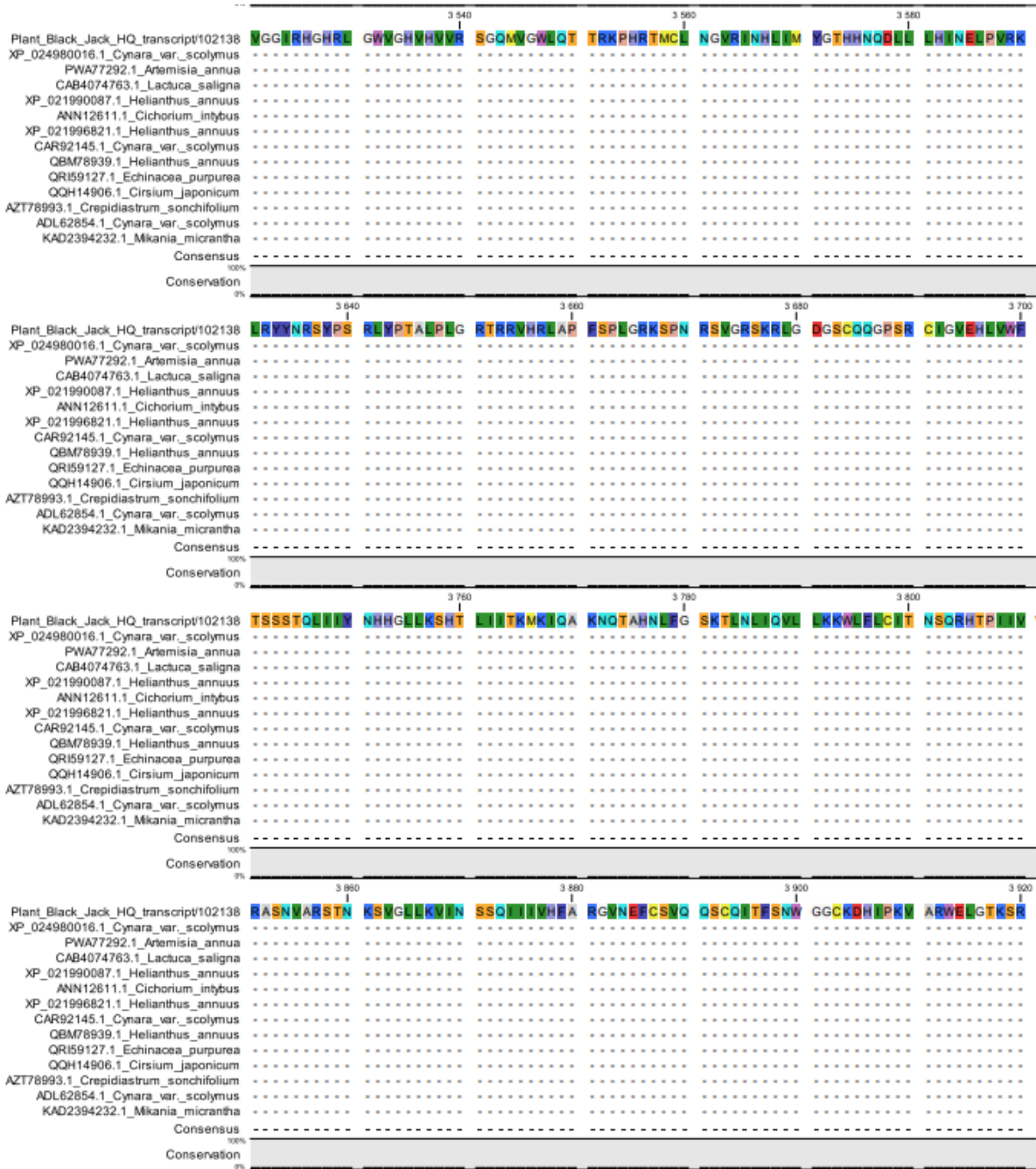


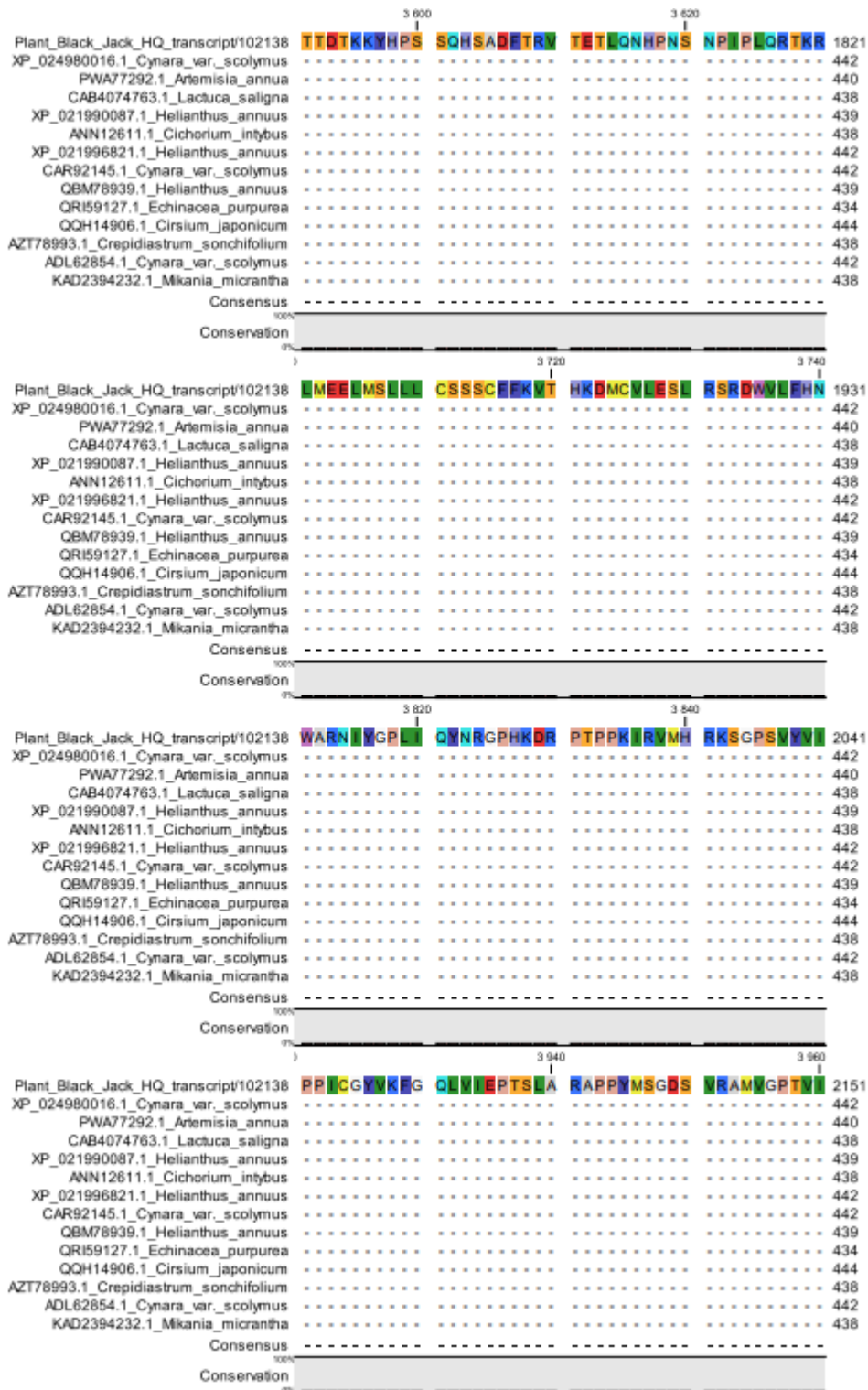


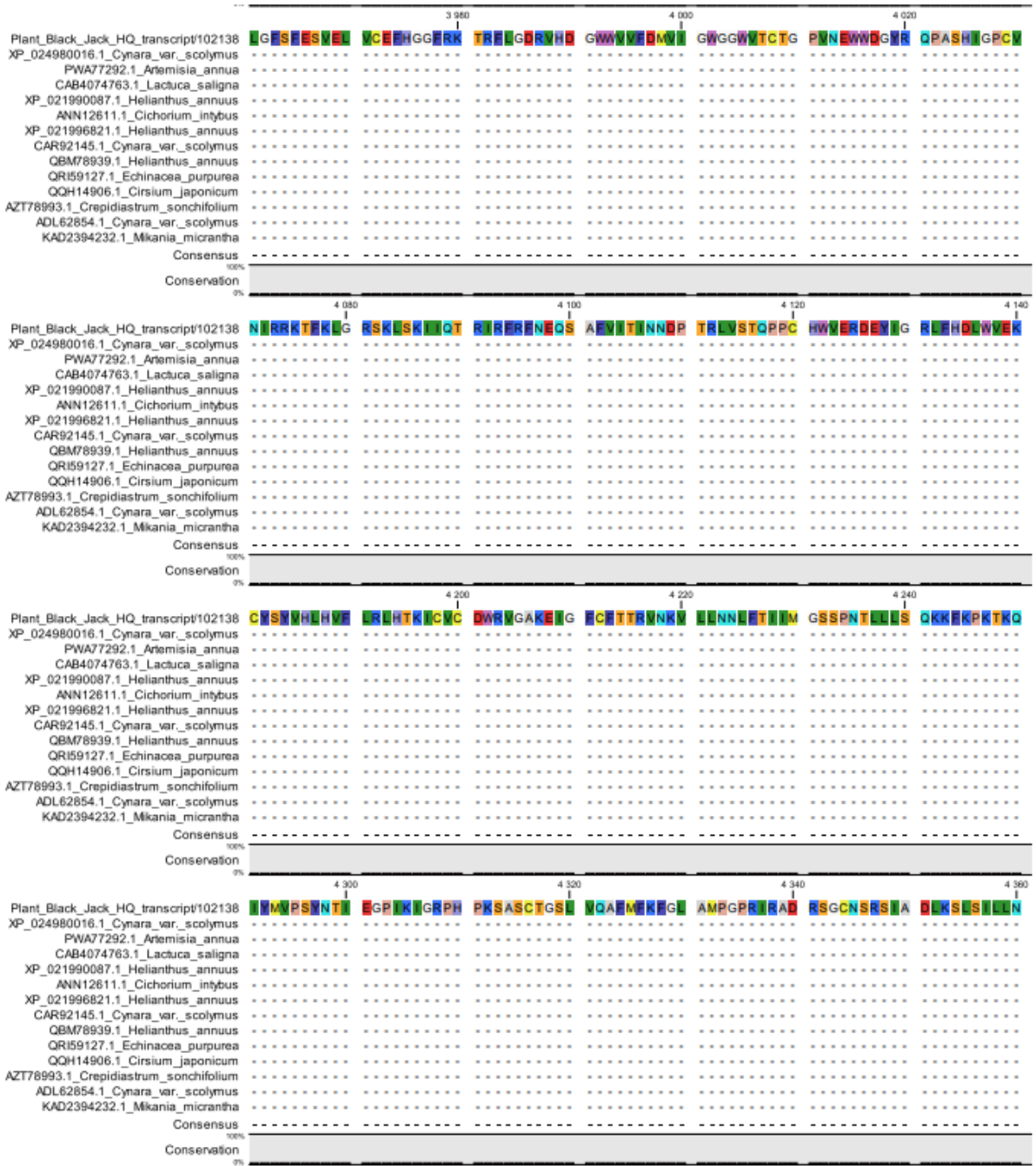


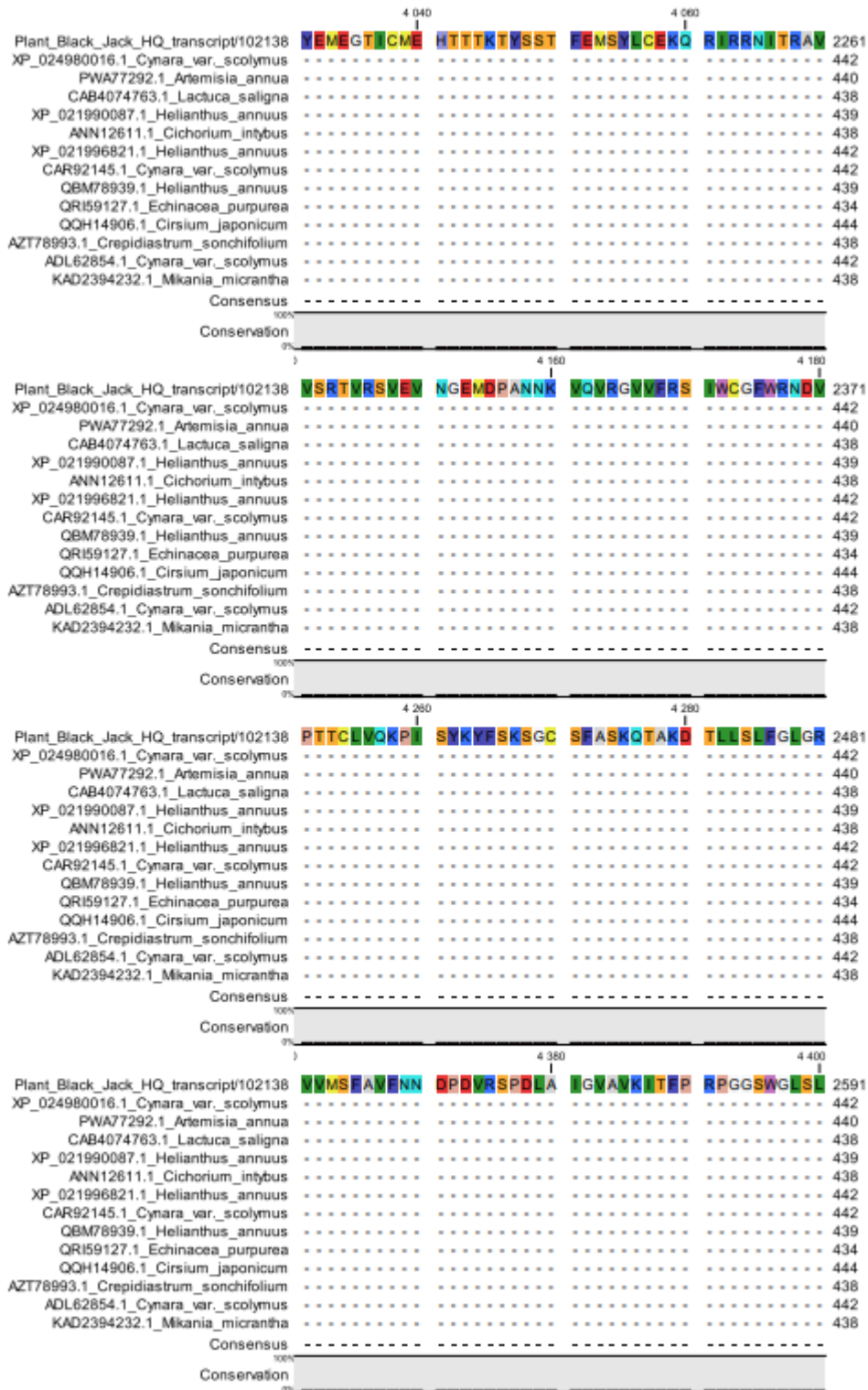












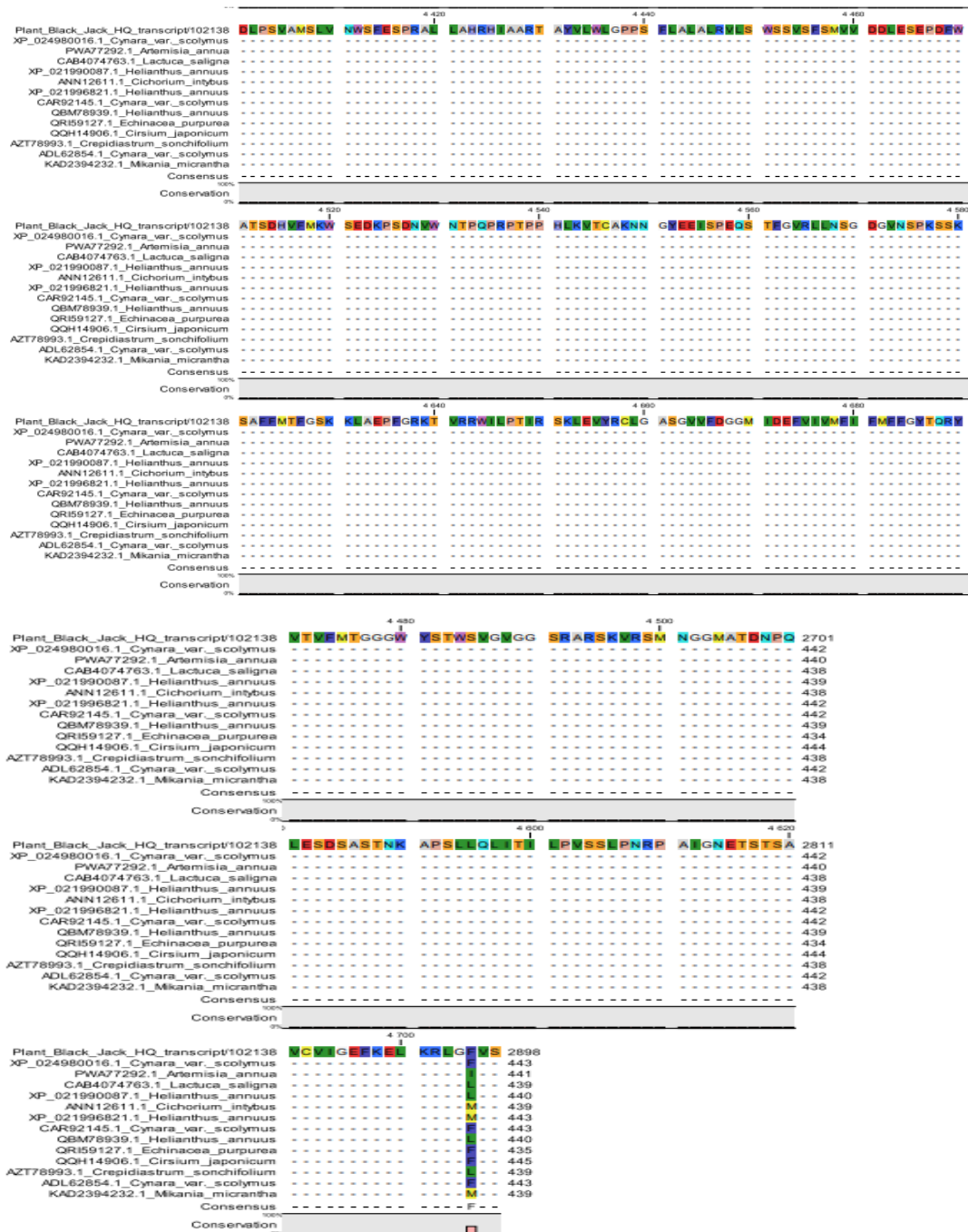
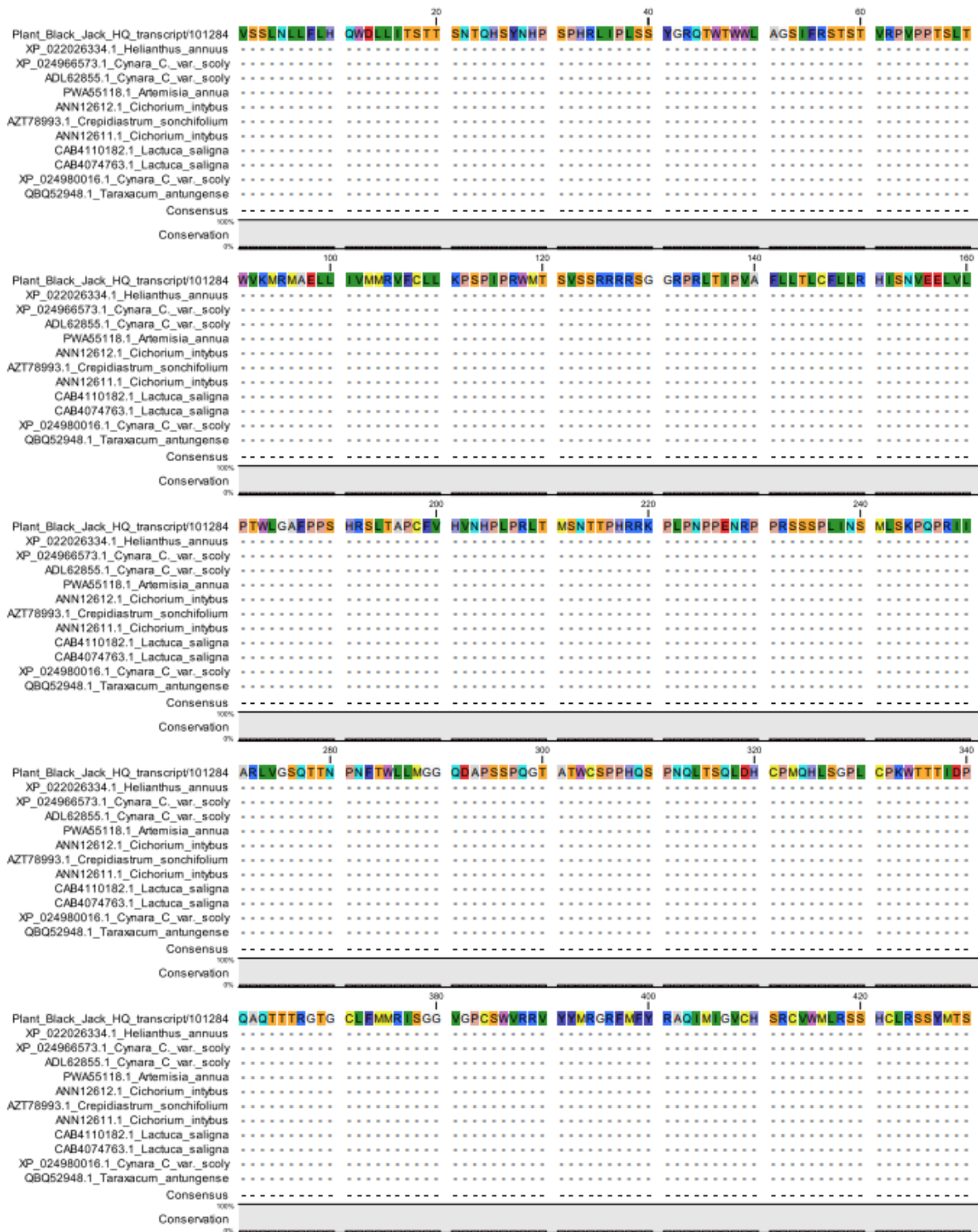
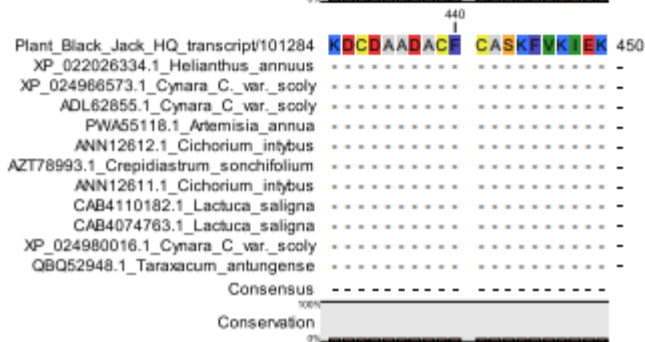
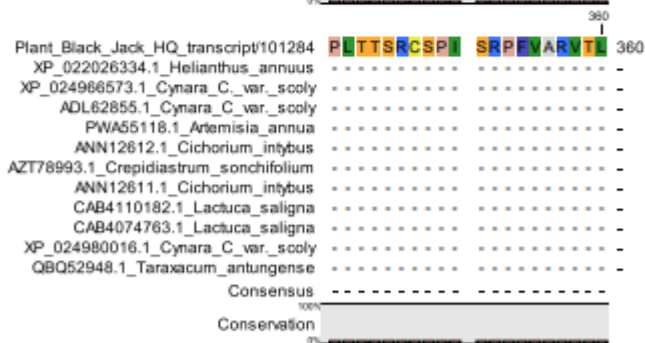
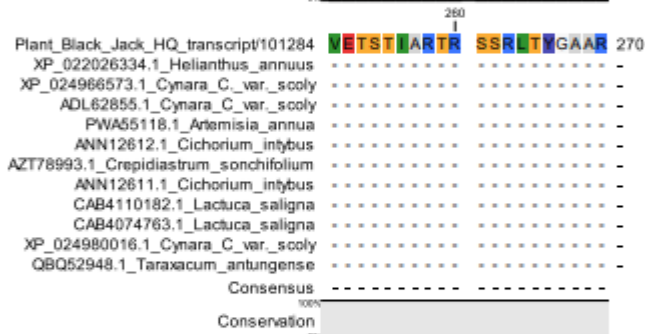
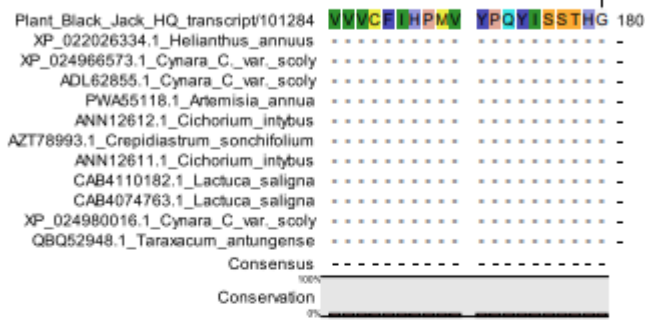
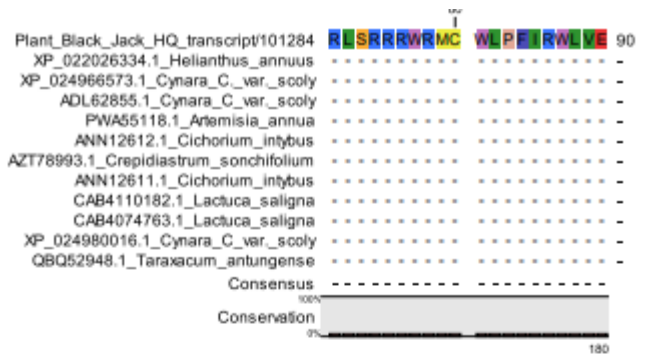
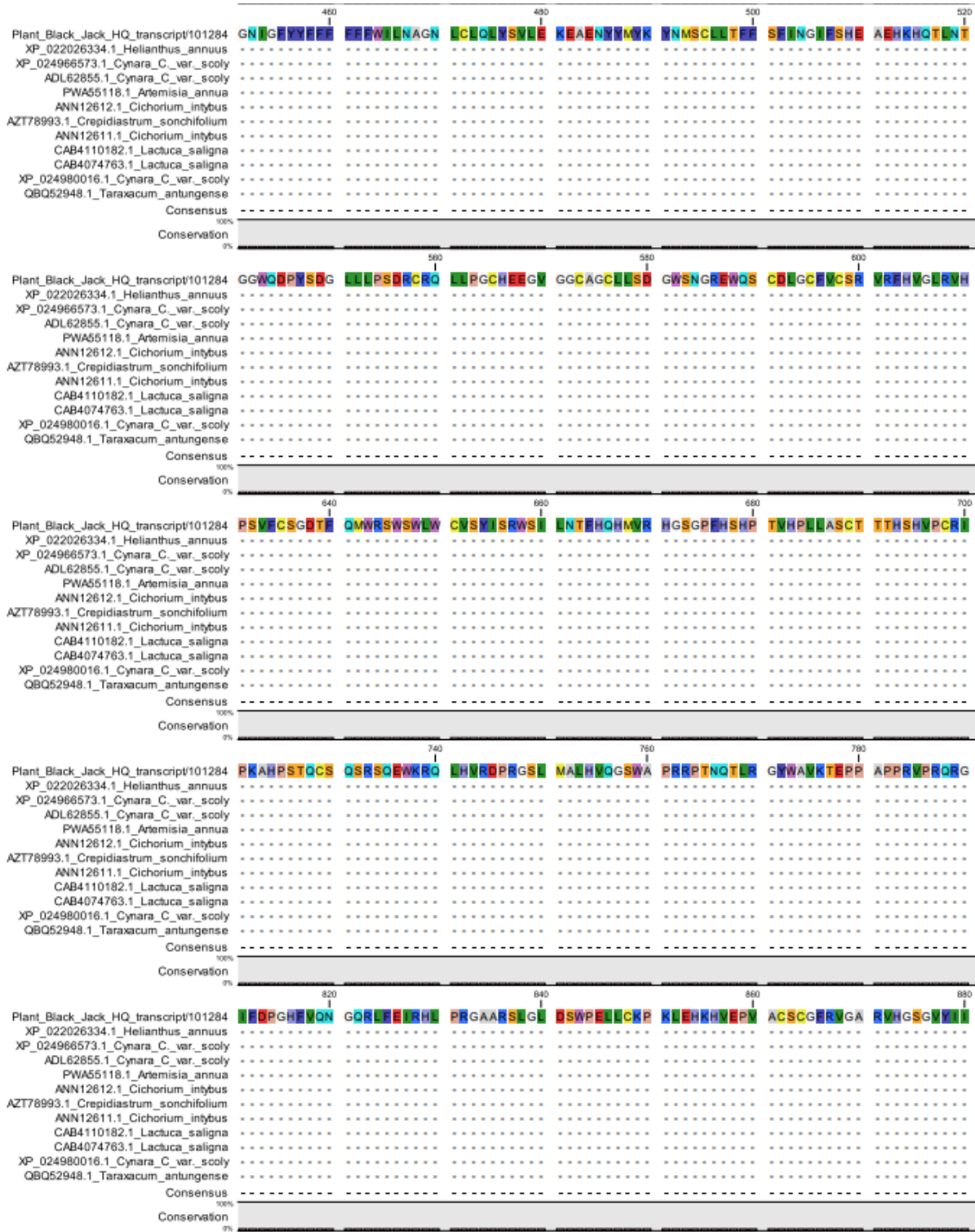


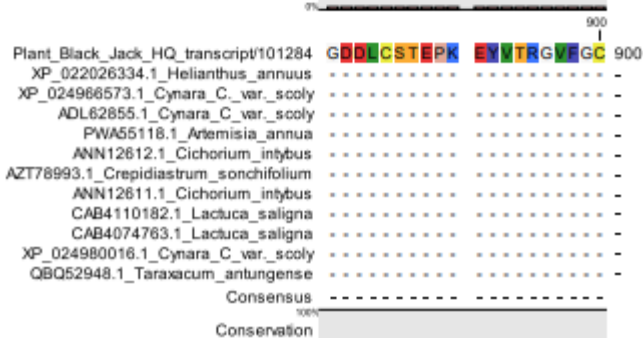
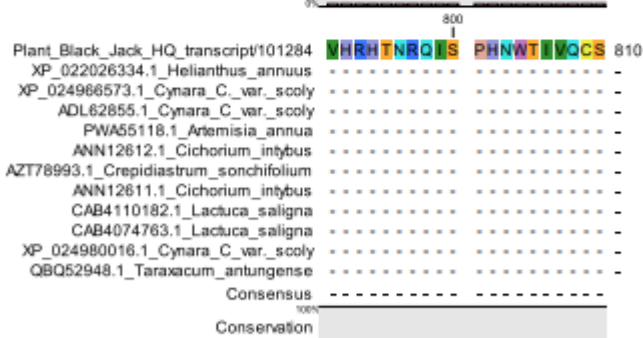
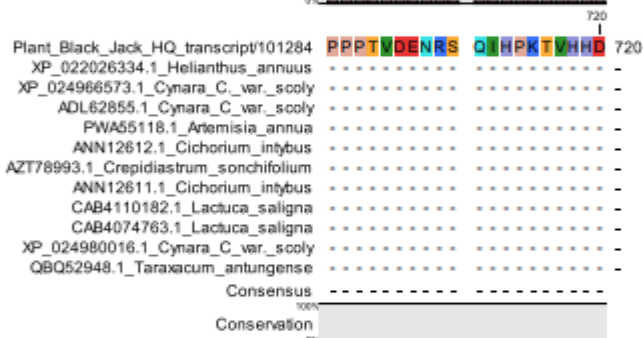
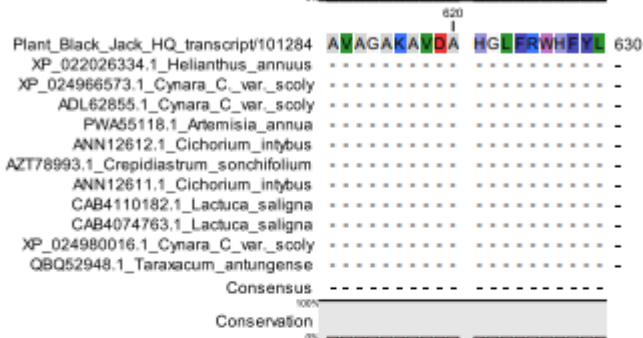
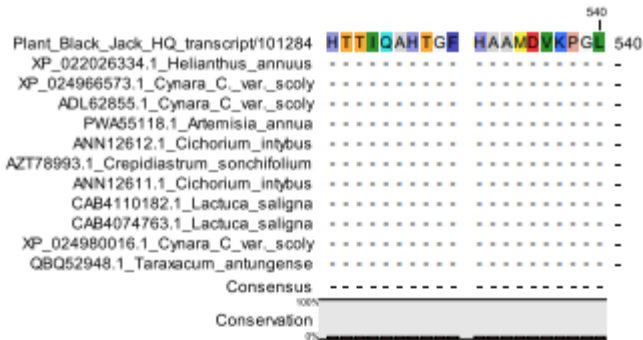
Figure S10. Full multiple sequence alignment of *B. pilosa* HQT2 gene.

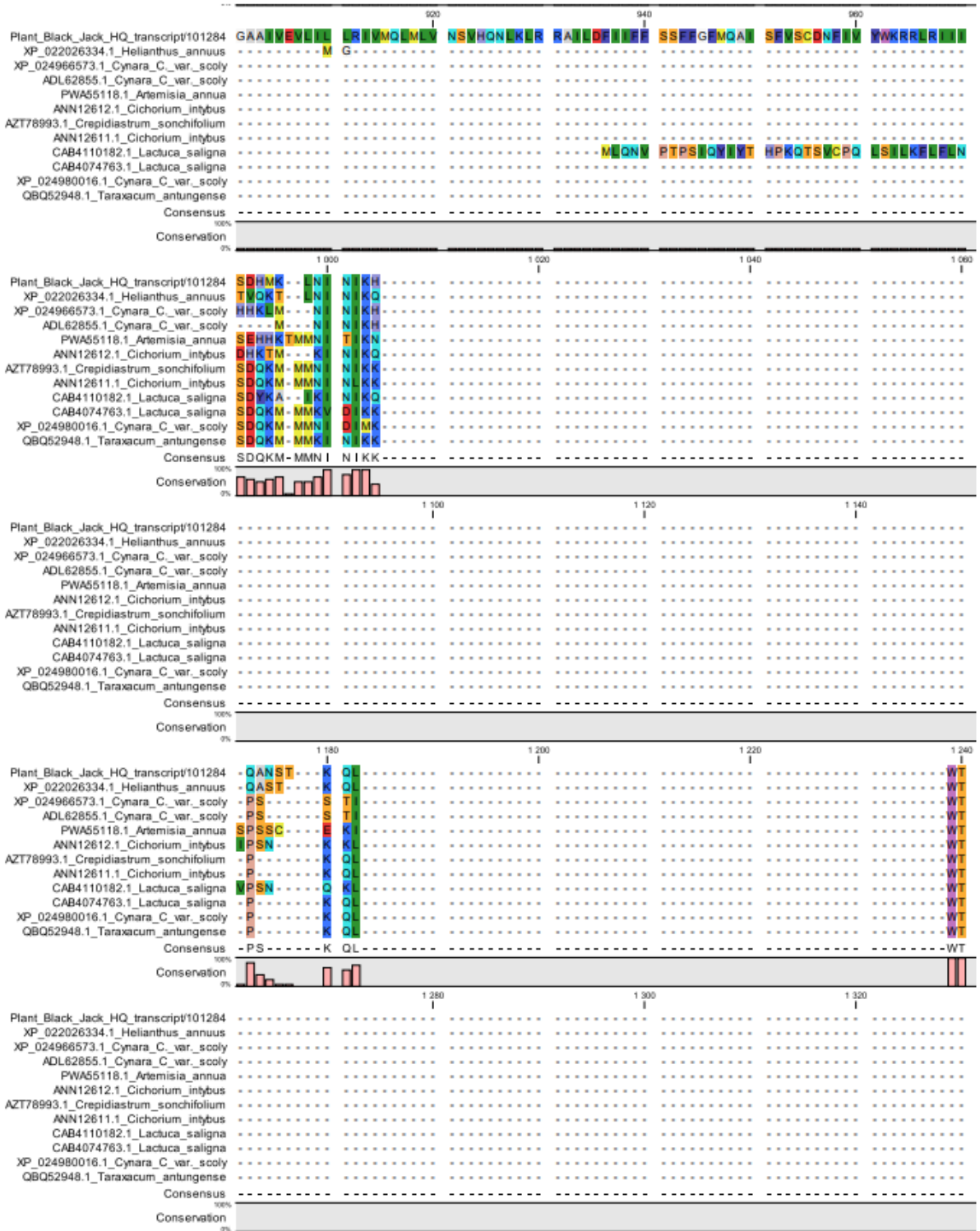
Multiple sequence alignment of *B. pilosa* HQT1 with its homologues from *Helianthus annuus* (QBM78938.1), *Cynara cardunculus var scolymus* (AFL93687.1), *Artemisia annua* (PWA39281.1), *Lactuca sativa* (XP_023733842.1), *Mikania micrantha* (KAD5794970.1), *Lonicera japonica* (AEK80405.1), *Chicorium intybus* (ANN12610.1) and *Tanacetum cinerariifolium* (GEV77257.1). Residues are grouped according to colours, for instance same colour represent similar residues across all genes from different plants. The alignment was generated using MUSCLE algorithm of the MEGA software. The position of the residue is shown by the number on the right.

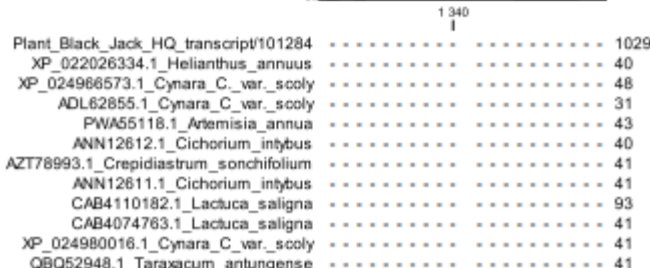
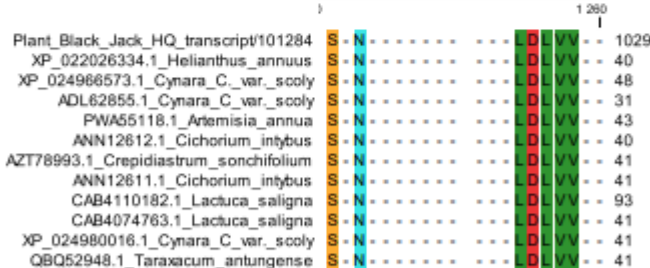
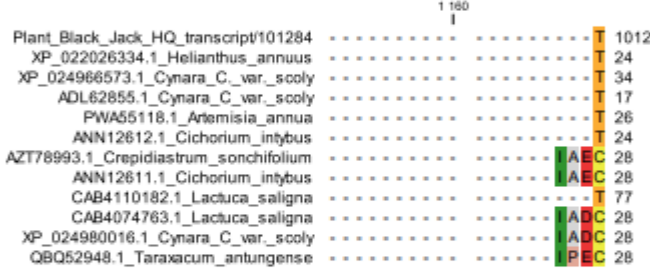
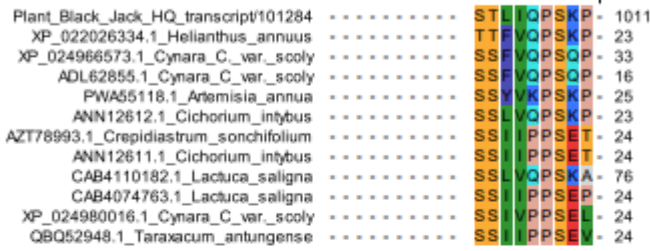
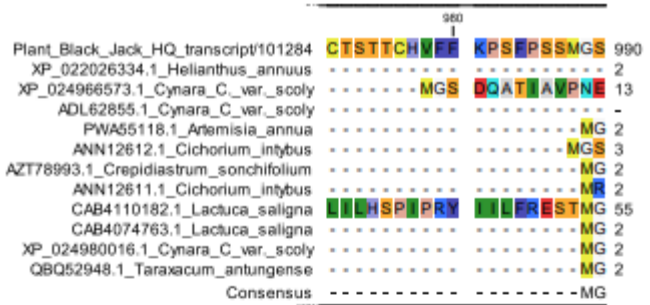


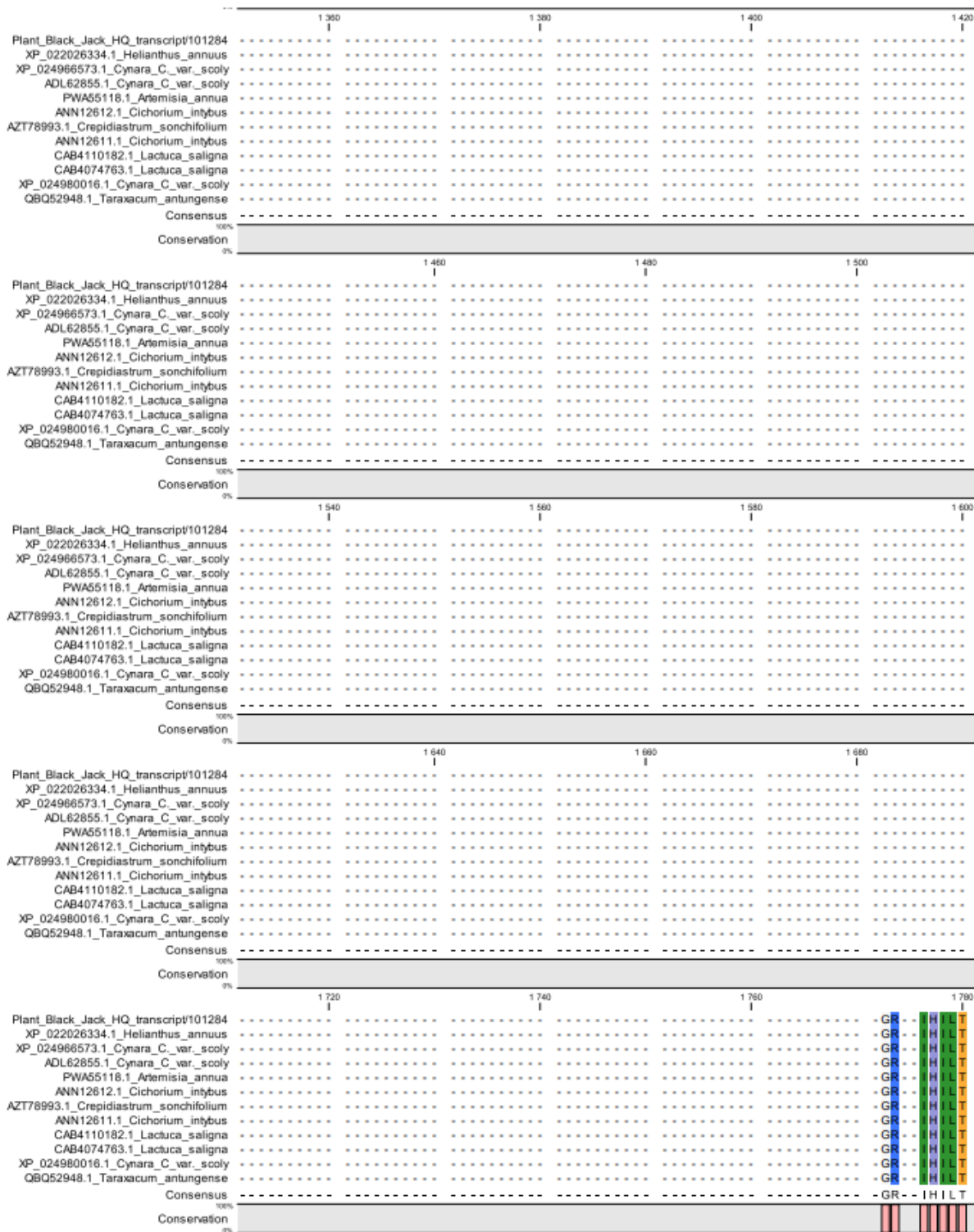


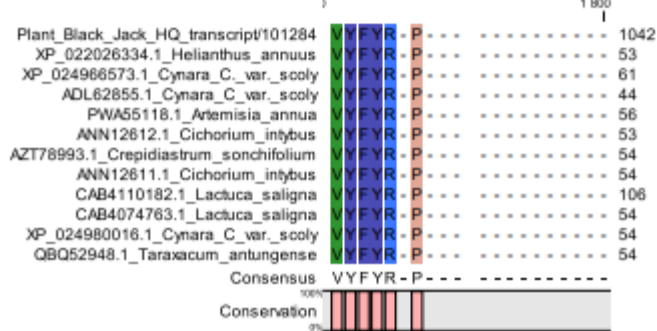
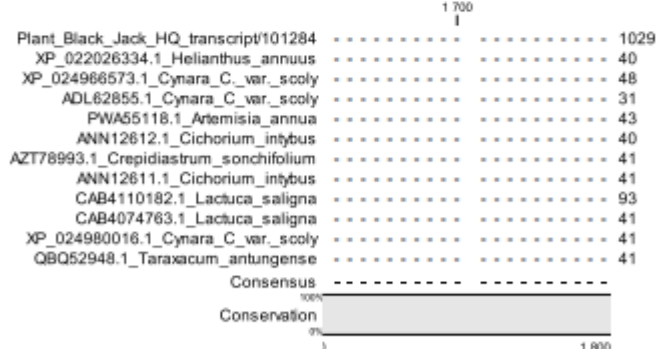
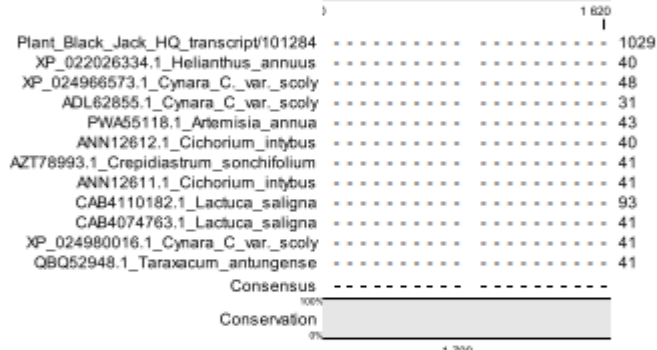
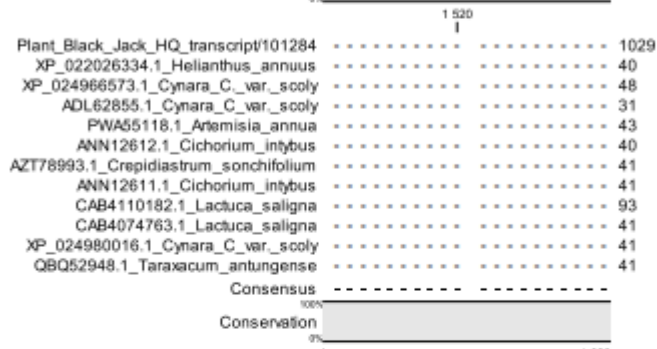
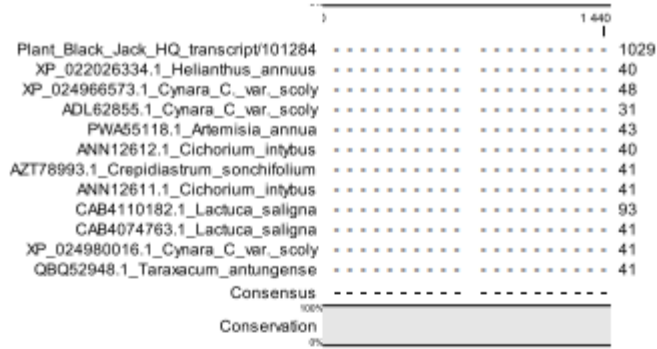


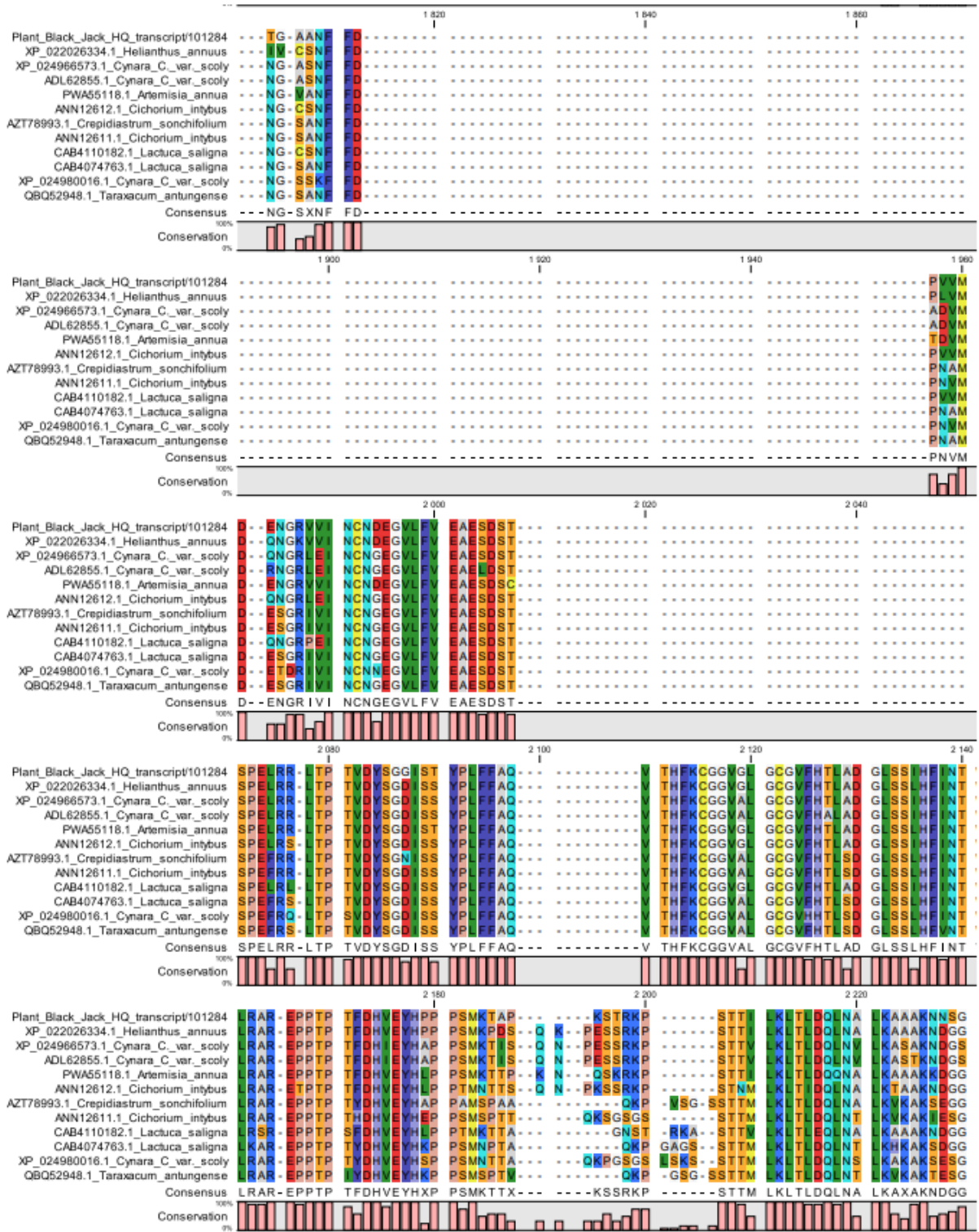


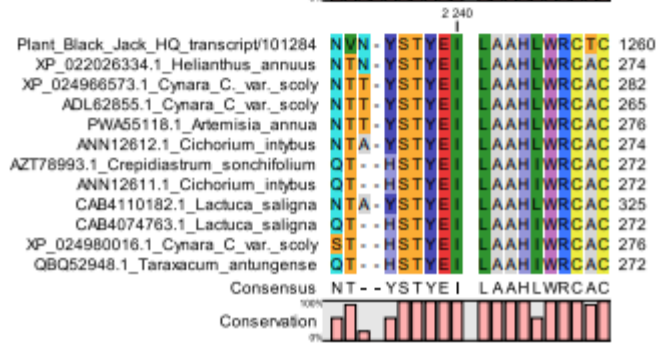
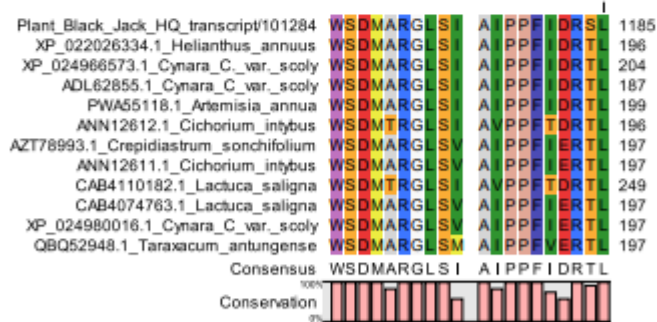
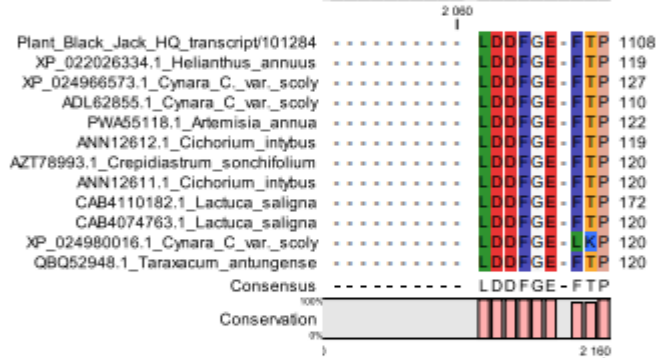
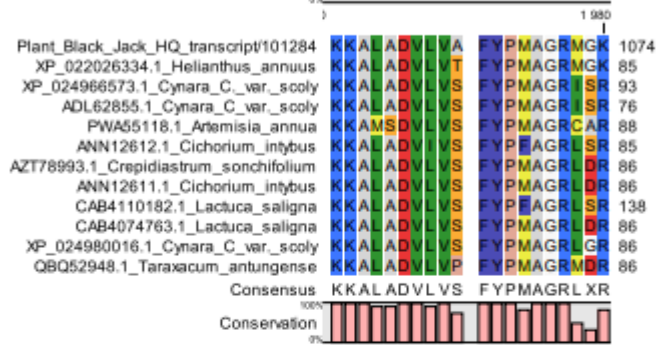
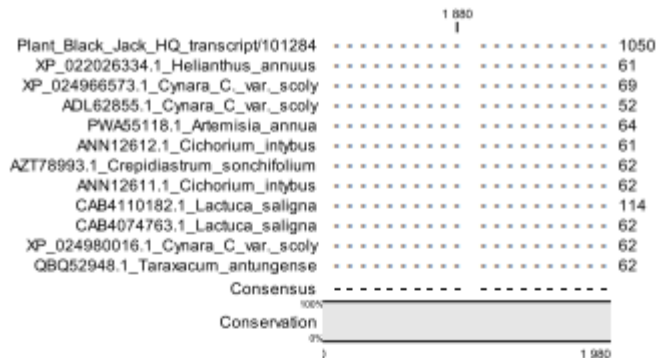


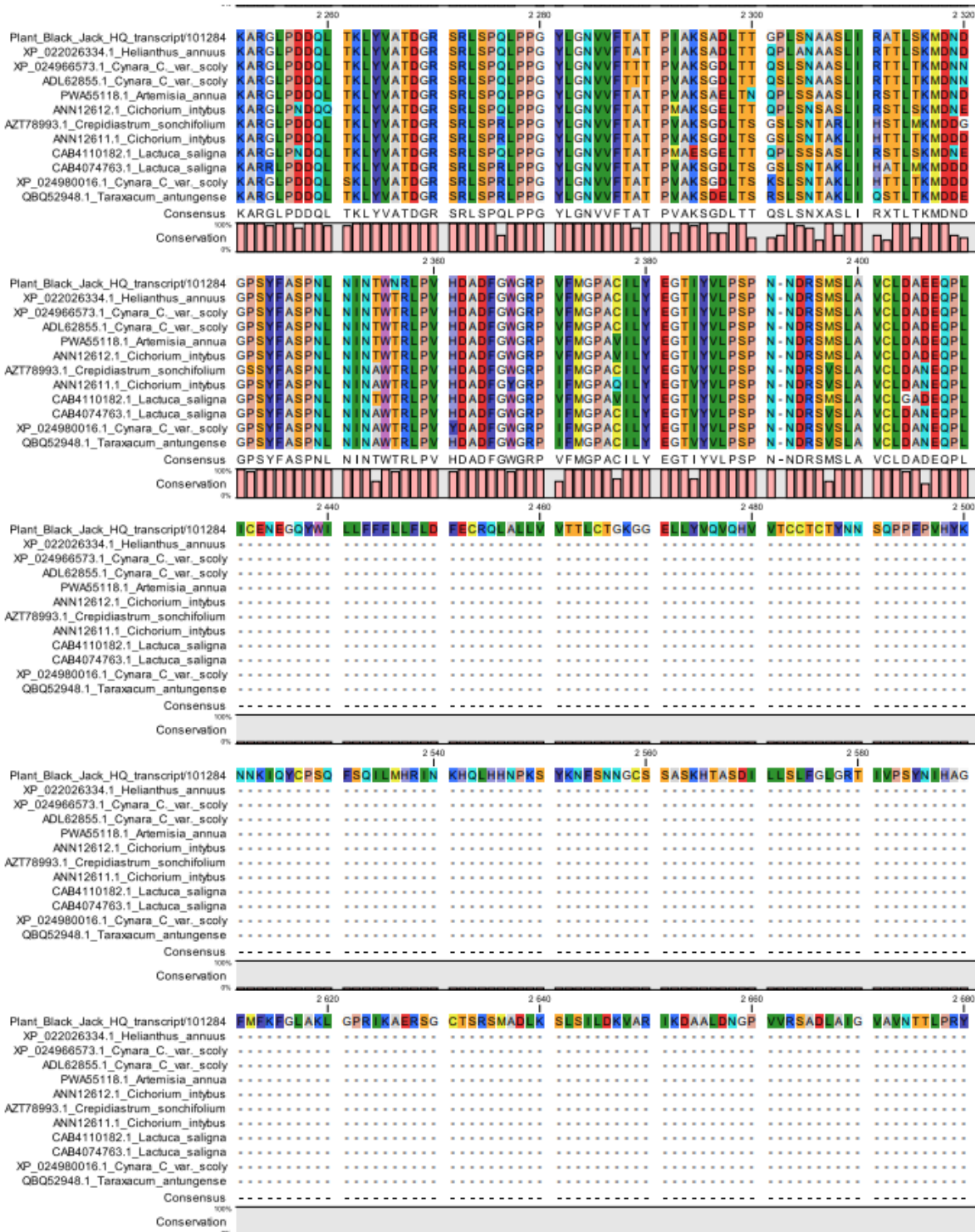


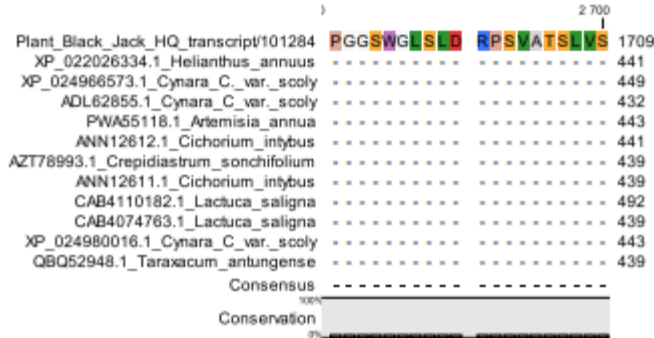
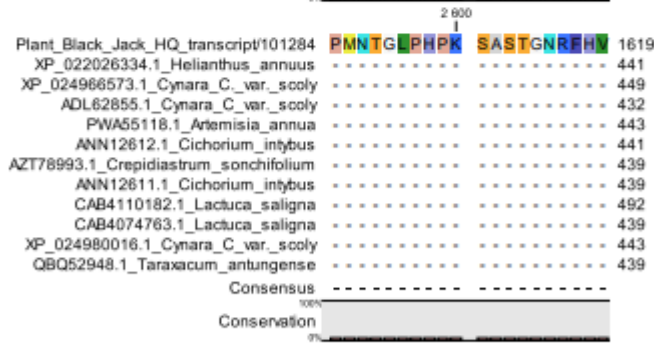
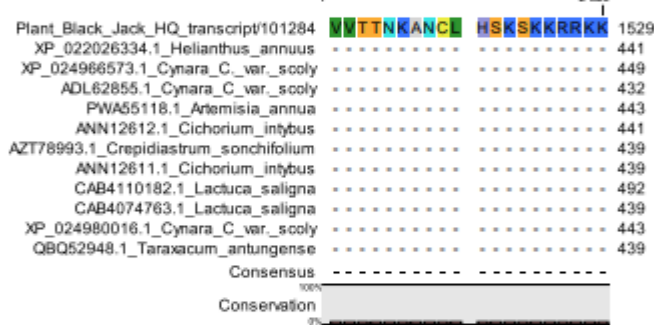
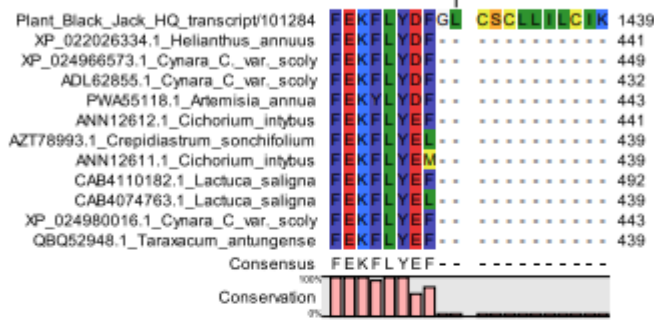
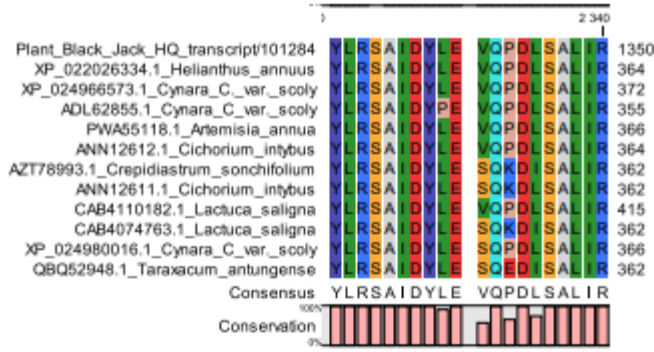


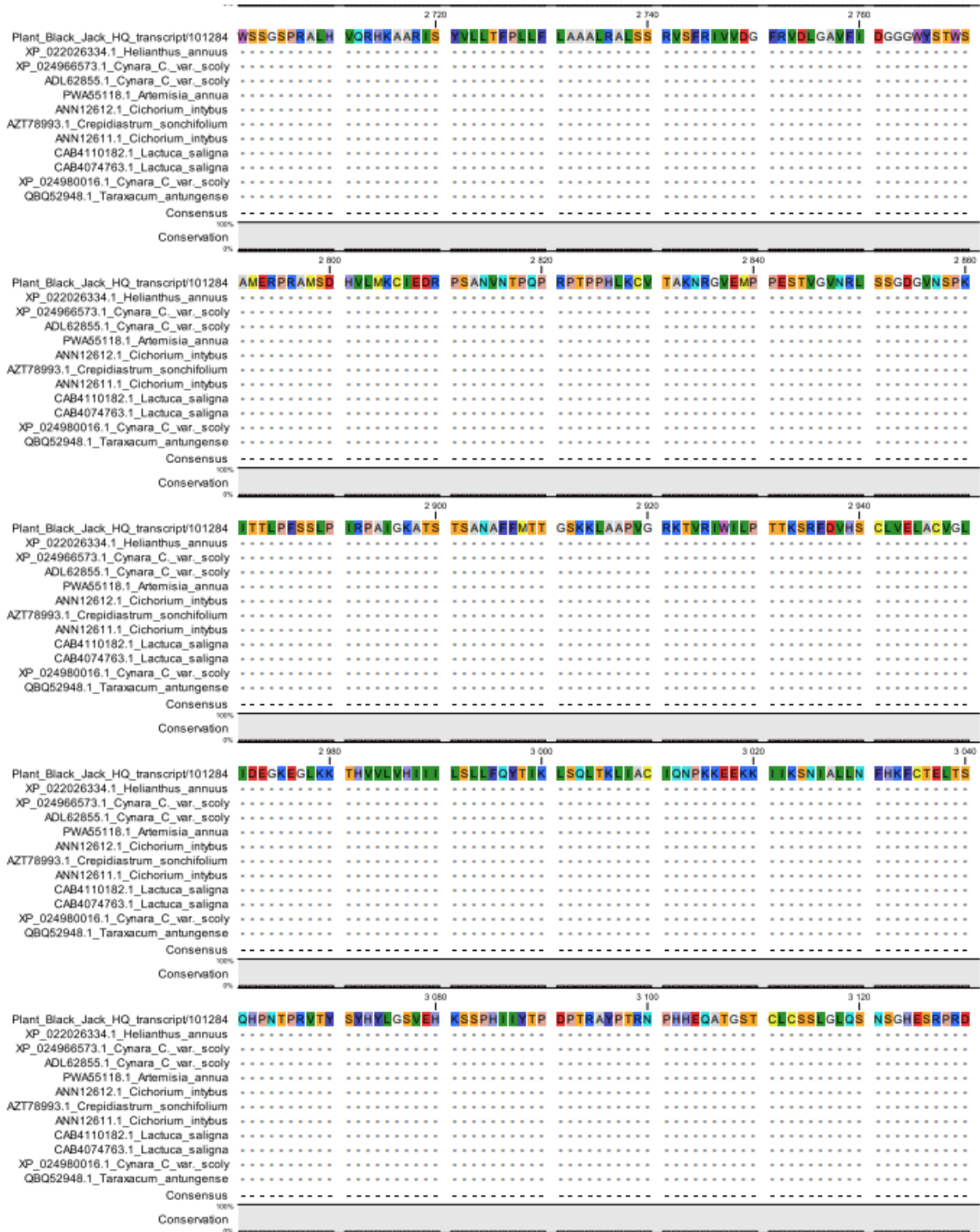












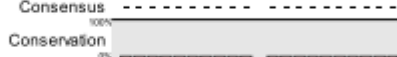
2 750
↓

Plant_Black_Jack_HQ_transcript101284	N	Y	G	G	S	R	A	R	S	K	E	R	S	M	N	G	G	M	1795
XP_022026334.1_Helianthus_annuus	441
XP_024966573.1_Cynara_C_var_scoly	449
ADL62855.1_Cynara_C_var_scoly	432
PWA55118.1_Artemisia_annua	443
ANN12612.1_Cichorium_intybus	441
AZT78993.1_Crepidiastrium_sonchifolium	439
ANN12611.1_Cichorium_intybus	439
CAB4110182.1_Lactuca_saligna	492
CAB4074763.1_Lactuca_saligna	439
XP_024980016.1_Cynara_C_var_scoly	443
QBQ52948.1_Taraxacum_antungense	439



2 850
↓

Plant_Black_Jack_HQ_transcript101284	S	S	N	E	S	S	A	S	T	N	K	T	P	S	S	L	C	1885
XP_022026334.1_Helianthus_annuus	441
XP_024966573.1_Cynara_C_var_scoly	449
ADL62855.1_Cynara_C_var_scoly	432
PWA55118.1_Artemisia_annua	443
ANN12612.1_Cichorium_intybus	441
AZT78993.1_Crepidiastrium_sonchifolium	439
ANN12611.1_Cichorium_intybus	439
CAB4110182.1_Lactuca_saligna	492
CAB4074763.1_Lactuca_saligna	439
XP_024980016.1_Cynara_C_var_scoly	443
QBQ52948.1_Taraxacum_antungense	439



2 950
↓

Plant_Black_Jack_HQ_transcript101284	G	C	M	S	V	E	C	M	E	M	S	F	M	S	E	D	P	1975
XP_022026334.1_Helianthus_annuus	441
XP_024966573.1_Cynara_C_var_scoly	449
ADL62855.1_Cynara_C_var_scoly	432
PWA55118.1_Artemisia_annua	443
ANN12612.1_Cichorium_intybus	441
AZT78993.1_Crepidiastrium_sonchifolium	439
ANN12611.1_Cichorium_intybus	439
CAB4110182.1_Lactuca_saligna	492
CAB4074763.1_Lactuca_saligna	439
XP_024980016.1_Cynara_C_var_scoly	443
QBQ52948.1_Taraxacum_antungense	439



3 050
↓

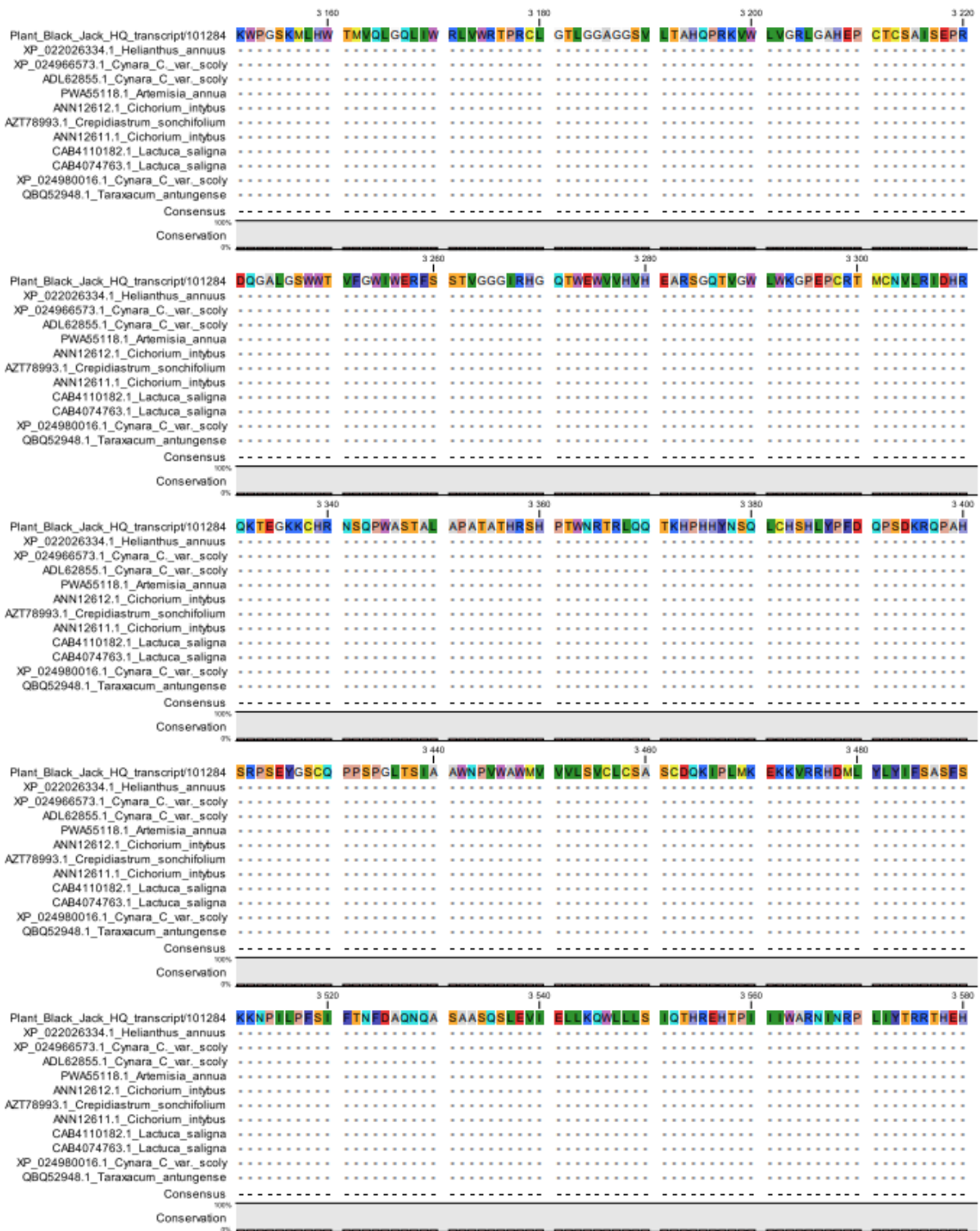
Plant_Black_Jack_HQ_transcript101284	I	S	C	T	I	L	R	S	I	R	T	S	O	T	M	A	A	P	2085
XP_022026334.1_Helianthus_annuus	441
XP_024966573.1_Cynara_C_var_scoly	449
ADL62855.1_Cynara_C_var_scoly	432
PWA55118.1_Artemisia_annua	443
ANN12612.1_Cichorium_intybus	441
AZT78993.1_Crepidiastrium_sonchifolium	439
ANN12611.1_Cichorium_intybus	439
CAB4110182.1_Lactuca_saligna	492
CAB4074763.1_Lactuca_saligna	439
XP_024980016.1_Cynara_C_var_scoly	443
QBQ52948.1_Taraxacum_antungense	439

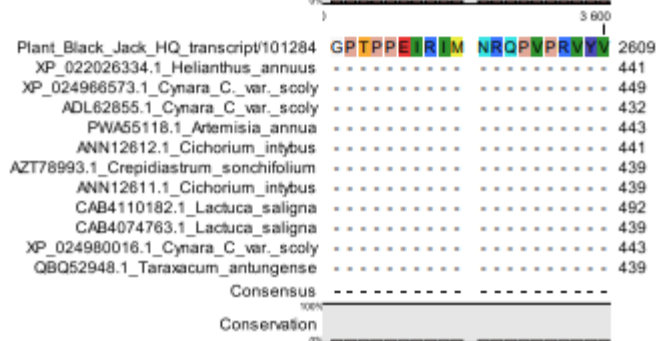
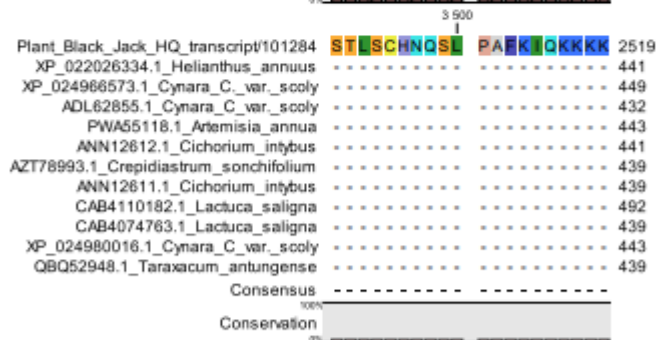
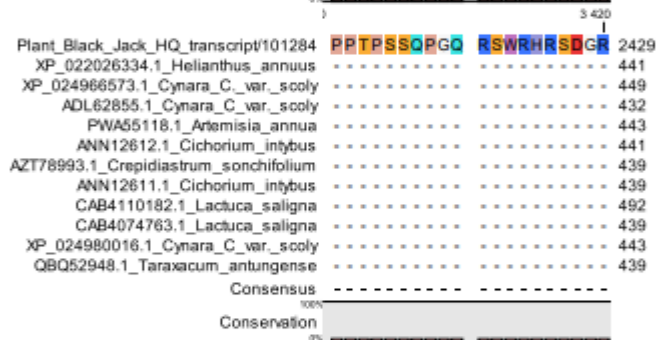
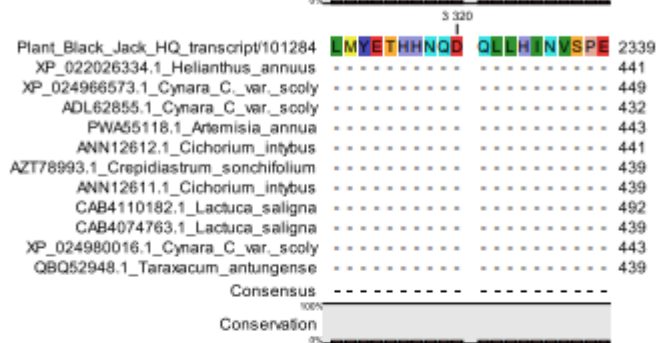
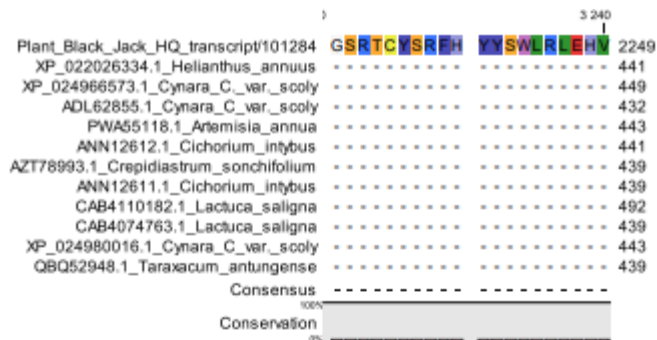


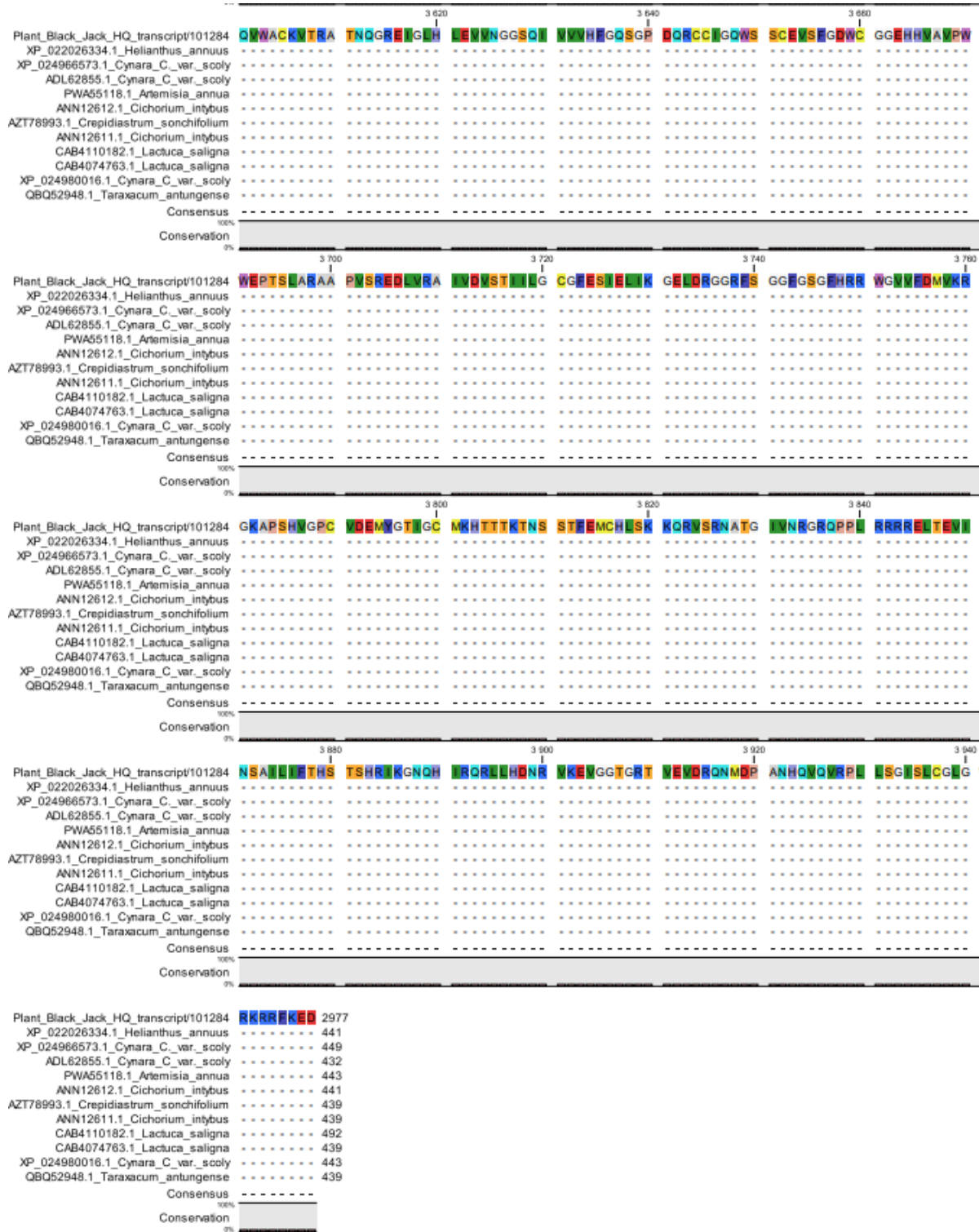
3 140
↓

Plant_Black_Jack_HQ_transcript101284	R	A	A	P	R	G	S	Q	W	R	I	S	N	S	R	C	P	W	T	2155
XP_022026334.1_Helianthus_annuus	441
XP_024966573.1_Cynara_C_var_scoly	449
ADL62855.1_Cynara_C_var_scoly	432
PWA55118.1_Artemisia_annua	443
ANN12612.1_Cichorium_intybus	441
AZT78993.1_Crepidiastrium_sonchifolium	439
ANN12611.1_Cichorium_intybus	439
CAB4110182.1_Lactuca_saligna	492
CAB4074763.1_Lactuca_saligna	439
XP_024980016.1_Cynara_C_var_scoly	443
QBQ52948.1_Taraxacum_antungense	439









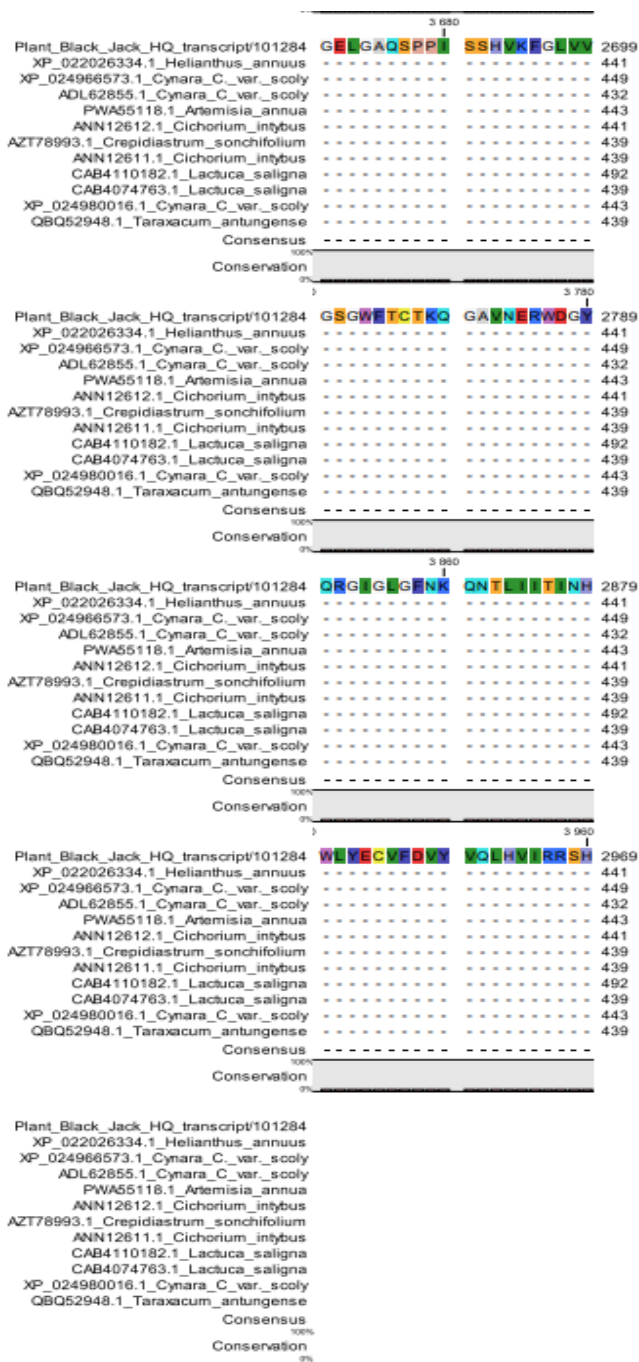


Figure S11. Full multiple sequence alignment of *B. pilosa* HQT3 gene.

Multiple sequence alignment of *B. pilosa* HQT1 with its homologues from *Helianthus annuus* (QBM78938.1), *Cynara cardunculus var scolymus* (AFL93687.1), *Artemisia annua* (PWA39281.1), *Lactuca sativa* (XP_023733842.1), *Mikania micrantha* (KAD5794970.1), *Lonicera japonica* (AEK80405.1), *Chicorium intybus* (ANN12610.1) and *Tanacetum cinerariifolium* (GEV77257.1). Residues are grouped according to colours, for instance same colour represent similar residues across all genes from different plants. The alignment was generated using MUSCLE algorithm of the MEGA software. The position of the residue is shown by the number on the right.

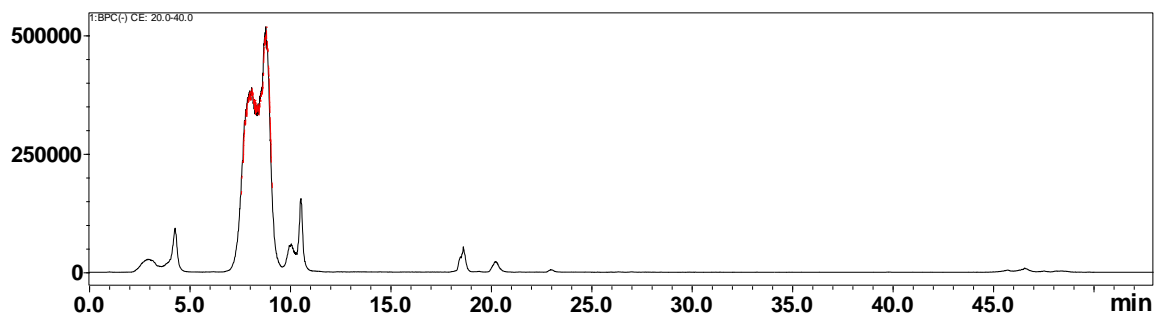


Figure S12. Representative UHPLC-qTOF-MS/MS chromatogram showing distribution patterns of chlorogenic acid derivatives in *Bidens pilosa*, with Y-axis showing peak intensity and X-axis showing retention time (min). The distribution pattern is as per the mass at m/z 353.

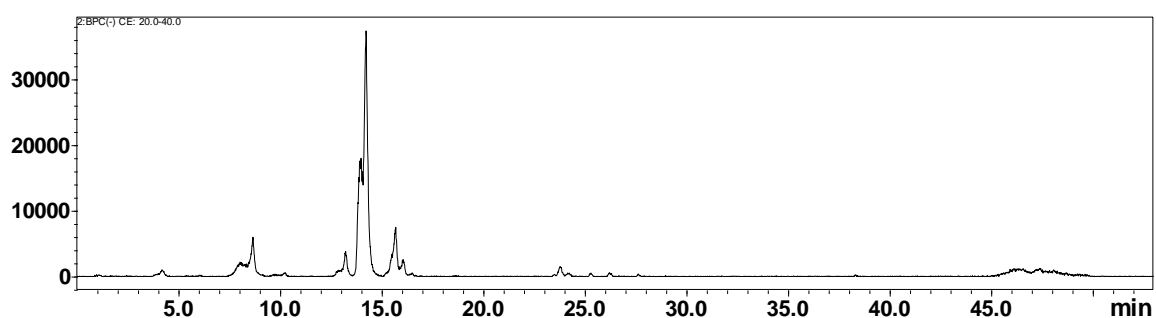


Figure S13. Representative UHPLC-qTOF-MS/MS chromatogram showing distribution patterns of chlorogenic acid derivatives in *Bidens pilosa*, with Y-axis showing peak intensity and X-axis showing retention time (min). The distribution pattern is as per the mass at m/z 367.

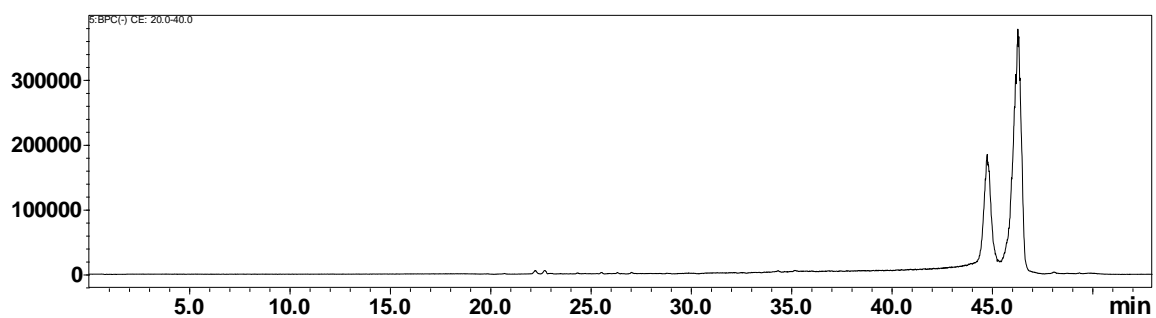


Figure S14. Representative UHPLC-qTOF-MS/MS chromatogram showing distribution patterns of chlorogenic acid derivatives in *Bidens pilosa*, with Y-axis showing peak intensity and X-axis showing retention time (min). The distribution pattern is as per the mass at m/z 499

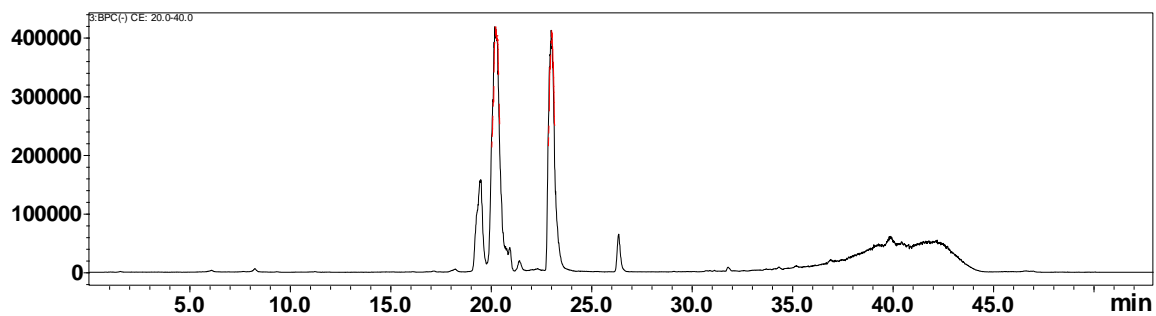


Figure S15. Representative UHPLC-qTOF-MS/MS chromatogram showing distribution patterns of chlorogenic acid derivatives in *Bidens pilosa*, with Y-axis showing peak intensity and X-axis showing retention time (min). The distribution pattern is as per the mass at m/z 515

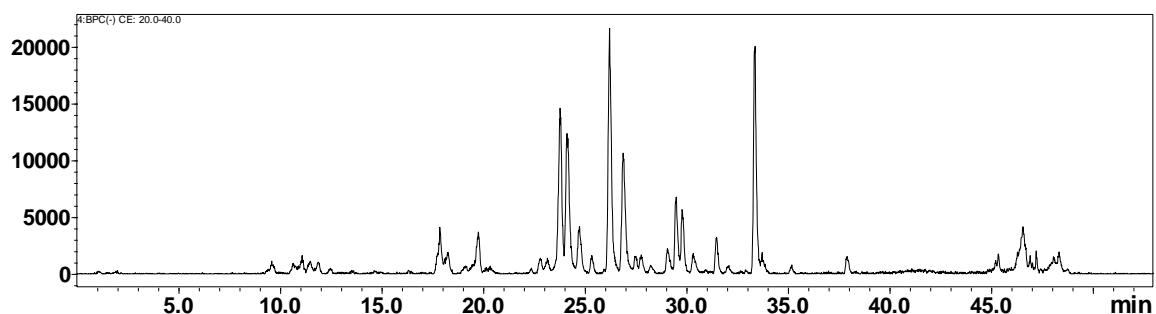


Figure S16. Representative UHPLC-qTOF-MS/MS chromatogram showing distribution patterns of chlorogenic acid derivatives in *Bidens pilosa*, with Y-axis showing peak intensity and X-axis showing retention time (min). The distribution pattern is as per the mass at m/z 529

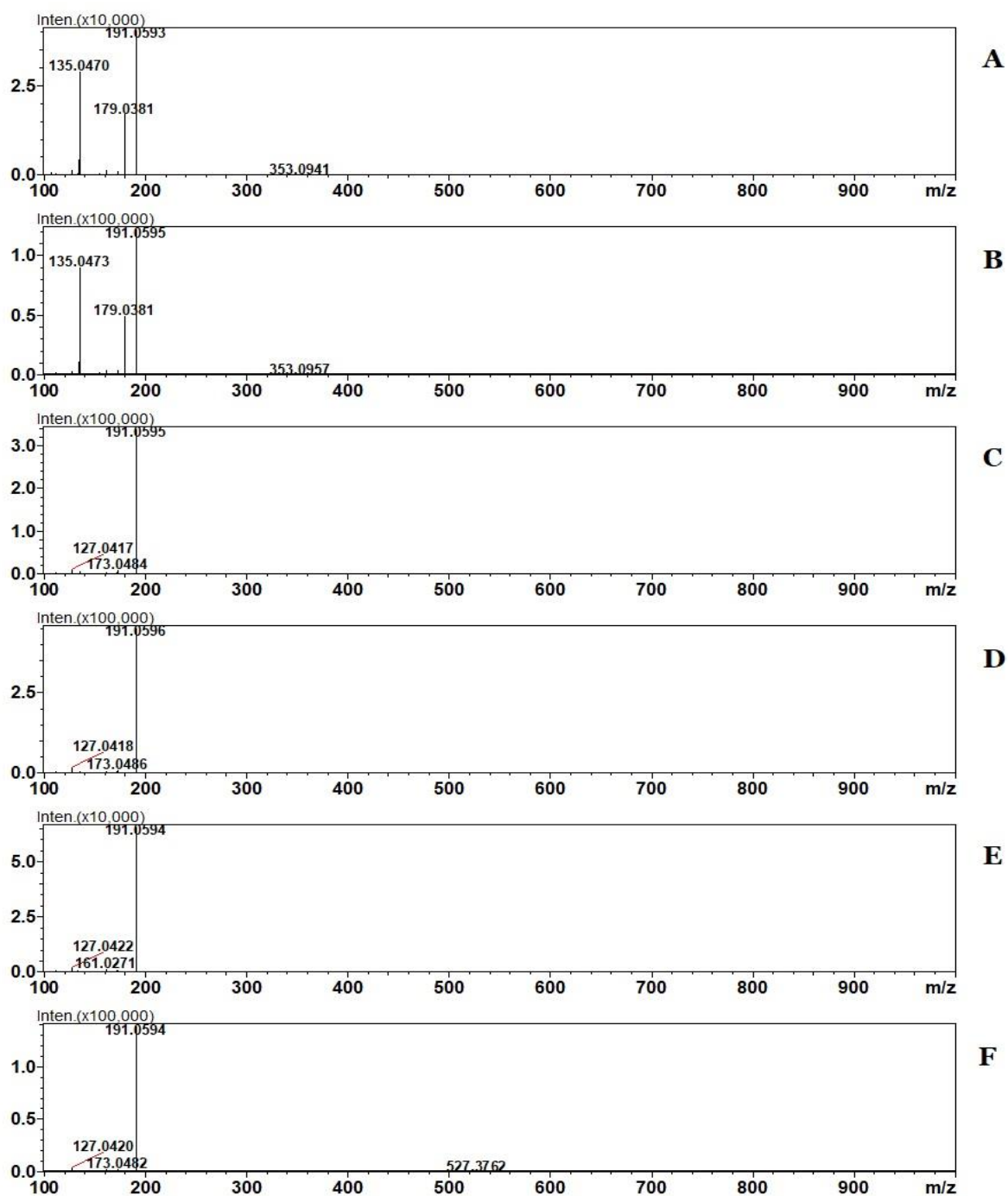


Figure S17. Typical mass spectra showing fragmentation of *cis*-3-Caffeoylquinic acid (A), *trans*-3-Caffeoylquinic acid (B), *cis*-4-Caffeoylquinic acid (C), *trans*-4-Caffeoylquinic acid (D), *trans*-5-Caffeoylquinic acid and *cis*-5-Caffeoylquinic acid (E)

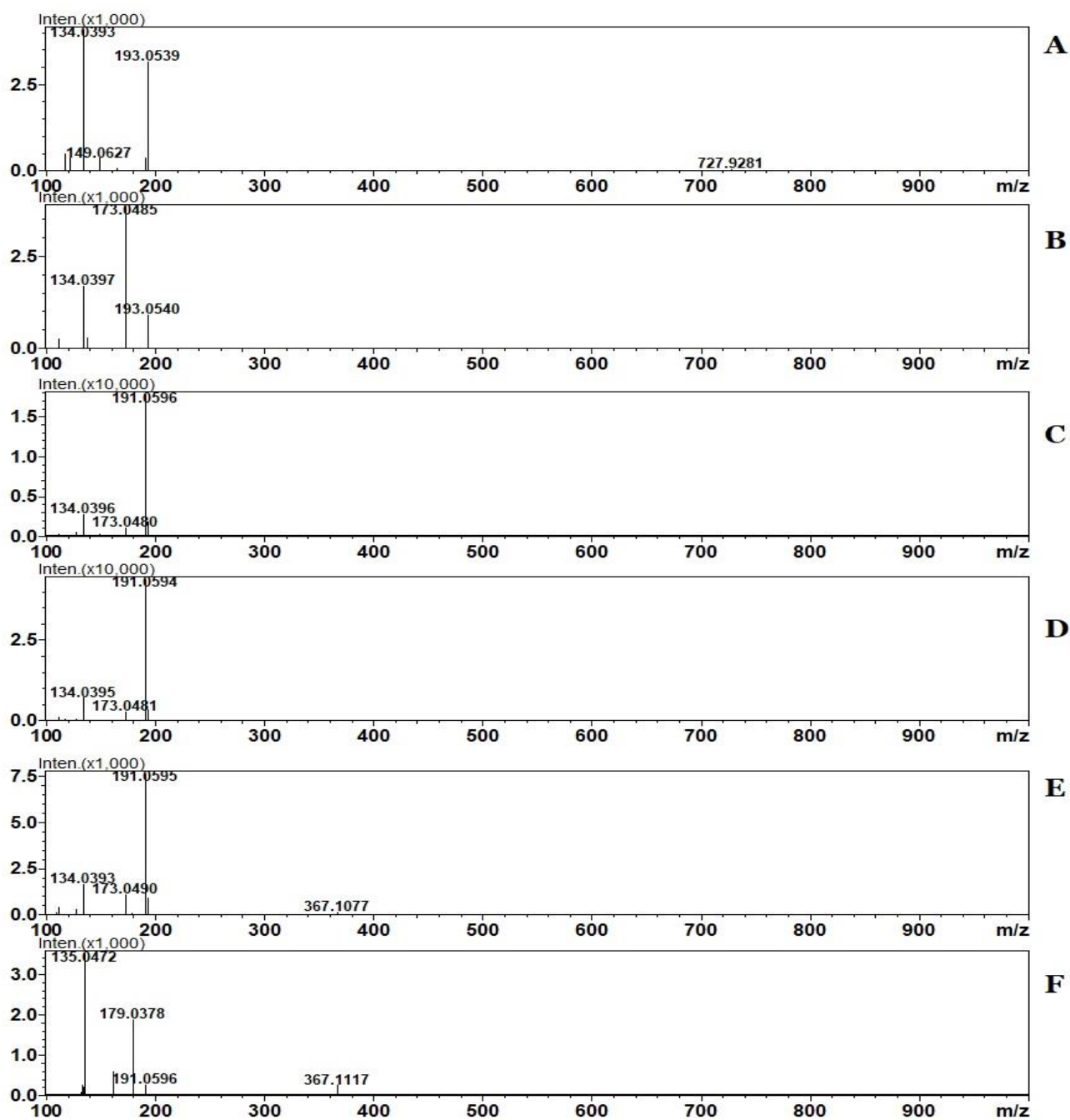


Figure S18. Typical mass spectra showing fragmentation of *trans*-3-Feruloylquinic acid (A), 3-Feruloylquinic acid (B), 5-Feruloylquinic acid (C), 5-Feruloylquinic acid (D), Feruloylquinic acid isomer (E) and Feruloylquinic acid isomer (F)

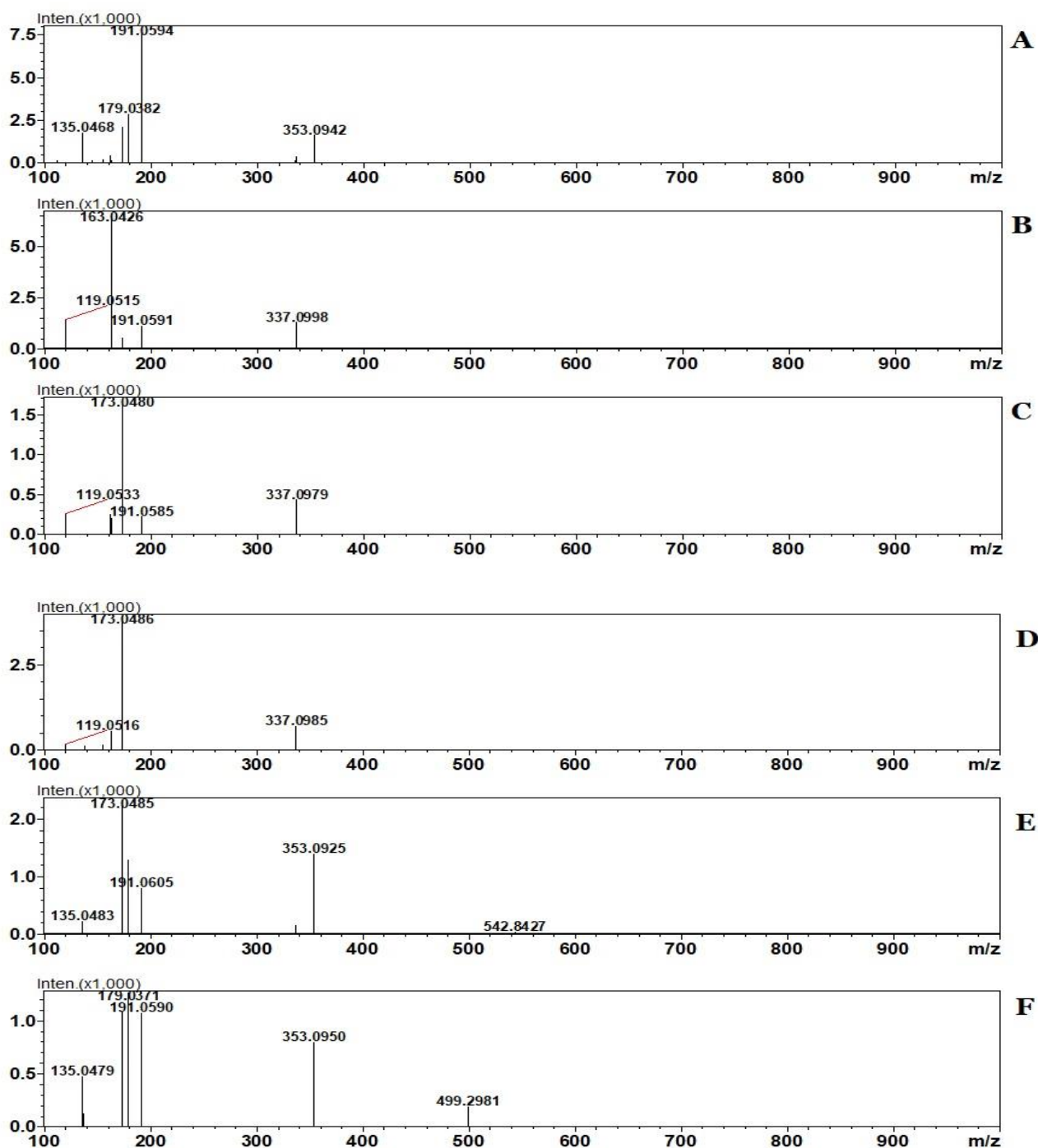


Figure S19. Typical mass spectra showing fragmentation of 3-Coumaroyl-4-caffeoylquinic acid (A), 3-Coumaroyl-5-caffeoylquinic acid (B), 3-Coumaroyl-4-caffeoylquinic acid (C), 3-Coumaroyl-4-caffeoylquinic acid (D), 4-Coumaroyl-5-caffeoylquinic acid (E) and 4-Coumaroyl-5-caffeoylquinic acid (F).

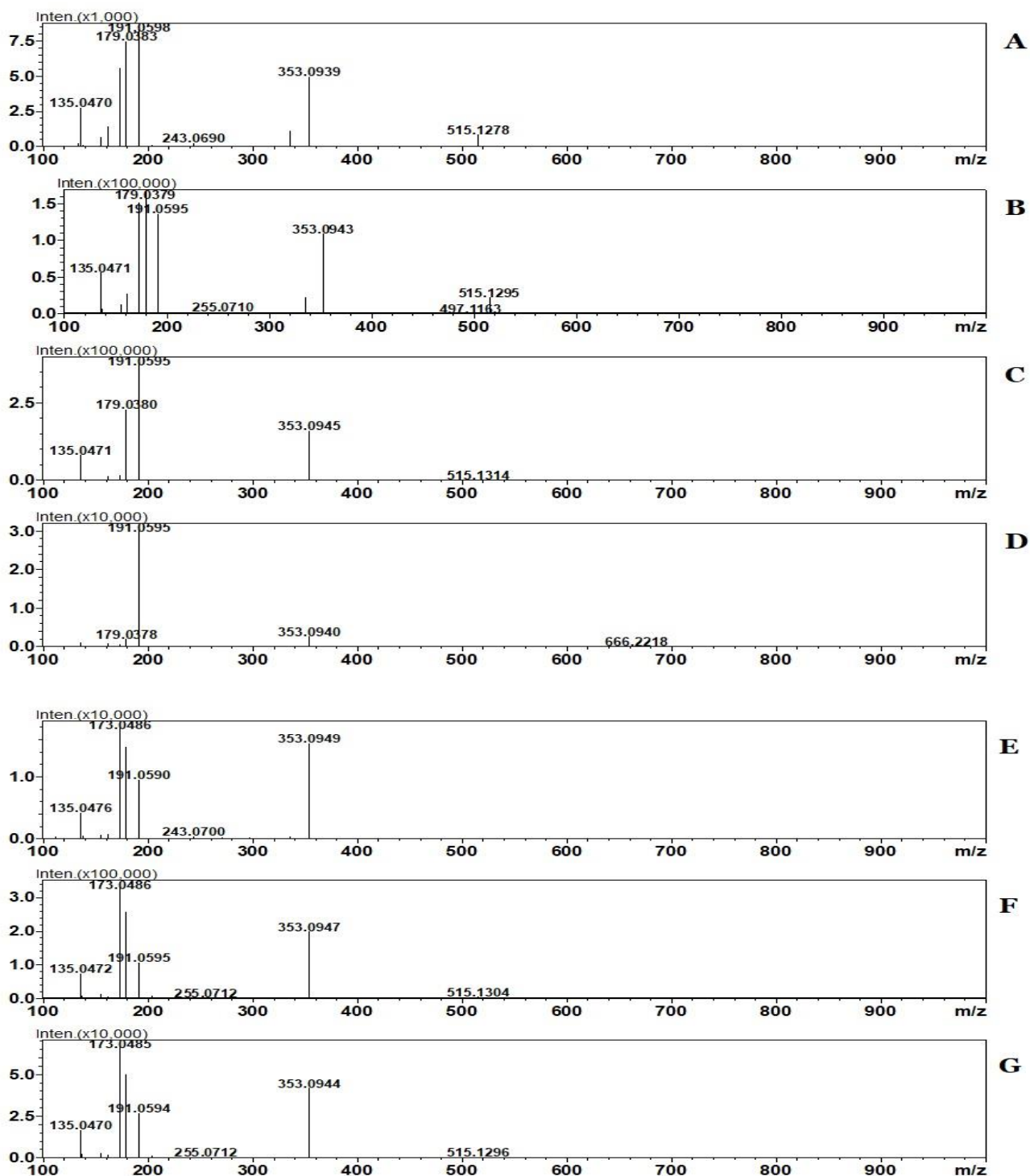
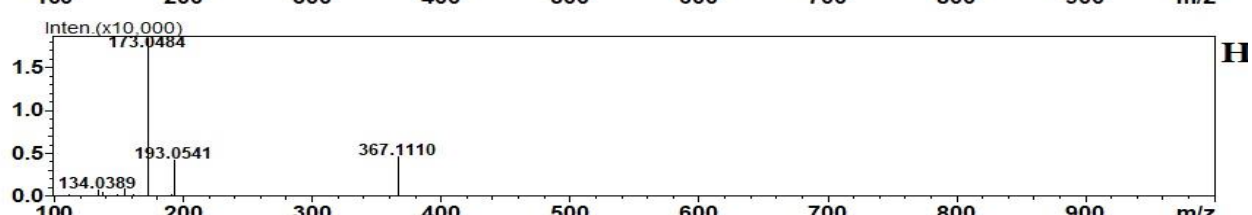
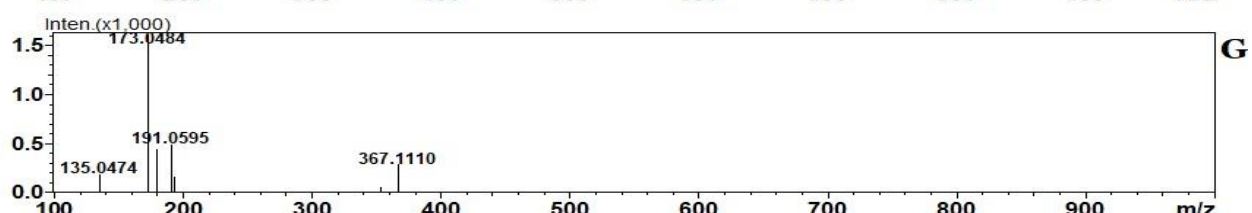
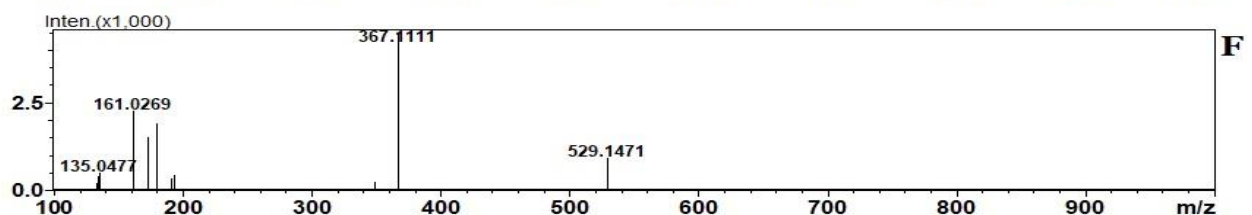
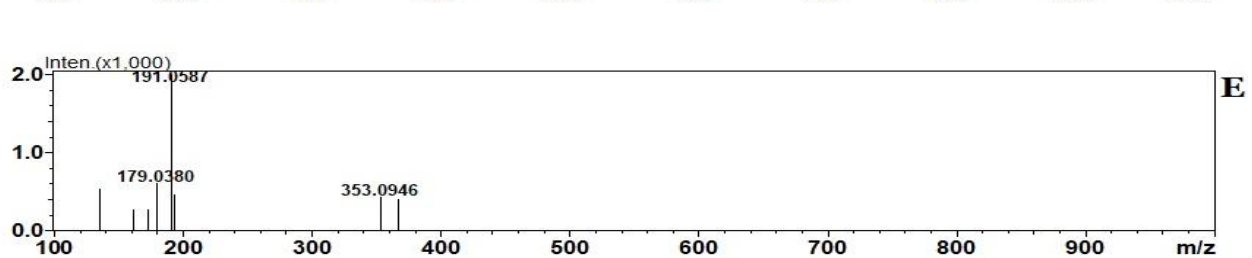
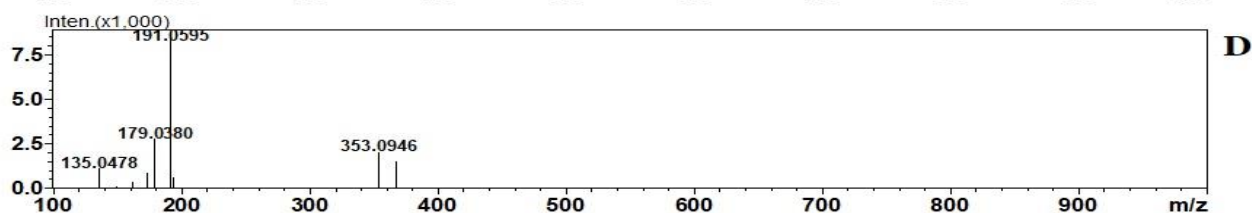
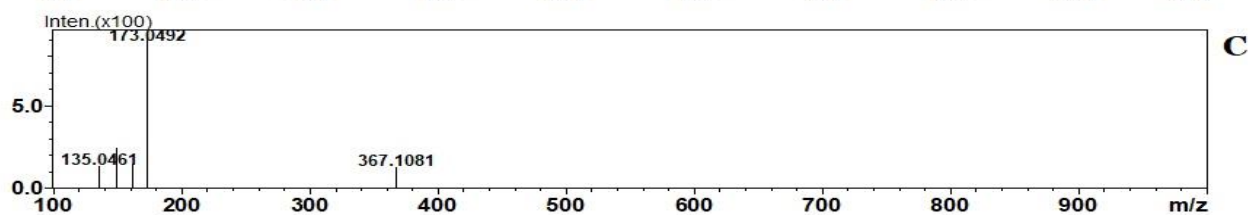
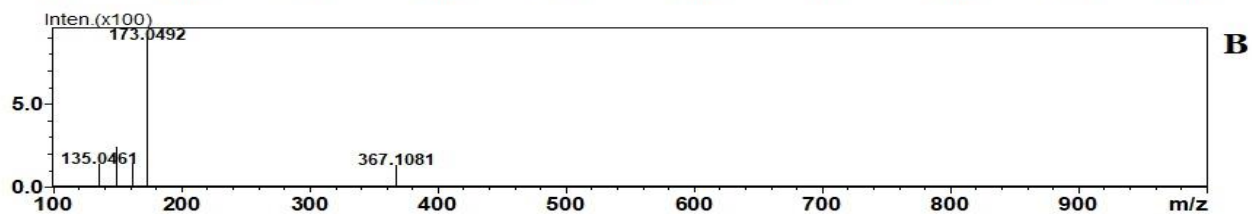
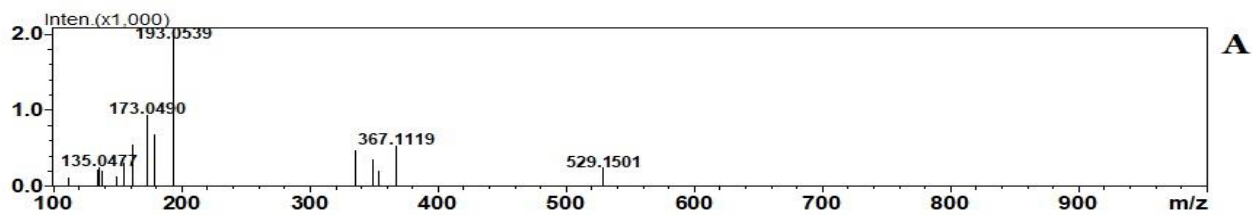


Figure S20. Typical mass spectra showing fragmentation of 3,4-di-Caffeoylquinic acid (A), 3,4-*di*-Caffeoylquinic acid (B), 3,5-*di*-Caffeoylquinic acid (C), 3,5-*di*-Caffeoylquinic acid (D), 4,5-*di*-Caffeoylquinic acid (E), 4,5-*di*-Caffeoylquinic acid (F) and 4,5-*di*-Caffeoylquinic acid (G)



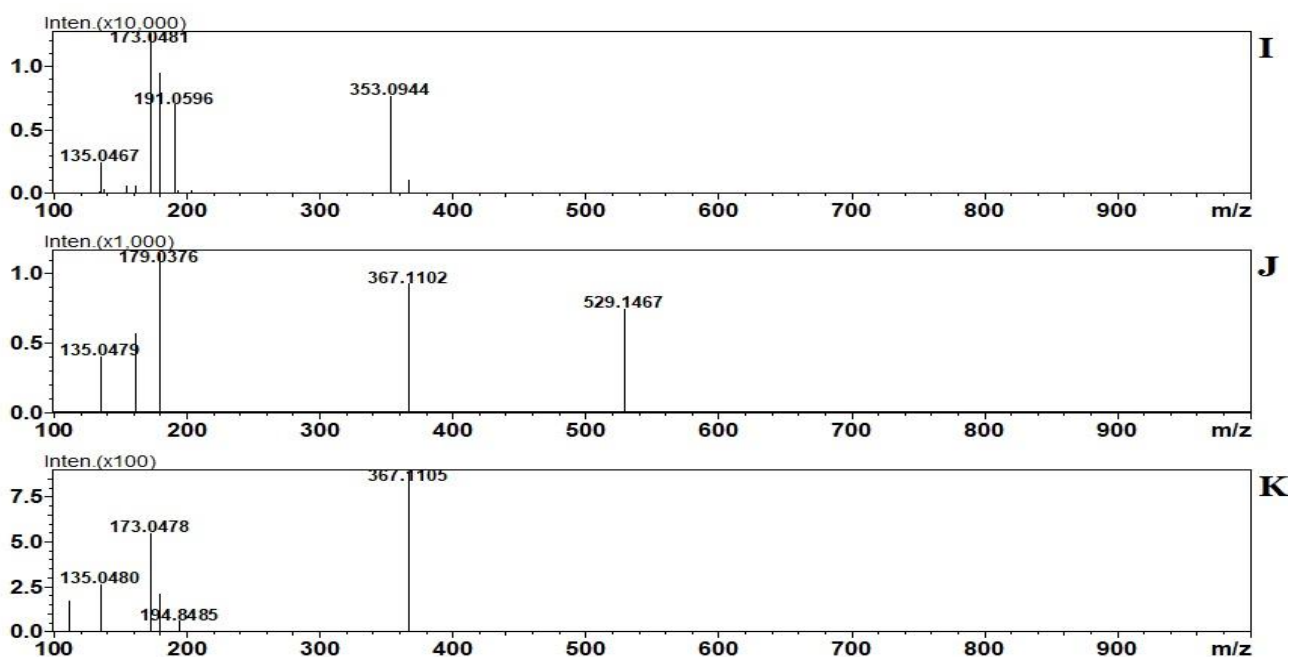


Figure S21. Typical mass spectra showing fragmentation of 3-Caffeoyl-4-feruloylquinic acid (A), 3-Caffeoyl-4-feruloylquinic acid (B), 3-Caffeoyl-4-feruloylquinic acid (C), 3-Feruloyl-5-caffeoylquinic acid (D), 3-Feruloyl-5-caffeoylquinic acid (E), Feruloyl-caffeoylquinic acid (F), 4-Caffeoyl-5-feruloylquinic acid (G), 4-Caffeoyl-5-feruloylquinic acid (H), 4-Caffeoyl-5-feruloylquinic acid (I), Feruloyl-caffeoylquinic acid (J) and Feruloyl-caffeoylquinic acid (K)

Chapter 3

Identification of putative acyl transferase genes responsible for biosynthesis of homogenous and heterogenous hydroxycinnamoyl-tartaric acid esters from *Bidens pilosa*

K Mathatha¹, A Khwathisi¹, A-T Ramabulana², I Mwaba¹ L.M Mathomu¹, and N.E Madala^{1*}

¹Department of Biochemistry and Microbiology, Faculty of Science, Engineering and Agriculture, University of Venda, Private Bag X5050, Thohoyandou, 0950 Limpopo, South Africa.

²Department of Biochemistry, University of Johannesburg, Auckland Park, Johannesburg, 2006 Gauteng, South Africa

Abstract

Bidens pilosa is an edible plant with highly sought-after nutraceutical properties. The purported bioactivities of this plant can be correlated to the high number of metabolites. Amongst these metabolites, different derivatives of hydroxy-cinnamoyl esters have been shown to exist in high proportions. However, the enzymatic machinery, thus the biosynthetic pathways responsible for the accumulation of these compounds in the plant have not yet been identified. We therefore report for the first time, the putative identification of two genes with sequence homology to hydroxycinnamoyl-CoA: tartaric acid hydroxycinnamoyl transferase (HTT) in *B. pilosa*. The full-length sequence of the two isoforms of HTT gene were achieved using single molecule real time (SMRT) sequencing approach. Analyses of methanolic extracts of *B. pilosa* through Liquid-chromatography hyphenated with mass spectrometry (LC-MS) technique revealed the existence of heterogeneous hydroxycinnamoyl-tartaric acid esters, consisting of mixed and different hydroxycinnamoyl derivatives. To the best of our knowledge, this is a first report on these molecules from *B. pilosa*. Taken altogether, this plant utilises hydroxycinnamoyl-CoA tartaric hydroxycinnamoyl transferase (HTT) genes to diversify its metabolite composition through esterification of tartaric acid acceptor molecule by acylating it with either homogeneous (same) or heterogenous (different) hydroxycinnamic acids (HCA) derivatives. Therefore, *B. pilosa* is a source of structurally diverse isomeric compounds with purported nutraceutical values. The enzyme products of the two identified HTT genes are therefore pointed out as possible catalysts which can be further exploited by incorporating them in other economically viable plants to enhance the nutraceutical values thereof.

Keywords: *Bidens pilosa*, Hydroxycinnamic acids, Hydroxycinnamoyl-CoA: tartaric acid hydroxycinnamoyl transferase gene, Single molecule real time sequencing, Liquid chromatography mass spectrometry

3.1 Introduction

For decades, plants have formed the backbone of complicated traditional medicine systems. One example of medicinal plants is *Bidens pilosa* from the *Asteraceae* family (Bartolome *et al.*, 2013), which is an erect, perennial herbaceous plant (Bashir *et al.*, 2018). *B. pilosa* is known to be a valuable source of nutraceuticals for both humans and animals (Feugang *et al.*, 2006, Sonnante *et al.*, 2010). *B. pilosa* is an underutilised plant often regarded as a weed whereas it possesses a wide array of important secondary metabolites (Lepelley *et al.*, 2007). The role of these secondary metabolites is to help the plant in defence and to adapt to the immediate environment (Tiwari and Rana, 2015; Bartwal *et al.*, 2013). To date, over 300 compounds of *B. pilosa* have been reported (Xuan and Khanh, 2016). The phytochemistry of *B. pilosa* comprises of aliphatic compounds, flavonoids, polyacetylenes, terpenoids, aromatics and hydroxycinnamic acids (Xuan and Khanh, 2016). Due to their structural diversity there is an increasing interest in the hydroxycinnamic acids and derivatives found in *B. pilosa* (Ramabulana *et al.*, 2020). Hydroxycinnamic acids (HCAs) comprise of a C6-C3 carbon skeleton and may occur as free forms (Garrido and Borges, 2013) or conjugated forms (Masike *et al.*, 2017). HCA compounds and their derivatives are ranked among the most valuable dietary phenolic compounds (Naveed *et al.*, 2018). Studies have shown that consumption of foods with high content of HCA-containing compounds such as coffee plays a pivotal role alleviating complications associated with diabetes (Lai *et al.*, 2015; Sonnante *et al.*, 2010; Scalbert *et al.*, 2005), skin cancer (Siddiqui *et al.*, 2018; Kang *et al.*, 2009), obesity and cardiovascular diseases (Li *et al.*, 2020; Gökçen and Şanlıer, 2019). These HCAs can be conjugated to sugar derivatives (Jaiswal *et al.*, 2014), quinic acid (Ncube *et al.*, 2014), citric acid (Masike *et al.*, 2017), polyamines (Mhlongo *et al.*, 2016) and tartaric acid (Nobela *et al.*, 2018). The structural diversity of these compounds arises from the regional-isomerisation owing to differential acylation of these acceptor molecules at different positions, which then diversify the metabolome pool of *B. pilosa* (Ramabulana *et al.*, 2020). Moreover, Molecules that contains HCA are also known to undergo geometrical isomerization (due to UV-light exposure) which then further amplifies the already diverse structural forms of these compounds (Clifford *et al.*, 2008; Zheng *et al.*, 2017). In *B. pilosa*, a common HCA derivative is chicoric acid, a molecule comprising of two caffeic acids

attached to a tartaric acid (Lee and Scagel, 2013). The chicoric acid from *B. pilosa* has been shown to form additional three geometrical isomers post UV-exposure owing to the stereochemical uniqueness of the tartaric acid found in *B. pilosa* (Nobela *et al.*, 2018). Consequently, the chicoric acid from this plant is a *meso*-chicoric acid which is rarely found in plants as compared to the most abundant L-chicoric acid form (Lee and Scagel, 2013). Chicoric acid is produced through the phenylpropanoid pathway and has been identified in different plants such as *Lactiva sativa* (Lee and Scagel, 2013), *Echinacea purpurea* (Fu *et al.*, 2021) and *Equisetum arvense*. The biosynthetic pathway for production of chicoric acid in *E. purpurea* has been elucidated (Lee and Scagel, 2013). BAHD acyltransferase gene, hydroxycinnamoyl CoA: tartaric acid hydroxycinnamoyl transferase (HTT) plays a pivotal role in the synthesis of chicoric acid and has been identified in *E. purpurea* (Fu *et al.*, 2021).

As stated above, the metabolites in *B. pilosa* are structurally diverse and complex, and their differentiation and annotation pose an undisputed analytical challenge. However, through Liquid-chromatography hyphenated with mass spectrometry (LC-MS) technique, many metabolites including HCA derivatives have been identified *in B. pilosa* (Ramabulana *et al.*, 2021). Apart from the few examples provided above, there is very few literature on the genetic machinery responsible for the biosynthesis of tartaric acid derivatives, and more research is needed in order to place these metabolites on the phenylpropanoid pathway, similarly to what is known about their structural counterparts, the quinic acid derivatives (Mudau *et al.*, 2018). Transcriptomic analyses through next generation sequencing (NGS) approach have been found to be a feasible approach to identify genes responsible for most metabolites found in different plants, such as transcriptomic data of *Aconitum heterophyllum* (Pal *et al.*, 2013), microRNA of *Picrorhiza kurroa* (*P. kurroa*) (Vashisht *et al.*, 2015), eleven *Rosa roxburghii* metabolism gene (Lu *et al.*, 2016). Elsewhere, genes responsible for chlorogenic acids biosynthesis have been identified through NGS sequencing (Kim *et al.*, 2013). There are several limitations associated with traditional NGS approach such as short read length (Rhoads and Au, 2015), which makes it difficult to assemble and to detect isoforms (Ma *et al.*, 2019). To overcome these challenges, a new way of sequencing which allows for a full-length transcriptome known as SMRT sequencing has been developed (Rhoads and Au, 2015), through this approach multiple genes

from various plants have been identified (Cui *et al.*, 2020; Xiang *et al.*, 2016; Malar *et al.*, 2019). In this study, SMRT sequencing approach and LC-MS technique were used to identify potential genes responsible for the biosynthesis of hydroxycinnamoyl-tartaric acid esters and their compositions respectively.

3.2. Methodology

3.2.1. Total RNA isolation

B. pilosa plants were grown at the university of Venda greenhouse. Four weeks old plants were taken to inqaba biotech in Pretoria, South Africa in separate pots. Total RNA was extracted from 1 gram leaf material using *Quick* – RNA plant mini prep kit was used for extraction of total RNA following manufacture’s protocol. Main steps followed were: Whole plants were cut with a razor to small pieces. Bashing beads with lysis buffer was used for RNA extraction. Zymo-spin IICG column were used to precipitate total RNA. Zymo-spin IICR column was used to filter unwanted biological components in the tube. RNA wash buffer was used to wash RNA and DNase/RNase free water was used to resuspend total RNA. Total RNA was quantified using Nanodrop 2000c spectrophotometer (Thermofischer Scientific).

3.2.2. cDNA preparation and sequencing

cDNA Library preparation for sequencing was done using the Iso-Seq™ Express Template Preparation for Sequel® and Sequel II Systems’ procedure and checklist as recommended. The main processes were as follows:

Total RNA was reverse transcribed into cDNA using a NEBNext® Single Cell cDNA Synthesis & Amplification Module and Iso-Seq Express Oligo Kit that was optimized for preparing high-quality, full-length cDNAs. The Oligo dT primer was used for the synthesis of cDNA and a random oligomer. The obtained full-length cDNA was amplified by PCR using random oligomers. The amplified product was purified by PB magnetic beads and quantified using Qubit 2.0. The amplified cDNA fragments were amplified by PCR for the second time and the full-length cDNA was purified by PB magnetic beads. The full-length cDNA were terminally repaired and the SMRT dumbbell adapter was connected. Exonuclease were used to digest the fragments that were not connected to the jointer. The PB magnetic beads were used for purification to obtain a sequencing library. After the library was constructed, Qubit 2.0 was used for accurate quantification. Then used Agilent 2100 to detect the library size. Sequencing was performed after the library size was qualified.

3.2.3. Identification of potential HTT genes from *B. pilosa*

Genes in phenylpropanoid pathway were retrieved (<https://www.ncbi.nlm.nih.gov/biosystems/493811>), and an inhouse database was created. A BLAST+ command line was used to search for similar genes in the *B. pilosa* full transcriptome sequences. Results were then subjected to filtering process as follows: *Echinacea purpurea*' HTT sequence was retrieved from NCBI, and it was used as a custom database against gene sequences found in *B. pilosa*. An excel file was created and all the 312 sequences from *B. pilosa* were further scrutinised to only leave the full-length sequences, thus those with start and stop codon. Furthermore, sequences that showed high similarity index to the HTT gene of *Echinacea purpurea* were marked as homologues on the excel file. This allowed the identification of only the genes of interest which are potential HTT genes in *B. pilosa*.

3.2.4. Percentage similarity

Percentage matrix index was computed with Multiple Alignment using Fast Fourier Transform (MAFFT) CD-hit of identified HTT gene sequences and homologue sequences retrieved from NCBI. The results were saved in a separate excel file where different colours were used to mark the most identical ones (>95%). The identical sequences were eliminated, preference given to sequences >1000 bp.

3.2.5. Integrity of Sequences

To check the integrity of the sequences, the Expert Protein Analysis System (ExpASy) was used. The identified HTT sequences were subjected to this computational tool to check if whether the sequence is truncated or not. The sequences were first mapped to the reference gene to establish the start and stop codons. Thereafter were taken to ExpASy to compute an open reading frame(s). The results were recorded in excel as truncated if the sequence could not generate an ORF that can be used downstream.

3.2.6. Multiple sequence alignment (MSA)

Identified potential HTT genes were further aligned with those of other plants obtained from NCBI in order to establish homology with sequences from other plants of different/same family. These sequences were aligned using MUSCLE built in MEGA (Hall, 2013) and viewed using GLC workbench.

3.2.7. Phylogenetic Analysis

Identified HTT protein sequences from *B. pilosa* together with the sequences retrieved from NCBI were exported to MEGA version 10.1.7 (Hall, 2013) and the neighbour joining (NJ) phylogenetic tree was constructed with the bootstrap value of 1000 (Mao *et al.*, 2019).

3.2.8. Metabolite extraction

Two grams (2 g) of fine ground leaves of *B. pilosa* were dissolved in 20 mL of 80% aqueous methanol. The mixture was spun throughout the night at 70 rpm to enhance the metabolite extraction. Tubes were placed in a rack to allow separation of supernatant (solvent) and pellet by gravity. The supernatant (1 mL) was transferred to a 2 mL Eppendorf tube. The samples were filtered twice using 1 mL syringe fitted with a 0.22 µm nylon filter (Agela Technologies, China) into a 2 mL vial fitted with 0.2 mL conical bottom glass insert. The samples were stored at 4 °C until analysis.

3.2.9. Liquid chromatography mass spectrometry

The extracts were analysed on an LC-qTOF-MS, model LC-MS 9030 instrument (Shimadzu, Kyoto Japan), fitted with a Shim Pack Velox C18 column (100 mm x 2.1 mm with particle size of 2.7 µm) (Shimadzu, Kyoto, Japan), placed in a column oven thermo-stated at 55 °C. A binary solvent system consisting of solvent A: 0.1% formic acid in water and solvent B: 0.1% formic acid in acetonitrile (UHPLC grade, Romil SpS, Cambridge, UK) was used with a total flow rate of 0.4 mL/min. Chromatographic separation of the analytes was achieved through a 53 min long gradient method consisting of the following steps: initial, 10% B for 3 min, followed by a steep gradient to 60% B over 40 min, constant at 60% for 3 min, increased to 90% in 2 min, kept at

90% over 3 min, and returned to 10% for 2 min and finally the initial conditions (10% B) were re-established and column was allowed to re-equilibrate for 3 min.

The MS detection parameters were set as follows: ESI negative ionization mode; interface voltage of 3.5 kV; nebulizer gas flow at 3 L/min; heating gas flow at 10 L/min; heat block temperature at 400 °C; CDL temperature at 250 °C; detector voltage at 1.70 kV; TOF tube temperature at 42 °C. Sodium iodide (NaI) was used as a mass calibration solution to ensure acquisition of high accurate masses. The m/z range of 100–1000 was used for both high-resolution MS and tandem MS (MS/MS) experiments. For MS/MS experiments, argon gas was used as collision gas, and MS^E mode using collision energy ramp of 15 to 25 eV was used to generate possible fragments.

3.3. Results and Discussion

In the current study, two putative HTT genes from *B. pilosa* were identified and the metabolites which are possibly products of phenylpropanoid pathway facilitated by two hydroxycinnamoyl-CoA tartaric hydroxycinnamoyl transferase genes were also identified herein. This study initially identified five potential HTT gene from *B. pilosa*, and after filtering process to check the quality of the sequences, only two were identified as potential HTT in *B. pilosa*. The filtering process involved percentage matrix index to eliminate all potential HTT gene within *B. pilosa* plant (Fig. S5). ExPASy was used to check sequence integrity by assessing if identified sequences will have an ORF that can be used to design primers. The ORF identified interestingly possesses both conserved regions DFGWG and HXXXD respectively (Figure S1 & S2). Sequences aligned share the same conserved motifs (DFGWG and HXXXG) highlighted with a red colour (Figure 3.1). The only conserved residues on the HXXXG motif are histidine and glycine, this then means variations are expected in the -XXX- region. Herein, the identified HTT genes contains HRVLD (HTT1) and HRVAD (HTT2) respectively. Literature has reported on different HXXXD motifs such as HYVVD, HVVAD, HVMCD, HRVVD and HKIAD in *Aster spathulifolius* (Park *et al.*, 2021). The MSA shows that this motif is different from one plant to another, as it is the case in *B. pilosa* (Figure 3.1 & Figure S1). These genes fall under the BAHD acyltransferase family of genes because of the conserved motifs that they possess (Pierre and De luca, 2000). HTT gene from *Echinacea purpurea* and *Helianthus annuus* showed a very high similarity to the two genes isolated from *B. Pilosa* in this study. Interestingly, all these three plants (*B. pilosa*, *Echinacea purpurea* and *Helianthus annuus*) are from the *Asteraceae* family, this gives confidence to the two putative hydroxycinnamoyl-CoA tartaric hydroxycinnamoyl transferase identified in this study because these genes are showing high similarity to same genes from plants of the same family. Homology of HTT genes from *Asteraceae* was further established using phylogeny and the two hydroxycinnamoyl-CoA tartaric hydroxycinnamoyl transferase identified showed a common evolutionary history by 100% (Figure 3.2). The bootstrap values show close similarity between the putative genes in *B. pilosa*, *H. annuus* and *E. purpurea*. Hydroxycinnamoyl-CoA tartaric hydroxycinnamoyl transferase 1 & 2 (HTT1 and HTT2) from *B. pilosa* is in the same clade as *E. purpurea* and the bootstrap value

is 87 which validates these sequences to be very similar that the split between was not significant. *B. pilosa* (HTT1 & HTT2) and *E. purpurea* (HTT) shares the same clade as *H. annuus* (sunflower), and the bootstrap value was found to be 59%. This shows that *B pilosa* HTT genes are more closely related to *E. purpurea* than *H. annuus* HTT gene. *B. pilosa* produce a wide variety of structurally diverse HCA compounds (Ramabulana *et al.*, 2021). Enzymes responsible for biosynthesis of HCAs such as chlorogenic acids (CGA) have been identified in other plants such as *Coffea canephora* (Koshiro *et al.*, 2007; Lepelley *et al.*, 2007). Overexpression of hydroxycinnamoyl-CoA quinate hydroxycinnamoyl transferase (HQT) gene in *Solanum lycopersicum* (tomato) increased the CGA content (Niggeweg *et al.*, 2004), this phenomenon was also observed in *taraxacum* (Qun *et al.*, 2018). The principal pathway for CGAs biosynthesis involves the synthesis *via trans*-esterification of the caffeoyl-CoA and quinic acid catalysed by the hydroxycinnamoyl-CoA quinate hydroxycinnamoyl transferase (HQT) enzyme. Interestingly, genes coding for HQT were shown to differ from one plant to another and their existence can be positively correlated with CGA content in various plants, although certain plants such as *Arabidopsis thaliana* are known to have HQT gene but are unable to produce CGAs (Lepelley *et al.*, 2007).

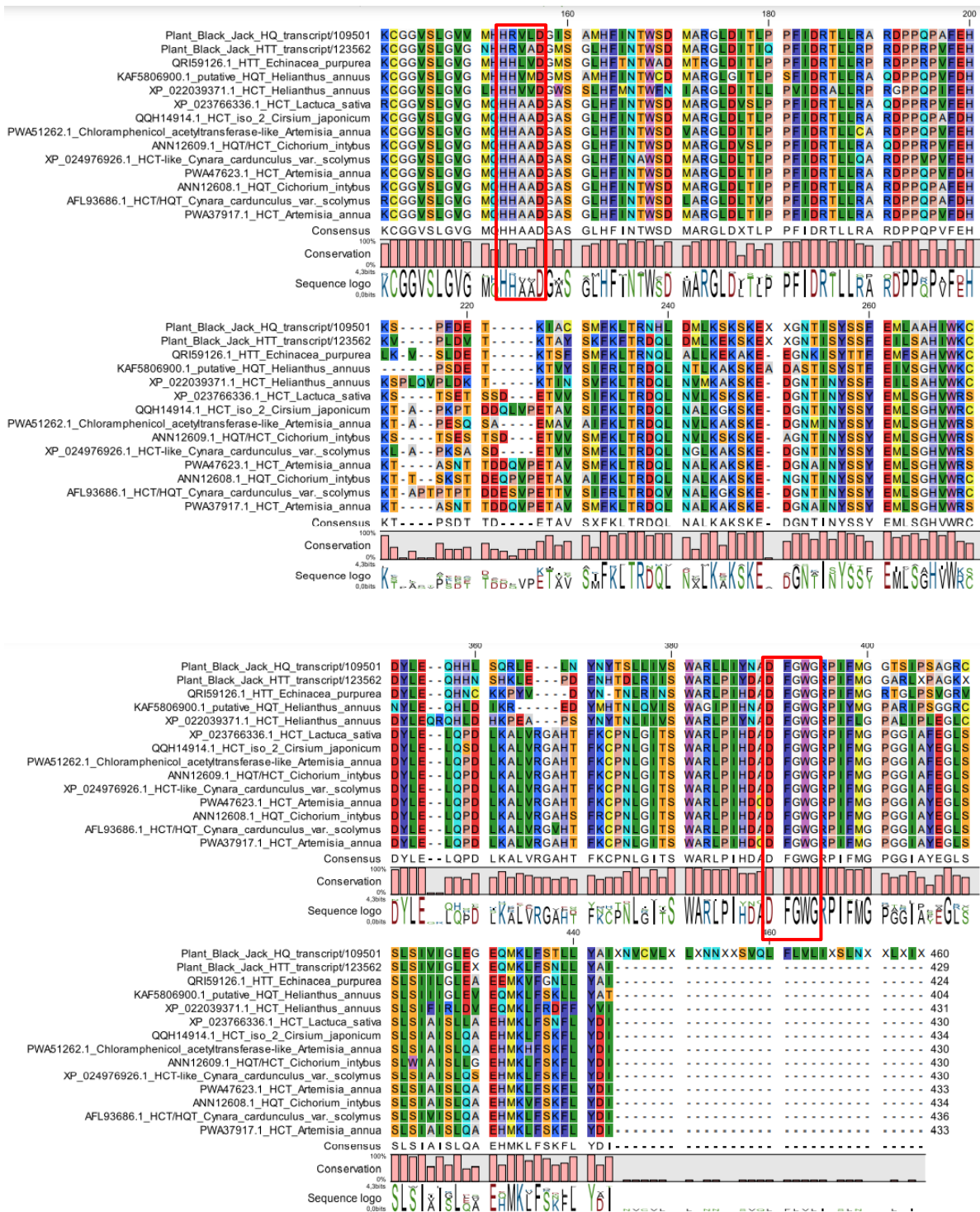


Figure 3.1 HTT gene sequences from different *Asteraceae* family plants showing emphasis to the conserved motifs, (A) DFGWG and (B) HXXXD highlighted in red. The first two sequences (*Plant_Black_Jack_HQ_transcript/109501* and *Plant_Black_Jack_HQ_transcript/123562*) are HTT1 and HTT2 genes respectively from *B. pilosa* obtained from Pacbio sequencing, and the other sequences were retrieved from NCBI (shown by NCBI accession numbers).

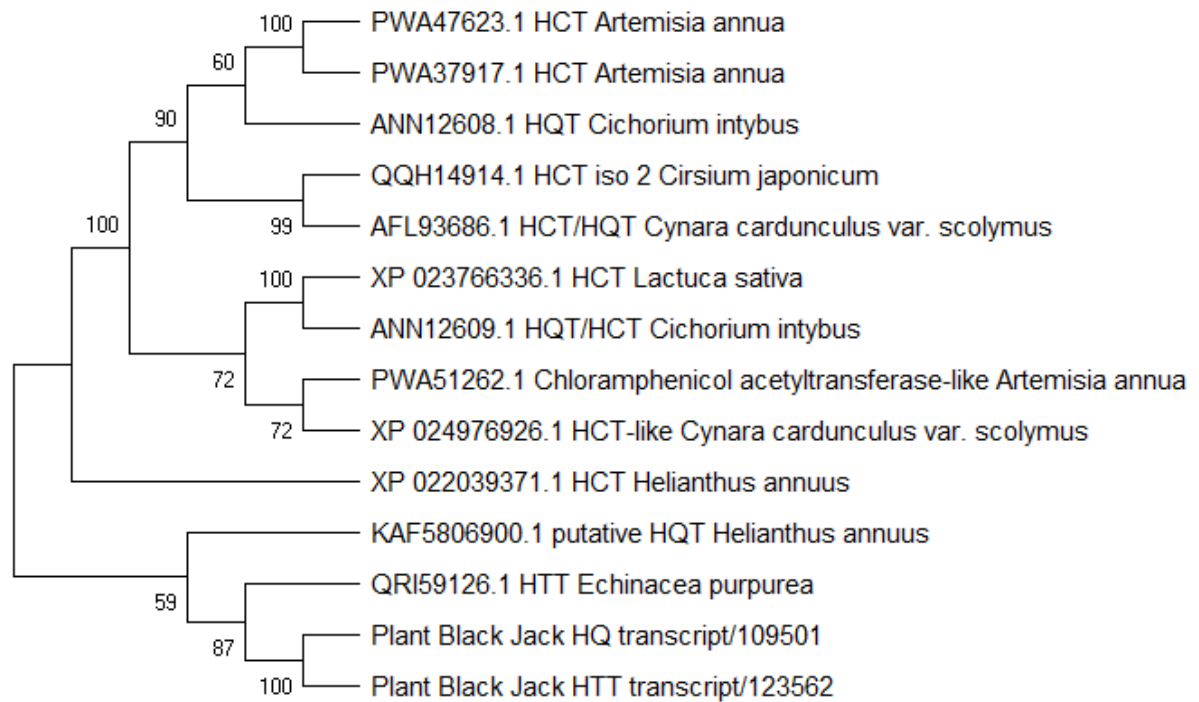


Figure 3.2 Neighbour joining phylogenetic analysis of HTT genes from *B. pilosa* and other HTT genes from *Asteraceae* family were retrieved from NCBI database. The evolutionary history was inferred using the Neighbor-Joining method. The optimal tree with the sum of branch length = 4,11446114 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site. This analysis involved 15 amino acid sequences. Evolutionary analyses were conducted in MEGA X.

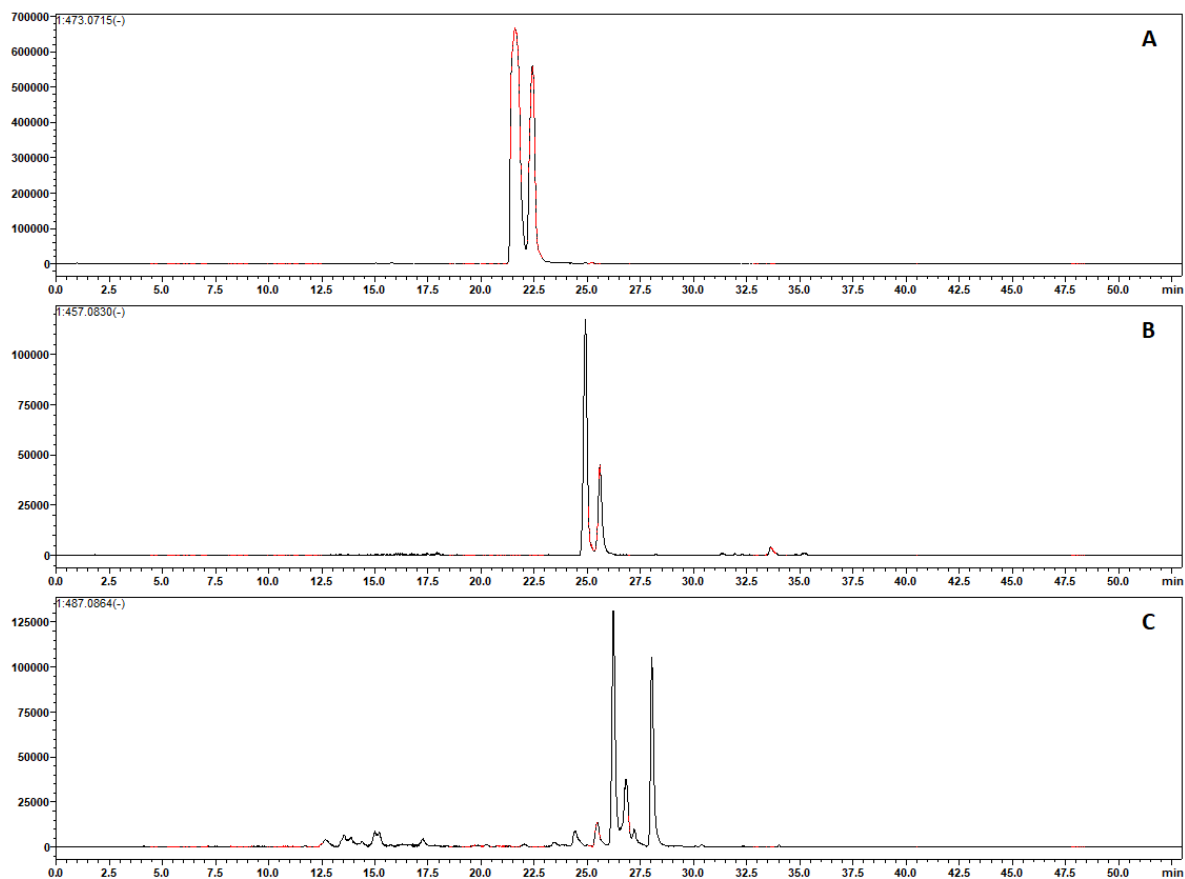


Figure 3.3 Representative UHPLC-qTOF-MS/MS chromatogram showing distribution patterns of tartaric acid derivatives in *B. pilosa*, with Y-axis showing peak intensity and X-axis showing retention time. The distribution pattern is as per the masses, at m/z 473 (A), at m/z 457 (B) and at m/z 487 (C).

Recent studies have reported the use of LC-MS in characterisation of diverse hydroxycinnamic acid tartaric acid esters in *B. pilosa* such as caftaric acid, chicoric acid, coutaric acid, fertaric acid, and caftaric acid glycoside (Khoza *et al.*, 2016; Nobela *et al.*, 2018; Ramabulana *et al.*, 2020).

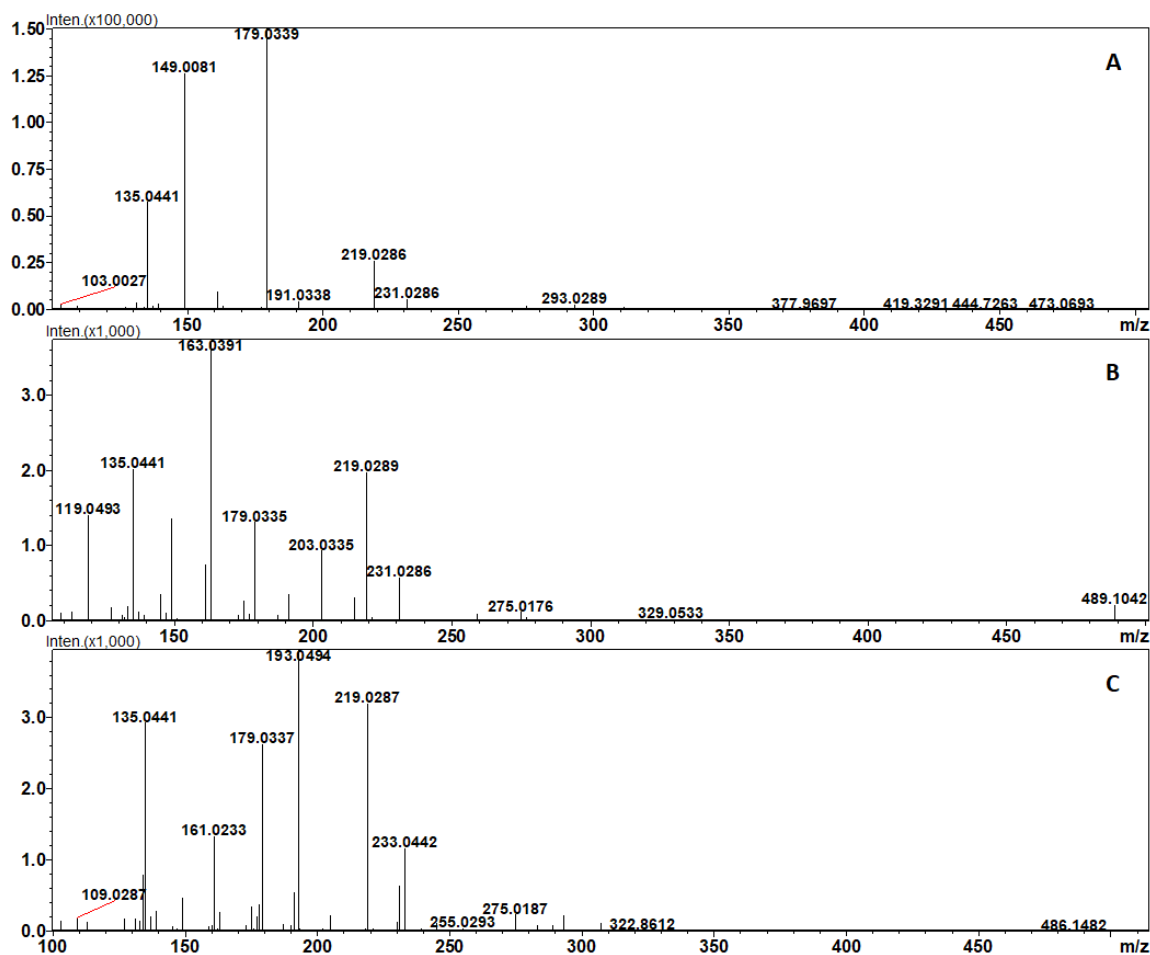


Figure 3.4 Typical mass spectra of the fragmentation patterns of dicaffeoyltartaric acid (A), p-coumaroyl-caffeoyl tartaric acid (B), feruloyl caffeoyl tartaric acid (C).

Metabolite 1 eluted at the retention time of 21.266 min with a parent ion at m/z 473. Fragmentation of this metabolite produced daughter ions at m/z 311, at m/z 219, at m/z 149, at m/z 135 (Figure 3.4A). Metabolite 1 was then identified as chicoric acid (CA) by its molecular ion $[M-H]^-$, at m/z 473 which fragmented to give product ions at m/z 311 $[CTA-H]^-$ due to the loss of a second caffeic acid residue. Other fragmented ions were produced at m/z 149 $[TA-H]^-$ and m/z 135 $[CFA-CO_2]$. Figure 3.5A shows the structure of this metabolite. Due to two caffeoyl acid attached to tartaric acid; the

structure show that it is homogeneous. This fragmentation pattern has also been noted elsewhere (Buiarelli *et al.*, 2010; Chen *et al.*, 2012; Ramabulana *et al.*, 2020). CA fragmented and produced daughter ions at m/z 311 (caftaric acid) due to the loss of 162Da which is the mass representing hexose. It also produced product ions at m/z 179 due to the loss of tartaric acid residue and other daughter ions were produced at m/z 149 (Caffeic acid) and m/z 135 due to the decarboxylation of caffeic acid residue. Chicoric acid and caffeic acid are biological active compounds which have been reported to have various nutraceutical benefits such as being an antioxidant, anti-cancer and many other benefits (Sunetha *et al.*, 2013). Importantly, recent studies demonstrated that hydroxycinnamic acid and tartaric acid esters i.e., chicoric acid, a chlorogenic acid derivative inhibits HIV-1 integrase enzyme at low concentrations (Healy *et al.*, 2009). Chlorogenic acids have also shown anti-viral activity against influenza virus (H1N1) (Ding *et al.*, 2017). This highlights the need to further explore the metabolome of *B. pilosa* to characterise other HCA derivatives that could have potential biological activities. These molecules are hydroxycinnamic acid derivatives and are found in plants like *E. purpurea* (Lee & Scagel., 2010), *C. intybus* (Jaiswal *et al.*, 2011), *Taraxacum antungense* (Chen *et al.*, 2012), *Lactuca sativa var. capitata* (Schütz *et al.*, 2004) and have also been positively identified in *B. pilosa* (Ramabulana *et al.*, 2020). These molecules have also been identified in wine to investigate their variations during different wine ageing stages (Buiarelli *et al.*, 2010).

Metabolite 2 eluted at retention time of 21.266 with a parent ion at m/z 457 [M-H]⁻ as shown in figure 3.4B above. Fragmentation pattern of this molecule produced daughter ions at m/z 135, at m/z 149, at m/z 179 and at m/z 311. The fragmentation pattern of this molecule shows tartaric and caffeic acid peaks and a peak that shows decarboxylation of caffeic acid which is caftaric acid. The parent ion of this molecule is [M-H]⁻ – at m/z 457 and as stated above it showed to have caffeic, coumaric and tartaric acid moiety. This molecule gave product ions at m/z 295 (coutaric acid) due to loss of caffeic acid, it further fragmented to produce coumaric acid at m/z 163 and its decarboxylated ion at m/z 119. Two signals of low abundance at m/z 311 and at m/z 179 proved the presence of a caffeoyltartaric acid. Due to the fragmentation pattern displayed above, this metabolite was identified to be *p*-coumaroyl caffeoyl tartaric acid and is represented by its structure in figure 3.6B. The structure in figure 3.6B shows

that *p*-coumaroyl caffeoyl tartaric acid is a heterogenous compound since it contains caffeoyl and coumaroyl. Fragmentation pattern of this molecule has also been noted elsewhere in *Taraxacum officinale* (Schütz *et al.*, 2004).

Metabolite 3 was eluted at a retention time of 21.266 with a parent ion at *m/z* 487. The fragmentation of the parent ion produced fragments at *m/z* 135, at *m/z* 179, at *m/z* 193, at *m/z* 161, at *m/z* 219, at *m/z* 149 as shown in figure 3.4C. Fragmentation pattern of this molecule showed at *m/z* 487 as the parent ion, which fragmented to give a peak at *m/z* 293 (caffeoyl tartaric acid), at *m/z* 193 (mono feruloyl tartaric acid), at *m/z* 179 (caffeic acid), at *m/z* 135 (caffeic acid). Based on this fragmentation pattern this molecule was identified as caffeoyl-feruloyl tartaric acid (Figure 3.5C). This fragmentation pattern has been noted also in *C. intybus* (Buiarelli *et al.*, 2010). Based on literature, plants that commonly produce HHT enzymes are also known to accumulate the tartaric acid esters, especially chicoric acid (Fu *et al.*, 2021). and, as such, identification of these acyl transferases genes in *B. pilosa* could be correlated to the existence of the tartaric acid esters of HCA reported herein (Figure 3.5), and those reported previously (Khoza *et al.*, 2016). More importantly, expression studies through recombinant DNA technology could also be carried out in order to establish the expression of these enzymes with the accumulation of these esters.

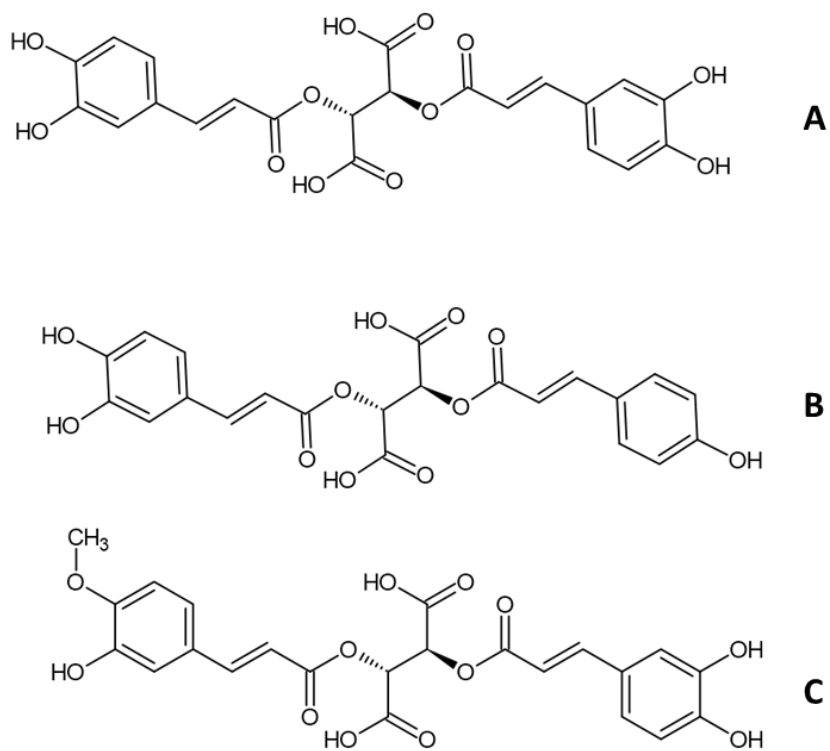


Figure 3.5 Chemical structures of HTT derivatives from *B. pilosa*. (A) dicaffeoyltartaric acid, (B) *p*-coumaroyl caffeoyl tartaric acid, (C) Feruloyl-caffeoyl tartaric acid.

3.4. Conclusion

In this study, *B. pilosa* was shown to synthesize hydroxycinnamoyl tartaric acid which are HCA derivatives and known to have many health benefits such as antioxidant, anti-cancer (Sunetha *et al.*, 2013) and many other benefits. These metabolites have been reported in other plants, but herein their existence in *B. pilosa* is reported for the first time. Hitherto, the enzymatic machinery responsible for the production of wide variety of HCA metabolites derivatives in *B. pilosa* is not known. As such the current transcriptomic data, through phylogenetic analysis (through comparison with other known genes) revealed genes that play a role in the phenylpropanoid pathway for production of HCA-tartarate derivatives in *B. pilosa*. Furthermore, the results of the current study highlight the two identified HTT genes of which their enzyme products are possible catalysts in biosynthesis of HCA derivatives which can be further exploited by incorporating them in other economically viable plants in order to enhance the nutraceutical values thereof.

References

- Bartolome, A. P., Villaseñor, I. M., Yang, W.C., 2013. *Bidens pilosa* L. (asteraceae): botanical properties, traditional uses, phytochemistry, and pharmacology. *Evidence-based complementary and alternative medicine*. <https://doi.org/10.1155/2013/340215>
- Bartwal, A., Mall, R., Lohani, P., Guru, S., Arora, S., 2013. Role of secondary metabolites and brassino steroids in plant defense against environmental stresses. *Journal of plant growth regulation* 32, 216–232. <https://doi.org/10.1007/s00344-012-9272-x>
- Bashir, A., Singh, C., Chauhan, N., Rani, A., 2018. A review: ethnobotanical study on medicinal plants of kargil district, ladakh, india. *Journal of Emerging Technologies and Innovative Research* 5, 181–196.
- Buiarelli, F., Coccioli, F., Merolle, M., Jasionowska, R., Terracciano, A., 2010. Analytical Methods Identification of hydroxycinnamic acid–tartaric acid esters in wine by HPLC–tandem mass spectrometry. *Food chemistry* 123, 827–833. <https://doi.org/10.1016/j.foodchem.2010.05.017>
- Chen, H., Inbaraj B., Chen, B., 2012. Determination of Phenolic Acids and Flavonoids in *Taraxacum formosanum Kitam* by Liquid Chromatography-Tandem Mass Spectrometry Coupled with a Post-Column Derivatization Technique. *International Journal of Molecular Sciences* 13, 260–285. <https://doi:10.3390/ijms13010260>
- Cui, J., Lu, Z., Xu, G., Wang, Y., Jin, B., 2020. Analysis and comprehensive comparison of PacBio and nanopore-based RNA sequencing of the *Arabidopsis* transcriptome. *Plant Methods* 16, 1–13. <https://doi.org/10.1186/s13007-020-00629-x>
- Ding, Y., Cao Z., Cao L., Ging G., Wang, Z., Xiao, W., 2017. Antiviral activity of chlorogenic acid against influenza a (h1n1/h3n2) virus and its inhibition of neuraminidase. *Scientific reports* 7, 1–11. <https://doi.org/10.1038/srep45723>

Feugang, J. M., Konarski, P., Zou, D., Stintzing, F. C., Zou, C., 2006. Nutritional and medicinal use of cactus pear (*Opuntia spp.*) cladodes and fruits. *Frontiers in Biosciences* 11, 2574–2589. <https://doi.org/10.2741/1992>

Fu, R., Zhang, P., Jin, G., Wang, L., Qi, S., Cao, Y., Martin, C. and Zhang, Y., 2021. Versatility in acyltransferase activity completes chicoric acid biosynthesis in purple coneflower. *Nature communications* 12(1), 1-13. <https://doi.org/10.1038/s41467-021-21853-6>

Garrido, J., Borges, F., 2013. Wine and grape polyphenols—A chemical perspective. *Food research international* 54, 1844–1858. <https://doi.org/10.1016/j.foodres.2013.08.002>

Gökçen, B.B., Sanlier, N., 2019. Coffee consumption and disease correlations. *Critical reviews in food science and nutrition* 59, 336–348. <https://doi.org/10.1080/10408398.2017.1369391>

Healy, E.F., Sanders, J., King, P.J., Robinson Jr, W.E., 2009. A docking study of L-chicoric acid with HIV-1 integrase. *Journal of Molecular Graphics and Modelling* 27, 584-589. <https://doi.org/10.1016/j.jmglm.2008.09.011>

Jaiswal, R., Kiprotich, J., Kuhnert, N., 2011. Determination of the hydroxycinnamate profile of 12 members of the *Asteraceae* family. *Phytochemistry* 72, 781 – 790. <https://doi.org/10.1016/j.phytochem.2011.02.027>

Kang, N. J., Lee, K. W., Shin, B. J., Jung, S. K., Hwang, M. K., Bode, A. M., Heo, Y.S., Lee, H. J., Dong, Z., 2009. Caffeic acid, a phenolic phytochemical in coffee, directly inhibits Fyn kinase activity and UVB-induced COX-2 expression. *Carcinogenesis* 30, 321-330. <https://doi.org/10.1093/carcin/bgn282>

Khoza, B.S., Gbashi, S., Steenkamp, P.A., Njobeh, P.B., Madala, N.E., 2016. Identification of hydroxycinnamoyl tartaric acid esters in *Bidens pilosa* by UPLC-tandem mass spectrometry. *South African Journal of Botany* 103, 95-100. <https://doi.org/10.1016/j.sajb.2015.08.018>

- Kim, Y.B., Thwe, A.A., Kim, Y.J., Li, X., Kim, H.H., Park, P.B., Suzuki, T., Kim, S.J., Park, S.U., 2013. Characterization of genes for a putative hydroxycinnamoyl-coenzyme A quinate transferase and *p*-coumarate 3'-hydroxylase and chlorogenic acid accumulation in tartary buckwheat. *Journal of agricultural and food chemistry* 61, 4120-4126. <https://doi.org/10.1021/jf4000659>
- Koshiro, Y., Jackson, M. C., Katahira, R., Wang, M. L., Nagai, C., Ashihara, H., 2007. Biosynthesis of chlorogenic acids in growing and ripening fruits of *coffea arabica* and *coffea canephora* plants. *Zeitschrift für naturforschung c* 62, 731-742. <https://doi.org/10.1515/znc-2007-9-1017>
- Lai, B. Y., Chen, T. Y., Huangetal, S. H., 2015. *Bidens Pilosa* formulation improves blood homeostasis and β -cell function in men: A Pilot Study. *Evidence-Based Complementary and Alternative Medicine* 5. <https://doi.org/10.1155/2015/832314>
- Lee, J., Scagel, C., 2010. Chicoric acid levels in commercial basil (*Ocimum basilicum*) and *Echinacea purpurea* products. *Journal of functional foods* 2, 77–84. <https://doi.org/10.1016/j.jff.2009.11.004>
- Lee, J., Scagel, C.F., 2013. Chicoric acid: chemistry, distribution, and production. *Frontiers in chemistry* 1, 40. <https://doi.org/10.3389/fchem.2013.00040>
- Lepelley, M., Cheminade, G., Tremillon, N., Simkin, A., Caillet, V., McCarthy, J., 2007. Chlorogenic acid synthesis in coffee: an analysis of CGA content and real-time RT-PCR expression of HCT, HQT, C3H1, and CCoAOMT1 genes during grain development in *C. canephora*. *Plant Science* 172, 978–996. <https://doi.org/10.1016/j.plantsci.2007.02.004>
- Li, L., Su, C., Chen, X., Wang, Q., Jiao, W., Luo, H., Tang, J., Wang, W., Li, S., Guo, S., 2020. Chlorogenic acids in cardiovascular disease: a review of dietary consumption, pharmacology, and pharmacokinetics. *Journal of agricultural and food chemistry* 68, 6464–6484. <https://doi.org/10.1021/acs.jafc.0c01554>

Lu, M., An, H., Li, L., 2016. Genome survey sequencing for the characterization of the genetic background of *Rosa roxburghii* Tratt and leaf ascorbate metabolism genes. *PLoS One*, 11(2): e0147530. <https://doi.org/10.1371/journal.pone.0147530>

Malar, C, M., Yuzon, J.D., Panda, A., Kasuga, T., Tripathy, S., 2019. Updated assembly of *Phytophthora ramorum* pr102 isolate incorporating long reads from PacBio sequencing. *Molecular Plant-Microbe Interactions* 32, 1472–1474. <https://doi.org/10.1094/MPMI-05-19-0147-A>

Mao, D.J., Xie J.F., Quan, G.M., Zhang, J., 2010. Effects of *Bidens pilosa* aqueous extracts on germination and seedling growth of two pastures. *Pharmacology* 1, 34–87. [https://doi.org/10.6165/tai.2009.54\(3\).255](https://doi.org/10.6165/tai.2009.54(3).255)

Masike, K., Mhlongo, M.I., Mudau, S.P., Nobela, O., Ncube, E.N., Tugizimana, F., George, M.J., Madala, N.E., 2017. Highlighting mass spectrometric fragmentation differences and similarities between hydroxycinnamoyl-quinic acids and hydroxycinnamoyl-isocitric acids. *Chemistry Central Journal* 11, 1–7. <https://doi.org/10.1186/s13065-017-0262-8>

Mhlongo, M.I., Steenkamp, P.A., Piater, L.A., Madala, N.E., Dubery, I.A., 2016. Profiling of altered metabolomic states in *Nicotiana tabacum* cells induced by priming agents. *Frontiers in plant science* 7, 1527. <https://doi.org/10.3389/fpls.2016.01527>

Mudau, S.P., Steenkamp, P.A., Piater, L.A., De Palma, M., Tucci, M., Madala, N.E., Dubery, I.A., 2018. Metabolomics-guided investigations of unintended effects of the expression of the hydroxycinnamoyl quinate hydroxycinnamoyltransferase (hqt1) gene from *Cynara cardunculus* var. *scolymus* in *Nicotiana tabacum* cell cultures. *Plant Physiology and Biochemistry* 127, 287-298. <https://doi.org/10.1016/j.plaphy.2018.04.005>

Naveed, M., Hejazi, V., Abbas, M., Kamboh, A. A., Khan, G. J., Shumzaid, M., Ahmad, F., Babazadeh, D., Fangfang, X., Modarresi-ghazani, F., 2018. Chlorogenic acid (CGA): a pharmacological review and call for further research. *Biomedicine & pharmacotherapy* 97, 67–74. <https://doi.org/10.1016/j.biopha.2017.10.064>

Niggeweg, R., Michael, A.J., Martin, C., 2004. Engineering plants with increased levels of the antioxidant chlorogenic acid. *Nature biotechnology* 22, 746–754. <https://doi.org/10.1038/nbt966>

Pal, T., Malhotra, N., Chanumolu, S.K., Chauhan R. S., 2015. Next-generation sequencing (NGS) transcriptomes reveal association of multiple genes and pathways contributing to secondary metabolites accumulation in tuberous roots of *Aconitum heterophyllum* Wall. *Planta* 242, 239–258. <https://doi.org/10.1007/s00425-015-2304-6>

Park, S., Sivagami, J.C., Park, S., 2021. Transcriptome-wide identification and quantification of Caffeoylquinic acid biosynthesis pathway and prediction of their putative BAHDs gene complex in *A. spathulifolius* 22, 633. <https://doi:10.20944/preprints202105.0527.v1>

Qun, Liu., Yue, Liu., Yachen, Xu., Lixiang, Yao., Zijia, Liu., Haitao, Cheng., Ming, Ma., Jie, Wu., Weiting, Wang., Wei, Ning., 2018. Overexpression of and RNA interference with hydroxycinnamoyl-CoA quinate hydroxycinnamoyl transferase affect the chlorogenic acid metabolic pathway and enhance salt tolerance in *taraxacum antungense* kitag. *Phytochemistry letters* 28, 116–123. <https://doi.org/10.1016/j.phytol.2018.10.003>

Ramabulana, A., Steenkamp, P., Madala, N., Dubery, A., 2020. Profiling of Chlorogenic Acids from *Bidens pilosa* and Differentiation of Closely Related Positional isomers with the Aid of UHPLC-QTOF-MS/MS-Based In-Source Collision-Induced Dissociation. *Metabolites* 10, 178. <https://doi:10.3390/metabo10050178>

Rhoads, A., Au, K., 2015. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics* 13, 278–289. <https://doi.org/10.1016/j.gpb.2015.08.002>

Scalbert, A., Manach, C., Morand, C., Rémésy, C., Jiménez, L., 2005. Dietary polyphenols and the prevention of diseases. *Critical reviews in food science and nutrition* 45, 287–306. <https://doi.org/10.1080/1040869059096>

Schütz, K., Kammerer, D. R., Carle, R., Schieber, A., 2004. Characterization of phenolic acids and flavonoids in dandelion (*taraxacum officinale* WEB. Ex WIGG.) Root and herb by high-performance liquid chromatography/electrospray ionization mass spectrometry. *Rapid Communications in Mass Spectrometry* 19, 179–186. <https://doi.org/10.1002/rcm.1767>

Siddiqui, Z. H., Hareramdas, B., Abbas, Z. K., Parween, T., Khan, M. N., 2018. Use of plant secondary metabolites as nutraceuticals for treatment and management of cancer: approaches and challenges. *Anticancer plants: properties and application*. Springer. https://doi.org/10.1007/978-981-10-8548-2_17

Sonnante, G., D'Amore, R., Blanco, E., Pierri, C.L., DePalma, M., Luo, J., Tucci, M., Martin, C., 2010. Novel hydroxycinnamoyl-coenzyme A quinate transferase genes from artichoke are involved in the synthesis of chlorogenic acid. *Plant Physiology* 53, 1224–1238. <https://doi.org/10.1104/pp.109.150144>

Tiwari, R., Rana, C., 2015. Plant secondary metabolites: a review. *International journal of engineering research and general science* 3, 661–670.

Vashisht, I., Mishra, P., Pal, T., Chanumolu, S., Singh, T.R., Chauhan, R.S., 2015. Mining NGS transcriptomes for miRNAs and dissecting their role in regulating growth, development, and secondary metabolites production in different organs of a medicinal herb, *Picrorhiza kurroa*. *Planta* 241, 1255–1268. <https://doi.org/10.1007/s00425-015-2255-y>

Xiang, B., Li, X., Qian, J., Wang, L., Ma, L., Tian, X., Wang, Y., 2016. The complete chloroplast genome sequence of the medicinal plant *Swertia mussoitii* using the PacBio RS II platform. *Molecules* 21, 1029. <https://doi.org/10.3390/molecules21081029>

Xuan, T. D., Khanh, T. D., 2016. Chemistry and pharmacology of *bidens pilosa*: an overview. *Journal of pharmaceutical investigation* 46, 91–132. <https://doi.org/10.1007/s40005-016-0231-6>

Zheng, X., Renslow, R.S., Makola, M.M., Webb, I.K., Deng, L., Thomas, D.G., Govind, N., Ibrahim, Y.M., Kabanda, M.M., Dubery, I.A., Heyman, H.M., 2017. Structural elucidation of *cis/trans* dicaffeoylquinic acid photoisomerization using ion mobility spectrometry-mass spectrometry. *The journal of physical chemistry letters* 8, 1381–1388. <https://doi.org/10.1021/acs.jpcllett.6b03015>

Supplementary Data

>Plant_Black_Jack_HQ_transcript/109501

GATATAAACCCAATATAAAAACATCTAATATATTCCAAAAGAATAAAGAAAAATCTTTTGTGACACC
TACATGGATTTTTAGACATATAATTCTCACGCGATCATTCTTCACTCTGTGTACGCATCCAATTAT
AGAGTTTCAGAAGACCATTGGTTAAAGTAGCTGTGAAGATCAAGATGAAGGTGGTCGTAAGAGAA
TCGACTATGGTGAGGCCAGCCGAGGAGACACCAACAACAAGCTTTGGATGTCTAGCCTTGATC
TAACTGCCTTCAACAACACTACACACAACAGTGTATTATTACCGGCACAACAGTGCTCCCAACTTCT
TTGACATAAAAAGTTATGAAGGACACTTTGAGCAGGGCGTTGGCTGCGTTTTACCCAATGGCAGG
GCGTTTTAGAAATGACGAAGATGGCCGATTGAAATCGATTGTCAAGGACAAGGGGCGTTGTTT
TTGGAAGCTGAGTCGGATGGCGTTATCGATGACTTCGGTGACTTTGAACCCACACCGGAATACTT
GAAACTCGTTCCAGTGATCGATTACTCTCTGGGAATTGAATCACTTCCTCTCCTAATGTTACAGGT
AACTTGCTTCAAATGCGGTGGGGTTTTCGCTAGGAGTTGTGATGCACCATCGTGTCTGGATGGG
ATATCTGCAATGCACTTCATCAACACATGGTCCGATATGGCGCGTGGCCTTGACATCACTCTTCC
ACCCTTCATAGACAGGACCCTCCTCCGTGCAAGGGACCCCCACAACCAGCCTTTGAGCACATC
GAATACCACTCAGATCCAAAAGTGAAGTCGCCCTTTGATGAAACCAAAATCGCTTGCTCAATGTT
TAAGTTAACACGAAACCATCTCGATATGCTCAAATCTAAATCAAAGGAAGATGGGAACACGATCA
GCTACAGCTCTTTTGAATGTTGGCGGCTCATATTTGGAAGTGCCTGTGTAAAGCTCGTGGGTTG
CCAGAAAACGTGAAACCAGTTTCGACTGCCAATCGATGGGCGGGCTCGCTTTGAACCACCAC
TTCCACCAGGCATTTTCGAAATGCTATCTTCAGAACCCTACTACAGCAACAGCAGGTGACATT
CAATCCAAACCTTTGTGGTACGTTGCAAGTAAATACACGATGCTGTAGCGAGGATGAATAATGA
CTATCTAAAATCATCACTCGATTATCTGGAACAACACCATCTTTCTCAAAGGCTCGAGCTTAATTA
TAACTACACGAGTCTTCTAATTGTAAGTTGGGCTAGGCTCCTGATTTACAATGCAGACTTTGGGT
GGGGTCGGCAATTTTCATGGGGGGTACGAGCATCCCATCCGCTGGTAGGTGTTATGTGTTACC
AAGCCCGGAAAACGATGGGAGCTTATCAATCGTCATTGGACTAGAAGGTGAACAAATGAAGCTAT
TCAGTACCTTGTTGTATGCAATCTGAAATGTATGTGTACTTTAATTGTAATAATAATCGGTACAAC
CTTTCTCGTATTGATCGAGTTTAAATAAACTTTGAATTGC

>|c||ORF

MKVVVRESTMVRPAEETPTTKLWMSSLDLTA FNNTQTVYYRHNSAPNFFDIKVMKDTLSRALAAF
YPMAGRFRNDEDGRIEIDCQQGALFLEAESDGVDDFGDFEPTPEYLKLVVIDYSLGIESLPLMLQ
VTCFKCGVSLGVVMH**HRVLD**GISAMHFINTWSDMARGLDITLPPFIDRTLLRARDPPQPAFEHIEYH
SDPKVKSPFDETKIACSMFKLTRNHLMLKSKSKEDGNTISYSSFEMLAHIWKCVCKARGLPENVET
RFDCPIDGRARFEPPLPPGIFGNAIFRTTTTATAGDIQSKPLWYVASKIHDAVARMNNDYLKSSLDYLE
QHLSQRLELNINYTSLLIVSWARLLIYNA**DFGWR**PIFMGGTSIPSAGRCYVLPSPENDGSLIVIGL
EGEQMKLFSTLLYAI



Figure S1: HTT1 nucleotide sequence from *B. pilosa* as obtained from the SMRT sequencing technique and the amino acid sequence of the ORF.

>Plant_Black_Jack_HQ_transcript/123562

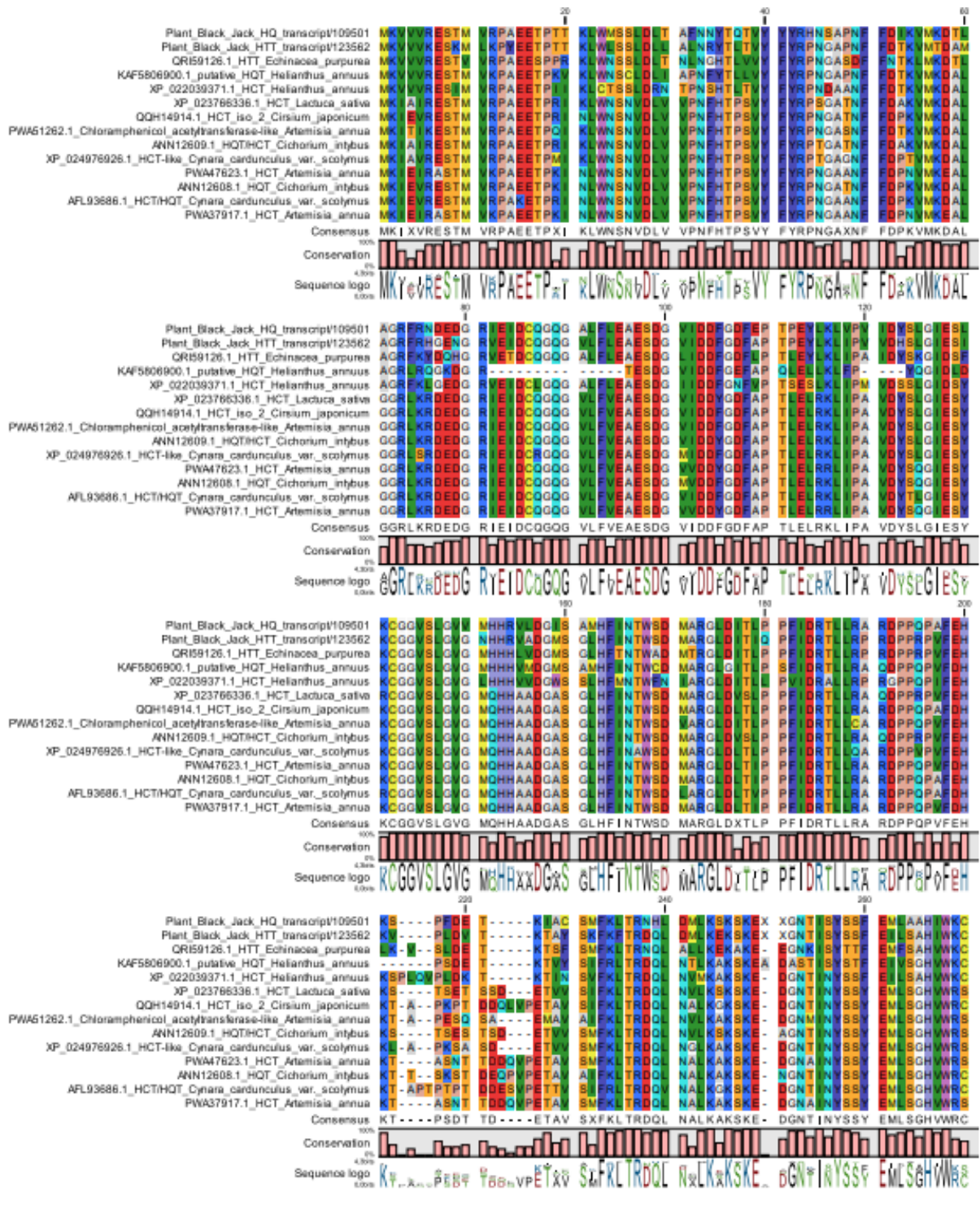
GTAAAGTAGAGAAGATAAGATGAAGGTGGTTCGTAAGAAGATCAAAGATGTTGAAGCCATATGAGG
 AACACCAACGACAAAGCTATGGCTCTCTAGCCTCGATCTCCTTGCCTTGAACAGATACACACTA
 ACAGTGTACTTTTTACCGGCCAACGGTCTCCCACTTCTTTGACACGAAAGTTATGACGGACGC
 TATGAGCAGGGCGTTAGTAGCGTTTTACCCAATGGCAGGCCGATTTAGACATGGTGAAGATGGA
 AGGGTTGAGATCGATTGTCAAGGACAAGGGGTGTTGTTTTTGAAGCCGAGTCAGATGGCGTTA
 TCGATGATTTCCGGTGACTTTGCACCCACACCAGAATACTTGAAACTCATTCCGGTGGTGCATCAC
 TCTTTGGGAATTGAATCAATTCCTTTCTGGTGTTCAGGTCCTGTTTCAAATGTGGCGGAGT
 TTCGCTAGGAGTTGGGAACCACCATCGTGTGGCGGATGGGATGTCTGGGTTGCACTTCATCAAC
 ACATGGTCCGATATGGCTCGTGGCCTTGACATCACGATCCAACCCTTCATAGACCGGACACTCC
 TCCGTCCGCGAGACCCACCACGACCAGTTTTGAACACATCGAATACCACCCATATCAAACAGTG
 AAGGTGCCCTTAGATGTAACGAAAACCGCCTACTCAAAGTTCAAGTTTACCCGAGACCAACTCGA
 TATGCTCAAAGAAAATCAAAGGAAGACGGGAACACGATCAGCTACAGCTCTTTTGAATTTTGT
 CAGCCCATATTTGAAATGCATGTGTATAGCAGCTCGTGTATCAGCTGCCAGAAAACGCTGAGAC
 CATGTTCCGACTGCCAATCGATGGGCGGGCTCGCTTTGAACCACCACTTCCGCGAGGCTTTTTTC
 GGAATGTTATCTTCAGAGCCACTACTTCAGCAACTGCAGGCGAAATTCAAACCAAACCTTTGTG
 GTACATTGCAAGTAAAATACATGATGCTTTAGCGAGGATGAATAGCGACTATCTAAGATCATCACT
 CGATTATCTGGAACAACACCATAATTCTCATAAGCTCGAGCCTGATTTTAAACCACCGGATCTTCG
 AATAATAAGTTGGGCTAGGCTCCCGATTTACGACGACGACTTTGGGTGGGGTCCGGCCCATCTTC
 ATGGGGGGTGCAGCCCTTACCCGAGGTAAGTGCATGTGTTACCAAGCCGACTAACGATGGG
 AGCTTATCAATCGTCATTGGACTAGAAGTGAACAAATGAAGCTCTTACAGTAACTTGTGTATGCAA
 TCTGAAATGTATGTACTTTAATCAATAATAATCAGTAAACTCTTTTTCTGATTGATCGAGTTT
 AAATAAATTTGAAGTG

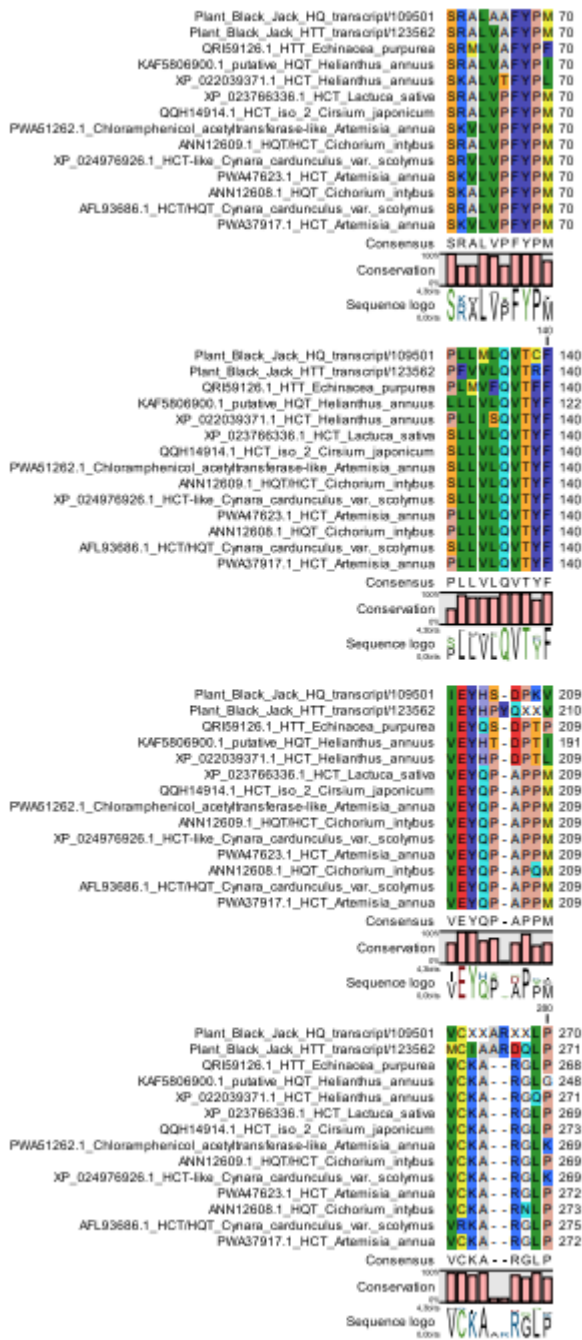
>|c|ORF2

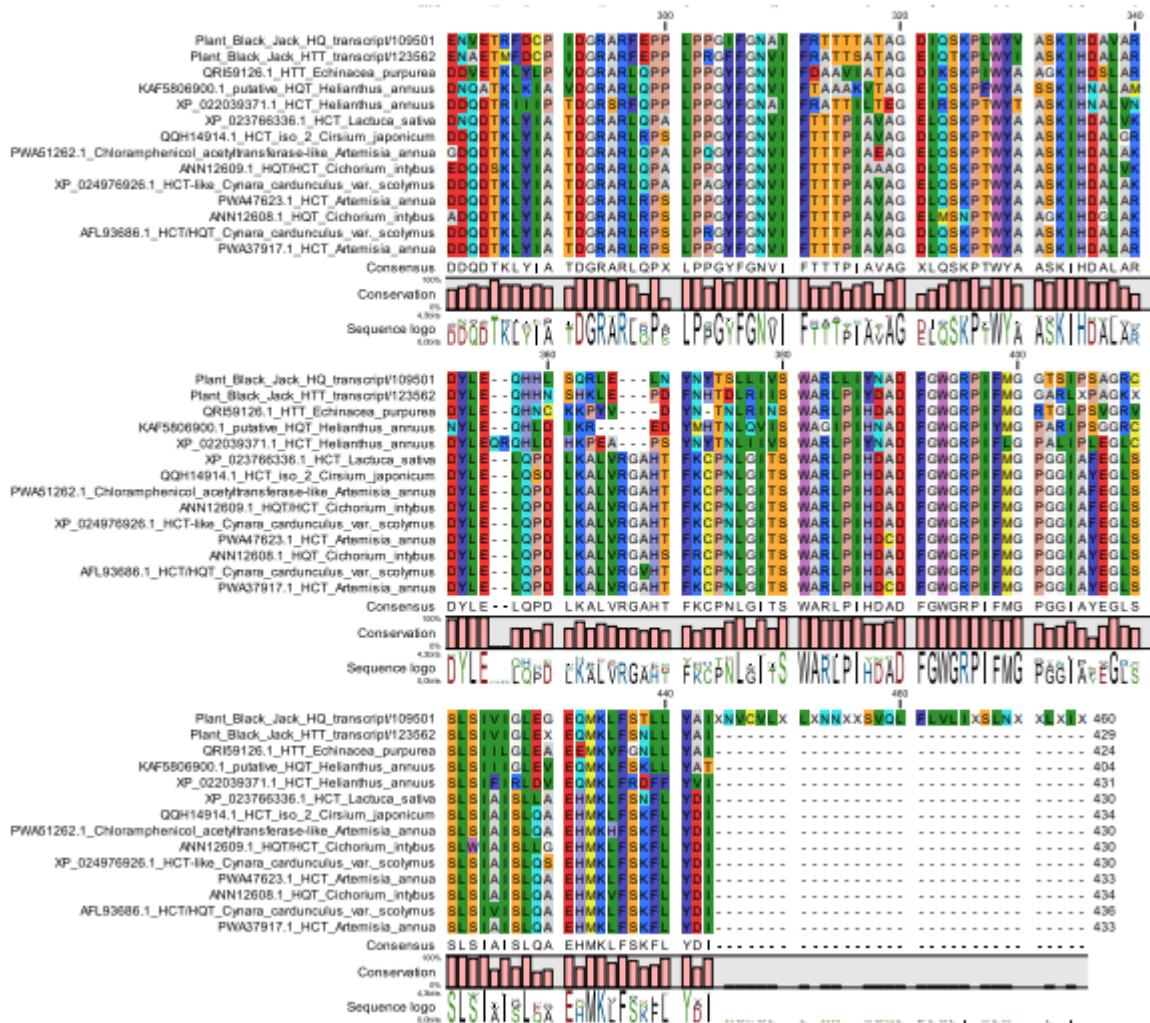
MKVVVKESKMLKPYEETPTTKLWSSLDLLALNRYTLTVYFYRPN GAPNFFDTKVMTDAMSRLVAF
 YPMAGRFRHGENRVEIDCQGGVLFLEAESDGVIDDGDFAPTPEYLKLVVVDHSLGIESIPFVVL
 QVTRFKCGGVS LGVGNHHRVADGMSGLHFINTWSDMARGLDITIQPFIDRTLLRPRDPPRPVFEHIEY
 HPYQTVKVPDVTKTAYSKFKFRDQLDMLKEKSKEDGNTISYSSFEILSAHIWKCMCIAARDQLPEN
 AETMFDPCPIDGRFVPEPLPRGFFGNVIFRATTSATAGEIQTKPLWYIASKIHDALARMNSDYLRSSLD
 YLEQHNSHKLEPDFNHTDLRIISWARLPIYDADFGWGRPIFMGGARLTRR



Figure S2: HTT2 nucleotide sequence from *B. pilosa* as obtained from the SMRT sequencing technique and the amino acid sequence of the ORF.







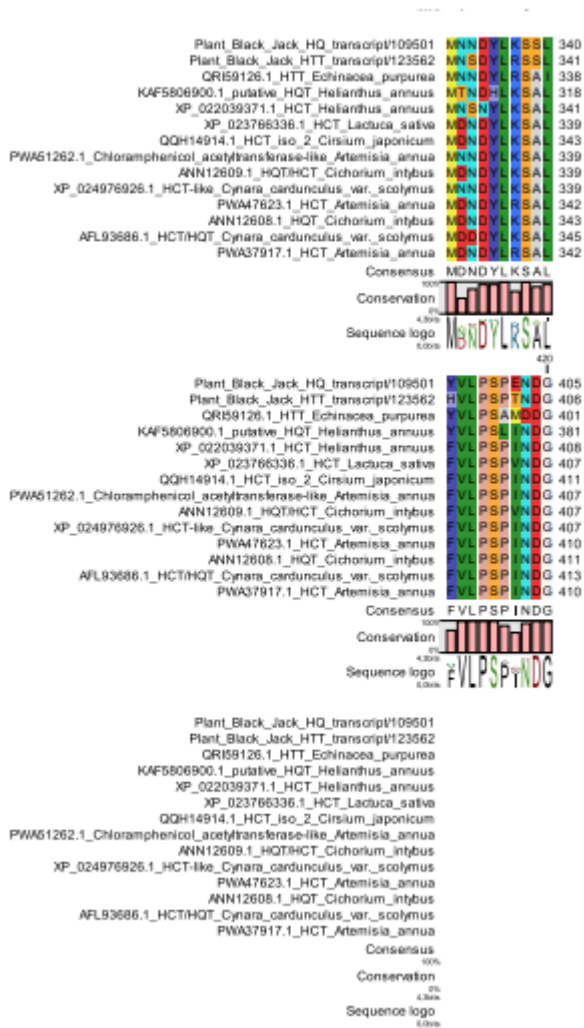


Figure S3: Multiple sequence alignment of *B. pilosa* HTT genes.

Multiple sequence alignment of *B. pilosa* HTT with its homologues from *Helianthus annuus*, *Echinacea purpurea* etc. Residues are grouped according to colours, for instance same colour represent similar residues across all genes from different plants. The alignment was generated using MUSCLE. The position of the residue is shown by the number on the right.

Sequence name	comme	similar	homolo	homolog 2
Plant_Black_Jack_HQ_transcript/105287	GGGAGGCTCGATTGCTCTTCGT	HST (MT936803)		
Plant_Black_Jack_HQ_transcript/109501	GATATAAACCCAATATAAAAAC	HST (MT936803)		
Plant_Black_Jack_HQ_transcript/119464	GATTCGT	Truncated/sequence	HST (MT936803)	
Plant_Black_Jack_HQ_transcript/123562	GTTAAGTAGAGAAGATAAGATG	HST (MT936803)		
Plant_Black_Jack_HQ_transcript/155837	GTCATTC	Truncated/sequence	HST (MT936803)	

Figure S4. Snapshot of five identified potential HTT gene sequences of *B. pilosa* during the filtering process using custom made filters on excel.

1	MT936803,1	100	78,77	77,59	78,94	78,91	70,63
2	Plant_Black_Jack_HQ_transcript/109501	78,77	100	84,75	86,08	75,19	71,94
3	Plant_Black_Jack_HQ_transcript/119464	77,59	84,75	100	96,59	74,98	81,11
4	Plant_Black_Jack_HQ_transcript/123562	78,94	86,08	96,59	100	77,1	82,87
5	Plant_Black_Jack_HQ_transcript/105287	78,91	75,19	74,98	77,1	100	67,87
6	Plant_Black_Jack_HQ_transcript/155837	70,63	71,94	81,11	82,87	67,87	100

Figure S5. This figure is showing a percentage matrix index of acyltransferases (HTT) from *B. pilosa*, the colours herein represent the following: Red = >95% similarity

Chapter 4

General conclusion

4. Conclusion

Single molecule real time sequencing approach allowed identification of acyltransferases in *B. pilosa*. Series of bioinformatics analysis revealed that *B. pilosa* has 312 acyltransferase genes that play a role in the phenylpropanoid pathway. Out of the 312 genes identified, 3 HQT and 1 HCT genes were characterised. Furthermore, HTT genes were also identified and characterised. These newly described genes were found to have overwhelming homology to the already published and authenticated gene sequences from plants within the *Asteraceae* family. Moreover, these genes were also found to contain sequences encoding highly conserved motifs synonymous with BADH acyl transferases. However, the HXXXD motif identified herein found to have a single amino acid difference from one gene to another. The percentage similarity index obtained herein further validates the SMRT sequencing approach as a feasible method for establishing sequences of genes of which the reference template is unavailable. Analysis of the methanol extracts of *B. pilosa* through LC-MS indicated a very complex pool of metabolites, from mono-, di-, tri- acyls of chlorogenic acids. Through LC-MS, tartaric acid esters of hydroxy cinnamic acids were also part of *B. pilosa*'s metabolome pool. This study also shows that *B. pilosa* is an alternative source of these highly sought-after chlorogenic acids compounds since it is easy to grow and at no cost. Future studies should aim to clone these transcripts in plant systems that do not produce CGAs in attempt to enhance their nutraceutical attributes.